

# Computational Systems Pharmacology and Toxicology

## Issues in Toxicology

### *Series editors:*

Diana Anderson, *University of Bradford, UK*

Michael D. Waters, *Michael Waters Consulting, USA*

Timothy C. Marrs, *Edentox Associates, UK*

### *Editorial advisor:*

Alok Dhawan, *CSIR-Indian Institute of Toxicology Research, Lucknow, India*

### *Titles in the Series:*

- 1: Hair in Toxicology: An Important Bio-Monitor
- 2: Male-mediated Developmental Toxicity
- 3: Cytochrome P450: Role in the Metabolism and Toxicity of Drugs and other Xenobiotics
- 4: Bile Acids: Toxicology and Bioactivity
- 5: The Comet Assay in Toxicology
- 6: Silver in Healthcare
- 7: *In Silico* Toxicology: Principles and Applications
- 8: Environmental Cardiology
- 9: Biomarkers and Human Biomonitoring, Volume 1: Ongoing Programs and Exposures
- 10: Biomarkers and Human Biomonitoring, Volume 2: Selected Biomarkers of Current Interest
- 11: Hormone-Disruptive Chemical Contaminants in Food
- 12: Mammalian Toxicology of Insecticides
- 13: The Cellular Response to the Genotoxic Insult: The Question of Threshold for Genotoxic Carcinogens
- 14: Toxicological Effects of Veterinary Medicinal Products in Humans: Volume 1
- 15: Toxicological Effects of Veterinary Medicinal Products in Humans: Volume 2
- 16: Aging and Vulnerability to Environmental Chemicals: Age-related Disorders and their Origins in Environmental Exposures
- 17: Chemical Toxicity Prediction: Category Formation and Read-Across
- 18: The Carcinogenicity of Metals: Human Risk Through Occupational and Environmental Exposure
- 19: Reducing, Refining and Replacing the Use of Animals in Toxicity Testing
- 20: Advances in Dermatological Sciences
- 21: Metabolic Profiling: Disease and Xenobiotics
- 22: Manganese in Health and Disease
- 23: Toxicology, Survival and Health Hazards of Combustion Products
- 24: Masked Mycotoxins in Food: Formation, Occurrence and Toxicological Relevance
- 25: Aerobiology: The Toxicology of Airborne Pathogens and Toxins

- 26: Chemical Warfare Toxicology, Volume 1: Fundamental Aspects
- 27: Chemical Warfare Toxicology, Volume 2: Management of Poisoning
- 28: Toxicogenomics in Predictive Carcinogenicity
- 29: Human Stem Cell Toxicology
- 30: The Comet Assay in Toxicology, 2nd edition
- 31: Computational Systems Pharmacology and Toxicology

*How to obtain future titles on publication:*

A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

*For further information please contact:*

Book Sales Department, Royal Society of Chemistry, Thomas Graham House,  
Science Park, Milton Road, Cambridge, CB4 0WF, UK

Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247

Email: [booksales@rsc.org](mailto:booksales@rsc.org)

Visit our website at [www.rsc.org/books](http://www.rsc.org/books)





# *Computational Systems Pharmacology and Toxicology*

Edited by

**Dale E. Johnson**

*University of Michigan, USA*

*Email: daleej@umich.edu*

**Rudy J. Richardson**

*University of Michigan, USA*

*Email: rjrich@umich.edu*



Issues in Toxicology No. 31

Print ISBN: 978-1-78262-332-8

PDF eISBN: 978-1-78262-373-1

EPUB eISBN: 978-1-78801-120-4

ISSN: 1757-7179

A catalogue record for this book is available from the British Library

© The Royal Society of Chemistry 2017

*All rights reserved*

*Apart from fair dealing for the purposes of research for non-commercial purposes or for private study, criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulations 2003, this publication may not be reproduced, stored or transmitted, in any form or by any means, without the prior permission in writing of The Royal Society of Chemistry or the copyright owner, or in the case of reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the licences issued by the appropriate Reproduction Rights Organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to The Royal Society of Chemistry at the address printed on this page.*

*Whilst this material has been produced with all due care, The Royal Society of Chemistry cannot be held responsible or liable for its accuracy and completeness, nor for any consequences arising from any errors or the use of the information contained in this publication. The publication of advertisements does not constitute any endorsement by The Royal Society of Chemistry or Authors of any products advertised. The views and opinions advanced by contributors do not necessarily reflect those of The Royal Society of Chemistry which shall not be liable for any resulting loss or damage arising as a result of reliance upon this material.*

The Royal Society of Chemistry is a charity, registered in England and Wales, Number 207890, and a company incorporated in England by Royal Charter (Registered No. RC000524), registered office: Burlington House, Piccadilly, London W1J 0BA, UK, Telephone: +44 (0) 207 4378 6556.

For further information see our web site at [www.rsc.org](http://www.rsc.org)

Printed in the United Kingdom by CPI Group (UK) Ltd, Croydon, CR0 4YY, UK

# *Preface*

Our motivation for bringing about a book on systems computational pharmacology and toxicology was a natural development from teaching courses on these subjects, first at the University of California in Berkeley and later at the University of Michigan in Ann Arbor. Our courses and this book address a critical need to modernize pharmacology and toxicology—to transform these fields from descriptive disciplines to predictive sciences. This transformation is necessary, because classic descriptive approaches are far too inefficient and expensive to assess the medical efficacy or toxicity of the many thousands of synthetic chemicals or natural products to which humans and other species are or will be exposed.

Not long ago, the approaches set forth in this book were either not possible or performed by specialists using such tools as quantum mechanics and mainframe computers. Now, because of rapid advances in technology, software, and theory, coupled with the public availability of large chemical and biomedical data sets through the internet, it is possible for non-specialist bench scientists to undertake sophisticated molecular modeling, bioinformatics, cheminformatics, and systems biology procedures on desktop computers as well as mobile devices, including to some extent electronic tablets and smart phones. Thus, powerful computational tools have become highly accessible, but knowing how and when to use the right tools in the right way can be a daunting task. This book seeks to make the job easier to understand and implement.

Recognizing that we now have the capability to understand pharmacological and toxicological effects at multiple biological levels, our book highlights the process of integrating the elements of complex phenomena into a systems approach. Thus, whereas inverse docking and pharmacophore mapping can identify molecular targets of candidate drugs or toxicants, intelligent mining of databases can identify networks of genes and proteins involved

in the system-wide biological responses to chemicals. A pharmacological or toxicological effect may begin with atomic-level binding, but ultimately the intact organism responds in a holistic manner. It is necessary to continually readjust our focus by many orders of magnitude to encompass the spectrum from molecular orbitals to human populations.

The tools and models discussed in the book hold tremendous promise for advancing applied and basic science, streamlining drug efficacy and safety testing, and increasing the efficiency and effectiveness of risk assessment for environmental chemicals. The content of chapters is designed to provide readers with an understanding of the basic principles and current methods of computational pharmacology and toxicology. These principles and approaches are discussed in several chapters in order to show how to connect chemicals with diseases and associated genes, and how to create pharmacology/toxicology connectivity maps or networks.

Vital to these expositions of principles and methods are illustrations of modeling and/or predicting potential pharmacological or toxicological effects from multiple properties. These characteristics include chemical structure, inference from similar compounds, *in silico* target identification, exposure, bioaccumulation, environmental persistence, biomarkers, and networks of biological pathways affected by a chemical.

Systems toxicology approaches used in the safer design of chemicals and identification of safer alternatives, which are major parts of global green chemistry initiatives, are also discussed, along with the concept of the adverse outcome pathway and modeling approaches for hazard identification and risk assessments for large numbers of environmental chemicals for which supporting data are sparse.

The book also expands the conventional boundaries of research and development of pharmaceutical agents. Thus, traditional Chinese medicines that include recipes containing several pharmacologically active phytochemicals are becoming role models of polypharmacy research.

The final chapter describes an inquiry-based computational toxicology course. Students work in small cooperative groups and are given tools, data, and basic concepts to solve toxicity-related environmental, public health, and/or disease-oriented problems in novel ways. Several case studies serve both to educate the reader and to provide material for teaching.

As co-editors, we are each involved in research and education on the topics covered in the book. We have authored or co-authored several of the chapters ourselves, and the other chapters have been written by experts recruited from around the world.

Dale E. Johnson and Rudy J. Richardson

# Contents

<b>Chapter 1</b>	<b>Systems Biology Approaches in Pharmacology and Toxicology</b>	<b>1</b>
	<i>Dale E. Johnson</i>	
1.1	Introduction	1
1.2	Systems Toxicology	2
1.3	Chemical Toxicities	3
1.3.1	Single-Target Toxicity Concepts	3
1.3.2	Toxicological Profiling for Potential Adverse Reactions	5
1.3.3	Toxicological Concepts for Safer Chemical Design	6
1.3.4	Biomarkers	8
1.4	Environmental Toxicology	9
1.4.1	Adverse Outcome Pathway	9
1.4.2	Expanding Exposure Concepts	10
1.4.3	Exposome	11
1.5	Systems and Network Pharmacology	12
1.5.1	Secondary Pharmacology and Off-Target Effects	13
1.5.2	Prediction of Potential Adverse Effects	14
1.6	Conclusions	14
	References	14

<b>Chapter 2</b>	<b>Databases Facilitating Systems Biology Approaches in Toxicology</b>	<b>19</b>
	<i>Dale E. Johnson and Ann M. H. Heslin</i>	
2.1	Introduction	19
2.2	Categorized Lists of Databases for Systems Toxicology	21
2.2.1	TOXNET Databases (Including Those with Direct Links from TOXNET)	21
2.2.2	US EPA Chemical Toxicity Databases	24
2.2.3	National Toxicology Program Databases	24
2.2.4	Additional Toxicity Databases	25
2.2.5	Chemical–Gene–Protein Databases	26
2.2.6	Pathway–Network Databases	27
2.2.7	Chemistry, Structural Alert, and QSAR Databases and Tools	28
2.2.8	Drug and Drug Target Databases	30
2.3	Websites with Extensive Links to Databases and Tools	31
2.4	Conclusions	31
	References	32
<b>Chapter 3</b>	<b>Tools for Green Molecular Design to Reduce Toxicological Risk</b>	<b>36</b>
	<i>David Faulkner, Leah K. Rubin Shen, Vanessa Y. De La Rosa, Dale E. Johnson, Rachel Hemingway, Richard V. Williams, Philip N. Judson, John Arnold and Chris D. Vulpe</i>	
3.1	Introduction	37
3.2	Physiochemical, Genotoxicity, and Blood–Brain Barrier Passage Properties of Chemicals	38
3.3	Tools for Green Molecular Design	39
3.3.1	Expert Systems	39
3.3.2	Decision Trees	44
3.3.3	QSAR Tools	44
3.3.4	Representative Tools	45
3.4	Case Study	50
3.5	The Design of Ideal Tools for Chemists	51
3.6	Conclusions	54
	References	54
<b>Chapter 4</b>	<b>Linking Environmental Exposure to Toxicity</b>	<b>60</b>
	<i>Noffisat Oki, Jeremy Leonard, Mark Nelms, Shannon Bell, Yu-Mei Tan, Lyle Burgoon and Stephen Edwards</i>	
4.1	Introduction	60
4.2	The AOP Framework: An Organizing Principle for Toxicological Data	64

<i>Contents</i>	xi
4.2.1 AOP Knowledge Management	67
4.2.2 Phases of AOP Development	68
4.2.3 Data Resources for AOP Development	70
4.3 Environmental Exposure and Pharmacokinetic Considerations for Adverse Outcome Development	73
4.4 The AEP Framework: An Organizing Principle for Exposure Data	75
4.4.1 Data resources for AEP development	79
4.5 AEP–AOP Integration for Linking Toxicity to Exposure: Applications of the AOP and AEP Frameworks for Risk Assessments and Chemical Management Decision Making	80
4.6 Conclusions and Future Directions	82
References	83
<b>Chapter 5 Linking Drug or Phytochemical Exposure to Toxicity</b>	<b>89</b>
<i>C. A. Rodríguez, N. S. Teuscher and J. A. Uchizono</i>	
5.1 Introduction	89
5.2 Pharmacokinetic and Toxicokinetic Models	91
5.2.1 Structural Models	91
5.2.2 Variance Models	98
5.3 PK/PD Relationships	102
5.3.1 Mathematical Description of Pharmacodynamic Effects	104
5.3.2 Combined PK/PD and TK/TD Modeling	109
5.3.3 Modeling Pharmacodynamics in the Absence of Pharmacokinetic Data: K-PD Models	110
5.4 Modeling Drug Interactions with Phytochemicals	111
5.4.1 Inhibition of Metabolism	112
5.4.2 Induction of Metabolism	113
5.4.3 Enhancement of Absorption	114
5.4.4 Inhibition of Absorption	115
5.4.5 Modeling of Pharmacodynamic Interactions	115
5.5 Conclusions	116
References	116
<b>Chapter 6 Chemical Similarity, Shape Matching and QSAR</b>	<b>120</b>
<i>E. V. Radchenko, G. F. Makhaeva, V. A. Palyulin and N. S. Zefirov</i>	
6.1 Introduction	120
6.2 Molecular Similarity, Chemical Spaces and Activity Landscapes	121
6.2.1 Molecular Similarity: Concept and Definitions	121
6.2.2 Chemical Spaces and Activity Landscapes	131
6.2.3 Applications of Molecular Similarity Analysis	138

6.3 Quantitative Structure–Activity/Property Relationships (QSAR/QSPR)	142
6.3.1 Congeneric Series and Consistent Mechanisms	143
6.3.2 Diverse Series and Big Data	148
6.4 Conclusion	153
Acknowledgements	154
References	154
<b>Chapter 7 <i>In silico</i> Chemical–Protein Docking and Molecular Dynamics</b>	<b>174</b>
<i>Sanjeeva J. Wijeyesakere and Rudy J. Richardson</i>	
7.1 Introduction	174
7.2 Molecular Docking: Overview and Applications	175
7.2.1 Genetic Algorithms	177
7.2.2 Monte Carlo Procedure	177
7.2.3 Matching Algorithms	177
7.3 Scoring Ligand Poses	178
7.4 Inverse Docking	179
7.5 Case Study: Using <i>In silico</i> Docking to Investigate Interactions of 1,3-Dinitrobenzene with Adenosine Deaminase	179
7.6 Case Study: Using <i>In silico</i> Docking to Assess Binding of Bisphenol-A to Estrogen-Related Receptor- $\gamma$	180
7.7 Molecular Dynamics	182
7.7.1 Running MD Simulations	182
7.7.2 Analysis of MD Trajectories	183
7.7.3 Case Study: Gaining Insights into the Conformational Dynamics of Human Neuropathy Target Esterase <i>via</i> MD Simulations of its Catalytic Domain Homologue Patatin-17 in Complex with Organophosphorous Compounds	185
References	187
<b>Chapter 8 Computational Tools for Chemical Toxicity Testing and Risk Assessment Under the Framework of Adverse Outcome Pathways</b>	<b>191</b>
<i>M. Mumtaz, P. Ruiz and Q. Zhang</i>	
8.1 Introduction	191
8.2 The AOP Concept	193
8.3 Quantitative Methods in Traditional Apical Endpoints Testing	194
8.4 PBPK Modeling and <i>In vitro</i> to <i>In vivo</i> Extrapolation	196



<i>Contents</i>	xiii
8.5 SAR Modeling	197
8.6 Computational Modeling of Toxicity Pathways	198
8.6.1 Concept of Toxicity Pathways	198
8.6.2 Purpose of Modeling Toxicity Pathways	199
8.6.3 How to Model Toxicity Pathways	200
8.6.4 Case Studies	203
8.6.5 Education on Computational Toxicology	205
8.6.6 Pathway Modeling Software Tools	205
Acknowledgements	206
References	206
<b>Chapter 9 <i>In silico</i> Toxicology: An Overview of Toxicity Databases, Prediction Methodologies, and Expert Review</b>	<b>209</b>
<i>D. Bower, K. P. Cross, S. Escher, G. J. Myatt and D. P. Quigley</i>	
9.1 Introduction	209
9.2 Toxicity Databases	215
9.2.1 Overview	215
9.2.2 Database Organization	216
9.2.3 Genetic Toxicity and Carcinogenicity	217
9.2.4 Reproductive and Developmental Toxicity	219
9.2.5 Acute and Repeated Dose Toxicity	220
9.3 <i>In silico</i> Methodologies	221
9.3.1 Overview	221
9.3.2 Expert Alerts	221
9.3.3 QSARs	223
9.3.4 Read-Across	226
9.4 Expert Reviews	227
9.4.1 Assessing Experimental Data	227
9.4.2 Drawing Conclusions from Multiple Systems	228
9.4.3 Reviews Accepting or Refuting An <i>In silico</i> Result	230
9.4.4 Documenting <i>In silico</i> Results	233
9.5 Conclusions	233
Acknowledgements	236
References	236
<b>Chapter 10 Data Sources for Herbal and Traditional Medicines</b>	<b>243</b>
<i>Hsueh-Fen Juan</i>	
10.1 Introduction	243
10.2 TCM Databases	244
10.2.1 Chem-TCM (Chemical Database of Traditional Chinese Medicine)	244
10.2.2 HIT (Linking Herbal Active Ingredients to Targets)	244

10.2.3	TCMSP (Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform)	246
10.2.4	TCMGeneDIT (A Database for Associated TCM, Gene and Disease Information Using Text Mining)	248
10.2.5	TCMID (Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis)	249
10.2.6	TTD (Therapeutic Target Database)	251
10.3	Omics Data in TCM	253
10.3.1	Genomics in TCM	253
10.3.2	Transcriptomics in TCM	254
10.3.3	Proteomics in TCM	256
10.3.4	Metabonomics in TCM	256
10.4	Summary	257
	Acknowledgements	257
	References	257

## **Chapter 11 Network Pharmacology Research Approaches for Chinese Herbal Medicines** 261

*Dale E. Johnson*

11.1	Introduction	261
11.1.1	Modernization of TCM	263
11.1.2	Concept of Network Pharmacology	263
11.2	Network Pharmacology in TCM Research	265
11.3	Network Pharmacology in the Understanding of Herb–Drug Interactions	268
11.4	Pharmacogenomics in TCM	269
11.5	TCMs in Clinical Trials	271
11.6	The Future of Network Pharmacology in Traditional Medicine	273
11.7	Conclusion	274
	References	274

## **Chapter 12 Chemical–Disease Category Linkage (CDCL): Computational Methods Linking Traditional Chinese Medicines and Western Therapeutics** 279

*Dale E. Johnson and Kit Wun Kathy Cheung*

12.1	Introduction	279
12.1.1	Databases for CDCL Information and Study	281
12.1.2	TCM Classifications	283
12.1.3	Active Ingredients in Herbs	286
12.2	Open Access Tools for CDCL Informatics	287

12.3 Computational CDCL Studies with Commercial Tools	289
12.4 Herb–Drug Interactions	292
12.4.1 Pharmacokinetic Interactions	292
12.4.2 Pharmacogenomic-Related Interactions	294
12.5 Combination Therapies and Future Directions	294
12.6 Conclusions	295
References	296

## **Chapter 13 Educational Programs for Computational Toxicology and Pharmacology 300**

*Dale E. Johnson and Rudy J. Richardson*

13.1 Introduction	300
13.2 Historical Context: Computational Toxicology	301
13.2.1 Background	301
13.2.2 Programs at University of California Berkeley and University of Michigan	302
13.3 Inquiry-Based Science Courses	303
13.4 Current Computational Toxicology Courses	304
13.4.1 Toxicology Tutorials	305
13.4.2 Course Concepts	305
13.4.3 Case Studies	306
13.5 Course Projects	310
13.5.1 Starting Projects	310
13.5.2 Typical Project Categories	310
13.5.3 Therapeutics vs. Environmental Chemicals	311
13.5.4 Challenges in Computational Toxicology	311
13.6 Sample Project: The Chemical of Concern Question	312
13.6.1 Health Effects Inquiry	313
13.6.2 Endpoints for Breast Cancer	313
13.6.3 Project Question	314
13.7 Course Projects Presented and Published	314
13.7.1 Projects Presented at National Scientific Meetings	314
13.7.2 Projects from the Courses Published in Journals	316
13.8 Computational Pharmacology as Part of the Principles of Drug Action	316
13.9 Conclusion	318
References	319

## **Subject Index 324**

## CHAPTER 1

# *Systems Biology Approaches in Pharmacology and Toxicology*

DALE E. JOHNSON<sup>\*a,b</sup>

<sup>a</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA;

<sup>b</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Morgan Hall, Berkeley, CA 94720-3104, USA

\*E-mail: [daleej@umich.edu](mailto:daleej@umich.edu)

## 1.1 Introduction

The science and practical field of toxicology has been changing dramatically over the last 15–20 years, transitioning into a more systems biology and network-based approach.<sup>1–4</sup> Several factors have been involved, including the developing genomics era where the understanding of genetic changes has enhanced the ability to understand diseases and chemically-induced toxicities at the molecular level. The genomics era has also ushered in “omics” technologies and approaches such as transcriptomics, metabolomics, proteomics, and epigenomics, which have changed the way we view mechanisms of toxicity and the perturbation of biological systems that lead to adverse outcomes.<sup>5</sup> These advances have been coupled with the public availability of large datasets of information and new modeling approaches that have enhanced the ability to understand toxicological events and effects at multiple biological levels.<sup>6</sup> Since our scientific approaches, inquiries, and visions aimed at understanding toxicological events and outcomes have

---

Issues in Toxicology No. 31

Computational Systems Pharmacology and Toxicology

Edited by Dale E. Johnson and Rudy J. Richardson

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

been broadened tremendously, this reinforces our need for new and better ways to assess toxicity and risk. The large numbers of uncharacterized chemicals already present in the environment and new chemicals that continue to enter it has required hazard and risk assessments to be made with very few data. These factors have had a major influence on the need to accelerate new approaches and move away from an overdependence on *in vivo* animal testing and make better use of computational, molecular, and *in vitro* tools.<sup>6,7</sup> The identification of the majority of toxic events in *in vivo* animal toxicology studies rely on high-dose exposure to the animals and default linear extrapolation procedures,<sup>8</sup> with the incorporation of newer technologies absent in the vast majority of animal studies. This has been considered a shortcoming in risk assessment and several weaknesses in this process include the comparative shape of the dose–response relationship after relevant levels of human exposure, whether biological and/or toxicological thresholds do in fact exist and for what toxicological endpoints, and potential population variability in response.<sup>5</sup>

## 1.2 Systems Toxicology

Accordingly, research in toxicology has moved into a new systems-oriented phase called systems toxicology, which involves the study of complex molecular response networks initiated by exposure (both intentional and unintentional) to chemical substances. At the heart of systems toxicology approaches are the development and usage of quantitative mechanistic models that create a predictive toxicology aspect relevant to all toxicology fields, including drug research and development and environmental research. The overall approach involves the integration of classical toxicology with the quantitative analysis of large networks of chemically-induced molecular and functional changes, which occur across multiple levels of biological organization.<sup>5</sup> Examples of key influential events in this transition since the year 2000 include the release of human genome sequencing data including specific signal transduction domains, the development and issuance of the report *Toxicity Testing in the Twenty-first Century* by the National Research Council (NRC),<sup>9</sup> which has influenced all sectors of the toxicology field, and the development and publication of the adverse outcome pathway (AOP) approach,<sup>6,10,11</sup> which has highlighted the realities that exist as the science moves away from an overdependence on *in vivo* testing and makes greater use of computational, molecular, and focused *in vitro* tools. Additional drivers of change include the European Union (EU) report from the Scientific Committee on Health and Environmental Risks, the EU's Registration, Evaluation, Authorisation and Restriction of Chemical Substances (REACH) program, and the International Programme on Chemical Safety (IPCS).<sup>7,12</sup> The paradigm shift can also be seen in the drug research and development sector, but rather than focusing on drugs during late stages of development or on marketed drugs, the systems-related efforts are positioned at the front end of research, both on safer chemical design and extensive target research.

While the drug industry is required to conduct animal toxicology studies by regulatory agencies and international guidelines, the major effort underway is to determine chemical liabilities early in the drug discovery pipeline, both to reduce the time and cost of failures later in the process, but also to avoid costly failures once a drug reaches the market.<sup>5</sup> Currently, there is an International Consortium for Innovation and Quality in Pharmaceutical Development (IQ), where several pharmaceutical and biotechnology companies have created a Nonclinical to Clinical Translational Database (WG1) to allow analysis of the reliability and potential limitations of nonclinical data in predicting clinical outcomes, including the evaluation of conventional biomarkers of toxicity.<sup>13</sup> Current screening approaches applied to the front end of drug research are described below.

## 1.3 Chemical Toxicities

### 1.3.1 Single-Target Toxicity Concepts

The science and practice of toxicology over the past several decades have consistently used classic toxicological approaches, such as *in vivo* and *in vitro* toxicology studies, combined with predictive toxicological methodologies. The desired endpoints of the *in vivo* animal research efforts have been the determination of a toxic dose where a chemical could be shown to induce pathologic effects after a specified duration of treatment or exposure. Where appropriate, these studies have included the estimate of the lowest observed adverse effect level, the no observed adverse effect level, and the maximally tolerated dose (MTD).<sup>5,14</sup> These adverse effect level estimates are traditionally used in drug research and development to predict the first dose in humans and to predict margins of safety estimates based on delivered dose and/or internal exposure from pharmacokinetic/pharmacodynamic (PK/PD) modeling with extrapolations into clinical trial subjects. By regulatory requirements, all potential drugs undergoing research and development will undergo both *in vitro* and *in vivo* studies, and, if the compound reaches the clinical trial stage successfully, data from human exposure to judge the adequacy of nonclinical data in predicting clinical outcomes. Uncertainties in these estimates include the definition of adverse, which is specific for each organ system in each study and typically determined by the study pathologist; the accuracy of cross-species extrapolations (particularly rodent-to-human); and the true definition of risk–benefit for each individual drug. However, the generation of classical toxicology data does not assure the accurate prediction of potential human toxicity. Sundqvist and colleagues<sup>15</sup> have reported on a human dose prediction process, supplemented by case studies, to integrate uncertainties into simplified plots for quantification. Drug safety is recognized as one of the primary causes of attrition during the clinical phases of development; however, in numerous instances the actual determination of serious adverse effects only occurs after the drug reaches the market.

In the United States, ~2 million patients are affected with drug-mediated adverse effects per year, of which ~5% are fatal.<sup>16</sup> This places drug toxicity as one of the top five causes of death in the United States, and the costs to the health care system worldwide are estimated at US\$40–50 billion per year.<sup>16</sup> In drug development there are always risk–benefit considerations, which will weigh any potential toxicity against the benefit expected to be gained by a patient taking the drug. An example of the uncertainty of these estimates can be seen in the methods used for carcinogenicity testing and evaluation for drug approval. The design of these studies rely on high-dose exposure to animals and default linear extrapolation procedures, while little consideration is given to many of the new advances in the toxicological sciences.<sup>17</sup> Carcinogenicity studies are typically 2-year studies in rodents conducted with three dosage groups (low, mid, and high dose) and one or two concurrent control groups. Dose levels are established from previous studies, such as 13-week toxicity studies, where a MTD has been estimated. Each group in the carcinogenicity study has 60–70 animals of each sex, and the analysis of whether there is a potential carcinogenicity concern is based on an analysis of each tumor in each tissue or organ system individually by gender; certain tumors are combined *via* standardized procedures for statistical analysis. The analysis uses the historical database from the laboratory where the studies are conducted to determine whether each tumor is considered common or rare, using the background incidence of 1% as the standard. Common tumors are those with a background incidence of 1% or over and rare tumors are those with a background incidence below 1%. In the statistical analysis, *p*-values for rare and common tumors are evaluated for pair-wise significance at 0.05 (for rare) and 0.01 (for common). The rare *vs.* common tumor classification is an arbitrary tumor threshold and adjustments to the specific classifications by individual tumor, which can occur from laboratory to laboratory and *via* analyses of different control groups, can have consequences in the overall tumor evaluation outcome.<sup>8</sup> Applying a “weight of evidence” approach into the evaluation procedures, particularly during regulatory review, attempts to alleviate some of the uncertainties; however, after more than 50 years of on-going experience, these studies still fail to bring the 21st century mindset to carcinogenicity testing. The classic toxicological process for drug development assumes that a chemical interacts with a higher affinity to a single macromolecule (the toxicological target), and therefore a single biological pathway may be perturbed at the initial target modulation. This would be followed by downstream activation of secondary and possibly tertiary pathways that result in the tissue or organ effect as indicated by key biomarkers.<sup>2</sup> In this concept, the magnitude of toxicological effects are related to the concentration of altered molecular targets (at the site of interest), which in turn is related to the concentration of the active form of the chemical (parent compound or metabolite) at the site where the molecular targets are located. Also included in this concept is the unique susceptibility of the organism exposed to the compound.

### 1.3.2 Toxicological Profiling for Potential Adverse Reactions

Predictive toxicology efforts in drug research and development involve the use of multiple sources of legacy data including data generated by chemical and pharmaceutical companies and data submitted to regulatory agencies. These efforts have led to the “data warehouse” model which includes data generated through high throughput and targeted screening, and *in vitro* and *in vivo* toxicology studies on thousands of compounds and structural analogues. In a majority of cases these data also include findings from clinical trials where an experimental drug was tested on humans.

The information is applied in a “backward” fashion to predict potential findings where data do not yet exist or where decisions are being made on new potential drug candidates. Bowes and colleagues<sup>18</sup> have described a pharmacological profiling effort by four large pharmaceutical companies: AstraZeneca, GlaxoSmithKline, Novartis, and Pfizer. The companies suggest that ~75% of adverse drug reactions can be predicted by studying pharmacological profiles of candidate drugs. The pharmacological screening identifies primary effects related to the intended action of the candidate drug, whereas identification of secondary effects due to interactions with targets other than the primary (intended) target could be related to off-target adverse events. The groups have identified 44 screening targets including 24 G-protein coupled receptors, eight ion channels, six intracellular enzymes, three neurotransmitter transporters, two nuclear receptors, and one kinase. These types of screening data are used in the data warehouse model, typically configured in a proprietary fashion within each company. Other collaborative efforts have been developed and data from these sources would also be incorporated.

Blomme and Will<sup>19</sup> have reviewed the current and past efforts by the pharmaceutical industry to optimize safety into molecules at the earliest stage of drug research. They conclude that new and emerging technologies in the past two decades have had limited impact on nonclinical attrition rates associated with safety issues. In addition, they point out that front-loading series of toxicology assays to “kill early, kill often” have been challenged due to high false-positive rates and an overall low positive predictive value. The primary issue cited is the lack of information on an efficacious exposures (PK/PD) and the fact that the assays are more likely to represent hazard identification and not risk assessment. Therefore, it is suggested that these data be used as alerts rather than discontinuance criteria. In a more systems toxicology approach, a large effort is now being directed towards understanding the extent of pharmacological modulation of both preceded and unpreceded targets in relation to potential safety liabilities and developing technologies to determine achievable therapeutic windows. Blomme and Will<sup>19</sup> discuss efforts at AbbVie and Pfizer where target safety assessments are explored. The assessments include the biology of the target, tissue expression maps, messenger RNA and proteins, human genetic data, phenotypes from genetically engineered animal models, historical data from on-going



and past clinical trials targeting similar targets and associated pathways, extensive datamining *via* biomedical databases, and *in silico* simulation of the various consequences of target modulation. The majority of systems research in drug safety use specific toxicities as the starting point. Ivanov and colleagues<sup>20</sup> discuss the use of specific methods to counteract ventricular tachyarrhythmia (VT). While the *in vitro* HERG potassium channel assay is used universally as a predictor of VT, this is only one mechanism of action and other targets must also be identified and explored. These researchers have used the following approach: (1) creation of VT positive and negative compound libraries; (2) *in silico* prediction of extensive drug–target interaction profiles of chemical libraries identifying potential VT-related targets; (3) gene ontology and pathway enrichment on these potential VT targets to elucidate potential biological processes; (4) creation of a cardiomyocyte regulatory network based on general and heart-specific signaling and regulatory pathways; and (5) simulation of the changes in the regulatory network caused by the inhibition at each node in order to define potential VT-related targets. These are the type of studies that lead to more refined *in vitro* and *in silico* assessments of potential drug adverse effects at the early stage of drug research.

Verbist and colleagues<sup>21</sup> have outlined another type of systems toxicology proposal at Janssen involving QSTAR (quantitative structure-transcription-assay relationships) by integrating high-throughput gene expression profiling data; chemical information, particularly detailed analogue analysis; and bioassay data. Using several compounds from a single chemical scaffold targeting PDE10A, a target of pharmacological interest at Janssen, changes in tubulin gene expression were identified in a subset of compounds. Therefore a screening process was developed involving multiple cell lines, gene expression profiling, *in vitro* micronucleus assays, and high-content imaging to show microtubule aggregates as compared to other phenotypes. Besides the chemical series of interest, known positive and negative compounds were included in the process. This study presents a valuable proof-of-concept of how to link and potentially improve the risk assessment in early drug discovery using several technologies in a drug research systems toxicology approach.

### 1.3.3 Toxicological Concepts for Safer Chemical Design

Vouthkova and colleagues<sup>22</sup> have outlined an extensive framework for safer chemical design using multiple data and modeling resources. These types of data generation and modeling approaches are the basis of the process for a green chemistry model for specific series of chemicals used or proposed for use as reagents, solvents, or chemical intermediates in chemical synthesis. The simplified scheme involves a model building process, where chemical structures of interest are evaluated for chemical motifs (structural alerts) known to be associated with human health or environmental hazards, chemicals are clustered into hazard categories and specific high- or higher-throughput

and targeted assays are identified for each hazard category. Analogue series directly applicable to the chemistry under evaluation are prepared or obtained and screened in the relevant assays. With homologous or structurally similar series, local chemical-toxicity models can be developed, validated and incorporated into the initial computational screening process. This general method can be applied to any specific hazard, any series of chemicals, and any assay methodology. Examples of hazard categories include carcinogenicity; reproductive and developmental; mutagenicity; neurotoxicity; endocrine disruption; cardiovascular; dermatotoxicity; digestive system toxicity; hematotoxicity; hepatotoxicity; immunotoxicity; muscular toxicity; nephrotoxicity; ocular toxicity; ototoxicity; respiratory toxicity; persistence in the environment; bioaccumulative in the environment; toxic to water organisms; water contaminant; and air pollutant. Structural motif alert and expert predictions can be achieved using OpenTox,<sup>23</sup> an open access system used to predict potential hazards from chemical structures and known chemical motifs associated with human health and environmental endpoints, and Derek from Lhasa,<sup>24</sup> a rule-based expert system that de-convolutes a chemical structure into sub-structural fragments and addresses potential toxicity consistent with the above hazard categories. The software is also used to create specific local expert predictions from screening data. Meteor (Lhasa) predicts potential metabolites and the metabolite structures can be used in Derek predictions. This type of inquiry is highly useful for establishing basic information for rank-ordering compounds, as in early candidate selection, and in the process of safer chemical synthesis. Multiple screening approaches have been used for evaluation of chemical toxicity using high-throughput technology and multiple assays. These include the United States Environmental Protection Agency (EPA) ToxCast program<sup>25</sup> where over 2000 chemicals have been evaluated in over 700 high-throughput assays. This is a section of the Tox21 testing program, a collaboration among EPA, the National Institutes of Health (NIH), including the National Center for Advancing Translational Sciences at the National Toxicology Program at the National Institute of Environmental Health Sciences, and the United States Food and Drug Administration. The Tox21 program involves high-throughput screening of more than 10 000 environmental chemicals and approved drugs using more than 100 assays. All data are publicly available, as discussed later. Wink and colleagues<sup>26</sup> discuss a quantitative high-content imaging *in vitro* process to elucidate chemical interactions with cellular adaptive stress response pathways to gain a better insight into chemical toxicities at a phenotypic cellular level. The key to their reported technology is a panel of reporter cell lines to monitor multiple key nodes of the adaptive stress response pathways. Examples include cellular redox homeostasis, unfolded protein response, endoplasmic reticulum damage, inflammatory signaling, and DNA damage response. These assays hold the potential to be incorporated into multiple large-scale screens to evaluate health-related chemically-induced biological phenomena in drug research as well as hazard identification.

### 1.3.4 Biomarkers

Biomarkers are typically used to define the onset, continuation, and either positive or negative characteristics of the induced biological effects of the drug (chemical) under research. Biomarkers have been classified as biomarkers of exposure, susceptibility, and outcome. The definition of biomarker as used in drug discovery and development is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic response(s) to a therapeutic intervention.<sup>27</sup> In pharmacological studies, where a relevant therapeutic target is identified and pursued, biomarkers are developed that correlate with the proof of concept for the drug candidate. Biomarkers are developed to show (1) that a desired modulation of the target occurs as anticipated by the chemical therapeutic; (2) that the chemical-induced target modulation produces a desired biological effect; (3) that the induced biological effect alters the disease under study; and (4) that there may be increased susceptibility to the therapeutic candidate by certain individuals, such as those based on pharmacogenetic predispositions. In toxicology studies, biomarkers are objectively measured and evaluated as indicators of (1) normal biological processes; (2) pathogenic processes; (3) pharmacologic response(s) to a therapeutic intervention, which in some cases could mean excessive or nonspecific pharmacologic activity; and (4) exposure–response relationships. Pharmacogenetic markers are also studied from a toxicological standpoint, particularly in relation to drug metabolism. In environmental research and risk assessment, biomarkers are frequently referred to as indicators of human or environmental hazards. Discovering and implementing new biomarkers for toxicity caused by exposure to a chemical from a therapeutic intervention or in some cases through unintentional exposure continues to be pursued through the use of animal models to predict potential human effects, from human studies (clinical or epidemiological) or from biobanked human tissue samples, or the combination of these approaches.<sup>27</sup> In addition, several omics technologies such as transcriptomics, metabolomics, and proteomics have added an important aspect to biomarker research.<sup>12</sup> More recently, epigenomics, which is the study of changes in gene activity not attributed to DNA sequence alterations, has been shown to have increased importance in disease causality research.<sup>12</sup> These technologies and data produced, along with large datasets of high-throughput screening data, essentially changed the process of defining biomarkers. The process of discovering or inferring biomarkers through computational means involves the identification or prediction of the molecular target(s) of the chemical, which in many cases can be secondary or undesirable targets and the association of these targets with perturbed biological pathways. The integration of these approaches with the quantitative analysis of chemically induced molecular and functional changes has brought to fruition the goals originally outlined in the 2007 NRC report.<sup>9</sup>

## 1.4 Environmental Toxicology

The concepts of environmental chemical toxicity are different to the concepts of drug toxicity, as with environmental exposures there are only risk considerations, with virtually no benefit associated with chemical exposure. Currently there are more than 80 000 chemicals on the market and/or reach the environment, and ~2000 new chemicals are introduced each year. Unlike drugs, for the majority of these compounds, there is limited or inadequate toxicological information with which to make rational evaluations of risk.<sup>2</sup> Where information exists or is associated from similar structures of analogous compounds, a risk assessor may arrive at various hazard reference values, such as a derived no-effect level, to estimate an acceptable level of protection of human or wildlife health and the environment.<sup>5</sup> Depending on the context and urgency of the risk assessment, which could include the classification of a chemical in terms of hazard severity and risk management assessment, several assumptions must be made which could cause the level of uncertainty to increase.

### 1.4.1 Adverse Outcome Pathway

The concept of the AOP was a necessary enhancement to the *Toxicity Testing in the Twenty-first Century* report, made to more adequately support ecological risk assessment.<sup>6</sup> After the first publication in 2010, and subsequent publications by scientists at the EPA,<sup>10,11</sup> the AOP concept and developing case studies have become a primary force in the progression of the computational systems toxicology approach in environmental risk assessment. Unlike work in drug discovery and development, which always has a confidential business information component and therefore results are not fully publicly available, AOPs are developed in a fully open-access mode and are supported by publicly available databases updated by EPA scientists. The AOP concept highlights existing knowledge that links the direct molecular initiating event, of which in theory is the interaction or modulation of a molecular biomolecule or target with a xenobiotic, and an adverse outcome at a biological organization that spans multiple levels of biological organization including the following general examples from Ankley and colleagues.<sup>6</sup> These events are outlined below.

- (1) Macro-molecular interactions, such as receptor–ligand interactions including agonism and antagonism, DNA binding, and protein oxidation.
- (2) Cellular responses, such as gene activation, protein production, alterations in signaling, and protein depletion.
- (3) Organ responses, such as disrupted homeostasis, alterations in tissue development and/or function, and altered physiology.
- (4) Organism responses, such as mortality, impaired reproductive and developmental function, and development of diseases, such as cancer.

- (5) Population responses, such as alterations in the structure of a population and potential extinction of species within regional and global environments.

In defining the key aspects of an AOP, macro-molecular interactions (1) are considered the initiating event, which is called “anchor 1”, and the organ and population responses (4 and 5) are considered adverse outcomes at the organism or population level, collectively called “anchor 2”. The aspects of connecting the initiating events to outcomes can take various forms, depending on the chemical itself and the amount of information available, including *in vivo*, *in vitro*, and computational sources. These various linkages also help to define key assays and technologies to enhance information collection and usage for each individual AOP. Ankley and colleagues<sup>6</sup> provide five case studies that illustrate these points. Events occurring in the information flow between the molecular initiating event and adverse outcome are called intermediate events.<sup>28</sup> When an intermediate event represents a biological event that is necessary for an adverse outcome to occur and is quantitatively measurable, it is considered to be a key event. In a systems toxicology approach, key event relationships between adjacent molecular initiating events, key events and adverse outcomes help define alternative approaches to assess environmental hazards. In a signaling pathway, this could be upstream or downstream events that help define more suitable assays or test systems and provide a faster quantitative evaluation of potential adverse outcomes. One of the challenges in the AOP process is to define or estimate the exposure of the xenobiotic in the relevant species under consideration for risk assessment.

### 1.4.2 Expanding Exposure Concepts

As discussed earlier, measurement of exposure, toxicokinetics, systemically and importantly, at the critical site of action (anchor 1) is an essential piece of information. Unlike pharmaceutical compounds, it is highly unlikely that there would ever be controlled human toxicokinetics data for industrial and environmental chemicals.<sup>29</sup> Extrapolating toxicokinetic models from *in vitro* data has been used with pharmaceutical compounds, and this is termed *in vitro* to *in vivo* extrapolation methodology. These models have been used with some success for environmental and industrial chemicals using high-throughput toxicokinetics (HTTK) models. Several examples of HTTK methodology have been published, including lecture series by scientists from the United States EPA, the National Center for Computational Toxicology, and the Hamner Institutes for Health Sciences.<sup>29,30</sup> The primary methodology is termed “reverse dosimetry”, which uses concentrations that produce bioactivity in *in vitro* assays to estimate doses (mg kg<sup>-1</sup> per day) sufficient to produce steady-state plasma concentrations (C<sub>ss</sub>) in μM. These approaches assume 100% bioavailability and a linear relationship between C<sub>ss</sub> and dose. Another approach, called probabilistic reverse dosimetry approaches for estimating exposure distributions (PROCEED)<sup>30</sup> uses

biomarkers of exposure, such as blood and urine levels from biomonitoring studies, to model the most likely exposure concentrations—or intake doses or dose levels—experienced by the study participants that would result in the concentrations measured. These modeling procedures are considered a work in progress (as of early 2016); however, they represent a critical piece of the puzzle in chemical-associated environmental risk assessment. Certain drivers of the *in vivo* toxicokinetics process may be inaccurately estimated by *in vitro* assays such as extra-hepatic metabolism (particularly when using *in vitro* hepatic cell lines), secretion in bile and enterohepatic recirculation, absorption and bioavailability, and active transport in several tissues including renal and hepatic. In addition, a process to include metabolites into the high-throughput screening process for all chemicals will be a necessary part of the functional HTTK process.<sup>29</sup>

### 1.4.3 Exposome

The exposome is currently defined as the totality of all human environmental exposures (exogenous and endogenous) from conception to death.<sup>31</sup> The National Institute of Environmental Health Sciences<sup>32</sup> has developed a broad definition of environmental exposures, which includes chemical exposures, diet, physical activity, stress, pre-existing disease, and the use of substances that could lead to addictive consequences. The concepts of measuring all exposure events over time is certainly difficult, particularly considering the dynamic aspects of exposures leading to adverse outcomes; however, much effort is being given to establishing biomarkers related to the exposome on both a population and individual basis. These biomarkers are being evaluated for refining exposure assessments in risk assessments; providing correlations leading to exposure–disease associations, particularly in data from epidemiological studies; the potential identification of susceptible individuals or groups; using human data rather than extrapolations from animal data; and potentially identifying interventions in reducing certain exposures and/or treating the adverse outcome.<sup>32</sup> One of the major efforts to define and understand environmental exposures is the published biomonitoring studies, National Health and Nutrition Examination Survey, from the Centers for Disease Control and Prevention National Center for Health Statistics. The *Fourth National Report on Human Exposure to Environmental Chemicals* with updated tables<sup>33</sup> provides national (USA) biomonitoring data (serum and urinary levels) on 265 chemicals from subsets of the population. The website contains details of data sources and data analysis, interpretation of report data, and chemical and toxicological information. Bell and Edwards<sup>34</sup> have described a workflow, a frequent itemset mining approach, to identify relationships between chemicals and health biomarkers and disease. Currently, the most complete information source for toxicology information and exposure identification, including the exposome, is the Toxin and Toxin-Target Database<sup>31</sup> (T3DB; [www.t3db.ca](http://www.t3db.ca)). The details of T3DB are discussed in Chapter 2.

## 1.5 Systems and Network Pharmacology

Systems pharmacology is defined as a translational science that aims to examine all the biological activities in the body related to internal exposure of a drug or drug candidate and the resultant drug responses and pharmacological activities.<sup>35</sup> Systems pharmacology uses both experimental approaches and computational analyses to examine and understand drug action across multiple levels including molecular, cellular, tissue, and whole organisms<sup>36</sup> with consideration to the presence of several interacting pathways.<sup>37</sup> The field has grown and developed rapidly because of the emergence of omics technologies and network analysis capabilities, and the increased number of computer scientists, engineers, and mathematicians involved in addressing and solving complex biological problems.<sup>35</sup> In an NIH white paper by the Quantitative Systems Pharmacology (QSP) workshop group in 2011, QSP was defined as providing an integrated approach to determining and understanding mechanisms of action of drugs and drug candidates in preclinical models (*in vitro* and *in vivo*) and in patients eventually receiving the drugs.<sup>38</sup> The stated goals were to create a knowledge base to facilitate the change of complex cellular networks in pre-determined ways with mono and/or combination therapies; maximize therapeutic benefit by altering the pathophysiology of the disease being treated; and minimize toxicity.<sup>38</sup> Given that the mammalian signaling and regulatory pathways are complex, drug–target interactions can potentially lead to adverse effects due to the propagation of signal flow to distal effectors (off-targets) in multiple cells and tissues.<sup>39</sup> However, using complex pharmacological and toxicological network analyses, both positive and negative effects can be predicted. Zhao and Iyengar<sup>39</sup> have identified key questions that highlight the importance of identifying and pursuing a systems pharmacology approach in drug research as a starting point: (1) what are characteristics of specific diseases where drugs modulating a single target may not provide therapeutic efficacy; (2) how do adverse events arise from intra- and intercellular networking; (3) how does the genomic status of an individual relate to potential drug efficacy particularly when poly-pharmacy (combination) is anticipated; (4) how do combinations of targets and/or signaling nodes in complex diseases predict efficacious outcomes with drug combinations; and (5) can detailed usage of the interactome and genetic status of an individual predict therapeutic efficacy or toxicity? Practically, systems pharmacology allows the application of model-based thinking during target selection and target validation before a lead compound is selected for development.<sup>40</sup> QSP models can incorporate details of single and multiple drug plasma concentrations, systems biology models, pertinent regulatory networks and motifs of upstream and downstream loops including feedback and feedforward processes, and individual genomic and epigenetic characteristics important for individualized patient therapies.<sup>41</sup> Visser and colleagues<sup>42</sup> describe the use of QSP models and the creation of a flexible tool kit at Merck, which has enhanced key drug discovery and development decisions. The tool kit includes PK/PD models, disease



models, comparator models, model-based meta-analysis approaches, clinical trial design simulations, criteria for quantitative decision-making, and overall performance metrics. As an example, these approaches have been used to quantify anticancer drug synergy in resistant cells and predicting effective drug combinations. Models have also been effective in predicting and understanding positive off-target activities that could require early risk-benefit considerations. These activities include endocrine disruptors, peroxisome proliferator-activated-receptor-agonists, 5-HT<sub>2B</sub> serotonin receptor agonists, and ligand-gated ion channel protein agonists.<sup>43</sup> An example of a tool for simulation and evaluation of QSP models, is the MatVPC, which incorporates visual predictive checks as a diagnostic to evaluate the structural and stochastic parts of a QSP model.<sup>44</sup> Biliouris and colleagues<sup>44</sup> illustrate the use with three models: (1) a three-compartment pharmacokinetics model with oral and intravenous bolus dosing; (2) a two-compartment pharmacokinetics model with multidose intravenous infusion; and (3) a pharmacodynamics model describing the time-course of body weight. Zhang and colleagues<sup>41</sup> describe a Sobol sensitivity analysis that determines how much of the variability in QSP models relates to each input parameter including the interactions of multiple parameters as they relate to the overall model output variability. This is a highly important aspect of QSP model building, refinement, and use, as it identifies the important and influential parameters that drive model output and, therefore, the inherent uncertainty of model predictions.

### 1.5.1 Secondary Pharmacology and Off-Target Effects

Secondary pharmacology has been described as off-target pharmacology where a drug interacts with other targets as well as the intended target, and multi-target drug research where drugs can interact effectively with multiple targets increasing the therapeutic efficacy in certain diseases.<sup>36,45–49</sup> These effects can provide both beneficial and adverse outcomes, and in some cases these drug qualities define several adverse effects seen with drugs in development and those marketed. Liu and colleagues<sup>49</sup> proposed a drug surveillance network for adverse drug reaction prediction through the integration of chemical compound signatures; biological targets including proteins, transporters, and enzymes, along with pathways; and phenotypic properties. Wang and colleagues<sup>45</sup> report on a protein pharmacology interaction network database, PhIN, where users can generate interacting target networks within and across human biological pathways by defining shared chemical compounds or scaffolds using a defined activity cutoff. The database also defines interactions between human-virus and virus-virus pathways. The database contains ~1 350 000 compounds; ~9400 targets with more than 12 400 000 activity measurements (as of March 2015). This type of database provides information and evidence-based predictions of chemical structures that interact with multiple targets, which would be useful in multi-target drug design and side effect predictions.



### 1.5.2 Prediction of Potential Adverse Effects

A predictive pharmaco-safety network has been proposed by Cami and colleagues,<sup>50</sup> where known drug–safety relationship networks are combined with adverse event information and detailed biological network information on several drugs as a means to predict likely but prospectively unknown adverse events. In this approach more directed surveillance programs can be instituted for drugs under development and those marketed. A multiple evidence fusion method for both approved and novel molecules was developed by Cao and colleagues<sup>51</sup>. In this approach the authors assumed that drug behavior at different biological levels would provide predictive information on adverse effects, and that semantic relationships between adverse reactions would aid in predicting new or unknown adverse reactions for certain drugs. They also found that drug–adverse-effect networks would allow the inference of unknown associations. These evaluations used similarity measures with drug and adverse event pairs. The authors concluded that these methods are inherently beneficial especially in drug discovery during target selection, drug repositioning, and multi-target inquiry and development. In addition, the methods provide a better focus for large-scale clinical trials, and more focused post-marketing drug surveillance.<sup>50–52</sup>

## 1.6 Conclusions

Systems biology approaches as applied to the fields of toxicology and pharmacology have increased our abilities to both visualize and understand complex chemical–biological interactions at the molecular, organ, susceptible individual, and species levels. Applying quantitative mechanistic models into a network-based analysis has not only improved our knowledge base on both chemically-induced pharmacological and toxicological effects, but also has allowed new approaches to emerge that rely less on *in vivo* animal testing. The ever growing abundance of databases and tools have caused the practitioners of toxicology, in particular, to step forward out of the proverbial animal toxicity box and approach solutions to problems in new ways. This has improved the understanding of adverse events, hazards, and risk assessment in all related fields: research and development of therapeutics; environmental, workplace, and household chemical exposures; and the design of safer chemicals in green chemistry endeavours.

## References

1. N. Plant, An Introduction to Systems Toxicology, *Toxicol. Res.*, 2015, **4**, 9–22.
2. T. Hartung, E. van Vliet, J. Jaworska, L. Bonilla and N. Skinner, Food for thought...Systems Toxicology, *ALTEX*, 2012, **29**, 119.
3. J. Bai and D. Abernethy, Systems Pharmacology to Predict Drug Toxicity: Integration across Levels of Biological Organization, *Annu. Rev. Pharmacol. Toxicol.*, 2013, **53**, 451–473.

4. K. Kongsbak, N. Hadrup, K. Audouze and A. Vinggaard, Applicability of Computational Systems Biology in Toxicology, *Basic Clin. Pharmacol. Toxicol.*, 2014, **115**, 45–49.
5. S. Sturla, A. Boobis, R. FitzGerald, J. Hoeng, R. Kavlock, K. Schirmer, M. Whelan, M. Wilks and M. Peitsch, Systems Toxicology: From Basic Research to Risk Assessment, *Chem. Res. Toxicol.*, 2014, **27**, 314–329.
6. G. Ankley, R. Bennett, R. Erickson, D. Hoff, M. Hornung, R. Johnson, D. Mount, J. Nichols, C. Russom, P. Schmieder, J. Serano, J. Tietge and D. Villeneuve, Adverse Outcome Pathways: A Conceptual Framework to Support Ecotoxicology Research and Risk Assessment, *Environ. Toxicol. Chem.*, 2010, **29**(3), 730–741.
7. C. Willett, J. Rae, K. Goyak, B. Landesmann, G. Minsavage and C. Westmoreland, Pathway-Based Toxicity: History, Current Approaches and Liver Fibrosis and Steatosis as Prototypes, *ALTEX*, 2014, **31**, 407–421.
8. D. Johnson, Estimating human cancer risk from rodent carcinogenicity studies: the changing paradigm for pharmaceuticals, *J. Drug Metab. Toxicol.*, 2012, **3**, e114, DOI: 10.4172/2157-7609.1000e114.
9. *Tox21*, NRC, 2007, <http://dels.nas.edu/Report/Toxicity-Testing-Twenty-first/11970>.
10. D. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. Hutchinson, C. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, M. Ottinger, L. Vergauwen and M. Whelan, Adverse Outcome Pathway (AOP) Development I: Strategies and Principles, *Toxicol. Sci.*, 2014, **142**(2), 312–320.
11. D. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. Hutchinson, C. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, M. Ottinger, L. Vergauwen and M. Whelan, Adverse Outcome Pathway (AOP) Development II: Best Practices, *Toxicol. Sci.*, 2014, **142**(2), 321–330.
12. G. Langley, C. Austin, A. Balapure, L. Birnbaum, J. Bucher, J. Fentem, S. Fitzpatrick, J. Fowle, R. Kavlock, H. Kitano, B. Lidbury, A. Muotri, S.-Q. Peng, D. Sakharov, T. Seidle, T. Trez, A. Tonevitsky, A. van de Stolpe, M. Whelan and C. Willett, Lessons from Toxicology: Developing a 21st Century Paradigm for Medical Research, *Environ. Health Perspect.*, 2015, **123**(11), A268–A272.
13. T. Monticello, Drug Development and Nonclinical to Clinical Translational Databases: Past and Current Efforts, *Toxicol. Pathol.*, 2015, **43**, 57–61.
14. C. Keenan, S. Elmore, S. Francke-Carroll, R. Kemp, S. Peddada, J. Pletcher, M. Rinke, S. Schmidt, I. Taylor and D. Wolf, Best practices for the use of historical control data of proliferative rodent lesions, *Toxicol. Pathol.*, 2009, **37**, 679–693.
15. M. Sundqvist, A. Lundahl, M. Någård, U. Bredberg and P. Gennemark, Quantifying and communicating uncertainty in preclinical to clinical human dose-prediction, *CPT: Pharmacometrics Syst. Pharmacol.*, 2015, **4**, 243–254.
16. R. Garcia-Serna, D. Vidal, N. Remez and J. Mestres, Large-scale Predictive Drug Safety: From Structural Alerts to Biological Mechanisms, *Chem. Res. Toxicol.*, 2015, **28**, 1875–1887.

17. A. Jacobs, Prediction of 2-year carcinogenicity study results for pharmaceutical products. How are we doing? *Toxicol. Sci.*, 2005, **88**, 18–23.
18. J. Bowes, A. Brown, J. Hamon, W. Jarolimek, A. Sridhar, G. Waldon and S. Whitebread, Reducing safety-related drug attrition: the use of *in vitro* pharmacological profiling, *Nat. Rev. Drug Discovery*, 2012, **11**, 909–922.
19. E. Blomme and Y. Will, Toxicology Strategies for Drug Discovery: Present and Future, *Chem. Res. Toxicol.*, 2016, **29**(4), 473–504.
20. S. Ivanov, A. Lagunin, P. Pogodin, D. Filimonov and V. Poroikov, Identification of Drug Targets Related to the Induction of Ventricular Tachyarrhythmia through a Systems Chemical Biology Approach, *Toxicol. Sci.*, 2015, **145**(2), 321–336.
21. B. Verbist, G. Verheyen, L. Vervoort, M. Crabbe, D. Beerens, C. Bosmans, S. Jaensch, S. Osselaer, W. Talloen, I. Van den Wyngaert, G. Van Hecke, D. Wuyts, QSTAR Consortium, F. Van Goethem and H. Göhlmann, Integrating High-Dimensional Transcriptomics and Image Analysis Tools into Early Safety Screening: Proof of Concept for a New Early Drug Development Strategy, *Chem. Res. Toxicol.*, 2015, **28**, 1914–1925.
22. A. Voutchkova, T. Osimitz and P. Anastas, Toward a Comprehensive Molecular Design Framework for Reduced Hazard, *Chem. Rev.*, 2010, **110**, 5845–5882.
23. [www.opentox.org/toxicity-prediction](http://www.opentox.org/toxicity-prediction).
24. Lhasa Derek and Meteor Suites, <http://www.lhasalimited.org>.
25. Toxcast, <http://www.epa.gov/ncct/toxcast/>.
26. S. Wink, S. Hiemstra, S. Huppelschoten, E. Danen, M. Niemeijer, G. Hendriks, H. Vrieling, B. Herpers and B. van de Water, Quantitative High Content Imaging of Cellular Adaptive Stress Response Pathways in Toxicity for Chemical Safety Assessment, *Chem. Res. Toxicol.*, 2014, **27**, 338–355.
27. H. Larson, E. Chan, S. Sudarsanam and D. Johnson, Biomarkers, in *Computational Toxicology, Methods in Molecular Biology*, ed. B. Reisfeld and A. Mayeno, Humana Press, Springer Science, NY, 2013, ch. 11, pp. 253–273.
28. K. Groh, R. Carvalho, J. Chipman, N. Denslow, M. Halder, C. Murphy, D. Roelofs, A. Rolaki, K. Schirmer and K. Watanabe, Development and application of the Adverse Outcome Pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology, *Chemosphere*, 2015, **120**, 764–777.
29. J. Wambaugh, B. Wetmore, R. Pearce, C. Strobe, R. Goldsmith, J. Sluka, A. Sedykh, A. Tropsha, S. Bosgra, I. Shah, R. Judson, R. Thomas and R. Setzer, Toxicokinetic Triage for Environmental Chemicals, *Toxicol. Sci.*, 2015, **147**(1), 55–67.
30. C. Grulke, K. Holm, M.-R. Goldsmith and Y.-M. Tan, PROCEED: Probabilistic reverse dosimetry approaches for estimating exposure distributions, *Bioinformation*, 2013, **9**(13), 707–709.
31. D. Wishart, D. Arndt, A. Pon, T. Sajed, A. Guo, Y. Djoumbou, C. Knox, M. Wilson, Y. Liang, J. Grant, Y. Liu, S. Goldansaz and S. Rappaport, T3DB: the toxic exposome database, *Nucleic Acids Res.*, 2015, **43**, D928–D934, DOI: 10.1093/nar/gku1004.

32. T. Adler, K. Sawyer, M. Shelton–Davenport and N. Grossblatt, The Exposome: A Powerful Approach for Evaluating Environmental Exposures and their Influences on Human Disease, *Emerging Science for Environmental Health Decisions Newsletter*, National Academies, 2010, Issue 2, ISSN 2376–1679.
33. NHANES, 2015. <http://www.cdc.gov/exposurereport>.
34. S. Bell and S. Edwards, Identification and Prioritization of Relationships between Environmental Stressors and Adverse Human Health Impacts, *Environ. Health Perspect.*, 2015, **123**(11), 1193–1199.
35. R. Turner, B. Park and M. Pirmohamed, Parsing interindividual drug variability: an emerging role for systems pharmacology, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2015, **7**(4), 221–241.
36. S. Berger and R. Iyengar, Network Analysis in Systems Pharmacology, *Bioinformatics*, 2009, **25**(19), 2466–2472.
37. S. Berger and R. Iyengar, Role of systems pharmacology in understanding adverse drug reactions, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2011, **3**(2), 129–135.
38. *Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms. An NIH White Paper by the QSP Workshop Group*, 2011, [www.nigms.nih.gov/training/documents/systemspharmawpsorger2011.pdf](http://www.nigms.nih.gov/training/documents/systemspharmawpsorger2011.pdf).
39. S. Zhao and R. Iyengar, Systems Pharmacology: Network Analysis to Identify Multiscale Mechanisms of Drug Action, *Annu. Rev. Pharmacol. Toxicol.*, 2012, **52**, 505–521.
40. P. Vicini and P. van der Graaf, Systems Pharmacology for Drug Discovery and Development: Paradigm Shift or Flash in the Pan? *Clin. Pharmacol. Ther.*, 2013, **93**(5), 379–381.
41. X.-Y. Zhang, M. Trame, L. Lesko and S. Schmidt, Sobol Sensitivity Analysis. A Tool to Guide the Development and Evaluation of Systems Pharmacology Models, *CPT: Pharmacometrics Syst. Pharmacol.*, 2015, **4**, 69–79.
42. S. Visser, D. de Alwis, T. Kerbusch, J. Stone and S. Allerheiligen, Implementation of Quantitative and Systems Pharmacology in Big Pharma, *CPT: Pharmacometrics Syst. Pharmacol.*, 2014, **3**, e142, DOI: 10.1038/psp.2014.40, 1–10.
43. T. Papoian, H.-J. Chiu, I. Elayan, G. Jagadeesh, I. Khan, A. Laniyonu, C. Li, M. Saulnier, N. Simpson and B. Yang, Secondary pharmacology data to assess potential off-target activity of new drugs: a regulatory perspective, *Nat. Rev. Drug Discovery*, 2015, **14**(4), 294.
44. K. Biliouris, M. Lavielle and M. Trame, MatVPC: A user-friendly MATLAB-based tool for the simulation and evaluation of Systems Pharmacology models, *CPT: Pharmacometrics Syst. Pharmacol.*, 2015, **4**(9), 547–557.
45. Z. Wang, J. Li, R. Dang, L. Liang and J. Lin, PhIN: A Protein Pharmacology Interaction Network Database, *CPT: Pharmacometrics Syst. Pharmacol.*, 2015, **4**, 160–166.
46. L. Espinoza-Fonseca, The benefits of the multi-target approach in drug design and discovery, *Bioorg. Med. Chem.*, 2006, **14**, 896–897.

47. E. Leung, Z.-W. Cao, Z.-H. Jiang, H. Zhou and L. Liu, Network-based drug discovery by integrating systems biology and computational technologies, *Briefings Bioinf.*, 2013, **14**, 491–505.
48. J. Dudley, E. Schadt, M. Sirota, A. Butte and E. Ashley, Drug discovery in a multidimensional world: systems, patterns, and networks, *J. Cardiovasc. Transl. Res.*, 2010, **3**, 438–447.
49. T. Liu, Y. Lin, X. Ween, R. Jorissen and M. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.*, 2007, **35**(suppl 1), D198–D201. Published online 2006.
50. A. Cami, A. Arnold, S. Manzi and B. Reis, Predicting Adverse Events using Pharmacological Network Models, *Sci. Transl. Med.*, 2011, **3**(114), 1–10.
51. D.-S. Cao, N. Xiao, Y.-J. Li, W.-B. Zeng, Y.-Z. Liang, A.-P. Lu, Q.-S. Xu and A. Chen, Integrating Multiple Evidence Sources to Predict Adverse Drug Reactions Based on a Systems Pharmacology Model, *CPT: Pharmacometrics Syst. Pharmacol.*, 2015, **4**, 498–506.
52. T. Lorderbaum, M. Nasir, M. Keiser, S. Vilar, G. Hripcsak and N. Tatonetti, Systems Pharmacology Augments Drug Safety Surveillance, *Clin. Pharmacol. Ther.*, 2015, **97**(2), 151–158.

## CHAPTER 2

# *Databases Facilitating Systems Biology Approaches in Toxicology*

DALE E. JOHNSON<sup>\*a,b</sup> AND ANN M. H. HESLIN<sup>b</sup>

<sup>a</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA;

<sup>b</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Morgan Hall, Berkeley, CA 94720-3104, USA

\*E-mail: [daleej@umich.edu](mailto:daleej@umich.edu)

## 2.1 Introduction

Several authors have published reviews of the application of systems biology approaches to toxicology, all of which require the use of multiple data sets and resources to draw novel inferences about mechanisms and modes of action of chemicals on biological targets, networks, and systems. These approaches include those in therapeutics research, particularly in understanding and predicting adverse drug reactions, and sometimes environmental or ecotoxicology approaches where information on the chemical(s) in question is sparse.<sup>1–7</sup> As discussed by Sturla and colleagues<sup>5</sup>, the core objective of systems toxicology is to uncover and hopefully elucidate mechanisms that causally link exposure to active substances with chemically induced adverse events and disease. The process requires the collection of quantifiable experimental

data typically coupled with extensive information gained by processing large sets of data positioned within biological networks and pathways. These data are garnered from accessible databases to allow the reflection of molecular changes in the context of cellular, tissue-level, or physiological changes that are linked to disease phenotypes or adverse events at the organism level.<sup>7</sup> Several authors have published lists of relevant databases for toxicology data, both for therapeutic and environmental research.<sup>1,8–10</sup> Accordingly, systems toxicology relies heavily on computational approaches to manage, analyze, and interpret these data with the ultimate goal to aid in the development of predictive *in silico* models that can be used in risk assessment. Computational systems toxicology has the following major areas of focus.<sup>6</sup>

- (1) Analyzing the massive amounts of *in vitro* and *in vivo* data contained in databases generated by multiple methods and correlating structural features of the compounds with levels of exposure and outcome.
- (2) Representing the relevant mechanisms leading to an adverse outcome as biological network models that describe the normal state and the causal effect of their perturbations upon exposure to chemical compounds.
- (3) Quantifying the dose-dependent and time-resolved perturbations of these biological networks their overall biological impact upon exposure and assessing risk.
- (4) Building and validating adequate computational models with predictive power that can be applied to risk assessment.

In the adverse outcome pathway model, the sequence of events that lead to an adverse outcome span multiple levels of biological organization,<sup>8</sup> but always contain a molecular initiating event, which is defined as the initial interaction between a chemical molecule and a biomolecule or biosystem that can be causally linked to an outcome *via* a pathway.<sup>8</sup> In the therapeutics toxicology field, systems pharmacology, an emerging interdisciplinary field combining network and chemical biology, provides important tools to uncover and understand adverse drug reactions and may mitigate the drawbacks of traditional methods. In particular, network analysis allows researchers to integrate heterogeneous data sources and quantify the interactions between biological and chemical entities. Recent work in this area has combined chemical, biological, and large-scale observational health data to predict adverse drug reactions in individual patients and global populations.<sup>11</sup>

As mentioned in Chapter 1, several factors have been involved in the rapid changes seen in the toxicology field, including the understanding of genetic changes, which have enhanced the ability to understand diseases, and chemically-induced toxicities at the molecular level. “Omics” technologies and approaches such as transcriptomics, metabolomics, proteomics, and epigenomics have changed the way mechanisms of toxicity and the perturbation of biological systems that lead to adverse outcomes are viewed.<sup>5</sup> These advances have been coupled with the public availability of large data sets of



information and new modeling approaches that have enhanced the ability to understand toxicological events and effects at multiple biological levels.<sup>6</sup> Since scientific approaches, inquiries, and visions aimed at understanding toxicological events and outcomes have been broadened tremendously, this reinforces the need for new and better ways to assess toxicity and risk.

## 2.2 Categorized Lists of Databases for Systems Toxicology

The following list includes free on-line data sources and tools placed into categories based on source and content.

### 2.2.1 TOXNET Databases (Including Those with Direct Links from TOXNET)

TOXicology Data NETwork (TOXNET)<sup>12</sup> is a central website hub with links to several toxicology data files that report on several chemicals, primarily those of toxicological and environmental concern. The breadth of information includes chemical nomenclature and current literature that gives evidence and/or speculation on a chemical's toxicological effects.

- *HSDB*<sup>13</sup> (*Hazardous Substances Data Bank*) Peer-reviewed toxicology data for >5000 hazardous chemicals. Data can be searched by relevance or filter by larger category groupings: human health effects, emergency medical treatment, animal toxicity studies, metabolism/pharmacokinetics, pharmacology, environmental fate/exposure, environmental standards & regulations, chemical/physical properties, chemical safety & handling, occupational exposure, standards, manufacturing/use information, laboratory methods, special references, synonyms and identifiers, and administrative information. Additional features of HSDB include “review status tags” that indicate the level of quality review: peer reviewed, QC reviewed (quality control review that has not yet been officially reviewed), and un-reviewed (statements that do not necessarily need scientific review). A complete list of chemicals in the HSDB is available at <https://sis.nlm.nih.gov/enviro/hsdbchemicalslist.html>.
- *TOXLINE*<sup>14</sup> 5 million references from specialized journals, government reports, meeting abstracts, and other relevant collections of toxicology information. The collection of information includes biochemical, pharmacological, physiological, and toxicological effects of drugs and other chemicals.
- *ChemIDPlus*<sup>15,16</sup> Dictionary of >400 000 chemicals including names, synonyms, and structures; a chemical searching system generated from more than 100 sources. National Library of Medicine databases serve as its primary source of information; however, it also compiles



information from the Canadian Domestic Substances List, European Inventory of Existing Commercial Chemical Substances (EINECS), Environmental Protection Agency (EPA) Toxic Substances Control Act (TSCA) Chemical Substance Inventory, the SUPERLIST set of regulatory resources, and other internet databases such as EPA Substance Registry System, the Food and Drug Administration (FDA) Drugs@FDA system, International Agency for Research on Cancer (IARC), National Institute of Allergy and Infectious Diseases (NIAID), and the National Institute of Standards and Technology (NIST) Chemistry WebBook. All chemicals are searchable by name, synonym, Chemical Abstracts Service (CAS) registry number, molecular formula, classification code, locator code, structure, and/or physical properties. Two versions of ChemID*plus* exist: ChemID*plus*Lite<sup>15</sup> and ChemID*plus*Advanced.<sup>16</sup> ChemID*plus*Lite provides chemical information searching and links to other resources. Unlike ChemID*plus*Advanced, ChemID*plus*Lite does not require plugins or applets. As such, ChemID*plus*Advanced has more advanced search capabilities. The default structure editor is Marvin for JavaScript (ChemAxon) that allows a user to download a single structure Mol file in ChemID*plus*. This applet enables users to conduct advanced chemical structure queries (substructure search, similarity search, exact structure search, flex search, and flexplus search), and filter similar and substructure chemical scaffolds. Structure descriptors include *InChI*<sup>TM</sup> (International Chemical Identifier), InChIKey, and *SMILES*<sup>TM</sup> (simplified molecular input line entry system) notations, all of which may be downloaded. InChIKeys link to other search engines to find a structure in other systems. Both Lite and Advanced ChemID*plus* records are updated daily. Although ChemID*plus* no longer supports Chime (it's previous free chemical display application), another structure-drawing package, Accelrys Draw No Fee, is now publicly accessible.

- *LactMed*<sup>17</sup> drugs and lactation database lists drugs and other chemicals to which breastfeeding mothers may be exposed. It includes information on the levels of such substances in breastmilk and infant blood, and the possible adverse effects in the nursing infant. Suggested therapeutic alternatives to those drugs are provided, where appropriate. All data are derived from the scientific literature and fully referenced.
- *DART*<sup>18</sup> Developmental and Reproductive Toxicology Database and references. It provides >200 000 journal references covering teratology and other aspects of developmental and reproductive toxicology. DART is created from a search profile using PubMed.
- *TOXMAP*<sup>19</sup> TOXMAP is a geographic information system from the Division of Specialized Information Services of the US National Library of Medicine that uses maps of the United States to show the amount and location of toxic chemicals released into the environment. Users can visually explore data derived from the EPA's Toxics Release Inventory (TRI), which provides information on the releases of toxic chemicals into the environment as reported annually by industrial facilities

around the United States. TOXMAP also contains information from the EPA's superfund program, as well as some non-EPA datasets such as the US Census and National Cancer Institute health data.

- *TRI*<sup>20</sup> Toxics Release Inventory. Annual environmental releases of more than 600 toxic chemicals by US facilities as reported annually to the EPA by US industrial and federal facilities. TRI's data reports, beginning with the 1987 reporting year, contain information about the types and amounts of toxic chemicals that are released each year to the air, water, land and by underground injection, as well as information on the quantities of toxic chemicals sent to other facilities for further waste management. In agreement with the Pollution Prevention Act of 1990, source reduction and recycling data are also included in TRI.
- *Household Products Database*<sup>21</sup> Potential health effects of chemicals in >10 000 common household products. Information is also available for some industrial-grade products. Products can be searched by brand name, product type, manufacturer, ingredient/chemical, and by health effects. The record for each product shows the ingredients as reported by the manufacturer. For many products, a link to the manufacturer's material safety data sheet is provided, which includes more information such as handling, disposal, and health effects.
- *Haz-Map*<sup>22</sup> Links jobs and hazardous tasks with occupational diseases and their symptoms, in which causality from chemical and/or biological agents has been established based on current scientific evidence.
- *IRIS*<sup>23</sup> Integrated Risk Information System. IRIS contains data in support of human health risk assessment, including hazard identification and dose-response assessments of more than 550 chemicals (as of mid-2014) that evaluate information on health effects (cancer and non-cancer) resulting from exposure to environmental contaminants. IRIS data are reviewed by EPA scientists several times a year and represents EPA consensus.
- *ITER*<sup>24</sup> International Toxicity Estimates for Risk. Risk information for more than 600 chemicals of environmental concern from authoritative groups worldwide. ITER integrates data from Centers for Disease Control Agency for Toxic Substances and Disease Registry, Health Canada, RIVM, US EPA, IARC, NSF International, and independent parties offering peer-reviewed risk values. It is compiled by Toxicology Excellence for Risk Assessment (TERA) and its records that are updated multiple times a year. The *Risk Information Exchange* (RiskIE; <http://www.alliance-forrisk.org/RiskIE.htm>) is a companion database to ITER. It includes in-progress or recently completed risk assessment projects. RiskIE is a database of notifications about a variety of human health risk assessment projects that are underway or recently completed. Projects listed on RiskIE are both chemical- and nonchemical-specific, and range from many types of risk value development to risk methods document development. RiskIE currently tracks >4000 in-progress or recently completed risk assessment projects conducted by 35 different organizations representing 13 different countries.

- *ALTBIB*<sup>25</sup> Resources on Alternatives to the Use of Live Vertebrates in Biomedical Research and Testing.
- *CCRIS*<sup>26</sup> Chemical Carcinogenesis Research Information System. Carcinogenicity and mutagenicity test results for >8000 chemicals (no longer updated).
- *CPDB*<sup>27</sup> Carcinogenic Potency Database. Standardized analyses of the results of 6540 chronic, long-term animal cancer tests (no longer updated).
- *GENE-TOX*<sup>28</sup> Genetic Toxicology Data Bank. Peer-reviewed genetic toxicology test data for more than 3000 chemicals (no longer updated).

### 2.2.2 US EPA Chemical Toxicity Databases

Chemical safety research data are made publicly available, including rapid, automated (high-throughput) chemical screening data; aggregated public sources of chemical toxicity data; animal toxicity studies; chemical exposure data and prediction models; and high quality chemical structures and annotations.

- *ToxCast and Tox21 Data*<sup>29</sup> Data on >2000 chemicals evaluated in more than 700 high-throughput assays.
- *ACToR*<sup>30</sup> ACToR (Aggregated Computational Toxicology Resource) is the EPA's online warehouse of all publicly available chemical toxicity data and can be used to find all publicly available data about potential chemical risks to human health and the environment.
- *ToxRefDB*<sup>31</sup> The Toxicity Reference Database (ToxRefDB) contains thousands of animal toxicity testing results, currently on 474 chemicals.
- *DSSTox*<sup>32</sup> DSSTox provides the accurate mapping of bioassay and physicochemical property data associated with chemical substances to their corresponding chemical structures.
- *CHAD*<sup>33</sup> Consolidated Human Activity Database; exposure and time-use studies. Data can be downloaded.
- *CPCat*<sup>34</sup> The Chemical and Product Categories database; catalogs the use of >40 000 chemicals and their presence in different consumer products.
- *iCSS Dashboard*<sup>35</sup> Developing tool expected to be the online portal for all chemical research data and studies.

### 2.2.3 National Toxicology Program Databases

- *CEBS*<sup>36</sup> The Chemical Effects in Biological Systems database houses data of interest to environmental health scientists. CEBS is a public resource, and has received depositions of data from academic, industrial, and governmental laboratories. CEBS is designed to display data

in the context of biology and study design, and to permit data integration across studies for novel meta-analysis. Users are required to install scripts for using Adobe Flex applications with JAWS.

- *DrugMatrix*<sup>37</sup> Comprehensive results of thousands of highly controlled and standardized toxicological experiments in which rats or primary rat hepatocytes were systematically treated with therapeutic, industrial, and environmental chemicals at both non-toxic and toxic doses.
- *ToxFx*<sup>38</sup> An automated toxicogenomics analysis application utilizing toxicogenomic signatures, specially curated biochemical pathways, and other relevant data to interpret toxicity-related transcriptomic data and present the results as a detailed, customized report.
- *Historical Control Databases*<sup>39–41</sup> Historical control data from NTP toxicity studies. Tumor incidences and growth and survival curves for control animals from NTP's 2 year carcinogenesis studies are summarized by species, sex, route of administration, and vehicle. The historical controls for the genetically modified models are also reported.
- *NICEATM LLMA Database*<sup>42</sup> (NTP Interagency Center for the Evaluation of Alternative Toxicological Methods) Analyses to evaluate the usefulness of the murine local lymph node assay (LLNA) to identify potential skin sensitizers.

## 2.2.4 Additional Toxicity Databases

- *T3DB*<sup>43</sup> The Toxin and Toxin Target Database is a bioinformatics resource that as its future name, Toxic Exposome Database, implies, is specifically designed to capture information about the toxic exposome. The focus of the T3DB is providing mechanisms of toxicity and target proteins for each toxin interactively linked in both directions. It is also fully searchable and supports extensive text, sequence, chemical structure, and relational query searches. The user can also hyperlink information into other databases without re-entering chemical information. The database currently houses 3673 toxins described by 41 733 synonyms, including pollutants, pesticides, drugs, and food toxins, which are linked to 2087 corresponding toxin target records. Altogether there are 42 471 toxin and toxin target associations. Each toxin record (ToxCard) contains more than 90 data fields and holds information such as chemical properties and descriptors, toxicity values, molecular and cellular interactions, and medical information. This information has been extracted from >18 143 sources which include other databases, government documents, books, and scientific literature. It is both modeled after and closely linked to the Human Metabolome Database and DrugBank.
- *FAERS*<sup>44</sup> Adverse Effects Reporting system of post-market safety surveillance for all approved drug and therapeutic biological products.
- *SIDER*<sup>45</sup> (Side Effect Resource) Information on marketed drugs and their recorded adverse drug reactions.

- *VAERS*<sup>46</sup> (Vaccine Adverse Event Reporting System).
- *JECDB*<sup>47</sup> (Japan Existing Chemical Data Base) Safety examination of existing chemicals and safety programs in Japan.
- *Offsides*<sup>48</sup> Finds different associations from adverse events reported during clinical trials before drug approval. The Offsides database is a resource of 438 801 off-label—those effects not listed on the FDA's official drug label—side effects for 1332 drugs and 10 097 adverse events.
- *Twosides*<sup>49</sup> A resource of polypharmacy side effects for pairs of drugs. This database contains 868 221 significant associations between 59 220 pairs of drugs and 1301 adverse events. These associations are limited to those that cannot be clearly attributed to either drug alone.
- *DITOP*<sup>50</sup> A comprehensive database providing Drug-Induced Toxicity Related Protein information. The related toxicities include overdose toxicity, idiosyncratic toxicity, drug–drug interactions, and genetic toxicity.

### 2.2.5 Chemical–Gene–Protein Databases

- *CTD*<sup>51</sup> Comparative Toxicogenomics Database. Provides access to scientific data describing relationships between chemicals, genes, and human diseases. The database contains curated data that describes cross-species interactions for chemical–gene, chemical–protein, and gene–disease. KEGG (Kyoto Encyclopedia of Genes and Genomes) and Reactome pathway data describe known molecular interaction and reaction networks. These data are integrated with chemicals, genes, and diseases in CTD to provide insights into molecular networks that may be affected by chemicals, and possible mechanisms underlying environmental diseases. CTD has a hierarchical arrangement of interactions that characterize physical, regulatory, and biochemical interactions. This vocabulary comprises 70 terms, including *actions* (e.g. “binds to”, “imports”), *operators* that describe the degree of a chemical's effect (e.g. “increases”), and *qualifiers* that specify the form of the gene or chemical involved in an interaction (e.g. “protein” or “chemical metabolite”, respectively). The chemical category integrates a chemical subset of the Medical Subject Headings (MeSH®), the hierarchical vocabulary from the US National Library of Medicine. The information about chemicals includes chemical structures, curated interacting genes and proteins, curated and inferred disease relationships, and enriched pathways and functional annotations. CTD contains curated and inferred chemical–disease and gene–disease associations. Inferred associations are established *via* CTD-curated chemical–gene interactions and inference scores are calculated for all inferred relationships. In gene–gene interactions, CTD represents gene–gene interactions from BioGRID (see later) that consist of genetic and protein interactions curated from primary literature for all major model organisms by BioGRID curators. These interactions are available for each gene and reference, and for the inference networks underlying each chemical–disease association.

In addition, the user can generate pathways for custom collections of genes using the set analyzer tool. Several other databases and tools that associate chemicals–genes–diseases use CTD as the primary source of information.

- *ChemDIS*<sup>52</sup> A chemical–disease inference system based on chemical–protein interactions.
- *STITCH*<sup>53</sup> (Search Tool for Interacting Chemicals) A resource to explore known and predicted interactions of chemicals and proteins. Chemicals are linked to other chemicals and proteins by evidence derived from experiments, databases, and the literature.
- *String*<sup>54</sup> Protein–protein interactive networks.
- *Chemprot*<sup>55</sup> A publicly available compilation of chemical–protein–disease annotation resources that enables the study of systems pharmacology for a small molecule across multiple layers of complexity from molecular to clinical levels.
- *Human Protein Atlas*<sup>56</sup> An interactive database and visualization tool for the human proteome based on antibody methods and transcriptomics analysis across all major tissues and organs of the human body.
- *BioGrid*<sup>57</sup> (Biological General Repository for Interaction Datasets) A searchable interaction repository with data compiled through comprehensive curation efforts. Includes protein and genetic interactions and post-translational modifications from major model organisms. All data can be freely downloaded.

## 2.2.6 Pathway-Network Databases

- *KEGG*<sup>58</sup> KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. The KEGG home page has links to several data-oriented entries including: KEGG pathway maps, BRITE functional hierarchies, KEGG modules, ortholog groups, genomes, genes and proteins, small molecules, glycans, biochemical reactions, enzymes, human diseases, drugs, and health information resources. There are also several analytical tools including mapping tools, pathogen checker of antimicrobial resistance genes, sequence similarity search, and chemical similarity search. The KEGG is a primary source of information for most databases that include pathway information.
- *Reactome*<sup>59</sup> Reactome is a curated and peer-reviewed pathway database that provides intuitive bioinformatics tools for visualization, interpretation, and analysis of pathway knowledge. Data mining and analysis tools include a pathway browser, tools that merge information into a portal, species comparisons, cystoscope plug-in, small molecule search, and literature citation searching.

- *Cytoscape*<sup>60</sup> Software platform for visualizing molecular interaction networks and biological pathways. Networks can be integrated with annotations, gene expression profiles, and other data. Additional features include apps, formerly called plug-ins (mostly free) which allow network and molecular profiling, different layouts, and connection with databases. Cytoscape links are frequently available with several other databases.
- *Pathway Commons*<sup>61</sup> Pathway Commons stores and disseminates knowledge about biological pathways: >42 000 pathways and 1 350 000 interactions from 22 data sources.
- *NDEX*<sup>62</sup> The Network Data Exchange provides an open-source framework where scientists and organizations can share, store, manipulate, and publish biological network knowledge.

### 2.2.7 Chemistry, Structural Alert, and QSAR Databases and Tools

- *PubChem*<sup>63</sup> Provides information on the biological activities of small molecules. PubChem is organized as three linked databases within the National Center for Biotechnology Information's Entrez information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem also provides a fast chemical structure similarity search tool.
- *ChemSpider*<sup>64</sup> ChemSpider is a free chemical structure database providing fast text and structure search access to over 50 million structures from hundreds of data sources.
- *ChemProp*<sup>65</sup> Several modules including structural alerts for electrophilic reactivity.
- *CORAL*<sup>66</sup> Quantitative structure–property relationships (QSPR)/quantitative structure–activity relationships (QSAR) analysis for several toxicity endpoints.
- *OECD Toolbox*<sup>67</sup> QSAR toolbox for grouping chemicals into categories.
- *TEST*<sup>68</sup> The Toxicity Estimation Software Tool allows users to easily estimate the toxicity of chemicals using QSAR methodologies.
- *Virtual Computational Chemistry Laboratory*<sup>69</sup> On-line cheminformatics tools to calculate chemical properties including ALogP.
- *Danish (Q)SAR Database*<sup>70</sup> This Danish (Q)SAR database is a repository of estimates from over 70 (Q)SAR models for 166 072 chemicals. The (Q)SAR models encompass endpoints for physicochemical properties, fate, eco-toxicity, absorption, metabolism and toxicity.
- *Advaitabio: iPathwayGuide*<sup>71</sup> presents an advanced pathway analysis platform for high-throughput sequencing data.
- *LAZAR*<sup>72</sup> (Lazy Structure–Activity Relationships) Takes a chemical structure as input and provides several toxicity predictions. LAZAR is built on top of OpenTox [www.opentox.org/](http://www.opentox.org/).



- *Toxtree*<sup>73</sup> Toxtree is a full-featured and flexible user-friendly open source application, which is able to estimate toxic hazard by applying a decision-tree approach. Toxtree could be applied to datasets from various compatible file types. User-defined molecular structures are also supported—they could be entered by SMILES, or by using the built-in 2D structure diagram editor. Toxtree currently has the following plug-ins:
  - Cramer rules and Cramer rules with extensions
  - Verhaar scheme and modified verhaar scheme
  - skin irritation prediction
  - eye irritation prediction
  - START biodegradation and persistence
  - Benigni/Bossa rulebase for mutagenicity and carcinogenicity
  - *in vitro* mutagenicity (Ames test) alerts by ISS
  - structure alerts for the *in vivo* micronucleus assay in rodents (ISSMIC)
  - structural alerts for functional group identification (ISSFUNC)
  - structure alerts for identification of Michael acceptors
  - structure alerts for skin sensitization reactivity domains
  - DNA binding alerts
  - Protein binding alerts
  - Kroes thresholds of toxicological concern decision tree
  - SMARTCyp: cytochrome P450-mediated drug metabolism and metabolites prediction
- *CAESAR*<sup>74</sup> QSAR models supporting the REACH legislation including bioconcentration, skin sensitization, carcinogenicity, and developmental toxicity.
- *ToxAlerts*<sup>75</sup> A web-based platform for collecting and storing toxicological structural alerts from literature and for virtual screening of chemical libraries to flag potentially toxic chemicals and compounds that can cause adverse side effects. An alert is uniquely identified by a SMARTS template, a toxicological endpoint, and a publication where the alert was described. Additionally, the system allows storing complementary information such as name, comments, and mechanism of action, as well as other data.
- *Online Chemical Database*<sup>76</sup> The Online Chemical Modeling Environment is a web-based platform that aims to automate and simplify the typical steps required for QSAR modeling. The platform consists of two major subsystems: the database of experimental measurements and the modeling framework. A user-contributed database contains a set of tools for easy input, search and modification of thousands of records. Includes chemical property predictions, ToxAlert screening, and optimization of different properties with MolOptimizer.
- *ToxRead*<sup>77</sup> Software to assist in making reproducible read-across evaluations. The software shows similar chemicals to the input chemical, structural alerts, and other relevant features in common between chemicals.



## 2.2.8 Drug and Drug Target Databases

- *DrugBank*<sup>78</sup> The DrugBank database is a bioinformatics and cheminformatics resource that contains detailed drug data with correlating comprehensive drug target, including sequence, structure, and pathway information. The database includes >8000 drug entries including approved small molecules, biologics, nutraceuticals, and >6000 experimental drugs. DrugBank is the source of drug information for most databases that incorporate drug information.
- *TTD*<sup>79</sup> (Therapeutic Target Database) A database that provides information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and corresponding drugs directed at each target. Included are links to all relevant databases where detailed information exists.
- *Chemmapper*<sup>80</sup> An online platform to predict polypharmacy effects and mode of action for small molecules based on 3D similarity computation. ChemMapper collects >350 000 chemical structures with bioactivities and associated target annotations (as well as >3 000 000 non-annotated compounds for virtual screening).
- *Pharmmapper*<sup>81</sup> An updated integrated pharmacophore-matching platform with statistical methods for potential target identification. A pharmacophore database extracted from all targets in TargetBank, DrugBank, BindDB, and PDTD (Tripos mol2 or MDL SDF formats).
- *PDTD*<sup>82</sup> (Potential Drug Target Database) A dual function database of known and potential drug targets focusing on targets with known 3D structures. PDTD contains 1207 entries covering 841 known and potential drug targets with structures from the Protein Data Bank. This is connected to a docking program, *Tarfishdock*<sup>83</sup> that docks a small molecule into protein targets in PDTD (mol2 formats).
- *PK/DB*<sup>84</sup> (Database for pharmacokinetic properties) was designed with the aim of creating robust databases for pharmacokinetic studies and *in silico* absorption, distribution, metabolism, and excretion (ADME) prediction. The database contains high-quality data for structurally diverse compounds associated with known ADME properties, including human oral bioavailability, human intestinal absorption, plasma protein binding, blood–brain barrier, among others. PK/DB manages 1389 compounds incorporating structurally diverse drug-like and lead-like molecules which represent 4141 pharmacokinetic measurements, including five validated models for *in silico* ADME prediction.
- *BindingDB Binding Database*<sup>85</sup> is a database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules. The database contains 1 233 342 binding data, for 6352 protein targets and 541 006 small molecules. There are 2907 protein–ligand crystal structures with BindingDB affinity measurements for proteins with 100% sequence identity, and 7392 crystal structures for proteins with 85% sequence identity.

## 2.3 Websites with Extensive Links to Databases and Tools

The most comprehensive listing and linking website for drug-related tools and databases is Click2Drug<sup>86</sup> maintained by the Swiss Institute of Bioinformatics. The website contains a comprehensive list and links to computer-aided drug design software, databases, and web services. It currently includes 777 links categorized into the following sections:

- databases
- chemical structure representations
- molecular modeling
- homology modeling
- binding site prediction
- docking
- screening
- target prediction
- ligand design
- binding free energy estimation
- QSAR
- ADME toxicity

The links direct the user to and commercial tools (typically by subscription) and free on-line tools. Included in the links are those tools with applications on mobile devices; typically iPhone, iPad, and Android.

Another direct link to several databases and tools is from MD Yousuf Ansari, a PhD research scholar at the Centre of Pharmacoinformatics.<sup>87</sup> Extensive links include software<sup>88</sup> (32 links), databases<sup>89</sup> (four links), and web services<sup>90</sup> (nine links).

## 2.4 Conclusions

Continuing changes in the field and practice of the toxicology augmented by the introduction of new technologies and multiple information sources have enhanced the way researchers and risk assessors approach detailed questions on mechanisms and modes of action of chemicals on biological targets, networks, and systems. This is true in every aspect of toxicology, from therapeutic drug research and development to ecotoxicology and environmental safety. These are often thought of as “top-down” and “bottom-up” processes. Top down refers to the ability to understand the properties and risks of a chemical itself that enters the human and/or environmental systems and relate this to potential adverse effects and disease. Bottom up refers to creating safer chemicals from starting blocks such as in medicinal chemistry based on known toxicities of other chemicals or structural analogues. This is done to avoid certain adverse reactions in drug research, or in the

field of green chemistry, to design safer chemicals or avoid regrettable substitutions when introducing alternative chemicals into products. In each case, detailed information must be gleaned from several up-to-date information sources along with establishing known and/or inferred biological networks that connect exposure to mechanism to outcome. In this chapter, several free on-line databases are outlined which can provide a clear entry into systems toxicology field. The information sources are categorized to allow a more targeted interest approach.

## References

1. N. Plant, An Introduction to Systems Toxicology, *Toxicol. Res.*, 2015, **4**, 9–22.
2. T. Hartung, E. van Vliet, J. Jaworska, L. Bonilla and N. Skinner, Food for thought...Systems Toxicology, *ALTEX*, 2012, **29**, 119.
3. J. Bai and D. Abernethy, Systems Pharmacology to Predict Drug Toxicity: Integration across Levels of Biological Organization, *Annu. Rev. Pharmacol. Toxicol.*, 2013, **53**, 451–473.
4. K. Kongsbak, N. Hadrup, K. Audouze and A. Vinggaard, Applicability of Computational Systems Biology in Toxicology, *Basic Clin. Pharmacol. Toxicol.*, 2014, **115**, 45–49.
5. S. Sturla, A. Boobis, R. FitzGerald, J. Hoeng, R. Kavlock, K. Schirmer, M. Whelan, M. Wilks and M. Peitsch, Systems Toxicology: From Basic Research to Risk Assessment, *Chem. Res. Toxicol.*, 2014, **27**, 314–329.
6. G. Ankley, R. Bennett, R. Erickson, D. Hoff, M. Hornung, R. Johnson, D. Mount, J. Nichols, C. Russom, P. Schmieder, J. Serano, J. Tietge and D. Villeneuve, Adverse Outcome Pathways: A Conceptual Framework to Support Ecotoxicology Research and Risk Assessment, *Environ. Toxicol. Chem.*, 2010, **29**(3), 730–741.
7. C. Willett, J. Rae, K. Goyak, B. Landesmann, G. Minsavage and C. Westmoreland, Pathway-Based Toxicity: History, Current Approaches and Liver Fibrosis and Steatosis as Prototypes, *ALTEX*, 2014, **31**, 407–421.
8. A. Toropov, A. Toropova, I. Raska, D. Leszczynska and J. Leszczynski, Comprehension of drug toxicity: software and databases, *Comput. Biol. Med.*, 2014, **45**, 20–25.
9. L. Valerio, *In silico* toxicology for the pharmaceutical sciences, *Toxicol. Appl. Pharmacol.*, 2009, **241**, 356–370.
10. R. Judson, Public Databases Supporting Computational Toxicology, *J. Toxicol. Environ. Health, Part B*, 2010, **13**, 218–231.
11. M. Boland, A. Jacunski, T. Lorberbaum, J. Romano, R. Moskovitch and N. Tatonett, Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2016, **8**, 104–122.
12. *ToxNet*, <https://toxnet.nlm.nih.gov/>.
13. *HSDB*, <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>.

14. *TOXLINE*, <https://toxnet.nlm.nih.gov/newtoxnet/toxline.htm>.
15. *ChemIDPlus Lite*, <http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp>.
16. *ChemIDPlus Advanced*, <http://chem.sis.nlm.nih.gov/chemidplus/>.
17. *LactMed*, <https://toxnet.nlm.nih.gov/newtoxnet/lactmed.htm>.
18. *DART*, <https://toxnet.nlm.nih.gov/newtoxnet/dart.htm>.
19. *TOXMAP*, <http://toxmap-classic.nlm.nih.gov/toxmap/main/index.jsp>.
20. *TRI*, <https://www.epa.gov/toxics-release-inventory-tri-program>.
21. *Household Products Database*, <https://householdproducts.nlm.nih.gov/>.
22. *Haz-Map*, <https://hazmap.nlm.nih.gov/>.
23. *IRIS*, <https://www.epa.gov/iris>.
24. *ITER*, <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?iter>.
25. *ALTBIB*, <https://www.nlm.nih.gov/pubs/factsheets/altbibfs.html>.
26. *CCRIS*, <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>.
27. *CPDB*, <https://toxnet.nlm.nih.gov/cpdb/>.
28. *GENE-TOX*, <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX>.
29. *ToxCast & Tox21 Data*, <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data>.
30. *ACToR*, <https://actor.epa.gov>.
31. *ToxRefDB*, <https://www.epa.gov/chemical-research/toxcast-toxrefdb-data-release>.
32. *DSSTox*, <https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database>.
33. *CHAD*, <https://www.epa.gov/healthresearch/consolidated-human-activity-database-chad-use-human-exposure-and-health-studies-and>.
34. *CPCat*, <http://actor.epa.gov/cpcat/faces/home.xhtml>.
35. *iCSS Dashboard*, <http://actor.epa.gov/dashboard/>.
36. *CEBS*, <http://www.niehs.nih.gov/research/resources/databases/cebs/>.
37. *DrugMatrix*, <https://ntp.niehs.nih.gov/drugmatrix/index.html>.
38. *ToxFx*, <https://ntp.niehs.nih.gov/toxfx/>.
39. *NTP Historical Control Database (for studies using NTP-07 Diet)*, <https://ntp.niehs.nih.gov/results/dbsearch/historical/>.
40. *NTP Historical Control Database (for studies using NTP-2000 diet)*, <https://ntp.niehs.nih.gov/results/dbsearch/historical/>.
41. *NTP Historical Control Database for Genetically modified models*, <http://ntp.niehs.nih.gov/results/dbsearch/historical/gmm/index.html>.
42. *NICEATM LLMA Database*, <https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/immunotoxicity/nonanimal/index.html>.
43. *T3DB*, <http://www.t3db.ca/>, D. Wishart, D. Arndt, A. Pon, T. Sajed, A. Guo, Y. Djoumbou, C. Knox, M. Wilson, Y. Liang, J. Grant, Y. Liu, S. Goldansaz and S. Rappaport, *Nucleic Acids Res.*, 2015, **43**, D928–D934, DOI: 10.1093/nar/gku1004.
44. *FAERS*, <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm083765.htm>.
45. *SIDER*, <http://sideeffects.embl.de/>, M. Kuhn, I. Letunic, L. Jensen and P. Bork, The SIDER database of drugs and side effects, *Nucleic Acids Res.*, 2015, **44**(D1), D1075–D1079. Epub 2015 Oct 19.

46. VAERS, <https://vaers.hhs.gov/index>.
47. JECDB, [http://dra4.nihs.go.jp/mhlw\\_data/jsp/SearchPageENG.jsp](http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp).
48. Offsides, <http://omictools.com/offsides-tool>.
49. Twosides, <http://tatonetttilab.org/resources/tatonetti-stm.html>.
50. DITOP, <http://bioinf.xmu.edu.cn:8080/databases/DITOP/>.
51. CTD, <http://ctdbase.org/>.
52. ChemDIS, <http://cwtung.kmu.edu.tw/chemdis/>.
53. STITCH, <http://stitch.embl.de/>.
54. STRING, <http://string.embl.de/>.
55. Chemprot, <http://omictools.com/chemprot-tool>.
56. Human Protein Atlas, <http://www.proteinatlas.org/>.
57. BioGrid, <http://thebiogrid.org/>.
58. KEGG, <http://www.genome.jp/kegg/>.
59. Reactome, <http://www.reactome.org>, R. Haw and L. Stein, Using the Reactome Database, *Curr. Protoc. Bioinform.*, 2012, **38**(8.7), 8.7.1–8.7.23.
60. Cytoscape, <http://www.cytoscape.org/>.
61. Pathway Commons, <http://www.pathwaycommons.org>.
62. NDEX, <http://www.ndexbio.org>.
63. PubChem, <http://pubchem.ncbi.nlm.nih.gov>.
64. Chemspider, <http://www.chemspider.com/>.
65. ChemProp, <http://www.ufz.de/index.php?en=34598>.
66. CORAL, <http://www.insilico.eu/coral/>.
67. OECD Toolbox, <http://oecd.org/chemicalsafety/risk-assessment> and <http://www.qsartoolbox.org>.
68. T.E.S.T., <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.
69. Virtual Computational Chemistry Laboratory, <http://www.vcclab.org>.
70. Danish (Q)SAR Database, <http://eurl.jrc.ec.europa.eu/laboratories-research/predictive-toxicology/qsar.tools/DDB>.
71. Advaitabio: iPathwayGuide, <http://www.advaitabio.com/ipathwayguide.html>.
72. LAZAR, <http://www.in-silico.de/>.
73. Toxtree, <http://toxtree.sourceforge.net/>.
74. CAESAR, <http://www.caesar-project.eu/>, available at, [www.vega-qsar.eu](http://www.vega-qsar.eu).
75. ToxAlerts, <http://omictools.com/toxalerts-tool>.
76. Online Chemical Database, <https://ochem.eu/home/show.do>.
77. ToxRead, <http://www.toxgate.eu/>.
78. DrugBank, <http://www.drugbank.ca>.
79. TTD, <http://bidd.nus.edu.sg/group/cjttd/>.
80. Chemmapper, <http://omictools.com/chemmapper-tool>.
81. Pharmmapper, [https://bioinformatics.ca/links\\_directory/tool/10622/pharmmapper](https://bioinformatics.ca/links_directory/tool/10622/pharmmapper).
82. PDTT, <http://www.dddc.ac.cn/pdtd/>.
83. Tarfisdock, [www.dddc.ac.cn/tarfisdock/](http://www.dddc.ac.cn/tarfisdock/).
84. PK/DB, <http://miro.ifsc.usp.br/pkdb/index.php>.

85. *BindingDB*, <https://www.bindingdb.org/bind/index.jsp>, T. Liu, Y. Lin, X. Ween, R. Jorissen and M. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.*, 2007, **40**, D1100, DOI: 10.1093/nar/gk1999.
86. *Click2Drug*, [www.click2drug.org](http://www.click2drug.org).
87. M. D. Yousuf Ansari, <https://sites.google.com/site/yousufniper/home>.
88. *Software*, <https://sites.google.com/site/pharmacoinformaticshajipur/adme-toxicity/software>.
89. *Databases*, <https://sites.google.com/site/pharmacoinformaticshajipur/adme-toxicity/databases>.
90. *Web services*, <https://sites.google.com/site/pharmacoinformaticshajipur/adme-toxicity/web-services>.

## CHAPTER 3

# *Tools for Green Molecular Design to Reduce Toxicological Risk*

DAVID FAULKNER<sup>\*a,b</sup>, LEAH K. RUBIN SHEN<sup>b,c</sup>, VANESSA Y. DE LA ROSA<sup>d,h</sup>, DALE E. JOHNSON<sup>a,e</sup>, RACHEL HEMINGWAY<sup>f</sup>, RICHARD V. WILLIAMS<sup>f</sup>, PHILIP N. JUDSON<sup>f</sup>, JOHN ARNOLD<sup>b,c</sup> AND CHRIS D. VULPE<sup>a,g</sup>

<sup>a</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Berkeley, CA 94720-3104, USA; <sup>b</sup>University of California, Berkeley, Berkeley Center for Green Chemistry, Berkeley, CA 94720-3104, USA; <sup>c</sup>University of California, Berkeley, Department of Chemistry, Berkeley, CA 94720-1460, USA; <sup>d</sup>Northeastern University Social Science Environmental Health Research Institute, 360 Huntington Avenue, 318 INV, Boston, MA 02115, USA; <sup>e</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA; <sup>f</sup>Lhasa Limited, Leeds, LS11 5PS, UK; <sup>g</sup>University of Florida, College of Veterinary Medicine, Department of Physiological Sciences, Gainesville, FL 32610-0103, USA; <sup>h</sup>Silent Spring Institute, 320 Nevada Street, Suite #302, Newton, MA 02460, USA

\*E-mail: [dfaulkner@berkeley.edu](mailto:dfaulkner@berkeley.edu)

### 3.1 Introduction

The green chemistry design philosophy has proven useful in both academic and business settings,<sup>1,2</sup> but some aspects have yet to be fully exploited during the earliest stages of chemical design. The fourth principle of green chemistry, which states that chemicals should be “benign by design”, is possibly the most difficult of them all, because it relies on the ability to predict the behavior of a given compound in a biological system, including the degree and rate at which a chemical substance enters the systemic circulation and/or is present at the site of physiological or biochemical activity. Voutchkova and colleagues<sup>3</sup> have discussed the three fundamental requirements for chemical toxicity in a living system: (1) there must be exposure of the chemical to the living system; (2) the chemical substance must be bioavailable (the degree and rate at which a chemical substance enters the systemic circulation and/or is present at the site of physiological or biochemical activity) by the route of exposure; and (3) the chemical and/or metabolites must be capable of directly or indirectly causing an alteration (initiating event) which leads to an adverse outcome. The physicochemical properties of the compound(s) play a large role in these factors. As an example, reducing or blocking potential bioavailability through chemical design has been shown to be a successful process in green molecular design (GMD).<sup>3</sup>

Fortunately for chemists, the field of computational toxicology has grown tremendously in recent years in response to several initiatives to anticipate and reduce toxic hazard, improve and accelerate toxicity testing of new chemicals, and reduce the use of animals in the testing process.<sup>4,5</sup> While the field of computational toxicology owes its genesis to the fields of medicinal chemistry and drug development, both new and existing platforms can be harnessed to evaluate industrial chemicals as well. Importantly, in June 2016, the Frank R. Lautenberg Chemical Safety for the 21st Century Act was signed into law, which overhauls the United States’ 40-year-old statute governing chemicals, The Toxic Substances Control Act. In summary, the law requires the US Environmental Protection Agency (EPA) to ensure that no chemical in the US commerce poses an unreasonable risk to human health or the environment. The law also aims to reduce the use of laboratory animals in toxicology studies, and to develop, validate, and use reliable alternatives to animal studies. The alternatives, already in process, include computational toxicology and high-throughput cell-based assays. In addition, new technologies such as organs-on-chips assays will also be validated and used, which in several cases allows human endpoints to be incorporated into both research and regulatory efforts. These efforts will eventually create a large transparent information source (frequently called “big data”) that can be harnessed into the GMD process. Data transparency is the key to these efforts, and it is anticipated that information will be available to create new and/or update existing platforms expanding past the chemical space of pharmaceutical compounds. In an effort to propose a new innovative tool for GMD, the authors have analyzed several on-line and subscription based tools involving chemical design and human health and



environmental toxicological endpoints. Several of these tools are discussed in this chapter and a brief case study is presented using three tools individually. Based on these assessments, a new proposed GMD tool is outlined.

## 3.2 Physiochemical, Genotoxicity, and Blood–Brain Barrier Passage Properties of Chemicals

The traditional model used by medicinal chemists, pharmacologists, and toxicologists to understand how xenobiotics (compounds not naturally produced by the body) enter and leave the body considers the absorption, metabolism, distribution, excretion, and toxicity (ADMET) of the compound. Although remarkable progress has been made in recent years regarding the development of computational tools for predicting toxicological endpoints,<sup>6</sup> accurately anticipating toxicological hazard remains a significant challenge.<sup>7</sup> Currently, the most effective means to proactively reduce toxicological risk of environmental or industrial chemicals is to design molecules that are not readily absorbed by a biological system.<sup>8,9</sup> An additional factor is to redesign structural motifs to avoid metabolism of compounds into more toxic intermediates.<sup>10,11</sup> A review by Mangiatordi *et al.* describes *in silico* procedures for predicting metabolic processes.<sup>12</sup> In drug research, medicinal chemists have developed models to predict which physicochemical properties make a molecule more “drug-like”: that is, likely to be absorbed into the body, to be distributed within the body to certain targets, and to produce an effect.<sup>13</sup> Those same principles can provide guidance to design molecules that are less likely to interact with biological targets. These guiding principles are primarily concerned with the likelihood that a compound will be absorbed if it is administered orally, but they overlap significantly with the rules of dermal and respiratory absorption as well. Generally speaking, the rules that govern the likelihood of a compound crossing the cellular membrane are consistent across routes of exposure. Lipinski’s rules are a commonly used set of five properties that are used to predict whether a compound is likely to be readily absorbed through oral ingestion. The rules are as follows: chemicals with (1) more than five hydrogen bond donors, (2) more than 10 hydrogen bond acceptors, (3) a molecular weight >500 Da, or (4) a logP value (sometimes called  $\log K_{ow}$ ) >5 are unlikely to be well absorbed, unless the compound is (5) a substrate for a biological transporter, in which case it can be an exception to the previous rules.<sup>14</sup> A comparable analysis by Veber and colleagues similarly found that compounds with >10 rotatable bonds, >12 total hydrogen bond acceptors and donors, or a polar surface area >140 Å<sup>2</sup> were unlikely to be orally bioavailable in rats.<sup>15</sup> While numerous properties are instrumental in the absorption of exogenous compounds, lipophilicity, charge, similarity to endogenous substances, blood-to-gas partition, molecular weight, and polar surface area appear to be of the greatest value when designing molecules.<sup>8,9</sup> That is, molecules that are large, hydrophilic, charged at neutral pH, and that possess a large polar surface area are not readily absorbed in the gastrointestinal tract. If the molecule of interest is a substrate for a one of the body’s

many biological transport proteins, it may be actively transported into cells, and possibly distributed to the rest of the body. More complex and sophisticated models have been developed for predicting intestinal absorption based on *in vitro* data collected in the Madin-Darby Canine Kidney Epithelial cells (MDCK) or Caco-2 cell lines, which are used as models for intestinal absorption.<sup>16,17</sup> MDCK and Caco-2 permeability predictions are not included in many computational platforms, but are notable when they appear, as they provide some direct suggestions about bioavailability. Given the importance of genetic toxicity in the evaluation of commercial compounds, computational tools that predict whether a chemical has mutagenic potential are common.<sup>11,18</sup> The Ames test is an *in vitro* test which principally uses the bacterium *Salmonella typhimurium* to determine the potential mutagenicity of chemicals.<sup>19–21</sup> A positive Ames result should prompt reconsideration of structural features (such as reactive nucleophiles) likely to elicit mutagenic activity, as well as additional chemical biotransformations that can change mutagenic potential. Currently, assessing the potential mutagenicity of metabolites is limited to specific computational platforms and in most cases, potential metabolite structures have to be assessed separately.

Knowing if a chemical can cross the blood–brain barrier (BBB) is a significant aspect of drug design. The highly selective permeability of the BBB can make it challenging to deliver some drugs to the brain as intended, but it can also allow other compounds through that are not meant to affect the central nervous system (CNS), resulting in undesirable effects.<sup>22–24</sup> Predictive models for BBB permeability use lipophilicity, polar surface area, and whether the compound is a substrate for specific transporters to identify compounds that are likely to cross the BBB and potentially interact with neurological pathways. However, computational predictions of BBB permeability must be interpreted cautiously, because the BBB is a complex membrane that is difficult to model,<sup>25,26</sup> and because poor penetration alone may not be sufficient to prevent a chemical with chronic systemic exposure from achieving significant levels in the brain. The key knowledge required for a chemist wishing to use such predictive toxicology tools is to understand which molecular interactions and characteristics that are most benign or worrisome for a physiological system—concepts which are central to the design of medicinal compounds.

### 3.3 Tools for Green Molecular Design

Currently, several types of tools exist to predict the physicochemical properties and potential toxic effects of compounds; they may be broadly grouped into three categories: expert systems, decision trees, and (quantitative) structure–activity relationships (QSARs).

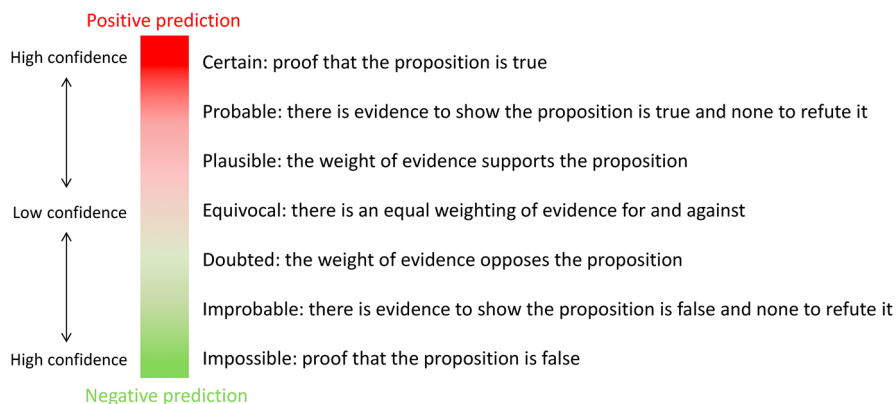
#### 3.3.1 Expert Systems

Expert systems use known relationships between chemical structures and toxicological outcomes to build rule-based predictive systems based on

structure–activity relationships.<sup>27,28</sup> Assembly of an expert system requires development and curation of a large database of “toxicophores”—specific functional groups or fragments of molecular structures known to cause toxicity—and implementing a set of rules that connects known toxins and toxicophores to appropriate toxic endpoints. If a user searches for a chemical that isn’t in the database, expert systems may employ “read-across” a technique which predicts hazard based on comparison to known toxins or toxicophores with similar structures.<sup>29</sup> It is anticipated that as new data become available, expert systems will be continually updated and may provide a more rapid path to validation of computational toxicology tools. Derek Nexus is an extensively used example of an expert system<sup>30</sup> and a detailed analysis of the system follows.

It is a proprietary reasoning-based expert system for the prediction of toxicity. Knowledge-based expert systems can be broadly placed into two categories: rule-based and reasoning based, with Derek Nexus being the latter. A rule-based system relies upon a series of rules that use “IF” and “THEN” statements to present the user with a prediction and the justification for it. However, there are limitations regarding the complexity of rules than can be encoded and the extent to which uncertainty can be handled thus limiting moderation of the outcome from a model. Reasoning based systems are capable of handling uncertainty and make predictions that account for the interactions between rules—which may themselves be imprecise.<sup>31</sup> The interactions between rules that may agree or disagree result in a prediction being strengthened, weakened, overturned, or contradicted,<sup>32</sup> and users are given information about confidence in the prediction Figure 3.1.

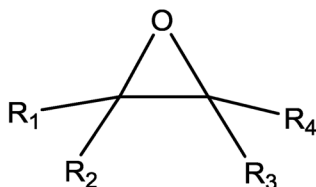
Derek Nexus is able to emulate the decision making of a human brain by using alerts that have been developed using appropriate data sources. The alerts are described by “patterns” or “Markush structures”, which define the scope of the alert and in combination with reasoning rules allow Derek Nexus



**Figure 3.1** Derek Nexus can assess the level of confidence in a prediction; expressed by a likelihood level (from certain to impossible).<sup>32</sup> The darker shades on the colour scale represent the increased confidence of a prediction being correct.

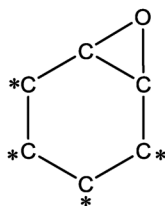
to make a prediction for or against toxicity for a given query compound and to advise on the level of confidence in it. They form a knowledge base covering more than 50 toxicological endpoints including mutagenicity, carcinogenicity, teratogenicity, and skin irritation. Each alert covers a specific area of chemical space by describing a chemical substructure, often referred to as a toxicophore, which is believed to be responsible for inducing a specific toxicological outcome. An alert often contains mitigating features or exclusions to ensure it is triggered only by the relevant toxicophore in the right structural environment. This can result in an alert containing many sub-structural patterns (Figure 3.2).

### Alert 019: Epoxide



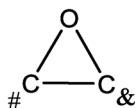
R1-R4 = Any atom but with exclusions including cyclohexyl epoxides with aliphatic fusions and tri- and tetra-alkyl substituted epoxides

#### Exclusion patterns:



\* ring bond count = 3,4

Cyclohexyl epoxides that are sterically hindered by other ring fusions generally do not give a positive response in the Ames test.<sup>33,34</sup> It is possible that the high molecular weight associated with these classes of compounds may also contribute to the attenuation of mutagenic activity



# hydrogen count = 0,1  
& hydrogen count = 0

Tri- and tetra-alkyl substituted epoxides are unlikely to possess Ames test activity<sup>35</sup> because under physiological conditions they generally undergo nucleophilic addition by an SN2 mechanism in a reaction which is primarily susceptible to steric influences<sup>36</sup>

**Figure 3.2** Alert 019 for epoxides demonstrates the compounds that would be predicted as Ames-active and those which would be excluded based on literature evidence.<sup>33,34,36</sup>

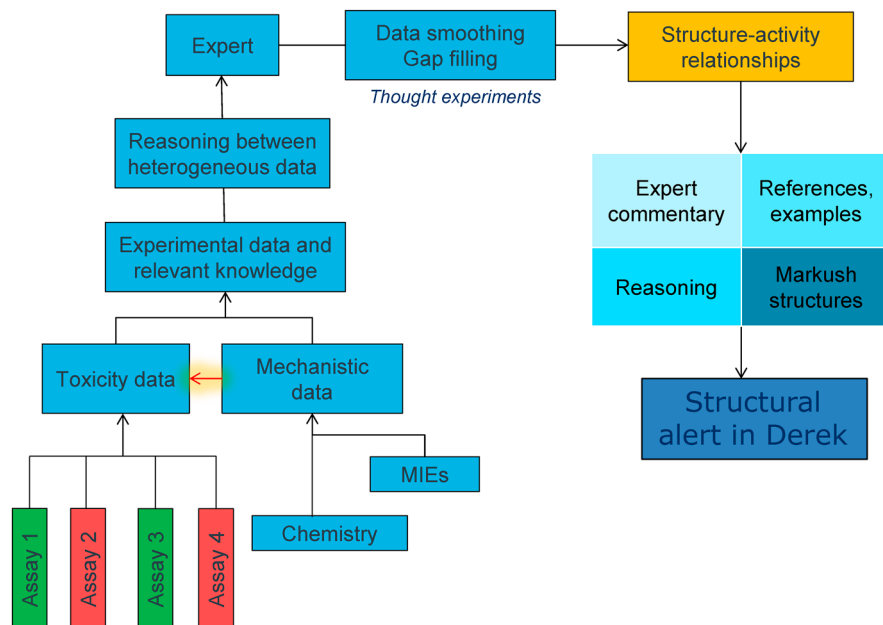
The data and information are appraised and interpreted by experts, who may be able to use their judgement to fill gaps in the data to derive a structure–activity relationship, which then forms the basis of the alert (Figure 3.4).

In addition, rules may be written about the species for which the alert is relevant and for factors that influence the manifestation of toxicity such as physico-chemical properties (*e.g.* octanol–water partition coefficients). These rules allow the inference engine to modulate the prediction for a query containing an alert and the result can range from a high expectation of activity all the way down to an explicit prediction of inactivity (as distinct from merely the absence of a prediction of activity). The use of human expertise is the key difference between a knowledge based system and a machine-learned statistical system such as Sarah Nexus.<sup>37</sup> Like a human expert, a knowledge-based system can assess the level of confidence in predictions from an

Diagram illustrating the metabolic pathways of an aromatic amine:

- Aromatic amine** (benzene ring with  $\text{NH}_2$ ) can be converted to an **N-hydroxylated metabolite** (benzene ring with  $\text{NH-OH}$ ).
- Aromatic amine** can also be converted to an **O-esterified metabolite** (benzene ring with  $\text{NH-OAc}$ ).
- The **O-esterified metabolite** can further be converted to a **Nitrenium ion** (benzene ring with  $\text{NH}^+$ ).

**Figure 3.3** Alert 351 for aromatic amines and amides illustrates how toxicity is predicted for a compound that is not directly toxic but is metabolised into an active species.<sup>38</sup>



**Figure 3.4** Workflow showing the synthesis of data and key stages for implementation of a structural alert in Derek.

alert for a given type of query structure on the basis of a wide range of relevant information such as chemical reactivity, characteristics of an assay, or an understanding of the likely mechanism, whereas a machine-learned system bases its overall call only on a statistical analysis of the training data and the descriptors employed. The derivation of each alert in Derek Nexus is described in the alert commentary along with supporting references and example compounds to provide a transparent prediction. Historically, Derek Nexus alerts have been restricted to qualitative prediction of the hazards posed by a chemical. Advances in both human and machine-learned knowledge mean that it is now possible for alerts to make at least semi-quantitative predictions. Derek Nexus can predict the skin sensitization potency class for a query compound in the local lymph node assay EC<sub>3</sub>, based on experimental data observed for the nearest neighbors, weighted by the Tanimoto similarity score, and selected from a set of compounds that exclusively fire the same alert as the query compound. The use of structural alerts for the prediction of toxicity is widely understood<sup>39,40</sup> and has found application in many settings.<sup>41–45</sup> Recent regulatory acceptance of *in silico* toxicity predictions is a significant milestone in the use of (Q)SAR for risk assessment. The International Conference on Harmonisation (ICH) M7 guidelines allow the acceptance of negative computer predictions for the genotoxic risk assessment of low-level pharmaceutical impurities, provided that they come from two complementary (Q)SAR methodologies; one expert rule-based and

one statistically based, for example Derek Nexus can be used in conjunction with Sarah Nexus. This dual approach reduces the risk that mutagens will be missed. In many cases, the development of new alerts and rules is supported by data-sharing initiatives in which organizations pool relevant data. These consortia share knowledge and data with the aim of advancing scientific knowledge as opposed to a specific product. ICH M7 has prompted new data- and knowledge-sharing initiatives for which Derek Nexus has been utilized as a tool.<sup>43</sup> Derek Nexus has also been used as part of an integrated testing strategy to predict for compounds outside the applicability domain of *in chemico/in vitro* assays<sup>44</sup> and as a tool to develop novel methods for making negative predictions for mutagenicity.<sup>45</sup> Derek Nexus is an example of computer prediction as a fast and green method with which to assess the toxicity of chemicals. The advent of stricter regulations on animal testing will increase the importance of computer models as predictive tools. As the models continue to improve and with increased sharing of data between companies, their predictions can be expected to become increasingly accurate and their use more commonplace.

### 3.3.2 Decision Trees

Decision trees use a series of Yes or No questions to classify and prioritize compounds based on their structural properties. The data for developing decision trees comes from mining the available literature and data of existing compounds and subsequently categorizing chemicals based on structural features and known toxic outcomes. An example is the Cramer classification scheme, a common decision tree for ranking chemicals in terms of their expected oral toxicity. It consists of 33 questions that place chemicals in one of three classes: class 1: low oral toxicity; class 2: moderate oral toxicity; and class 3: high oral toxicity. The original decision tree was developed by Cramer and Ford and colleagues<sup>46</sup> in the late 1970s, but modifications were proposed in 2002 to improve the accuracy of classifications. Although some consider the Cramer scheme to be in need of additional revision,<sup>47–49</sup> this decision tree has proven useful enough to be included in many computational tools such as Toxtree and the Organisation for Economic Co-operation and Development (OECD) QSAR Toolbox. Verhaar's scheme is yet another decision tree, developed to predict the likelihood of a chemical causing environmental toxicity.<sup>50</sup>

### 3.3.3 QSAR Tools

QSARs (sometimes called quantitative structure–property relationships (QSPR)) model the relationship between a chemical structure and a specific biological endpoint.<sup>27</sup> QSAR processes are discussed in detail in Chapter 6. There are many ways to create a QSAR, but the basic strategy is as follows: starting with a “training set” of molecules (compounds with known positive and negative values for an endpoint of interest), generate a set of properties



(descriptors) from the structure and apply one of a variety of mathematical techniques. These include multiple linear regression, partial least squares, neural networks, logistic regression, or linear discriminant analysis (Yee and Wei provide an excellent overview of various methods in a chapter on statistical modeling of molecular descriptors in QSAR/QSPR<sup>51</sup> to explain the variation in molecules in accordance with their values for that endpoint). QSARs are then validated with molecules that have known values for the endpoint, but which are not in the training set. The best computational tools provide information about the training sets used to build their QSARs, so that it is possible to determine how accurate their predictions are for a given molecule. If a molecule is very dissimilar from the compounds used in the training set, the QSAR will not generate a statistically meaningful prediction for that molecule. The chemical space for which a QSAR can provide statistically meaningful output is referred to as the “applicability domain”. Examples of tools that use QSARs are programs such as ACD Percepta, ADMET Predictor, Sarah Nexus, Medchem Designer, The OECD QSAR Toolbox, Moby@RPBS, and QikProp.

### 3.3.4 Representative Tools

To assess the utility and accessibility of computational tools currently available for molecular designers, a representative list of available computational tools was generated using [www.click2drug.org](http://www.click2drug.org), an online “directory of computer-aided drug design tools” maintained by the Swiss Institute of Bioinformatics.<sup>52</sup> The list includes single-function QSARs and collections of ADMET prediction models, with a mix of downloadable software, databases, and web-based platforms. The list includes the tools that the authors considered most useful and accommodating to general trends for the overlapping fields of computational toxicology, computational chemistry, and computational medicinal chemistry.

#### 3.3.4.1 ACD Percepta ([www.acdlabs.com](http://www.acdlabs.com))

This medicinal chemistry software<sup>42</sup> is distributed by Advanced Chemistry Development.<sup>53</sup> The program accepts single and batch .sdf or .mol files and also allows users to draw structures directly in the software. The outputs include basic physicochemical parameters (logP, H-donors/acceptors, rotational bonds, rings, Lipinski's rules violations, and solubility) as well as ADME (Caco-2 permeability, plasma protein binding, and CNS penetration), and some drug safety information (some cytochrome P450 (CYP) inhibition tests, Ames, and hERG). The program provides color-coded interpretations of those outputs and indicates on sliding scales how drug-like a molecule is. Navigation is easy, and different compounds can be compared in convenient tables or tabular format. Saving and sharing data is simple through export into either .pdf or .csv format, and the user interface is aesthetically pleasing



and easy to work with. Of particular utility is the Structure Design Engine module, as it allows the chemist to edit a molecule in a drawing window and observe how structural modifications affect predicted physicochemical and ADME properties in real time: as the molecule is modified, sliders move and alerts appear based on structural changes. Additionally, this module is capable of proposing a set of analogs for a given molecule based on a set of desired physicochemical parameters—a feature which greatly simplifies the “iteration” phase of the workflows. It is worth noting that this (like most of the programs discussed herein) is intended for the development of pharmaceuticals, not industrial chemicals, so appropriate care must be taken when interpreting the results generated by ACD Percepta, as the molecule of interest may fall outside the training sets used to develop the predictive algorithms. For chemists looking to improve the ADMET profile for a compound, this software offers most of what is needed.

#### 3.3.4.2 ADMET Predictor ([www.simulations-plus.com](http://www.simulations-plus.com))

A medicinal chemistry program distributed by Simulations Plus, this program contains a large number of QSARs capable of predicting ADME values, a plethora of toxicity endpoints, and numerous metabolism parameters including CYP metabolism kinetics.<sup>54</sup> Chemicals may be input as SMILES, .sdf, .rdf, or .mdl files, and may be entered in single or batch format. Predictions are returned in the form of a table, and hovering over a prediction produces an explanatory tooltip to ease interpretation. ADMET Predictor includes several “summary” toxicity prediction models, which assign rankings and codes indicating the specific toxicity concern(s) to chemicals based on the data outputs from certain physicochemical or QSAR models. These indicate if, for example, a compound is predicted to cause acute toxicity or carcinogenicity in rodents, to cause hepatotoxicity in humans, or to be Ames-positive. Aside from the physicochemical property predictions, these summary predictions are likely to be of the most interest to chemists because they greatly simplify data analysis. Prediction data is easily shared in the form of .tsv files. Like ACD Percepta, ADMET Predictor is intended for use in the development of pharmaceuticals, so predictions for the properties of industrial chemicals must be considered thoughtfully. The .tsv output files include “applicability of domain” statistics that indicate if the submitted molecule falls within the scope of the predictive model (based on similarity to the chemical structures used in the model training set). This feature provides a degree of confidence of the prediction, and incorporates a degree of transparency into the software.

#### 3.3.4.3 Medchem Designer

A chemical drawing program coupled with ADMET Predictor or available as a free standalone program, Medchem Designer features basic predictions of drug-likeness and bioavailability of molecules.<sup>55</sup> Structures may be drawn

or uploaded in SMILES, .mol, and .sdf formats. Predictions appear in table format below the workspace and are very easily exported to Excel. The interface is simple, clean, and can be learnt quickly, making it ideal for the novice user. The program uses the same codes to indicate predicted toxicity as are used in ADMET Predictor. Although Medchem Designer has far fewer QSARs than ADMET Predictor, it is sufficient for collecting enough physicochemical and ADMET parameters to make iterative molecular design adjustments. One of the useful features of this program is the Optical Structure Recognition tool, which allows the user to draw a box around an onscreen chemical structure and import that structure into the program. This can be helpful when the .mol or .sdf file of a compound is not immediately available for import.

#### 3.3.4.4 *Derek and Meteor Nexus from Lhasa Limited* ([www.lhasalimited.org](http://www.lhasalimited.org))

Lhasa Limited distributes the Nexus software through which the Derek and Meteor products are licensed.<sup>56,57</sup> The Derek platform is detailed earlier in Section 3.3.1.

As mentioned earlier, Derek provides expert, knowledge-based toxicity predictions using the Lhasa Knowledge Base, a curated database derived from literature and proprietary sources.<sup>30,56</sup> Structures may be drawn in the workspace, imported singly, or imported in batch format from .mol files, SMILES strings, or any delimited structure file. Derek is capable of providing predictions for charged or metal-containing compounds, which few medicinal chemistry tools are capable of. Outputs are tabular and the program provides a list of species and endpoints, including both the plausibility of toxicity and a note for which structure features triggered the alert. This data can then be easily exported to a .tsv file for viewing in Excel. Toxicity plausibility predictions are classified according to likelihood from “IMPOSSIBLE” to “CERTAIN”. If a structural feature does trigger an alert, the software highlights the offending moiety and offers detailed reasoning behind the alert, which may be used to inform design decisions. Derek also includes documentation to aid interpretation of predicted toxicophores or outcomes.

Meteor provides metabolism predictions in the form of tree diagrams, including the evidence-based rule for the reaction, a score indicating the likelihood of it occurring, and any intermediates that may be generated along the way.<sup>30,57</sup> Toxic products and intermediates are indicated, as are moieties on the molecule that are most likely to be sites of metabolism. Clicking on a metabolite opens a series of tabs containing information about the series of transformations which lead to it, details about biotransformations, and the “nearest neighbors”, chemical transformations documented in the literature which most closely resemble the reaction of interest. These predictions are helpful for identifying toxic biotransformation products and can be used to help make the necessary structural modifications to avoid them.

### 3.3.4.5 Qikprop ([www.schrodinger.com/QikProp](http://www.schrodinger.com/QikProp))

Qikprop is a predictive ADME module within the Maestro suite produced by Schrödinger.<sup>58</sup> It accepts singular .sdf, .mol, and .pdb files, and batch files may be imported in a variety of formats. A variety of output formats, including .csv, are also available. The program offers several predictors for ADME including CNS penetration, QPPCAco (permeability across the gut–blood barrier), QPPMDCK (kidney permeability), human oral absorption, Lipinski's rule of five, and JM (predicted maximum transdermal transport rate). Qikprop has an option to rank compounds on the basis of how drug-like they are. There is also an option that simplifies molecule-to-molecule comparisons by indicating how many of the predicted ADME values for the molecule fall outside the 95% range of similar values for known drugs—this is known as the "stars" mechanic, where more stars indicates a less drug-like molecule. Like ADMET Predictor, this program includes an array of predictive tools. Additionally, the documentation is extensive and thorough, and provides excellent information on the methods and training sets used to build the predictive models, so domain of applicability questions are easily answered.

### 3.3.4.6 OECD QSAR Toolbox

([www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm](http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm))

Produced by the OECD for evaluating industrial chemicals, this platform includes a blend of decision trees, QSARS, and predictive metabolism modules.<sup>59</sup> Structures may be uploaded as .mol, .sdf, or SMILES files, searched by a name or Chemical Abstracts Service (CAS) number, or drawn as a structure, and may be uploaded in batch format. Output is in a tree-based layout, indicating violations of various rules and sources of evidence for why the rule violations may contribute to toxicity. Modules for simulating metabolism, oxidation, and hydrolysis are included as well; the latter two may be used as indirect measures of environmental persistence. Notably, the software is capable of profiling both the chemical structure entered and the predicted metabolites of that structure in the same run. Attempting to run all decision trees and QSAR predictions on a batch of compounds requires extra time, but can expedite the process of evaluating a set of chemicals. Each output is hot-linked and clicking through links provides information about the QSARS or decision trees used to generate the outputs, their interpretations, and what part of the molecular structure triggered them. If the compound of interest is in the European Union database or has a CAS number, it is possible to search for charged compounds or metals—an advantage over medicinal chemistry software suites that do not handle any metals at all.

### 3.3.4.7 Toxtree (<http://toxtree.sourceforge.net/>)

Toxtree<sup>60</sup> provides basic toxicity evaluations, generally in the form of a binary "toxic" or "non-toxic" format, or a simple ranked format, usually with three to five hazard rankings.<sup>61</sup> The Toxtree models include the two Cramer's rule

sets,<sup>46</sup> Verhaar's schema,<sup>50</sup> the ISS decision trees, structural alerts for the *in vivo* micronucleus assay (a predictor of genetic toxicity), and the binding alerts trees.

#### 3.3.4.8 Chemaxon Suite (*Marvin Sketch and Metabolizer*) ([www.chemaxon.com/](http://www.chemaxon.com/))

The Marvin Sketch and Marvin Space<sup>62</sup> are useful for creating, viewing, and editing chemical structure files. In addition to its drawing function, Marvin Sketch also includes models for calculating a variety of physicochemical properties with assessment utility for bioavailability: pKa, logP, logD, aqueous solubility, H-bond donor/acceptor, and polar surface area. The Metabolizer program included in the Chemaxon suite provides metabolism predictions.

#### 3.3.4.9 Chemicalize ([www.chemicalize.com](http://www.chemicalize.com))

An online tool powered by Chemaxon's predictive algorithms, Chemicalize<sup>63</sup> can be used to generate an array of physicochemical properties and a few drug-likeness parameters. The data are easily exported and the layout allows a user to view multiple predictions at the same time. The interface allows the user to quickly modify chemical structures and generate physicochemical parameters.

#### 3.3.4.10 AIM (*Analog Identification Methodology*) (<http://www.epa.gov/tsca-screening-tools/analog-identification-methodology-aim-tool>)

Produced and distributed by EPA, this tool<sup>64</sup> accepts CAS numbers, chemical names, SMILES, or structural drawings, and returns a .pdf report with links to any information publically available for that compound or its analogs in US or Canadian chemical hazard databases.<sup>55</sup> This can be particularly useful when modifying structures, as well as for gathering information about the chemical space surrounding a compound of interest.

#### 3.3.4.11 Chemspider ([www.chemspider.com](http://www.chemspider.com))

A chemistry search engine with built in physicochemical prediction and other useful capabilities. Several options for modeling are available: users may choose between models from ACD, ChemAxon, or EPA's EpiSuite.<sup>65</sup>

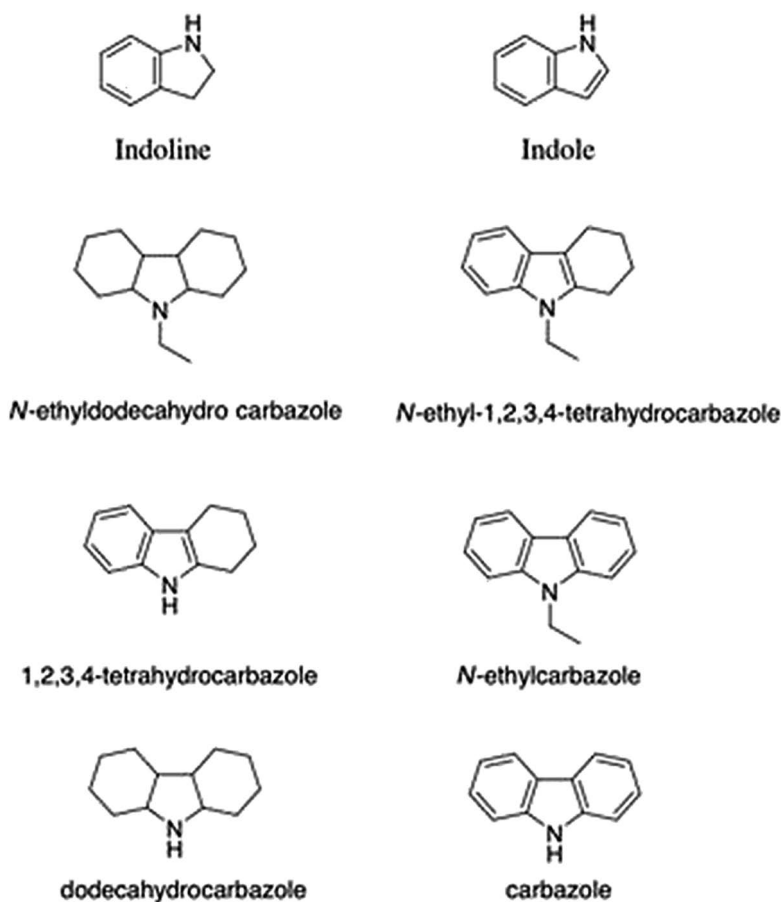
#### 3.3.4.12 Mobyle@RPBS (<http://mobyle.rpbs.univ-paris-diderot.fr>)

An online physicochemical profiler managed by the University of Paris Diderot,<sup>66</sup> this modeling tool accepts .sdf or .mol files for individual compounds.<sup>67,68</sup> Predictions can be saved in online user profiles or downloaded as a text file. The various physicochemical property predictive models are nested in a series

of menu trees. Mobyle@RPBS provides references to source papers used to program the physicochemical models, which provides transparency to the tool.

### 3.4 Case Study

As an example of how these tools may be put to use, we discuss a case study in which a collection of compounds with potential as hydrogen storage materials for proton-exchange membrane (PEM) fuel cells<sup>69–72</sup> are evaluated for their potential for human toxicity. The case study is centered on reduction of bioavailability and human hazard. The efficacy and performance of these compounds in PEM fuel cells was not considered as part of this example. Eight compounds were chosen for evaluation (Figure 3.5). The following tools were used to evaluate the compounds of interest: Lhasa Derek and Meteor Nexus, ADMET Predictor, and ACD Percepta.



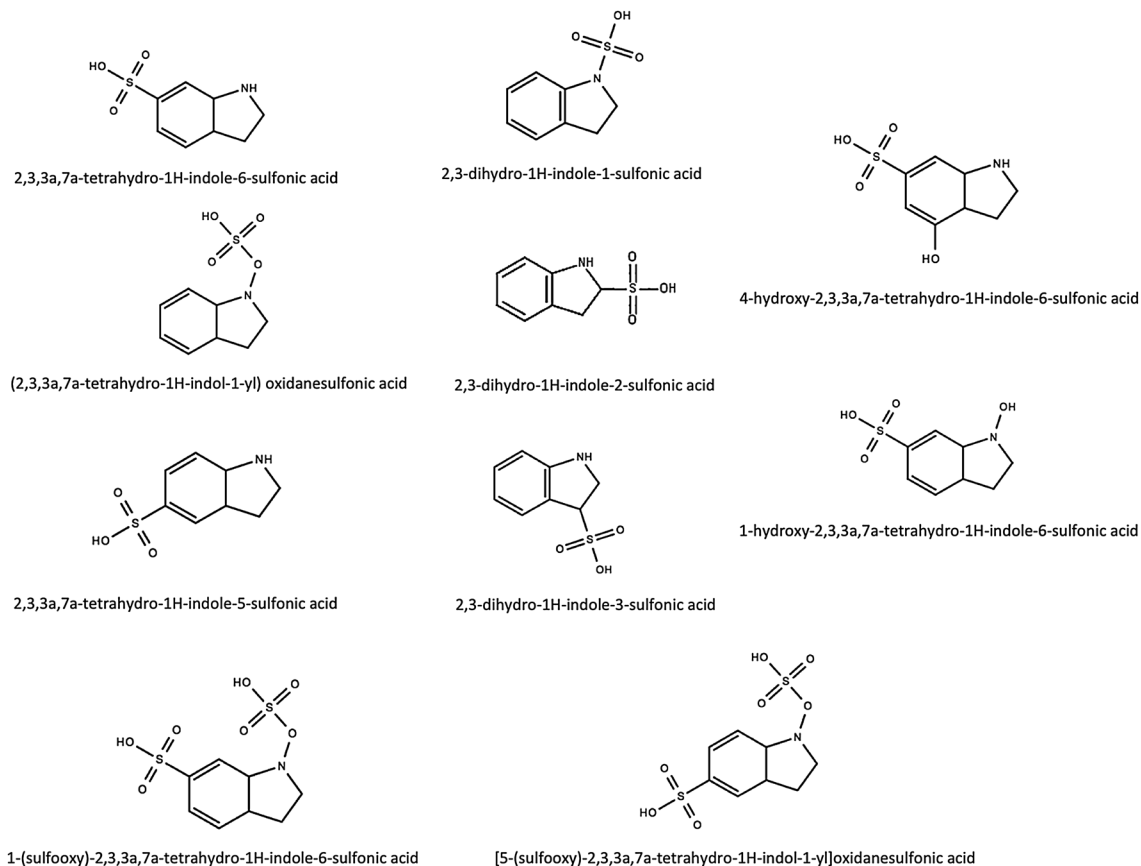
**Figure 3.5** Hydrogen compounds utilized in case study to evaluate predictive tools and workflow.

ACD Percepta and ADMET Predictor agreed that all the compounds were predicted to penetrate the BBB and were likely to be orally bioavailable due to favorable logP values (roughly between 2 and 5), low polarized surface area (all  $<16 \text{ \AA}^2$ ), and small size (all  $<400 \text{ Da}$ ). The Derek analysis highlighted *N*-ethylcarbazole as a CERTAIN Ames-positive compound, and gave an EQUIVOCAL rating to *N*-ethyldodecahydrocarbazole and dodecahydrocarbazole for causing phospholipidosis (a tissue-specific lipid metabolism disorder) in multiple species, including humans. Indoline was selected for further iteration, as it was the only compound among the initial eight that was without negative predicted effects, was predicted to be Ames-negative, and did not inhibit any CYPs. Using indoline as the base structure, a series of molecules was generated using the additions of alcohol, sulfite, or sulfate groups at various positions on the indoline core to generate less toxic and less bioavailable compounds. These groups and sites were chosen after interpreting the metabolism of indoline in humans as predicted by Meteor Nexus, and with consideration of the metabolism of indole. The second iteration of compounds is indicated in Figure 3.6. Running the compounds through ADMET Predictor, ACD Profiler, and Derek a second time revealed an improved overall profile for hazard and bioavailability for most of the compounds.

The Derek evaluation of the iterated compounds found no structural alerts for mutagenicity at EQUIVOCAL or PROBABLE levels, although four out of the 10 compounds did trigger 12% EC3 predictions (indicating that the compound might be a weak skin sensitizer), each compound only had a single supporting EC3 prediction to provide evidence. The second pass at the ACD Profiler indicated reduced gastrointestinal absorption, fewer CYP interactions, reduced BBB absorption, and “hydrophilic” ratings for half of the compounds. An iterative run through ADMET Predictor yielded an improved profile for many of the compounds, although some were still predicted to cross the BBB or cause genotoxicity. 4-Hydroxy-2,3,3a,7a-tetrahydro-1H-indole-6-sulfonic acid, however, had a very promising profile: it was predicted to be poorly permeable to the BBB and the gastrointestinal tract, charged at physiological pH, and was hydrophilic, suggesting that it would be cleared quickly. While this is not a complete redesign of the selected compound, it illustrates the value of the predictive tools selected for GMD.

### 3.5 The Design of Ideal Tools for Chemists

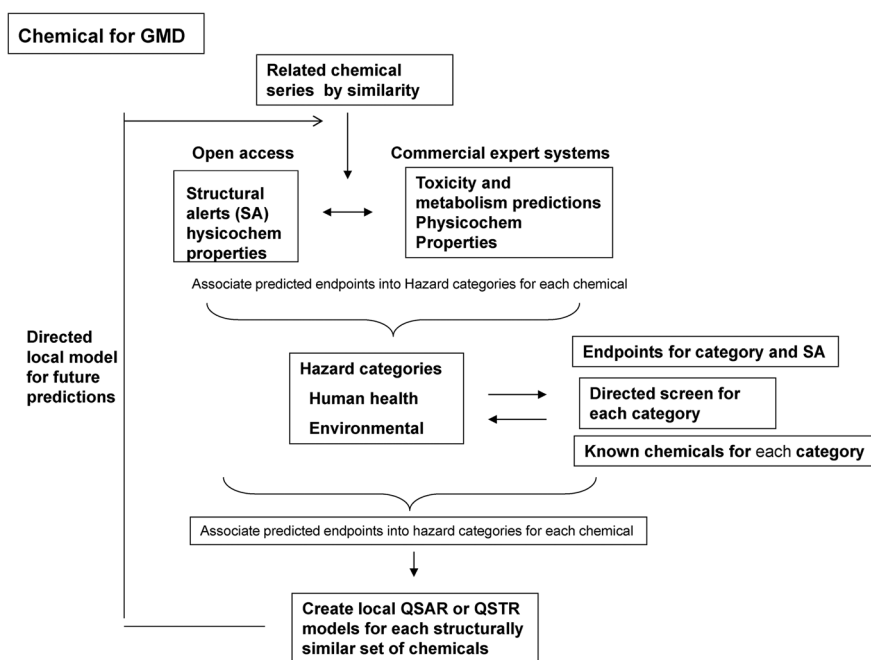
The ideal GMD tool for chemists of varying degrees of experience and knowledge, particularly of toxicological principles, would be an open-access front-end with the ability to incorporate subscription-based tools for specific functionalities. This is typically referred to as a “freemium” model. The premium or subscription-based tools would depend on the institution or business where the chemist works or studies. In this chapter, we have highlighted tools available at the University of California, Berkeley. The GMD freemium tool would allow all possible entries of chemical structures, predict species-centric metabolites, recommend multiple compounds with similar



**Figure 3.6** Compounds generated from second iteration of indoline.

structures, and allow instant analysis of all compounds for structural alerts, physicochemical properties, and hazard identification. Based on these evaluations, structural motifs could be modified in the tool to provide multiple options for GMD based on the hazard profiles predicted. It is anticipated that several screens are/ or will be available for rapid evaluation of hazard categories. Ultimately, the predicted results and screening results could be used to create local QSAR or quantitative structure–toxicity relationship (QSTR) models similar to models based on close structural analogs. These models could be used along with the initial structural alert and physicochemical property predictions. This process would supersede the use of global QSAR models where chemical space and applicability domain can become an issue. The overall scheme is presented in Figure 3.7.

The goals of eliminating or reducing animal toxicity testing as the primary source of information for risk evaluation has increased the development of new *in vitro* models, and the derived data will be key to the GMD process. Hutson and colleagues<sup>73</sup> have stated that organs-on-chips, which are 3D cultures of heterotypic cells that approximate the *in vivo* cellular microenvironment, are the next generation of tools for chemical toxicity testing. These tools are expected to improve computational modeling and the understanding of chemical exposure responses using several technologies, including high-content screening. As new sources of information become available, it is important that current tools stay updated and incorporate new findings.



**Figure 3.7** Scheme for green molecular design (GMD) tool.



For chemists involved in GMD, it is extremely important that the tools available will maximize the ability to predict modifications to structures that retain the desired functionality but also reduce toxicological risk.

## 3.6 Conclusions

GMD of chemicals involves the ability to use several tools both in chemical design and toxicological evaluation. There is a need for a new and innovative tool that allows chemists the ability to add proposed chemical structures in a variety of formats and to automatically calculate and predict key endpoints to aid in new chemical design. A proposed “freemium” model that provides an on-line interface with the ability to be coupled with various open-access and subscription based tools would be ideal. The tool could be used both in academics and industry and would provide an exceptional educational model for chemistry students.

## References

1. P. J. Dunn, The importance of green chemistry in process research and development, *Chem. Soc. Rev.*, 2012, **41**, 1452–1461, DOI: 10.1039/c1cs15041c.
2. M. J. Mulvihill, E. S. Beach, J. B. Zimmerman and P. T. Anastas, Green chemistry and green engineering: a framework for sustainable technology development, *Annu. Rev. Environ. Resour.*, 2011, **36**, 271–293, DOI: 10.1146/annurev-environ-032009-095500.
3. A. M. Voutchkova, T. G. Osimitz and P. T. Anastas, Toward a comprehensive molecular design framework for reduced hazard, *Chem. Rev.*, 2010, **110**, 5845–5882, DOI: 10.1021/cr9003105.
4. M. E. Andersen and D. Krewski, The vision of toxicity testing in the 21st century: moving from discussion to action, *Toxicol. Sci.*, 2010, **117**, 17–24, DOI: 10.1093/toxsci/kfq188.
5. C. O. T. Testing, A. O. E. Agents, B. O. E. Studies, Toxicology, Institute for Laboratory Animal Research, D. O. Earth and L. Studies National Research Council, *Toxicity Testing in the 21st Century: a Vision and a Strategy*, National Academies Press, Washington, D.C., 2007, DOI: 10.17226/11970.
6. A. Worth, J. Barroso, S. Bremer, J. Burton, S. Casati, S. Coecke, R. Corvi, B. Desprez, C. Dumont, V. Gouliarmou, M. Goumenou, R. Gräpel, C. Griesinger, M. Halder, A. Janusch Roi, A. Kienzler, F. Madia, S. Munn, M. Nepelska, A. Paini, A. Price, P. Prieto, A. Rolaki, M. Schäffer, J. Triebe, M. Whelan, C. Wittwehr, V. Zuang and Institute for Health and C. Protection, *Alternative Methods for Regulatory Toxicology a State-of-the-Art Review*, Publications Office, Luxembourg, 2014, DOI: 10.2788/11111.
7. C. L. Russom, S. P. Bradbury, S. J. Broderius, D. J. Hammermeister, R. A. Drummond and G. D. Veith, Predicting modes of toxic action from chemical structure, *Environ. Toxicol. Chem.*, 2013, **32**, 1441–1442, DOI: 10.1002/etc.2249.

8. C. Dumont, P. Prieto, D. Asturiol and A. Worth, Review of the availability of *in vitro* and *in silico* methods for assessing dermal bioavailability, *Appl. In Vitro Toxicol.*, 2015, **1**(2), 147–164, DOI: 10.1089/aivt.2015.0003.
9. A. M. Voutchkova, L. A. Ferris, J. B. Zimmerman and P. T. Anastas, Toward molecular design for hazard reduction-fundamental relationships between chemical properties and toxicity, *Tetrahedron*, 2010, **66**, 1031–1039.
10. D. K. Curtis, O. A. Mary and D. John, *Casarett and Doull's Toxicology: the Basic Science of Poisons*, Mc Graw-Hill, USA, 2001.
11. N. G. Bakhtyari, G. Raitano, E. Benfenati, T. Martin and D. Young, Comparison of *in silico* models for prediction of Mutagenicity, *J. Environ. Sci. Health, Part C*, 2013, **31**, 45–66, DOI: 10.1080/10590501.2013.763576.
12. G. F. Mangiatordi, D. Alberga, C. D. Altomare, A. Carotti, M. Catto, S. Cellamare, D. Gadaleta, G. Lattanzi, F. Leonetti, L. Pisani, A. Stefanachi, D. Trisciuzzi and O. Nicolotti, Mind the Gap! A Journey towards Computational Toxicology, *Mol. Inf.*, 2016, **35**, 294.
13. T. Hou, J. Wang, W. Zhang and X. Xu, ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules?, *J. Chem. Inf. Model.*, 2007, **47**, 460–463, DOI: 10.1021/ci6003515.
14. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Delivery Rev.*, 1997, **64**, 4–17, DOI: 10.1016/j.addr.2012.09.019.
15. D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, *Med. Chem.*, 2002, **45**, 2615–2623, DOI: 10.1021/jm020017n.
16. J. D. Irvine, L. Takahashi, K. Lockhart, J. Cheong, J. W. Tolan, H. E. Selick and J. R. Grove, MCDK (Madin-Darby Canine Kidney) cells: a tool for membrane permeability screening, *J. Pharm. Sci.*, 1999, **88**, 28–33, DOI: 10.1021/js9803205.
17. A.-L. B. Ungell, Caco-2 replace or refine? *Drug Discovery Today: Technol.*, 2004, **1**, 423–430, DOI: 10.1016/j.ddtec.2004.11.003.
18. A. Cassano, G. Raitano, E. Mombelli, A. Fernández, J. Cester, A. Roncaglioni and E. Benfenati, Evaluation of QSAR models for the prediction of Ames genotoxicity: a retrospective exercise on the chemical substances registered under the EU REACH regulation, *J. Environ. Sci. Health, Part C*, 2014, **32**, 273–298, DOI: 10.1080/10590501.2014.938955.
19. J. McCann and B. N. Ames, Detection of carcinogens as mutagens in the salmonella/microsome test: assay of 300 chemicals: discussion, *Proc. Natl. Acad. Sci. U. S. A.*, 1976, **73**, 50–54.
20. R. M. Walmsley and N. Billinton, How accurate is *in vitro* prediction of carcinogenicity? *Br. J. Pharmacol.*, 2011, **162**, 1250–1258, DOI: 10.1111/j.1476-5381.2010.01131.x.
21. D. Kirkland, M. Aardema, L. Henderson and L. Müller, Evaluation of the ability of a battery of three *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens: I. Sensitivity, specificity

- and relative predictivity, *Mutat. Res.*, 2005, **584**, 1–256, DOI: 10.1016/j.mrgentox.2005.02.004.
22. N. J. Abbott, Blood-brain barrier structure and function and the challenges for CNS drug delivery, *J. Inherited Metab. Dis.*, 2013, **36**, 437–449, DOI: 10.1007/s10545-013-9608-0.
  23. B. J. Aungst, Absorption enhancers: applications and advances, *AAPS J.*, 2012, **14**, 10–18, DOI: 10.1208/s12248-011-9307-4.
  24. B. Obermeier, R. Daneman and R. M. Ransohoff, Development, maintenance and disruption of the blood-brain barrier, *Nat. Med.*, 2013, **19**, 584–596, DOI: 10.1038/nm.3407.
  25. N. Strazielle and J.-F. O. Ghersi-Egea, Factors affecting delivery of antiviral drugs to the brain, *Rev. Med. Virol.*, 2005, **15**, 105–133, DOI: 10.1002/rmv.454.
  26. C. Suenderhauf, F. Hammann and J. Huwyler, Computational prediction of blood-brain barrier permeability using decision tree induction, *Molecules*, 2012, **17**, 10429–10445, DOI: 10.3390/molecules170910429.
  27. S. Modi, M. Hughes, A. Garrow and A. White, The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries, *Drug Discovery Today*, 2012, **17**, 135–142, DOI: 10.1016/j.drudis.2011.10.022.
  28. A. Sutter, A. Amberg, S. Boyer, A. Brigo, J. F. Contrera, L. L. Custer, K. L. Dobo, V. Gervais, S. Glowienke, J. V. Gompel, N. Greene, W. Muster, J. Nicolette, M. V. Reddy, V. Thybaud, E. Vock, A. T. White and L. Müller, Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities, *Regul. Toxicol. Pharmacol.*, 2013, **67**, 39–52, DOI: 10.1016/j.yrtph.2013.05.001.
  29. Y. Low, A. Sedykh, D. Fourches, A. Golbraikh, M. Whelan, I. Rusyn and A. Tropsha, Integrative chemical-biological read-across approach for chemical hazard classification, *Chem. Res. Toxicol.*, 2013, **26**, 1199–1208, DOI: 10.1021/tx400110f.
  30. C. A. Marchant, K. A. Briggs and A. Long, In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for Windows, Meteor, and Vitic, *Toxicol. Mech. Methods*, 2008, **18**, 177–187, DOI: 10.1080/15376510701857320.
  31. P. N. Judson and J. D. Vessey, A comprehensive approach to argumentation, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1356–1365, DOI: 10.1021/ci020271o.
  32. P. N. Judson, C. A. Marchant and J. D. Vessey, Using argumentation for absolute reasoning about the potential toxicity of chemicals, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1364–1370, DOI: 10.1021/ci020272g.
  33. H. Glatt, R. Jung and F. Oesch, Bacterial mutagenicity investigation of epoxides: drugs, drug metabolites, steroids and pesticides, *Mutat. Res.*, 1983, **111**, 99–118, DOI: 10.1016/0027-5107(83)90056-8.
  34. R. Lang and R. Reinmann, Studies for a genotoxic potential of some endogenous and exogenous sex steroids. I. Communication: examination for the induction of gene mutations using Ames Salmonella/

- microsome test and the HGPRT test in V79 cells, *Environ. Mol. Mutagen.*, 1992, **21**, 272–304, DOI: 10.1002/em.2850210311.
35. C. De Meester, M. Mercier and F. Poncelet, Non-mutagenicity of 2-methyl-2,3-epoxybutane and factors influencing the mutagenicity of 2,3-epoxybutane, *J. Appl. Toxicol.*, 1982, **2**, 284–288, DOI: 10.1002/em.2850210311.
36. R. E. Parker and N. S. Issacs, Mechanisms of epoxide reactions, *Chem. Rev.*, 1959, **59**, 737–799, DOI: 10.1002/em.2850210311.
37. T. Hanser, C. Barber, E. Rosser, J. D. Vessey, S. J. Webb and S. Werner, Self organising hypothesis: a new approach for representing and structuring SAR knowledge, *J. Cheminf.*, 2014, **6**, 1–21, DOI: 10.1186/1758-2946-6-21.
38. M. E. Colvin, F. T. Hatch and J. S. Felton, Chemical and biological factors affecting mutagen potency, *Mutat. Res.*, 1998, **400**, 479–492, DOI: 10.1016/S0027-5107(98)00073-6.
39. P. N. Judson, S. Stalford and J. Vessey, Assessing confidence in predictions made by knowledge-based systems, *Toxicol. Res.*, 2013, **2**, 70–79, DOI: 10.1039/C2TX20037F.
40. K. L. Dobo, N. Greene, M. O. Cyr, S. Caron and W. K. Ku, The application of structure-based assessment to support safety and chemistry diligence to manage genotoxic impurities in active pharmaceutical ingredients during drug development, *Regul. Toxicol. Pharmacol.*, 2006, **44**, 282–293, DOI: 10.1016/j.yrtph.2006.01.004.
41. N. Greene, K. L. Dobo, M. O. Kenyon, J. Cheung, J. Munzner, Z. Sobol, G. Sluggett, T. Zelesky, A. Sutter and J. Wichard, A practical application of two in silico systems for identification of potentially mutagenic impurities, *Regul. Toxicol. Pharmacol.*, 2015, **72**, 335–349, DOI: 10.1016/j.yrtph.2015.05.008.
42. A. Sutter, A. Amberg, S. Boyer, A. Brigo, J. F. Contrera, L. L. Custer, K. L. Dobo, V. Gervais, S. Glowienke, J. van Gompel, N. Greene, W. Muster, J. Nicolette, M. V. Reddy, V. Thybaud, E. Vock, A. T. White and L. Muller, Use of in silico systems and expert knowledge for structure-based assessment of potentially mutagenic impurities, *Regul. Toxicol. Pharmacol.*, 2013, **67**, 39–52, DOI: 10.1016/j.yrtph.2013.05.001.
43. D. P. Elder, a. White, J. Harvey, A. Teasdale, R. Williams and E. Covey-Crump, Mutagenic impurities: Precompetitive/competitive collaborative and data sharing initiatives, *Org. Process Res. Dev.*, 2015, **19**, 1486–1494, DOI: 10.1021/acs.oprd.5b00128.
44. D. S. Macmillan, S. J. Canipa, M. L. Chilton, R. V. Williams and C. G. Barber, Predicting skin sensitisation using a decision tree integrated testing strategy with an in silico model and in chemico/in vitro assays, *Regul. Toxicol. Pharmacol.*, 2016, **76**, 30–38, DOI: 10.1016/j.yrtph.2016.01.009.
45. R. V. Williams, A. Amberg, A. Brigo, L. Coquin, A. Giddings, S. Glowienke, N. Greene, R. Jolly, R. Kemper, C. O'Leary-Steele, A. Parenty, H. P. Spirk, S. Stalford, S. K. Weiner and J. Wichard, It's difficult, but important, to make negative predictions, *Regul. Toxicol. Pharmacol.*, 2016, **76**, 79–86, DOI: 10.1016/j.yrtph.2016.01.008.

46. G. M. Cramer, R. A. Ford and R. L. Hall, Estimation of toxic hazard—a decision tree approach, *Food Cosmet. Toxicol.*, 1976, **16**, 255–276, DOI: 10.1016/S0015-6264(76)80522-6.
47. S. Lapenna, A. Worth and Institute for Health and C. Protection, *Analysis of the Cramer Classification Scheme for Oral Systemic Toxicity Implications for its Implementation in Toxtree*, Publications Office, Luxembourg, 2011, DOI: 10.2788/39716.
48. S. Bhatia, T. Schultz, D. Roberts, J. Shen, L. Kromidas and A. Marie Api, Comparison of Cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgement, *Regul. Toxicol. Pharmacol.*, 2015, **71**, 52–62, DOI: 10.1016/j.yrtph.2014.11.005.
49. H. Kalkhof, M. Herzler, R. Stahlmann and U. Gundert-Remy, Threshold of toxicological concern values for non-genotoxic effects in industrial chemicals: re-evaluation of the Cramer classification, *Arch. Toxicol.*, 2012, **86**, 17–25, DOI: 10.1007/s00204-011-0732-z.
50. H. J. M. Verhaar, C. J. van Leeuwen and J. L. M. Hermens, Classifying environmental pollutants, *Chemosphere*, 1992, **25**, 471–491, DOI: 10.1016/0045-6535(92)90280-5.
51. L. C. Yee and Y. C. Wei, *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, 2012, Current modeling methods used in QSAR/QSPR, vol. 10, p. 1., DOI: 10.1002/9783527645121.ch1.
52. Click2Drug, [www.click2drug.org](http://www.click2drug.org), Jan. 2016.
53. G. Gasser and N. Metzler-Nolte, *Curr. Opin. Chem. Biol.*, 2012, **16**, 84–91.
54. ADMET Predictor, Version 7.2, Simulations Plus Inc., 2016.
55. Medchem Designer, Version 3.0, Simulations Plus Inc., 2014.
56. Derek, Version 5.0.1, Lhasa Limited, 2016.
57. Meteor, Version 3.0.0, Lhasa Limited, 2016.
58. QikProp, Version 4.6, Schrödinger, LLC, New York, NY, 2015.
59. OECD QSAR Toolbox, Version 3.3.5, Organization for Economic Co-operation and Development, 2015.
60. Toxtree, Version 2.6.13, Ideacon Ltd, 2015.
61. G. Patlewicz, N. Jeliaskova, R. J. Safford, A. P. Worth and B. Aleksiev, An evaluation of the implementation of the Cramer classification scheme in the Toxtree software, *Environ. Res.*, 2008, **19**, 495–524, DOI: 10.1080/10629360802083871.
62. Chemaxon Suite, <https://www.chemaxon.com/products/marvin/marvin-sketch/>, Jan. 2016.
63. Chemicalize, <http://www.chemicalize.org/>, Jan. 2016.
64. AIM: Analog Identification Methodology, Version 1.01, U.S. EPA, Risk Assessment Division, 2016.
65. *Chemspider*, Royal Society of Chemistry, [www.chemspider.com](http://www.chemspider.com), Jan. 2016.
66. Mobylye@RPBS, <http://mobylye.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#welcome>, Jan. 2016.
67. B. Néron, H. Ménager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery and C. Letondal, Mobylye: a new full web

- bioinformatics framework, *Bioinformatics*, 2009, **25**, 3005–3011, DOI: 10.1093/bioinformatics/btp493.
68. C. Alland, F. Moreews, D. Boens, M. Carpentier, S. Chiusa, M. Lonquety, N. Renault, Y. Wong, H. Cantalloube, J. Chomilier, J. Hochez, J. Pothier, B. O. Villoutreix, J. F. Zagury and P. Tufféry, RPBS: a web resource for structural bioinformatics, *Nucleic Acids Res.*, 2005, **33**, W44–W49, DOI: 10.1093/nar/gki477.
69. A. Moores, M. Poyatos, Y. Luo and R. H. Crabtree, *New J. Chem.*, 2006, **30**, 1675, DOI: 10.1039/b608914c.
70. E. Clot, O. Eisenstein and R. H. Crabtree, Computational structure-activity relationships in H<sub>2</sub> storage: how placement of N atoms affects release temperatures in organic liquid storage materials, *Chem. Commun.*, 2007, 2231, DOI: 10.1039/B705037B.
71. C. M. Araujo, D. L. Simone, S. J. Konezny, A. Shim, R. H. Crabtree, G. L. Soloveichik and V. S. Batista, Fuel selection for a regenerative organic fuel cell/flow battery: thermodynamic considerations, *Energy Environ. Sci.*, 2012, **5**, 9534, DOI: 10.1039/C2EE22749E.
72. P. F. Driscoll, E. Deunf, L. Rubin, J. Arnold and J. B. Kerr, Electrochemical redox catalysis for electrochemical dehydrogenation of liquid hydrogen carrier fuels for energy storage and conversion, *J. Electrochem. Soc.*, 2013, **160**, G3152–G3158, DOI: 10.1149/2.024307jes.
73. S. Hutson, P. Alexander, V. Allwardt, D. Aronoff, K. Bruner-Tran, D. Cliffl, J. Davidson, A. Gough, D. Markov, L. McCawley, J. McKenzie, J. McLean, K. Osteen, V. Pensabene, P. Samson, N. Senutovitch, S. Sherrod, M. Shotwell, D. Taylor, L. Tetz, R. Tuan, L. Veretti and J. Wikswo, Organs-on-Chips as Bridges for Predictive Toxicology, *Appl In Vitro Toxicol.*, 2016, **2**(2), 97–102, DOI: 10.1089/aivt.2016.0003.

## CHAPTER 4

# *Linking Environmental Exposure to Toxicity*<sup>†</sup>

NOFFISAT OKI<sup>a,b</sup>, JEREMY LEONARD<sup>a,c</sup>, MARK NELMS<sup>a,b</sup>,  
SHANNON BELL<sup>d</sup>, YU-MEI TAN<sup>c</sup>, LYLE BURGOON<sup>e</sup> AND  
STEPHEN EDWARDS<sup>\*b</sup>

<sup>a</sup>Oak Ridge Institute for Science and Education, Oak Ridge, TN, 37831, USA; <sup>b</sup>National Health and Environmental Effects Research Laboratory, United States Environmental Protection Agency, 109 TW Alexander Dr Mail Code B305-01, Research Triangle Park, NC 27709, USA; <sup>c</sup>National Exposure Research Laboratory, United States Environmental Protection Agency, Research Triangle Park, NC, 27709, USA; <sup>d</sup>National Institute of Environmental Health Sciences, Research Triangle Park, NC, 27709, USA; <sup>e</sup>Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS, USA

\*E-mail: [Edwards.stephen@epa.gov](mailto:Edwards.stephen@epa.gov)

## 4.1 Introduction

Exposure to chemicals that have been released into the environment contribute to the overall disease burden along with genetic causes and lifestyle

---

<sup>†</sup>The information in this document has been funded wholly (or in part) by the US Environmental Protection Agency. It has been subjected to review by the National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

---

Issues in Toxicology No. 31

Computational Systems Pharmacology and Toxicology

Edited by Dale E. Johnson and Rudy J. Richardson

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)



decisions. Addressing the disease through pharmaceutical intervention can positively impact all sources of disease, but exposure to environmental stressors and lifestyle decisions represent areas where disease prevention can potentially avoid the need for medical treatment in many cases. While people are ultimately responsible for their lifestyle decisions when informed about how those decisions impact their health and well-being, governmental bodies around the world have assumed the responsibility for controlling exposure to environmental stressors (primarily, but not exclusively, chemicals) to ensure the safety of their citizens. This requires science to support those risk-based decisions as well as mechanisms for translating the scientific information into a form that can be readily used for this purpose. The field of toxicology has always been the scientific core of this process and continues in this role. However, an understanding of the processes leading to the exposure of humans and wildlife to the chemicals released in the environment is critical for determining the ultimate impact on health and well-being. Coupling this information with an understanding of the toxicology of these chemicals allows for the prevention of disease by eliminating the source of toxicity rather than treatment of the disease after the fact.

Traditionally, the toxicity testing paradigm involved the use of large and expensive batteries of animal testing studies to determine the risk of adverse outcomes upon exposure to chemicals and other environmental toxicants. However, this traditional approach is quickly giving way to new approaches such as *in vitro*, *in silico*, and other alternative and more humane testing strategies for toxicity screening to support risk-based decision making. This paradigm shift has occurred, due in part, to the recognition that current *in vivo* testing methods are unable to keep pace with the ever-increasing number of chemicals in commerce and the environment, for which toxicity testing is needed. These issues have long been recognized, and toxicologists have operated under the principle of the three “R”s (reduction, replacement, and refinement of animal-based tests)<sup>1</sup> in toxicity testing for decades. More recently, legislative mandates in the United States<sup>2</sup> and Europe<sup>3</sup> have aimed at setting timelines for implementation of these principles into future toxicity testing strategies.

In 2007, the National Research Council (NRC) published a report outlining a new vision and strategy for toxicity testing in the 21st century.<sup>4</sup> This report recommended a switch to *in vitro* based approaches, with the additional benefit of not only leading to the reduction of animals used, but also reduction in the time and cost of toxicity testing. The basis for these tests is the concept of a toxicity pathway, which is defined as a cellular/molecular pathway that when sufficiently perturbed, can lead to an adverse health outcome. The NRC also made suggestions for the type of technological tools and computational methods that might be used to achieve their vision and that would aid in identifying critical toxicity pathways.<sup>4</sup> These tools and methods include high-throughput screening tests, microarrays, genomic data, the use of physiologically based pharmacokinetic (PBPK) models, and computational biology analysis methods. This new paradigm has led to an era of advances in



toxicological testing methods driven by improved and increasingly cost-efficient analytical and high-throughput *in vitro* technologies that are capable of measuring cellular pathway responses.

In contrast to animal-based methods, *in vitro* high-throughput toxicity testing (HTT) technologies are able to achieve broader biological coverage at relatively lower cost, as these assays are capable of screening thousands of chemicals across several hundred toxicity pathway endpoints.<sup>5,6</sup> In addition, these HTT technologies also aid in the prioritization of chemicals<sup>7</sup> for more rigorous follow-up *in vivo* testing. Therefore, HTT technologies are expected to reduce the heavy footprints of traditional toxicity testing methods and are better suited to keep pace with the number of chemicals for which toxicity testing is needed. However, these new technologies and methods present challenges including the need to understand (1) the toxicity pathways and/or mechanistic processes for which these *in vitro* technologies are acting as *in vivo* surrogates; (2) the other potential mechanistic processes required for manifestation of an adverse effect; and (3) the type and magnitude of exposure needed to generate the adverse reaction. With respect to challenge 3, animal-based tests have the advantage of including the ADME (absorption, distribution, metabolism, and excretion) processes that are needed to link external exposures to internal target tissue doses, albeit with no guarantee that the dosimetry will be consistent across different species. *In vitro* tests, in contrast, require greater understanding of the ADME processes in order to match the chemical concentrations that show effects in the *in vitro* assay with chemical concentrations to which humans and wildlife are exposed. Fortunately, advances in HTT have been occurring in parallel with advances in mechanistic toxicology and predictive exposure science such that these challenges are being addressed.

The mode of action (MOA) framework was initially described in 2001<sup>67</sup> and has been extensively refined over the past decade.<sup>68</sup> The US Environmental Protection Agency (EPA) guidelines for carcinogen risk assessment ([www.epa.gov/sites/production/files/2013-09/documents/cancer\\_guidelines\\_final\\_3-25-05.pdf](http://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf)) defines an MOA as the sequence of key events that begin with the interaction of an agent (chemical or environmental toxicant) and result in cancer formation. In 2008, the World Health Organization (WHO) International Programme on Chemical Safety (IPCS) expanded this framework to include non-cancer adverse outcomes.<sup>8</sup> Historically, MOA analyses, through the application of a weight-of-evidence approach, have been used to determine whether an MOA effect observed in an animal-based study will translate to the same effect in humans;<sup>9</sup> and as a means to incorporate mechanistic data in human health risk assessment.

The adverse outcome pathway (AOP) framework emerged out of similar concepts that had been developed to support ecological risk assessment,<sup>10</sup> and it provided a slightly more general framework because of the need to consider many species rather than a specific focus on humans. In a way, MOA could be thought of as a species-specific implementation of the AOP framework. More recently the AOP framework has been further generalized with an

emphasis on integration of toxicological data including HTT to understand the common mechanisms of toxicity rather than the assembly of chemical-specific data for a single risk assessment.<sup>7,11</sup> The AOP framework maintains the concept of a key event as the MOA and separates key events into different levels of biological organization (molecular/cellular, tissue, organ, individual, and population), that connect events that are measurable in HTT with adverse outcomes of relevance to human or ecological risk assessment.

Being chemically agnostic<sup>7</sup> allows the AOP to represent HTT in general and allow for increasing reliance on data from these assays for determining the potential for chemical toxicity as the confidence in the AOP increases. This is in contrast with traditional MOA analysis, where data used to build the MOA is typically derived exclusively from the chemical of interest. The chemically agnostic nature of AOPs also allows key events to be shared among AOPs, resulting in reusable components that allow parts of one AOP to be broadly applied to several adverse outcomes. For AOPs to be applicable to risk assessment, the chemical-specific data from an HTT can only inform potential for hazard. Any attempt at dose–response analysis requires that *in vitro* concentrations capable of perturbing a molecular target should be extrapolated to biologically effective doses, which can then be converted to external exposure levels through reverse dosimetry.<sup>12–14</sup> Chemical-specific ADME behaviors, which are usually lacking in HTT assays, mediate the relationships between external concentrations, biologically effective doses, and resulting chemical toxicity. This chemical-specific information can then be integrated with the broad biological knowledge contained within the AOP framework to better inform risk assessment strategies.

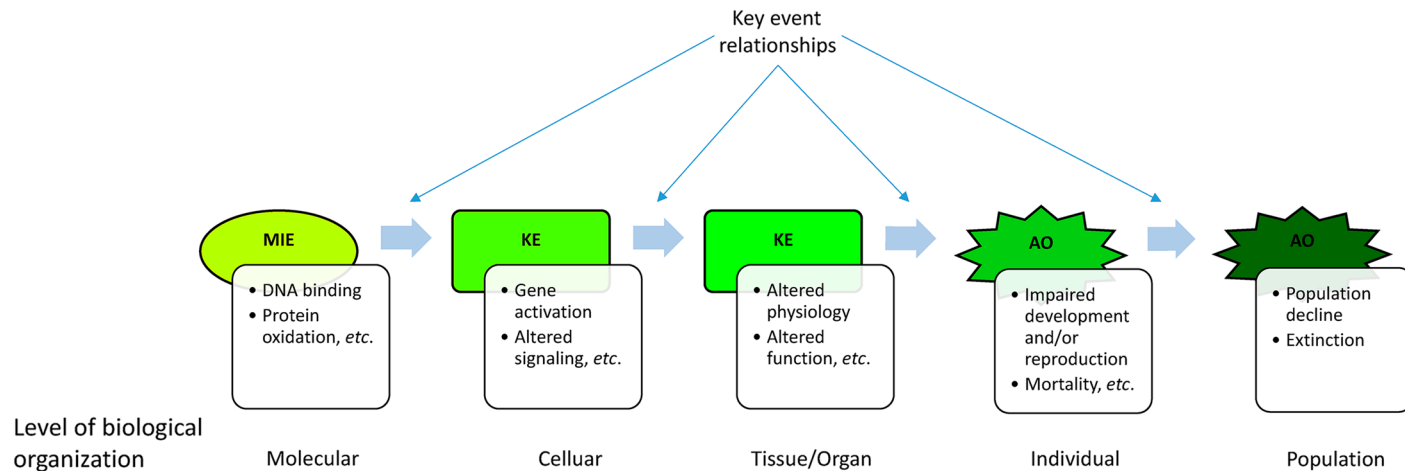
Another important component of the NRC's vision for toxicity testing was the incorporation of exposure and population-based data in the testing strategy. The IPCS harmonization project document on risk assessment terminology ([www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf](http://www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf)) defines exposure as the contact between an agent (*e.g.* chemical or environmental stressor) and a target (*e.g.* human or ecological receptor). Exposure science seeks to understand and characterize this interaction for the purpose of human and ecological health protection. Exposure assessments typically include measurements or estimates of the magnitude, frequency, route, pathway, and duration of exposure to the chemical agent, along with other population characteristics in the identification of public health hazards. Under the traditional risk assessment paradigm, exposure science is generally relegated to a supporting role, providing exposure estimates for comparison against hazard-based guidance values to determine whether unacceptable risks to public health exist. Also, historically, exposure data have been observational, which has limited them to serving as accompaniments to epidemiological studies for understanding impacts on individuals as well as populations. However, given the NRC's vision and the goal of more comprehensive assessments of thousands of chemicals/stressors for use in predictive toxicology, the need arose to complement toxicity testing with new strategies for exposure science in the 21st century.<sup>15</sup>

Major advances in analytical methods, biomonitoring, and computational tools have aided in the rapid transition of exposure science to a field that is more predictive, as well as data- and knowledge-driven. The need for an organizational and predictive framework for exposure science that furthers the application of systems-based approaches and fits the evolved exposure paradigm led to the conception of the aggregate exposure pathway (AEP) framework. The AEP is defined as a framework for organizing existing knowledge concerning biologically, chemically, and physically plausible, as well as empirically supported, links between the introduction of a chemical or stressor into the environment and its concentration at a site of action or target site of exposure.<sup>16</sup> AEPs allow for the organization of data and information emerging from an invigorated and expanding field of exposure science. The AEP framework is a layered structure that describes the elements of an exposure pathway, as well as the relationship between those elements.

Altogether, the aforementioned frameworks form a construct for understanding the processes that occur from the release of a stressor into the environment, subsequent exposure to that stressor, and the mechanism underlying any adverse outcome associated with the exposure. These frameworks also serve as a basis for informing risk assessment and decision making for endpoints of regulatory significance.

## 4.2 The AOP Framework: An Organizing Principle for Toxicological Data

An AOP consists of key events and key event relationships (KERs) as its basic components (Figure 4.1). These key events consist of essential and measurable occurrences at each level of biological organization and may be of three types: the molecular initiating event (MIE), an intermediary key event, and the adverse outcome.<sup>7</sup> The AOP typically begins with the MIE, a specialized key event that is defined as the initial interaction between the chemical/environmental stressor and the molecular/cellular agent. It serves as an anchor point for the AOP, and upon sufficient perturbation, may lead to further downstream events in the AOP. The intermediate key events are those key events downstream of the MIE which may span the molecular/cellular levels of biological organization up to the tissue/organ/individual level. The final key event in the AOP, which serves as the second anchor point at the other end of the spectrum, is the adverse outcome itself, which may extend through the individual and population levels of biological organization. For outcomes of human and public health concern (cancer, impaired development or reproduction, mortality, *etc.*) where regulations are based on the incidence of an apical endpoint within a target population, the adverse outcome within the AOP would be captured at the individual level. However, for ecotoxicological endpoints (population decline, extinction, *etc.*) the population level adverse outcome may be more appropriate. The KER, as its name suggests, describes the causal relationship between the two key events it connects and how



**Figure 4.1** The components of the adverse outcome pathway framework at the different levels of biological organization. MIE: molecular initiating event; KE: key event; AO: adverse outcome.

perturbations of the upstream key event affect the downstream key event in the AOP. Whenever possible, quantitative information (*e.g.* expected/predicted magnitude of change in one key event in relation to a corresponding change in the other key event) is emphasized when describing the relationship between connected key events. However, qualitative descriptions of an AOP can be assembled from existing literature with targeted studies to fill data gaps. These AOPs can be useful in many risk-assessment contexts where a quantitative risk assessment is not required or as support for a quantitative risk assessment using methods such as read-across.

In 2012, the Organisation for Economic Co-operation and Development (OECD) launched the AOP development program ([www.oecd.org/chemical-safety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm](http://www.oecd.org/chemical-safety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm)). Through the course of several meetings, workshops, and ancillary events, this program has established formal standards and guidance for the development and management of AOPs across the scientific and chemical management contexts in which they are used.<sup>17–19</sup> Currently, AOP development should be guided by five fundamental principles<sup>7</sup> that streamline the development process. These principles are stated as follows: (1) an AOP should not be specific to any one chemical; (2) AOPs are expected to be modular and reusable; (3) each individual AOP is a unit for development and evaluation; (4) for most applications, AOP networks are expected to be the functional unit of prediction; and (5) AOPs are living documents and, as such, may be refined/revised as new evidence is collected.

The first two principles have been discussed previously. Chemically agnostic AOPs maximize the “return on investment” for the extensive experimentation required to fully describe an AOP by allowing that information to be used for interpreting toxicological data for many chemicals. When coupled with reusable key events and KERs, which require little extra effort when used across multiple AOPs, this can greatly reduce the time and effort required when compared with a chemical-by-chemical approach, such as a traditional MOA analysis. The reuse of key events and KERs also results in an automatic assembly of AOP networks to address principles 3 and 4. Since MIEs and adverse outcomes are simply specialized key events, this can include the possibility of an MIE from one AOP being an intermediate key event in another AOP. For instance, in an example described by Ankley *et al.*,<sup>10</sup> estrogen receptor binding is considered to be an MIE within one AOP, regardless of which chemical may cause this event. However, within another AOP, this same event (estrogen binding) may be an intermediate key event as it occurs further downstream in the pathway, such as in one where it is describing a perturbation impacting endogenous circulating estrogen levels at a downstream point in the AOP that is not caused by a direct chemical perturbation. The KERs for a specific pair of key events are also expected to be reusable in other AOPs where those key events may occur.

During the development process, each AOP is a discrete unit described by a developer who may have an in-depth knowledge in the specific area of biology related to the AOP being developed, and as such, the complete picture of

biological complexity is often unknown. While this linear development process allows for less burdensome assembly of AOPs, it is expected that in reality these discrete AOPs actually form a network of inter-related key events, which give an estimate of the actual level of biological complexity that exists in nature. The final principle takes into account the limitations of our knowledge, along with additional data and information that may be gathered over time; therefore, the AOP development process is designed to be an iterative one. Defining AOPs such that they can evolve, be refined, or become modified as new data/evidence becomes available is important for their continued usefulness and applicability for predictive and risk-based toxicology.

### 4.2.1 AOP Knowledge Management

A major component of the OECD AOP development program has been the development of a resource referred to as the AOP knowledgebase (AOP-KB). The AOP-KB has matured as part of the OECD AOP development program and now houses all AOPs developed under the program (<http://www.aopkb.org/>). This publicly accessible repository assists in streamlining the development process by reducing/eliminating duplicate development efforts and allowing for community development. It introduces uniformity in the qualitative and quantitative terms used, thus reducing/eliminating inconsistencies through implementation of a controlled vocabulary, enabling easier and more efficient cross-referencing across AOPs. It also serves as a site to manage/resolve any other AOP knowledge management related issues. The AOP-KB project consists of four independent but connected modules: the AOP-Wiki, AOPXplorer, Effectopedia, and the Intermediate Effects Database (IEDB).

The AOP-Wiki (<http://www.aopwiki.org/>), created by a joint effort between the US EPA and the European Commission's Joint Research Centre, serves as the site for development, organization, and evaluation of formal AOPs. The wiki, which was released in 2014, supports AOP development based on guidelines provided in the OECD AOP development handbook.<sup>18</sup> The wiki allows users to assemble information for developing putative and formal AOPs in a structured format on interconnected wiki pages. The information that can be entered includes graphical and textual descriptions for an AOP, as well as any underlying empirical evidence. As of 2016, the wiki contained more than 100 AOPs with an ever-growing number undergoing formal OECD review.

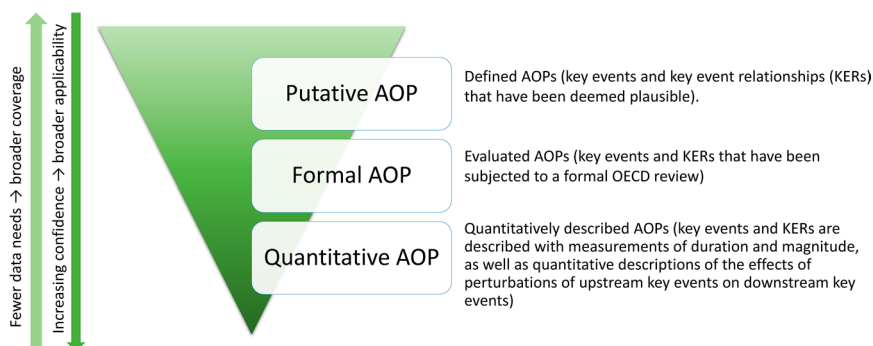
The remaining modules were released in 2016 along with the e.AOP.portal, which aggregates all AOP information from the AOP-KB and provides links to the appropriate modules for additional information (<http://www.aopkb.org/>). Effectopedia ([www.effectopedia.org/](http://www.effectopedia.org/)) is a collaborative platform for modeling and aggregate information for AOP development and use. It provides a graphical interface for the visualization of AOPs based on parameters such as sex, species, level of biological organization, *etc.* It also incorporates information about chemicals and quantitative information about dose and relationship between pairs of key events into the network view. The AOPXplorer (<http://www.aopxplorer.org/>) integrates AOPs from the AOP-Wiki and

Effectopedia with other data, such as information on biological processes or data from computationally predicted AOPs (cpAOPs) (described later). It is designed as a computational tool to facilitate the automated graphical representations of AOPs in a global network context. It gives the user the flexibility of assembling all AOPs relevant for a specific decision-making context in a single network view. Lastly, the IEDB connects chemical-related data that may inform how chemicals or individual compounds give rise to MIEs or key events. These connections, which can be cross-referenced, can also serve as a mechanism for submission of AOPs for risk-assessment purposes.

The knowledge management tools that have been highlighted above create an open-source environment for AOP development, thereby fostering collaborations and helping to continue the refinement process. This highlights a need for controlled vocabularies based on formal ontologies to eliminate/reduce duplicate events or AOPs due to semantic differences in describing the same processes or events. When developed, these ontology-supported controlled vocabularies for the AOP-KB will guide users during the process of describing key terms for their AOPs, thereby allowing for easier cross-referencing. The ontologies will also allow for identifying similar entities that may provide the same function even if the specific object is different. This latter application is particularly important when considering the taxonomic applicability (*i.e.*, ability to extrapolate across different species) of an AOP. This will eliminate redundancies in the network of AOPs in the AOP-KB, highlight connections that may otherwise be missed, and aid any future AOP refinements or revisions.

### 4.2.2 Phases of AOP Development

Three operationally defined stages or phases have been defined for the AOP development process: putative, formal, and quantitative<sup>7</sup> (Figure 4.2). All three phases are expert-driven, resulting in manually curated AOPs. As an



**Figure 4.2** Overlapping phases of adverse outcome pathway (AOP) development and their characteristics. KERs: key event relationships; OECD: Organisation for Economic Co-operation and Development.



AOP progresses through these stages it will include increasing detail about the mechanisms of the key events and KERs contained within the AOP, as well as better coverage, confidence, and applicability of the AOP. A putative AOP is assembled from manually curated information as part of ongoing research or previously published literature. Compared to formal AOPs, the putative AOP requires less time and effort to assemble and can include hypothetical key events and KERs that are based on limited biological or statistical plausibility without a formal assembly and evaluation of the experimental evidence. The formal AOPs are those putative AOPs that have been subjected to a formal OECD review process, and the information included within them meets the international OECD guidance.<sup>17</sup> These AOPs have undergone a more extensive evaluation of the biological plausibility, which will include information from the medical and biological literature rather than just toxicology. The KERs have also undergone a formal review of the empirical evidence, which specifically considers the dose–response and temporal concordance of the upstream and downstream key events in the relationship for the reference chemicals used in defining the KER. Collectively, this information allows users to determine the strength of the evidence available to support the AOP and to understand where confidence gaps may exist in the AOP. Quantitative AOPs are those AOPs assembled to include descriptions that quantify the accuracy and precision, as well as magnitude and/or duration, of how changes in an upstream key event affect a downstream key event. Such quantitative AOPs typically require more data for making these inferences and are usually supported by computational modeling.

The AOP development principles coupled with the AOP knowledge management tools provide tremendous efficiencies over a chemical-by-chemical approach. An AOP can be defined once by incorporating evidence from multiple chemicals and then used for any additional chemical that operates through that pathway based on limited chemical-specific data. Ideally, once the AOP has been extensively characterized, HTT data alone for any given chemical should be sufficient when coupled with the AOP. The phased AOP development process also allows for the prioritization of AOP development to meet the needs of decision makers. In cases where decisions do not require high quantitative precision, the relative impact of creating quantitative AOPs will be low. In cases where the AOP is needed to support quantitative risk assessments based on HTT data alone, the impact of a fully described quantitative AOP is tremendous. By establishing this process, AOP developers can provide limited information for as many scenarios as possible while also providing detailed information for that subset where the impact is greatest. However, the manual curation process for AOP development is time-intensive and laborious. Coverage of the biological space and range of application of AOPs is also limited by the knowledge area of the domain expert.

Computational approaches can be employed to not only accelerate the development of AOPs, but also to widen the biological space covered by these AOPs through data-integration strategies. Computationally driven methods can harness and leverage the large amounts of pre-existing and publicly

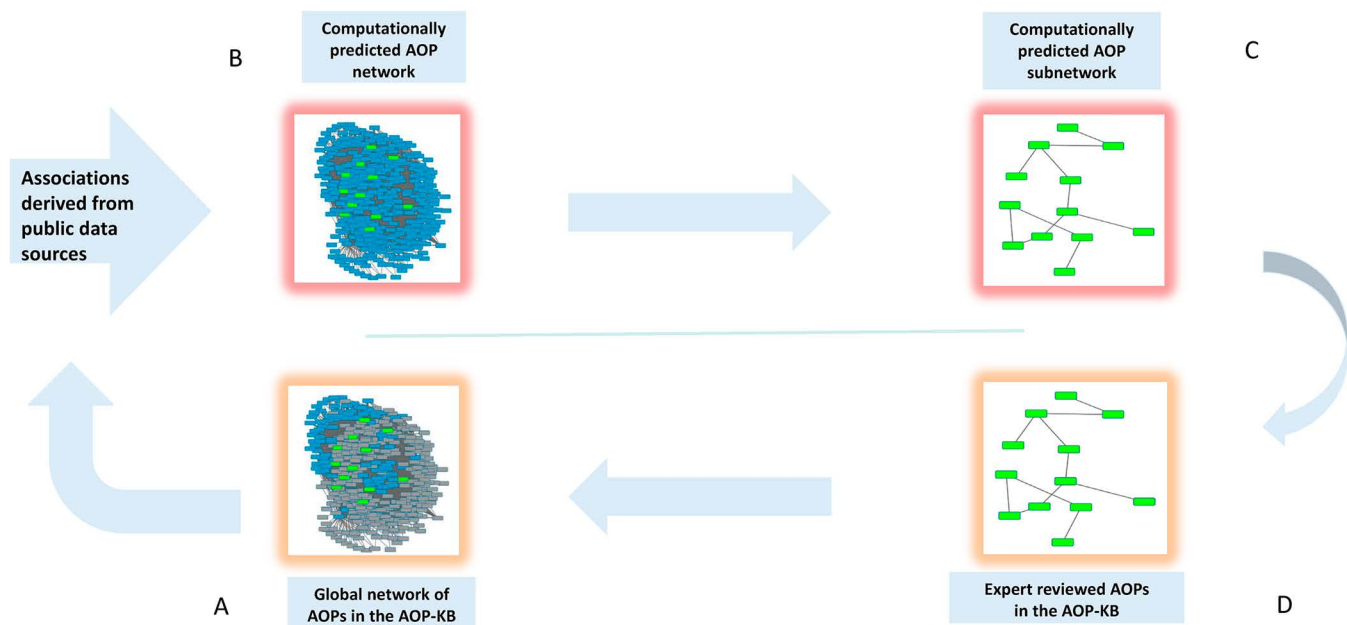


available toxicogenomics data as well as other data types (metabolomics, transcriptomics, biological pathway, *etc.*) for the purpose of AOP generation. These computational methods may be geared towards expanding pre-existing AOPs by incorporating prior knowledge or they may be used for *de novo* AOP generation through data-driven assembly (Figure 4.3). Examples for the former include an AOP for embryonic vascular disruption by 5HPP-33 developed by Kleinstreuer *et al.*,<sup>20</sup> which built upon an existing putative AOP<sup>21</sup> and employed the use of a combination of computational methods, and additional data from ToxCast.<sup>6</sup> Perkins *et al.*<sup>22</sup> also used prior knowledge about the adverse outcome of interest and toxicity pathways to link genes of known function to expand previously undefined toxicity pathways.

A strategy for *de novo* AOP development and assembly has been proposed by Bell *et al.*<sup>23</sup> and Oki and Edwards.<sup>24</sup> The authors generated cpAOP networks through the computational integration of high-throughput screening (HTS) *in vitro* data, gene expression data, biological pathways, clinical markers, pathology, and other *in vivo* phenotype data. Smaller subnetworks of cpAOPs were then extracted from these networks for specified adverse outcomes. Frequent itemset mining<sup>25</sup> was used by both Bell *et al.*<sup>23</sup> and Oki and Edwards<sup>24</sup> to make connections between the different data types representative of different levels of biological organization. The cpAOP networks assume that the events (biological processes or environmental interactions) leading up to an adverse outcome are interconnected sets of entities and, as such, should not be viewed in isolation. This is a more pragmatic view of AOP development, and the networks generated from these analyses may be viewed as an estimate of the universe of information from which subnetworks of specific AOPs may be extracted for expert review. Upon expert review, these cpAOP subnetworks may then serve as scaffolds for putative, formal, or quantitative AOP development, where further review or more targeted testing may be needed (Figure 4.3). These cpAOP networks have demonstrated their potential for accelerating the AOP development process, by providing hypothetical AOPs as a starting point for the development process. It is also expected that this process would be an iterative one, where the networks are continuously refined as new information becomes available and expanded as more data are added (Figure 4.3). Therefore, another important component to this effort is in identifying data-rich sources for AOP development.

### 4.2.3 Data Resources for AOP Development

As many of the computational methods available for AOP development are often data driven, they typically require large amounts of reliable data to be able to make accurate estimates. The cpAOPs produced are a reflection of the data types used; therefore, gaps or shortfalls in the data will have an effect on the coverage of the AOPs. For example, Bell *et al.*<sup>23</sup> and Oki and Edwards<sup>24</sup> assembled cpAOPs using *in vitro* and *in vivo* experimental data in conjunction with information from biological pathway databases. Bell *et al.*<sup>23</sup> developed a cpAOP network for hepatic steatosis using phenotype data from



**Figure 4.3** Iterative process of development and refinement of computationally predicted/generated adverse outcome pathway (cpAOPs) to increase the global network of known AOPs. (A) The global network of possible AOPs. The blue nodes indicate known AOPs while the grey indicate those that are yet to be discovered. Green nodes indicate key events for a hypothetical adverse outcome of interest. (B) Using data mining and integration coupled with other computational approaches, cpAOPs are generated from publicly available data and include more AOPs in the universe of possible AOPs than is currently available in the AOP-knowledgebase (KB). (C) Subnetwork for specific adverse outcome of interest is extracted from larger network of cpAOPs. (D) This subnetwork is subjected to expert review and may be refined upon review. Once this review process is complete and cpAOP is deemed plausible, the cpAOP subnetwork is entered into the AOP-KB as a putative AOP and used to refine the global network of known AOPs.

TG-GATES (Toxicogenomics Project Genomics Assisted Toxicity Evaluation System) rat liver data,<sup>26</sup> Reactome<sup>27</sup> pathways, and HTS assays from the US EPA ToxCast project data.<sup>6</sup> Oki and Edwards also built a cpAOP network connecting assays from the ToxCast project with gene expression data from the Comparative Toxicogenomics Database (CTD),<sup>28</sup> along with curated disease information also from the CTD. In these studies, (1) the HTS *in vitro* data represent events happening at a molecular/cellular level; (2) biological pathway data from Reactome<sup>27</sup> may represent molecular/cellular events; (3) liver pathology and clinical chemistry data may represent tissue/organ level information; and (4) disease data represent events occurring at an organ/individual level of organization.

The resources listed above are just a few of the sources of data that may be employed in building AOPs. Aside from those already mentioned, there are several publicly available data repositories providing a wealth of information at the different biological levels of organization from which key events may be extracted for AOP development. A review by Oki *et al.*<sup>29</sup> lists resources available for AOP development along with the levels of biological organization they cover. A brief list of some of these include transcriptomics, gene expression and metabolomics data incorporated into databases such as ArrayExpress,<sup>30</sup> Gene Expression Omnibus (GEO),<sup>31</sup> and the Human Metabolome Database (HMDB),<sup>32</sup> respectively, from which molecular/cellular level events can be retrieved. Information on chemical initiators or environmental stressors can be found in resources such as DrugBank<sup>33</sup> and PharmGKB,<sup>34</sup> and combining these data with pathway data may give indications of connections occurring for the MIE up to the cellular levels. There are also several phenotype databases (*e.g.* ToxRefDB,<sup>35</sup> PhenomicDB,<sup>36</sup> NHANES (National Health and Nutrition Examination Survey; [www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)), OMIM (Online Mendelian Inheritance in Man),<sup>37</sup> *etc.*) containing *in vivo* data from the tissue up to the individual levels of concern that may be used to provide anchoring for the adverse outcome. Model organism databases such as ZFIN (Zebrafish Model Organism Database)<sup>38</sup> and MGI (Mouse Genome Informatics)<sup>39–41</sup> also contain extensive amounts of *in vitro* and *in vivo* genetic and toxicogenomic data. Given the wealth of information already available for model organisms, species extrapolation also becomes an important part of the AOP development process, as data is pulled from different biological systems.

As with the knowledge management aspect of AOP development, the computational approaches to AOP development would also benefit from the incorporation of ontologies. These ontologies, when included in data-driven development, can increase the efficiency of automation in the process of AOP assembly, as they will help facilitate better data integration between data from different sources. Finally, if the same ontologies are used for assembling computationally predicted AOPs that are used in the assembly of the expert-curated AOPs in the AOP-KB, the expert review and migration of computationally predicted AOPs into the AOP-KB becomes much simpler.

### 4.3 Environmental Exposure and Pharmacokinetic Considerations for Adverse Outcome Development

Inclusion of exposure data is not only important for risk assessment, but is also integral to understanding the public health implications of interactions between environmental factors and genetic factors leading to disease. The field of exposure science focuses on understanding and characterizing the interactions between receptors and stressors of interest. In the 2012 NRC report titled *Exposure Science in the 21st Century: a Vision and a Strategy*, the NRC defined exposure as the amount of a stressor (e.g. chemical) that comes in contact with a receptor (e.g. cell, tissue, soil microbe, building, etc.) without any determination of dose.<sup>42</sup> Exposure assessments involve the estimation or measurement of the magnitude, frequency, and duration of exposure that a receptor might have with a chemical or non-chemical stressor, along with other characteristics (exposure source, exposure routes, time–location activities, etc.) related to the exposed population ([www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf](http://www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf)). To expand upon exposure science and its utility in epidemiological studies in a broader sense, the concept of an “exposome” was proposed, in which every possible exposure is considered for an individual from conception until death.<sup>43</sup> The exposome is comprised of internal mechanisms (e.g. oxidative stress and hormonal signaling), external chemical and non-chemical stressors (e.g. environmental contaminants, radiation, or infectious microbes), lifestyle activities (e.g. smoking, drinking, or occupation), and socio-economic status (e.g. education, financial stability, or mental state). It is thought that integrating these various factors can help elucidate the link between non-genetic factors and disease. Thus, a weight-of-evidence approach can aid in the assessment of exposure–disease causality relationships, and facilitate disease prevention through mediation of identified key contributors of that disease.

A host of complex diseases are influenced by environmental factors, and biologically relevant exposures should be studied in order to better understand the contributions of the environment to disease etiology. The strategy outlined by the NRC in the prior 2007 report<sup>4</sup> highlighted the need for human exposure and population-based studies to provide data for understanding the following: (1) perturbations in cellular response networks and toxicity pathways; (2) how susceptibility and background exposures can be used for interpreting and extrapolating *in vitro* results; and (3) appropriate dose selection for environmentally relevant exposures. Exposure science is no longer primarily composed of hypothesis-driven research that investigates exposures of chemicals that are already known to be problematic throughout the environment. Rather, advances in remote sensors and monitoring instrumentation, as well as high-throughput -omics approaches, have allowed untargeted exposure analysis to facilitate exploratory research, whereby focus can be placed on investigating chemicals that might pose

risks previously unidentified. In addition, biomonitoring and population health surveillance through population-based studies can help in identifying biomarkers, health risks, and potentially important exposures not previously found through toxicity testing. There are several examples of adverse outcomes that are not detected in the lab or during clinical testing but that are discovered through population surveillance, especially in relation to pharmaceutical exposures.<sup>44,45</sup> In other cases, the agent is tested and found to be safe under certain circumstances (*e.g.* oral exposures), but may still pose risks if exposure occurs through a different untested (*e.g.* inhalation) or unknown route. Such was the case in 2000, when several instances of a severe and rare pulmonary disorder now referred to as “popcorn lung” were reported among workers in a food flavoring processing and manufacturing facility.<sup>46</sup>

As mentioned above, traditional toxicity testing using whole animal models in well controlled experimental conditions offers several advantages, such as providing direct measurement of dose–response for apical endpoints relevant for risk assessment and including ADME processes for the test species. As we move to reduce the use of animals in toxicity testing and shift to more sustainable *in silico* and *in vitro* based methods (*e.g.* HTT screening or transcriptomic technologies), extrapolation and scaling up of *in vitro* results to biologically effective contexts becomes necessary, including use of our knowledge regarding chemical-specific exposure and ADME information. For example, even though a chemical appears inactive in an *in vitro* assay, it may be metabolized within a living system to an active moiety that is capable of perturbing a molecular target. Conversely, those chemicals that appear active in the same *in vitro* assay may have limited exposure or may possess ADME properties that prevent them from reaching the *in vivo* molecular target. High-throughput screening assays may be used to identify potential hazards,<sup>47,48</sup> while *in silico* models can be used to estimate exposure,<sup>49,50</sup> and finally quantitative structure–activity relationship (QSAR) models may be utilized for predicting or evaluating ADME properties of a chemical.<sup>47</sup>

The most comprehensive approach to risk assessment involves the traditional paradigm of investigating only one to a few chemicals at a time, which allows for acquisition of large amounts of exposure and/or ADME data. Because this paradigm is no longer sustainable, exposure and ADME information must be obtained as rapidly as possible for hundreds to thousands of chemicals. The rapid measurement and characterization of ADME parameters will likely require computational chemistry methods such as read-across, QSAR, and models for extrapolation from cell to tissue and between organisms. When evaluating the risk of a chemical based on its exposure or ADME properties, the approaches that can be taken will depend on the quality and availability of data, as well as the needs of the investigator.

Qualitative screening can be conducted when data are sparse or when researchers are primarily interested in selecting chemicals of higher priority based on their exposure and ADME properties in addition to bioactivities.<sup>51</sup> Those chemicals that lack optimal exposure or ADME characteristics to allow them to interact with a molecular target *in vivo* would have a lower priority in

regard to carrying them on to more extensive *in vivo* testing. In data-poor situations, QSAR models and exposure predictions can reveal whether exposure and ADME properties are optimal for reaching the *in vivo* target,<sup>52,53</sup> *i.e.* if the chemical is of the right solubility or charge for absorption and distribution or if the chemical can penetrate the blood–brain barrier. Such models can also be used to identify those chemicals inactive *in vitro* that may be of risk *in vivo* due to bioactivation or because of issues related to assay design that limit binding to the technological target.

As data becomes more abundant, pharmacokinetic, PBPK, and pharmacodynamic models can be built to incorporate chemical-specific *in vitro* dose–response data, along with exposure and ADME data to estimate *in vivo* toxicological activity through *in vitro* to *in vivo* extrapolation. These activities can then be ranked after sorting into priority bins of increasing concern.<sup>54</sup> Alternatively, the concentrations that are able to elicit a biological response *in vivo* can be linked to external concentrations *via* reverse dosimetry to derive a point of departure that will result in biologically effective internal concentrations. This point-of-departure concentration can then be compared to predicted or actual population exposure levels in order to estimate a margin of exposure.<sup>14,47,55</sup> Elucidating the ADME and exposure characteristics that allow for triggering of the MIE is critical in understanding the hazard posed by chemicals in order to inform risk-based decision making as it relates to application of AOPs.

## 4.4 The AEP Framework: An Organizing Principle for Exposure Data

Through emerging technologies and techniques, new streams of data are now being generated at higher volumes and speed so as to allow exposures to be more accurately characterized for all endogenous and exogenous chemical and non-chemical stressors throughout the lifetime of an individual (*i.e.* the exposome). The evolution of exposure science also offers new opportunities to transition risk assessment approaches from investigating a specific exposure source that leads to a certain disease, to understanding the intricate interactions among chemical and non-chemical stressors, the environment, and biology in the real world. To enable these opportunities and to integrate exposure science data with human and environmental health communities, the AEP framework was conceptualized.<sup>16</sup> The AEP framework organizes the increasing flow of dispersed exposure information, and also provides common nomenclature to facilitate communications and innovations and interoperability across numerous resources (*e.g.* multi-media exposure models, conceptual site models, and exposure databases).

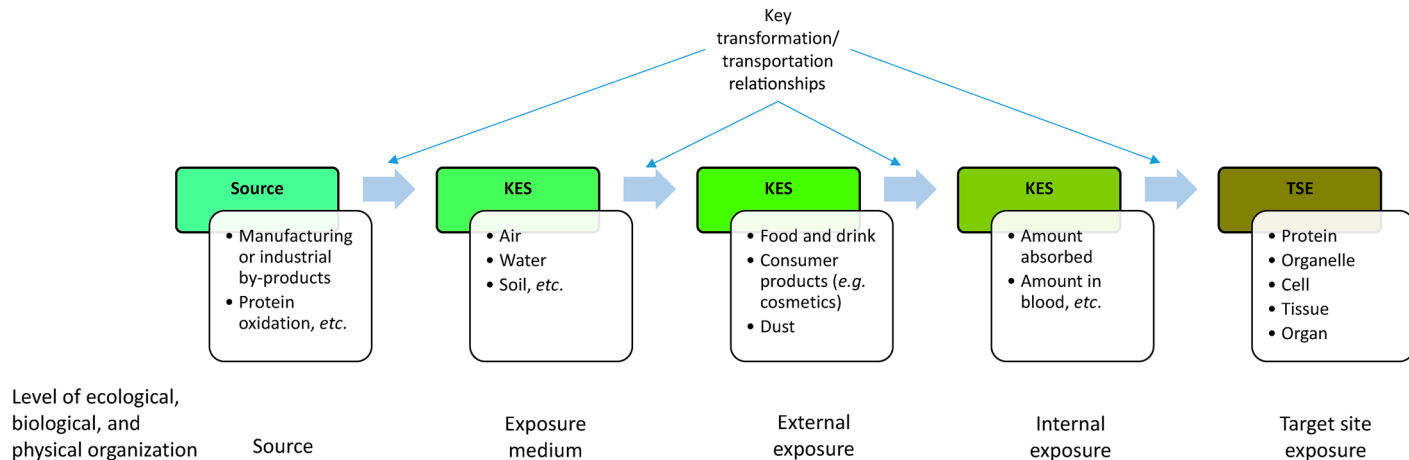
Just as the AOP framework organizes knowledge on biologically plausible and empirically supported links between perturbations on a biological system at the molecular level to adverse outcomes, the AEP framework offers an intuitive approach to organize exposure data and predictions. By linking data

and predictions from sources to concentrations at internal target sites, AEPs serve as natural companions to AOPs, and can help support exposure assessments, risk assessments, and epidemiological studies. The coupling of these frameworks will ensure that biologically relevant exposure information is integrated into toxicity testing strategies; and effects of multiple, cumulative, and chronic low dose exposures can be evaluated. Together, the AEP and AOP concepts form a flexible framework that allows for risk-based, hazard-based, and exposure-based decision making.

The basic building blocks of an AEP are derived from the general scheme of a maturing AOP. Similar to an AOP, an AEP includes two key components: key exposure states (KES) describe measurable, obligate steps that occur throughout the pathway, and key transformation/transportation relationships (KTR) describe the linkages between each pair of KESs. An AEP begins with the source(s) of a stressor and terminates at the target site exposure (TSE), which is a specialized KES (Figure 4.4). These components of the AEP framework are defined by Teeguarden *et al.*<sup>16</sup> as follows: the KES is a measurable change in a chemical state or concentration that is essential, but not necessarily sufficient for the delivery of a chemical to the TSE. It is represented by a node in the AEP network, and it provides verifiability to an AEP description (*e.g.* increased mycotoxins in corn). The KTR, in addition to describing the transition from one KES to the next, also establishes the order of the events and is supported by chemical, physical, and biological plausibility, as well as empirical evidence. It is represented as an arrow between nodes within the AEP and is a unit of inference within the AEP (*e.g.* production of aflatoxin by the fungus *Aspergillus flavus*). The TSE connects the AEP and the AOP by describing the concentration and timing or duration of a stressor at a site of action that corresponds to an MIE. These measured or predicted concentrations at the TSE can aid in informing the relevance of biological effects observed or expected at the target site and at downstream levels of biological organization related to the AOP(s) resulting from exposure to a stressor.

By describing the path of a stressor from its source to its internal site of action, at which the MIE occurs that triggers an AOP, integration of AOPs with AEPs provides a natural linkage from source to outcome. Teeguarden *et al.*<sup>16</sup> state that the TSE is measured at the level of biological organization corresponding to a defined protection goal, and that it may be biotic (*e.g.* concentration of lead in bone) or abiotic (*e.g.* concentration of methane in air) depending on the goal of the assessment. Unlike an AOP, AEPs are chemical-centric, though not necessarily chemical-specific, as they do explicitly incorporate the chemical properties in the definition. The aggregation of such chemical-centric AEPs can lead to a network of interacting exposure pathways that share common KESs, TSEs, and in turn, link to common MIEs from AOPs (Figure 4.5). Elements of AEPs also can potentially be chemically agnostic and reusable, such as KTRs that describe transformation processes between more than one type of media (*e.g.* evaporation out of a large lake into the surrounding air) that can apply to multiple chemicals.





**Figure 4.4** Components of the aggregate exposure pathway framework along with the levels of biological, ecological, and physical organization. KES: key exposure states; TSE: target site of exposure.





**Figure 4.5** Integration of the aggregate exposure pathway and adverse outcome pathway frameworks to complete the source to outcome continuum. MIE: molecular initiating event; AO: adverse outcome. Adapted with permission from J. G. Teeguarden, Y.-M. Tan, S. W. Edwards, J. A. Leonard, K. A. Anderson, R. A. Corley, A. K. Harding, M. L. Kile, S. M. Simonich, D. Stone, R. L. Tanguay, K. M. Waters, S. L. Harper and D. E. Williams. Completing the link between exposure science and toxicology for improved environmental health decision making: the aggregate exposure pathway framework, *Environmental Science & Technology*, 2016, 50, 4579–4586. Copyright 2016 American Chemical Society.

The applications for AEPs in human and environmental health are numerous. Use of the AEP framework can aid in determining and prioritizing critical or common exposure pathways and predicting chemical concentrations and transformations within and across biological and physical components of the pathway. Source-mitigation strategies can be developed through combined use of AOPs and AEPs that identify key sources, environmental processes, and exposure routes that lead to internal concentrations at target sites that are linked to AOPs of regulatory concern. In addition, new target sites of concern can be identified that can support AOPs in need of development. Although there are several exposure models in existence (see, for example<sup>49,50</sup>), their availability and capabilities may be unknown due to a scattering across the literature. AEPs offer an organizational framework, through modular descriptions of exposure pathways, that researchers can use to share and integrate data and predictions that are generated from each model resource. In doing so, data gaps can then be identified so that focus can be placed on acquiring the most relevant substance-specific data, which are often costly to generate. Cumulative risk assessment of multiple chemicals acting through a common AOP can be accomplished by determining concentrations and transformation processes throughout their individual AEPs (e.g. concentrations of perfluorinated compounds in clothing, furniture, or house dust). More importantly, by completing the source-to-outcome continuum through linkages between the AEP and AOP frameworks, the integration of toxicity testing information with exposure information can lead to opportunities for developing computational tools that enable more rapid and comprehensive assessment of exposure and hazard for informing risk assessment strategies.<sup>56</sup>

The new paradigms of high-throughput exposure predictions and HTT testing in the 21st century can benefit significantly from the integration of

AEPs with AOPs. Computational models able to predict external exposure concentrations can be linked to pharmacokinetic and PBPK models to determine internal concentrations in blood or tissue. Comparison of these TSE concentrations with high-throughput *in vitro* assay concentrations that are shown to trigger an MIE can then allow for chemical/stressor prioritization based on the margin of exposure given an HTT and the TSE. For example, the topical agent anthralin is used to treat psoriasis<sup>57</sup> and is found to be active in an assay related to acetylcholinesterase (AChE) inhibition in the US EPA's ToxCast program. Though anthralin is easily absorbed into the skin, it is highly lipophilic and cannot enter into the systemic circulation.<sup>57</sup> Therefore, AOPs associated with molecular targets in the skin (*e.g.* mitochondria of keratinocytes) are likely of more concern than AOPs that have an internal target site of action, *i.e.* the TSE, in systemic tissues (*e.g.* the brain). Thus, AEP–AOP integration is informative regarding the plausibility and appropriateness of an AOP for a specific stressor, by providing essential information on the potential (qualitative) or amounts (quantitative) of the stressor that might reach the TSE and lead to an MIE. In addition, AEPs that characterize both environmental (*e.g.* oxidation or reduction) transformations and biological metabolism of parent compounds could aid in identifying the complete set of metabolites and degradation products for subsequent bioactivity/toxicity testing. Even if a parent compound or its metabolites are not considered to initiate the primary AOP of interest, knowledge of their ability to reach the TSE can inform testing for alternative AOPs that are related to that TSE. For example, chemicals that are able to reach the brain but whose exposure does not lead to AChE inhibition may still be capable of triggering other MIEs associated at that site, such as binding to nicotinic cholinergic receptors or sodium-gated ion channels.

#### 4.4.1 Data resources for AEP development

The recent advances and innovations in exposure science are enabling a shift to a distinct discipline that revolutionizes toxicity testing strategies and conventional risk assessment by facilitating AEP development. Using emerging monitoring instrumentation, large numbers of chemicals can now be detected at extremely low concentrations in environmental and biological samples, providing relevant exposure information that can be integrated into toxicity testing strategies. Novel biomarkers are being developed to inform our understanding of exposures, early biological effects, and susceptibility to chemical and non-chemical stressors leading to adverse outcomes. One such example is in how biomarkers in tooth layers are able to provide information regarding prenatal and early life exposures to reconstruct and establish relationships between these exposure episodes and neurodevelopmental outcomes in children.<sup>58,59</sup> Non-targeted approaches (*e.g.* chemical analysis of house dust samples) can also be used to advance our understanding of the human exposome by serving as indicators of indoor exposure and direct exposure.<sup>60</sup> In addition to exposure characterization, these advancements in

the exposure science field allow for integration of exposure data with hazard information obtained from *in vitro* toxicity screening assays and/or disease outcomes found in epidemiologic studies to provide insight into the complex relationship between multiple exposures and multiple health outcomes.

#### **4.5 AEP–AOP Integration for Linking Toxicity to Exposure: Applications of the AOP and AEP Frameworks for Risk Assessments and Chemical Management Decision Making**

The potential for a chemical or other environmental agent to adversely impact human health is a source of concern for many regulatory agencies and the general public. The ever-increasing amounts of chemicals being used in commerce require more efficient approaches to testing and risk assessment. Increasing public awareness of the potential public health hazards posed by these chemicals, along with societal demands for safer chemicals and reduced animal testing, have acted as catalysts and prompted the shift in how toxicity testing and public health risk assessments are undertaken. The old model is unable to balance the expectation of fewer animal tests while also keeping pace with the number of chemicals for which testing is needed. This problem presents a challenge and opportunity for toxicologists and risk assessors to develop methods that can meet both requirements. The AOP and AEP frameworks, when coupled with the toxicity pathway concept and HTT, provide immense value to predictive toxicology, risk-based toxicology, and pharmacology. This value comes from the ability of these frameworks to bridge the gap between exposure and toxicity by providing the necessary linkages connecting events from the source of exposure, up to a target site exposure, and on through to a resulting adverse outcome.

Similar to the MOA framework that has been incorporated by regulatory agencies to allow for the inclusion of mechanistic information in risk assessments,<sup>61</sup> the AOP framework is also being considered for use in several areas of risk-based toxicology. These include chemical grouping and screening, prioritization for further testing, and read-across, as well as hazard identification and risk assessment. The chemically agnostic nature of AOPs allows them to be applied for chemical grouping purposes and read-across, thereby allowing for more efficient classification of chemicals that are not as well characterized with well-characterized chemicals of similar qualities/properties. This facilitates identification of chemicals that may present public health hazards, such as carcinogenicity or reproductive effects, through examination of shared pathways leading to the same adverse outcomes in the AOP network. The regulatory application and decision context both play a significant role when considering the potential use of an AOP, as well as the amount of evidence for and confidence in its required components. Another driving force behind AOP development is the recognition of its potential to provide support for mechanism-based risk assessment, which may reduce

the need for whole organism tests, as individual key events and KERs may be tested systematically *in vitro* or *in silico*. The AOP structure is also flexible enough to allow for its application in situations of abundant or limited data, and in the case of the later situation, it will allow for easier and more efficient identification of important data gaps or needs thereby prompting the development of relevant assays and tests.

Integrated approaches to testing and assessment (IATA) have been proposed as a vehicle for integrating AOPs for use in risk assessments and decision making. Tollefsen *et al.*<sup>19</sup> described an IATA as a structured approach that gathers, integrates, and weighs different types of data for the purposes of hazard identification, hazard characterization, and safety assessment for a chemical or chemical group. Given sufficient information, an IATA-informed AOP may be applied in making risk-based decisions or in driving hypothesis generation and the development of new tests in order to make such risk-based decisions. Perkins *et al.*,<sup>62</sup> in their examination and analysis of various AOPs with differing levels of completeness and scientific confidence, also highlighted the utility of AOPs for predictive toxicology, chemical prioritization, read-across, hazard assessment and identification, and risk assessment. Guidelines have been developed for the application of AOPs for chemical screening and prioritization, ecological hazard and risk assessment, and for IATA. Examples of AOPs that have been utilized for these purposes include application of the skin sensitization AOP, membrane disruption leading to respiratory failure, and growth impairment in fish.

Skin sensitization may be defined as the potential of a substance to cause dermatitis or an allergic skin reaction on contact with the skin. The skin sensitization AOP published by the OECD<sup>63</sup> is an example of a well-studied endpoint for which abundant data were available for AOP development. As a practical application of this AOP for hazard identification and assessment, Patlewicz *et al.*<sup>64</sup> proposed an IATA for skin sensitization using already existing information available for the substances tested along with data from computational approaches such as QSAR. The IATA workflow developed may be used to evaluate the skin sensitization potential of a substance or to identify what extra data may be needed to make a decision or eliminate uncertainty in the process. Jaworska *et al.*<sup>65</sup> also developed a quantitative approach to assess skin sensitization by employing the use of a Bayesian integrated testing strategy and a murine local lymph node assay (LLNA). While the Jaworska *et al.* approach originated independently of the skin sensitization AOP, their approach in effect uses the LLNA assay as a key event to give a quantitative measure to prediction/identification of skin sensitization hazards. This may be then extrapolated for human risk assessment, given exposure information (including site of exposure and duration of exposure) and consideration of other factors, such as variation in response sensitivity across individuals and populations.

Perkins *et al.*<sup>62</sup> have also elucidated ways in which AOPs with differing degrees of completeness and scientific confidence may be used for chemical management applications. One such example of the application of a low confidence AOP is in the utilization of the membrane disruption (narcosis)

leading to respiratory failure AOP for chemical categorization, prioritization, and read-across. Despite the low confidence and incompleteness (this AOP has a general MIE and the exact nature of the relationship between the key events leading to the adverse outcome are not fully characterized), the narcosis AOP still contained events which were testable by assays and could be linked statistically to the adverse outcome.<sup>62</sup> The AOP was also used to develop a QSAR that is the basis for model testing within the US EPA's EPI Suite software designed for compliance under the Toxic Substances Control Act.<sup>66</sup> This narcosis model, in addition to the skin sensitization models, has also been incorporated into the OECD QSAR application toolbox<sup>66</sup> for implementing the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)<sup>3</sup> regulations of the European Union.

In an ecotoxicology application, Groh *et al.*<sup>14</sup> developed AOPs to illustrate toxicity mechanisms leading to growth impairment in fish. Their work highlighted different pathways taken by three agents (pyrethroids, cadmium, and selective serotonin reuptake inhibitors) leading to the same apical endpoint of growth impairment. They were able to identify predictive key events for which further testing is needed, in addition to those for which alternative and more refined tests improving animal welfare could be developed. Their work also highlighted missing exposure data needs, including information (such as exposure routes, concentrations, and duration) that may be provided by having a companion AEP and that may lead to better-defined effects and allow for better extrapolation across species. The area of species extrapolation, in particular, benefits from this exposure information. In addition, a better mechanistic understanding of the biological systems underlying the key events and KERs can be used to highlight ADME differences between species, thus elucidating the different effects in observed in differing species.

## 4.6 Conclusions and Future Directions

Effective integration of toxicological and exposure science through the adoption of the AEP and AOP frameworks leads to a more holistic approach to toxicity testing, thereby shrinking the knowledge gap that currently exists between exposure and toxicity outcomes. The AEP–AOP continuum offers a valuable framework for aggregating the different biological (mechanistic, qualitative, quantitative, *etc.*) and exposure data available for establishing these linkages and organizes them in a manner to best support risk-based decisions. These approaches also put into consideration the non-linear nature of the biological processes that may be involved in the manifestation of an adverse outcome and allow for the integration of multiple pathways and networks of exposure related events leading to the same adverse outcome. The implementation of these frameworks will be important advancements for hazard identification, risk assessment, and risk characterization, which translate to better decisions from regulators and improved outcomes for public and environmental health.

## References

1. W. M. S. Russell and R. L. Burch, *The Principles of Humane Experimental Technique*, Methuen, 1959.
2. United States Congress, *ICCVAM Authorization Act of 2000*, 2000, cited 2015 08/03/2015, available from: <https://www.congress.gov/bill/106th-congress/house-bill/4281>.
3. European Commission, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), *Off. J. Eur. Communities: Inf. Not.*, 2006, **EC 1907/2006**, .
4. National Research Council (NRC), *Toxicity Testing in the 21st Century: A Vision and a Strategy*, The National Academies Press, Washington, DC, 2007.
5. M. S. Attene-Ramos, N. Miller, R. Huang, S. Michael, M. Itkin, R. J. Kavlock, C. P. Austin, P. Shinn, A. Simeonov, R. R. Tice and M. Xia, The Tox21 robotic platform for the assessment of environmental chemicals – from vision to reality, *Drug Discovery Today*, 2013, **18**, 716–723.
6. R. S. Judson, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah, A. M. Richard and D. J. Dix, In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project, *Environ. Health Perspect.*, 2010, **118**, 485–492.
7. D. L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. H. Hutchinson, C. A. LaLone, B. Landesmann, T. Lettieri, S. Munn and M. Nepelska, Adverse outcome pathway (AOP) development I: strategies and principles, *Toxicol. Sci.*, 2014, **142**, 312–320.
8. C. Sonich-Mullin, R. Fielder, J. Wiltse, K. Baetcke, J. Dempsey, P. Fenner-Crisp, D. Grant, M. Hartley, A. Knaap, D. Kroese, I. Mangelsdorf, E. Meek, J. M. Rice and M. Younes, IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis, *Regul. Toxicol. Pharmacol.*, 2001, **34**, 146–152.
9. M. E. Meek, A. Boobis, I. Cote, V. Dellarco, G. Fotakis, S. Munn, J. Seed and C. Vickers, New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis, *J. Appl. Toxicol.*, 2014, **34**, 1–18.
10. A. R. Boobis, J. E. Doe, B. Heinrich-Hirsch, M. E. Meek, S. Munn, M. Ruchirawat, J. Schlatter, J. Seed and C. Vickers, IPCS Framework for Analyzing the Relevance of a Noncancer Mode of Action for Humans, *Crit. Rev. Toxicol.*, 2008, **38**, 87–96.
11. R. A. Becker, G. T. Ankley, S. W. Edwards, S. Kennedy, I. Linkov, B. Meek, M. Sachana, H. Segner, B. Van Der Burg and D. L. Villeneuve, Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence, *Regul. Toxicol. Pharmacol.*, 2015.



12. G. T. Ankley, R. S. Bennett, R. J. Erickson, D. J. Hoff, M. W. Hornung, R. D. Johnson, D. R. Mount, J. W. Nichols, C. L. Russom and P. K. Schmieder, Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment, *Environ. Toxicol. Chem.*, 2010, **29**, 730–741.
13. D. L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. H. Hutchinson, C. A. LaLone, B. Landesmann, T. Lettieri, S. Munn and M. Nepelska, Adverse outcome pathway development II: Best practices, *Toxicol. Sci.*, 2014, **142**, 321–330.
14. J. E. Simmons, M. V. Evans and W. K. Boyes, Moving From External Exposure Concentration to Internal Dose: Duration Extrapolation Based on Physiologically Based Pharmacokinetic Derived Estimates of Internal Dose, *J. Toxicol. Environ. Health, Part A*, 2005, **68**, 927–950.
15. J. Stadnicka-Michalak, K. Tanneberger, K. Schirmer and R. Ashauer, Measured and Modeled Toxicokinetics in Cultured Fish Cells and Application to *In Vitro* – *In Vivo* Toxicity Extrapolation, *PLoS One*, 2014, **9**, e92303.
16. K. J. Groh, R. N. Carvalho, J. K. Chipman, N. D. Denslow, M. Halder, C. A. Murphy, D. Roelofs, A. Rolaki, K. Schirmer and K. H. Watanabe, Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology, *Chemosphere*, 2015, **120**, 764–777.
17. P. J. Lioy and K. R. Smith, A Discussion of Exposure Science in the 21st Century: A Vision and a Strategy, *Environ. Health Perspect.*, 2013, **121**, 405–409.
18. J. G. Teeguarden, Y.-M. Tan, S. W. Edwards, J. A. Leonard, K. A. Anderson, R. A. Corley, A. K. Harding, M. L. Kile, S. M. Simonich, D. Stone, R. L. Tanguay, K. M. Waters, S. L. Harper and D. E. Williams, Completing the Link between Exposure Science and Toxicology for Improved Environmental Health Decision Making: The Aggregate Exposure Pathway Framework, *Environ. Sci. Technol.*, 2016, **50**, 4579–4586.
19. OECD Organisation for Economic Co-operation and Development, in *Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Organisation for Economic Co-operation and Development*, ed. P. A. B. Environment Directorate, Paris, FR, 184th edn, 2013.
20. OECD Organisation for Economic Co-operation and Development, *Users' Handbook Supplement to the Guidance Document for Developing and Assessing Aops*, [https://aopkb.org/common/AOP\\_Handbook.pdf](https://aopkb.org/common/AOP_Handbook.pdf), accessed 7/13/2015, 2014.
21. K. E. Tollefsen, S. Scholz, M. T. Cronin, S. W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth and G. Patlewicz, Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA), *Regul. Toxicol. Pharmacol.*, 2014, **70**, 629–640.
22. N. Kleinstreuer, D. Dix, M. Rountree, N. Baker, N. Sipes, D. Reif, R. Spencer and T. Knudsen, A Computational Model Predicting Disruption of Blood Vessel Development, *PLoS Comput. Biol.*, 2013, **9**, e1002996.

23. T. B. Knudsen and N. C. Kleinstreuer, Disruption of embryonic vascular development in predictive toxicology, *Birth Defects Res., Part C*, 2011, **93**, 312–323.
24. E. J. Perkins, J. K. Chipman, S. Edwards, T. Habib, F. Falciani, R. Taylor, G. Van Aggelen, C. Vulpe, P. Antczak and A. Loguinov, Reverse engineering adverse outcome pathways, *Environ. Toxicol. Chem.*, 2011, **30**, 22–38.
25. S. M. Bell, M. M. Angrish, C. E. Wood and S. W. Edwards, Integrating Publicly Available Data to Generate Computationally Predicted Adverse Outcome Pathways for Fatty Liver, *Toxicol. Sci.*, 2016, **150**, 510–520.
26. N. O. Oki and S. W. Edwards, An integrative data mining approach to identifying adverse outcome pathway signatures, *Toxicology*, 2016, **350–352**, 49–61.
27. R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, *ACM SIGMOD Rec.*, 1993, **22**, 207–216.
28. Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani and H. Yamada, Open TG-GATEs: a large-scale toxicogenomics database, *Nucleic Acids Res.*, 2015, **43**, D921–D927.
29. D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein and P. D'Eustachio, The Reactome Pathway Knowledgebase, *Nucleic Acids Res.*, 2014, **42**, D472–D477.
30. A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegiers and C. J. Mattingly, The Comparative Toxicogenomics Database's 10th Year Anniversary: Update 2015, *Nucleic Acids Res.*, 2015, **43**, D914–D920.
31. N. O. Oki, M. D. Nelms, S. M. Bell, H. M. Mortensen and S. W. Edwards, Accelerating Adverse Outcome Pathway Development Using Publicly Available Data Sources, *Curr. Environ. Health Rep.*, 2016, **3**, 53–63.
32. N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans and A. Brazma, ArrayExpress update—simplifying data submissions, *Nucleic Acids Res.*, 2015, **43**, D1113–D1116.
33. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis and A. Soboleva, NCBI GEO: archive for functional genomics data sets—update, *Nucleic Acids Res.*, 2013, **41**, D991–D995.
34. D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorn Dahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert, HMDB 3.0—The Human Metabolome Database in 2013, *Nucleic Acids Res.*, 2013, **41**, D801–D807.



35. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
36. M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman and T. E. Klein, Pharmacogenomics Knowledge for Personalized Medicine, *Clin. Pharmacol. Ther.*, 2012, **92**, 414–417.
37. M. T. Martin, R. S. Judson, D. M. Reif, R. J. Kavlock and D. J. Dix, Profiling chemicals based on chronic toxicity results from the US EPA ToxRef Database, *Environ. Health Perspect.*, 2009, **117**, 392–399.
38. P. Groth, N. Pavlova, I. Kaley, S. Tonov, G. Georgiev, H.-D. Pohlenz and B. Weiss, PhenomicDB: a new cross-species genotype/phenotype resource, *Nucleic Acids Res.*, 2007, **35**, D696–D699.
39. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res.*, 2005, **33**, D514–D517.
40. D. G. Howe, Y. M. Bradford, T. Conlin, A. E. Eagle, D. Fashena, K. Frazer, J. Knight, P. Mani, R. Martin and S. A. T. Moxon, ZFIN, the Zebrafish Model Organism Database: Increased Support for Mutants and Transgenics, *Nucleic Acids Res.*, 2013, **41**, D854–D860.
41. J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson and G. The Mouse Genome Database, The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease, *Nucleic Acids Res.*, 2015, **43**, D726–D736.
42. C. M. Smith, J. H. Finger, T. F. Hayamizu, I. J. McCright, J. Xu, J. Berghout, J. Campbell, L. E. Corbani, K. L. Forthofer, P. J. Frost, D. Miers, D. R. Shaw, K. R. Stone, J. T. Eppig, J. A. Kadin, J. E. Richardson and M. Ringwald, The mouse Gene Expression Database (GXD): 2014 update, *Nucleic Acids Res.*, 2014, **42**, D818–D824.
43. C. M. Smith, J. H. Finger, T. F. Hayamizu, I. J. McCright, J. Xu, J. T. Eppig, J. A. Kadin, J. E. Richardson and M. Ringwald, GXD: a community resource of mouse Gene Expression Data, *Mamm. Genome*, 2015.
44. National Research Council (NRC), *Exposure Science in the 21st Century: A Vision and a Strategy*, The National Academies Press, Washington, D.C., 2012.
45. C. P. Wild, Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology, *Cancer Epidemiol., Biomarkers Prev.*, 2005, **14**, 1847–1850.
46. J. Lexchin, Drug withdrawals from the Canadian market for safety reasons, 1963–2004, *Can. Med. Assoc. J.*, 2005, **172**, 765–767.
47. IOM (Institute of medicine), *The Future of Drug Safety: Promoting and Protecting the Health of the Public*, The National Academies Press, Washington, DC, 2007.

48. K. Kreiss, A. Gomaa, G. Kullman, K. Fedan, E. J. Simoes and P. L. Enright, Clinical Bronchiolitis Obliterans in Workers at a Microwave-Popcorn Plant, *N. Engl. J. Med.*, 2002, **347**, 330–338.
49. R. S. Judson, R. J. Kavlock, R. W. Setzer, E. A. Cohen Hubal, M. T. Martin, T. B. Knudsen, K. A. Houck, R. S. Thomas, B. A. Wetmore and D. J. Dix, Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment, *Chem. Res. Toxicol.*, 2011, **24**, 451–462.
50. M. Vinken, The adverse outcome pathway concept: A pragmatic tool in toxicology, *Toxicology*, 2013, **312**, 158–165.
51. K. K. Isaacs, W. G. Glen, P. Egeghy, M.-R. Goldsmith, L. Smith, D. Valero, R. Brooks, C. M. Grulke and H. Özkaynak, SHEDS-HT: An Integrated Probabilistic Exposure Model for Prioritizing Exposures to Chemicals with Near-Field and Dietary Sources, *Environ. Sci. Technol.*, 2014, **48**, 12750–12759.
52. J. F. Wambaugh, A. Wang, K. L. Dionisio, A. Frame, P. Egeghy, R. Judson and R. W. Setzer, High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals, *Environ. Sci. Technol.*, 2014, **48**, 12760–12767.
53. M. B. Phillips, J. A. Leonard, C. M. Grulke, D. T. Chang, S. W. Edwards, R. Brooks, M.-R. Goldsmith, H. El-Masri and Y.-M. Tan, A Workflow to Investigate Exposure and Pharmacokinetic Influences on High-Throughput In Vitro Chemical Screening Based on Adverse Outcome Pathways, *Environ. Health Perspect.*, 2016, **124**, 53–60.
54. H. Sun, Predicting ADMET Properties by Projecting onto Chemical Space? Benefits and Pitfalls, *Curr. Comput.-Aided Drug Des.*, 2005, **1**, 179–193.
55. J. K. Wegner, H. Fröhlich, H. M. Mielenz and A. Zell, Data and Graph Mining in Chemical Space for ADME and Activity Data Sets, *QSAR Comb. Sci.*, 2006, **25**, 205–220.
56. J. A. Leonard, A. Sobel Leonard, D. T. Chang, S. Edwards, J. Lu, S. Scholle, P. Key, M. Winter, K. Isaacs and Y.-M. Tan, Evaluating the Impact of Uncertainties in Clearance and Exposure When Prioritizing Chemicals Screened in High-Throughput Assays, *Environ. Sci. Technol.*, 2016, **50**, 5961–5971.
57. J. A. Leonard, Y.-M. Tan, M. Gilbert, K. Isaacs and H. El-Masri, Estimating Margin of Exposure to Thyroid Peroxidase Inhibitors Using High-Throughput in vitro Data, High-Throughput Exposure Modeling, and Physiologically Based Pharmacokinetic/Pharmacodynamic Modeling, *Toxicol. Sci.*, 2016, **151**, 57–70.
58. E. A. Cohen Hubal, A. Richard, L. Aylward, S. Edwards, J. Gallagher, M.-R. Goldsmith, S. Isukapalli, R. Tornero-Velez, E. Weber and R. Kavlock, Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment, *J. Toxicol. Environ. Health, Part B*, 2010, **13**, 299–313.
59. V. N. Sehgal, P. Verma and A. Khurana, Anthralin/dithranol in dermatology, *Int. J. Dermatol.*, 2014, **53**, e449–e460.

60. A. Modabbernia, E. Velthorst, C. Gennings, L. De Haan, C. Austin, A. Sutherland, J. Mollon, S. Frangou, R. Wright, M. Arora and A. Reichenberg, Early-life metal exposure and schizophrenia: A proof-of-concept study using novel tooth-matrix biomarkers, *Eur. Psychiatry*, 2016, **36**, 1–6.
61. S. S. Andra, C. Austin and M. Arora, The tooth exposome in children's health research, *Curr. Opin. Pediatr.*, 2016, **28**, 221–227.
62. J. E. Rager, M. J. Strynar, S. Liang, R. L. McMahan, A. M. Richard, C. M. Grulke, J. F. Wambaugh, K. K. Isaacs, R. Judson, A. J. Williams and J. R. Sobus, Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, *Environ. Int.*, 2016, **88**, 269–280.
63. M. E. Meek and V. C. Armstrong, in *Risk Assessment of Chemicals*, ed. C. J. V. Leeuwen and T. G. Vermeire, Springer, Netherlands, 2007, pp. 591–621.
64. E. J. Perkins, P. Antczak, L. Burgoon, F. Falciani, N. Garcia-Reyero, S. Gutsell, G. Hodges, A. Kienzler, D. Knapen, M. McBride and C. Willett, Adverse Outcome Pathways for Regulatory Applications: Examination of Four Case Studies with Different Degrees of Completeness and Scientific Confidence, *Toxicol. Sci.*, 2015, **148**, 14–25.
65. OECD Organisation for Economic Co-operation and Development, *Series on Testing and Assessment No.168*, Paris, France, 2012.
66. G. Patlewicz, C. Kuseva, A. Kesova, I. Popova, T. Zhechev, T. Pavlov, D. W. Roberts and O. Mekenyan, Towards AOP application – Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization, *Regul. Toxicol. Pharmacol.*, 2014, **69**, 529–545.
67. J. Jaworska, Y. Dancik, P. Kern, F. Gerberick and A. Natsch, Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice, *J. Appl. Toxicol.*, 2013, **33**, 1353–1364.
68. OECD Organisation for Economic Co-operation and Development, *Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Organisation for Economic Co-operation and Development*, ed. P. A. B. Environment Directorate, Paris, FR, 102nd edn, 2009.

## CHAPTER 5

# *Linking Drug or Phytochemical Exposure to Toxicity*

C. A. RODRÍGUEZ<sup>\*a,b</sup>, N. S. TEUSCHER<sup>c</sup> AND J. A. UCHIZONO<sup>d</sup>

<sup>a</sup>Adjunct Associate Professor of Pharmaceutical Sciences, The University of Michigan, Ann Arbor, MI, USA; <sup>b</sup>Adjunct Professor of Pharmaceutics and Medicinal Chemistry, University of the Pacific, Stockton, CA, USA; <sup>c</sup>Vice President of Scientific Training and Director of Strategic Consulting, Certara, Princeton, NJ, USA; <sup>d</sup>Professor of Pharmaceutics and Medicinal Chemistry, Associate Provost of Research, and Interim Dean of Graduate Studies, University of the Pacific, Stockton, CA, USA

\*E-mail: [carodrig@umich.edu](mailto:carodrig@umich.edu)

## 5.1 Introduction

Toxicokinetic studies are an integral part of the overall evaluation of the toxicity of a new chemical entity. By definition,<sup>1</sup> these studies are designed to generate pharmacokinetic data that may be used in the interpretation of toxicology findings and their relevance in clinical safety. At a very minimum, toxicokinetic data includes measurements of the systemic exposure in different animals at various dose levels to attempt a correlation between these exposure measurements and toxicology findings. In a typical toxicology study with toxicokinetic support, serial measurements of plasma concentrations of parent drug or metabolites (whether total or unbound) at specific time points yield exposure pharmacokinetic parameters (for example,  $C_{\max}$ , AUC,  $C_{ss}$ ,  $C$  at some time  $t$ ), that quantify the magnitude of the exposure in

---

Issues in Toxicology No. 31

Computational Systems Pharmacology and Toxicology

Edited by Dale E. Johnson and Rudy J. Richardson

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

that particular study. The collection of toxicokinetic data from multiple toxicology studies allows for the determination of exposures associated with the absence of toxicological findings, and as such, provide a margin of safety based on concentrations when clinical studies are planned.<sup>2</sup>

The predictive value of toxicokinetic data is enhanced when a suitable statistical model can be used to summarize the collective findings from different studies (and in some cases, even different compounds). These *pharmacostatistical* models contain two parts: a *structural* (or systematic) component which describes the exact mathematical relationship between different exposure and toxicity parameters, and a *variance* (or stochastic) component, which attempts to explain the statistical variability in the observed data. In general, biological data tend to have more noise than models found in the physical sciences. For that reason, pharmacostatistical models tend to be more complex.

The general approach to developing pharmacostatistical models is no different to other areas of science. First, the modeler scientist studies and identifies the problem at hand. In this process, some exposure or toxicity measurements may have already been collected. The modeler determines which variables are relevant, and make plans to measure additional variables if necessary. The relevant experiments are performed and the data are collected as planned. Part of data collection involves a documentation of the procedures followed in these experiments, a check in the quality of the data in which uncertainties in the measurements are assessed, and a data preparation step in which the data are placed in a suitable format for modeling. At this stage, the data are examined for trends, and a base model relating the toxicokinetic (or toxicology) parameters and the observations is postulated. An interactive process starts in which the base model gets refined with the addition of other parameters, helping to explain more of the variability observed in the data. These additional models are compared to the original base model, and only those models which statistically explain more of the observed data are subsequently retained, until the addition of more parameters does not significantly change the predictive value of the model. At this stage, a final model is chosen based on statistical goodness-of-fit criteria (such as residual analysis, coefficients of determination, and others), and among competing models with similar goodness-of-fit measures, the simplest one is chosen (Occam's razor). The predictive ability of the final model is checked by generating simulated predictions based on the model and comparing these predictions to the experimental observations. The result of this *visual predictive check* not only support the appropriateness of the final model, but can also suggest model improvements.

The final step in the model building process is the communication of the results by the modeler. The results of the model should be presented in relation to the original problem that was sought to be answered by the modeler. However, the message needs to be tailored to the intended recipient of the information. The level of detail in the communication of the results will be greater if the recipient is a regulatory agency or another scientist modeler; whereas if the recipient is a project team, the results of the modeling process

and the impact of the model in future decisions are more important than the details of the model. Regardless of the level of detail, the results should be presented in a clear and succinct manner. This will aid in understanding the science behind the modeling, its implications, and will avoid rejection of the model based on a lack of knowledge. This is assuming, of course, that the features and benefits of pharmacostatistical modeling are recognized by all who are the intended recipients. If not, models that are scientifically sound will be discarded, and considerable time would be spent on convincing others that the whole modeling approach is worthwhile.

Computing advances in the late 20th century have led to an increase in the use of pharmacostatistical models, not only in the traditional areas of drug development (that is, those related to human pharmacokinetics and approval of drugs), but also in other areas, such as toxicokinetics<sup>3</sup> and in veterinary medicine.<sup>4</sup> Regulatory agencies have recognized the value of pharmacostatistical modeling and have required the presentation of such data, not only as an integral part of a submission,<sup>5,6</sup> but also during the drug development process.<sup>7</sup> In fact, the United States Food and Drug Administration (FDA) encourages sponsors “to discuss the use of quantitative drug development methods (*e.g.* trial simulation using disease, drug, placebo, and dropout models) before conducting phase 2B and phase 3 clinical trials”.<sup>7</sup> For that reason alone, an understanding of these models is important for any drug development scientist, and, as the use of these models increases in toxicology areas, it is relevant for toxicologists to understand them as well. The following sections present an overview of the models used in drug development related to pharmacokinetic/toxicokinetic (PK/TK) measurements, PK/TK responses, and drug interactions.

## 5.2 Pharmacokinetic and Toxicokinetic Models

### 5.2.1 Structural Models

Traditionally, the structural component of PK/TK models have been described by a system of ordinary differential equations (ODE) that describe the time course of the concentration of the compound(s) of interest in the body. The level of complexity of these equations will depend on many factors, but in general they seek to relate the changes in the observed concentrations with time (*output*) as a function of how the compound was administered (*input*) and how the body handled it (*disposition*). This disposition component is perhaps the one that imparts the most complexity in the structural model. On one hand, the modeler could assume that the compound is distributed through different organs at different transfer rates. On the other hand, the modeler can simply assume that body in general is a collection of *compartments*, not necessarily physiologically based, where the amount of material acts kinetically like a distinct, homogeneous, well-mixed amount of the material.<sup>8</sup> While it may seem appealing at first to model the disposition of a drug using a physiological basis, the mathematical complexity is reduced

considerably by using compartmental modeling instead, and for this reason, compartmental models are the mainstay of PK/TK modeling.

### 5.2.1.1 Compartmental Models

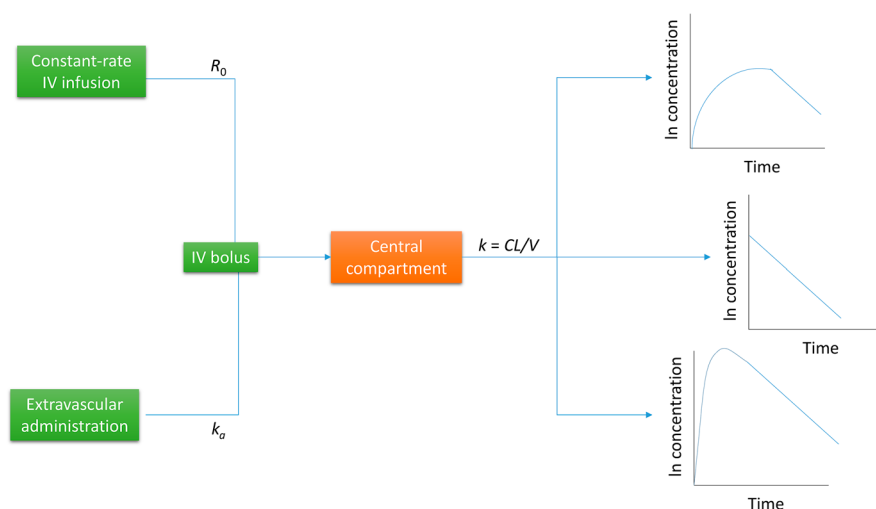
The choice of the number of compartments is based on both visual and statistical considerations. The modeler follows the “let the data speak” approach and examines the shape of the plasma concentration–time curve. If, once the drug is administered, the body appears to eliminate the drug following simple exponential decay, the modeler can assume that the drug is instantaneously distributed to all tissues in the body. Hence, the body can be considered as a single compartment (one-compartment model).

The differential equations associated with this the one-compartment model depend on how the drug was administered. After intravenous bolus administration (instantaneous input, Figure 5.1), the ODE that describes the rate of change of the amount of drug ( $A$ ) in the body with respect to time ( $t$ ) is

$$\frac{dA}{dt} = -kA \quad (5.1)$$

where  $k$  is the elimination rate constant. In this model, the rate of elimination is assumed to be proportional to the amount of drug in the body. The solution to that differential equation describes the relationship between the amount of drug in the body and time

$$A = A_0 e^{-kt} \quad (5.2)$$



**Figure 5.1** Schematic representation of the one-compartment model after different routes of administration and typical plasma concentration–time profiles.

where  $A_0$  is the initial amount, or the dose after intravenous bolus administration. However, instead of amounts of drug in the body, we measure concentrations in plasma. Hence, the same two equations in terms of concentration are

$$\frac{dC}{dt} = -kC \quad (5.3)$$

and

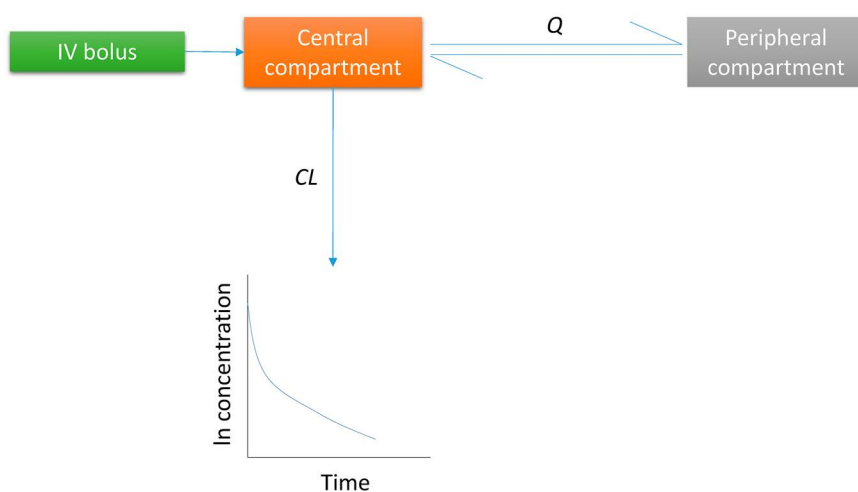
$$C = C_0 e^{-kt} \quad (5.4)$$

where  $C_0$  is now the initial concentration in plasma (Figure 5.2). The amount of drug in the body  $A$  is related to the concentration in plasma  $C$  by the relationship  $A = CV$ , where  $V$  is the volume of distribution, a proportionality constant with no physiological meaning. If, instead of an intravenous bolus administration, we have an intravenous infusion with constant rate  $R_0$  (Figure 5.1), the rate of change in plasma concentrations with time is

$$\frac{dC}{dt} = \frac{R_0}{V} - kt \quad (5.5)$$

and the concentration–time profile can be explained using the equation

$$C = \frac{R_0}{CL} (1 - e^{-kt}) \quad (5.6)$$



**Figure 5.2** Schematic representation of the two-compartment model after intravenous bolus administration and a typical plasma concentration–time profile showing the inflection point.  $CL$  represents the total body clearance and  $Q$  is the intercompartmental clearance.



where  $CL$  is the total body clearance of the compound. The ratio  $CL/V$  is simply  $k$ . Lastly, the equations associated with extravascular administration are

$$\frac{dC}{dt} = \frac{FDk_a}{V} - kC \quad (5.7)$$

and

$$C = \frac{FDk_a}{V(k_a - k)}(e^{-kt} - e^{-k_a t}) \quad (5.8)$$

where  $D$  is the administered dose,  $F$  is the absolute bioavailability, and  $k_a$  is the absorption rate constant. These are all one-compartment model equations, since there is only one elimination rate constant (Figure 5.1).

Sometimes it is not possible to use a one-compartment model. This is expected, because compounds do not distribute instantaneously. Two or more compartments representing sets of equally equilibrating tissues are needed to fully characterize the distribution characteristics of these compounds. In turn, the presence of multicompartment pharmacokinetics is reflected in the concentration–time profile, where the curve appears to have one or more inflection points (Figure 5.2). These inflection points separate the concentration–time curve into phases. In the two-compartment model, two phases are present: an initial phase where distribution predominates, and a terminal phase where elimination predominates after distribution equilibrium has been achieved. In these multicompartment models, a system of two or more differential equations is needed to fully describe the pharmacokinetics of the compound. As an example, in the two-compartment model represented by central (plasma, compartment 1) and peripheral (tissue, compartment 2) compartments, two differential equations are required, each describing the concentration–time profile of the drug in each one of the compartments. In the case where the compound is administered as an intravenous bolus:

$$\begin{cases} \frac{dC_1}{dt} = -(k_{12} + k_{10})C_1 + k_{21}C_2 \\ \frac{dC_2}{dt} = k_{12}C_1 - k_{21}C_2 \end{cases} \quad (5.9)$$

where  $k_{ij}$  is the transfer rate constant from the  $i$ th to the  $j$ th compartment, and  $C_i$  is the concentration in the  $i$ th compartment (in this notation, compartment 0 represents excreta). The solution to this system of equations is also a system of equations describing the concentration of drug in the central ( $C_1$ ) and peripheral ( $C_2$ ) compartments. Since in most PK/TK studies, tissue concentrations are not routinely monitored, it is the concentration in the central compartment that is modeled:

$$C_1 = Ae^{-\lambda_1 t} + Be^{-\lambda_2 t} \quad (5.10)$$

where

$$\begin{aligned} A &= \frac{D(k_{21} - \lambda_1)}{V_c(\lambda_2 - \lambda_1)}, \\ B &= \frac{D(k_{21} - \lambda_2)}{V_c(\lambda_1 - \lambda_2)} \end{aligned} \quad (5.11)$$

$V_c$  is the volume of the central compartment,  $D$  is the dose, and the rate constants  $\lambda_1$  and  $\lambda_2$  (the *macroconstants*) are functions of the transfer constants  $k_{12}$ ,  $k_{21}$ , and  $k_{10}$  (the *microconstants*). The sum of  $A$  and  $B$  is the initial concentration in the central compartment ( $D/V_c$ ). As the number of compartments increases, the number of terms in the equation describing increases and, in turn, the complexity of the model increases. Furthermore, if the compound is administered as a constant-rate infusion, or by oral administration, further terms representing these dosing events have to be added to the differential equation, and these would also be reflected in the final equation. For example, for an orally dosed drug that obeys a two-compartment model, the concentration–time profile can be modeled using the equation

$$\begin{aligned} C_1 = \frac{FKk_a}{V_c} \left[ \frac{(k_{21} - \lambda_1)}{(\lambda_2 - \lambda_1)(k_a - \lambda_1)} e^{-\lambda_1 t} + \frac{(k_{21} - \lambda_2)}{(\lambda_1 - \lambda_2)(k_a - \lambda_2)} e^{-\lambda_2 t} \right. \\ \left. + \frac{(k_{21} - k_a)}{(\lambda_1 - k_a)(\lambda_2 - k_a)} e^{-k_a t} \right] \end{aligned} \quad (5.12)$$

Wagner<sup>9</sup> has listed these compartmental models and readers are referred to this valuable reference.

In the derivation and use of these models, it was assumed that the all kinetic processes involved are first-order processes, that is, with rates directly proportional with concentration (linear pharmacokinetics). While this is the most common approach, it is not a necessary assumption. In fact, toxicokinetic data usually exhibit nonlinearities due to some capacity limitation in the absorption, distribution, metabolism, and excretion (ADME) of the compound as higher doses of the compound are administered. These nonlinearities can be incorporated in the model as well. They can take the form of a dose-dependent change in a pharmacokinetic parameter such  $F$ ,  $CL$ ,  $V$ , or intercompartmental rate constants. The literature is replete with examples of these situations. Dissolution limitations have been reported for griseofulvin, danazol, and the thiazides; first-pass metabolism has been reported for propranolol and salicylamide; permeability changes in transporters have been documented for some vitamins, cephalosporins, and aminopenicillins; saturable excretion has been associated with riboflavin and vitamin C; and saturable metabolism is present in phenytoin, acetaminophen, carbamazepine, alcohol, and theophylline.<sup>10</sup> These nonlinear processes make the modeling process more challenging. For example, parallel first-order and

saturable processes are evident in the elimination of aspirin. The presence of a distribution phase in addition to nonlinear elimination is a feature of monoclonal antibodies. Some metabolites can inhibit the metabolism of the parent drug. Enzymes can be induced or inhibited. In other cases, circadian rhythms affect the pharmacokinetics of some drugs. The modeler has to be aware of these potential situations and incorporate them in the model as appropriate.

### 5.2.1.2 Physiological Models

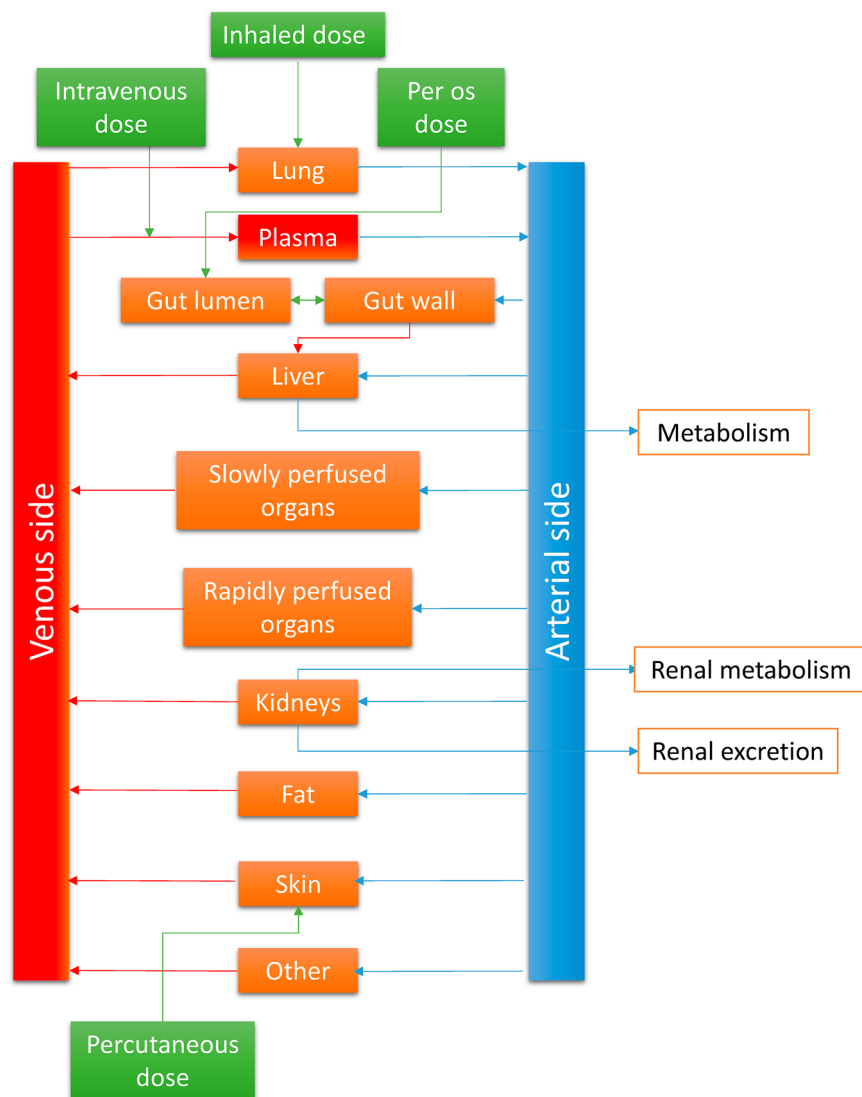
While compartmental models are used routinely in modeling, they suffer from two drawbacks. First, they give little insight into the physiological determinants of pharmacokinetics. It is not easy to predict the effect of a change in physiological status, or change in the pathophysiology, or a change of species, in the overall model. Second, they suffer from an identifiability problem. This means that multiple kinds of models can also explain the data. This last setback is partially remedied by the use of Occam's razor, but the modeler needs to be aware of the limitations of the model.

Physiologically based pharmacokinetic modeling (PBPK) is an attempt to remedy this situation. In these models, organ and tissue physiological and biochemical variables are related to the physicochemical characteristics of the compound under study (protein binding, lipid solubility, ionization, permeability characteristics, *etc.*; Figure 5.2). They are similar to the compartmental models in the sense that the "compartments" in this kind of modeling are the organs, and the transfer rate constants between compartments are the blood flows through those organs (Figure 5.3). However, they are unique in that both physiological and physicochemical variables are part of the differential equations that describe the system.

PBPK models are of two general kinds. In *flow-limited* models, it is assumed that the compound is taken up rapidly by the organ, hence only organ blood flow determines the disposition of the compound in that organ. In *permeability-limited* (or membrane-limited) models, there is a barrier within the organ or tissue that impedes the uptake of the compound. Hence, in these models, the organs are subdivided into two sub-compartments, each representing the extracellular and intracellular concentrations. As a consequence, the decline in intracellular concentrations does not parallel the decline in blood concentrations.

One general advantage of physiological modeling is scalability. In principle, if the physiological parameters are known for a certain species, the model can be scaled to consider other mammalian species by the use of allometry.<sup>11</sup> Hence, once a PBPK model is postulated, the PK/TK in other species may be predicted by the use of allometric scaling.

Physiological models have been used successfully in environmental toxicology,<sup>12</sup> in particular in the toxicological human risk assessment of compounds.<sup>13,14</sup> As an example, Clewell *et al.*<sup>13</sup> investigated the comparative carcinogenic potency of vinyl chloride (VC) in rodents and humans.



**Figure 5.3** A hypothetical physiologically based pharmacokinetic model, showing the direction of organ blood flow and routes of administration.

They developed a PBPK model that described the pharmacokinetics of VC to support the dosimetry needed for the risk assessment. Features of the model included two saturable metabolism pathways of VC, as well as glutathione depletion by VC metabolites. Their model described the pharmacokinetic data well across species. In addition, the animal and human risk assessment were consistent, and, in comparison to environmental data, their risk calculations were lower than those currently used by the US Environmental Protection Agency by a factor of 80.

Physiological modeling has been primarily used to make prior predictions of the distribution of a compound in various organs and tissues, and the agreement between theory and experiment has been good, in general. The main drawback of these models is their mathematical complexity. It is not unusual to see PBPK models with differential equations containing five or more terms, each representing the possible organs and tissues where the drug distributes. Furthermore, in the case of permeability-limited models, this number is even larger because the compound concentration in both extracellular and intracellular fluids needs to be accounted for. Attempts have been made to simplify the modeling by lumping together similarly behaved organs (in the kinetic sense), and shift the focus to the particular organ of interest. These hybrid or minimal-PBPK approaches have been used successfully,<sup>15,16</sup> when the lumping of organs and subsequent simplifications in the model are justified.

The applications of PBPK models in toxicology are varied. PBPK can help integrate findings across studies with different routes of administration, different dosing regimens, and different species.<sup>17</sup> Moreover, this technique can be used for the estimation of human equivalent doses.<sup>18</sup> By combining the pharmacodynamics of the compound, PBPK can relate the toxicological effect to particular exposures at the target organs.<sup>19</sup> Furthermore, these models could be modified to include lactation and fetal transfer, and thus estimate exposure in breast milk and in the fetus.<sup>20,21</sup>

### 5.2.2 Variance Models

An important component of pharmacostatistical models is the model used to explain the variability of the data. The choice and the form of these variance models relies in statistical theory. Typical assumptions in the simplest of models are that (1) the values of the independent variable are known without error and (2) the random errors are additive, have zero mean, common variance  $\sigma^2$  (the dependent variable has constant variability), and are uncorrelated. Furthermore, for hypothesis testing and construction of confidence intervals, normality and independence of errors are assumed. The estimation of the PK/TK parameters in the model is usually achieved *via* a nonlinear least-squares procedure, or by a maximum-likelihood procedure. After the modeling is complete, the statistical assumptions of the model are verified. If the common variance assumption is violated, remedies such as weighted least-squares, generalized-least squares, or transformation of the data are performed.<sup>22</sup> If the additive error assumption is violated, other error structures (Table 5.1) can be considered.

In PK/TK models, it is common to split the error into three components: an intersubject variability component, due to differences in the pharmacokinetic parameters between the subjects as a consequence the effect of variables such as age, gender, weight, and others, and interoccasion variability due to differences in PK parameters between periods. *Mixed-effect* modeling allows for the estimation of pharmacokinetic parameters in populations

**Table 5.1** Common error structures used in pharmacostatistical modeling.<sup>a</sup>

Type of error	Formula
Additive	$Y_{ijk} = f(X'_{ij}, \theta) + \varepsilon_{ijk}$
Poisson	$Y_{ijk} = f(X'_{ij}, \theta) + \sqrt{f(X'_{ij}, \theta)} \cdot \varepsilon_{ijk}$
Multiplicative	$Y_{ijk} = f(X'_{ij}, \theta) \varepsilon_{ijk}$
Proportional	$Y_{ijk} = f(X'_{ij}, \theta)(1 + \varepsilon_{ijk})$
Exponential	$Y_{ijk} = f(X'_{ij}, \theta) \exp(\varepsilon_{ijk})$
Combination	$Y_{ijk} = f(X'_{ij}, \theta)(1 + \varepsilon_{1ijk}) + \varepsilon_{2ijk}$

<sup>a</sup> $Y_{ijk}$  is the observed variable,  $f(X'_{ij}, \theta)$  is the regression function, and  $\varepsilon_{ijk}$  is the residual error.

where the pharmacokinetic parameters are not known. The acronym of the commonly used nonlinear estimation program NONMEM<sup>23</sup> comes from NONlinear Mixed-Effect Modeling. *Population pharmacokinetics* studies the sources and correlates of variability in plasma drug concentration between individuals representative of those in which the drug will be used clinically. The techniques of population pharmacokinetics are generally applicable to other areas, including toxicokinetics.

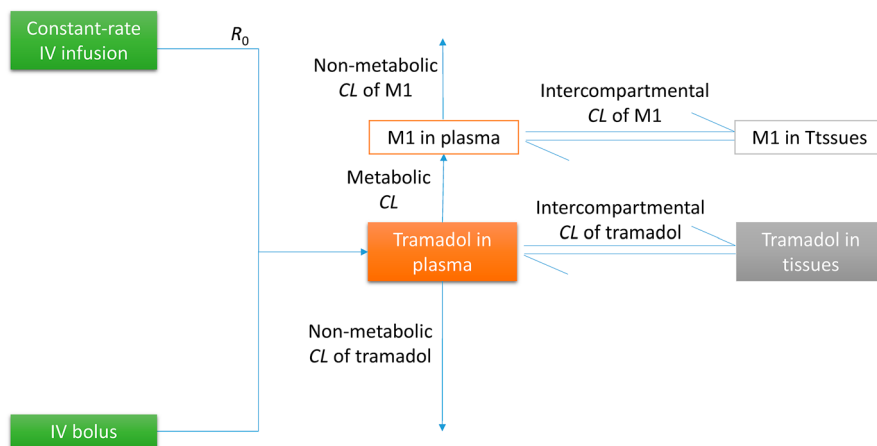
In general, this is how it works. The ideal situation would be to collect a large number of samples from each patient (rich dataset) in a large study and characterize the pharmacokinetics of the complete population in the study. However, this is impractical, especially in phase 2 and phase 3 clinical trials due to cost (assay, time, and staff), patient and medical staff inconvenience, and the possible added effect of sampling burden on the pharmacodynamic response. Luckily, this is not needed to achieve an appropriate population model. Rather, we can collect few points from many subjects. This procedure is known as *sparse sampling*. Each sample carries with it not only the concentration and time of the sample, but also information about the subject's characteristic (age, gender, body weight, *etc.*), since these are large trials in which these variables are either not controlled, or controlled between a wider range than in healthy volunteer trials. Hence, the structural parameters of the model may be functions of other random variables that reflect the individual characteristics of the subjects. These additional random variables are called *covariates*. The covariates exert their influence in the dependent variable through the structural parameters of the model, hence their effect on response is indirect. In other words, the covariates are not part of the structural model itself, but form a covariate model. As a consequence, the covariate model defines the interindividual variability, and since it is a random element, the form of the covariate model is very similar to the variance model.

As an example, consider a compound that obeys a one-compartment model with first-order elimination after intravenous bolus administration. Let us assume that only *CL* exhibits intersubject variability, and *V* is a constant only subjected to random error. We can assume that *CL* is a power function of body weight (BW). In this case, the population model could be written as

$$\begin{aligned}
 Y_{ij} &= (D/V_i) \exp[(-CL_i/V_i)t_{ij}] + \varepsilon_{ij} \\
 CL_i &= \theta_1 (BW)^{\theta_2} + \eta_{i,CL} \\
 V_i &= V \\
 \text{Var}(\varepsilon_{ij}) &= \sigma^2, \text{Var}(\eta_{i,CL}) = \omega_{CL}^2
 \end{aligned}
 \tag{5.13}$$

where  $\eta_k$  and  $\varepsilon_{ij}$  are the estimates of interindividual (covariate) and residual errors, respectively. In this past example,  $CL_{cr}$  is the covariate. The variance in the random errors is just  $\sigma^2$ , and the variance of the interindividual errors is called  $\omega^2$ . Other covariates may be considered, such as body weight, age, dose, race, body mass index, and other laboratory parameters. Notice the combination of fixed ( $t$ ) and random ( $CL_{cr}$ ) effects in the model (hence the “mixed-effect” nomenclature). In the past example, both the interindividual and random errors were assumed to be additive. Additional error structures could be tested: proportional or combination errors for the residual error and exponential errors for the interindividual error are common approaches (Table 5.1). If there were individual measurements performed on more than one occasion, additional covariates could be added to account for the different sampling periods.

A recent article by Holford *et al.*<sup>3</sup> illustrates the principles of PK/TK modeling presented in this section. In their article, the authors attempted to build a single model that would explain the pharmacokinetics of tramadol in different species, including dogs, rats, goats, horses, cats, donkeys, and humans, and predict whether current doses given to nonhuman species would be effective in treating pain. Tramadol is metabolized *via* CYP2D6 to *O*-desmethyl-tramadol (M1). Tramadol is a racemic mixture, and it exerts its analgesic mechanism by multiple mechanisms. Both (+)-tramadol and (+)-M1 are weak  $\mu$ -opioid agonists. Both (+)-tramadol and (+)-M1 stimulate serotonin release, (+)-tramadol inhibits serotonin reuptake, and (–)-tramadol inhibits the reuptake of norepinephrine in the spinal cord. Subjects with reduced M1 levels (poor metabolizers) have worse analgesia than extensive metabolizers. In nonhuman models of analgesia, M1 is six times more potent than tramadol. The authors were able to show that the concentration–time data followed a model shown in Figure 5.4. This structural model consisted of two two-compartment models, one for tramadol and another for M1, connected by a first-order transfer constant from the central compartment of tramadol to the central compartment of M1. The transfer constants between the central and peripheral compartments, as well as the exit rate constants from the central compartments, were all assumed to be first-order. Covariates considered in the covariate model included body weight (modeled as a power function) on all clearance and volume parameters for both tramadol and M1, and CYP2D6 metabolizer status (slow *vs.* extensive metabolizer) on the  $CL$  of tramadol to M1. The variance model consisted of exponential interindividual errors, and a combination model (additive and proportional) for the residual errors. The authors were able to quantify differences in tramadol  $CL$  to M1 between the species, as this value was lower in rats, dogs, and



**Figure 5.4** Compartmental model describing tramadol pharmacokinetics across species.

horses, compared to humans. In spite of accounting for body weight and metabolizer status differences between subjects and species, the authors were not able to further account for the residual variability of the data. Other covariates unrelated to body size (such as genotype, diet, and environment) may have contributed to the variability in the pharmacokinetic parameters of the model. Hence, substantial differences between species in the pharmacokinetics of tramadol and its primary metabolite remained, which were not explained by differences in body weight alone. The authors were able to show that volumes of distribution across species were not similar. Furthermore, based on the predicted steady-state M1 concentrations by the model, effective pain relief would not be achieved in dogs or horses with typically used doses in these two species, as the predicted M1 levels were 43.4% and  $\leq 10\%$  of the M1 concentrations in humans, respectively.

### 5.2.2.1 Other Modeling Methodologies

In the pharmacostatistical models considered so far, the structural mathematical model is supplemented by independent covariance and error models. This is the most commonly used approach. Stochastic models differ from deterministic mathematical models in that the inherent variability of the model parameters (or the model itself) is built into the mathematical model *a priori*, in other words, as part of the ODEs that describe the system. Hence, in contrast to the usual methodology where the variability in the data is assumed to come from interindividual and residual error components, in the stochastic model a noise component is introduced in the ODEs that allows for uncertainties in the system under study. The stochastic approach has been used successfully in pharmacokinetics and pharmacodynamics (PK/PD)<sup>24–26</sup> and represents another tool at the modeler's disposal.



Artificial neural networks (ANN) methodology is an alternative approach in the analysis of PK/TK data that does not rely on standard statistical methodologies. An ANN is a computer program that simulates the way the human brain processes information.<sup>27</sup> These networks study the data at hand, detecting patterns and relationships between dependent and independent variables without the use of complicated equations. Hence, ANN can be used to recognize patterns, predict outcomes, and model observed data. The data can be literature-based or derived from experiments. To achieve these goals, the networks are “trained” to recognize patterns by the use of a model dataset. However, the training requires relatively large datasets. Neural networks have been used in the description of PK/PD relationships.<sup>28–31</sup>

### 5.3 PK/PD Relationships

As commonly defined, “pharmacokinetics is what the body does to the drug; and pharmacodynamics is what the drug does to the body”; this chapter also defines the relationship between the concentration of drug and its resulting outcome or effect. The translation of how different levels of drug concentrations resulted in different scales of effects was quantified by many, with Gerhard Levy and colleagues proposing quantifiable relationships between doses and effects in the 1960s.<sup>32–34</sup> If we consider dose (or drug in the dosage form) as the input to the body and the resulting effect or response as the output, then the relationship between pharmacokinetics and pharmacodynamics becomes clearer. Pharmacokinetic models help characterize and correlate drug in the dosage form to a quantifiable concentration in tissues of interest, usually those tissues responsible for desired physiological effects and those tissues that may be responsible for undesired or toxic effects. Pharmacodynamic models will provide the quantitative relationship between the drug concentrations determined by the pharmacokinetic model and the measurable physiological outcomes resulting from those concentrations. Tying the pharmacokinetic model to the pharmacodynamic model creates a quantitative time–course relationship that is sensitive to size of dose and dosing interval to its resulting outcome; or, the PK/PD model quantitatively connects the input (dose) to the output (physiological effect). Generally, the more interesting relationship is the reverse, that is, for a specific output, how does one need to control the input; or, “if some physiological outcome,  $E$ , is desired, what size of dose(s) and at what interval(s) should the drug be given?” PK/PD modeling can provide the answer to this important question. Therefore, this section of the chapter will focus on the pharmacodynamics with a smaller subsection mentioning toxicokinetics/toxicodynamics (TK/TD). The size of this subsection is not indicative of its importance, rather it is mentioned because of its importance.

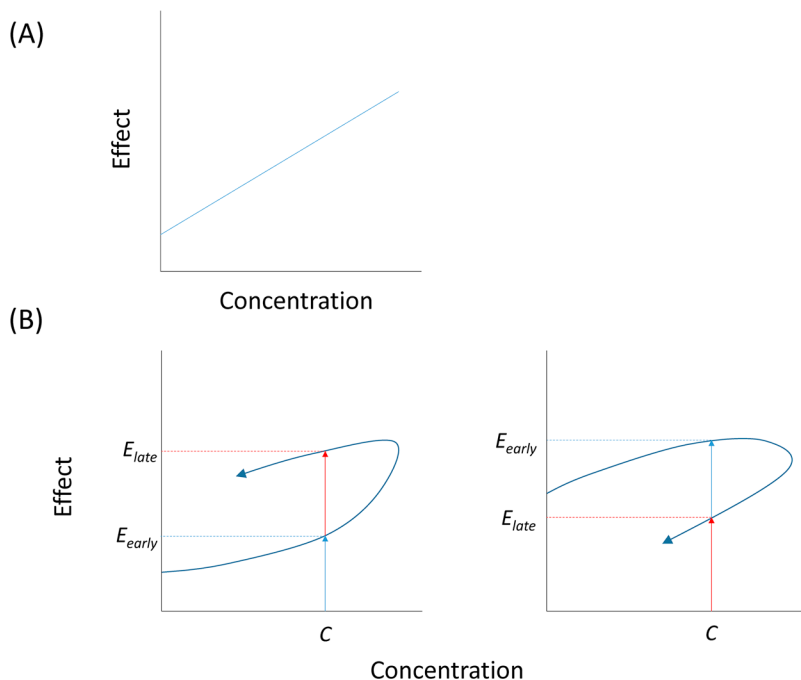
Although a pharmacokinetic description of a compound is necessary for an understanding of the relationship between exposure and response, pharmacokinetic models are most useful when they are used as the driving force for models that describe observed PD/TD effects. Fortunately, the mathematics

describing pharmacokinetics tend to be relatively straightforward and, for the most part, time-invariant (except in dose-dependent/nonlinear pharmacokinetics, for example, metabolism or transporters). Additionally, dose (input) tends to be proportional to the concentration of the compound of interest.

In pharmacodynamics, the mathematics are not necessarily more difficult than in pharmacokinetic modeling, but sometimes relationships between drug concentrations and corresponding physiological endpoints exhibit more complex dose-dependencies due to the influence of the compound on rate constants (coefficients) of formation or degradation of intermediates necessary to produce effect, *vs.* simply each compound linked to a specific level of necessary intermediate in a steady-state manner. Hence, drug effects in pharmacodynamic modeling (whether reversible or irreversible) are of two general classes, according to how quickly changes in plasma concentration are translated into effect: some compounds have immediate effects, where others have a lag between plasma concentration and effect. The immediacy of the effect is evidence of the attainment of a steady state or pseudo-equilibrium between concentration and response. Pharmacodynamic modeling provides the tools to quantitatively address both steady-state (pseudo-equilibrium) and nonequilibrium translation of drug into physiological effect. Direct models will be used to address the former and indirect models will be used to address the latter.

*Direct models* assume that the delay between changes in plasma concentration and the resulting effect is relatively small compared to the delay between changes in dose and the resulting plasma concentration. In these models, the maximum effect is observed at  $C_{\max}$ . In contrast, *indirect models* generally assume that the drug response is rate-limited by the biology of the pharmacodynamic system. In such models, maximum effects occur later than  $C_{\max}$ , and the effect can persist long after plasma concentrations are below the limit of quantitation. This delay may occur either because it takes time for the compound to reach the active site, or because the compound affects some other process that occurs before the observed effect (for example, inhibition of a pathway).

Direct and indirect effects can also be distinguished by using response *vs.* concentration plots (Figure 5.5). In contrast to direct effects in which each concentration has a unique value of response (no delay between plasma concentration and observed effect; Figure 5.5A), indirect effects show a *hysteresis loop*, the direction of which will depend on the specific mechanism of action (Figure 5.5B). In other words, two responses will be observed in indirect models: an early response corresponding to the first time the concentration is observed, and a late response corresponding to the second time the concentration is observed. *Counterclockwise* hysteresis occurs when the compound stimulates pharmacodynamic input (for example, stimulating the release of an endogenous substance) or when the compound inhibits pharmacodynamic output (such as inhibiting the release of an endogenous substance), since in both cases the effect increases with time for a given



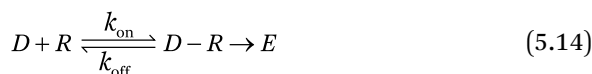
**Figure 5.5** Effect vs. concentration plots in (A) direct and (B) indirect models.

concentration. On the other hand, if the compound inhibits pharmacodynamic input (for example, inhibition of acid secretion by proton-pump inhibitors) or if it stimulates pharmacodynamic output (such as in the secretion of electrolytes in urine by diuretics), a *clockwise hysteresis loop* (or *proteresis*) will be observed. In this case, the drug effect decreases with time for a given concentration. It is important to note that other mechanisms can produce hysteresis. For example, clockwise hysteresis can also result from (1) drug tolerance, (2) formation of antagonistic metabolites, (3) down-regulation of receptors, and (4) feedback regulation, whereas counterclockwise hysteresis can also be observed from (1) distribution delays between the plasma and active site, (2) response delays, (3) sensitization of receptors, (4) formation and accumulation of active metabolites, and (5) up-regulation of receptors after chronic exposure to the drug. Furthermore, in some cases, changes in formulation of a drug can also elicit changes in the direction of the hysteresis loops.<sup>35</sup>

### 5.3.1 Mathematical Description of Pharmacodynamic Effects

The mathematical description of pharmacodynamic effects results from the law of mass action, where free drug in the active site  $D$  binds to a free receptor  $R$  producing a drug–receptor complex ( $D - R$ ). The formation of the

drug–receptor complex, in turn, elicits a sequence of events (either immediate or delayed) that leads to the observed effect ( $E$ )



where  $k_{\text{on}}$  is the rate constant of drug and receptor binding and  $k_{\text{off}}$  is the rate constant of drug–receptor complex dissociation. If the binding is reversible, then equilibrium constants of association ( $K_a$ ) and dissociation ( $K_d$ ) can be defined as

$$\begin{aligned} K_a &= \frac{k_{\text{on}}}{k_{\text{off}}} = \frac{[D - R]}{[D][R]} \\ K_d &= \frac{1}{K_a} = \frac{[D][R]}{[D - R]} \end{aligned} \quad (5.15)$$

where  $[D]$  is the concentration of the free drug at the active site,  $[R]$  is the concentration of receptor,  $[D - R]$  is the concentration of drug–receptor complex. If the binding is irreversible, then  $k_{\text{off}}$  (and  $K_d$ )  $\approx 0$ . The consequences of the law of mass action are obvious: increases in drug concentration at the active site lead to increases in drug–receptor complex, and therefore, increasing drug effect. Since the number of receptors is finite, additional increases in concentration beyond the saturation point do not produce additional increases in drug effect; hence a theoretical maximum effect ( $E_{\text{max}}$ ) exists for all compounds. This is expressed by the *Hill equation*

$$E = \frac{E_{\text{max}}[D]^\gamma}{EC_{50}^\gamma + [D]^\gamma} \quad (5.16)$$

where  $EC_{50}$  is the active site concentration at half the maximum effect. The parameter  $\gamma$  is a sigmoidicity parameter (the *Hill coefficient*) which describes the steepness of the curve: the greater the number, the steeper the curve in its ascending portion (Figure 5.6). This equation is reminiscent of the Michaelis–Menten equation of enzyme kinetics.

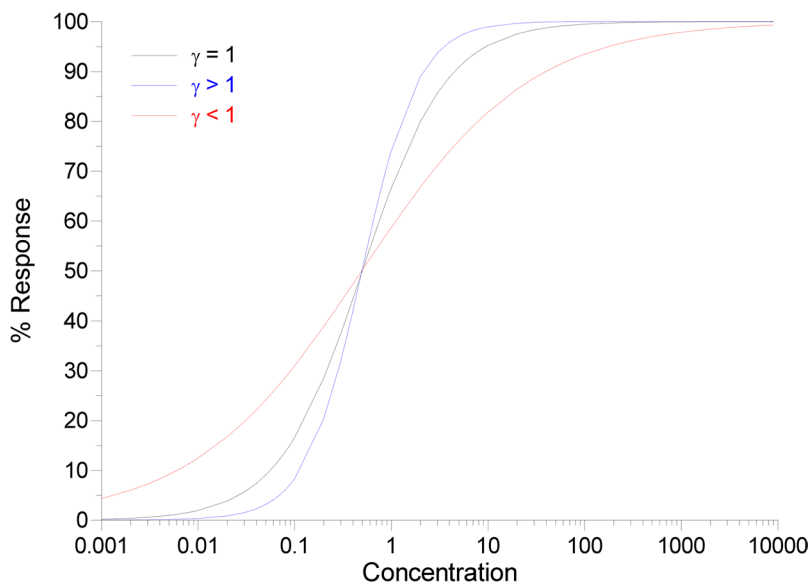
Additional pharmacodynamic models can be derived from the Hill equation. If we assume that  $\gamma = 1$ , the simple  $E_{\text{max}}$  model is obtained

$$E = \frac{E_{\text{max}}[D]}{EC_{50} + [D]} \quad (5.17)$$

If the  $EC_{50}$  is much higher than the concentration at the active site  $[D]$ , the Hill equation reduces to a simple linear model

$$E = \frac{E_{\text{max}}}{EC_{50}^\lambda} [D] \quad (5.18)$$

However, the linear model is not likely to be a reasonable model unless the physiological effect is over a very small linear range: it predicts that any large effect can be achieved as long as the dose is increased enough.



**Figure 5.6** Typical effect vs. concentration plot for a compound that follows the Hill equation. The coefficient  $\gamma$  is related to the steepness of the curve.

A baseline effect  $E_0$  can be added to the Hill equation

$$E_{\text{total}} = E_0 \pm \frac{E_{\text{max}}[D]^\gamma}{EC_{50}^\gamma + [D]^\gamma} \quad (5.19)$$

where  $E_{\text{total}}$  is algebraic sum of the baseline and drug effects, and the  $\pm$  indicates that the drug effect can be stimulatory (+) or inhibitory (−).

### 5.3.1.1 Direct Effects and Effect Compartments

Pharmacologists have long recognized the direct relationship between drug concentration and response (or effect). There has also been a long-running understanding that drug concentration as a function of time was related to measured response as a function of time, leading to numerous studies designed specifically to assure that steady-state concentrations were achieved before the measurement of the response measurement. These steady-state experiments led to a host of models (still used today) where the assumption of steady-state drug concentration was or is experimentally upheld.

If the lag between effect and concentration is very short, or if the experiment is performed at steady-state conditions, plasma concentrations can be used as a surrogate for concentrations at the active site. Hence,  $[D]$  can be substituted with  $C$ , the plasma concentration, and the Hill equation becomes

$$E_{\text{total}} = E_0 \pm \frac{E_{\text{max}}C^\gamma}{EC_{50}^\gamma + C^\gamma} \quad (5.20)$$

Sometimes there is a disconnect between the concentration at the active site and the concentration in plasma, in particular when equilibration between the two sites does not occur rapidly. In those cases, it is convenient to include an *effect compartment* (also called a *link compartment*), in which a negligible amount of the drug in plasma (proportional to the administered dose) distributes with a specific rate constant  $k_{e0}$ ; negligible because the compartment should have no influence on the overall pharmacokinetics of the drug. Hence, the effect compartment is a hypothetical kinetic drug compartment, described by

$$\frac{dC_e}{dt} = -k_{e0}(C - C_e) \quad (5.21)$$

where  $C_e$  is the concentration in the effect compartment. The Hill equation then becomes

$$E_{\text{total}} = E_0 \pm \frac{E_{\text{max}} C_e^\gamma}{EC_{50}^\gamma + C_e^\gamma} \quad (5.22)$$

Notice that the effect is still direct with respect to  $C_e$ .

Since the maximum effect is observed at  $C_{\text{max}}$  in direct-effect models, the maximum effect always occurs at  $t_{\text{max}}$ , and a change in dose should not affect the time of maximum effect, assuming that the drug obeys linear pharmacokinetics.

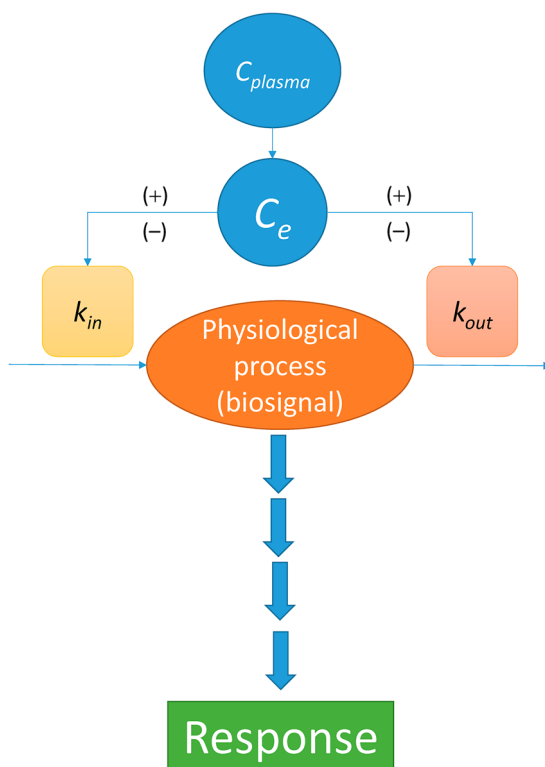
The Michaelis–Menten and Hill equations provide physiological reality, in that both have asymptotic maximum effects ( $E_{\text{max}}$ ): despite increasing the dose to very large sizes, these models return an effect that is capped at  $E_{\text{max}}$ , and very little gain is achieved by increasing the dose to large sizes. While these pharmacodynamics models are quite useful, the parameters ( $E_{\text{max}}$ ,  $EC_{50}$ , and  $\gamma$ ) are only a function of pharmacological and physiological cascades; the parameters are not influenced by dose or time.

### 5.3.1.2 Indirect Effects

The presence of a lag between the observed effect and plasma concentrations that cannot be accounted by the presence of an effect compartment alone is indicative of a mechanism that involves indirect effects. In this case, the overall effect can be thought of as the result of the interaction between a compound and one or more physiological processes at steady state that produce mediators (for example, biomarker, hormones, signaling proteins, *etc.*) that modulate the effect (Figure 5.7). The production of these mediators is a zero-order process, whereas the output is a first-order process dependent on the mediator concentration  $S$ .

$$\frac{dS}{dt} = k_{\text{in}} - k_{\text{out}}S \quad (5.23)$$

where  $k_{\text{in}}$  and  $k_{\text{out}}$  are the specific input and output rate constants of the mediator, respectively. In the absence of drug and at steady state, the baseline level



**Figure 5.7** General scheme of an indirect effect model. At the top, the plasma concentration is related to the concentration at the site of action ( $C_e$ ), which in turn modulates the rate of change of the concentration of a mediator, either by stimulating (+) or inhibiting (–) the production (or input,  $k_{in}$ ) or consumption (output,  $k_{out}$ ) of a biosignal. The thick blue arrows represent pharmacodynamic signal transduction leading to the final response.

of  $S$  ( $S_0$ ) is  $S_0 = k_{in}/k_{out}$ , and the baseline effect ( $E_0$ ) is proportional to  $S_0$ . Compounds can act by increasing or decreasing the input ( $k_{in}$ ) or the output ( $k_{out}$ ) of those mediators *via* the functions  $H(t)$  and  $G(t)$ , respectively:

$$\frac{dS}{dt} = H(t)k_{in} - G(t)k_{out}S \quad (5.24)$$

For example, the oral hypoglycemic tolbutamide stimulates the release of insulin, which causes a decrease in blood glucose levels, whereas atorvastatin inhibits the production of mevalonic acid by HMG-coA reductase, resulting in a decrease in blood cholesterol.

Notice that in contrast to direct models, where there is a direct effect of the concentration of the compound in the concentration of the mediator ( $S$ ), in indirect models the concentration of drug affects the *rate* of change of the concentration of the mediator ( $dS/dt$ ), and because of that, the effect takes longer to occur.

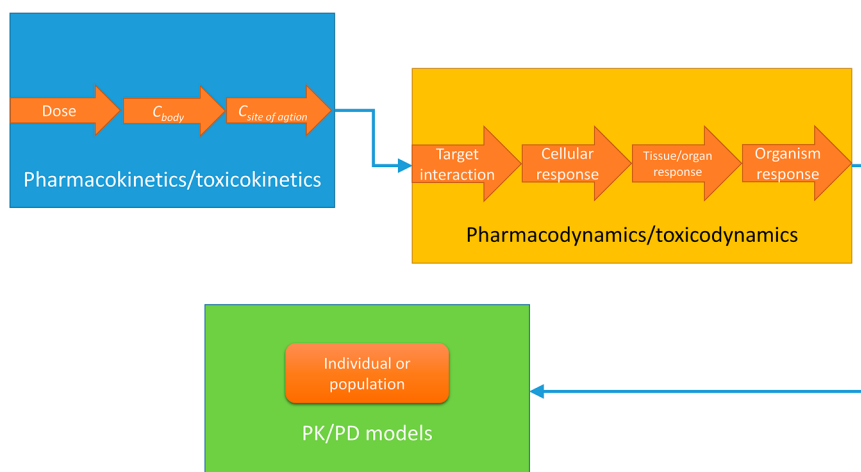
Indirect models can be refined by describing the signal transduction processes that occur between the pharmacokinetics and the final effect. These additional processes can be modeled by the use of transit compartments between the activation of the receptor and the observed final effect, and can be described by additional differential equations. These transit compartment models have been used successfully in anticancer drugs and in describing the time course of myelosuppression in clinical studies.<sup>36,37</sup>

A feature of indirect models is the effect of dose on the time of the observed maximum effect. In contrast to direct-effect models, the time of the observed maximum effect is shifted to later times as the dose is increased.

### 5.3.2 Combined PK/PD and TK/TD Modeling

While the level of sophistication of PK/PD and TK/TD models has continued to improve as research characterizing translational biological cascades of beneficial and adverse/toxic effects continues, PK/PD models have tended to be “agency”-driven, and TK/TD relationships have been much broader in scope. Yet, PK/PD and TK/TD models are simultaneously modeling input and outcomes (Figure 5.8).<sup>38</sup> PK/PD modeling strategies have been the focus up to this point because of regulatory requirements for new drug entities and the wealth of literature coverage. On the other hand, while TK/TD literature is more difficult to find due to its breadth of systems, many of the PK/PD modeling strategies can be easily applied to address TK/TD challenges.

The broader scope of TK/TD literature, which includes phytochemical and TK/TD models in animals<sup>39–41</sup> and general modeling strategies,<sup>42–44</sup> demonstrate that defining of a toxicological endpoint is difficult; creating models to describe those difficult endpoints is even more difficult; and determining



**Figure 5.8** Strategy and scope of toxicokinetic/toxicodynamic modeling. Information from ref. 38.



model parameter estimates is even more difficult. Two models of TK/TD are highlighted in this section to illustrate an essential element for any well-designed TK/TD model and/or experiment; both link the dose or environmental disturbance (usually concentration of pollutant) time-course to adverse or toxic effects to the entity exposed to the drug or disturbance. The TK/TD model's ability to predict outcomes outside the experimental design are highly dependent on the TK/TD model's linking of the time-course input of dose or disturbance to the resulting outcome.

Lobo and Balthasar<sup>39</sup> created three TK/TD time-dependent models from time-static models, for mice exposed to the cancer chemotherapy agent methotrexate. To preserve the dose relationship of the static models, the time derivative was obtained for each of the static models to create the time-dependent models. The adverse/toxicological event of concern was the nadir of the time-, dose-, and exposure-dependent nadir of percentage body weight. Their study is an excellent example of PK/PD and simultaneous modeling of TK/TD for a drug with complex time-independent and time-dependent measurable outcomes.

As described earlier, TK/TD tends to be broader in scope than PK/PD with regards to the system under examination. Ashauer and co-workers<sup>41</sup> modeled the influence of environmental chemical on reduced fish growth. This work is analogous to *in vitro*–*in vivo* correlation (IVIVC) in PK/PD modeling. The IVIVC aspect of this model creates a more time-static dosing regimen as input, but the linking of *in vitro* to *in vivo* is reasonably done and points out the dose-dependency incorporated into the TK/TD model. Eqn (5.25) gives the relationship between *in vitro* concentrations,  $C_{\text{int}}$ , (cell culture) and environmental chemical concentrations,  $C_{\text{ext}}$ . Note the equation similarity between first-order input and first-order elimination used in PK/PD modeling.

$$\frac{dC_{\text{int}}(t)}{dt} = C_{\text{ext}}(t)k_{\text{in}} - C_{\text{int}}(t)k_{\text{out}} \quad (5.25)$$

This equation is coupled through  $C_{\text{int}}$  to two additional differential equations modeling a stochastic death hazard function, which describes the pharmacodynamics.

### 5.3.3 Modeling Pharmacodynamics in the Absence of Pharmacokinetic Data: K-PD Models

A full implementation of a PK/PD model is not always possible because of a lack of sufficient concentration–time points to fully characterize the pharmacokinetics of the drug. One strategy is to simplify the pharmacodynamics model by assuming that the driving force of the pharmacodynamic effect is the amount of drug in the body,<sup>45</sup> instead of the drug concentrations. These K-PD models (the omission of the P represents the absence of pharmacokinetics data) have been used successfully, in particular in cases where previous knowledge of the pharmacokinetics of the drug is available, or in relatively rich pharmacodynamics datasets (with wider dose ranges, longer

observation periods, higher number of samples, and/or higher number of subjects than a regular PK/PD analysis), since information about the pharmacokinetics of the drug could be deduced in part from the pharmacodynamics response. Although the models are generally well behaved, they tend to produce large biases when high interindividual variabilities are present. Hence, K-PD models are by no means a replacement for full PK-PD models; they should be used cautiously in cases where such pharmacokinetics data are unreliable or missing, as they rely on a sound knowledge of the underlying pharmacokinetics and pharmacodynamics of the compound.

## 5.4 Modeling Drug Interactions with Phytochemicals

The current environment of polypharmacy results in single individual patients receiving many different combinations of drug products to treat single or multiple medical conditions. Many efforts have been made by regulators to establish consistent labeling standards with respect to drug–drug interactions that might occur. These labels can then be used by prescribing physicians or pharmacists to make recommendations on the choice of therapies for each patient. However, the increasing use of herbal supplements is leading to a new challenge of addressing the potential for drug–herb interactions that might lead to adverse drug reactions. Unfortunately, there is very sparse information available on possible drug–herb interactions, and in many countries the number of herbal preparations available to patients is greater than the number of drug products. Thus the evaluation of the potential effects of herbal preparations on drug pharmacokinetics is an important part of the drug development paradigm and can affect patient safety.

While the number of herbal preparations and drug combinations are nearly infinite, the methods that can be used to model those interactions can be reduced to four major categories: (1) inhibition of metabolism; (2) induction of metabolism; (3) enhancement of absorption; or (4) inhibition of absorption. The inhibition of metabolism (1) and enhancement of absorption (4) will lead to higher drug levels and potential for toxicity while the induction of metabolism (2) and inhibition of absorption (3) will lead to reduced drug levels and potential loss of clinical effect. Although the end result may be similar across effects, the methods used to model the changes are very different. Categories 1 and 2 represent effects on the systemic pharmacokinetic parameters of the drug, while categories 3 and 4 induce an effect through pre-systemic processes.

Little experimental work has been performed in the area of clinical drug–herb interactions, nevertheless, modeling and simulation can play a significant role in evaluating potential interactions. In the following sections, methodologies for modeling the four categories of interactions will be presented. These models can be coupled with preclinical and *in vitro* data to allow for simulation of potential changes in drug levels that may impact clinical usage.

### 5.4.1 Inhibition of Metabolism

Inhibition of metabolism is one of the most common types of drug-herb interaction. All exogenous chemicals are handled by the body's metabolic system, regardless of the origin of the chemical species. Thus, many phytochemicals or herbs are handled by the same metabolic machinery as the pharmaceutical agents. Therefore, inhibition of key metabolic enzymes such as the cytochrome P450 is not uncommon. One of the most common examples of inhibition of metabolism is the concomitant use of citrus juice (for example, grapefruit juice) and a drug that is metabolized by CYP3A4. For example, the label for the lipid-lowering drug atorvastatin<sup>46</sup> contains the following recommendation: "Grapefruit juice contains one or more components that inhibit CYP3A4 and can increase plasma concentrations of atorvastatin, especially with excessive grapefruit juice consumption (more than 1.2 liters per day)".<sup>46</sup> The inhibition of drug metabolism leads to increases in drug levels which can cause adverse drug reactions. Thus inhibition is of specific interest related to the safety of drug-herb combinations.

Modeling the inhibition of metabolism can be achieved with small modifications to existing pharmacokinetic equations. As an example, let us assume we have a theoretical drug is administered orally and follows one-compartment pharmacokinetics. Eqn (5.8) can be written as follows:

$$C = \frac{1}{V} \left( \frac{FDk_a}{k_a - CL/V} \right) \left[ e^{-tCL/V} - e^{-k_a t} \right] \quad (5.26)$$

Inhibition of metabolism means that the presence of the herbal preparation will reduce the clearance of the drug from the body. Therefore, the parameter  $CL$  will be modified in the presence of the herbal preparation. One way to model this interaction would be to use a fraction term, such that in the presence of an herbal preparation is as follows

$$CL_{\text{herb}} = f_{\text{herb}} CL \quad (5.27)$$

where  $f_{\text{herb}}$  represents the fractional reduction in clearance due to the presence of the herb. The value for  $f_{\text{herb}}$  is between 0 and 1. This assumes that the inhibition of metabolism is constant over time and that the herb is at steady-state levels that lead to maximal inhibition. Another way to model this interaction would be to use a saturable equation as follows

$$CL_{\text{herb}} = \left( 1 - \frac{I_{\text{max}} C_{\text{herb}}}{IC_{50} + C_{\text{herb}}} \right) CL \quad (5.28)$$

where  $I_{\text{max}}$  represents the maximal inhibition of the clearance of the drug and can range from 0 (no inhibition) to 1 (complete inhibition),  $C_{\text{herb}}$  is the concentration of herb in the systemic circulation. This concentration can be represented using a separate pharmacokinetic equation for the herb, or set to a constant value to approximate steady-state kinetics of the herb, and  $IC_{50}$

is the concentration of herb required to elicit 50% of the maximal inhibition. This model assumes that there is a fractional reduction in drug clearance that is related to the concentration of herb present. As herb levels rise, inhibition increases until you hit a maximum inhibition of  $1 - I_{\max}$ . At steady-state levels of an herb, this essentially collapses to the fractional model presented previously. A third way to model the interaction would be to separate the drug clearance into two separate components, one that is susceptible to inhibition and one that it not. The portion that is susceptible to inhibition would then be modified using one of the methods shown above. For example,

$$CL_{\text{herb}} = CL_1 + f_{\text{herb}} CL_2 \quad (5.29)$$

where  $CL_1$  represents the drug clearance that is independent of the inhibition,  $CL_2$  represents the drug clearance that is susceptible to inhibition, and  $f_{\text{herb}}$  is the fractional reduction in clearance due to the presence of the herb and can range between 0 and 1. A similar model could be constructed using a saturable equation to model the drug clearance that is susceptible to inhibition.

The inhibition of metabolism by herbs can be modeled using a series of simple equations that account for reductions in drug clearance either through constant fraction reductions or saturable mechanisms. However, all of these methods are empirical: each attempts to quantify the extent of the inhibition without specifying the mechanism. Inclusion of the drug-herb interaction terms in a pharmacokinetic model for a drug would allow simulation of the potential effects of herbs. *In vitro* data could be used to calculate fractional reductions and/or saturable inhibition properties of different herbs. Then this information could be fed into a pharmacokinetic model to predict the expected drug levels in the presence and absence of a specific herbal preparation.

### 5.4.2 Induction of Metabolism

Induction of metabolism results in the increased biological activity of a metabolic enzyme which can lead to increased drug metabolism and decreased drug levels in systemic circulation. One example of an inducer is St John's Wort (*Hypericum perforatum*), a potent inducer of CYP3A4 and P-glycoprotein,<sup>47</sup> which is used for the treatment of anxiety, depression, stomach upset, insomnia, fluid retention, and hemorrhoids. This herbal preparation is widely used by many people; however, 802 drugs have reported interactions with St John's Wort.<sup>48</sup> The potential impact of induction is the possibility of reduced efficacy or breakthrough symptoms due to low drug levels.

As with inhibition, the induction of metabolism can be modeled with simple modifications to the clearance parameters. The first method is to use a fraction term in the presence of an herbal preparation as in eqn (5.27), but with  $f_{\text{herb}}$  now representing the fractional increase in clearance due to the presence of the herb and takes a value  $>1$ . This assumes that the induction of

metabolism is constant over time and that the herb is at steady-state levels that lead to maximal induction. This simple method is often accurate, as induction requires upregulation of metabolic enzymes which circulate for long periods of time relative to drug administration. Another way to model the interaction would be to separate the drug clearance into two separate components, one that is susceptible to induction and one that it not, as in eqn (5.29), where  $CL_1$  now represents the drug clearance that is independent of the induction,  $CL_2$  represents the drug clearance that is susceptible to induction, and  $f_{\text{herb}}$  is the fractional increase in clearance due to the presence of the herb ( $>1$ ).

The induction of metabolism by herbs is simpler than inhibition because the time course for induction is much longer than the time course for the affected drug product. Thus, steady-state assumptions regarding the amount of upregulation of the metabolism is supported. The fractional increase methodology is easy to implement and intuitive to interpret when explaining the effect of inducing a drug-metabolizing enzyme.

### 5.4.3 Enhancement of Absorption

The process by which drugs administered orally are absorbed into the systemic circulation are complex and involve a variety of mechanisms. Processes including dissolution of the dosage form, solubility in environments with changing pH levels, cell wall permeability and transport, and even pre-systemic metabolism all play a role in the absorption of a drug following oral administration. Most herbal preparations are administered at large doses relative to pharmaceutical agents. For example, most pharmaceutical agents are administered at a dose  $<1000$  mg, while many herbal preparations are administered at doses  $>1000$  mg. Thus concomitant administration of an herbal preparation and a drug can lead to significant effects in the gastrointestinal tract. These effects can include altering pH levels, inhibiting cellular uptake or transport, or even blocking metabolic processes. While the potential interactions are complex and varied, the mathematical models that can describe the effects are quite straightforward. For an orally administered drug, there are two parameters that are related to the absorption process, as shown in eqn (5.26). Bioavailability ( $F$ ) represents the amount of drug absorbed, and the absorption rate ( $k_a$ ) represents the rate at which the drug is absorbed into the systemic circulation.

If the presence of the herbal preparation enhances the extent of absorption, then the herb will increase the effective bioavailability of the drug. This can be modeled using a simple additive model such as

$$F_{\text{herb}} = F + F1 \quad (5.30)$$

where  $F$  represents the absolute bioavailability of the drug in the absence of the herbal preparation, and  $F1$  represents the increase in bioavailability due to the herb. It is important to note that the modeled value of  $F_{\text{herb}}$  must

be between 0 and 1, therefore  $F1$  must be less than  $1 - F$  in all cases. Setting appropriate boundaries is essential for the modeling to be successful.

If the herbal preparation enhances the rate of drug absorption, a fractional increase in the rate constant  $k_a$  can be modeled as

$$k_{a,\text{herb}} = f_{\text{herb}} k_a \quad (5.31)$$

where  $f_{\text{herb}}$  is the fractional increase in absorption rate constant, and is  $>1$ . Modeling increases in the absorption rate require multiple concentration measurements between time of dose administration and time at which the peak drug concentration is observed ( $t_{\text{max}}$ ). It is also possible to have a compound effect where both the amount of drug absorbed and the rate at which absorption occurs can be enhanced. These increases in the rate and extent of absorption can result in elevated drug levels that might lead to toxicity. Therefore, modeling changes in these parameters can help guard against unexpected adverse drug reactions.

#### 5.4.4 Inhibition of Absorption

When a drug and herbal preparation are administered concomitantly, inhibition of absorption is a possible scenario. As discussed previously, the absorption process is complex; however, we normally model it with only two pharmacokinetics parameters ( $F$  and  $k_a$ ). Modeling the inhibition of absorption is very similar to modeling enhancements. If the amount of drug absorbed is reduced (either due to increased pre-systemic metabolism or lack of absorption from the gut lumen), it can be modeled as a reduction in  $F$  as shown below:

$$F_{\text{herb}} = F - F1 \quad (5.32)$$

where  $F_{\text{herb}}$  is the bioavailability in the presence of the herb, and  $F1$  is the reduction in bioavailability due to the presence of the herb. The value for  $F_{\text{herb}}$  must be between 0 and 1, and the value of  $F1$  must always be less than or equal to  $F$ , the bioavailability of the drug in the absence of the herb.

If the herbal preparation slows the rate of drug absorption, then a change in the absorption rate constant,  $k_a$ , can be modeled, as shown in eqn (5.31). As noted, to model changes in the absorption rate constant, several concentration measurements are needed between the time of dose administration and the  $t_{\text{max}}$ . The herbal preparation may cause a reduction in both the extent of absorption ( $F$ ) and the rate of drug absorption ( $k_a$ ), in which case you would include both eqn (5.31) and (5.32) in the pharmacokinetic model for the drug.

#### 5.4.5 Modeling of Pharmacodynamic Interactions

If two compounds act on the same receptor, or if they act on the same component of the pharmacodynamics cascade, they have the potential to elicit

a pharmacodynamic interaction, whether additive or synergistic, or antagonistic. Modifications to the Hill equation have been used alongside response surface methodology to mathematically describe these interactions. For example, if a drug  $D$  and a partial receptor agonist  $A$  interact, the overall effect can be described as

$$E = \frac{E_{\max} IC_{50}[D] + I_{\max} EC_{50}[A]}{EC_{50} IC_{50} + [D] IC_{50} + [A] EC_{50}} \quad (5.33)$$

where  $[A]$  is the concentration of the agonist,  $IC_{50}$  is the concentration of the agonist at 50% inhibition, and  $I_{\max}$  is the maximum inhibitory effect.

## 5.5 Conclusions

TK/TD modeling rests on sound pharmacokinetic and pharmacodynamic principles. While the application of modeling techniques to explain toxicokinetic and toxicodynamic outcomes is not as widespread and abundant as the description of clinical pharmacology, efficacy, or safety outcomes, it is expected that this area will continue to grow, as regulatory agencies continue to encourage the application of these techniques, and as computer science continues to advance, both in hardware and software. Moreover, while little experimental work has been performed in the area of clinical drug–herb interactions using either experimental or modeling and simulation analyses, the potential effects of drug–herb interactions are significant: reductions in drug levels can cause lack of efficacy which may result in breakthrough symptoms and increases in drug levels can lead to unexpected adverse drug reactions. Here, modeling and simulation could be used to estimate these potential effects.

## References

1. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatory-information/guidances/ucm074937.pdf>, last accessed January 2016.
2. <http://www.fda.gov/downloads/Drugs/./Guidances/UCM078932.pdf>, last accessed January 2016.
3. S. Holford, K. Allegaert, B. J. Anderson, B. Kukanich, A. B. Sousa and A. Steinman, *et al.*, Parent-metabolite pharmacokinetic models for tramadol—tests of assumptions and predictions, *J. Pharmacol. Clin. Toxicol.*, 2014, 2(1), 1023.
4. D. H. Zhao, Z. Zhang, C. Y. Zhang, Z. C. Liu, H. Deng and J. J. Yu, *et al.*, Population pharmacokinetics of valnemulin in swine, *J. Vet. Pharmacol. Ther.*, 2014, 37(1), 59–65.
5. <http://www.fda.gov/ScienceResearch/SpecialTopics/WomensHealthResearch/default.htm>, last accessed January 2016.
6. <http://www.fda.gov/downloads/drugs/guidances/UCM072137.pdf>, last accessed January 2017.



7. <http://www.fda.gov/downloads/Drugs/.Guidances/ucm079690.pdf>, last accessed January 2016.
8. J. A. Jacquez, *Compartmental Analysis in Biology and Medicine*, University of Michigan Press, Ann Arbor, 2nd edn, 1985.
9. J. G. Wagner, *Pharmacokinetics for the Pharmaceutical Scientist*, Technomic Publishing Company, Inc., Lancaster, 1st edn, 1993.
10. M. Rowland and T. N. Tozer, *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications*, Lippincott, Williams & Wilkins, Philadelphia, 4th edn, 2010.
11. Q. Huang and J. R. Riviere, The application of allometric scaling principles to predict pharmacokinetic parameters across species, *Expert Opin. Drug Metab. Toxicol.*, 2014, **10**(9), 1241–1253.
12. D. Liu, L. Pan, H. Yang and J. Wang, A physiologically based toxicokinetic and toxicodynamic model links the tissue distribution of benzo[a]pyrene and toxic effects in the scallop *Chlamys farreri*, *Environ. Toxicol. Pharmacol.*, 2014, **37**, 493–504.
13. H. J. Clewell, P. R. Gentry, J. M. Gearhart, B. C. Allen and M. E. Andersen, Comparison of cancer risk estimates for vinyl chloride using animal and human data with a PBPK model, *Sci. Total Environ.*, 2001, **274**, 37–66.
14. M. Pelekis and C. Emond, Physiological modeling and derivation of the rat to human toxicokinetic uncertainty factor for the carbamate pesticide aldicarb, *Environ. Toxicol. Pharmacol.*, 2009, **28**, 179–191.
15. H. Sayama, H. Komura and M. Kogayu, Application of hybrid approach based on empirical and physiological concept for predicting pharmacokinetics in humans—usefulness of exponent on prospective evaluation of predictability, *Drug Metab. Dispos.*, 2013, **41**, 498–507.
16. H. Sayama, H. Komura, M. Kogayu and M. Iwaki, Development of a hybrid physiologically based pharmacokinetic model with drug-specific scaling factors in rat to improve prediction of human pharmacokinetics, *J. Pharm. Sci.*, 2013, **102**(11), 4193–4204.
17. C. S. Kim, J. A. Sandberg, W. Slikker, Z. Binienda, P. M. Schlosser and T. A. Patterson, Quantitative exposure assessment: application of physiologically-based pharmacokinetic (PBPK) modeling of low-dose, long-term exposures of organic acid toxicant in the brain, *Environ. Toxicol. Pharmacol.*, 2001, **9**(4), 153–160.
18. L. M. Sweeney, C. R. Kirman, S. A. Gannon, K. D. Thrall, M. L. Gargas and J. H. Kinzell, Development of a physiologically based pharmacokinetic (PBPK) model for methyl iodide in rats, rabbits, and humans, *Inhalation Toxicol.*, 2009, **21**(6), 552–582.
19. C. Timchalk, R. J. Nolan, A. L. Mendrala, D. A. Dittenber, K. A. Brzak and J. L. Mattsson, A Physiologically based pharmacokinetic and pharmacodynamic (PBPK/PD) model for the organophosphate insecticide chlorpyrifos in rats and humans, *Toxicol. Sci.*, 2002, **66**(1), 34–53.
20. A. E. Loccisano, M. P. Longnecker, J. L. Campbell, M. E. Andersen and H. J. Clewell, Development of PBPK models for PFOA and PFOS for human pregnancy and lactation life stages, *J. Toxicol. Environ. Health, Part A*, 2013, **76**(1), 25–57.



21. M. Yoon, J. D. Schroeter, A. Nong, M. D. Taylor, D. C. Dorman and M. E. Andersen, *et al.*, Physiologically based pharmacokinetic modeling of fetal and neonatal manganese exposure in humans: describing manganese homeostasis during development, *Toxicol. Sci.*, 2011, **122**(2), 297–316.
22. D. G. Kleinbaum, L. L. Kupper and K. E. Muller, *Applied Regression Analysis and Other Multivariate Methods*, PWS-KENT Publishing Company, Boston, 2nd edn, 1988.
23. S. Beal, L. B. Sheiner, A. Boeckmann and R. J. Bauer, *NONMEM User's Guides (1989–2009)*, Icon Development Solutions, Ellicott City, 2009.
24. S. Donnet and A. Samson, A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models, *Adv. Drug Delivery Rev.*, 2013, **65**(7), 929–939.
25. J. Leander, J. Almquist, C. Ahlström, J. Gabrielsson and M. Jirstrand, Mixed effects modeling using stochastic differential equations: illustrated by pharmacokinetic data of nicotinic acid in obese Zucker rats, *AAPS J.*, 2015, **17**(3), 586–596.
26. F.-R. Yan, P. Zhang, J.-L. Liu, Y.-X. Tao, X. Lin and T. Lu, *et al.*, Parameter Estimation of Population Pharmacokinetic Models with Stochastic Differential Equations: Implementation of an Estimation Algorithm, *J. Probab. Stat.*, 2014, **2014**, 836518.
27. S. Agatonovic-Kustrin and R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, *J. Pharm. Biomed. Anal.*, 2000, **22**, 717–727.
28. K. M. Tolle, H. Chen and H.-H. Chow, Estimating drug/plasma concentration levels by applying neural networks to pharmacokinetic data sets, *Decis. Support Syst.*, 2000, **30**, 139–151.
29. J. V. S. Gobburu and E. P. Chen, Artificial Neural Networks As a Novel Approach to Integrated Pharmacokinetic-Pharmacodynamic Analysis, *J. Pharm. Sci.*, 1996, **85**(5), 505–510.
30. S. H. Haidar, S. B. Johnson, M. J. Fossler and A. S. Hussain, Modeling the Pharmacokinetics and Pharmacodynamics of a Unique Oral Hypoglycemic Agent Using Neural Networks, *Pharm. Res.*, 2002, **19**(1), 87–91.
31. A. Belič, I. Grabnar, I. Belič, R. Karba and A. Mrhar, Predicting the anti-hypertensive effect of nitrendipine plasma concentration profiles using artificial neural networks, *Comput. Biol. Med.*, 2005, **35**, 892–904.
32. G. Levy, Relationship between elimination rate of drugs and rate of decline of their pharmacologic effects, *J. Pharm. Sci.*, 1964, **53**, 342–343.
33. G. Levy, Kinetics of pharmacologic effects, *Clin. Pharmacol. Ther.*, 1966, **7**(3), 362–372.
34. R. Nagashima, R. A. O'Reilly and G. Levy, Kinetics of pharmacologic effects in man: the anticoagulant action of warfarin, *Clin. Pharmacol. Ther.*, 1969, **10**(1), 22–35.
35. C. Louizos, J. A. Yáñez, M. L. Forrest and N. M. Davies, Understanding the hysteresis loop conundrum in pharmacokinetic/pharmacodynamic relationships, *J. Pharm. Pharm. Sci.*, 2014, **17**(1), 34–91.

36. B. C. Bender, F. Schaedeli-Stark, R. Koch, A. Joshi, Y. W. Chu and H. Rugo, *et al.*, A population pharmacokinetic/pharmacodynamic model of thrombocytopenia characterizing the effect of trastuzumab emtansine (T-DM1) on platelet counts in patients with HER2-positive metastatic breast cancer, *Cancer Chemother. Pharmacol.*, 2012, **70**(4), 591–601.
37. S. Hayes, P. N. Mudd, Jr., D. Ouellet, B. M. Johnson, D. Williams and E. Gibiansky, Population PK/PD modeling of eltrombopag in subjects with advanced solid tumors with chemotherapy-induced thrombocytopenia, *Cancer Chemother. Pharmacol.*, 2013, **71**, 1507–1520.
38. <http://www.ecotoxmodels.org/hot-topics/toxicokinetic-toxicodynamic-models/>, last accessed February 2016.
39. E. D. Lobo and J. P. Balthasar, Pharmacokinetic-pharmacodynamic modeling of methotrexate-induced toxicity in mice, *J. Pharm. Sci.*, 2003, **92**(8), 1654–1664.
40. R. Ashauer and C. D. Brown, Toxicodynamic assumptions in ecotoxicological hazard models, *Environ. Toxicol. Chem.*, 2008, **27**(8), 1817–1821.
41. J. Stadnicka-Michalak, K. Schirmer and R. Ashauer, Toxicology across scales: Cell population growth *in vitro* predicts reduced fish growth, *Sci. Adv.*, 2015, **1**(7), e1500302.
42. K. K. Rozman and J. Doull, Dose and time as variables of toxicity, *Toxicology*, 2000, **144**(1–3), 169–178.
43. J. M. Frazier, Predictive toxicodynamics: Empirical/mechanistic approaches, *Toxicol. In Vitro*, 1997, **11**(5), 465–472.
44. T. Jager, C. Albert, T. G. Preuss and R. Ashauer, General unified threshold model of survival—a toxicokinetic-toxicodynamic framework for ecotoxicology, *Environ. Sci. Technol.*, 2011, **45**(7), 2529–2540.
45. P. Jacqmin, E. Snoeck, E. A. van Schaick, R. Gieschke, P. Pillai and J. L. Steimer, *et al.*, Modelling response time profiles in the absence of drug concentrations: definition and performance evaluation of the K-PD model, *J. Pharmacokinet. Pharmacodyn.*, 2007, **34**(1), 57–85.
46. [http://www.accessdata.fda.gov/drugsatfda\\_docs/label/2009/020702s056lbl.pdf](http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/020702s056lbl.pdf), last accessed February 2016.
47. S. Zhou, E. Chan, S. Q. Pan, M. Huang and E. J. Lee, Pharmacokinetic interactions of drugs with St. John's wort, *J. Psychopharmacol.*, 2004, **18**(2), 262–276.
48. <http://www.drugs.com>, last accessed February 2016.

## CHAPTER 6

# *Chemical Similarity, Shape Matching and QSAR*

E. V. RADCHENKO<sup>\*a,b</sup>, G. F. MAKHAEVA<sup>b</sup>, V. A. PALYULIN<sup>a,b</sup>  
AND N. S. ZEFIROV<sup>a,b</sup>

<sup>a</sup>Lomonosov Moscow State University, Department of Chemistry, Leninskie gory 1/3, Moscow, 119991, Russia; <sup>b</sup>Institute of Physiologically Active Compounds RAS, Severniy proezd 1, Chernogolovka, Moscow region, 142432, Russia

\*E-mail: [genie@qsar.chem.msu.ru](mailto:genie@qsar.chem.msu.ru)

## 6.1 Introduction

One of the fundamental paradigms of chemistry is the *similarity property principle*: “similar structures possess similar properties”.<sup>1,2</sup> It follows naturally from the chemical structure theory (structure determines the properties of a compound)<sup>3,4</sup> and the intuitive expectation that such structure-to-property functions are continuous (‘Nature makes no leaps’). This principle lays the basis for the detection, analysis, and interpretation of patterns in the known data on the properties (including biological activities) of the compounds as well as for using these patterns to predict the properties for novel structures or to design the structures with desired properties. Although, as always, the devil is in the detail, and the specific applications of these approaches often have their own prerequisites and limitations (as we will see later), they

provide extremely useful tools for many areas of chemistry, pharmacology, and toxicology.

In addition to the purely scientific applications, the question of whether a given compound is substantially (or sufficiently) similar to some other compound may also have legal implications, in particular in the intellectual property and drug enforcement contexts. Unfortunately, the criteria for solving it set forth in various jurisdictions and laws have significant differences and often tend to be too vague, too over-reaching and/or too restrictive. To some extent, this is caused by the attempts to prescribe simple solutions for complex *structural* problems that are, in turn, only indirectly related to the critical issue of whether a given compound will be substantially (or sufficiently) similar *in properties* to some other compound.

Due to the limited size of this chapter, we cannot hope to present a complete picture of all the methods developed over more than five decades in the fields of quantitative structure–activity relationships (QSAR)/quantitative structure–property relationships (QSPR) and molecular similarity analysis. Instead, the choice of the topics covered in the chapter will be a little subjective, with some focus on basic ideas and representative examples, on more recent results and on the techniques and services that are immediately available for solving some of the practical problems of computational pharmacology and toxicology. Among many excellent texts and reviews looking at somewhat different aspects of these areas, several publications<sup>1,5–12</sup> deserve a particular mention. The references to more detailed reviews on specific topics are also provided throughout the chapter when appropriate.

In the following sections, we first consider the concept of molecular similarity as a measure of closeness in chemical space and some approaches to the analysis of activity landscapes defined in such chemical spaces (Section 6.2). Then the approaches to the analysis of structure–activity/property relationships in congeneric and/or compact series of compounds that can be expected to share a consistent mechanism of action are discussed (Section 6.3.1). Finally, the approaches to the analysis of diverse and/or large data sets are presented (Section 6.3.2). This distinction reflects significant differences in the goals and methods of the analysis in these two situations.

## 6.2 Molecular Similarity, Chemical Spaces and Activity Landscapes

### 6.2.1 Molecular Similarity: Concept and Definitions

#### 6.2.1.1 Molecular Similarity Concept

The idea that some molecules are more or less similar to some other molecules may seem obvious to anyone who has ever studied chemistry. Indeed, it was recognized in the earliest years of chemical science<sup>5</sup> and forms an important framework for the systematic analysis and presentation of the accumulated chemical knowledge. However, as is often the case with

'intuitively obvious' concepts, the concept of molecular similarity has many aspects, variations, definitions, and applications, as evidenced by the vast body of literature on the subject.<sup>1,5</sup> In most general terms, we can consider three components of a similarity measure: *applicability domain*, *structure representation*, and *similarity (comparison) function*. These components are not independent but more or less closely related to each other, and a choice for one of them usually limits the possible (or desirable) choices for others. Nevertheless, within these limits one can combine different approaches almost as freely as the units of a model railway. This opens immense possibilities for the molecular similarity analysis but also greatly complicates the field in practice.

### 6.2.1.2 Applicability of Molecular Similarity Measures

Firstly, each molecular similarity measure involves an explicit or, more often, implicit specification of its *applicability domain* – a set of structures (in other words, a region in chemical space) and a set of problems for which its application is possible and meaningful. If a similarity measure is used beyond its applicability domain, the results are usually not very predictable or relevant. It is probably not possible and not necessary to define some 'true' or 'universal' similarity measure. In fact, chemical structures (and compounds) are complex objects that have many different facets. In different situations, we may be focused on similarities and differences in scaffolds, substituents, and functional groups or in low-level patterns of atoms and bonds, in physical or physico-chemical properties, in chemical reactivity, in interactions with specific biological targets or in overall physiological effects. Each of these facets is objective, but any similarity measure reflecting them would also involve some elements of subjective perceptions and cognitive processes. Thus, a similarity measure should be selected (or constructed) in such a way as to properly capture the features of a structure that are important for a specific problem. In addition, some kind of testing or validation<sup>5</sup> is desirable in order to confirm that the important features are indeed captured.

### 6.2.1.3 Structure Representations for Molecular Similarity Analysis

The different facets of the chemical structures relevant for particular applications are reflected in the structure representations used to define various application-focused similarity measures. The most commonly used representations are considered below.

Perhaps the simplest and most natural representation is directly based on the *common (matching) and different (mismatching) parts or fragments of the structures*. In chemoinformatics, they correspond to various subgraphs (in particular, the maximum common subgraph) of their molecular graphs.<sup>5</sup> This approach works best when the differences between the structures are relatively small and localized, such as the variations in only one (or very few)

atoms, substituents or functional groups attached to or present within a common structural scaffold,<sup>13</sup> and will not be useful in more general cases.

The standard approach to the estimation of topological similarity for more different and diverse structures (where the differences cannot be easily isolated) is based on the consideration of *large numbers of very simple and often overlapping substructural fragments and/or other structural features*. Most commonly only the presence of each feature in a given structure is taken into account (thus, from a mathematical point of view, structures are represented by sets of features that are encoded as bit vectors often called *molecular fingerprints* or *molecular keys*),<sup>14</sup> although the fragment/feature occurrence counts can also be used (giving rise to fragment multisets).<sup>5</sup> The term 'molecular fingerprint' reflects the fact that these entities provide unique representations of each molecule, but individual bit values in them are not characteristic and may be difficult to interpret in isolation. From a practical viewpoint, the fingerprint bit vectors should have manageable size and reasonable population density (proportion of the 'on' bits), since comparing too sparse or too dense vectors (where all or almost all positions are zeroes or ones, respectively) is not very informative. In order to achieve this and still cover the most representative set of fragments, hashing algorithms may be used to map each fragment to specific bit vector positions that are associated with several different fragments,<sup>15</sup> making the fingerprint/hologram nature of the representation even more pronounced. Nevertheless, an important property of all types of fingerprints is that the bit positions associated with a certain substructure will also be set in the fingerprints of larger structures containing it.<sup>15</sup>

Among the most commonly used molecular fingerprint schemes one should mention the MACCS keys,<sup>16</sup> CACTVS/PubChem keys,<sup>17</sup> Extended Connectivity Fingerprints (ECFP),<sup>18</sup> Daylight fingerprints,<sup>15</sup> and the proprietary procedures such as the ChemAxon<sup>19</sup> fingerprints. Major differences between them are related to the sets or classes of fragments and other structural features included in the fingerprint generation as well as to the hashing algorithms. In addition, some of the schemes are not sufficiently documented, and some discrepancies between different implementations may exist. All these differences lead to some uncertainties complicating the use of the resulting similarity measures (in particular, the values obtained from different schemes are not directly comparable). Even more importantly, due to the universal 'common denominator' nature of the fingerprint-based similarity measures, they may mask or distort the importance of the fine structural differences relevant for a particular problem. Numerous attempts have been made to optimize the fingerprint schemes and/or select the best scheme and parameter set for a specific task (*e.g.* similarity-based search for active compounds, see Section 6.2.3.1).<sup>20–22</sup> Nevertheless, we believe that the fingerprint-based similarity measures probably can be used most safely to recognize very similar or very dissimilar structures, while any interpretation of small changes in similarity (especially outside of the high or low similarity ranges) should be very cautious.

The *physical, physico-chemical* or *topological characteristics* of the compound that are expected to be relevant for a particular problem can be represented by simple vectors of molecular descriptor values.<sup>5</sup> Such descriptors may be bulk or macroscopic properties (*e.g.* solubility, lipophilicity, acidity, density, or heat of vaporization) measured experimentally or predicted by means of the quantitative structure–property models, as well as molecular graph invariants (from molecular weight, number of rotatable bonds, or number of hydrogen bond donors and acceptors to topological indices reflecting various structure properties) or quantum chemical parameters (*e.g.* molecular orbital or ionization energies). In order to simplify the processing, eliminate the effect of different parameter scales, and enable visual representation of the similarity relationships, the dimensionality reduction techniques such as the principal component analysis may be used.<sup>23,24</sup>

A similar approach can be used to estimate the molecular similarity based on the *chemical reactivity* or *biological activity profiles* represented by suitable parameters such as the equilibrium or rate constants, effective concentrations, or induced changes in the ‘omics’ fingerprints.

The analysis of similarities in the patterns of *potential intermolecular interactions* (in particular, ligand interactions with a biological target) requires some representation of the molecular interaction fields. In the 3D QSAR, the sampling of these fields at the nodes of a rectangular grid is traditionally used.<sup>25,26</sup> However, the representations based on a set of Gaussian functions positioned on atoms and/or interaction centers<sup>27–29</sup> enable the analysis of similarity both in molecular shape and specific interactions. They also have important advantages in terms of computational performance as well as stability with respect to grid changes, imperfect alignment, and minor conformational variations.

Another approach considers the potential interactions of small molecules on a topological (2D) level of chemical structure defined by their structural formulas. In order to represent them in a common frame of reference, the Molecular Field Topology Analysis (MFTA)<sup>30,31</sup> builds a molecular supergraph, that is, a kind of superstructure such that all structures belonging to a series under study can be superimposed onto the supergraph and characterized in a uniform way. As molecular descriptors, MFTA employs local physico-chemical parameters (atom and bond properties) that can be quickly evaluated from a structural formula. In particular, electrostatic (*e.g.* effective atomic charge), steric (van der Waals radii of atoms and groups), lipophilicity, and hydrogen-bonding descriptors may be considered. Although this method was developed primarily for the QSAR analysis (it and its applications will be discussed in more detail in Section 6.3.1 below), such supergraph-based uniform representations of potential atom-centered interactions (or other relevant local parameters) can also be used to estimate the similarity between compounds or to match their structures in the most meaningful way (*i.e.* most similar with respect to specific parameters).



### 6.2.1.4 Molecular Similarity Functions

A *molecular similarity function* maps a pair of molecules  $M_i$  and  $M_j$  (described by some suitable representation that is relevant to the intended application) to a real-valued *similarity measure*  $S(M_i, M_j)$  reflecting a ‘degree’ or ‘amount’ of similarity between them. In most cases, such functions are *symmetric*, i.e.  $S(M_i, M_j) \equiv S(M_j, M_i)$ , although asymmetric similarity measures have been proposed (see later) and may be more efficient in some applications where there is a natural asymmetry in the roles of the two molecules (e.g., a query/template molecule vs. library molecules, or a parent structure vs. its derivatives).

The similarity functions are usually constructed as normalized ratios, with the numerator expressing some measure of size or count of the common elements and the denominator expressing a measure of total size or element count of the molecules (or other objects). Thus, their values lie in the interval from 0 to 1. The unity value is taken when a molecule is compared to itself ( $S(M_i, M_i) \equiv 1$ ) or any other molecule that is identical in terms of a representation used for the similarity calculation. The zero value means that the molecules are ‘completely’ different (although such situations are quite rare in practice).

The *distance* or *dissimilarity* between molecules is commonly defined as a similarity complement to unity<sup>32</sup> according to eqn (6.1).

$$D(M_i, M_j) = 1 - S(M_i, M_j) \quad (6.1)$$

This distance is also limited to the  $[0, 1]$  interval, which may seem counterintuitive (after all, the chemical space is unlimited). Moreover, for many similarity measures the corresponding distance is not a metric in the mathematical sense. In particular, the identity of indiscernibles axiom is usually violated (as noted above, different molecules may have the same representation and thus  $S = 1$  and  $D = 0$ ). The triangle inequality is also often violated (although it may be satisfied in certain special cases).<sup>32</sup> Thus, it is best to view this parameter as a ‘distance coefficient’ or ‘dissimilarity index’, which is rather informal but adequate in most applications. Although some nonlinear transformations of the similarity values may be constructed in order to overcome the ‘limited distance’ problem, in most cases such complexity is not necessary.

Numerous similarity and distance measures have been proposed in the literature.<sup>1,5,32,33</sup> Sometimes the same or closely related parameters in different or even the same fields were developed by different authors independently. In this chapter, we consider only a few similarity functions that have found a widespread use in chemoinformatics (see Table 6.1). It is remarkable that the functions intended for different types of molecular representations have a substantial similarity that allows us to formulate them in a uniform way. For two molecules  $A$  and  $B$ , they are defined in terms of the generalized ‘size’ values for their intersection (common part)  $|A \cap B|$ , their union  $|A \cup B|$ , as well as individual molecules  $|A|$  and  $|B|$  (in some cases, the union size is more conveniently expressed as a sum of individual molecule sizes corrected by the intersection size according to eqn (6.2)). Various quantities used in these



calculations are illustrated in Figure 6.1. Naturally, the way they are defined depends on the molecular representation selected for the study.

$$|A \cup B| = |A| + |B| - |A \cap B| \quad (6.2)$$

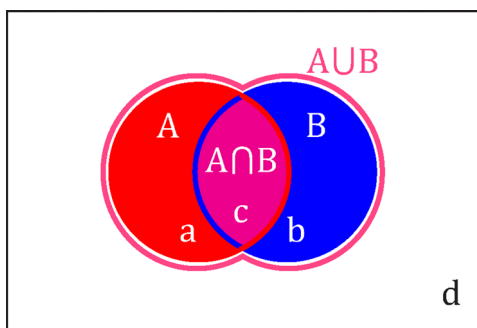
For the *set-based representations* (in particular, *molecular fingerprints*) of the molecules *A* and *B*, we can denote their individual components by  $A_k$  and  $B_k$ , respectively. They can represent the presence (0/1) or occurrence count of a particular fragment or a structural feature, or simply the value in the *k*-th position of the fingerprint vector. Then the sizes for the similarity calculation are defined by eqn (6.3)–(6.6). In the most common case where *A* and *B* are bit vectors, the equations can be simplified by introducing the bit counts *a*, *b* and *c* representing respectively the number of bits that are set only in *A*, only in *B* and in both vectors (see Figure 6.1). In practical implementations, optimized bitwise operations are often used.

$$|A| = \sum_k A_k = a + c \quad (6.3)$$

$$|B| = \sum_k B_k = b + c \quad (6.4)$$

$$|A \cap B| = \sum_k \min[A_k, B_k] = c \quad (6.5)$$

$$|A \cup B| = \sum_k \max[A_k, B_k] = a + b + c \quad (6.6)$$



**Figure 6.1** Size quantities used in the calculation of similarity functions for two molecules. Molecule *A* is represented by the red circle, molecule *B* is represented by the blue circle. Their intersection (common part)  $A \cap B$  is shown in magenta, their union  $A \cup B$  is indicated by the pink contour. Lowercase letters denote the sizes of individual areas: the red crescent *a* is the difference  $A \setminus B$  (part of *A* not present in *B*), the blue crescent *b* is the difference  $B \setminus A$  (part of *B* not present in *A*), the magenta lens *c* is the intersection  $A \cap B$ , and the white area *d* represents elements of the representation that are not present in either *A* or *B* (this quantity is not used in the similarity functions considered here).

For the *graph-based representations* of the molecules  $A$  and  $B$ , the size is usually defined as a number of edges (bonds) in the corresponding molecular graphs (cardinality of the edge set), denoted by  $|E(A)|$  and  $|E(B)|$ . The intersection between the molecules is then defined as their maximum common edge substructure (MCES), *i.e.* a common subgraph containing maximum number of edges (for more rigorous graph-theoretical terminology and overview of algorithmic approaches see ref. 34 and 35). The focus on the matching bonds generally provides a more ‘chemically meaningful’ picture of structural relations.<sup>5,36</sup> The size quantities for the similarity calculation are defined by eqn (6.7)–(6.10).

$$|A| = |E(A)| \quad (6.7)$$

$$|B| = |E(B)| \quad (6.8)$$

$$|A \cap B| = |\text{MCES}(A, B)| = |E[\text{MCES}(A, B)]| \quad (6.9)$$

$$|A \cup B| = |A| + |B| - |A \cap B| = |E(A)| + |E(B)| - |E[\text{MCES}(A, B)]| \quad (6.10)$$

In the *vector-based representations* the molecules  $A$  and  $B$  are characterized by the descriptor vectors  $\mathbf{A}$  and  $\mathbf{B}$  with individual components  $\mathbf{A}_k$  and  $\mathbf{B}_k$ , respectively. The descriptors may also be derived by sampling the continuous 3D molecular fields at the nodes of a rectangular grid. The individual molecule size is then defined as a squared norm of the corresponding vector (denoted by  $\|\mathbf{A}\|^2$  and  $\|\mathbf{B}\|^2$ ), while the intersection size is defined as a dot (inner) product of the vectors  $\langle \mathbf{A}, \mathbf{B} \rangle$ . The size quantities for the similarity calculation are defined by eqn (6.11)–(6.14).

$$|A| = \|\mathbf{A}\|^2 = \langle \mathbf{A}, \mathbf{A} \rangle = \sum_k \mathbf{A}_k^2 \quad (6.11)$$

$$|B| = \|\mathbf{B}\|^2 = \langle \mathbf{B}, \mathbf{B} \rangle = \sum_k \mathbf{B}_k^2 \quad (6.12)$$

$$|A \cap B| = \langle \mathbf{A}, \mathbf{B} \rangle = \sum_k \mathbf{A}_k \mathbf{B}_k \quad (6.13)$$

$$|A \cup B| = |A| + |B| - |A \cap B| = \|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 - \langle \mathbf{A}, \mathbf{B} \rangle \quad (6.14)$$

For the *representations based on a continuous 3D molecular field*, the molecules  $A$  and  $B$  are characterized by point functions  $A(x, y, z)$  and  $B(x, y, z)$ , respectively. The size quantities for the similarity calculation are defined in basically the same way as for the vector-based representations. However, the dot products of vectors are replaced by the inner products of functions in a function space (*i.e.* their overlap integrals over volume) according to eqn (6.15)–(6.18). As mentioned above, the field functions can be represented as combinations of Gaussians to enhance the performance and stability.<sup>27–29</sup>

$$|A| = \langle A, A \rangle = \int [A(x, y, z)]^2 dV \quad (6.15)$$

$$|B| = \langle B, B \rangle = \int [B(x, y, z)]^2 dV \quad (6.16)$$

$$|A \cap B| = \langle A, B \rangle = \int A(x, y, z) B(x, y, z) dV \quad (6.17)$$

$$|A \cup B| = |A| + |B| - |A \cap B| = \langle A, A \rangle + \langle B, B \rangle - \langle A, B \rangle \quad (6.18)$$

Using the appropriate size quantities for the selected molecular representation, various similarity functions defined in Table 6.1 are calculated. For the most part, the application of these measures is quite straightforward. However, some points deserve additional discussion.

- (1) The combination of the Tanimoto similarity function with some kind of the molecular fingerprint (bit vector) representation works rather well in most applications and seems to be the most popular approach to the structural similarity analysis. In fact, it may be regarded as the '*default similarity measure*' that is implied in the absence of additional qualifications. Nevertheless, we have shown that many other approaches are possible in this field and may be more suitable to each particular problem (see also the discussion of the applicability domains of molecular similarity measures in Section 6.2.1.2).
- (2) For the representations involving bit vectors (*i.e.* vectors having 0 or 1 as component values), the same size quantities and thus the same similarity measures are obtained when using the equations defined for the set- and vector-based representations (except the potentially diminished performance in the latter case). However, for the multisets (*i.e.* vectors with non-negative integer components) the results from the set- and vector-oriented equations will be different.
- (3) If a representation involves only non-negative components (binary, integer, real, or functional), it can be shown that a specific ordering of various similarity measures is observed (eqn (6.19)).<sup>5,33,37</sup> However, care should be taken if negative component values are possible (*e.g.* for the lipophilicity, atomic charge, or molecular electrostatic potential). In this case, the intersection size  $|A \cap B|$  may also become negative, leading to more complicated relations (even the  $[0,1]$  bounds on the similarity values can be violated).<sup>29,37</sup>

$$0 \leq S_{\text{Tan}} \leq S_{\text{Min}} \leq S_{\text{Dice}} \leq S_{\text{Cos}} \leq S_{\text{Max}} \leq 1 \quad (6.19)$$

- (4) The Pearson correlation coefficient is closely related to the Cosine/Carbó similarity (it is calculated in the same way after centering the vectors by subtracting their mean values) and can also be used as a

**Table 6.1** Definitions of the commonly used similarity functions.<sup>a</sup>

Function	Definition	Notes
Tanimoto (Jaccard)	$S_{\text{Tan}} = T_c = \frac{ A \cap B }{ A \cup B } = \frac{ A \cap B }{ A  +  B  -  A \cap B } = \frac{c}{a + b + c}$	Traditionally called Tanimoto similarity in chemoinformatics context, although the same measure was proposed much earlier by Jaccard
Dice Hodgkin–Richards	$S_{\text{Dice}} = S_{\text{H-R}} = \frac{ A \cap B }{\frac{1}{2}( A  +  B )}$	Named after Dice for representations based on sets and graphs, after Hodgkin and Richards for representations based on vectors and functions. Denominator is the arithmetic mean of individual sizes
Cosine Carbó	$S_{\text{Cos}} = S_{\text{Car}} = \frac{ A \cap B }{\sqrt{ A  \cdot  B }}$	Equals to cosine of the angle between vectors. Named after Carbó for representations based on vectors and functions. Denominator is the geometric mean of individual sizes
Min	$S_{\text{Min}} = \frac{ A \cap B }{\max( A ,  B )}$	Denominator is the maximum of individual sizes
Max	$S_{\text{Max}} = \frac{ A \cap B }{\min( A ,  B )}$	Denominator is the minimum of individual sizes
Tversky	$S_{\text{Tve}} = \frac{ A \cap B }{\alpha A  + \beta B  + (1 - \alpha - \beta) A \cap B } = \frac{c}{\alpha a + \beta b + c}$ $S_{\text{Tve}} = \frac{ A \cap B }{\alpha A  + (1 - \alpha) B } = \frac{c}{\alpha a + (1 - \alpha)b + c}$	Asymmetric similarity function proposed by Tversky (depends on parameters $\alpha, \beta \geq 0$ )

<sup>a</sup>Calculation of the similarity functions for two molecules *A* and *B* is based on the following size quantities:  $|A|$ , size of molecule *A*;  $|B|$ , size of molecule *B*;  $|A \cap B|$ , size of their intersection (common part);  $|A \cup B|$ , size of their union. For the bit vector representations, the bit counts *a*, *b* and *c* represent the number of bits that are set only for molecule *A*, only for molecule *B* and for both molecules, respectively.

similarity measure, especially when the representation vector components are similar in nature and scale.

- (5) The Tanimoto similarity and most of the other symmetric similarity functions (with the possible exception of  $S_{\text{Max}}$ ) exhibit the so-called size-dependent behavior.<sup>5,38,39</sup> When the two molecules differ significantly in their size and/or complexity, the denominator of the similarity ratio becomes dominated by the contribution from the larger molecule. The resulting similarity values are not only lower than intuitively expected but also less sensitive to changes in the smaller molecule.
- (6) The *asymmetric similarity function* proposed by Tversky<sup>40</sup> (eqn (6.20)) allows control of the issues arising from the asymmetry between the molecules (which may be caused by differences in their sizes or roles). In fact, it is not a single function but a family of similarity functions dependent on two non-negative parameters ( $\alpha$  and  $\beta$ ) that specify the relative weights of the molecules  $A$  and  $B$  in the comparison. If  $\alpha = \beta$ , the resulting similarity function is symmetric. For example, the Tanimoto function is obtained by letting  $\alpha = \beta = 1$  and the Dice function by letting  $\alpha = \beta = 1/2$ . Often, a simplified form of the Tversky function is used that is obtained by letting  $\beta = 1 - \alpha$ , and thus involves only one asymmetry parameter (eqn (6.21)).<sup>41</sup> In the similarity-based virtual screening context (see Section 6.2.3.1), the computational experiments indicate that better results (*i.e.* higher hit rates) are obtained when the weight for the query molecule is close but not equal to unity.<sup>41</sup> One common option is using  $\alpha = 0.95$ .<sup>29</sup>

$$S_{\text{Tve}} = \frac{|A \cap B|}{\alpha|A| + \beta|B| + (1 - \alpha - \beta)|A \cap B|} \quad (6.20)$$

$$S_{\text{Tve}} = \frac{|A \cap B|}{\alpha|A| + (1 - \alpha)|B|} \quad (6.21)$$

Recently a new ‘second-order’ *Hausdorff-like similarity measure* was proposed,<sup>42</sup> which is applicable to complex chemical systems (*e.g.* mixtures or ionic liquids), to individual compounds represented by the sets of their metabolites or substructures, or to arbitrary sets of structures. For two sets  $A = \{a_i\}$  and  $B = \{b_j\}$  it is defined as a normalized sum of pairwise similarities between component structures according to eqn (6.22). This approach takes into account all components of the sets, providing more informative and intuitively desirable comparison.

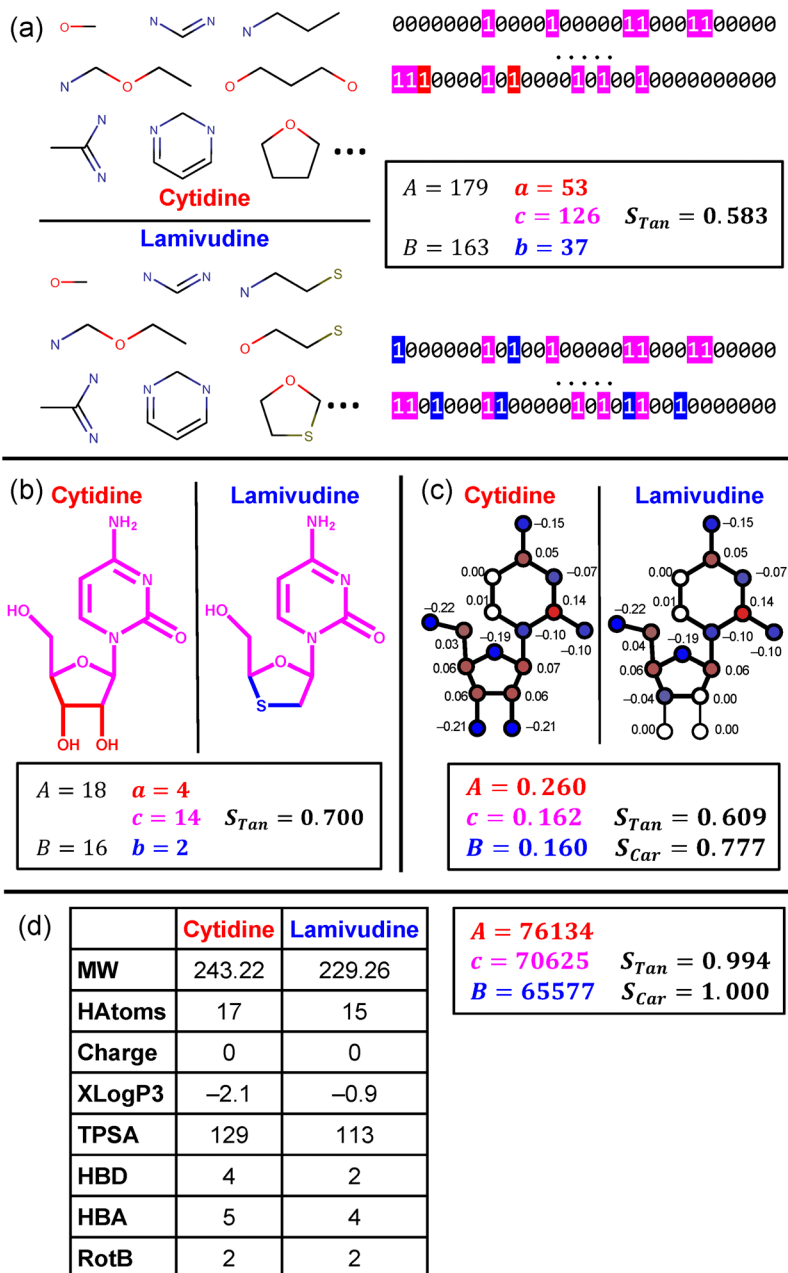
$$S_{\text{Hs}} = \frac{\sum_a \max_j S(a, b_j) + \sum_b \max_i S(b, a_i)}{|A| + |B|} \quad (6.22)$$

Concluding this section devoted to the definitions of molecular similarity measures, we will present an illustrative example. Figure 6.2 demonstrates the calculation of some measures discussed above for two compounds,

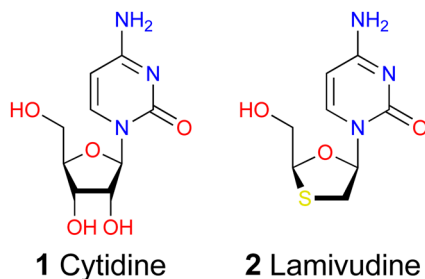
cytidine 1 and lamivudine 2 (see the structures in Scheme 6.1). In particular, the Tanimoto similarity based on the molecular fingerprint representation is shown in Figure 6.2a and the Tanimoto similarity based on the molecular graph representation is shown in Figure 6.2b. In Figure 6.2c, the Tanimoto and Carbó similarity functions are calculated for the vector representation of atomic charges on the topological (2D) structure level based on the MFTA supergraph. Finally, in Figure 6.2d, the Tanimoto and Carbó similarity functions are calculated for the vector representation involving a set of calculated 1D (global) molecular descriptors and predicted properties. It should be noted that most of the similarity values are moderate, although lamivudine would normally be perceived by a medicinal chemist as a quite close analog of cytidine (and was specifically developed as such). In addition, different values are obtained when using different similarity functions, representations and underlying parameters. In practical applications of the molecular similarity analysis (see Section 6.2.3) it is important to calibrate and/or validate a similarity measure before using it to make any decisions. The apparent very high similarity based on the 1D descriptors may be regarded as a kind of ‘artifact’ caused by using a small set of descriptors dominated by only two parameters with large absolute values (molecular weight and topological polar surface area). In practice, some prescaling or dimensionality reduction procedure based on a representative training set should be applied before using such descriptors in similarity calculations (see Section 6.2.1.3).

## 6.2.2 Chemical Spaces and Activity Landscapes

As noted in Section 6.1, the similarity property principle stating that similar structures (should) possess similar properties<sup>5,6</sup> plays a very important role in many areas of chemistry, pharmacology, and toxicology. However, despite its usefulness, this principle is a generalization rather than a fundamental law of nature. Thus, it has limitations and exceptions that may be caused by various factors. Recognition and analysis of this fact led to the introduction of the *activity landscape* (or, in general, *property landscape*) metaphor.<sup>47,48</sup> (The term *structure–activity landscape* is also commonly used.) It represents the relationship between the structures and activity values as a hypersurface over chemical space in the same way as the earth surface in a real landscape defines the relationship between the geographical locations and altitude values.<sup>49,50</sup> Using suitable dimensionality reduction and visualization techniques, this hypersurface can be represented graphically in 2D or 3D for convenient human perception.<sup>24,47</sup> Similar to real terrain, the activity landscape is far from being uniform. In some regions of chemical space it can resemble flat prairies or gently rolling hills where the structure-to-activity function is smooth or continuous and the similarity property principle is obeyed. However, in other regions this function can be steep or discontinuous, resembling rugged gorges or peaks. Such discontinuities have been termed *activity cliffs*, defined as pairs of structurally similar compounds having a large difference in activity.<sup>49,51,52</sup>



**Figure 6.2** Calculation of some similarity measures for the cytidine ( $A$ ) and lamiivudine ( $B$ ) molecules (Scheme 6.1). (a) Tanimoto similarity based on the molecular fingerprint representation implemented in ChemAxon Instant JChem.<sup>19,43</sup> Some of the substructural patterns (linear paths, branches, cycles) found in these molecules are shown on the left, the fragments of the 512-bit hashed fingerprint vectors are shown on the

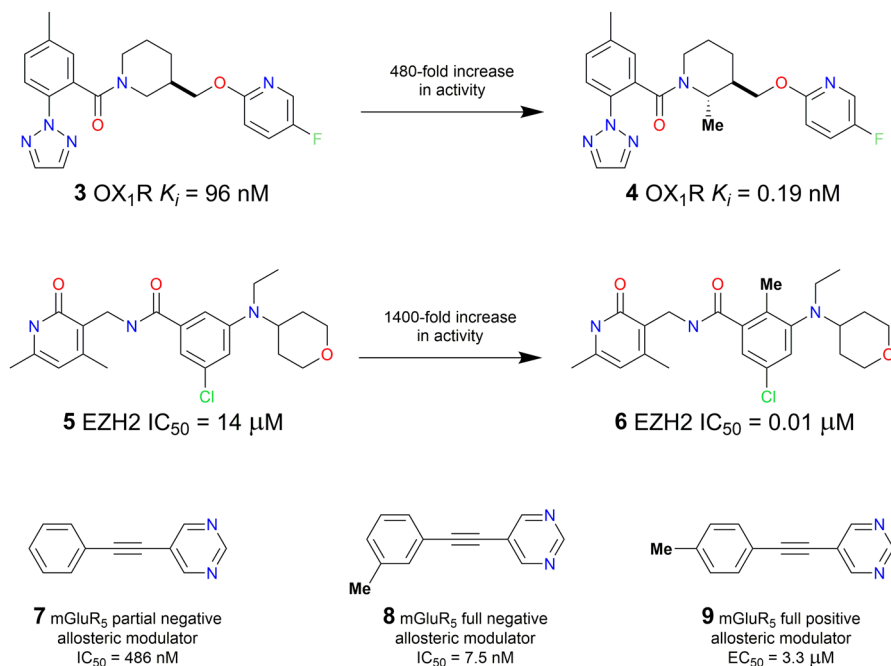


**Scheme 6.1** Structures of the compounds cytidine and lamivudine for the example of similarity calculations.

Among the most striking examples of activity cliffs is the so-called *magic methyl effect*, *i.e.* large changes in activity due to the introduction of only one (or few) methyl groups.<sup>53</sup> In some cases, the increase in potency (commonly attributed to modified conformational behavior, shape complementarity or solvation) can be very significant, up to 2 or 3 orders of magnitude. In Scheme 6.2 this is illustrated by the orexin receptor antagonists<sup>54</sup> 3 and 4, as well as by the histone methyltransferase EZH2 inhibitors<sup>55</sup> 5 and 6. It should be noted that usually additional methyls lead to more modest potency boosts and can just as likely cause a decrease in activity.<sup>53</sup> In other cases, their introduction

right. The bits set for both molecules are highlighted in magenta, the bits set only for molecules *A* or *B* are highlighted in red and blue, respectively. The size quantities (bit counts) and the similarity function value are shown in the box. (b) Tanimoto similarity based on the molecular graph representation. The intersection of the molecular graphs (maximum common edge substructure) is shown in magenta, the fragments present only in molecules *A* or *B* are shown in red and blue, respectively. The size quantities (edge counts) and the similarity function value are shown in the box. (c) Tanimoto and Carbó similarity functions based on the vector representation of atomic charges on the topological (2D) structure level using the Molecular Field Topology Analysis (MFTA).<sup>30,31,44</sup> The molecular supergraph and the Gasteiger charge values for the two molecules are shown. Atoms in the supergraph are colored according to the sign of the charge values: red for positive, blue for negative, and white for values close to zero. The size quantities (vector dot products) and the similarity function values are shown in the box. (d) Tanimoto and Carbó similarity functions based on the vector representation of calculated 1D (global) molecular descriptors and predicted properties. The parameter values obtained from the PubChem Compound Database for cytidine (CID 6175)<sup>45</sup> and lamivudine (CID 73339)<sup>46</sup> are shown in the table. (MW: molecular weight, g mol<sup>-1</sup>; HAtoms: heavy atom count; Charge: formal charge; XLogP3: calculated logarithm of octanol–water partition coefficient; TPSA: topological polar surface area, Å<sup>2</sup>; HBD: hydrogen bond donor count; HBA: hydrogen bond acceptor count; RotB: rotatable bond count). The size quantities (vector dot products) and the similarity function values are shown in the box.

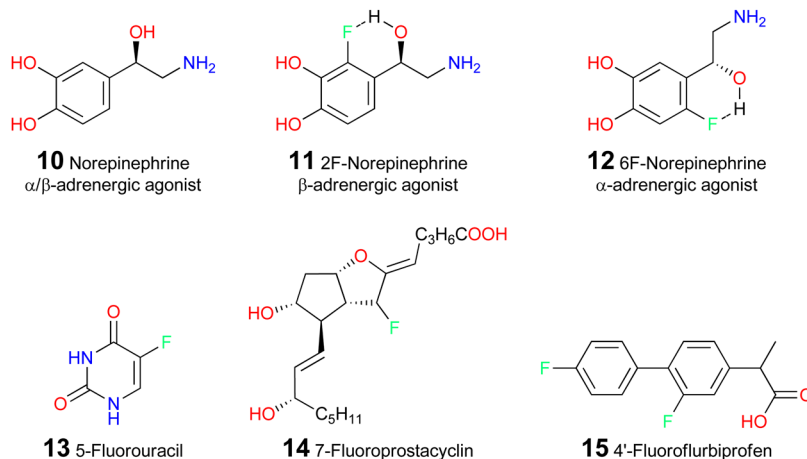




**Scheme 6.2** Examples of the activity cliffs caused by methyl substitution.

can lead to changes in selectivity profile, activity type (*e.g.* negative and positive allosteric modulators of mGluR<sub>5</sub> metabotropic glutamate receptor<sup>56</sup> 7–9), metabolism, or solubility.<sup>53</sup> The *fluorine substitution* can produce even more versatile effects mediated by its steric similarity to hydrogen and by the changes in electronic effects, conformational behavior, physico-chemical properties, metabolic stability, or reaction mechanisms.<sup>57,58</sup> For example, norepinephrine **10** and its fluorinated analogs **11** and **12** (Scheme 6.3) have different selectivity profiles towards  $\alpha$ - and  $\beta$ -adrenergic receptors,<sup>57,59</sup> probably due to the stabilization of different conformers by the OH...F hydrogen bond. 5-Fluorouracil **13** is a mechanism-based suicide inhibitor of thymidylate synthase. Similar to uracil, it is incorporated into the intermediate covalently bound to the enzyme, but cannot eliminate the oxidized cofactor dihydrofolate due to the lack of the C5 proton.<sup>57</sup> The fluorinated derivatives of drugs are also commonly used to control their metabolism. In 7-fluoroprostacyclin **14**, the electron-withdrawing fluorine atom helps to minimize the enol ether hydrolysis, increasing its half-life to >1 month, compared to 10 min for the parent compound.<sup>57</sup> Substitution of fluorine for hydrogen at metabolically labile sites prevents their oxidation by the cytochrome P450 (*e.g.* in 4'-fluoroflurbiphen **15**).<sup>60</sup>

The activity cliffs usually have both 'good' and 'bad' sides.<sup>49</sup> The most significant problems they cause occur during the modeling of the structure–activity relationships.<sup>48</sup> When the property landscape is steep or discontinuous,



**Scheme 6.3** Examples of the activity cliffs caused by fluorine substitution.

the approaches used in this field usually have poor performance or require models of much higher complexity, resulting in significant prediction errors for cliff and other compounds.<sup>23</sup> The presence of activity cliffs also complicates the identification of active hit compounds during the screening of compound libraries. Only a diverse subset of limited size is usually tested to save the resources, and the compounds with small structural differences (such as methyl substituent in compounds 5 and 6) may easily be missed.<sup>55</sup> In contrast, the activity cliffs (*i.e.* structures with much higher activity than might be expected based on other known data) open interesting possibilities for further discovery and optimization of promising compounds as well as for better understanding of their mechanism of action and modeling of the structure–activity relationships.<sup>49,51,61,62</sup>

Although the concept of activity cliffs is widely accepted, the detailed criteria for their recognition can be debated and numerous approaches have been proposed. One of them involves the calculation of the *Structure–Activity Landscape Index* (SALI) for each pair of compounds.<sup>63,64</sup> Somewhat reminiscent of the estimated gradient, it is defined as the absolute difference in activity divided by the similarity-based distance between molecules (eqn (6.23)), usually normalized relative to the largest SALI value in the data set. If this parameter exceeds a specified threshold, the pair is considered as an activity cliff. In addition, it is possible to analyze the cliff relationships in the data set as a whole (represented by the so-called SALI networks) and estimate how well these relationships are predicted by different models.<sup>64,65</sup> The SALI parameters can also be used to estimate the statistical significance of the activity cliffs detected from a particular molecular representation.<sup>66</sup> An alternative approach to the classification of the structure–activity relationships and landscapes (continuous, discontinuous, and heterogeneous) is based on the *structure–activity relationship index*.<sup>67</sup>

$$\text{SALI}(A,B) = \frac{|\text{Act}(A) - \text{Act}(B)|}{1 - S(A,B)} \quad (6.23)$$

The SALI and similar parameters are quite useful for the analysis of activity cliffs. However, they also have certain disadvantages. Firstly, they are based on the actual numerical values of activity and similarity measures. However, as shown in Section 6.2.1 above, there is no single ‘true’ measure of molecular similarity (or molecular description in general). Different descriptions and similarity measures focusing on different facets of a structure may be more or less relevant to different properties (in particular, structural characteristics critical for a property or activity of interest may be missing from the description or masked by other, insignificant features). Thus, compound pairs classified as activity cliffs using one combination of molecular representation, similarity function and activity measure may not be detected as such using some other combinations. In addition, the calculated similarity values (especially based on the molecular fingerprint representations) are not always intuitive and may be difficult to interpret from a chemical perspective.<sup>68</sup> This has even lead to a (somewhat provocative) question of whether all activity cliffs might actually be the artifacts of using inadequate molecular description.<sup>49,69</sup> Indeed, in many cases we can hope to select and/or optimize the description and modeling techniques (using available information on the targets and mechanisms of action, previous experience, educated guess, trial-and-error, and/or automated learning techniques) in order to better capture the relevant structural features and correctly predict the activity cliffs that were mispredicted using other approaches.<sup>52,70–73</sup>

Additional problems with the SALI parameter are caused by its relative scale and the singularity at  $S = 1$ . As a result, it can distort the actual magnitude of the activity differences, leading to the detection of irrelevant (minor) cliffs. For compounds having the same representation the SALI values are infinite, and for highly similar compounds they are extremely sensitive to small variations in the similarity value which are not really significant (see Section 6.2.1).<sup>49,68</sup> Thus, the discrete activity cliff criteria were proposed that are directly based on the general definition.<sup>13,49,51,61</sup> A pair of compounds should be considered as activity cliff if it meets the following conditions (which can of course be adjusted to a particular problem but need to be clearly specified):

- (1) both compounds are active, their potency is characterized by  $K_i$  (preferably) or  $\text{IC}_{50}$  values and at least one compound has potency in the nanomolar range. (Alternatively, confirmed inactive compounds may also be considered);<sup>74</sup>
- (2) the compounds satisfy a pre-established similarity criterion; and
- (3) the potency of compounds differs by at least 2 orders of magnitude (2 logarithmic units).

These conditions allow us to define the similarity criterion explicitly, taking into account the features of a problem and the desired level of interpretability.<sup>13,49,61,75</sup> One option is to require the value of the fingerprint-based Tanimoto similarity (or other similarity measure) above a certain threshold, but limits on the structural modification or even on the similarity of 3D structures or binding modes can also be specified.<sup>76,77</sup> The so-called MMP cliffs<sup>68,75</sup> (inspired by the *matched molecular pair* concept, see Section 6.3.2) proved to be especially useful, easily interpretable, and chemically intuitive. In this approach, the two structures are considered similar if they differ only in one substructural fragment (terminal substituent or central core), defining a structural transformation. The size of the transformations is restricted to keep them chemically acceptable: the transformed fragments cannot exceed 13 non-hydrogen atoms and must be at least two-fold smaller than the unchanged part of the molecule, and the difference in fragment size cannot exceed eight non-hydrogen atoms. These limits were chosen to allow the addition of a substituted six-membered ring (*e.g.* a phenolic substituent) or the replacement of a five- or six-membered ring by a substituted condensed two-ring systems containing up to 10 ring atoms. (If the activity difference in a matched molecular pair is small, the transformation corresponds to a bioisosteric replacement.)

The distribution of the activity cliffs was thoroughly analyzed<sup>61,68,75</sup> using the MMP and fingerprint/Tanimoto similarity criteria and the high-confidence activity data from the ChEMBL<sup>78</sup> database. The results indicate that the activity cliffs are in fact fairly common, rather than rare and exceptional. For all analyzed targets the proportion of the cliff-forming compounds determined from the different fingerprint similarity measures ranges from 34% to 41%. Based on the MMP similarity, it was lower (close to 28%), but approximately one out of three compounds was still involved in one or more activity cliffs. Interestingly, the consistency between different measures was rather poor (only ~15% of the compounds are recognized as cliff-forming by all fingerprint criteria and ~11% by all fingerprint and MMP-based criteria). Thus, the application of structurally conservative and chemically interpretable MMP similarity measures seems preferable.<sup>61</sup> For different targets some variability in the activity cliff distributions is present, but the differences are not substantial. Based on the MMP similarity, the proportion of cliff-forming compounds ranges from 20% to 38% and the proportion of the cliffs (relative to all compound pairs meeting the similarity criteria) ranges from 5% to 15%.<sup>75</sup> It should be noted that most of the activity cliffs (>95%) are not isolated, but form disjoint clusters of various topologies that may include up to dozens of compounds (so-called coordinated cliffs).<sup>79</sup> Some techniques to extract the structure–activity relationship information from the coordinated cliff networks have been proposed.<sup>52,80</sup> This approach seems quite promising, since historical analysis shows that the compound optimization paths originating from the cliff compounds are much more likely to reach the most potent compounds in the series,<sup>52,61,81</sup> but these possibilities are severely underutilized in the practice of medicinal chemistry.<sup>61,82</sup>

### 6.2.3 Applications of Molecular Similarity Analysis

In this section, we will consider several areas of the computational pharmacology and toxicology where the approaches based on the molecular similarity analysis have been successfully used. Most of them are explicitly or implicitly based on the similarity property principle discussed in Section 6.2.2. Thus, it is advisable to bear in mind all the limitations and caveats mentioned there. It should also be noted that the similarity-based approaches, in contrast to the QSAR analysis and related techniques, do not construct specialized prediction models (*e.g.* using machine learning methods). Instead, their ‘model’ is simply a set of suitably represented molecules, sometimes with associated property/activity information.

#### 6.2.3.1 Similarity-Based Virtual Screening

The *virtual screening* (often also called *in silico screening*) can be defined as the application of chemoinformatics techniques to analyze large libraries of chemical compounds in order to identify potentially promising structures during drug discovery and development. The virtual screening workflows may involve a variety of approaches based on the available target and ligand information, as well as the filters taking into account the predicted ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, drug-likeness estimates, and other data.<sup>83</sup> Importantly, the virtual screening is not expected to provide any guarantee that the screened compounds will indeed be active or have other desired properties, or that no active compounds will be missed. Instead, it ranks the library (database) compounds according to some estimated preference in such a way that the compounds closer to the top of the list (should) have a higher chance of being active. In other words, the top (priority) list is *enriched* in active compounds. In addition to the correctly predicted active compounds (*true positives*) and inactive compounds (*true negatives*), two kinds of errors are possible. *False positives* are the inactive compounds erroneously ranked as active, and *false negatives* are the active compounds erroneously predicted to be inactive. During the design of the virtual screening procedures, an important step is their *validation*, *i.e.* estimation of the recognition accuracy for active and inactive compounds and of the *enrichment factors* for top-ranking hits.

In the *similarity-based virtual screening*, the ranking of the library compounds is based on their calculated similarity to some known active compound (reference or probe molecule). This process was traditionally called the *similarity search*. It is assumed that the similarity property principle is obeyed, so the compounds sufficiently close to the reference in chemical space should also possess some activity.<sup>84,85</sup> Historically, this was one of the first and most prominent applications of the molecular similarity analysis in chemistry, and most of the research on the development and optimization of the similarity measures (see Section 6.2.1) was aimed at improving its performance and efficiency. Of course, compared to other, more sophisticated

virtual screening methods, the similarity search is a rather rough technique that is likely to produce a lot of both false positives and false negatives (caused by the irregular activity landscapes and/or by the insufficiently relevant similarity measures). Nevertheless, it can be quite useful in the beginning of a drug discovery project when no additional information on the target and ligands is available beyond the structures of one or a handful of hit compounds.<sup>85</sup> In such situations any guidance in choosing the following steps of research to map the structure–activity relationship would be valuable. In addition, the similarity search often has better computational performance.

The parameters of the similarity search (in particular, the molecular representation, similarity function, and the desired similarity level) should be chosen in such a way as to minimize the number of errors and to avoid a trivial situation where only the reference molecule itself will be found. The optimization and calibration of these parameters was a topic of extensive research.<sup>5,6,20–22</sup> A common rule of thumb based on an early seminal study<sup>86</sup> suggests that the Tanimoto similarity value  $S_{\text{Tan}} \geq 0.85$  reflects a high probability that two compounds would have the same activity. However, subsequent research has shown that the problem of *activity-relevant similarity values* is much more complicated (and even led to calling this rule a ‘0.85 myth’).<sup>6</sup> As discussed in Section 6.2.1, different molecular representations and different fingerprint schemes result in substantially different similarity values for the same molecules.<sup>6,22</sup> In addition, the distribution of the similarity values (diversity) and the sensitivity of activity to structural changes significantly depends on a particular activity class and data set.<sup>6,22</sup> In a more thorough statistical analysis it is shown that the values of  $S_{\text{Tan}} \geq 0.8$  (for MACCS fingerprints) or  $S_{\text{Tan}} \geq 0.3$  (for ECFP4 fingerprints) are much more likely for compounds sharing the same activity than for the comparison of two randomly selected compounds or one active and one randomly selected compound. Nevertheless, these values could not be reliably applied as distinct thresholds for the similarity search as the number of false positives and false negatives would be high and hardly predictable.<sup>22</sup> An interesting approach aiming to improve the virtual screening performance by augmenting the usual activity-independent similarity measures with background knowledge was proposed.<sup>87</sup> An extension component in such ‘hybrid’ measures is calculated as a fingerprint-based similarity using structural fragments that are believed to be required or beneficial for activity.

### 6.2.3.2 Activity Prediction

A number of approaches directly employ the molecular similarity estimates to predict the activity or property values based on the known data for similar compounds. One of them is the *k nearest neighbors* (KNN) method,<sup>23,88</sup> which can be used for discrete (classification) and continuous endpoints. For a compound of interest, the *k* nearest compounds (most similar in terms of the selected molecular representation and similarity function) are found in a dataset of compounds with known activity. The predicted activity is then

determined by some kind of a consensus, voting or averaging procedure. A lot of variations of this approach using different techniques for similarity estimation and final prediction have been proposed.<sup>88</sup> High prediction accuracy can be achieved if the activity landscape is sufficiently smooth and the desired applicability domain in chemical space is adequately covered by the compounds with known activity. In addition, unreliable predictions due to the absence of sufficiently close analogs can be detected. On the other hand, the KNN method is quite sensitive to the choice of similarity measure<sup>88</sup> and the prediction model has to include all the known structure–activity data, which can have computational performance and intellectual property implications. Nevertheless, the method has been used successfully for the prediction of various compound properties, for example, G-protein coupled receptor ligand binding,<sup>89</sup> total drug clearance in humans,<sup>90</sup> and acute toxicity in fish.<sup>91</sup> The conceptually related but more general *read-across* approach<sup>92,93</sup> predicts an activity or property endpoint for a compound by defining a category of similar chemicals and interpolating the endpoint from known data within this category. In contrast to the KNN, in different problems the formation of a category may be based on various criteria such as molecular similarity, structural congenerity, common mode of action, *etc.* The interpolation methods also may range from simple consensus, voting, or averaging procedures to ‘local’ (category-specific) structure–property models.

In the *Similarity Ensemble Approach* (SEA) the similarity relationships between the ligand sets of different targets are analyzed.<sup>94–96</sup> The fingerprint-based Tanimoto similarity values for all ligand pairs having similarity above a specified threshold are summed up to produce a raw score, which is then adjusted for the ligand set sizes and transformed into the expectation value that estimates the statistical significance of observed similarity compared to random data. These parameters can be useful in two ways. Firstly, they define a network (or a tree) of similarity relationships between the ligand classes that is independent of the sequence similarity trees for target proteins (analyzed in bioinformatics) and can complement them, providing valuable insights into the pharmacologically interesting chemical space. For example, in the SEA tree the ligands of different ionotropic glutamate receptor ligands are closely related, as are the different classes of antifolate drugs (even though their targets are dissimilar by sequence). In contrast, the ligands of the ionotropic and metabotropic serotonin receptors are separated. Secondly, the similarity expectation values calculated for a single compound with respect to different ligand classes can be used to estimate the likelihood of its activity towards their respective targets. This allows us to predict potential off-targets and adverse effects for known drugs or new indications for drug repositioning, as well as to identify possible mechanisms of action for compounds found by phenotypic screening.<sup>97,98</sup> In many cases the predicted activities were confirmed experimentally.<sup>99–103</sup> The SEA web service is available online.<sup>104</sup> Recently a conceptually related method called *CSNAP* (*Chemical Similarity Network Analysis Pulldown*) was proposed and its predictions of the targets for new antimitotic compounds were validated



in the experiment.<sup>105</sup> This tool is also available online.<sup>106</sup> (Of course, other approaches to target identification based on the compound structures are also possible. Some of them are considered in Section 6.3.2.2).

The similarity-based activity estimation is also used implicitly to solve an important problem during the validation of the virtual screening procedures (see Section 6.2.3.1). Since the data on the experimentally confirmed inactive compounds are scarce or unavailable for most of the targets, a common practice employs the so-called *decoy* compounds that are structurally dissimilar to the known actives but close to them in basic physico-chemical properties and structural descriptors (*e.g.*, molecular weight, lipophilicity, number of hydrogen bond donors and acceptors, *etc.*).<sup>107,108</sup> In the original benchmarking database *Directory of Useful Decoys (DUD)*, a simple threshold for the fingerprint-based Tanimoto similarity value was used ( $S_{\text{Tan}} < 0.9$  for CACTVS fingerprints).<sup>107</sup> However, often this could not eliminate all the fragments and functional groups critical for the target binding, leading to underestimated screening performance. Thus, the refined procedure used in the *Directory of Useful Decoys, Enhanced (DUD-E)* database includes a more restrictive filter, removing the 75% most similar compounds by the ECFP4 fingerprints.<sup>108</sup>

### 6.2.3.3 Clustering, Networks and Diversity

The molecular similarity measures reflect the closeness of compounds in chemical space and can be used to analyze the intrinsic structure of the compound data sets. Two basic approaches have been proposed. In a more traditional *clustering* approach, the similarity values (or the equivalent distances) are used to build a 'natural' hierarchical classification in such a way that the compounds in each class (cluster) are more similar to each other than to any compounds in other clusters.<sup>23,109,110</sup> The activity information may be overlaid onto tree-like classification diagrams (dendrograms) to analyze how the structural changes in chemical space translate into the biological space.<sup>110</sup> Alternatively, both the structure and activity similarities may be combined during cluster analysis.<sup>110</sup> The clustering information for a data set (*e.g.* a compound library) is commonly used to select the most *diverse and/or representative subset* comprising a minimal (or specified) number of compounds, *e.g.* by taking one representative from each cluster.<sup>109</sup> For example, this procedure can be applied to select a library of compounds for experimental screening in such a way as to cover the desired region of chemical space with minimal resource expenditures<sup>83</sup> or to select a representative training set for QSAR analysis and estimate the predictivity and the applicability domain of the resulting model.<sup>111</sup> Apart from the cluster-based approaches, some algorithms for the diversity subset selection operate directly on the pairwise distances between compounds,<sup>109</sup> *e.g.* by sequentially choosing the compounds most distant to all the previously selected compounds.

Another approach to the analysis and visualization of the similarity structure in a set of compounds involves the construction of the so-called *Chemical Space Networks (CSN)*.<sup>24</sup> They provide a coordinate-free (graph-based)



representation of chemical space wherein the nodes represent compounds and the edges represent pairwise similarity relationships. An edge connecting two compound nodes may be introduced if their similarity value is above a certain threshold (usually determined in such a way as to attain a specified edge density, *i.e.* proportion of the present and possible edges). The Tanimoto and Tversky measures based on the molecular fingerprints or the MCES size have been used (see Section 6.2.1.4). Alternatively, an edge may simply encode that the compounds form a restricted-size matched molecular pair (see Section 6.2.2). The CSNs (especially with the overlaid activity information) can support the navigation and visualization of the biologically relevant chemical space.<sup>112–115</sup>

### 6.3 Quantitative Structure–Activity/Property Relationships (QSAR/QSPR)

This section presents a number of approaches to the analysis of the QSAR/QSPR and their applications to some problems in computational pharmacology and toxicology. Observing the current trends in the field, we could define QSAR/QSPR in a broad sense as the *use of machine learning and other computational techniques to derive predictive models from the experimental activity or property data for chemical compounds*. (Strictly speaking, this definition also covers some of the similarity-based techniques discussed in Section 6.2.3). This process usually employs an intermediate representation of chemical structures by a suitable set of numerical *molecular descriptors* characterizing relevant facets of a structure. In general, the QSAR/QSPR workflow involves six basic steps: (1) collection and curation of a *training set* containing structures of the compounds, experimental activity or property *endpoint* values and other associated data; (2) calculation of molecular descriptors; (3) generation of machine learning models; (4) evaluation of model quality and acceptability; (5) interpretation of the model; (6) prediction of endpoint values for ‘new’ compounds of interest (not used in derivation of the model).

In the drug design context, the training and prediction set compounds are the ligands interacting with a biological target. However, the target structure information is only implied rather than used in the QSAR analysis (and may not be available at all). Thus, this approach is often called *ligand-based design*. Some of the QSAR/QSPR approaches are closely related to the molecular modeling as well as to the *structure-based drug design* methods that directly operate with the target structure. In a real research project, all the applicable tools can—and should—be used together, reflecting the interdisciplinary power of chemoinformatics.

For a more detailed and systematic account of the QSAR/QSPR field in general, we can suggest the excellent introductory<sup>116</sup> and intermediate<sup>117</sup> texts. In addition, a recent review<sup>118</sup> by some of the leading chemoinformatics experts presents a short history of QSAR evolution, an overview of the current trends, challenges, and prospects, as well as the guidelines for development,

validation and application of the QSAR models. Another review<sup>119</sup> presents current state-of-the-art and future trends of QSAR and discusses its role in the field of drug discovery and design.

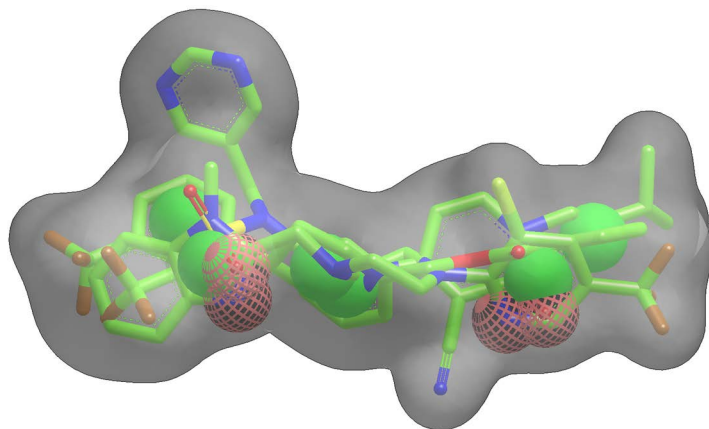
### 6.3.1 Congeneric Series and Consistent Mechanisms

The approaches to the analysis of structure–activity relationships discussed in this section are designed primarily for reasonably congeneric and/or compact series of compounds that can be expected to share a consistent mechanism of action mediated by binding to the same target in basically the same binding mode. Of course, these requirements are not strict, and good or at least acceptable results can sometimes be obtained for relatively large and diverse datasets<sup>120,121</sup> and even for physico-chemical properties,<sup>122</sup> but this usually requires additional effort and success is not guaranteed. On the other hand, the models derived by these methods are often more readily interpretable and consistent with the structure of a biological target.

The *pharmacophore* concept<sup>117,123</sup> has evolved substantially since the early 1900s, when it was first proposed.<sup>124</sup> Originally inspired by the chromophore concept in the dye theory, this term was used to denote the chemical groups (in modern terms, scaffolds, substituents, or functional groups) responsible for the activity. By the 1960s, this approach (now called 2D or topological pharmacophores) was recognized as overly simplistic for drugs, although it remains quite useful in the form of the *toxicophores*, *i.e.* functional groups and/or structural fragments related to the toxic (*e.g.* mutagenic) action of a compound<sup>125</sup> (for detailed discussion of the toxicophore concept see Chapter 7). For target-mediated drug action, the 3D pharmacophore approach was proposed, and the pharmacophore is now defined by the International Union of Pure and Applied Chemistry as an ‘ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response’.<sup>126</sup> The pharmacophore model comprises a 3D arrangement of features (often called pharmacophore centers or elements) responsible for various interaction types, for example, positive and negative charges, donors and acceptors of hydrogen bonds, hydrophobic groups, ring systems, *etc.* During the model construction, a common or optimal pattern of the centers for a set of active molecules is identified. Traditionally, models with a rather limited number of centers (often three to five) represented by simple tolerance spheres were used.<sup>127,128</sup> The steric requirements were either ignored or represented by exclusion spheres.<sup>123,127</sup> Nevertheless, even with such limitations, the pharmacophore models can provide useful insights into the ligand structural features important for activity, as well as an approximate but very fast virtual screening technique. A valuable advantage of the pharmacophore-based screening is its support for the so-called *scaffold hopping*, *i.e.* the identification of potential ligands that match the same pharmacophore model (and thus could bind the same target), but have a new chemical scaffold compared to the previously known active compounds.

A significant breakthrough in these areas can be attributed to the use of the ‘soft’ matching techniques based on the Gaussian functions (see Sections 6.2.1.3 and 6.2.1.4) in the *Rapid Overlay of Chemical Structures* (ROCS) approach.<sup>27–29,129</sup> Originally developed for fast matching of molecules by shape, it now takes into account not only the common steric features of the active molecules but also their chemical features (interaction centers represented by ‘color atoms’). Depending on a particular problem, molecules can be matched, aligned, and screened using the Tanimoto or Tversky similarity functions taking into account only shape features, only color features or their combination. Although the ROCS authors still prefer to view it as primarily a ‘shape-based similarity tool’, in our opinion it can also be considered as a new level in the development of the pharmacophore concept that solves some of its issues outlined above by combining some elements of the molecular similarity analysis, QSAR (ligand-based modeling), and molecular modeling. In a recent virtual screening benchmark study,<sup>130</sup> it demonstrated the best recognition quality and enrichment among the tested 3D ligand shape comparison methods. An example of the pharmacophore model (shape-based virtual screening query) for the positive allosteric modulators of the mGluR<sub>2</sub> metabotropic glutamate receptor<sup>131</sup> built using the OpenEye vROCS 3.1.2 software<sup>132</sup> is shown in Figure 6.3.

Another promising extension of the pharmacophore concept derives a consensus model from the snapshots obtained during the molecular dynamics simulations of the protein–ligand complexes. This approach can take into account the *dynamic flexibility* of a system, eliminate the dependence on a

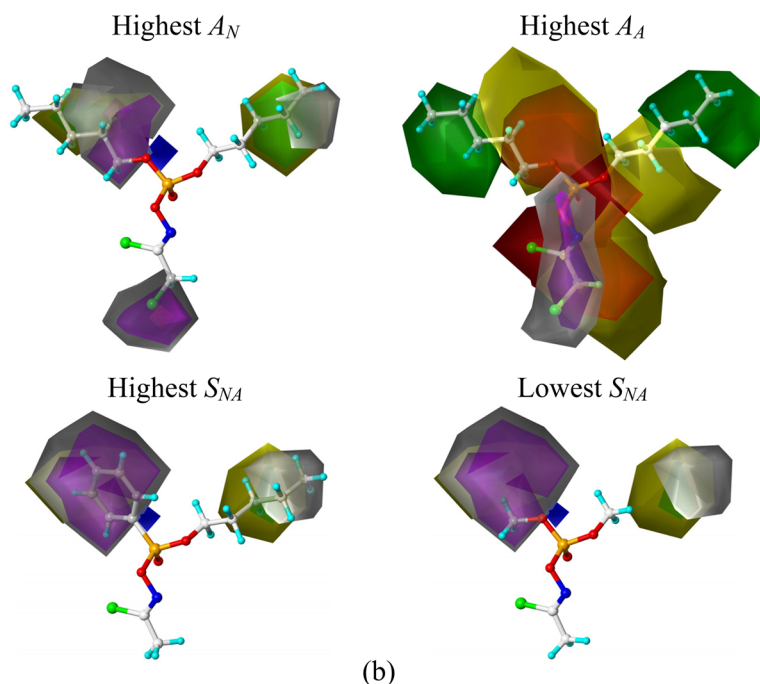
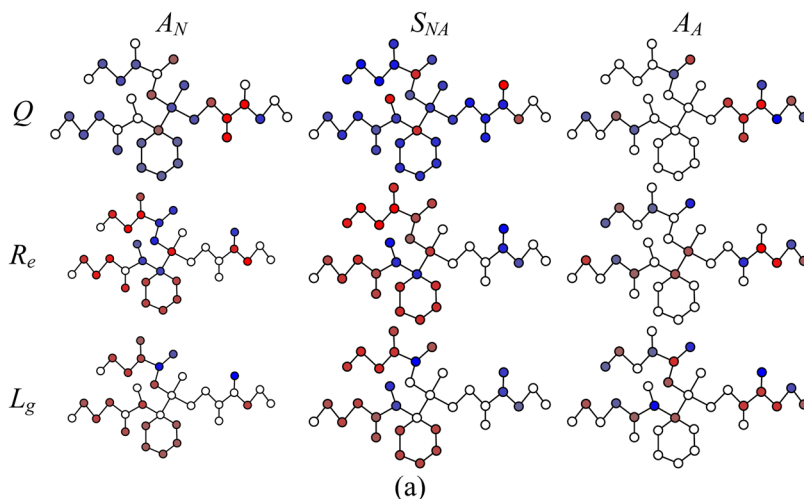


**Figure 6.3** Pharmacophore model (ROCS shape-based query) of positive allosteric modulators of mGluR<sub>2</sub> metabotropic glutamate receptor. The structures of the reference compounds are represented by the stick models. The shape of a sterically favorable region is shown in light gray. The green balls mark the positions of the centers of ring systems, and the pink mesh balls show the positions of the hydrogen bond acceptors. (Reproduced with permission from ref. 131.)

single source structure and identify the most stable and significant pharmacophore features.<sup>133</sup>

In *Molecular Field Topology Analysis (MFTA)*,<sup>30,31</sup> a quantitative structure–activity model is based on the local physico-chemical parameters that reflect major types of interactions involved in binding of small molecule ligands to biotargets (and can be quickly evaluated from a structural formula of a ligand). In particular, effective atomic charges are usually used to characterize the potential electrostatic interactions, van der Waals radii of atoms and groups for steric interactions, local lipophilicity for hydrophobic interactions, and the hydrogen bond donor and acceptor ability for hydrogen bonding. To bring these molecular descriptors for different structures into a common frame of reference, it builds a so-called molecular supergraph, *i.e.* a kind of topological network such that all structures belonging to a series under study can be superimposed onto it. Thus, the MFTA method is free from the conformational flexibility and molecular alignment problems that often arise in the 3D QSAR approaches (see later). Nevertheless, the training set should contain reasonably congeneric compounds with a common basic scaffold. From the supergraph-based uniform descriptor set, a predictive QSAR model is derived using the partial least squares regression (PLSR)<sup>23</sup> or other machine learning methods. In addition, graphical activity and selectivity maps help to visualize and analyze the effects of various local properties in different structure positions. This information can be used to suggest significant ligand–target interactions and guide the design of novel promising structures in a chemically intuitive way. The structural variability information encoded in a supergraph can also be used to generate a representative structure library for virtual screening within the model applicability domain.<sup>134</sup> The MFTA method was successfully used by our group<sup>120,134–143</sup> and others<sup>144–147</sup> in the activity and selectivity modeling, design, and virtual screening of promising ligands of various enzymes and receptors as well as viral fusion inhibitors. An example of the MFTA activity and selectivity maps for the acute and delayed neurotoxicity of *O*-phosphorylated oximes mediated by the acetylcholinesterase and neuropathy target esterase (NTE) inhibition<sup>138,140</sup> is shown in Figure 6.4a, indicating an increase in selectivity to NTE due to the presence of the phenyl and longer alkoxy groups attached to the phosphorus atom and to the increase in the total lipophilicity.

The 3D QSAR analysis, as the name suggests, aims to build a QSAR model on the basis of descriptors characterizing the spatial (three-dimensional) structures of molecules.<sup>117</sup> Instead of compressing them into a few parameters such as molecular volume or length of a substituent, it operates directly with the molecular interaction fields. In the classical and immensely popular *Comparative Molecular Field Analysis (CoMFA)*<sup>25</sup> approach the descriptors are the steric (van der Waals) and electrostatic (Coulomb) interaction energies of aligned molecules with an imaginary probe atom or group (often  $\text{CH}_3^+$ ) located at the nodes of a rectangular 3D grid. A predictive partial least squares regression (PLSR)<sup>23</sup> model built from these values can be used to estimate the activity for other congeneric and aligned molecules. In addition, the key



**Figure 6.4** Activity and selectivity maps for acute and delayed neurotoxicity of *O*-phosphorylated oximes mediated by the acetylcholinesterase (AChE) and neuropathy target esterase (NTE) inhibition. Endpoints:  $A_A$ ,  $A_N$ : inhibitor activity towards AChE and NTE, respectively;  $S_{NA}$ : selectivity to NTE over AChE. (a) Molecular Field Topology Analysis (MFTA) models. In the positions of molecular supergraph marked with the red circles an increase in the descriptor values is favorable for the activity (selectivity); conversely, in positions marked with the blue circles an increase in the descriptor tends to decrease the activity. Intensity of colors reflects

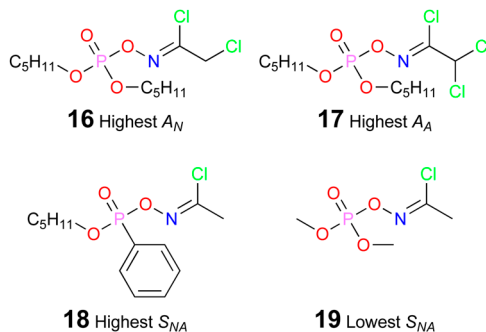
model features may be visualized in convenient activity maps representing favorable and unfavorable regions for steric interactions and positive or negative charges that are helpful during the model interpretation and activity optimization.<sup>25,117,148</sup> The utility of this approach is confirmed by hundreds of CoMFA models published in the literature, often with experimental validation of the predicted activities and/or good agreement with known target structures. Among its extensions, the *Comparative Molecular Similarity Indices Analysis (CoMSIA)* method should be mentioned, which follows the same workflow but uses uniform descriptors for the steric, electrostatic, hydrogen bonding, and hydrophobic interaction fields.<sup>26,117</sup> An example of the CoMSIA activity and selectivity maps for acute and delayed neurotoxicity of *O*-phosphorylated oximes<sup>140</sup> obtained using the Certara Sybyl<sup>149</sup> software is shown in Figure 6.4b. Consistent with the MFTA results, they also indicate the NTE preference for the phenylphosphonate groups and longer alkylphosphate chains.

As noted earlier, the correct and productive application of the CoMFA/CoMSIA approach implies a consistent alignment of all fragments of the molecules. Attaining it can be problematic, especially for conformationally flexible structures. Historically this required a careful and tedious manual preparation of each structure<sup>140</sup> or using an alignment based on the pharmacophore search or docking<sup>121</sup> procedures (with all the uncertainty inherent in these approximate methods). To solve this problem, the *topomer CoMFA*<sup>150,151</sup> and *template CoMFA*<sup>152–155</sup> approaches based on robust rules for conformer generation and alignment were proposed. The published results of their application seem very promising. In some cases, the alignment consistency and model quality were even better than the results obtained from alignments based on the X-ray structures of ligand–target complexes (which can involve substantial noise caused by the ligand and protein flexibility as well as by the changes in binding mode).<sup>153</sup>

The *Open3DQSAR* project (including *Open3DALIGN* and *Open3DGRID*) is a promising open-source initiative that could be used to implement the classical CoMFA workflow including conformer generation, structure alignment, calculation of molecular field descriptors and derivation of the machine

---

the magnitude of the influence. Local descriptors:  $Q$ : effective atomic charge;  $R_c$ : effective van der Waals radius taking into account the steric requirements of the central non-hydrogen atom and other attached atoms;  $L_c$ : group lipophilicity taking into account the contributions of the central non-hydrogen atom and the attached hydrogen atoms. (b) Comparative Molecular Similarity Indices Analysis (CoMSIA) models. In green-colored regions additional steric bulk is favorable for activity and in yellow-colored regions it is unfavorable. In blue-colored regions the activity tends to be increased by additional positive charges and in red-colored regions by additional negative charges. In magenta-colored regions the activity tends to be increased by hydrophobic groups and in gray-colored regions by hydrophilic groups. The structures with the highest activity or selectivity towards NTE and AChE (16–19, see Scheme 6.4) are shown for reference.



**Scheme 6.4** Structures of the *O*-phosphorylated oximes with the highest activity or selectivity towards neuropathy target esterase (NTE) and acetylcholinesterase (AChE).

learning models.<sup>156–158</sup> A number of interesting algorithms as well as interfaces to popular software are provided, although the alignment approaches as well as the integration and ease of use of these tools should be improved.

### 6.3.2 Diverse Series and Big Data

This section is focused on the approaches to the analysis of structure–activity relationships that usually work best for large and diverse sets of compounds. Of course, nothing prevents their application to compact or congeneric sets (provided that the model complexity is properly controlled to avoid overfitting and chance correlations),<sup>23</sup> but it involves some trade-offs that could lead to suboptimal results. The methods specifically designed for congeneric sets of compounds with consistent mechanisms of action and binding modes generally would provide more simple and easily interpretable models. Conversely, the methods oriented towards diverse compounds commonly use more generic molecular descriptors (in many cases, various fragmental descriptors, see Section 6.2.1.3) and sophisticated machine learning techniques that need sufficiently rich and representative data. In addition, lower prediction accuracy can be tolerated in exchange for broader applicability domain that is more likely to contain activity cliffs (however, good models still can achieve the level of accuracy comparable to the accuracy of the experimental data).<sup>159,160</sup>

In view of these factors, in the following subsections we will discuss some of the applications of the QSAR/QSPR modeling in four broad (and partially overlapping) areas.

- (1) Prediction of pharmacokinetic properties and toxicity of drugs, drug-like compounds and organic chemicals in general.
- (2) Prediction of potential drug targets and activity spectra.
- (3) Prediction of activity towards individual targets.
- (4) Prediction of relevant physico-chemical properties.



In the final subsection, several open web-based QSAR/QSPR services are presented.

The quality of the *source structure–activity data* (including their accuracy, consistency, and representativeness) is crucial for any QSAR analysis.<sup>118</sup> Over the past decade, an enormous wealth of raw data was generated and made freely available to the scientific community.<sup>161,162</sup> The public domain ‘Big Data’ databases created and populated through the concerted efforts of the government and public institutions, industry and academia, including ChEMBL,<sup>78</sup> BindingDB,<sup>163</sup> PubChem,<sup>164,165</sup> and many others currently provide hundreds of millions of activity data points for millions of compounds and thousands of targets. However, even the ‘good old’ compact QSAR datasets (which were usually obtained in the same laboratory or manually compiled from a few literature publications) often contained up to 10% of errors in compound structures, activity values, and associated metadata.<sup>118,166,167</sup> In the big data world the situation is aggravated by heterogeneous and inconsistent data, incomplete and/or inconsistent annotation (adequate specifications of the assay, species, and target are especially problematic), data duplication, *etc.*<sup>168–172</sup> Unfortunately, the data collection initiatives usually do not have sufficient resources for the complete curation of all available data. Thus, any QSAR study must start with careful data filtering and curation.<sup>167,168,170,171</sup> On the other hand, in spite of the great volume of the open databases, they may sometimes provide only limited coverage of specific practically relevant endpoints and/or regions of chemical space. In an attempt to extend the applicability domain of the public predictive models, approaches for incorporating proprietary data into the QSAR/QSPR analysis workflows were proposed.<sup>173</sup> Instead of the actual structure information, only a set of generic descriptors (from which the sensitive chemical structures cannot be easily reconstructed) is calculated and transferred in a controlled manner.

The *molecular descriptors* and *machine learning approaches* employed in the QSAR/QSPR analysis have been thoroughly discussed both from a general methodological perspective<sup>117,118</sup> and in the specific contexts of the ADMET properties<sup>174,175</sup> and toxicity<sup>176–180</sup> prediction. In fact, it may be difficult to name a method that has not been tried in QSAR/QSPR at some point or the other. Nevertheless, various physico-chemical, substructural (fragmental), and topological descriptors seem to be most prominent. Among the machine learning methods, the partial least squares regression, artificial neural networks, support vector machines and support vector regression, decision trees and random forests, KNN, and Bayesian classifiers are frequently used in recent studies, although this list is far from being exhaustive.

Another approach, the *Matched Molecular Pair (MMP) analysis*,<sup>181–186</sup> should also be mentioned. In a sense, it can be viewed as a combination of the QSAR and molecular similarity methods (see also Section 6.2.2). Extending and generalizing the ‘naïve’ approach intuitively used by chemists to detect trends and patterns in a structure–activity relationships table, the MMP analysis identifies the matched molecular pairs in a data set, *i.e.* pairs of similar compounds that differ only in one substructural fragment (terminal substituent or central



core). If such structural transformation is consistently associated with a specific change in activity, this can be used to predict the activity in other pairs with the same transformation as well as to guide the optimization of activity, toxicity or other properties in a series of compounds. The *Matching Molecular Series* (MMS) concept further extends this approach from matched pairs to series of transformations associated with consistent trends in activity.<sup>187</sup>

### 6.3.2.1 Prediction of ADMET Properties

This section is focused on the pharmacokinetic properties such as absorption, distribution, metabolism and excretion, as well as on the toxicity (including adverse reactions) and potential drug–drug (or compound–drug) interactions of drugs, drug-like compounds and organic chemicals in general. This multitude of properties independent of the intended drug target are often collectively called *ADMET* (*absorption, distribution, metabolism, excretion and toxicity*) *properties*. Of course, many of them are in fact mediated by targets (or anti-targets) of their own, so the distinction between target activity and ADMET properties is not clear-cut. These properties have a profound influence on the activity, pharmacological profile, and mode of use of a drug or on the potential risks associated with other xenobiotics. Thus, the ADMET prediction and optimization is an important and broad topic attracting a significant interest. As any attempt to provide its complete coverage is beyond the scope and volume of this chapter (even the relevant endpoints are numbered in dozens, if not hundreds), only some notable examples of the current state-of-the-art and recent publications are presented. Wang and Urban<sup>188</sup> provide a more detailed account of this field, as do the texts and reviews mentioned above.<sup>7–12</sup>

Among the *absorption* properties, most of the research was focused on the human intestinal absorption (HIA) and oral bioavailability modeling.<sup>189,190</sup> A recently published HIA model based on the fragmental descriptors and artificial neural networks has quite high predictivity and a broad applicability domain.<sup>191</sup> For the *distribution* stage, the studies concerning plasma protein binding<sup>192</sup> and the blood–brain barrier permeability<sup>193</sup> of drug compounds should be mentioned. A recent model for the blood–brain barrier permeability prediction<sup>160</sup> is derived from the most complete and accurate data set based on the open quantitative data. Using the fragmental descriptors and artificial neural networks, it provides high prediction accuracy (comparable to the experimental precision) and a broad applicability domain. At the *metabolism* stage, research was focused primarily on the binding and metabolism site predictions for various isoforms of the cytochrome P450.<sup>194</sup> Although the interest in *excretion* modeling seems to be smaller, some results on renal clearance prediction have been published.<sup>195</sup>

The *toxicity* of chemicals to humans, animals, other organisms, and the environment is a broad phenomenon with many facets and mechanisms, from the non-specific narcosis by volatile organic compounds and the mutagenicity of reactive chemicals to the cholinesterase inhibition by organophosphorus compounds and the metabolic disruptions caused by the nuclear

receptor ligands. Toxicity prediction models not only help to optimize the structures to avoid or minimize the risks but also support non-animal hazard assessment for regulatory purposes.<sup>117,176,177</sup> QSAR models are now accepted in the European Union under Registration, Evaluation and Authorisation of Chemicals (REACH) regulations,<sup>196</sup> in the United States under the Toxic Substances Control Act (TSCA),<sup>197</sup> by the Organisation for Economic Co-operation and Development,<sup>198</sup> and other bodies. Naturally, certain requirements must be met before the QSAR predictions indicating the presence or absence of a certain dangerous property may be used instead of testing, as follows.<sup>199</sup>

- (1) Results are derived from a (Q)SAR model whose scientific *validity* has been established. In particular, it should have (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness, and predictivity; and (v) a mechanistic interpretation, if possible.
- (2) The substance falls within the *applicability domain* of the (Q)SAR model.
- (3) Results are *adequate* for the purpose of classification and labelling and/or risk assessment.
- (4) Adequate and reliable *documentation* of the applied method is provided.

(In fact, these principles are quite reasonable and applicable not only to the regulatory risk assessment, but to any use or publication of the QSAR predictions.)

From the very first years of QSAR modeling, the prediction of *acute toxicity* has been a significant topic of research, and many models have been published.<sup>200</sup> Among recent results, the models of acute rat toxicity<sup>201</sup> and aquatic toxicity<sup>202,203</sup> can be mentioned. Studies aimed at predicting *mutagenicity* and *carcinogenicity*<sup>204,205</sup> also have a long tradition,<sup>125</sup> but continue to attract interest.<sup>159,173,206</sup> Among the more target-specific types of toxicity, *cardiac toxicity* mediated by the hERG potassium channel (QT interval prolongation leading to ventricular tachyarrhythmia) is a very important and sometimes lethal adverse drug reaction. The prediction of the hERG inhibition plays a significant role in ensuring drug safety and has been a topic of many studies.<sup>207</sup> A recently published model based on the fragmental descriptors and artificial neural networks has quite high predictivity and a broad applicability domain.<sup>208</sup> Significant progress was made in the prediction of *drug-induced liver injury* and hepatotoxicity risks.<sup>209</sup> The prediction of potential *reproductive toxicants* and *endocrine disruptors* is also possible.<sup>210,211</sup>

The *Toxicology in the 21st Century (Tox21)* program<sup>212</sup> aims to develop better toxicity assessment methods to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to negative health effects. Using the quantitative high-throughput screening techniques, a collection of ~10 000 environmental chemicals and approved drugs (called the Tox21 10K library) is screened in a large panel of assays for their potential to disrupt biological pathways that may result in toxicity. Currently, the panel includes the cell viability

and apoptosis, membrane integrity, mitochondrial toxicity, DNA damage, cytokine (proinflammatory), hERG channel, nuclear receptor, and stress response pathway assays. These *in vitro* results have been shown to be predictive of the *in vivo* toxicity outcomes.<sup>212</sup> During the Tox21Challenge, the predictive models for a subset of the nuclear receptor and stress response targets were successfully derived by a number of teams using a variety of approaches ranging from molecular similarity analysis to the deep learning artificial neural networks.<sup>213–222</sup>

*Drug–drug (or compound–drug) interactions*<sup>223</sup> mediated by drug metabolizing enzymes such as cytochrome P450s as well as drug transporters, ion channels, and other proteins may lead to unexpected changes in the pharmacokinetic behavior of a drug and cause severe adverse effects that may result in drug withdrawal or failure to reach the market. On the other hand, they may be used to control drug metabolism (*e.g.* ritonavir is commonly used to boost the activity of the HIV protease inhibitors by preventing their cytochrome P450-mediated clearance). Thus, the prediction of potential drug–drug interactions is an important goal which has been a topic of many studies<sup>223–227</sup> (see also Section 6.3.2.2).

### 6.3.2.2 Prediction of Potential Drug Targets

In this section we focus on the approaches to the prediction of *potential target profiles and activity spectra* of drugs and chemicals. They can be used to discover new potential pharmacological indications for registered drugs (so-called *drug repositioning* or *drug repurposing*)<sup>228</sup> and organic compounds developed by synthetic chemists, as well as to obtain a preliminary estimate of adverse reaction risks and identify potential off-targets responsible for such reactions.<sup>223,229</sup> In addition to the similarity-based approaches discussed in Section 6.2.3.2, a combination of chemical features and protein interaction networks<sup>230</sup> can be used. Another approach called the *Prediction of Activity Spectra for Substances* (PASS)<sup>231–233</sup> is based on the chemical similarity assessment using substructural<sup>234</sup> and electronic-substructural<sup>235</sup> descriptors. This method was successfully used to predict potential targets for natural compounds<sup>236</sup> and new types of activity for synthetic compounds. In many cases the predictions were confirmed in the experiment.<sup>237,238</sup>

### 6.3.2.3 Prediction of Activity Towards Individual Targets

Although approaches implying a consistent binding mode (see Section 6.3.1) are usually preferable for modeling activity towards a specific target, the methods discussed in this section can also be successfully employed, especially for large and/or diverse data sets comprising compounds with different binding modes or mechanisms of action. The models for some of the targets were discussed earlier, such as the human serum albumin playing a primary role in plasma protein binding, cytochrome P450 enzymes, hERG channel, nuclear receptors, *etc.* In addition, research on the P-glycoprotein

transporter modulating intestinal absorption, blood–brain barrier permeation, tumor multidrug resistance, and other processes,<sup>225,226</sup> as well as on different subtypes of chemokine receptors<sup>239</sup> can be mentioned.

#### 6.3.2.4 Prediction of Physico-Chemical Properties

The QSPR modeling of the physico-chemical properties of organic compounds has a long tradition starting with the pioneering work by C. Hansch and R. Rekker on lipophilicity prediction. Among the recent publications in this field, research into the lipophilicity,<sup>240,241</sup> solubility,<sup>242</sup> melting point,<sup>243,244</sup> and biodegradability<sup>245</sup> can be mentioned.

#### 6.3.2.5 Open Web-Based QSAR/QSPR Services

In this section we present a number of open publicly available web-based services for QSAR/QSPR prediction and modeling.

The *Way2Drug* portal<sup>246</sup> developed at the Institute of Biomedical Chemistry of the Russian Academy of Sciences offers a number of services for the prediction of activity spectra of substances (PASS),<sup>235</sup> acute toxicity,<sup>201</sup> anti-target activity,<sup>233</sup> ecotoxicity, cytochrome P450 and UDP-glucuronosyltransferase substrate/metabolite specificity, site of metabolism (SOMP),<sup>247</sup> etc.

The *Online Chemical Modeling Environment* (OCHEM)<sup>248,249</sup> developed at the German Research Center for Environmental Health (Helmholtz Zentrum München) is a web platform for data storage, model development and publishing of chemical information. It offers a large selection of curated QSAR/QSPR data sets and validated predictive models mostly focused on the ADMET endpoints. In addition, users can upload, analyze and model their own data using a variety of molecular descriptors and machine learning techniques, predict the activities and other relevant properties, screen the compounds against a set of structural alerts (toxicophores),<sup>250</sup> as well as publish the data sets and models for the benefit of the community (if desired).

An *integrated online service for the ADMET properties prediction*,<sup>251</sup> developed at the Laboratory of Medicinal Chemistry of the Department of Chemistry, Lomonosov Moscow State University uses a panel of validated predictive models based on the fragmental descriptors and artificial neural networks to estimate a number of important ADMET endpoints such as blood–brain barrier permeability (LogBB),<sup>160</sup> human intestinal absorption (HIA),<sup>191</sup> hERG-mediated cardiac toxicity,<sup>208</sup> etc. Convenient tools for the visualization and analysis of the results are available.

## 6.4 Conclusion

The similarity property principle, which implies that similar structures (should) possess similar properties, lays the basis for the detection, analysis, and interpretation of patterns in the known data on the properties (including biological activities) of compounds, as well as for using these patterns to

predict the properties of novel structures or to design structures with desired properties. A lot of approaches to molecular similarity analysis have been proposed, reflecting the rich and multifaceted nature of the molecular structures. A similarity measure should be selected or constructed in such a way as to properly capture the features of a structure that are important for a specific problem. Nevertheless, an activity landscape is usually not uniform, and the properties for some (seemingly) similar structures may differ substantially. Such activity cliffs somewhat complicate structure–activity relationship modeling, but can offer an opportunity for better understanding of the mechanisms of action and for the design of more potent compounds.

The Quantitative Structure–Activity/Property Relationships (QSAR/QSPR) analysis employs machine learning and other computational techniques to derive predictive models from the experimental activity or property data for chemical compounds using an intermediate representation of their structures by a suitable set of numerical molecular descriptors. A variety of QSAR/QSPR approaches can help to predict the biological activities, pharmacokinetic properties, and toxicities, as well as the relevant physico-chemical properties of drugs, drug-like compounds, and organic chemicals in general. Some of the predictive models are available in convenient open web-based services. QSAR/QSPR methods and other ligand-based techniques complement the structure-based drug design methods that directly operate with the biotarget structure. In real research projects in chemistry, pharmacology, and toxicology, all the applicable tools can—and should—be used together, reflecting the interdisciplinary power of chemoinformatics.

## Acknowledgements

This work was partially supported by the Russian Foundation for Basic Research (projects 14-03-01063 and 15-03-09084). The authors are grateful to the ChemAxon and OpenEye companies for kindly providing the academic licenses for the structural data processing and molecular modeling software used in preparing the illustrations.

## References

1. M. I. Skvortsova, I. V. Stankevich, V. A. Palyulin and N. S. Zefirov, Molecular similarity concept and its use for predicting the properties of chemical compounds, *Russ. Chem. Rev.*, 2006, 75(11), 961–979, DOI: 10.1070/RC2006v075n11ABEH003616.
2. *Concepts and Applications of Molecular Similarity*, ed. M. A. Johnson and G. M. Maggiora, Wiley, 1990, ISBN: 978-0-471-62175-1.
3. D. F. Larder and F. F. Kluge, Alexander Mikhailovich Butlerov's theory of chemical structure, *J. Chem. Educ.*, 1971, 48(5), 287–289, DOI: 10.1021/ed048p287.
4. D. E. Lewis, 150 Years of organic structures, in *Atoms in Chemistry: From Dalton's Predecessors to Complex Atoms and Beyond*, ed. C. J. Giunta, ACS, 2010, pp. 35–57, DOI: 10.1021/bk-2010-1044.ch004.

5. J. L. Medina-Franco and G. M. Maggiora, Molecular Similarity Analysis, in *Chemoinformatics for Drug Discovery*, ed. J. Bajorath, Wiley, 2014, pp. 343–399, DOI: 10.1002/9781118742785.ch15.
6. G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular similarity in medicinal chemistry, *J. Med. Chem.*, 2014, 57(8), 3186–3204, DOI: 10.1021/jm401411z.
7. *Recent Advances in QSAR Studies: Methods and Applications*, ed. T. Puzyn, J. Leszczynski and M. T. Cronin, Springer, 2010, DOI: 10.1007/978-1-4020-9783-6.
8. *Computational Toxicology*, ed. B. Reisfeld and A. Mayeno, Springer, 2013, vol. I, DOI: 10.1007/978-1-62703-050-2.
9. *Computational Toxicology*, ed. B. Reisfeld and A. Mayeno, Springer, 2013, vol. II, DOI: 10.1007/978-1-62703-059-5.
10. S. Ekins, S. Kortagere, M. D. Krasowski, A. J. Williams, J. J. Xu and M. Zientek, Ligand-based modeling of toxicity, in *Drug Design Strategies: Quantitative Approaches*, ed. D. J. Livingstone and A. M. Davis, RSC, 2012, pp. 312–344, DOI: 10.1039/9781849733410-00312.
11. A. M. Davis and R. J. Riley, ADME(T) predictions in drug discovery, in *Drug Design Strategies: Quantitative Approaches*, ed. D. J. Livingstone and A. M. Davis, RSC, 2012, pp. 345–366, DOI: 10.1039/9781849733410-00345.
12. T. Steger-Hartmann, *In silico* toxicology – current approaches and future perspectives to predict toxic effects with computational tools, in *Predictive Toxicology: From Vision to Reality*, ed. F. Pfannkuch and L. Suter-Dick, Wiley, 2015, pp. 11–32, DOI: 10.1002/9783527674183.ch02.
13. Y. Hu, D. Stumpfe and J. Bajorath, Advancing the activity cliff concept, *F1000Research*, 2013, 2, 199, DOI: 10.12688/f1000research.2-199.v1.
14. A. Dalke, Molecular fingerprints background, *Dalke Scientific Software*, LLC, 2008, [http://www.dalkescientific.com/writings/diary/archive/2008/06/26/fingerprint\\_background.html](http://www.dalkescientific.com/writings/diary/archive/2008/06/26/fingerprint_background.html), accessed 01.06.2016.
15. Fingerprints – screening and similarity, *Daylight theory manual*, Daylight 4.9, Daylight Chemical Information Systems, Inc., 2011, <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html>, accessed 01.06.2016.
16. *MACCS Structural Keys*, ISIS/Base 2.4, Accelrys, San Diego, CA, 2002.
17. *PubChem Substructure Fingerprint V1.3*, [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.txt](ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt), accessed 01.06.2016.
18. D. Rogers and M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.*, 2010, 50(5), 742–754, DOI: 10.1021/ci100050t.
19. *Chemical Hashed Fingerprint*, ChemAxon Ltd, <https://docs.chemaxon.com/display/docs/Chemical+Hashed+Fingerprint>, accessed 01.06.2016.
20. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, Reoptimization of MDL keys for use in drug discovery, *J. Chem. Inf. Comput. Sci.*, 2002, 42(6), 1273–1280, DOI: 10.1021/ci010132r.
21. K. Heikamp and J. Bajorat, Fingerprint design and engineering strategies: rationalizing and improving similarity search performance, *Future Med. Chem.*, 2012, 4(15), 1945–1959, DOI: 10.4155/fmc.12.126.



22. S. Jasial, Y. Hu, M. Vogt and J. Bajorath, Activity-relevant similarity values for fingerprints and implications for similarity searching, *F1000-Research*, 2016, **5**, 591, DOI: 10.12688/f1000research.8357.2.
23. D. Livingstone, *A Practical Guide to Scientific Data Analysis*, Wiley, 2009, DOI: 10.1002/9780470017913.
24. D. I. Osolodkin, E. V. Radchenko, A. A. Orlov, A. E. Voronkov, V. A. Palyulin and N. S. Zefirov, Progress in visual representations of chemical space, *Expert Opin. Drug Discovery*, 2015, **10**(9), 959–973, DOI: 10.1517/17460441.2015.1060216.
25. R. D. Cramer, D. E. Patterson and J. D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, 1988, **110**(18), 5959–5967, DOI: 10.1021/ja00226a005.
26. G. Klebe, U. Abraham and T. Mietzner, Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity, *J. Med. Chem.*, 1994, **37**(24), 4130–4146, DOI: 10.1021/jm00050a010.
27. J. A. Grant, M. A. Gallardo and B. T. Pickup, A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape, *J. Comput. Chem.*, 1996, **17**(14), 1653–1666, DOI: 10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K.
28. P. C. D. Hawkins, A. G. Skillman and A. Nicholls, Comparison of shape-matching and docking as virtual screening tools, *J. Med. Chem.*, 2007, **50**(1), 74–82, DOI: 10.1021/jm0603365.
29. *ROCS Theory*, OpenEye Scientific Software, Santa Fe, NM, [https://docs.eyesopen.com/rocs/shape\\_theory.html](https://docs.eyesopen.com/rocs/shape_theory.html), accessed 01.06.2016.
30. V. A. Palyulin, E. V. Radchenko and N. S. Zefirov, Molecular Field Topology Analysis method in QSAR studies of organic compounds, *J. Chem. Inf. Comput. Sci.*, 2000, **40**(3), 659–667, DOI: 10.1021/ci980114i.
31. E. V. Radchenko, V. A. Palyulin and N. S. Zefirov, Molecular Field Topology Analysis in drug design and virtual screening, in *Cheminformatics Approaches to Virtual Screening*, ed. A. Varnek and A. Tropsha, RSC, 2008, pp. 150–181, DOI: 10.1039/9781847558879-00150.
32. P. Willett, J. M. Barnard and G. M. Downs, Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, 1998, **38**(6), 983–996, DOI: 10.1021/ci9800211.
33. S. J. Swamidass and P. Baldi, Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time, *J. Chem. Inf. Model.*, 2007, **47**(2), 302–317, DOI: 10.1021/ci600358f.
34. J. W. Raymond, E. J. Gardiner and P. Willett, Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm, *J. Chem. Inf. Comput. Sci.*, 2002, **42**(2), 305–316, DOI: 10.1021/ci010381f.
35. Y. Cao, T. Jiang and T. Girke, A maximum common substructure-based algorithm for searching and predicting drug-like compounds, *Bioinformatics*, 2008, **24**(13), i366–i374, DOI: 10.1093/bioinformatics/btn186.

36. J. J. McGregor and P. Willett, Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions, *J. Chem. Inf. Comput. Sci.*, 1981, **21**(3), 137–140, DOI: 10.1021/ci00031a005.
37. G. M. Maggiora, J. D. Petke and J. Mestres, A general analysis of field-based molecular similarity indices, *J. Math. Chem.*, 2002, **31**(3), 251–270, DOI: 10.1023/A:1020784004649.
38. Y. Wang and J. Bajorath, Balancing the influence of molecular complexity on fingerprint similarity searching, *J. Chem. Inf. Model.*, 2008, **48**(1), 75–84, DOI: 10.1021/ci700314x.
39. J. Mestres and G. M. Maggiora, Putting molecular similarity into context: asymmetric indices for field-based similarity measures, *J. Math. Chem.*, 2006, **39**(1), 107–118, DOI: 10.1007/s10910-005-9007-3.
40. A. Tversky, Features of similarity, *Psychol. Rev.*, 1977, **84**(4), 327–352, DOI: 10.1037/0033-295X.84.4.327.
41. X. Chen and F. K. Brown, Asymmetry of chemical similarity, *ChemMed-Chem*, 2007, **2**(2), 180–182, DOI: 10.1002/cmdc.200600161.
42. A. Mauri, D. Ballabio, R. Todeschini and V. Consonni, Mixtures, metabolites, ionic liquids: A new measure to evaluate similarity between complex chemical systems, *21st European Symposium on Quantitative Structure-Activity Relationship (EuroQSAR 2016)*, Verona, Italy, September 4–8, 2016, p. OC20.
43. *Instant JChem 16.5.23*, ChemAxon Ltd, 2016, <http://www.chemaxon.com>.
44. *MFTAWin 3.5*, Department of Chemistry, Lomonosov Moscow State University, 2015, <http://qsar.chem.msu.ru>.
45. *PubChem Compound Database*, National Center for Biotechnology Information, CID 6175, <https://pubchem.ncbi.nlm.nih.gov/compound/6175>, accessed 01.06.2016.
46. *PubChem Compound Database*, National Center for Biotechnology Information, CID 73339, <https://pubchem.ncbi.nlm.nih.gov/compound/73339>, accessed 01.06.2016.
47. R. Guha, Exploring structure–activity data using the landscape paradigm, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**(6), 829–841, DOI: 10.1002/wcms.1087.
48. G. M. Maggiora, On outliers and activity cliffs – Why QSAR often disappoints, *J. Chem. Inf. Model.*, 2006, **46**(4), 1535, DOI: 10.1021/ci060117s.
49. M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. Cordeiro and F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, 2014, **19**(8), 1069–1080, DOI: 10.1016/j.drudis.2014.02.003.
50. J. L. Medina-Franco, G. Navarrete-Vázquez and O. Méndez-Lucio, Activity and property landscape modeling is at the interface of chemoinformatics and medicinal chemistry, *Future Med. Chem.*, 2015, **7**(9), 1197–1211, DOI: 10.4155/fmc.15.51.
51. D. Stumpfe and J. Bajorath, Exploring activity cliffs in medicinal chemistry, *J. Med. Chem.*, 2012, **55**(7), 2932–2942, DOI: 10.1021/jm201706b.



52. D. Dimova and J. Bajorath, Advances in activity cliff research, *Mol. Inf.*, 2016, **35**(5), 181–191, DOI: 10.1002/minf.201600023.
53. H. Schönherr and T. Cernak, Profound methyl effects in drug discovery and a call for new C–H methylation reactions, *Angew. Chem., Int. Ed.*, 2013, **52**(47), 12256–12267, DOI: 10.1002/anie.201303207.
54. P. J. Coleman, J. D. Schreier, C. D. Cox, M. J. Breslin, D. B. Whitman, M. J. Bogusky, G. B. McGaughey, R. A. Bednar, W. Lemaire, S. M. Doran, S. V. Fox, S. L. Garson, A. L. Gotter, C. M. Harrell, D. R. Reiss, T. D. Cabalu, D. Cui, T. Prueksaritanont, J. Stevens, P. L. Tannenbaum, R. G. Ball, J. Stellabott, S. D. Young, G. D. Hartman, C. J. Winrow and J. J. Renger, Discovery of [(2R,5R)-5-[(5-fluoropyridin-2-yl)oxy]methyl]-2-methylpiperidin-1-yl][5-methyl-2-(pyrimidin-2-yl)phenyl]methanone (MK-6096): A dual orexin receptor antagonist with potent sleep-promoting properties, *ChemMedChem*, 2012, **7**(3), 415–424, DOI: 10.1002/cmdc.201200025.
55. K. W. Kuntz, J. E. Campbell, H. Keilhack, R. M. Pollock, S. K. Knutson, M. Porter-Scott, V. M. Richon, C. J. Sneeringer, T. J. Wigle, C. J. Allain, C. R. Majer, M. P. Moyer, R. A. Copeland and R. Chesworth, The importance of being Me: Magic methyls, methyltransferase inhibitors, and the discovery of tazemetostat, *J. Med. Chem.*, 2016, **59**(4), 1556–1564, DOI: 10.1021/acs.jmedchem.5b01501.
56. M. R. Wood, C. R. Hopkins, J. T. Brogan, P. J. Conn and C. W. Lindsley, “Molecular switches” on mGluR allosteric ligands that modulate modes of pharmacology, *Biochemistry*, 2011, **50**(13), 2403–2410, DOI: 10.1021/bi200129s.
57. S. Purser, P. R. Moore, S. Swallow and V. Gouverneur, Fluorine in medicinal chemistry, *Chem. Soc. Rev.*, 2008, **37**(2), 320–330, DOI: 10.1039/b610213c.
58. E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly and N. A. Meanwell, Applications of fluorine in medicinal chemistry, *J. Med. Chem.*, 2015, **58**(21), 8315–8359, DOI: 10.1021/acs.jmedchem.5b00258.
59. S.-F. Lu, B. Herbert, G. Haufe, K. W. Laue, W. L. Padgett, O. Oshunlet, J. W. Daly and K. L. Kirk, Syntheses of (R)- and (S)-2- and 6-fluoronorepinephrine and (R)- and (S)-2- and 6-fluoroepinephrine: Effect of stereochemistry on fluorine-induced adrenergic selectivities, *J. Med. Chem.*, 2000, **43**(8), 1611–1619, DOI: 10.1021/jm990599h.
60. M. J. Shaughnessy, A. Harsanyi, J. Li, T. Bright, C. D. Murphy and G. Sandford, Targeted fluorination of a nonsteroidal anti-inflammatory drug to prolong metabolic half-life, *ChemMedChem*, 2014, **9**(4), 733–736, DOI: 10.1002/cmdc.201300490.
61. D. Stumpfe, Y. Hu, D. Dimova and J. Bajorath, Recent progress in understanding activity cliffs and their utility in medicinal chemistry, *J. Med. Chem.*, 2014, **57**(1), 18–28, DOI: 10.1021/jm401120g.
62. D. Dimova, D. Stumpfe, Y. Hu and J. Bajorath, Activity cliff clusters as a source of structure-activity relationship information, *Expert Opin. Drug Discovery*, 2015, **10**(5), 441–447, DOI: 10.1517/17460441.2015.1019861.

63. R. Guha and J. H. Van Drie, Structure–Activity Landscape Index: Identifying and quantifying activity cliffs, *J. Chem. Inf. Model.*, 2008, **48**(3), 646–658, DOI: 10.1021/ci7004093.
64. R. Guha, The ups and downs of structure–activity landscapes, in *Chemoinformatics and Computational Chemical Biology*, ed. J. Bajorath, Springer, 2011, pp. 101–117, DOI: 10.1007/978-1-60761-839-3\_3.
65. R. Guha and J. H. Van Drie, Assessing how well a modeling protocol captures a structure–activity landscape, *J. Chem. Inf. Model.*, 2008, **48**(8), 1716–1728, DOI: 10.1021/ci8001414.
66. R. Guha and J. L. Medina-Franco, On the validity versus utility of activity landscapes: are all activity cliffs statistically significant? *J. Cheminf.*, 2014, **6**(1), 11, DOI: 10.1186/1758-2946-6-11.
67. L. Peltason and J. Bajorath, SAR Index: Quantifying the nature of structure–activity relationships, *J. Med. Chem.*, 2007, **50**(23), 5571–5578, DOI: 10.1021/jm0705713.
68. X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, MMP-Cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs, *J. Chem. Inf. Model.*, 2012, **52**(5), 1138–1145, DOI: 10.1021/ci3001138.
69. J. L. Medina-Franco, Activity cliffs: facts or artifacts?, *Chem. Biol. Drug Des.*, 2013, **81**(5), 553–556, DOI: 10.1111/cbdd.12115.
70. K. Heikamp, X. Hu, A. Yan and J. Bajorath, Prediction of activity cliffs using support vector machines, *J. Chem. Inf. Model.*, 2012, **52**(9), 2354–2365, DOI: 10.1021/ci300306a.
71. C. Tebby and E. Mombelli, Modelling structure activity landscapes with cliffs: A kernel regression-based approach, *Mol. Inf.*, 2013, **32**(7), 609–623, DOI: 10.1002/minf.201300016.
72. J. Husby, G. Bottegoni, I. Kufareva, R. Abagyan and A. Cavalli, Structure-based predictions of activity cliffs, *J. Chem. Inf. Model.*, 2015, **55**(5), 1062–1076, DOI: 10.1021/ci500742b.
73. K. Klimenko, G. Marcou, D. Horvath and A. Varnek, Chemical space mapping and structure–activity analysis of the ChEMBL antiviral compound set, *J. Chem. Inf. Model.*, 2016, **56**(8), 1438–1454, DOI: 10.1021/acs.jcim.6b00192.
74. Y. Hu, G. M. Maggiora and J. Bajorath, Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account, *J. Comput.-Aided Mol. Des.*, 2013, **27**(2), 115–124, DOI: 10.1007/s10822-012-9632-4.
75. D. Stumpfe and J. Bajorath, Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds, *J. Chem. Inf. Model.*, 2012, **52**(9), 2348–2353, DOI: 10.1021/ci300288f.
76. Y. Hu and J. Bajorath, Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database, *J. Chem. Inf. Model.*, 2012, **52**(7), 1806–1811, DOI: 10.1021/ci300274c.

77. Y. Hu and J. Bajorath, Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs, *J. Chem. Inf. Model.*, 2012, **52**(3), 670–677, DOI: 10.1021/ci300033e.
78. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107, DOI: 10.1093/nar/gkr777.
79. D. Stumpfe, A. de la Vega de León, D. Dimova and J. Bajorath, Advancing the activity cliff concept, part II, *F1000Research*, 2014, **3**, 75, DOI: 10.12688/f1000research.3788.1.
80. D. Dimova, D. Stumpfe and J. Bajorath, Method for the evaluation of structure–activity relationship information associated with coordinated activity cliffs, *J. Med. Chem.*, 2014, **57**(15), 6553–6563, DOI: 10.1021/jm500577n.
81. D. Stumpfe, D. Dimova, K. Heikamp and J. Bajorath, Compound pathway model to capture SAR progression: Comparison of activity cliff-dependent and -independent pathways, *J. Chem. Inf. Model.*, 2013, **53**(5), 1067–1072, DOI: 10.1021/ci400141w.
82. D. Dimova, K. Heikamp, D. Stumpfe and J. Bajorath, Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets, *J. Med. Chem.*, 2013, **56**(8), 3339–3345, DOI: 10.1021/jm400147j.
83. *Virtual Screening: Principles, Challenges, and Practical Guidelines*, ed. C. Sotriffer, Wiley, 2011, ISBN: 978-3-527-32636-5.
84. D. Stumpfe and J. Bajorath, Similarity searching, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**(2), 260–282, DOI: 10.1002/wcms.23.
85. I. Muegge and P. Mukherjee, An overview of molecular fingerprint similarity search in virtual screening, *Expert Opin. Drug Discovery*, 2016, **11**(2), 137–148, DOI: 10.1517/17460441.2016.1117070.
86. D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, Neighborhood behavior: A useful concept for validation of “molecular diversity” descriptors, *J. Med. Chem.*, 1996, **39**(16), 3049–3059, DOI: 10.1021/jm960290n.
87. T. Girschick, L. Puchbauer and S. Kramer, Improving structural similarity based virtual screening using background knowledge, *J. Cheminf.*, 2013, **5**(1), 50, DOI: 10.1186/1758-2946-5-50.
88. R. Todeschini, D. Ballabio, M. Cassotti and V. Consonni, N3 and BNN: Two new similarity based classification methods in comparison with other classifiers, *J. Chem. Inf. Model.*, 2015, **55**(11), 2365–2374, DOI: 10.1021/acs.jcim.5b00326.
89. M. Luo, X. S. Wang and A. Tropsha, Comparative analysis of QSAR-based vs. chemical similarity based predictors of GPCRs binding affinity, *Mol. Inf.*, 2016, **35**(1), 36–41, DOI: 10.1002/minf.201500038.
90. M. J. Yu, Predicting total clearance in humans from chemical structure, *J. Chem. Inf. Model.*, 2010, **50**(7), 1284–1295, DOI: 10.1021/ci1000295.

91. M. Cassotti, D. Ballabio, R. Todeschini and V. Consonni, A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*), *SAR QSAR Environ. Res.*, 2015, **26**(3), 217–243, DOI: 10.1080/1062936X.2015.1018938.
92. T. W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D. J. Knight, C. Mahony, M. Schwarz, A. White and M. T. D. Cronin, A strategy for structuring and reporting a read-across prediction of toxicity, *Regul. Toxicol. Pharmacol.*, 2015, **72**(3), 586–601, DOI: 10.1016/j.yrtph.2015.05.016.
93. M. Cronin, J. Madden, S. Enoch and D. Roberts, *Chemical Toxicity Prediction: Category Formation and Read-Across*, RSC, 2013, DOI: 10.1039/9781849734400, ISBN: 978-1-84973-384-7.
94. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, Relating protein pharmacology by ligand chemistry, *Nat. Biotechnol.*, 2007, **25**(2), 197–206, DOI: 10.1038/nbt1284.
95. J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea and B. K. Shoichet, Quantifying the relationships among drug classes, *J. Chem. Inf. Model.*, 2008, **48**(4), 755–765, DOI: 10.1021/ci8000259.
96. Z. Wang, L. Liang, Z. Yin and J. Lin, Improving chemical similarity ensemble approach in target prediction, *J. Cheminf.*, 2016, **8**, 20, DOI: 10.1186/s13321-016-0130-x.
97. F. Schmidt, H. Matter, G. Hessler and A. Czich, Predictive in silico off-target profiling in drug discovery, *Future Med. Chem.*, 2014, **6**(3), 295–317, DOI: 10.4155/fmc.13.202.
98. M. Schenone, V. Dančák, B. K. Wagner and P. A. Clemons, Target identification and mechanism of action in chemical biology and drug discovery, *Nat. Chem. Biol.*, 2013, **9**(4), 232–240, DOI: 10.1038/nchembio.1199.
99. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijter, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, Predicting new molecular targets for known drugs, *Nature*, 2009, **462**(7270), 175–181, DOI: 10.1038/nature08506.
100. A. J. DeGraw, M. J. Keiser, J. D. Ochocki, B. K. Shoichet and M. D. Distefano, Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs, *J. Med. Chem.*, 2010, **53**(6), 2464–2471, DOI: 10.1021/jm901613f.
101. C. Laggner, D. Kokel, V. Setola, A. Tolia, H. Lin, J. J. Irwin, M. J. Keiser, C. Y. Cheung, D. L. Minor, B. L. Roth, R. T. Peterson and B. K. Shoichet, Chemical informatics and target identification in a zebrafish phenotypic screen, *Nat. Chem. Biol.*, 2011, **8**(2), 144–146, DOI: 10.1038/nchembio.732.
102. E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet and L. Urban, Large-scale prediction and testing of drug activity on side-effect targets, *Nature*, 2012, **486**(7403), 361–367, DOI: 10.1038/nature11159.

103. G. Mugumbate, K. A. Abrahams, J. A. G. Cox, G. Papadatos, G. van Westen, J. Lelièvre, S. T. Calus, N. J. Loman, L. Ballell, D. Barros, J. P. Overington and G. S. Besra, Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and *in vitro* validation, *PLoS One*, 2015, **10**(3), e0121492, DOI: 10.1371/journal.pone.0121492.
104. <http://sea.docking.org/>, accessed 01.06.2016.
105. Y.-C. Lo, S. Senese, C.-M. Li, Q. Hu, Y. Huang, R. Damoiseaux and J. Z. Torres, Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens, *PLoS Comput. Biol.*, 2015, **11**(3), e1004153, DOI: 10.1371/journal.pcbi.1004153.
106. <http://services.mbi.ucla.edu/CSNAP/>, accessed 01.06.2016.
107. N. Huang, B. K. Shoichet and J. J. Irwin, Benchmarking sets for molecular docking, *J. Med. Chem.*, 2006, **49**(23), 6789–6801, DOI: 10.1021/jm0608356.
108. M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, Directory of Useful Decoys, Enhanced (DUD-E): Better ligands and decoys for better benchmarking, *J. Med. Chem.*, 2012, **55**(14), 6582–6594, DOI: 10.1021/jm300687e.
109. N. Brown, *In Silico Medicinal Chemistry: Computational Methods to Support Drug Design*, RSC, 2016, DOI: 10.1039/9781782622604, ISBN: 978-1-78262-163-8.
110. N. J. Perualila-Tan, Z. Shkedy, W. Talloen, H. W. Göhlmann, M. V. Moerbeke and A. Kasim, Weighted similarity-based clustering of chemical structures and bioactivity data in early drug discovery, *J. Bioinf. Comput. Biol.*, 2016, **14**(4), 1650018, DOI: 10.1142/S0219720016500189.
111. D. G. Sprous, Fingerprint-based clustering applied to define a QSAR model use radius, *J. Mol. Graphics Modell.*, 2008, **27**(2), 225–232, DOI: 10.1016/j.jmglm.2008.04.009.
112. M. Zwierzyna, M. Vogt, G. M. Maggiora and J. Bajorath, Design and characterization of chemical space networks for different compound data sets, *J. Comput.-Aided Mol. Des.*, 2015, **29**(2), 113–125, DOI: 10.1007/s10822-014-9821-4.
113. B. Zhang, M. Vogt, G. M. Maggiora and J. Bajorath, Comparison of bioactive chemical space networks generated using substructure- and fingerprint-based measures of molecular similarity, *J. Comput.-Aided Mol. Des.*, 2015, **29**(7), 595–608, DOI: 10.1007/s10822-015-9852-5.
114. B. Zhang, M. Vogt, G. M. Maggiora and J. Bajorath, Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures, *J. Comput.-Aided Mol. Des.*, 2015, **29**(10), 937–950, DOI: 10.1007/s10822-015-9872-1.
115. M. Wu, M. Vogt, G. M. Maggiora and J. Bajorath, Design of chemical space networks on the basis of Tversky similarity, *J. Comput.-Aided Mol. Des.*, 2016, **30**(1), 1–12, DOI: 10.1007/s10822-015-9891-y.
116. K. Roy, S. Kar and R. N. Das, *A Primer on QSAR/QSPR Modeling*, Springer, 2015, DOI: 10.1007/978-3-319-17281-1, ISBN: 978-3-319-17280-4.

117. K. Roy, S. Kar and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Elsevier, 2015, ISBN: 978-0-12-801505-6.
118. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.*, 2014, **57**(12), 4977–5010, DOI: 10.1021/jm4004285.
119. T. Wang, M. B. Wu, J. P. Lin and L. R. Yang, Quantitative structure-activity relationship: promising advances in drug discovery platforms, *Expert Opin. Drug Discovery*, 2015, **10**(12), 1283–1300, DOI: 10.1517/17460441.2015.1083006.
120. E. V. Radchenko, D. S. Karlov, V. A. Palyulin, N. S. Zefirov and V. M. Pentkovski, Computer-aided modeling of activity and selectivity of quinazolinones as noncompetitive NMDA receptor antagonists, *Dokl. Biochem. Biophys.*, 2012, **443**(1), 118–122, DOI: 10.1134/S1607672912020159.
121. E. V. Radchenko, D. S. Karlov, A. N. Zefirov, V. A. Palyulin, N. S. Zefirov and V. M. Pentkovski, Computer-aided design of negative allosteric modulators of NMDA receptor, *Dokl. Biochem. Biophys.*, 2013, **448**(1), 22–26, DOI: 10.1134/S1607672913010079.
122. K. H. Kim and Y. C. Martin, Direct prediction of dissociation constants (pK<sub>a</sub>'s) of clonidine-like imidazolines, 2-substituted imidazoles and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach, *J. Med. Chem.*, 1991, **34**(7), 2056–2060, DOI: 10.1021/jm00111a020.
123. T. Kaserer, K. R. Beck, M. Akram, A. Odermatt and D. Schuster, Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases, *Molecules*, 2015, **20**(12), 22799–22832, DOI: 10.3390/molecules201219880.
124. O. F. Güner and J. P. Bowen, Setting the record straight: The origin of the pharmacophore concept, *J. Chem. Inf. Model.*, 2014, **54**(5), 1269–1283, DOI: 10.1021/ci5000533.
125. J. Kazius, R. McGuire and R. Bursi, Derivation and validation of toxicophores for mutagenicity prediction, *J. Med. Chem.*, 2005, **48**(1), 312–320, DOI: 10.1021/jm040835a.
126. C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998), *Pure Appl. Chem.*, 1998, **70**(5), 1129–1143, DOI: 10.1351/pac199870051129.
127. G. Wolber, T. Seidel, F. Bendix and T. Langer, Molecule-pharmacophore superpositioning and pattern matching in computational drug design, *Drug Discovery Today*, 2008, **13**(1–2), 23–29, DOI: 10.1016/j.drudis.2007.09.007.



128. S. Y. Yang, Pharmacophore modeling and applications in drug discovery: challenges and recent advances, *Drug Discovery Today*, 2010, **15**(11–12), 444–450, DOI: 10.1016/j.drudis.2010.03.013.
129. T. S. Rush, J. A. Grant, L. Mosyak and A. Nicholls, A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction, *J. Med. Chem.*, 2005, **48**(5), 1489–1495, DOI: 10.1021/jm040163o.
130. W. H. Shin, X. Zhu, M. G. Bures and D. Kihara, Three-dimensional compound comparison methods and their application in drug discovery, *Molecules*, 2015, **20**(7), 12841–12862, DOI: 10.3390/molecules200712841.
131. E. V. Radchenko, D. S. Karlov, V. A. Palyulin and N. S. Zefirov, Molecular modeling of the transmembrane domain of mGluR<sub>2</sub> metabotropic glutamate receptor and the binding site of its positive allosteric modulators, *Dokl. Biochem. Biophys.*, 2014, **454**(1), 13–16, DOI: 10.1134/S1607672914010050.
132. *vROCS 3.1.2*, *OpenEye Scientific Software*, 2011, <http://www.eyesopen.com>.
133. M. Wieder, U. Perricone, S. Boresch, T. Seidel and T. Langer, Evaluating the stability of pharmacophore features using molecular dynamics simulations, *Biochem. Biophys. Res. Commun.*, 2016, **470**(3), 685–689, DOI: 10.1016/j.bbrc.2016.01.081.
134. A. A. Mel'nikov, V. A. Palyulin, E. V. Radchenko and N. S. Zefirov, Generation of chemical structures on the basis of QSAR models of molecular field topology analysis, *Dokl. Chem.*, 2007, **415**(2), 196–199, DOI: 10.1134/S0012500807080058.
135. A. Sun, A. Prussia, W. Zhan, E. E. Murray, J. Doyle, L.-T. Cheng, J.-J. Yoon, E. V. Radchenko, V. A. Palyulin, R. W. Compans, D. C. Liotta, R. K. Plemper and J. P. Snyder, Nonpeptide inhibitors of measles virus entry, *J. Med. Chem.*, 2006, **49**(17), 5080–5092, DOI: 10.1021/jm0602559.
136. V. I. Chupakhin, S. V. Bobrov, E. V. Radchenko, V. A. Palyulin and N. S. Zefirov, Computer-aided design of selective ligands of the benzodiazepine-binding site of the GABA<sub>A</sub> receptor, *Dokl. Chem.*, 2008, **422**(1), 227–230, DOI: 10.1134/S0012500808090073.
137. E. V. Radchenko, G. F. Makhaeva, V. V. Malygin, V. B. Sokolov, V. A. Palyulin and N. S. Zefirov, Modeling of the relationships between the structure of O-phosphorylated oximes and their anticholinesterase activity and selectivity using Molecular Field Topology Analysis (MFTA), *Dokl. Biochem. Biophys.*, 2008, **418**(1), 47–51, DOI: 10.1134/S1607672908010122.
138. E. V. Radchenko, G. F. Makhaeva, V. B. Sokolov, V. A. Palyulin and N. S. Zefirov, Study of the structural determinants of acute and delayed neurotoxicity of O-phosphorylated oximes by Molecular Field Topology Analysis (MFTA), *Dokl. Biochem. Biophys.*, 2009, **429**(1), 309–314, DOI: 10.1134/S1607672909060064.
139. E. V. Radchenko, S. O. Koshelev, D. A. Tsareva, A. E. Voronkov, V. A. Palyulin and N. S. Zefirov, Computer-aided design of arylphthalazines as potential Smoothed receptor antagonists, *Dokl. Chem.*, 2012, **443**(2), 97–100, DOI: 10.1134/S0012500812040027.

140. G. F. Makhaeva, E. V. Radchenko, I. I. Baskin, V. A. Palyulin, R. J. Richardson and N. S. Zefirov, Combined QSAR studies of inhibitor properties of O-phosphorylated oximes toward serine esterases involved in neurotoxicity, drug metabolism and Alzheimer's disease, *SAR QSAR Environ. Res.*, 2012, **23**(7–8), 627–647, DOI: 10.1080/1062936X.2012.679690.
141. E. V. Radchenko, A. A. Mel'nikov, G. F. Makhaeva, V. A. Palyulin and N. S. Zefirov, Molecular design of O-phosphorylated oximes – selective inhibitors of butyrylcholinesterase, *Dokl. Biochem. Biophys.*, 2012, **443**(1), 91–95, DOI: 10.1134/S1607672912020093.
142. G. F. Makhaeva, E. V. Radchenko, V. A. Palyulin, E. V. Rudakova, A. Y. Aksinenko, V. B. Sokolov, N. S. Zefirov and R. J. Richardson, Organophosphorus compound esterase profiles as predictors of therapeutic and toxic effects, *Chem.-Biol. Interact.*, 2013, **203**(1), 231–237, DOI: 10.1016/j.cbi.2012.10.012.
143. E. V. Radchenko, G. F. Makhaeva, N. P. Boltneva, O. G. Serebryakova, I. V. Serkov, A. N. Proshin, V. A. Palyulin and N. S. Zefirov, Molecular design of N,N-disubstituted 2-aminothiazolines as selective inhibitors of carboxylesterase, *Russ. Chem. Bull.*, 2016, **65**(2), 570–575, DOI: 10.1007/s11172-016-1339-6.
144. A. S. Girgis, S. R. Tala, P. V. Oliferenko, A. A. Oliferenko and A. R. Katritzky, Computer-assisted rational design, synthesis, and bioassay of non-steroidal anti-inflammatory agents, *Eur. J. Med. Chem.*, 2012, **50**, 1–8, DOI: 10.1016/j.ejmech.2011.11.034.
145. P. V. Oliferenko, A. A. Oliferenko, G. I. Poda, D. I. Osolodkin, G. G. Pillai, U. R. Bernier, M. Tsikolia, N. M. Agramonte, G. G. Clark, K. J. Linthicum and A. R. Katritzky, Promising *Aedes aegypti* repellent chemotypes identified through integrated QSAR, virtual screening, synthesis, and bioassay, *PLoS One*, 2013, **8**(9), e64547, DOI: 10.1371/journal.pone.0064547.
146. P. V. Oliferenko, A. A. Oliferenko, A. S. Girgis, D. O. Saleh, A. M. Srouf, R. F. George, G. G. Pillai, C. S. Panda, C. D. Hall and A. R. Katritzky, Synthesis, bioassay, and Molecular Field Topology Analysis of diverse vasodilatory heterocycles, *J. Chem. Inf. Model.*, 2014, **54**(4), 1103–1116, DOI: 10.1021/ci400723m.
147. F. Jabeen, P. V. Oliferenko, A. A. Oliferenko, G. G. Pillai, F. L. Ansari, C. D. Hall and A. R. Katritzky, Dual inhibition of the  $\alpha$ -glucosidase and butyrylcholinesterase studied by Molecular Field Topology Analysis, *Eur. J. Med. Chem.*, 2014, **80**, 228–242, DOI: 10.1016/j.ejmech.2014.04.018.
148. J. Verma, V. M. Khedkar and E. C. Coutinho, 3D-QSAR in drug design – a review, *Curr. Top. Med. Chem.*, 2010, **10**(1), 95–115, DOI: 10.2174/156802610790232260.
149. *Sybyl 8.0, Sybyl-X 2.1*, Certara L.P., 2008–2013, <http://www.certara.com>.
150. R. D. Cramer, Topomer CoMFA: A design methodology for rapid lead optimization, *J. Med. Chem.*, 2003, **46**(3), 374–388, DOI: 10.1021/jm020194o.
151. R. D. Cramer, Rethinking 3D-QSAR, *J. Comput.-Aided Mol. Des.*, 2011, **25**(3), 197–201, DOI: 10.1007/s10822-010-9403-z.
152. R. D. Cramer and B. Wendt, Template CoMFA: The 3D-QSAR grail? *J. Chem. Inf. Model.*, 2014, **54**(2), 660–671, DOI: 10.1021/ci400696v.



153. B. Wendt and R. D. Cramer, Challenging the gold standard for 3D-QSAR: template CoMFA versus X-ray alignment, *J. Comput.-Aided Mol. Des.*, 2014, **28**(8), 803–824, DOI: 10.1007/s10822-014-9761-z.
154. R. D. Cramer, Template CoMFA applied to 116 biological targets, *J. Chem. Inf. Model.*, 2014, **54**(7), 2147–2156, DOI: 10.1021/ci500230a.
155. R. D. Cramer, Template CoMFA generates single 3D-QSAR models that, for twelve of twelve biological targets, predict all ChEMBL-tabulated affinities, *PLoS One*, 2015, **10**(6), e0129307, DOI: 10.1371/journal.pone.0129307.
156. P. Tosco, T. Balle and F. Shiri, Open3DALIGN: an open-source software aimed at unsupervised ligand alignment, *J. Comput.-Aided Mol. Des.*, 2011, **25**(8), 777–783, DOI: 10.1007/s10822-011-9462-9.
157. P. Tosco and T. Balle, Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields, *J. Mol. Model.*, 2011, **17**(1), 201–208, DOI: 10.1007/s00894-010-0684-x.
158. <http://open3dqsar.org/>, accessed 01.06.2016.
159. S. B. Sosnin, E. V. Radchenko, V. A. Palyulin and N. S. Zefirov, Generalized fragmental approach in QSAR/QSPR studies, *Dokl. Chem.*, 2015, **463**(1), 185–188, DOI: 10.1134/S0012500815070071.
160. A. S. Dyabina, E. V. Radchenko, V. A. Palyulin and N. S. Zefirov, Prediction of blood-brain barrier permeability of organic compounds, *Dokl. Biochem. Biophys.*, 2016, **470**(1), 371–374, DOI: 10.1134/S1607672916050173.
161. S. J. Lusher and T. Ritschel, Finding the right approach to big data-driven medicinal chemistry, *Future Med. Chem.*, 2015, **7**(10), 1213–1216, DOI: 10.4155/fmc.15.58.
162. I. V. Tetko, O. Engkvist, U. Koch, J. L. Reymond and H. Chen, BIGCHEM: Challenges and opportunities for Big Data analysis in chemistry, *Mol. Inf.*, 2016, **35**(11–12), 615–621, DOI: 10.1002/minf.201600073.
163. M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic Acids Res.*, 2016, **44**(D1), D1045–D1053, DOI: 10.1093/nar/gkv1072.
164. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, PubChem Substance and Compound databases, *Nucleic Acids Res.*, 2016, **44**(D1), D1202–D1213, DOI: 10.1093/nar/gkv951.
165. Y. Wang, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B. A. Shoemaker, A. Gindulyte and S. H. Bryant, PubChem BioAssay: 2014 update, *Nucleic Acids Res.*, 2014, **42**(D1), D1075–D1082, DOI: 10.1093/nar/gkt978.
166. T. I. Oprea, M. Olah, L. Ostopovici, R. Rad and M. Mracec, On the propagation of errors in the QSAR literature, in *EuroQSAR 2002 – Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, ed. M. Ford, D. Livingstone, J. Dearden and H. van de Waterbeemd, Blackwell, 2003, pp. 314–315.
167. D. Fourches, E. Muratov and A. Tropsha, Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research, *J. Chem. Inf. Model.*, 2010, **50**(7), 1189–1204, DOI: 10.1021/ci100176x.

168. A. J. Williams and S. Ekins, A quality alert and call for improved curation of public chemistry databases, *Drug Discovery Today*, 2011, **16**(17–18), 747–750, DOI: 10.1016/j.drudis.2011.07.007.
169. B. Zdrazil, M. Pinto, P. Vasanthanathan, A. J. Williams, L. Z. Balderud, O. Engkvist, C. Chichester, A. Hersey, J. P. Overington and G. F. Ecker, Annotating human P-glycoprotein bioassay data, *Mol. Inf.*, 2012, **31**(8), 599–609, DOI: 10.1002/minf.201200059.
170. D. A. Tsareva and G. F. Ecker, How far could we go with open data – A case study for TRPV1 antagonists, *Mol. Inf.*, 2013, **32**(5–6), 555–562, DOI: 10.1002/minf.201300019.
171. O. A. Tarasova, A. F. Urusova, D. A. Filimonov, M. C. Nicklaus, A. V. Zakharov and V. V. Poroikov, QSAR modeling using large-scale databases: Case study for HIV-1 reverse transcriptase inhibitors, *J. Chem. Inf. Model.*, 2015, **55**(7), 1388–1399, DOI: 10.1021/acs.jcim.5b00019.
172. A. A. Nikitina, A. A. Orlov, D. I. Osolodkin, V. A. Palyulin and N. S. Zefirov, Analysis and visualization of antiviral chemical space, *21st European Symposium on Quantitative Structure-Activity Relationship (EuroQSAR 2016)*, Verona, Italy, September 4–8, 2016, p. P180.
173. E. Ahlberg, A. Amberg, L. D. Beilke, D. Bower, K. P. Cross, L. Custer, K. A. Ford, J. Van Gompel, J. Harvey, M. Honma, R. Jolly, E. Joossens, R. A. Kemper, M. Kenyon, N. Kruhlak, L. Kuhnke, P. Leavitt, R. Naven, C. Neilan, D. P. Quigley, D. Shuey, H.-P. Spirkel, L. Stavitskaya, A. Teasdale, A. White, J. Wichard, C. Zwickl and G. J. Myatt, Extending (Q)SARs to incorporate proprietary knowledge for regulatory purposes: A case study using aromatic amine mutagenicity, *Regul. Toxicol. Pharmacol.*, 2016, **77**, 1–12, DOI: 10.1016/j.yrtph.2016.02.003.
174. L. Tao, P. Zhang, C. Qin, S. Y. Chen, C. Zhang, Z. Chen, F. Zhu, S. Y. Yang, Y. Q. Wei and Y. Z. Chen, Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools, *Adv. Drug Delivery Rev.*, 2015, **86**, 83–100, DOI: 10.1016/j.addr.2015.03.014.
175. V. G. Maltarollo, J. C. Gertrudes, P. R. Oliveira and K. M. Honorio, Applying machine learning techniques for ADME-Tox prediction: a review, *Expert Opin. Drug Metab. Toxicol.*, 2015, **11**(2), 259–271, DOI: 10.1517/17425255.2015.980814.
176. A. B. Raies and V. B. Bajic, *In silico* toxicology: computational methods for the prediction of chemical toxicity, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2016, **6**(2), 147–172, DOI: 10.1002/wcms.1240.
177. G. F. Mangiardi, D. Alberga, C. D. Altomare, A. Carotti, M. Catto, S. Cellamare, D. Gadaleta, G. Lattanzi, F. Leonetti, L. Pisani, A. Stefanachi, D. Trisciuzzi and O. Nicolotti, Mind the gap! A journey towards computational toxicology, *Mol. Inf.*, 2016, **35**(8–9), 294–308, DOI: 10.1002/minf.201501017.
178. F. Sanz, P. Carrió, O. López, L. Capoferri, D. P. Kooi, N. P. Vermeulen, D. P. Geerke, F. Montanari, G. F. Ecker, C. H. Schwab, T. Kleinöder, T. Magdziarz and M. Pastor, Integrative modeling strategies for predicting drug toxicities at the eTOX project, *Mol. Inf.*, 2015, **34**(6–7), 477–484, DOI: 10.1002/minf.201400193.

179. M. Nendza, S. Gabbert, R. Kühne, A. Lombardo, A. Roncaglioni, E. Benfenati, R. Benigni, C. Bossa, S. Strempel, M. Scheringer, A. Fernández, R. Rallo, F. Giralt, S. Dimitrov, O. Mekenyan, F. Bringezu and G. Schüürmann, A comparative survey of chemistry-driven *in silico* methods to identify hazardous substances under REACH, *Regul. Toxicol. Pharmacol.*, 2013, **66**(3), 301–314, DOI: 10.1016/j.yrtph.2013.05.007.
180. A. Sutter, A. Amberg, S. Boyer, A. Brigo, J. F. Contrera, L. L. Custer, K. L. Dobo, V. Gervais, S. Glowienke, J. van Gompel, N. Greene, W. Muster, J. Nicolette, M. V. Reddy, V. Thybaud, E. Vock, A. T. White and L. Müller, Use of *in silico* systems and expert knowledge for structure-based assessment of potentially mutagenic impurities, *Regul. Toxicol. Pharmacol.*, 2013, **67**(1), 39–52, DOI: 10.1016/j.yrtph.2013.05.001.
181. A. G. Dossetter, E. J. Griffen and A. G. Leach, Matched Molecular Pair Analysis in drug discovery, *Drug Discovery Today*, 2013, **18**(15–16), 724–731, DOI: 10.1016/j.drudis.2013.03.003.
182. C. Kramer, J. E. Fuchs, S. Whitebread, P. Gedeck and K. R. Liedl, Matched Molecular Pair Analysis: Significance and the impact of experimental uncertainty, *J. Med. Chem.*, 2014, **57**(9), 3786–3802, DOI: 10.1021/jm500317a.
183. Y. Sushko, S. Novotarskyi, R. Körner, J. Vogt, A. Abdelaziz and I. V. Tetko, Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process, *J. Cheminf.*, 2014, **6**, 48, DOI: 10.1186/s13321-014-0048-0.
184. J. M. Beck and C. Springer, Quantitative structure-activity relationship models of chemical transformations from matched pairs analyses, *J. Chem. Inf. Model.*, 2014, **54**(4), 1226–1234, DOI: 10.1021/ci500012n.
185. A. de la Vega de León and J. Bajorath, Prediction of compound potency changes in matched molecular pairs using support vector regression, *J. Chem. Inf. Model.*, 2014, **54**(10), 2654–2663, DOI: 10.1021/ci5003944.
186. D. J. Warner, M. H. Bridgland-Taylor, C. E. Sefton and D. J. Wood, Prospective prediction of antitarget activity by Matched Molecular Pairs Analysis, *Mol. Inf.*, 2012, **31**(5), 365–368, DOI: 10.1002/minf.201200020.
187. A. de la Vega de León, Y. Hu and J. Bajorath, Systematic identification of matching molecular series and mapping of screening hits, *Mol. Inf.*, 2014, **33**(4), 257–263, DOI: 10.1002/minf.201400017.
188. *Predictive ADMET: Integrated Approaches in Drug Discovery and Development*, ed. J. Wang and L. Urban, Wiley, 2014, ISBN: 978-1-118-29992-0.
189. J. Wang and T. Hou, Advances in computationally modeling human oral bioavailability, *Adv. Drug Delivery Rev.*, 2015, **86**, 11–16, DOI: 10.1016/j.addr.2015.01.001.
190. F. T. Silva and G. H. Trossini, The survey of the use of QSAR methods to determine intestinal absorption and oral bioavailability during drug design, *Med. Chem.*, 2014, **10**(5), 441–448, DOI: 10.2174/1573406410666140415122115.
191. E. V. Radchenko, A. S. Dyabina, V. A. Palyulin and N. S. Zefirov, Prediction of human intestinal absorption of drug compounds, *Russ. Chem. Bull.*, 2016, **65**(2), 576–580, DOI: 10.1007/s11172-016-1340-0.

192. G. Lambrinidis, T. Vallianatou and A. Tsantili-Kakoulidou, *In vitro, in silico* and integrated strategies for the estimation of plasma protein binding. A review, *Adv. Drug Delivery Rev.*, 2015, **86**, 27–45, DOI: 10.1016/j.addr.2015.03.011.
193. K. Lanevskij, P. Japertas and R. Didziapetris, Improving the prediction of drug disposition in the brain, *Expert Opin. Drug Metab. Toxicol.*, 2013, **9**(4), 473–486, DOI: 10.1517/17425255.2013.754423.
194. L. Olsen, C. Oostenbrink and F. S. Jørgensen, Prediction of cytochrome P450 mediated metabolism, *Adv. Drug Delivery Rev.*, 2015, **86**, 61–71, DOI: 10.1016/j.addr.2015.04.020.
195. B. Feng, J. L. LaPerle, G. Chang and M. V. Varma, Renal clearance in drug discovery and development: molecular descriptors, drug transporters and disease state, *Expert Opin. Drug Metab. Toxicol.*, 2010, **6**(8), 939–952, DOI: 10.1517/17425255.2010.482930.
196. J. G. Hengstler, H. Foth, R. Kahl, P. J. Kramer, W. Lilienblum, T. Schulz and H. Schweinfurth, The REACH concept and its impact on toxicological sciences, *Toxicology*, 2006, **220**(2–3), 232–239, DOI: 10.1016/j.tox.2005.12.005.
197. C. M. Auer, J. V. Nabholz and K. P. Baetcke, Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure-activity relationships (SAR) under TSCA, Section 5, *Environ. Health Perspect.*, 1990, **87**, 183–197, DOI: 10.2307/3431024.
198. Organisation for Economic Co-Operation and Development, *OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure–Activity Relationship Models*, 2004, <http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>, accessed 01.06.2016.
199. European Chemicals Agency, *Guidance for the implementation of REACH. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.6: QSARs and Grouping of Chemicals*, May 2008, [https://echa.europa.eu/documents/10162/13632/information\\_requirements\\_r6\\_en.pdf](https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf), accessed 01.06.2016.
200. J. Devillers and H. Devillers, Prediction of acute mammalian toxicity from QSARs and interspecies correlations, *SAR QSAR Environ. Res.*, 2009, **20**(5–6), 467–500, DOI: 10.1080/10629360903278651.
201. A. Lagunin, A. Zakharov, D. Filimonov and V. Poroikov, QSAR modelling of rat acute toxicity on the basis of PASS prediction, *Mol. Inf.*, 2011, **30**(2–3), 241–250, DOI: 10.1002/minf.201000151.
202. M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I. V. Tetko and R. Todeschini, Prediction of acute aquatic toxicity toward *Daphnia magna* by using the GA-kNN method, *Altern. Lab. Anim.*, 2014, **42**(1), 31–41. <http://www.atla.org.uk/prediction-of-acute-aquatic-toxicity-toward-daphnia-magna-by-using-the-ga-knn-method/>, accessed 01.06.2016.
203. P. Gramatica, S. Cassania and A. Sangion, Aquatic ecotoxicity of personal care products: QSAR models and ranking for prioritization and safer alternatives' design, *Green Chem.*, 2016, **18**(16), 4393–4406, DOI: 10.1039/C5GC02818C.

204. R. Benigni and C. Bossa, Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology, *Chem. Rev.*, 2011, **111**(4), 2507–2536, DOI: 10.1021/cr100222q.
205. R. Benigni, C. Bossa, O. Tcheremenskaia and A. Giuliani, Alternatives to the carcinogenicity bioassay: *in silico* methods, and the *in vitro* and *in vivo* mutagenicity assays, *Expert Opin. Drug Metab. Toxicol.*, 2010, **6**(7), 809–819, DOI: 10.1517/17425255.2010.486400.
206. K. H. Hsu, B. H. Su, Y. S. Tu, O. A. Lin and Y. J. Tseng, Mutagenicity in a molecule: Identification of core structural features of mutagenicity using a scaffold analysis, *PLoS One*, 2016, **11**(2), e0148900, DOI: 10.1371/journal.pone.0148900.
207. B. O. Villoutreix and O. Taboureau, Computational investigations of hERG channel blockers: New insights and current predictive models, *Adv. Drug Delivery Rev.*, 2015, **86**, 72–82, DOI: 10.1016/j.addr.2015.03.003.
208. E. V. Radchenko, Yu. A. Rulev, A. Ya. Safanyaev, V. A. Palyulin and N. S. Zefirov, Computer-aided estimation of the hERG-mediated cardiotoxicity risk of potential drug compounds, *Dokl. Biochem. Biophys.*, 2017, **473**(1), in press.
209. M. Chen, H. Bisgin, L. Tong, H. Hong, H. Fang, J. Borlak and W. Tong, Toward predictive models for drug-induced liver injury in humans: are we there yet?, *Biomarkers Med.*, 2014, **8**(2), 201–213, DOI: 10.2217/bmm.13.146.
210. M. Novič and M. Vračko, QSAR models for reproductive toxicity and endocrine disruption activity, *Molecules*, 2010, **15**(3), 1987–1999, DOI: 10.3390/molecules15031987.
211. A. Rybacka, C. Rudén, I. V. Tetko and P. L. Andersson, Identifying potential endocrine disruptors among industrial chemicals and their metabolites – development and evaluation of *in silico* tools, *Chemosphere*, 2015, **139**, 372–378, DOI: 10.1016/j.chemosphere.2015.07.036.
212. R. Huang, M. Xia, S. Sakamuru, J. Zhao, S. A. Shahane, M. Attene-Ramos, T. Zhao, C. P. Austin and A. Simeonov, Modelling the Tox21 10 K chemical profiles for *in vivo* toxicity prediction and mechanism characterization, *Nat. Commun.*, 2016, **7**, 10425, DOI: 10.1038/ncomms10425.
213. R. Huang, M. Xia, D.-T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. A. Shahane, A. Rossoshek and A. Simeonov, Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs, *Front. Environ. Sci.*, 2016, **3**, 85, DOI: 10.3389/fenvs.2015.00085.
214. M. N. Drwal, V. B. Siramshetty, P. Banerjee, A. Goede, R. Preissner and M. Dunkel, Molecular similarity-based predictions of the Tox21 screening outcome, *Front. Environ. Sci.*, 2015, **3**, 54, DOI: 10.3389/fenvs.2015.00054.
215. F. Stefaniak, Prediction of compounds activity in nuclear receptor signaling and stress pathway assays using machine learning algorithms and low-dimensional molecular descriptors, *Front. Environ. Sci.*, 2015, **3**, 77, DOI: 10.3389/fenvs.2015.00077.

216. A. Mayr, G. Klambauer, T. Unterthiner and S. Hochreiter, DeepTox: Toxicity prediction using deep learning, *Front. Environ. Sci.*, 2016, **3**, 80, DOI: 10.3389/fenvs.2015.00080.
217. A. Abdelaziz, H. Spahn-Langguth, K.-W. Schramm and I. V. Tetko, Consensus modeling for HTS assays using *in silico* descriptors calculates the best balanced accuracy in Tox21 Challenge, *Front. Environ. Sci.*, 2016, **4**, 2, DOI: 10.3389/fenvs.2016.00002.
218. S. J. Capuzzi, R. Politi, O. Isayev, S. Farag and A. Tropsha, QSAR modeling of Tox21 Challenge stress response and nuclear receptor signaling toxicity assays, *Front. Environ. Sci.*, 2016, **4**, 3, DOI: 10.3389/fenvs.2016.00003.
219. Y. Uesawa, Rigorous selection of random forest models for identifying compounds that activate toxicity-related pathways, *Front. Environ. Sci.*, 2016, **4**, 9, DOI: 10.3389/fenvs.2016.00009.
220. A. Koutsoukas, J. St. Amand, M. Mishra and J. Huan, Predictive toxicology: Modeling chemical induced toxicological response combining circular fingerprints with random forest and support vector machine, *Front. Environ. Sci.*, 2016, **4**, 11, DOI: 10.3389/fenvs.2016.00011.
221. K. Ribay, M. T. Kim, W. Wang, D. Pinolini and H. Zhu, Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data, *Front. Environ. Sci.*, 2016, **4**, 12, DOI: 10.3389/fenvs.2016.00012.
222. G. Barta, Identifying biological pathway interrupting toxins using multi-tree ensembles, *Front. Environ. Sci.*, 2016, **4**, 52, DOI: 10.3389/fenvs.2016.00052.
223. S. M. Ivanov, A. A. Lagunin and V. V. Poroikov, *In silico* assessment of adverse drug reactions and associated mechanisms, *Drug Discovery Today*, 2016, **21**(1), 58–71, DOI: 10.1016/j.drudis.2015.07.018.
224. N. Ai, X. Fan and S. Ekins, *In silico* methods for predicting drug-drug interactions with cytochrome P-450s, transporters and beyond, *Adv. Drug Delivery Rev.*, 2015, **86**, 46–60, DOI: 10.1016/j.addr.2015.03.006.
225. F. Montanari and G. F. Ecker, Prediction of drug-ABC-transporter interaction – Recent advances and future challenges, *Adv. Drug Delivery Rev.*, 2015, **86**, 17–26, DOI: 10.1016/j.addr.2015.03.001.
226. F. Klepsch, P. Vasanthanathan and G. F. Ecker, Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors, *J. Chem. Inf. Model.*, 2014, **54**(1), 218–229, DOI: 10.1021/ci400289j.
227. A. V. Zakharov, E. V. Varlamova, A. A. Lagunin, A. V. Dmitriev, E. N. Muratov, D. Fourches, V. E. Kuz'min, V. V. Poroikov, A. Tropsha and M. C. Nicklaus, QSAR modeling and prediction of drug-drug interactions, *Mol. Pharmaceutics*, 2016, **13**(2), 545–556, DOI: 10.1021/acs.molpharmaceut.5b00762.
228. S. Ekins, A. J. Williams, M. D. Krasowski and J. S. Freundlich, *In silico* repositioning of approved drugs for rare and neglected diseases, *Drug Discovery Today*, 2011, **16**(7–8), 298–310, DOI: 10.1016/j.drudis.2011.02.016.



229. F. Schmidt, H. Matter, G. Hessler and A. Czich, Predictive *in silico* off-target profiling in drug discovery, *Future Med. Chem.*, 2014, **6**(3), 295–317, DOI: 10.4155/fmc.13.202.
230. L. C. Huang, X. Wu and J. Y. Chen, Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures, *Proteomics*, 2013, **13**(2), 313–324, DOI: 10.1002/pmic.201200337.
231. A. V. Stepanchikova, A. A. Lagunin, D. A. Filimonov and V. V. Poroikov, Prediction of biological activity spectra for substances: evaluation on the diverse sets of drug-like structures, *Curr. Med. Chem.*, 2003, **10**(3), 225–233, DOI: 10.2174/0929867033368510.
232. P. V. Pogodin, A. A. Lagunin, D. A. Filimonov and V. V. Poroikov, PASS Targets: Ligand-based multi-target computational system based on a public data and naïve Bayes approach, *SAR QSAR Environ. Res.*, 2015, **26**(10), 783–793, DOI: 10.1080/1062936X.2015.1078407.
233. A. V. Zakharov, A. A. Lagunin, D. A. Filimonov and V. V. Poroikov, Quantitative prediction of antitarget interaction profiles for chemical compounds, *Chem. Res. Toxicol.*, 2012, **25**(11), 2378–2385, DOI: 10.1021/tx300247r.
234. D. Filimonov, V. Poroikov, Y. Borodina and T. Glorizova, Chemical similarity assessment through Multilevel Neighborhoods of Atoms: Definition and comparison with the other descriptors, *J. Chem. Inf. Comput. Sci.*, 1999, **39**(4), 666–670, DOI: 10.1021/ci980335o.
235. D. A. Filimonov, A. V. Zakharov, A. A. Lagunin and V. V. Poroikov, QNA-based 'Star Track' QSAR approach, *SAR QSAR Environ. Res.*, 2009, **20**(7–8), 679–709, DOI: 10.1080/10629360903438370.
236. A. Lagunin, D. Filimonov and V. Poroikov, Multi-targeted natural products evaluation based on biological activity prediction with PASS, *Curr. Pharm. Des.*, 2010, **16**(15), 1703–1717, DOI: 10.2174/138161210791164063.
237. G. V. Kokurkina, M. D. Dutov, S. A. Shevelev, S. V. Popkov, A. V. Zakharov and V. V. Poroikov, Synthesis, antifungal activity and QSAR study of 2-arylhydroxynitroindoles, *Eur. J. Med. Chem.*, 2011, **46**(9), 4374–4382, DOI: 10.1016/j.ejmech.2011.07.008.
238. S. A. Kryzhanovskii, R. M. Salimov, A. A. Lagunin, D. A. Filimonov, T. A. Glorizova and V. V. Poroikov, Nootropic action of some antihypertensive drugs: computer predicting and experimental testing, *Pharm. Chem. J.*, 2012, **45**(10), 605–611, DOI: 10.1007/s11094-012-0689-0.
239. M. N. Kurilo, F. V. Ryzhkov, P. V. Karpov, E. V. Radchenko, V. A. Palyulin and N. S. Zefirov, Molecular design of selective ligands of chemokine receptors, *Dokl. Biochem. Biophys.*, 2015, **461**(1), 131–134, DOI: 10.1134/S1607672915020167.
240. R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96 000 compounds, *J. Pharm. Sci.*, 2009, **98**(3), 861–893, DOI: 10.1002/jps.21494.

241. I. V. Tetko, G. I. Poda, C. Ostermann and R. Mannhold, Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound, *Chem. Biodiversity*, 2009, **6**(11), 1837–1844, DOI: 10.1002/cbdv.200900075.
242. I. V. Tetko, S. Novotarskyi, I. Sushko, V. Ivanov, A. E. Petrenko, R. Dieden, F. Lebon and B. Mathieu, Development of dimethyl sulfoxide solubility models using 163 000 molecules: Using a domain applicability metric to select more reliable predictions, *J. Chem. Inf. Model.*, 2013, **53**(8), 1990–2000, DOI: 10.1021/ci400213d.
243. I. V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A. E. Petrenko, L. Charochkina and A. M. Asiri, How accurately can we predict the melting points of drug-like compounds? *J. Chem. Inf. Model.*, 2014, **54**(12), 3320–3329, DOI: 10.1021/ci5005288.
244. I. V. Tetko, D. M. Lowe and A. J. Williams, The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS, *J. Cheminf.*, 2016, **8**, 2, DOI: 10.1186/s13321-016-0113-y.
245. S. Vorberg and I. V. Tetko, Modeling the biodegradability of chemical compounds using the Online CHEMical Modeling Environment (OCHEM), *Mol. Inf.*, 2014, **33**(1), 73–85, DOI: 10.1002/minf.201300030.
246. <http://way2drug.com/>, accessed 01.06.2016.
247. A. V. Rudik, A. V. Dmitriev, A. A. Lagunin, D. A. Filimonov and V. V. Poroiakov, Metabolism site prediction based on xenobiotic structural formulas and PASS prediction algorithm, *J. Chem. Inf. Model.*, 2014, **54**(2), 498–507, DOI: 10.1021/ci400472j.
248. <https://www.ochem.eu/>, accessed 01.06.2016.
249. I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko and I. V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, *J. Comput.-Aided Mol. Des.*, 2011, **25**(6), 533–554, DOI: 10.1007/s10822-011-9440-2.
250. I. Sushko, E. Salmina, V. A. Potemkin, G. Poda and I. V. Tetko, ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions, *J. Chem. Inf. Model.*, 2012, **52**(8), 2310–2316, DOI: 10.1021/ci300245q.
251. <http://qsar.chem.msu.ru/admet/>, accessed 01.06.2016.



## CHAPTER 7

# ***In silico Chemical–Protein Docking and Molecular Dynamics***

SANJEEVA J. WIJEYESAKERE<sup>a</sup> AND RUDY J. RICHARDSON<sup>\*b,c</sup>

<sup>a</sup>Department of Microbiology and Immunology, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA; <sup>b</sup>Department of Environmental Health Sciences (Toxicology Program), University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; <sup>c</sup>Department of Neurology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

\*E-mail: [rjrich@umich.edu](mailto:rjrich@umich.edu)

## **7.1 Introduction**

This chapter briefly explores the principles and applications to the field of Toxicology of two computational structural biology techniques: (1) molecular docking of ligands (*e.g.* drugs or toxicants) with their biological receptors or targets (*e.g.* proteins or nucleic acids); and (2) molecular dynamics (MD) simulations of ligand–receptor complexes compared with receptors alone. In general, these *in silico* approaches enable us to do two important things: (1) gain insight into molecular mechanisms of toxicity; and (2) suggest mechanistic hypotheses to be tested experimentally. In addition, through the application of inverse docking and pharmacophore/toxicophore mapping, it is possible to identify potential macromolecular targets of toxicants, including

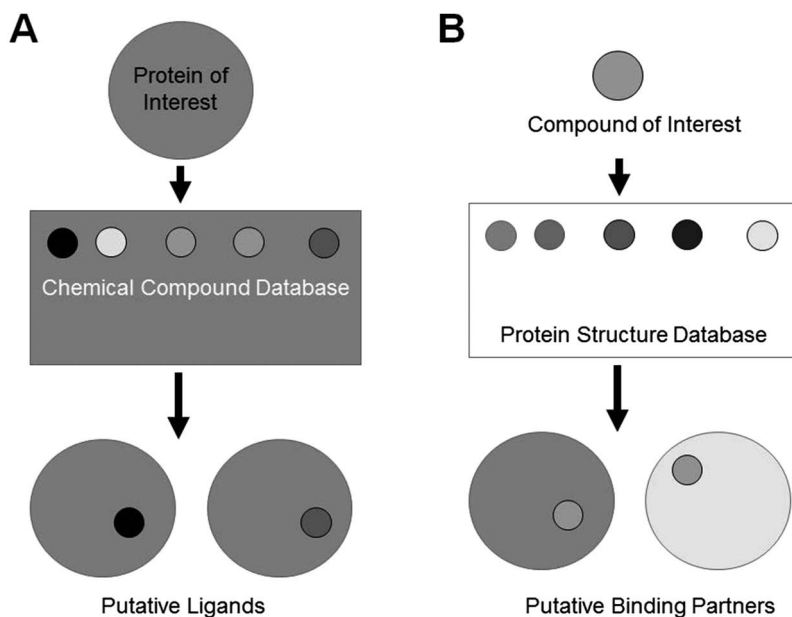
off-targets of pharmaceutical agents. Although docking and MD simulations can be used to examine protein-protein, protein-RNA/DNA, and protein-lipid interactions, we focus here on small-molecule toxicants and protein targets (*e.g.* enzymes or receptors).

Recent advances in the power of desktop computers have afforded scientists unprecedented opportunities to investigate the molecular basis of ligand-protein interactions using consumer-grade hardware. Indeed, together with experimental elucidation of structures *via* X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, which account for more than 95 000 and 9000 protein structures, respectively, in the Protein Data Bank ([www.pdb.org](http://www.pdb.org)),<sup>1</sup> molecular docking and MD simulations provide atomic-level descriptions of the interactions of toxicants with their biological targets. While these *in silico* techniques have been routinely employed in diverse fields such as Biophysics, Biochemistry, and Medicinal Chemistry, their use has been relatively sparse in the field of Toxicology. In this chapter, we briefly explore the theoretical basis and practical uses for computational techniques such as docking and MD simulations to understand the structural basis for toxicant-protein interactions and dynamic conformational changes within the protein associated with toxicant binding.

## 7.2 Molecular Docking: Overview and Applications

Molecular docking is a widely utilized tool in drug discovery that allows users to model ligand binding sites and conformations within a target of interest. In particular, docking allows the investigator to identify the orientation as well as the unique conformation(s) of a ligand in complex with a target of interest. Broadly speaking, docking consists of (1) determining the orientation of a ligand within a binding site (termed a 'pose'); and (2) assessment of the affinity of the identified pose (Figure 7.1A). This is undertaken *via* a scoring function, which comprises mathematical models that approximate the non-covalent binding energy of a ligand pose within the binding cavity of the receptor. In this manner, *via* iterative sampling of multiple ligand poses, multiple orientations of the ligand within a binding site can be modeled together with assessments of the favorability of each individual pose or cluster of poses within the interaction site.

While deriving from the same basic theory, and given the large number of possible conformational variants of a ligand within a binding site, docking algorithms employ distinct sampling techniques to identify energetically favorable ligand poses. The most commonly used sampling techniques are (1) genetic algorithms; (2) Monte Carlo techniques; and (3) matching algorithms. Furthermore, different software packages that employ the same docking algorithm may, in practice, yield different results based on differences in the scoring function used to delineate the thresholds that define the exclusion criteria for unfavorable ligand conformations. Overall, docking provides a fast, computationally inexpensive method to screen receptor-ligand interactions. The advantages and disadvantages of *in silico* docking are detailed in Table 7.1.



**Figure 7.1** Schematic highlighting the concepts that underlie (A) classical docking and (B) inverse docking. In general, classical docking allows for the screening of a library of compounds against a known receptor or target of interest, while inverse docking allows for the identification of potential interaction targets for a compound of interest.

**Table 7.1** Advantages and disadvantages of *in silico* docking.

Advantages	Disadvantages
Fast and computationally inexpensive	Potential for identification of incorrect binding sites or targets. This is especially the case with novel or previously uncharacterized xenobiotics
High throughput—can screen multiple ligands against a single target receptor (classical docking) or multiple target receptors against a single ligand (inverse docking)	Unreliable when used with ligands that covalently modify their target receptor(s)
Allows for rapid identification of novel and/or secondary ligand-binding sites within a receptor	
Identified ligand poses can be scored <i>in vacuo</i> or in implicit solvent to obtain theoretical binding affinities	
Results can be combined with pharmacophore mapping, shape-matching and molecular dynamic simulations to further investigate the nature and structural basis for a ligand–receptor interaction	

### 7.2.1 Genetic Algorithms

Genetic algorithm-based docking techniques are employed by widely used programs such as AutoDock<sup>2</sup> and GOLD<sup>3</sup> and are based on analogies to the principles of classical genetics and Darwinian evolution. In this technique, values corresponding to the orientation of a ligand within a receptor are encoded as 'genes'. Consequently, all factors that describe the orientation and conformation of a ligand within the system are denoted by its genotype and the interaction energy of the ligand with the receptor determines its fitness. Genetic algorithm-based docking techniques sample ligand conformations *via* mating random pairs of 'genes', with certain offspring (which correspond to newly generated ligand poses) undergoing pseudo-random mutations of single genes. Following each step, the fitness of the offspring is determined *via* the defined scoring function. Thus, if the fitness exceeds a pre-defined threshold, the corresponding offspring is selected to proceed through subsequent mating steps. In this manner, genetically unfit offspring (which denote energetically unfavorable ligand poses) are systematically screened out at each step, while favorable conformations are selected to proceed through subsequent generations.<sup>4,5</sup>

Recent versions of AutoDock have incorporated a Lamarckian genetic algorithm (LGA) that uses inverse mapping from a local search to yield a genotype from a phenotype. The LGA has proved to be more efficient and reliable than the earlier genetic algorithm.<sup>2</sup> More recently, the application of a gradient optimization approach in the local optimization procedure has yielded a new program called AutoDock Vina that is much faster and often more accurate than AutoDock.<sup>6</sup> However, there are certain situations, such as using metal ions as ligands, that Vina cannot handle; here AutoDock or GOLD could be used instead.

### 7.2.2 Monte Carlo Procedure

Monte Carlo-based docking techniques used by applications such as ICM (Molsoft)<sup>7</sup> start with an initial randomly generated ligand conformation that is subsequently modified *via* bond rotations or rigid body translations. As with genetic algorithm-based techniques, the resulting ligand conformations are scored using a pre-defined scoring function before the ligand is subjected to another round of modification. Energetically favorable conformations that exceed the defined scoring threshold are saved, while those that are energetically unfavorable are discarded. In this manner, multiple conformations can be sampled over the course of a docking run.

### 7.2.3 Matching Algorithms

Due to its inherent speed, matching algorithm-based docking can be used to screen compounds rapidly. Matching algorithm-based techniques combine analysis of the topology of the binding site with parameters derived

from pharmacophore modeling of the ligand and active site to identify putative binding conformations. To this end, the initial steps in a matching algorithm-based docking run involve generation of a 'ligand-accessible' molecular surface representation of the target receptor. Afterwards, the ligand is modeled into the identified binding site using pharmacophore-derived parameters as well as shape-fit algorithms.<sup>8</sup> Matching algorithm-based techniques account for numerous factors such as hydrogen-bonding and van der Waals interactions in evaluating the favorability of the ligand conformation<sup>9–11</sup> and are employed in applications such as Dock<sup>12</sup> and SanDock.<sup>8</sup>

### 7.3 Scoring Ligand Poses

All docking algorithms employ scoring functions to determine the energetic favorability of any given ligand pose. In general, the equations used to calculate scoring functions encapsulate key factors that mediate ligand–receptor interactions (such as hydrogen bonds, solvation effects, polar interactions, and non-polar interactions). Overall, there are three major classes of scoring functions:<sup>13</sup> (1) force-field based functions; (2) empirical functions; and (3) knowledge-based functions. In this section, we briefly describe the force-field and empirical scoring functions that are used by popular docking algorithms such as AutoDock, CHARMM and GOLD.

The docking algorithm in CHARMM<sup>14</sup> employs a force-field based scoring function that uses Lennard-Jones potentials to assess van der Waals interactions between the ligand and its receptor while also summing the effects of electrostatic interactions between the ligand and receptor in the ligand–receptor complex. Overall, the major drawback to the use of a force-field based scoring function is the presence of local minima that may skew the resulting energy calculations and which, in turn, necessitates the use of energy minimization steps prior to scoring a ligand pose.

In contrast, AutoDock,<sup>2,15</sup> AutoDock Vina<sup>6</sup> and GOLD<sup>16</sup> use semi-empirical or empirical scoring functions. Broadly speaking, these algorithms apply a weighting scheme to each interaction factor involved in describing the binding of a ligand to a receptor. These weights are calculated *via* analysis of a training data set and involve fitting the components of the scoring function to the experimentally derived binding constants. When using empirical or semi-empirical scoring functions, care must be taken to ensure that the ligand of interest is not overly dissimilar to the ligands used in the training set used to weight the various interaction terms.

Due to the inherent flaws in any single scoring function, alternate techniques can be used to reduce the likelihood of false-positives. These involve the use of consensus approaches where two or more scoring functions are used to independently rate any given pose.<sup>17</sup> In these analyses, only poses that are deemed favorable by two or more scoring functions are taken to signify favorable energetics of interaction.

## 7.4 Inverse Docking

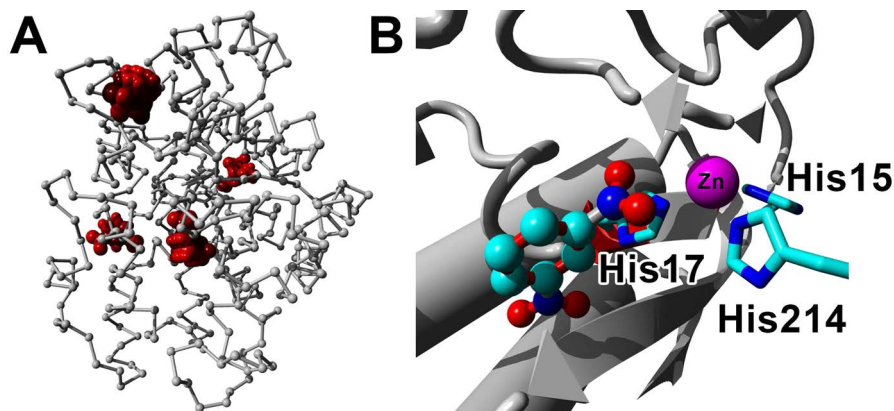
Thus far, the docking techniques described have consisted of modeling single or multiple ligands within the binding site(s) of a single receptor. In addition to this 'classical' docking technique, there exists an alternative technique called inverse docking (Figure 7.1B). Compared to classical docking, inverse docking is a relatively newer procedure and involves screening a single compound against a library of target receptors. This technique is of particular interest in fields such as Toxicology, where identification of 'off-targets' or secondary sites of toxicity could play a key role in the evaluation of the safety of novel compounds.

Similar to classical docking, inverse docking runs can be undertaken based on pharmacophore mapping<sup>18</sup> or 3D complementarity comparison.<sup>19</sup> By far the most computationally inexpensive inverse docking techniques employ inverse pharmacophore mapping. As such, tools such as PharmMapper utilize a database comprising pharmacophore models of target receptors (e.g. PharmDB) as the starting point for the modeling effort, with a pharmacophore model of the ligand of interest serving as the probe.<sup>20</sup>

In contrast to inverse pharmacophore mapping, inverse 3D complementarity comparisons are more akin to standard docking procedures, with the exception that docking scores for the ligand of interest are evaluated against a database of receptors. The earliest inverse 3D complementarity application, INVdock employed such an approach, with docking runs undertaken within multiple surface cavities of putative target receptors contained within a 3D structural database.<sup>21</sup> Newer programs such as MDock employ modified inverse docking algorithms with knowledge-based scoring functions to maximize speed, minimize computational overheads, and optimize scoring of the docking results.<sup>22,23</sup>

## 7.5 Case Study: Using *In silico* Docking to Investigate Interactions of 1,3-Dinitrobenzene with Adenosine Deaminase

1,3-dinitrobenzene (1,3-DNB) is a neurotoxicant that primarily affects astrocytes within the auditory and vestibular nuclei of the brainstem.<sup>24</sup> Adenosine deaminase (ADA) is a metalloprotein that plays a crucial role in the deamination of purines.<sup>25</sup> Given the neuroprotective effects of adenosine and structural similarities between 1,3-DNB and the adenine ring of adenosine,<sup>26</sup> interactions between 1,3-DNB and murine ADA were modeled *in silico* as part of a broader investigation to test the hypothesis that 1,3-DNB functions as an inhibitor of ADA.<sup>26</sup> Initial global docking of 1,3-DNB to murine ADA using AutoDock revealed the presence of four distinct clusters of energetically favorable ligand poses (defined as having a free energy of binding  $\Delta G_{\text{Binding}} < 0 \text{ kcal mol}^{-1}$ ) (Figure 7.2A). Furthermore, the cluster of poses of 1,3-DNB within the active site of ADA was found to have the lowest (most favorable)



**Figure 7.2** (A) Global docking of 1,3-dinitrobenzene (1,3-DNB) to the crystal structure of murine adenosine deaminase (ADA). ADA is rendered as a grey C- $\alpha$  atom trace with docked 1,3-DNB molecules shown as red spheres. (B) Analysis of the most energetically favorable pose of 1,3-DNB within the active site of ADA. 1,3-DNB is rendered as a ball-and-stick model with the active-site Zn<sup>2+</sup> moiety depicted as a magenta sphere with the coordinating histidine residues (His15, His17, and His214) rendered as sticks with carbons colored cyan. Hydrogen atoms are omitted for clarity. (Adapted from Wang *et al.*, Mixed inhibition of adenosine deaminase activity by 1,3-dinitrobenzene: a model for understanding cell-selective neurotoxicity in chemically-induced energy deprivation syndromes in brain. *Toxicol Sci*, 2012, 125(2): 509, with permission from Oxford University Press<sup>26</sup>).

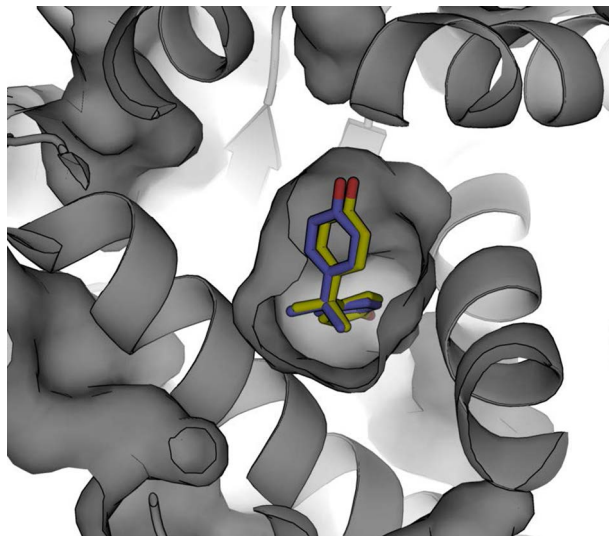
free energy of binding (Figure 7.2B), suggesting the inhibition of ADA by occupancy of the active site by 1,3-DNB.<sup>26</sup> Interaction of 1,3-DNB with the identified peripheral binding sites was consistent with kinetics results that fit a mixed model of inhibition, but experimental confirmation of these sites and the physiological consequences of ligand binding to these loci remain areas for further investigation.

## 7.6 Case Study: Using *In silico* Docking to Assess Binding of Bisphenol-A to Estrogen-Related Receptor- $\gamma$

Bisphenol-A (BPA) is a widely used plasticizer whose toxicological significance is currently the subject of public interest.<sup>27</sup> Previously, BPA has been shown to be a potent ligand for binding to the estrogen-related receptor- $\gamma$  (ERR $\gamma$ ), an orphan nuclear receptor whose precise physiological role is unclear.<sup>28,29</sup> Recently, the crystal structure of human ERR $\gamma$  in complex with BPA has been solved to a resolution of 1.6 Å.<sup>30</sup>

To assess the validity of *in silico* docking in modeling the orientation of a ligand within the binding site of a receptor, we docked BPA to ERR $\gamma$  in the





**Figure 7.3** Docking of bisphenol A (BPA) to human estrogen-related receptor- $\gamma$  (ERR $\gamma$ ). The crystal structure of ERR $\gamma$  (PDB ID 2E2R) (rendered as a grey cartoon with corresponding molecular surface) is shown in complex with BPA conformations determined *via* docking (rendered as sticks with cyan carbons) or the experimentally defined conformation (rendered as sticks with carbons colored yellow). The striking similarity between the computationally derived and experimentally defined conformations is evident. Hydrogen atoms have been omitted for clarity.

following manner. The crystal structure of human ERR $\gamma$  in complex with BPA (PDB ID 2E2R) was downloaded from the Protein Data Bank (PDB) ([www.pdb.org](http://www.pdb.org)). All water molecules and ligands were deleted, and hydrogens were added to the resulting apo-ERR $\gamma$  structure at a physiological pH (7.4). The receptor was then subjected to energy minimization *via* simulated annealing in YASARA-Structure (YASARA Biosciences; [www.yasara.org](http://www.yasara.org)). Hydrogens were added to the ligand (BPA) at pH 7.4 and its structure was independently optimized *via* semi-empirical techniques using the MOPAC plugin in YASARA-Structure. Next, BPA was docked to ERR $\gamma$  using YASARA-Structure as the graphical front-end to AutoDock Vina.<sup>6</sup> Docking was undertaken within a cubic simulation cell 75 Å per edge centered on the protein. This is a global ‘blind’ docking procedure that considers the entire protein molecule rather than constraining the search to a small region around the known binding site. The resulting poses were scored and the most favorable conformation was found to be identical to the ligand pose observed in the crystal structure of ERR $\gamma$  in complex with BPA (Figure 7.3). Taken together, these findings suggest that docking can predict the orientation of a ligand/toxicant within the binding site of a receptor. Such redocking procedures provide a good test of the reliability of any given docking program applied to a particular receptor and ligand.

## 7.7 Molecular Dynamics

Currently, due to technical limitations associated with NMR spectroscopy (protein size limitations) and electron microscopy (limited resolution), X-ray crystal structures constitute the vast majority of high-resolution experimentally determined macromolecular structures. However, a limitation of crystal structures of proteins is that the process of crystal formation and subsequent analysis of X-ray diffraction patterns results in an ‘averaging’ effect wherein molecular motions within the asymmetric unit are lost or unresolved. Thus, while we obtain structural data from X-ray crystallography, these gains are made at the expense of information concerning the conformational dynamics of the protein/macromolecule of interest.

Given that the dynamics of biological molecules in solution frequently determine or underlie alterations in their *in vivo* function, it is prudent to investigate the motions of proteins and protein–ligand complexes in addition to their static structures. To this end, the elucidation of the atomic motions within proteins of interest can be assessed *in silico via* MD simulations. The basic principles that define MD simulations were initially described in the mid-1970s,<sup>31</sup> and the field has advanced significantly over the ensuing years as computers have evolved to encompass ever greater computational power. In addition to intrinsic advances in individual central processing units (CPUs), multiple CPUs can be employed in parallel, either in high-performance computing clusters, or in multiple CPUs available as co-processor add-in cards for desktop computers. Computing power can also be greatly enhanced by using the graphical processing units on high-end video cards to perform computations in parallel. At their core, atomistic MD simulations employ the principles of Newtonian mechanics to simulate the motions of individual atoms over a simulated time period (termed a trajectory). Thus, MD simulations provide a mechanism for investigating the motion and conformational dynamics of macromolecules with or without bound ligands.

Other forms of MD involve the use of coarse-grained models to study very large systems or long time scales (reviewed in ref. 32). However, for the purposes of this chapter, we consider atomistic (all-atom) MD simulations, which provides the greatest detail in terms of the composition of the simulated system over time.

### 7.7.1 Running MD Simulations

There are numerous programs that can be used to run MD simulations. These include applications such as GROMACS,<sup>33</sup> NAMD,<sup>34</sup> Desmond (Schrodinger; [www.schrodinger.com/Desmond](http://www.schrodinger.com/Desmond)) and YASARA to name a few. Each of these implements its own MD algorithms and each has its advantages and disadvantages. Moreover, each program varies in the permissiveness of its licensing and ease of use. When running MD simulations, it is critical to ensure that trajectories are written in a common interoperable format (such as the GROMACS (\*.xtc) or CHARMM (\*.dcd) formats), which would allow

subsequent analyses to be performed using alternative trajectory analysis tools. It is noteworthy that trajectories obtained from MD simulations allow for the analysis of any atom within the system over the simulated time frame. As such, these analyses can be considered to constitute a unique class of 'big data' datasets wherein only a subset of atoms and atomic interactions are analyzed in any given study.

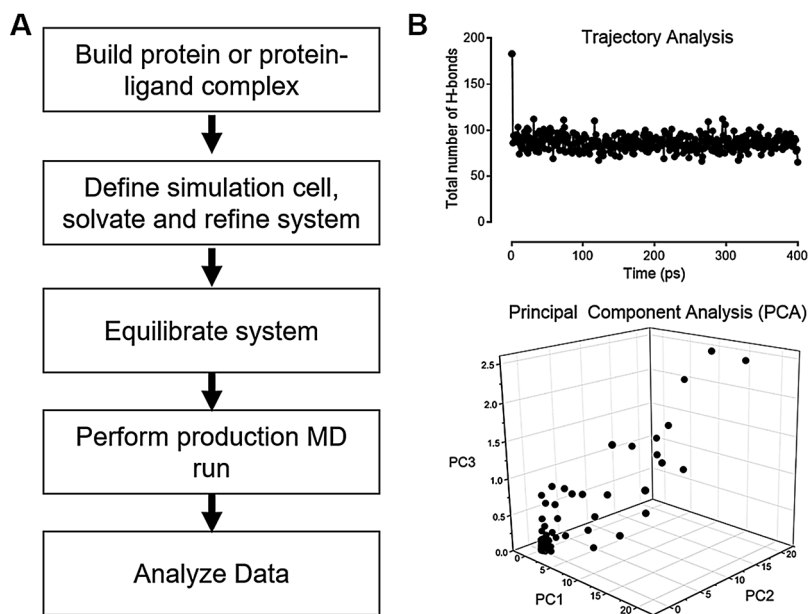
In order to begin an atomistic MD simulation, the macromolecule or macromolecule–ligand complex of interest is built and placed within a simulation cell whose walls are larger than the molecule of interest (typically extending 10–15 Å from the edges of the molecule). The resulting system is solvated (typically with explicit solvent consisting of TIP3 water molecules) and counter ions are added to neutralize charges within the cell. The system is then refined *via* multiple rounds of simulated annealing and allowed to equilibrate to the desired temperature (typically 298 K (24.85 °C) for a room-temperature run). Afterward, the production MD simulation is run, with the forces and acceleration of each atom in the system calculated at given intervals (termed time-steps) (Figure 7.4A). At each step, a ball-and-spring model is used with bonds treated as springs and the potential energies calculated from a pre-defined force-field.<sup>35</sup>

Once started, the system is allowed to run for the pre-determined number of steps, corresponding to a certain simulated time scale (typically within the nanosecond range). In this manner, the movement of each atom (and hence, the macromolecule) in the system is determined over the entire trajectory (Figure 7.4A). As expected, these calculations are much more computationally intensive than docking runs and consequently, the time-frames simulated over several days or weeks of computer time can be relatively modest (*e.g.* a few hundred nanoseconds). The advantages and disadvantages of MD simulations are detailed in Table 7.2.

### 7.7.2 Analysis of MD Trajectories

Each MD program brings with it its own analysis toolkit. However, trajectories written in interoperable formats allow users to import their data into an alternative analysis toolkit to evaluate the entire dataset or a subset thereof. Once the atomic motions of a simulated system have been recorded in a trajectory, the data can be analyzed in numerous ways. The simplest form of analysis involves measuring a single variable (such as inter-molecular distances, number of hydrogen bonds, or the radius of gyration of a macromolecule) over the course of the trajectory (Figure 7.4B). In general, these types of analyses can yield insights into potential conformational changes within the protein or major destabilization events during the simulation.

Additionally, more advanced analyses of MD trajectories involve the use of principal component analysis (PCA) to study the dynamics of specific regions of interest in a macromolecule. In these analyses, the trajectory is analyzed as a series of discrete components that measure the variance in the distribution of residues over the trajectory (Figure 7.4B). These analyses can



**Figure 7.4** Overview of molecular dynamics (MD) simulations. (A) Generalized schematic depicting the steps involved in setting up and running an MD simulation. (B) Example analyses from an MD simulation. Panels show examples of a trajectory (upper) and a principal component analysis (lower).

**Table 7.2** Advantages and disadvantages of molecular dynamics (MD) simulations.<sup>a</sup>

Advantages	Disadvantages
Allows for investigations into the dynamics of protein–ligand interactions	MD techniques are computationally expensive and relatively slow compared to other techniques (such as docking). However, multi-CPU/GPU systems can help accelerate MD runs
Allows for identification of conformational changes in proteins and tertiary structure changes over a simulated time frame	Classical MD simulations use Newtonian mechanics to study atomic motion. As such, it is not possible to model chemical reactions without incorporating quantum mechanical techniques into the simulation
Yields insights into protein stability following ligand binding	
MD simulations can run in explicit solvent and can simulate complex biological environments such as a lipid membrane	
When combined with quantum mechanical techniques, molecular dynamic simulation can be used to predict spectroscopic properties of the simulated system	

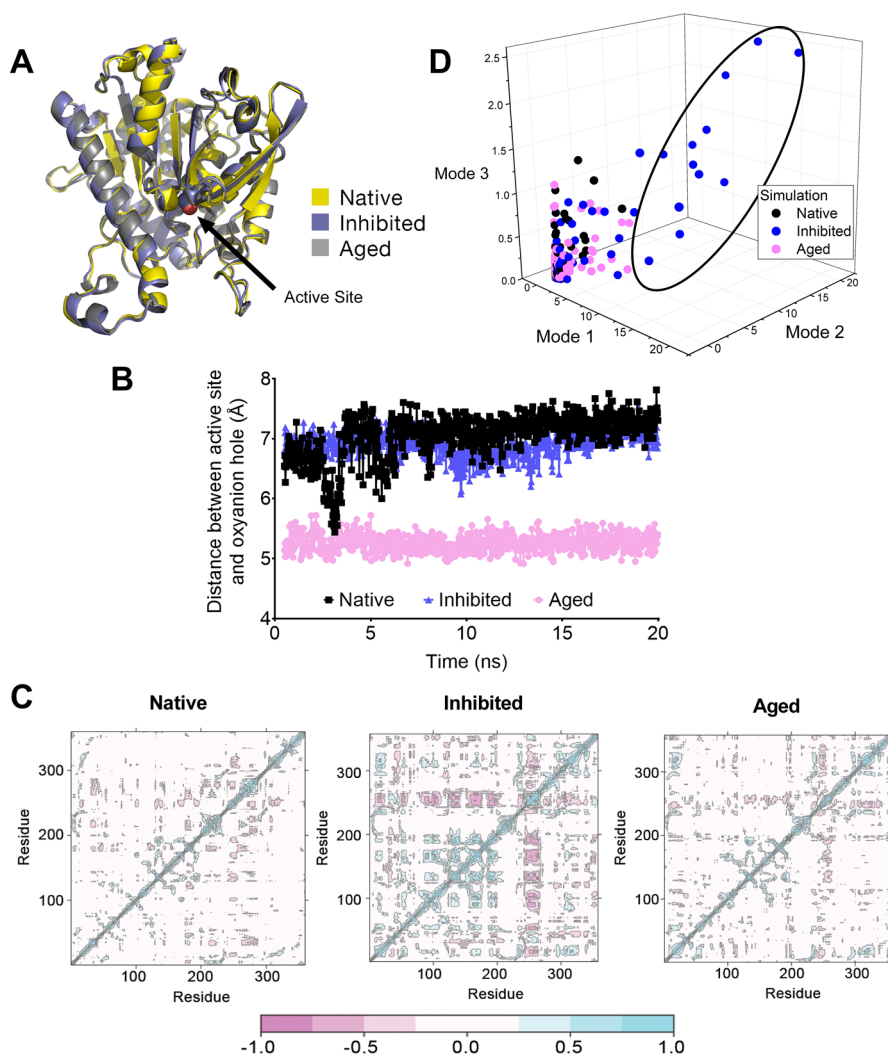
<sup>a</sup>CPU: central processing unit; GPU: graphical processing unit.

yield insights into regions of conformational space explored by the macromolecule over the course of the simulated time frame. Moreover, visualization of the principal component plots of a macromolecule of interest under different conditions (*e.g.* wild type *versus* mutant or native protein *versus* the protein in complex with ligands/toxicants of interest) can yield insights into differential sampling of conformational space of a modified protein relative to the native (or apo) state. Furthermore, a tighter cluster of components can be taken to indicate a more rigid structure, while a looser distribution characterized by multiple clusters can indicate a conformationally labile protein. Thus, while the crystal structure of a protein in complex with a ligand may appear to be indistinguishable from the crystal structure of the protein in its apo state, differences in flexibility and mobility can be accessed *via* PCA of MD trajectories of the native *versus* complexed states of a protein.

### 7.7.3 Case Study: Gaining Insights into the Conformational Dynamics of Human Neuropathy Target Esterase *via* MD Simulations of its Catalytic Domain Homologue Patatin-17 in Complex with Organophosphorous Compounds

Neuropathy target esterase (NTE; also known as patatin-like phospholipase domain-containing protein-6 (PNPLA6)) is a serine hydrolase whose structure is yet undetermined. NTE is member of a broad class of phospholipases termed the patatin-like phospholipase domain-containing proteins (reviewed in ref. 36 and 37) whose catalytic domains are homologous to patatin, a phospholipase found in potatoes and other members of the nightshade family.<sup>38,39</sup> NTE is the cellular target for a delayed neurotoxicity elicited by a subclass of organophosphorus compounds. Chemical modification of NTE by certain organophosphorous compounds triggers a delayed distal axonopathy with concomitant paralysis and sensory loss known as organophosphorous compound-induced delayed neuropathy, a condition for which there is no cure (reviewed in ref. 36 and 37).

Organophosphorous compounds can inhibit NTE and its catalytic domain homologue patatin by forming covalent adducts on the active site nucleophiles of the enzymes: Ser966 in NTE or Ser77 in patatin isoform 17 (pat17), thereby inhibiting the enzyme (reviewed in ref. 28). Following inhibition, these adducts can undergo a second, post-inhibitory reaction termed 'aging' wherein the organophosphorous adduct gains a negative charge *via* the net loss of an alkyl side chain (reviewed in ref. 36 and 37). While the structure of NTE is unknown, experimentally derived structures of pat17 its native state as well as in complex with aged and non-aged inhibitors have been described.<sup>40,41</sup> Interestingly, the overall static crystal structures of pat17 are virtually identical in all three states (ref. 41 and Figure 7.5A), with no major conformational changes being evident in the backbone C- $\alpha$  atoms of the protein.<sup>41</sup>



**Figure 7.5** Molecular Dynamics (MD) simulation of patatin-17 (pat17) in its native state as well as in complex with aged diisopropylfluorophosphate and non-aged methyl arachidonyl fluorophosphonate. (A) Comparison of the crystal structures of native, inhibited and aged pat17.<sup>41</sup> (B) Representative analysis of a 20 ns MD trajectory with respect to the total number of hydrogen bonds between the active site nucleophile (Ser77) and the oxyanion hole of pat17 in the native, inhibited and aged states. (C and D) Principal component analysis (PCA) of the MD trajectories. (C) Covariance matrices for the residues in native, inhibited and aged pat17 over the simulated time-frame with correlated (cyan) and anti-correlated (pink) residue motions indicated. (D) Plots of the top three PCA modes (denoted mode 1, mode 2, and mode 3), with each point representing a single residue. The ellipse indicates a region of conformational space that is exclusively sampled by the inhibited form of the enzyme.



To investigate changes in the conformational dynamics of pat17 (and, by analogy, the catalytic domain of NTE) in complex with aged and non-aged inhibitors, 20 ns MD simulations were run of crystal structures of native pat17 (PDB ID 4PK9) as well as in complex with aged diisopropylfluorophosphate (PDB ID 4PKA) and non-aged methyl arachidonyl fluorophosphonate (MAFP) (PDB ID 4PKB).<sup>41</sup> MD simulations were run in YASARA at a temperature of 300 K within a rectangular simulation cell extending 10 Å from the edges of the protein and in the presence of TIP3 water molecules, with residues hydrogenated at a pH of 7.4. The resulting trajectories were analyzed in VMD<sup>42</sup> and Bio3D.<sup>43</sup> Following completion of each MD simulation, various parameters (total number of hydrogen bonds, bonds and distances between specific atoms, *etc.*) were calculated in YASARA and VMD (Wijeyesakere, S. J. and Richardson, R. J., unpublished data). Initial trajectory analyses were undertaken of the distance between centers of mass of the active site nucleophile (Ser77) and Gly37 and Gly38, which form the oxyanion hole in pat17. No significant differences were observed between the distances of Ser77 and the oxyanion hole in native and inhibited pat17, while the aged state displayed a shorter distance with reduced variability over the simulated time frame (Figure 7.5B). Overall, these findings are consistent with observations from the crystal structure, which showed that the negatively charged DFP adduct is stabilized by the oxyanion hole in a manner similar to the tetrahedral intermediate during substrate catalysis.<sup>41</sup>

While the static structures of native, inhibited, and aged pat17 were relatively similar (Figure 7.5A), PCA of the trajectories revealed distinct differences in conformational dynamics of pat17. Overall, covariance analysis of the residue motions over the simulated time-frame revealed a marked increase in the correlated and anti-correlated motions of residues in the inhibited form of pat17 (Figure 7.4C). Furthermore, the conformational space sampled by the three states of the enzyme displayed distinct regions that were sampled exclusively by the inhibited state (Figure 7.5D, region within the ellipse). In this analysis, individual residues within each trajectory were calculated as the sum of 10 principal components, with the initial components (higher modes) corresponding to changes within the backbone of the protein. Thus, by plotting each of the higher modes as a function of the two other modes, differences in the conformational space sampled by differentially modified forms of the enzyme were visualized.

## References

1. H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki, The protein data bank, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2002, **58**, 899–907.
2. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, *J. Comput. Chem.*, 1998, **19**, 1639–1662.



3. M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, Improved protein-ligand docking using GOLD, *Proteins*, 2003, **52**, 609–623.
4. G. Bohm, New approaches in molecular structure prediction, *Biophys. Chem.*, 1996, **59**, 1–32.
5. G. Heberle and W. F. de Azevedo Jr, Bio-inspired algorithms applied to molecular docking simulations, *Curr. Med. Chem.*, 2011, **18**, 1339–1352.
6. O. Trott and A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.*, 2010, **31**, 455–461.
7. R. Abagyan, M. Totrov and D. Kuznetsov, ICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation, *J. Comput. Chem.*, 1994, **15**, 488–506.
8. P. Burkhard, P. Taylor and M. D. Walkinshaw, An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å X-ray structure of a Thrombin-Ligand complex, *J. Mol. Biol.*, 1998, **277**, 449–466.
9. A. T. Brint and P. Willett, Algorithms for the identification of three-dimensional maximal common substructures, *J. Chem. Inf. Model.*, 1987, **27**, 152–158.
10. I. Halperin, B. Ma, H. Wolfson and R. Nussinov, Principles of docking: An overview of search algorithms and a guide to scoring functions, *Proteins: Struct., Funct., Genet.*, 2002, **47**, 409–443.
11. R. Norel, D. Fischer, H. J. Wolfson and R. Nussinov, Molecular surface recognition by a computer vision-based technique, *Protein Eng. Des. Sel.*, 1994, **7**, 39–46.
12. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, A geometric approach to macromolecule-ligand interactions, *J. Mol. Biol.*, 1982, **161**, 269–288.
13. H. J. Bohm and M. Stahl, The use of scoring functions in drug discovery applications, *Rev. Comput. Chem.*, 2002, **18**, 41–87.
14. F. A. Momany and R. Rone, Validation of the general-purpose Quanta(R)3.2/Charmm(R) force-field, *J. Comput. Chem.*, 1992, **13**, 888–900.
15. D. S. Goodsell and A. J. Olson, Automated docking of substrates to proteins by simulated annealing, *Proteins: Struct., Funct., Genet.*, 1990, **8**, 195–202.
16. G. Jones, P. Willett and R. C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation, *J. Mol. Biol.*, 1995, **245**, 43–53.
17. P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins, *J. Med. Chem.*, 1999, **42**, 5100–5109.
18. B. Bhattacharjee and J. Chatterjee, Identification of Proapoptotic, Anti-Inflammatory, anti-proliferative, anti-invasive and anti-angiogenic targets of essential oils in cardamom by dual reverse virtual screening and binding pose analysis, *Asian Pac. J. Cancer Prev.*, 2013, **14**, 3735–3742.

19. S. L. Kinnings and R. M. Jackson, ReverseScreen3D: A structure-based ligand matching method to identify protein targets, *J. Chem. Inf. Model.*, 2011, **51**, 624–634.
20. X. Liu, S. Ouyang, B. Yu, Y. Liu, K. Huang, J. Gong, S. Zheng, Z. Li, H. Li and H. Jiang, PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach, *Nucleic Acids Res.*, 2010, **38**, W609–W614.
21. Y. Z. Chen and D. G. Zhi, Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule, *Proteins: Struct., Funct., Genet.*, 2001, **43**, 217–226.
22. S. Z. Grinter, Y. Liang, S.-Y. Huang, S. M. Hyder and X. Zou, An inverse docking approach for identifying new potential anti-cancer targets, *J. Mol. Graph. Model.*, 2011, **29**, 795–799.
23. S.-Y. Huang and X. Zou, An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials, *J. Comput. Chem.*, 2006, **27**, 1866–1875.
24. M. A. Philbert, C. C. Nolan, J. E. Cremer, D. Tucker and A. W. Brown, 1,3-Dinitrobenzene-induced encephalopathy in rats, *Neuropathol. Appl. Neurobiol.*, 1987, **13**, 371–389.
25. G. Cristalli, S. Costanzi, C. Lambertucci, G. Lupidi, S. Vittori, R. Volpini and E. Camaioni, Adenosine deaminase: functional implications and different classes of inhibitors, *Med. Res. Rev.*, 2001, **21**, 105–128.
26. Y. Wang, X. Liu, B. Schneider, E. A. Zverina, K. Russ, S. J. Wijeyesakere, C. A. Fierke, R. J. Richardson and M. A. Philbert, Mixed inhibition of adenosine deaminase activity by 1,3-dinitrobenzene: a model for understanding cell-selective neurotoxicity in chemically-induced energy deprivation syndromes in brain, *Toxicol. Sci.*, 2012, **125**, 509–521.
27. L. N. Vandenberg, R. Hauser, M. Marcus, N. Olea and W. V. Welshons, Human exposure to bisphenol A (BPA), *Reprod. Toxicol.*, 2007, **24**, 139–177.
28. S. Takayanagi, T. Tokunaga, X. Liu, H. Okada, A. Matsushima and Y. Shimohigashi, Endocrine disruptor bisphenol A strongly binds to human estrogen-related receptor  $\gamma$  (ERR $\gamma$ ) with high constitutive activity, *Toxicol. Lett.*, 2006, **167**, 95–105.
29. H. Okada, T. Tokunaga, X. Liu, S. Takayanagi, A. Matsushima and Y. Shimohigashi, Direct evidence revealing structural elements essential for the high binding ability of bisphenol A to human estrogen-related receptor- $\gamma$ , *Environ. Health Perspect.*, 2008, **116**, 32–38.
30. A. Matsushima, Y. Kakuta, T. Teramoto, T. Koshiba, X. Liu, H. Okada, T. Tokunaga, S. Kawabata, M. Kimura and Y. Shimohigashi, Structural evidence for endocrine disruptor bisphenol A binding to human nuclear receptor ERR $\gamma$ , *J. Biochem.*, 2007, **142**, 517–524.
31. A. Warshel and M. Levitt, Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme, *J. Mol. Biol.*, 1976, **103**, 227–249.
32. M. Orozco, L. Orellana, A. Hospital, A. N. Naganathan, A. Emperador, O. Carrillo and J. L. Gelpi, Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings, *Adv. Protein Chem. Struct. Biol.*, 2011, **85**, 183–215.

33. H. J. C. Berendsen, D. van der Spoel and R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun.*, 1995, **91**, 43–56.
34. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, Scalable molecular dynamics with NAMD, *J. Comput. Chem.*, 2005, **26**, 1781–1802.
35. M. Rueda, C. Ferrer-Costa, T. Meyer, A. Perez, J. Camps, A. Hospital, J. L. Gelpi and M. Orozco, A consensus view of protein dynamics, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 796–801.
36. R. J. Richardson, N. D. Hein, S. J. Wijeyesakere, J. K. Fink and G. F. Makhaeva, Neuropathy target esterase (NTE): overview and future, *Chem. Biol. Interact.*, 2013, **203**, 238–244.
37. S. J. Wijeyesakere and R. J. Richardson, Neuropathy target esterase, in *Hayes' Handbook of Pesticide Toxicology*, ed. R. Krieger, Elsevier/Academic Press, San Diego, 3rd edn, 2010, vol. 2, pp. 1435–1478.
38. M. W. Ganai, M. W. Bonierbale, M. S. Roeder, W. D. Park and S. D. Tanksley, Genetic and physical mapping of the patatin genes in potato and tomato, *Mol. Gen. Genet.*, 1991, **225**, 501–509.
39. G. Vancanneyt, U. Sonnewald, R. Hofgen and L. Willmitzer, Expression of a patatin-like protein in the anthers of potato and sweet pepper flowers, *Plant Cell*, 1989, **1**, 533–540.
40. T. J. Rydel, J. M. Williams, E. Krieger, F. Moshiri, W. C. Stallings, S. M. Brown, J. C. Pershing, J. P. Purcell and M. F. Alibhai, The crystal structure, mutagenesis, and activity studies reveal that patatin is a lipid acyl hydrolase with a Ser-Asp catalytic dyad, *Biochemistry*, 2003, **42**, 6696–6708.
41. S. J. Wijeyesakere, R. J. Richardson and J. A. Stuckey, Crystal structure of patatin-17 in complex with aged and non-aged organophosphorus compounds, *PLoS One*, 2014, **9**, e108245.
42. W. Humphrey, A. Dalke and K. Schulten, VMD: Visual molecular dynamics, *J. Mol. Graphics*, 1996, **14**, 33–38.
43. B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. D. Caves, Bio3d: an R package for the comparative analysis of protein structures, *Bioinformatics*, 2006, **22**, 2695–2696.

## CHAPTER 8

# *Computational Tools for Chemical Toxicity Testing and Risk Assessment Under the Framework of Adverse Outcome Pathways*

M. MUMTAZ<sup>\*a</sup>, P. RUIZ<sup>a</sup> AND Q. ZHANG<sup>b</sup>

<sup>a</sup>Agency for Toxic Substances and Disease Registry (ATSDR), Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA; <sup>b</sup>Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

\*E-mail: [mgm4@cdc.gov](mailto:mgm4@cdc.gov)

## 8.1 Introduction

We are exposed to hundreds of chemicals routinely, whether intentionally or unintentionally. That number increases each year as more chemicals are introduced in commerce and the world we live in. Chemical risk assessments estimate public health consequences from exposure to chemicals in the environment, workplace, and medicine. Traditional experimental determination of toxicity profiles using laboratory animals takes a great deal of time, money, and other resources. Classically, toxicity testing gathers data on set exposures

of 14 days, 90 days, and 2 years in animal bioassays. Increased access to and thorough analyses of these data during the past decade has brought a realization that the results have not significantly advanced quantitative assessment of chemical toxicity. These classical approaches were based on a coarse and rudimentary understanding of the mechanism of action(s) leading to toxicity; the amount of modern knowledge of biology and physiology used for risk assessment, particularly at the cellular and molecular levels, has been minimal.

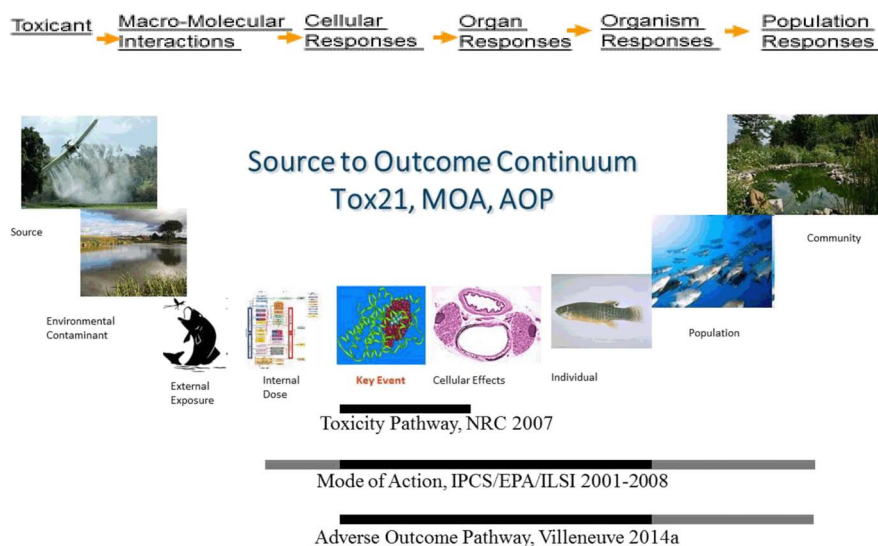
In response to such concerns about classical toxicity testing approaches, efforts have been made on multiple fronts to develop alternatives. These alternative approaches would make best use of modern science. They also would be economical, efficient, humane, and informative to real-world human exposures. The National Academy of Sciences (NAS) proposed a revolutionary new approach, TT21C, published in 2007.<sup>1</sup> NAS recommended rigorous investigations of perturbation of generally accepted toxicity pathways at the molecular and cellular levels. The Interagency Coordination Committee on the Validation of Alternative Methods Authorization Act, passed by the United States Congress in 2000, has authorized the adoption of tests that will also achieve the goal of humane treatment of animals by reducing, refining, and replacing animal toxicity testing.<sup>2</sup> Through these activities, toxicology is taking a more aggressive approach to overcome the paucity of data. Toxicology now accesses a broad panel of *in vitro* assays that the drug discovery industry has been using for years.<sup>3</sup> Several US government agencies and organizations around the world are integrating these methods for high-throughput screening (HTS) of chemicals. Noteworthy government-funded programs include the US Environmental Protection Agency (EPA) Toxicity ForeCaster (ToxCast) and Toxicology in the 21st Century (Tox21). Registration, Evaluation, Authorisation, and Restriction of Chemical substances (REACH) is a European initiative that will generate experimental data at a pace never witnessed in the history of the toxicology. The advent of numerous *in vitro* methods, rapid increase in associated databases, and the advancement in data analysis tools of large data sets has allowed us to holistically view the toxicity assessment process in light of systems biology and toxicity pathway analysis. Through these means, we might then overcome the shortcomings of the animal toxicity testing protocols.<sup>4,5</sup>

Such thinking has led to the development of a new “adverse outcome pathway” (AOP) framework that covers the gamut of biological processes involved in risk assessment, from the exposure source to population and community impacts.<sup>6,7</sup> Through these strategy and approach developments, we see that computational tools based on biological mechanisms would play an increasingly large role in chemical risk assessment, hand-in-hand with novel ways of experimental interrogations of the underlying biological systems. Some of these computational tools needed to implement the AOP framework have been in use and improved over many years. Examples include quantitative structure–activity relationship and physiologically based pharmacokinetic (PBPK) modeling. Others, such as systems biology pathways modeling,

are still emerging and being developed. AOPs piece together these various approaches under a single umbrella for a common cause. In this chapter, we give a brief introduction of the traditional quantitative approaches in risk assessment and how some of them can be adapted to the new AOP framework. We then focus on the computational tools for systems biology-based toxicity pathway modeling. These computational tools will be used to understand and predict cellular and tissue responses. They will be developed in conjunction with the animal-free movement of toxicity testing and risk assessment.

## 8.2 The AOP Concept

The AOP framework is a pragmatic simplification of biology. It provides a framework to collect, organize, and evaluate relevant information on chemical, biological, and toxicological effects of chemicals.<sup>6,7</sup> It is a construct that shows linkages between molecular interaction of an environmental chemical and specific biomolecules at various levels within an organism, and affected population in a broader context. It is developed and supported by data from a range of studies. These include *in vivo*, *in vitro*, HTS assays, omics, and *in silico* models. It represents a number of biochemical steps required for progression of a toxic response. These include molecular initiating events (MIEs) that serve as the origination of perturbation of the toxicity pathway and adjacent key events that are essential steps along the pathway representing the series of interactions in a system leading to an adverse effect (Figure 8.1). It is preferable to identify at least one key event at each level of the biological



**Figure 8.1** Key features of adverse outcome pathways.

organization (molecular, cellular, tissue, organ, organ system, and individual), without which no progression to the downstream events in the AOP occur.

Ideally, an AOP should include a relatively small or finite number of key events to establish the causal linkage between the MIE and the final adverse outcome. Identification of the MIE is a prerequisite for all subsequent steps. A single MIE can affect several signaling cascades (toxicity pathways). Toxicity is multi-modal, hence a particular MIE might lead to several final outcomes. Conversely, several MIEs might lead to the same final outcome. However, an AOP often supports an evaluation focusing on just one MIE and a single final adverse effect. Each component of an AOP might itself be influenced by other pathways within the biological system being modeled. Moreover, non-branched sequences of key events that formulate an AOP unit likely possess key events or key event relationships (KERs) that are shared with other AOPs, generating interactions that lead to the creation of AOP networks. These AOP networks provide a broader picture and encompass all available key events and KERs that potentially can be perturbed, leading to the same adverse outcomes. Such AOP networks can serve as key tools for predicting adverse outcomes based on mechanistic or pathway-based information and are thought to be applicable for most real-world scenarios.<sup>7</sup>

In developing AOPs and AOP networks, critical data gaps might exist in establishing the cause–effect relationships and in accurately extrapolating the adverse outcomes for environmental low-dose exposures based on available assays. These data gaps include the relationships between environmental pollutants, their environmental transport pathways, bioavailability, human population exposure, and the resulting change in health. Computational techniques have to be used to fill the data gaps, on an as-needed basis. These techniques use mathematical know-how and computer science advances to estimate or predict the inherent toxicity with quantifiable health risks.

Following the information flow in an AOP, three primary computational approaches can be used. The first of these is PBPK modeling to determine tissue concentrations and reversely predict exposure doses that result in tissue concentrations leading to adverse outcomes as determined by cell-based *in vitro* assays. The second approach is improved structure–activity relationships (SAR) modeling to determine MIEs. The third approach is systems biology modeling of toxicity pathways to determine and predict cellular responses and physiological-level modeling for organ and organism responses. Before we delve into each of the approaches, we will briefly review the mathematical methods traditionally used in quantifying toxicity and risk anchored on apical endpoints in experimental animals.

### 8.3 Quantitative Methods in Traditional Apical Endpoints Testing

Several quantitative methods have been used to determine the apical toxicity of chemicals. The earliest and the most often used test of acute potency and toxicity of chemicals is the LD<sub>50</sub> value. This is the dose that kills 50% of



test species (animals) in a single exposure through a given route of exposure. Although a crude test of toxicity, the common denominator being death, it allows comparison of acute toxicity among various chemicals that might cause different health effects. This test has given way to more subtle toxicity tests because of public concern about animal welfare.

One of them, maximum tolerable dose (MTD), is determined through a dose range-finding study. It is the highest administered dose that causes minimal toxicity, often a slight suppression in body weight gain, but other indicators of concerned toxicity can also be used. MTD, in turn, is used to space the doses to conduct complete dose-response testing studies. The principal goals of experimental dose-response studies, subchronic or chronic, are to identify a point of departure (POD), the point that marks the beginning of a low-dose extrapolation. PODs such as the no observed adverse effect levels (NOAELs) or lowest observed adverse effect levels (LOAELs) are used to derive acceptable levels. Various federal agencies, such as EPA, the Agency for Toxic Substances and Disease Registry (ATSDR), the National Institute for Occupational Safety and Health, and the Occupational Safety and Health Administration, derive acceptable levels to meet their mandates. Those levels go by various names: reference doses, reference concentrations, minimal risk levels, and permissible exposure limits. These levels are used to ascertain significant human exposure levels for acute, subchronic, and chronic health effects of chemicals to safeguard the humans and the environment from undue exposure to chemicals and their mixtures.

In real life, people are exposed to multiple chemicals through multiple routes for variable durations. Of necessity, the toxicity assessment process is often forced to rely on data derived from studies entailing time-weighted average exposures to single chemicals, single routes of exposure, and short-term exposure scenarios. Thus, acceptable levels are derived for limited exposure scenarios. Acceptable levels are derived for adverse health effects. These levels are supported by epidemiological data, for a specified duration of exposure, and typically entail the application of uncertainty factors to experimentally derived NOAELs or LOAELs.

The benchmark dose (BMD), an enhancement of the acceptable levels approach, has gained prominence.<sup>8</sup> BMD is a mathematical fitting method that allows us to determine a statistical lower confidence limit for a dose that produces a predetermined response rate of an adverse effect (the benchmark response) compared to background. BMD takes into account dose-response information by fitting a predetermined set of mathematical model equations to dose-response data. The benchmark response is generally set near the lower limit of responses that can be measured directly in animal experiments. Thus, unlike the risk assessment methods used for cancer effects, the BMD method does not extrapolate to doses far below the experimental range. The BMD approach presents a significant opportunity to improve the scientific basis of non-cancer risk assessment by overcoming limitations of the NOAEL approach, such as restricting the acceptable level derivation to the experimental tested doses or to a single NOAEL dose.

Despite the improvements in providing more informative risk assessment, quantitative approaches are still largely based on traditional toxicity testing studies. They use little modern biology, particularly at cellular and molecular levels. Their inflexibility affects their ability to make predictions for human exposure scenarios drastically different from the animal experimental setup. Some use for BMD has been found in describing *in vitro* response data and correlating them to *in vivo* data.<sup>9</sup> However, as animal-based toxicity testing is phased out, and as cell-based *in vitro* assays mature as alternatives, computational approaches are needed that mechanistically address key events in AOPs.

## 8.4 PBPK Modeling and *In vitro* to *In vivo* Extrapolation

PBPK models are a family of computational tools to relate external exposures to internal tissue concentrations of chemicals. Their potential applications in human health risk assessment have stirred considerable interest.<sup>10–13</sup> The salient feature of PBPK models is that they simulate and approximate the kinetic behavior of chemical(s) through integration of biological and physiological processes of a biological system.<sup>14,15</sup> PBPK models are designed to predict the internal tissue concentration of a chemical after exposure by one or more routes and doses in a particular species. They bridge major data gaps in chemical risk assessment.<sup>16,17</sup>

PBPK models can be adapted to achieve the reverse goal, as suggested by the TT21C approach.<sup>18,19</sup> For example, we might know the *in vitro* concentration that causes an adverse effect at a cellular level, as determined by *in vitro* cell-based assays anchored on toxicity pathways. We then could use PBPK models to extrapolate to *in vivo* tissue levels and ultimately to a human-allowable external dose. This process is called reverse dosimetry or *in vitro* to *in vivo* extrapolation (IVIVE).<sup>19,20</sup> Determinations of human external exposures to chemicals based on *in vitro* concentrations (or measured blood/tissue levels) are not trivial. Although they only represent the concentration at the sample time, they are a product of complex exposures from multiple routes and multiple sources.<sup>20–22</sup> Combining *in vitro* pharmacokinetics and pharmacodynamics information through interrogating cells with *in vitro* assays can produce a concentration that could be used for risk assessment.<sup>20,23</sup> Once identified, the *in vivo* exposure corresponding to the *in vitro* concentration can be estimated using IVIVE.<sup>24</sup> This is achieved by developing appropriate dose metrics, incorporating mode of action and other chemical-specific information to predict *in vivo* dose–response curves from *in vitro* data.<sup>20,24</sup> Thus, PBPK models can serve two purposes in future research endeavors. They can be used to (1) predict target organ concentrations of chemicals and (2) integrate such information to predict human exposures that are deemed safe or toxic.

Although multiple PBPK models are available, they sometimes seem too complex for health risk assessors to apply in the field. An additional challenge

for risk assessors is that the models are in multiple simulation languages that require advanced education and training. To overcome this challenge and promote field application and use of PBPK models, the ATSDR has converted and is recoding the best available, published PBPK models from multiple simulation languages into a single, easy to learn and use language, Berkeley Madonna (Kagi Shareware, Berkeley, CA, USA).<sup>25</sup> This library of models (tool-kit) currently houses models for several priority pollutants including some volatile organic compounds, metals/metalloids, and persistent organic pollutants. The toolkit will help risk assessors and researchers assess the potential health effects of chemicals collected during biological monitoring at waste sites. Another way to increase their use and acceptance is through joint efforts between the model developers and model users. Such interactions can provide common ground for understanding a model's advantages and limitations, thus optimizing the benefits from these models through information exchange and shared know-how.

The US Centers for Disease Control and Prevention's National Health and Nutrition Examination Survey provides representative data for hundreds of chemicals found in the general population across the United States. Several similar biomonitoring programs have been undertaken by states such as California, and other nations, such as Canada and Germany. Such programs give risk managers opportunities to use dosimetry techniques using PBPK models to interpret site-specific data on levels found in the general population.<sup>19</sup> Biomonitoring equivalents (estimates of the concentration of a chemical or its metabolite in a biological medium) also have been used to interpret environmental exposures.<sup>26–29</sup> The available computational tools have the potential to contribute to future toxicological testing strategies and become an integral part of the risk assessment process.<sup>11</sup>

## 8.5 SAR Modeling

A traditional way to predict the toxicity of new chemicals is through computational chemistry exploiting chemical structures. SAR models are a family of computational tools based on correlations between the molecular structures and biological activities/toxicities of chemicals.<sup>30–33</sup> A SAR model establishes qualitative association between a chemical's substructure, called alert, and its toxicity potential. In many cases, such models can also provide mechanistic information, as long as it was collected from the published literature during knowledge generation. An experienced user can then assess the relevance of a toxicological alert.

A quantitative structure–toxicity relationship (QSTR) is a mathematical relationship between a chemical's quantitative molecular descriptors and its toxicological endpoint. Molecular descriptors derived from atomic or molecular properties that encode physicochemical (*e.g.* octanol–water partition coefficient), topological (*e.g.* electrotopological states), and surface properties (*e.g.* polarity) of molecules form the backbone of a predictive QSTR. These descriptors are then correlated with a toxicological response

of interest through a suitable statistical approach, such as linear multiple regression, discriminant analysis, recursive partitioning, or artificial neural networks. QSTR provides discrete quantitatively predicted values. Such models help to estimate potential toxic effects of a large number of chemicals using data on individual chemicals and their relationship to other members of the same class. These models cannot replace standard bioassays or protocols, but aid decision making until necessary data become available. Many commercial and open source software packages, such as TOPKAT, CASE and MultiCASE, Derek (Lhasa), OncoLogic™, and ToxTree,<sup>34</sup> allow us to predict toxicity solely from chemical structure. Chemical and pharmaceutical industries and regulatory agencies, including the ATSDR, EPA, US Food and Drug Administration, Denmark's Environmental Protection Agency, and the United Kingdom's Health and Safety Executive, have used these tools to rapidly assess potential toxicity given just the drawing of a chemical structure.

SAR-based modeling can have several applications in the alternative testing era. As the TT21C efforts are being implemented, QSTR techniques, by providing rational models for interim estimate of the toxicity of chemicals with limited data, can help to set priorities for HTS. In addition, SAR modeling, along with simulations such as molecular docking, can provide crucial information regarding MIEs in the AOP framework.<sup>32</sup> Several other chapters in this book give a more in-depth introduction to this technique.

## 8.6 Computational Modeling of Toxicity Pathways

### 8.6.1 Concept of Toxicity Pathways

The pharmacokinetic process determines the concentrations of chemicals reaching various organs and tissues of the human body. These tissue concentrations are what cells “see” and are perturbed by. The concentrations fluctuate over time in accordance with constantly changing external exposure levels. The adverse biological effects of chemicals, starting from this point on, fall in the realm of the traditional pharmacodynamics or toxicodynamics, a field that studies what chemicals do to the body. As part of the new AOP framework, this toxicodynamic process now includes MIEs, toxicity pathway perturbations defining cellular responses, and subsequent key events leading to adverse individual and population outcomes.

Toxicity pathways, a key term coined in the 2007 National Research Council (NRC) report, conceptually encompass all biochemical networks operating in cells that can be perturbed by extrinsic chemicals.<sup>1</sup> These biochemical networks comprise properly integrated gene, protein, and metabolic pathways that maintain physiological cellular functions and responses. An MIE of a chemical corresponds to its interaction with one or more molecular targets (nodes) of these complex intracellular biochemical networks. If the chemical or its metabolite is reactive, many nodes can be targeted, as in the case of oxidative chemicals and electrophilic metabolites. The MIE of a chemical can also be specific, targeting one or two molecular nodes, such as endocrine

disruptors interacting with nuclear receptors. The molecular perturbations first occurring at the points of MIEs will spread to a subset of the intracellular biochemical networks, activating or suppressing their activities. If the scope and magnitude of network perturbation reach such an extent that it eventually leads to adverse health outcomes, then this subset of intracellular biochemical networks can be viewed as the toxicity pathway associated with that particular chemical. Traditionally, chemical toxicants are classified based on a number of features. These include their structures, physio-chemical properties, biological targets, hazards, and, at a higher level, endpoint outcomes. These classifications are too narrow or too broad. None reflects much of the biological mechanisms needed to understand and predict toxicity. Toxicity pathways give a new way of classifying chemicals, based on a sweeping scope of information of biological perturbations.

Although the 2007 NRC report provided a general definition of toxicity pathways, their nature and scope are still unclear. How many toxicity pathways are there, what are they, and how are they related to the traditionally defined biological pathways? Biological pathways are often invoked in the literature within the functional context of a particular field of interest. Examples include signal transduction, stress response, developmental, metabolic, and gene ontology pathways.<sup>35</sup> These pathways can be biochemically overlapping; they describe loosely a suite of physiological processes at the cellular and tissue levels. Toxicity pathways cut across these traditionally defined layers of pathways, pinning them together through a central theme: perturbation can lead to adverse outcomes. Depending on the concentrations and MIEs of a chemical, multiple traditionally defined biochemical pathways can be recruited and cross-activate or suppress one another, which can be collectively described as the toxicity pathway of the chemical in question.

### 8.6.2 Purpose of Modeling Toxicity Pathways

As a suite of cell-based *in vitro* assays is being developed to properly and informatively interrogate cellular responses anchored on toxicity pathways, computational systems biology models of these pathways must also be developed. The purpose of dose–response modeling of toxicity pathways is multi-faceted. First, cells are a highly nonlinear system as far as the perturbation–response relationship is concerned.<sup>36,37</sup> These non-linearities are crucial to the determination of the shapes of dose–response curves at cellular, tissue, and organism levels. Understanding them requires computational modeling of the toxicity pathways. The modeling is particularly important to help extrapolate cellular responses to low-level chemical exposures, adding weight to evidence for threshold responses. This cannot be determined by the statistics of assay data alone.<sup>38</sup> The second purpose of toxicity pathway modeling is to make predictions by linking dynamic exposure data or scenarios to cellular responses. *In vitro* cell assays are useful for determining POD concentrations of chemicals in the culture media, and IVIVE can relate these POD concentrations to idealized constant environmental exposure level.

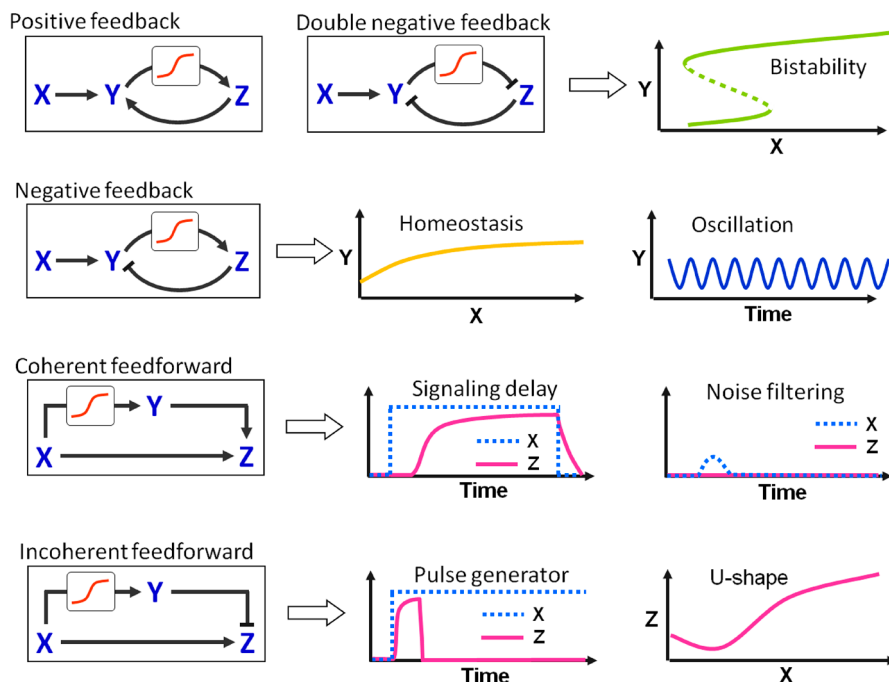
In reality, however, exposure levels often vary through time. We can conduct computer simulations to predict whether such fluctuating exposures can cause cellular adversity that leads to adverse health outcomes. We can do this through simulations of the computer models of the toxicity pathways calibrated with *in vitro* assay data. Lastly, the alternative cell-based toxicity testing approach does not directly address the issue of individual variability in MIE, toxicity pathways, and key events through AOP. Despite this, PBPK modeling has traditionally been used in a way that incorporates the kinetic variabilities in human population. Toxicity pathway modeling has the advantage of considering genetic and epigenetic variabilities to simulate quantitative differences in AOPs among individuals.

### 8.6.3 How to Model Toxicity Pathways

A modern integrated electronic circuit chip easily contains 1 billion transistors, which can be practically simulated in a circuit design computer program. Intracellular biochemical circuits contain a finite set of interacting molecular components, about 20 000 to 30 000 genes, their transcriptomic and proteomic products, and small-molecule metabolites. In theory, toxicity pathways, which are subsets of the entire intracellular biochemical networks, can be computed. Mapping these pathways is an ongoing scientific effort. With the ever-improving, modern high-throughput, and high-content omic analytic tools, we are moving closer to the day that all the molecular components, their interactions, component concentrations, and interacting strength are known.

These toxicity pathways will be modeled as dynamic systems, where RNAs, proteins, and metabolites are treated as dependent state variables, and their rates of change are governed by coupled differential equations. To come up with these equations, we need to draw on the types of biochemical interactions specific to the participating molecular components, which define the quantitative characteristics of the interactions and regulations. Elementary molecule–molecule interactions proceed on the principle of mass action (the rate of which is proportional to the concentrations of reactants). Certain combinations of these linear interactions can give us highly non-linear signaling properties. These small combined circuit structures are called network motifs.<sup>39</sup> They appear in biochemical networks repeatedly at frequencies higher than random combinations of reactions. Common network motifs include ultrasensitive response motifs, feedback loops, and feedforward loops (Figure 8.2). Logically connected network motifs constitute larger biochemical pathways or networks underpinning cellular-level functions, such as cell cycle progression, differentiation, stress response and homeostasis, and apoptosis.<sup>40–42</sup>

Signal amplification is an essential feature of signal propagation through biochemical networks. Amplification is required to compensate for attenuation of signals as they move along the pathways. It also provides properties needed for emergent, non-trivial behaviors. A suite of ultrasensitive response



**Figure 8.2** Network motifs, including positive feedback, negative feedback, coherent and incoherent feedforward, and their dynamical functions.

motifs, discovered over the past few decades, can amplify small percentage changes in input biochemical signals to larger percentage changes in output signals. These motifs include positive cooperative binding, homo-multimerization, multi-step signaling, zero-order covalent modification cycle, and molecular titration.<sup>36,43–45</sup> On a global input–output plot, ultrasensitive responses provided by these motifs usually appear as sigmoidal curves that can be approximated by a Hill equation. The Hill coefficient represents the degree of signal amplification. Multiple ultrasensitive response motifs can be further combined into more complex network motifs to get a higher degree of amplification. An example is the mitogen-activated protein kinase cascade.<sup>46</sup>

Although these ultrasensitive motifs can operate alone to amplify signals, they are usually embedded in feedback or feedforward network motifs to support more complex cellular behaviors. Positive feedback loops are the prototypical network structure necessary for irreversible switch-like behaviors. This is seen with cell differentiation in lineage specification and cell cycle progression through distinct phases. It is also seen in other cellular processes involving binary decision making based on whether the input signal surpass a threshold or not.<sup>42</sup> For instance, in B-lymphocyte terminal differentiation, a bistable circuit, comprising coupled transcriptional feedback loops involving multiple transcriptional repressors, operates to mediate the



irreversible transition of B-cells into antibody-secreting plasma cells.<sup>47</sup> Environmental immunotoxicant TCDD, acting *via* the aryl hydrocarbon receptor, can perturb this feedback circuit and suppress B-cell differentiation.

Negative feedback loops are a prototype network motif structure necessary for adaptation and homeostasis. It is a well-conserved motif operating in cellular stress responses to a variety of stressors.<sup>35,41</sup> These stress response pathways normally involve a sensor protein detecting cellular state changes, a transcription factor, and a suite of genes induced by the transcription factor to counteract the perturbed cellular state. For instance, in oxidative stress response, Keap1 is the sensor molecule detecting cellular redox state changes. The altered function of Keap1 as an E3 ligase adaptor protein then results in Nrf2 activation through protein stabilization. Nrf2 then translocates into the nucleus where it induces a number of antioxidant or related genes to bring the altered redox state back to normal.<sup>48</sup> Ultrasensitive motifs are invariably embedded in these negative feedback loops to sufficiently amplify signals to induce stress genes to levels high enough to normalize the perturbed cellular state. Negative feedback is an inherently nonlinear motif that can generate superlinearly shaped (concave downward) or threshold dose responses.<sup>48</sup> Post-translational feedback activation of stress proteins in the absence of transcriptional alteration of stress genes might be important for cells to adapt to low level and transient stresses.<sup>49</sup> When this post-translational adaptation is saturated and transcriptional control is invoked, cellular and even apical adversity will likely result.<sup>50</sup>

Feedforward loops are another network motif structure that can mediate cellular adaptation. This requires an “incoherent” feedforward loop, where the two arms of the loop regulate a common target molecule in opposite directions. In oxidative stress response, a number of chemical oxidants can directly modify the sensor molecule Keap1. The signaling path from a chemical oxidant to Keap1, Nrf2, antioxidant genes, and finally to reactive oxygen species (ROS) forms the negative feedforward arm of the feedforward loop.<sup>48</sup> The positive arm is from chemical oxidants to ROS, in which ROS production is increased by chemical oxidants. This incoherent forward loop can generate three different types of dose responses (superlinear, threshold, or hormetic) in the low-dose region, depending on the signaling strength of the feedforward arms.<sup>51</sup>

In the toxicology and risk assessment field, debates on whether biological systems have a threshold or not in response to external perturbations have been ongoing. Statistical arguments surrounding response data have been of little help. Such data can only support but not prove the existence of thresholds in the low-dose region where biological and measurement variability tend to dominate. Additional evidence supporting thresholds has to come from mechanistic studies where the underlying toxicity pathway being perturbed is expected to produce a threshold or threshold-like behaviors. Most of the common network motif structures that inherently produce threshold responses are summarized in a review article.<sup>38</sup> These motifs include, in addition to the feedback and feedforward structures discussed above, a

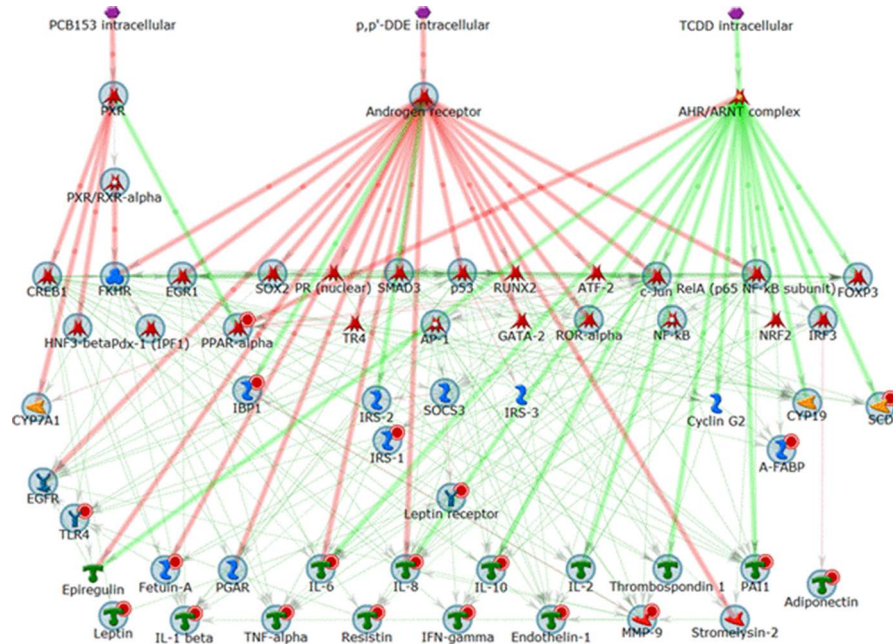
small number of bifurcation motifs (transcritical bifurcation and supercritical pitchfork bifurcations). Many biological examples are associated with these threshold motifs. Knowledge of the underlying network structure of toxicity pathways being perturbed will help better describe the shape of the dose–response curve in the low-dose region, reducing the uncertainty of whether or not there is a threshold.

### 8.6.4 Case Studies

The following case studies show the usefulness of toxicity pathway modeling in mechanistically interpreting and predicting threshold dose–response behaviors. In the first example, a minimal mathematical model of the DNA damage response by etoposide, an anticancer drug, was built to aid in developing *in vitro* assays for testing genotoxic chemicals.<sup>52</sup> In this model, a toxicity pathway based on experimental and published data was defined to include a positive feedback loop between ataxia telangiectasia mutated (ATM) and  $\gamma$ H2AX. ATM is a kinase sensing DNA double-strand breaks (DSB).  $\gamma$ H2AX is the phosphorylated form of DNA histone used as a biomarker for DSB formation. This positive feedback loop functions as a bistable switch that can be turned on or off through saddle-node bifurcations. When stochasticity is considered, where at low doses of etoposide the number of DSBs produced is small, this toxicity pathway model predicts a threshold response for  $\gamma$ H2AX at the break ends of DSBs and phosphorylated p53. The prediction provides a mechanistic interpretation for the similarly observed data in HT1080 cells and the graded responses above the threshold concentration of etoposide.

The second example focuses on the mitochondrial toxicity pathway. A peroxisome proliferator-activated receptor- $\gamma$  coactivator (PGC)-1 $\alpha$ -mediated coupled transcriptional negative feedback loop operates in cells to maintain cellular energetic and redox homeostasis in response to mitochondrial perturbation. PGC-1 $\alpha$  is a transcriptional coactivator. It can be activated by increasing ROS and decreasing ATP, enabling the transcriptional activity of a number of transcription factors. These transcription factors, including peroxisome proliferator-activated receptor (PPAR)- $\alpha$ , nuclear respiratory factor (NRF)1, and estrogen-related receptor, then upregulate a suite of genes responsible for increasing metabolic flux to the respiratory electron transfer chain, thus promoting mitochondrial biogenesis and enhancing antioxidant capacity. These transcriptional responses, acting in a negative feedback manner, counteract the redox and bioenergetics alterations, restoring mitochondrial homeostasis. Modeling such a coupled negative feedback toxicity pathway was shown to repeat the threshold toxicity behaviors observed in AC16 cardiomyocytes with doxorubicin, another cancer drug that has clinical adverse effects in the heart.<sup>53</sup> As with DNA damage response, without the support of modeling of the underlying toxicity pathways, the observed data will need to be interpreted through statistical analysis to suggest threshold responses. Computational modeling of these toxicity pathways provides further evidence from a mechanistic perspective for the shape of dose–response curves.

In the third example, systems biology pathway tools are used to map and prune the wiring diagram of toxicity pathways to study interaction relationships of individual chemicals and their mixtures in a system. Using commercially available MetaCore™/MetaDrug™ software, (Thomson Reuters; originally developed by Genego), we studied three well-known contaminants (2,3,7,8-tetrachlorodibenzodioxin [TCDD], 2,2',4,4',5,5'-hexachlorobiphenyl [PCB 153], and *p,p'*-dichlorodiphenyldichloroethylene [*p,p'*-DDE]).<sup>54</sup> Advanced search queries were conducted to identify underlying interactions for each contaminant and in combination within the database. This was followed by directional network construction to identify common mechanisms for the three chemicals within as many as two interaction steps downstream of their primary targets. This analysis brought together predictive chemical analyses based on compound structure and chemogenomics data, thus linking chemical, biological, and toxicological properties. The final network reveals converging pathways leading to activation of common targets downstream (Figure 8.3). This includes activating interactions (green arrows), inhibiting interactions (red arrows), and intermediate cross interactions that



**Figure 8.3** Network showing potential converging genes associated for three contaminants: 2,3,7,8-tetrachlorodibenzodioxin (TCDD), 2,2',4,4',5,5'-hexachlorobiphenyl (PCB 153), and *p,p'*-dichlorodiphenyldichloroethylene (*p,p'*-DDE) POPs. Symbols as defined by MetaCore™ at <http://lsresearch.thomsonreuters.com/static/uploads/files/2014-05/MetaCoreQuickReferenceGuide.pdf>. Reproduced from *Environmental Health Perspectives* with permission: <http://dx.doi.org/10.1289/ehp.1510308>.

could influence or activate each other, implying functional and mechanistic connectivity of the associated genes. These types of pathway analyses give support to the mechanistic hypothesis; they contribute to new interpretations linking published toxicology findings to various apical endpoints; and they are subject to dynamic modeling.

### 8.6.5 Education on Computational Toxicology

Biologists and toxicologists are not traditionally trained as quantitative scientists. The current demand for quantitative analysis and understanding of how biological systems work in response to perturbations is high, especially with the deluge of big omics data and the need to compute cells and organisms as dynamic systems. Many recent systems biology graduates are equipped with bioinformatic and machine learning knowledge and skills, which are applicable to classification and clustering for pattern recognition of big data. Even so, most professionals in the systems pharmacology and toxicology field still lack kinetic and dynamic analysis skills. Therefore, providing relevant training in academic and professional settings is necessary to promote the use of computational tools for pharmacokinetics and pharmacodynamics. PBPK courses have been offered nationwide and internationally for a couple of decades. They have been quite fruitful when considering the increasing number of scientists who have adopted this set of tools for risk assessment. In contrast, courses on dose–response modeling based on simulating toxicity pathways as dynamical systems are still emerging. Scientists from institutions such as the Hamner Institutes for Health Sciences have been working to expose scientists around the globe to the basic knowledge required to model cellular biochemical networks and toxicity pathways. The field is still in the early stage of exponential growth. More scientists will likely soon join in receiving quantitative training for pharmacology and toxicology applications.

### 8.6.6 Pathway Modeling Software Tools

Until recently, access to coupled ordinary differential equation (ODE)-based dynamic modeling has been limited to mathematicians, physicists, and engineers who have the mathematical and programming skills to implement a numerical solver and code a model. Traditional general-purpose simulation environments have included MatLab and Berkely Madonna. With the emergence of systems biology, simulation tools for biological pathway modeling have increased steadily. These tools, such as Simbiology, JDesigner, and Copasi, are often graphically interfaced. This greatly reduces the learning curve for biologists whose training is generally not quantitatively oriented. Many of the graphic tools allow users to build pathways by mouse clicking and dragging onscreen elements, essentially drawing a pathway picture. The coupled ODEs are automatically generated behind the scene. Systems biology markup language (SBML), a common modeling language, has been

developed to support model exchange between these different platforms.<sup>55</sup> Many of these tools are available through academia or commercially. (A full list of the tools supporting SBML is available at [www.sbml.org](http://www.sbml.org).)

In conclusion, toxicology started as a discipline seeking solutions and explanation of toxic effects. It has derived most of its knowledge from observational animal studies. It matured as a science, as experimental data were organized into databases and hypotheses were formulated for toxicity testing. Now, it is completely transforming into a predictive science that uses innovative methods and minimal use of animal toxicity testing. Several computational methods, including statistical and software programs, have and are being developed that can process large amounts of data and expedite this predictive transformation. However, several non-scientific considerations such as education, training, legality, and ethics play pivotal roles in acceptance of scientific and technological advances and this transformation will take some time before it is adopted by field practitioners of risk assessment.

## Acknowledgements

For editorial contributions the authors acknowledge Donald Meadows, writer-editor, Agency for Toxic Substances and Disease Registry, Atlanta, GA, USA.

## References

1. National Research Council, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, The National Academies Press, Washington, DC, USA, 2007.
2. Interagency Coordination Committee for Validation of Alternative Methods, *National Institutes of Environmental Health Sciences*, Research Technology Park, North Carolina, USA, 2010, <https://ntp.niehs.nih.gov/pubhealth/evalatm/iccvam/index.html>, last accessed April 30, 2016.
3. K. A. Houck and R. J. Kavlock, *Toxicol. Appl. Pharmacol.*, 2008, **227**, 163.
4. S. W. Edwards, Y. M. Tan, D. L. Villeneuve, M. E. Meek and C. A. McQueen, *J. Pharmacol. Exp. Ther.*, 2016, **356**, 170.
5. K. E. Tollefsen, S. Scholz, M. T. Cronin, S. W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth and G. Patlewicz, *Regul. Toxicol. Pharmacol.*, 2014, **70**, 629.
6. G. T. Ankley, R. S. Bennett, R. J. Erickson, D. J. Hoff, M. W. Hornung, R. D. Johnson, D. R. Mount, J. W. Nichols, C. L. Russom, P. K. Schmieder, J. A. Serrano, J. E. Tietge and D. L. Villeneuve, *Environ. Toxicol. Chem.*, 2010, **29**, 730.
7. D. L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. H. Hutchinson, C. A. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, M. A. Ottinger, L. Vergauwen and M. Whelan, *Toxicol. Sci.*, 2014, **142**, 312.
8. K. S. Crump, *Fundam. Appl. Toxicol.*, 1984, **4**, 854–871.
9. L. G. Soeteman-Hernandez, M. D. Fellows, G. E. Johnson and W. Slob, *Toxicol. Sci.*, 2015, **148**, 355.

10. M. Mumtaz, *Principles and Practice of Mixtures Toxicology*, Wiley-VCH, Weinheim, Germany, 2010.
11. H. A. El-Masri, M. M. Mumtaz and M. L. Yushak, *Environ. Toxicol. Pharmacol.*, 2004, **16**, 57.
12. R. S. H. Yang and Y. Lu, in *Risk Assessment for Environmental Health*, ed. M. G. Robson and W. A. Toscano, John Wiley & Sons, Hoboken, N.J., 2007, p. 85.
13. R. S. H. Yang and M. E. Andersen, in *Physiologically Based Pharmacokinetics: Science and Applications*, ed. M. B. Reddy, R. S. H. Yang, H. J. Clewell III and M. E. Andersen, John Wiley and Sons, Inc., New York, NY, 2004, pp. 349–373.
14. M. E. Andersen, *Toxicol. Lett.*, 1995, **79**, 35.
15. Y. M. Tan, H. Clewell, J. Campbell and M. Andersen, *Int. J. Environ. Res. Public Health*, 2011, **8**, 1613.
16. P. Ruiz, B. A. Fowler, J. D. Osterloh, J. Fisher and M. Mumtaz, *SAR QSAR Environ. Res.*, 2010, **21**, 603.
17. H. J. Clewell 3rd, P. R. Gentry, J. M. Gearhart, T. R. Covington, M. I. Banton and M. E. Andersen, *Toxicol. Sci.*, 2001, **63**, 160.
18. M. Mumtaz, J. Fisher, B. Blount and P. Ruiz, *J. Toxicol.*, 2012, **2012**, 904603.
19. Y. M. Tan, K. H. Liao and H. J. Clewell 3rd, *J. Exposure Sci. Environ. Epidemiol.*, 2007, **17**, 591.
20. H. J. Clewell, Y. M. Tan, J. L. Campbell and M. E. Andersen, *Toxicol. Appl. Pharmacol.*, 2008, **231**, 122.
21. L. E. Redding, M. D. Sohn, T. E. McKone, J. W. Chen, S. L. Wang, D. P. Hsieh and R. S. Yang, *Environ. Health Perspect.*, 2008, **116**, 1629.
22. T. E. McKone, *J. Exposure Anal. Environ. Epidemiol.*, 1993, **3**, 339.
23. B. J. Blaauboer, *Toxicol. Lett.*, 2003, **138**, 161.
24. B. J. Blaauboer, *J. Toxicol. Environ. Health, Part B*, 2010, **13**, 242.
25. M. M. Mumtaz, M. Ray, S. R. Crowell, D. Keys, J. Fisher and P. Ruiz, *J. Toxicol. Environ. Health, Part A*, 2012, **75**, 6.
26. L. L. Aylward, S. M. Hays, M. Gagne, A. Nong and K. Krishnan, *Regul. Toxicol. Pharmacol.*, 2010, **58**, 25.
27. C. R. Kirman, L. L. Aylward, S. M. Hays, K. Krishnan and A. Nong, *Regul. Toxicol. Pharmacol.*, 2011, **60**, 172.
28. S. M. Hays, L. L. Aylward, J. S. LaKind, M. J. Bartels, H. A. Barton, P. J. Boogaard, C. Brunk, S. DiZio, M. Dourson, D. A. Goldstein, J. Lipscomb, M. E. Kilpatrick, D. Krewski, K. Krishnan, M. Nordberg, M. Okino, Y. M. Tan, C. Viau and J. W. Yager, *Regul. Toxicol. Pharmacol.*, 2008, **51**, S4.
29. S. M. Hays, R. A. Becker, H. W. Leung, L. L. Aylward and D. W. Pyatt, *Regul. Toxicol. Pharmacol.*, 2007, **47**, 96.
30. P. Ruiz, M. Mumtaz and V. Gombar, *Toxicol. Appl. Pharmacol.*, 2011, **254**, 198.
31. J. P. Richard, C. K. Heo and M. M. Toteva, *J. Phys. Org. Chem.*, 2008, **21**, 531.
32. P. Ruiz, K. Ingale J. Wheeler and M. Mumtaz, *Chemosphere*, 2016, **144**, 2238.



33. N. Greene, *Adv. Drug Delivery Rev.*, 2002, **54**, 417.
34. R. Benigni, C. Bossa, N. Jeliaskova, T. Netzeva and A. Worth, *The Benigni/Bossa Rulebase for Mutagenicity and Carcinogenicity—a Module of Toxtree. European Commission report EUR 23241 EN*, 2008.
35. S. O. Simmons, C. Y. Fan and R. Ramabhadran, *Toxicol. Sci.*, 2009, **111**, 202.
36. Q. Zhang, S. Bhattacharya and M. E. Andersen, *Open Biol.*, 2013, **3**, 130031.
37. J. J. Tyson, K. C. Chen and B. Novak, *Curr. Opin. Cell Biol.*, 2003, **15**, 221.
38. Q. Zhang, S. Bhattacharya, R. B. Conolly, H. J. Clewell 3rd, N. E. Kaminski and M. E. Andersen, *Environ. Health Perspect.*, 2014, **122**, 1261.
39. U. Alon, *Nat. Rev. Genet.*, 2007, **8**, 450.
40. B. Novak and J. J. Tyson, *Nat. Rev. Mol. Cell Biol.*, 2008, **9**, 981.
41. Q. Zhang and M. E. Andersen, *PLoS Comput. Biol.*, 2007, **3**, e24.
42. J. E. Ferrell and W. Xiong, *Chaos*, 2001, **11**, 227.
43. J. E. Ferrell Jr and S. H. Ha, *Trends Biochem. Sci.*, 2014, **39**, 496.
44. J. E. Ferrell Jr and S. H. Ha, *Trends Biochem. Sci.*, 2014, **39**, 556.
45. J. E. Ferrell Jr and S. H. Ha, *Trends Biochem. Sci.*, 2014, **39**, 612.
46. C. Y. Huang and J. E. Ferrell Jr, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 10078.
47. S. Bhattacharya, R. B. Conolly, N. E. Kaminski, R. S. Thomas, M. E. Andersen and Q. Zhang, *Toxicol. Sci.*, 2010, **115**, 51.
48. Q. Zhang, J. Pi, C. G. Woods and M. E. Andersen, *Toxicol. Appl. Pharmacol.*, 2010, **244**, 84.
49. Q. Zhang, S. Bhattacharya, J. Pi, R. A. Clewell, P. L. Carmichael and M. E. Andersen, *Toxicol. Sci.*, 2015, **147**, 302.
50. R. S. Thomas, S. C. Wesselkamper, N. C. Wang, Q. J. Zhao, D. D. Petersen, J. C. Lambert, I. Cote, L. Yang, E. Healy, M. B. Black, H. J. Clewell 3rd, B. C. Allen and M. E. Andersen, *Toxicol. Sci.*, 2013, **134**, 180.
51. Q. Zhang, J. Pi, C. G. Woods and M. E. Andersen, *Toxicol. Appl. Pharmacol.*, 2009, **237**, 345.
52. Z. Li, B. Sun, R. A. Clewell, Y. Adeleye, M. E. Andersen and Q. Zhang, *Toxicol. Sci.*, 2014, **137**, 371.
53. H. Yuan, Q. Zhang, J. Guo, T. Zhang, J. Zhao, J. Li, A. White, P. L. Carmichael, C. Westmoreland and S. Peng, *Toxicol. Sci.*, 2016, **150**, 400.
54. P. Ruiz, A. Perlina, M. Mumtaz and B. A. Fowler, *Environ. Health Perspect.*, 2016, **124**, 1034.
55. M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang and S. Forum, *Bioinformatics*, 2003, **19**, 524.



## CHAPTER 9

# ***In silico Toxicology: An Overview of Toxicity Databases, Prediction Methodologies, and Expert Review***

D. BOWER<sup>a</sup>, K. P. CROSS<sup>a</sup>, S. ESCHER<sup>b</sup>, G. J. MYATT<sup>\*a</sup> AND D. P. QUIGLEY<sup>a</sup>

<sup>a</sup>Leadscope, 1393 Dublin Road, Columbus, OH 43215, USA; <sup>b</sup>Fraunhofer Institute of Toxicology and Experimental Medicine (ITEM), Chemical Risk Assessment, Nikolai-Fuchs-Strasse 1, 30625, Hannover, Germany

\*E-mail: [gmyatt@leadscope.com](mailto:gmyatt@leadscope.com)

## **9.1 Introduction**

Understanding chemical toxicity is a necessary part of the research and development (R&D) and regulatory approval process across many industries (*e.g.* pharmaceuticals, cosmetics, and pesticides). Toxicologists have an increasingly rich set of *in vivo* and *in vitro* methods with which to assess hazard and risk, which are being progressively supplemented with newer *in silico* approaches. There are some general advantages, disadvantages and issues using these different approaches, which are summarized in Table 9.1.

**Table 9.1** General assessment of current *in vivo*, *in vitro* and *in silico* approaches.<sup>a</sup>

	<i>In vivo</i>	<i>In vitro</i>	<i>In silico</i>
Coverage of toxicity endpoint	Extensive coverage	Limited coverage	Limited coverage
Time to generate results	Slow (months-to-years)	Faster (weeks-to-months)	Fast (seconds)
Cost to generate results	Expensive	Less expensive	Minimal cost (once model is acquired or built)
Need for compound	Large samples of the compound is needed to perform the test	Smaller sample sizes of the compound are needed to perform the test	No material requirements, only the chemical structure digital record
Understanding of mechanisms	Calls are based on findings, <i>e.g.</i> gross and histopathological findings, weight changes, clinical chemistry, urine analysis and hematology Understanding of biological mechanisms is not mandatory and often limited	Some assays can probe biological mechanisms	Some <i>in silico</i> approaches will suggest a plausible biological mechanism; however, they can only be as good as the underlying data. If the data (see <i>in vivo</i> or <i>in vitro</i> ) do not provide a good understanding of mechanisms, then the <i>in silico</i> method cannot be comprehensive
Understanding of structural basis	None	None	Will often identify the portion of the chemical responsible for the positive (or negative) prediction

Available guidelines to interpret the results	Thorough set of guidelines	Limited to those assays currently accepted by regulatory authorities	Practically none; however, under REACH there is a QSAR reporting format (QMRF) <sup>1</sup>
Accuracy of results	Inter- and intraspecies differences: there is a need to assess the human relevance of the findings Adversity: has to be assessed, <i>e.g.</i> reversible effects have to be evaluated for their relevance	There is a need to understand the limitation (coverage, reliability and accuracy) of the assay system as well as determine the human relevance	Able to predict <i>in vivo</i> and <i>in vitro</i> results, but accuracy is dependent on the training set and modeling methodology. They generally do not take account of dose or concentration
Chemical coverage	Applicable to all technically testable compounds	Applicable to all technically testable compounds	Only applicable to those in the applicability domain of the model
Qualitative assessment (hazard)	Yes	Yes	Yes
Quantitative assessment (NOAEL, point of departure)	Yes	Possible if <i>in vitro</i> to <i>in vivo</i> extrapolation is applicable	Limited

<sup>a</sup>NOAEL: no observed adverse effect level; REACH: Registration, Evaluation, Authorisation and Restriction of Chemicals; QSAR: quantitative structure–activity relationship; QMRF: QSAR model reporting format.

Today, there is a wide array of *in vivo* models to assess chemically induced toxic effects, including acute and repeated dose toxicity, reproductive and developmental toxicity, carcinogenicity, and skin and eye sensitivity and irritation. The principle advantage of these *in vivo* testing models is that they expose a living animal to the test chemical and in doing so account for the complex interplay of toxicodynamics and toxicokinetics across the organism. Above all, they give quantitative results that can be compared between studies or with other compounds. The wide acceptance of these tests results from many years of refinement along with the harmonization of their procedures and interpretation through guidelines developed by international organizations (e.g. the Organisation for Economic Co-operation and Development (OECD)<sup>2</sup> or the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)).<sup>3</sup> Safety factors are used to extrapolate the animal outcome to the human situation. These factors account for interspecies differences, e.g. differences in enzyme and receptor expression, as well as a higher human susceptibility, e.g. elderly or very young patient subpopulations. Although the concordance of animal toxicity to humans is only considered to be ~70%,<sup>4</sup> using *in vivo* toxicity testing models remains the gold-standard testing strategy.<sup>5</sup>

The use of these models to assess the hazard and risk of chemicals is well understood and can be classified either by their exposure duration or specific endpoint. Acute-exposure *in vivo* tests are used to estimate an appropriate dose range for further toxicity testing (acute toxicity range-finding tests), to establish the median lethal dose (LD<sub>50</sub>) or median lethal concentration (LC<sub>50</sub>). Subacute studies, e.g. 28 day *in vivo* studies, are in use to identify primary target organs and select dosing for longer term studies. The longer term, subchronic repeated-dose *in vivo* models are used to investigate target-organ toxicity, determine the bioaccumulation potential and establish no observed adverse effect levels (NOAELs). They may influence the dose range selection for subsequent long-term testing. Chronic and carcinogenicity *in vivo* tests are long-term investigations into the cumulative effect of exposure to the test chemical for at least 12 months across multiple doses. The endpoint-specific *in vivo* models include reproductive-developmental toxicity, neurotoxicity, and immunotoxicity.

There are numerous limitations associated with the *in vivo* toxicity testing approach. The amount of time (typically, months to years) and expense (often hundreds of thousands of dollars) to generate and interpret *in vivo* results are impediments to their use, in particular when evaluating a large set of chemicals. For example, it may be impractical to employ *in vivo* testing as a toxicity screen during the R&D discovery phase where many candidates (potentially tens of thousands of chemicals) are being considered. Another limitation of the *in vivo* strategy is the evaluation of potential or theoretical impurities or degradants in pharmaceuticals and other products, which may be difficult to synthesize in sufficient quantity and quality for testing. There are also some regulatory restrictions to using animal models, such as the Cosmetics Directive in the European Union.<sup>6</sup>

*In vitro* methods have been developed for certain endpoints. For genetic toxicity, many relevant *in vitro* assays have been developed and are being widely used to prioritize and assess the genetic toxicity of chemicals. These include the bacterial reverse mutation assay (often called the Ames assay),<sup>7</sup> *in vitro* mammalian chromosomal aberration test,<sup>8</sup> and *in vitro* mammalian cell micronucleus test.<sup>9</sup> A number of *in vitro* assays in other areas have also been developed, such as *in vitro* dermal absorption methods,<sup>10</sup> *in vitro* skin corrosion (human skin model test),<sup>11</sup> *in vitro* endocrine disruptor activity,<sup>12</sup> and so on. Many of these methods have been validated and standardized through international bodies and the generation and interpretation of the results has been documented as part of these guidelines (e.g. the European Union Reference Laboratory for Alternatives to Animal Testing,<sup>13</sup> Interagency Coordinating Committee on the Validation of Alternative Methods,<sup>14</sup> the OECD, and ICH). The time and cost to generate *in vitro* results is significantly less than *in vivo* tests. Despite an increasing number of tests that can maintain tissues or cells in culture for up to 14 days or longer, *in vitro* methods or combinations thereof are not yet available to assess complex *in vivo* endpoints such as local or systemic toxicity after repeated low exposure or developmental and reproductive toxicity. In an attempt to rectify this issue, there are currently a number of major initiatives to develop new *in vitro* methods to supplement and/or replace traditional *in vivo* models.<sup>15</sup> However, the inter-relationship between organs is not yet present and systems such as organ-on-a-chip with the use of microfluidics may strengthen the case for *in vitro* models. Most of these model systems are still under development or have limited validation and have no appropriate guidelines. Individually, they do not provide a complete replacement for *in vivo* tests, as they often only probe a specific biological event or mechanism, and therefore a battery of such tests is usually necessary. In the case of genetic toxicity, *i.e.* for which an understanding of the mode of action based on reactivity is available, the *in vivo* tests can be replaced by evidence from *in vitro* indicator tests such as the comet assay, the Ames test or micronucleus tests. They provide a helpful way of prioritizing *in vivo* follow-up tests as well as helping to understand the biological mechanism and relevance of any *in vivo* findings. The development of assessment frameworks/testing batteries for more complex endpoints such as systemic toxicology after repeated exposure that are relevant to the low doses humans are generally exposed to is still an active field of research.

There is an increasing number of *in silico* models for toxicological endpoints. Some of these models offer significant benefits over existing *in vivo* or *in vitro* models. First, they require no material and will make a prediction from the chemical structure alone. Once the models are built, they are usually fast and cheap to run. This supports their use in screening large volumes of chemicals. These methodologies often provide an indication of the structural basis (*i.e.* highlighting a portion of the molecule) for any positive (or negative) predictions that *in vivo* or *in vitro* models do not provide. This is particularly important in supporting the redesign of candidate chemicals to help avoid the projected toxicity (while retaining other desirable properties).

The structural basis of a toxicity prediction can be an important component in the weight-of-evidence for the projected toxicity. For example, an unexpected positive *in vivo* or *in vitro* result without any concurring structural alerts may possibly indicate an issue with how the test was performed, such as the existence of an impurity in the test material or an interaction between the test material and the solvent, such as DMSO and acid halides in the Ames test.<sup>16</sup> These *in silico* models can also encode knowledge related to mechanism or even deficiencies in the assay system to help avoid false positives that results from certain experimental conditions or artifacts of the assay.

There are a number of disadvantages with today's *in silico* methods and models. A specific model can only reliably predict an endpoint for a chemical in a known area of chemistry (*i.e.* only for chemicals within the applicability domain of the model). There are currently few models that provide an indication of dose or concentration or potency. Another major disadvantage of the *in silico* approach is the lack of internationally agreed guidelines for the use and interpretation of *in silico* results. Therefore many models appear to regulatory toxicologists as black-box approaches, as often the data quality of the underlying assays, the relevance of the assays with regard to the *in vivo* endpoint, and the uncertainty of the prediction is not documented. Furthermore, many toxicologists do not fully understand the validation approaches (such as cross validation) and cannot interpret the performance statistics provided by models that define its sensitivity or specificity. Moreover, even if the sensitivity and specificity of the model is provided, information on the reliability of the underlying experimental data for the endpoint is missing. Such insights would inform the risk assessors about how reliable the prediction is compared to the experimental value alone. This lack of documentation limits the widespread use of *in silico* approaches among practising toxicologists by inhibiting their ability to interpret the results in an accepted, consistent, and defensible manner. A few initiatives have started to address this issue, *e.g.* the database on quantitative structure–activity relationship (QSAR) models and QSAR model reporting formats developed by the European Commission's Joint Research Centre.<sup>17</sup> In addition, each model is built from different training sets and modeling techniques, leading to a wide variance in the predictivity of models.

The most important component in the development of *in silico* methods are high quality up-to-date toxicity databases. These databases can identify specific experimental data from an adequately performed study. Toxicological databases are often used to generate read-across predictions (see Section 9.3.4). Read-across identifies “similar” chemicals or analogs (often using a mechanistic-based category). Adequate toxicity data from these analogs are used to predict the qualitative toxicity of an untested target compound (*e.g.* type of effect or hazard) or quantitative toxicity (*e.g.* dose level or point of departure). Analysis of toxicity databases supports computational models such as expert alerts (often referred to as structural alerts) or QSAR models. The breadth and quality of the data is the most important factor influencing the predictivity of these models.

Rule-based expert alerts (described in Section 9.3.2) generate a prediction based upon the presence or absence of structural rules (usually encoded as one or more substructure searches) that flag chemicals for different types of toxicity. The predictions are often accompanied by an explanation of the mechanistic basis associated with the matching alert(s). QSAR models (described in Section 9.3.3) are constructed from experimental laboratory results (referred to as a training set) where molecular descriptors are calculated from the chemical structures in the training set and used in computational models to predict the target toxicological effect.

All the systems in use—*in vivo*, *in vitro*, and *in silico*—are predictive and each prediction should be validated. As with any other study, *in silico* results should be critically assessed and thoroughly documented. This expert review may include an assessment of any available and appropriate data from the literature, an assessment of the combined results from potentially more than one *in silico* methodology, expert reviews to refute (or accept) the results from any *in silico* analysis (including inconclusive predictions or out-of-domain results), and how to proceed after assessing the results (*e.g.* additional testing or controlling the exposure of the product).

This chapter covers approaches to organizing toxicology databases. Specific databases covering endpoints related to genetic toxicity, carcinogenicity, and reproductive and development toxicity, as well as acute and repeated dose toxicity are reviewed. The major *in silico* prediction methods and systems are outlined. The chapter discusses how to combine and document the results from these methodologies as part of an expert review. The chapter concludes with a discussion on key issues and future directions.

## 9.2 Toxicity Databases

### 9.2.1 Overview

Toxicology data are generated across many organizations, with the results being reported in the literature or in an online database. In some cases, the full study results may be reported, covering the identity of the tested chemical, the experimental protocol, the study findings (such as details on individual animals treated with histopathology, clinical chemistry, and hematology details recorded), and an overall outcome. In other cases, only summary information on high-level endpoint calls [such as NOAEL or LOAEL (lowest observed adverse effect level) values or positive/negative/equivocal calls] may be included. The format and level of detail that is collected across the public literature and online databases are highly variable. The experimental protocols, result findings, and description of the tested chemical are also reported using different non-standardized terms. These variables make it difficult to compare different studies.

To support the development and use of new *in silico* methods, it is highly desirable to organize the chemical and toxicity data in a harmonized manner. A chemical registration and database system alongside a toxicity database



where the content is organized in a consistent manner according to well-defined ontologies is necessary to support searching across chemicals for study types and experimental results. It is also essential for *in silico* model building, building expert alerts systems, and read-across.

The following sections outline several ways to organize toxicity databases in a harmonized manner and summarizes a number of databases covering different toxicity endpoints.

### 9.2.2 Database Organization

ToxML (<http://toxml.org>) supports the exchange, integration, and organization of toxicological study data.<sup>18</sup> It is an open standard that uses the Extensible Markup Language (XML)<sup>19</sup> format plus a customized controlled vocabulary. The use of XML supports the inherently hierarchical nature of toxicological data in a consistent manner. Currently, ToxML is able to handle many toxicology studies including genetic toxicity studies (bacterial mutagenesis, *in vitro* and *in vivo* chromosome aberration, micronucleus, and mammalian mutagenesis), single-(acute) and repeat-dose studies (carcinogenicity, chronic, and subchronic), reproductive and developmental studies, skin irritation, skin penetration, and skin sensitization studies. At the top level of the ToxML hierarchy is the chemical structure, from which all associated toxicology study data are recorded. Each study record is organized by study, test, and treatment level information. The ToxML website contains a series of tools to support the use and extension of the standard, including a specification editor that permits users to make suggestions and contribute to the development of ToxML.

As part of the implementation of the Clinical Data Interchange Standards Consortium Study Data Tabulation Model (SDTM),<sup>20,21</sup> the Standard for Exchange of Nonclinical Data (SEND) was developed. SEND was set up to support the electronic submission of these data to regulatory agencies. The SDTM describes how study results can be organized into three general observation classes: interventions (investigational, therapeutic, and other treatments administered), events (planned protocol milestones), and findings (the observations resulting from tests, usually at specific time points). SDTM uses the SAS version 5 transport format; however, there is currently an initiative to replace this format with XML. SEND currently covers toxicological data endpoints for single-dose toxicity, repeat-dose toxicity, and carcinogenicity studies using a controlled vocabulary to facilitate the exchange of data. Representation for additional endpoints is ongoing.

To support the transmission of data on chemicals to European regulators (such as the European Chemicals Agency), the OECD has developed a set of harmonized templates. These templates have been developed using XML and reflect properties or effects on human health and the environment.<sup>2</sup> They have been developed as part the International Uniform Chemical Information Database (IUCLID) software and database system.<sup>22</sup> This system is

used to prepare and submit data for regulatory purposes. This system is also used to gather information within the OECD High Production Volume (HPV) chemicals program.

The standardization of vocabulary is essential to assess and analyze data, especially for representing histopathological findings. The INHAND Project (International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice), a joint initiative of the societies of toxicologic pathology of Europe, Great Britain, Japan, and North America, is developing an internationally accepted nomenclature for proliferative and non-proliferative lesions in laboratory animals like rats and mice.<sup>23</sup> INHAND provides a standardized nomenclature and differential diagnosis for classifying mainly histopathological lesions observed in different organ systems such as the hepatobiliary system, glands (mammary, Zymbal's, preputial, and clitoral), male and female reproductive system, central and peripheral nervous system, urinary system, and respiratory tract.<sup>24–28</sup> The standardized nomenclature presented is also available for society members on the internet.<sup>29</sup>

In addition, the National Institutes of Health has recently made available an atlas of non-neoplastic lesions.<sup>30</sup>

Ontologies were developed to use these standardized terms, *e.g.* for data mining and knowledge generation. The main advantage of an ontology compared to a code list is the presence of a hierarchical structure, *e.g.* tree structure linking terms. When terms are coming from repeated-dose *in vivo* studies, this linkage can be performed according to the pathological process. Of note here is the Hpath ontology from the eTOX project, which established an ontology of >31 000 terms describing histopathological alterations and >17 000 terms describing the affected target organs in nine species. This ontology and further open source ontologies can be found at the OBO Foundry ([www.obofoundry.org/](http://www.obofoundry.org/)).

### 9.2.3 Genetic Toxicity and Carcinogenicity

The following databases include information on genetic toxicity and carcinogenicity.

- *Carcinogenic Potency Database (CPDB)* Curated genetic toxicity and rodent carcinogenicity data covering the period 1980–2011.<sup>31</sup>
- *Chemical Carcinogens: structures and experimental data from the Chemical Carcinogenesis Research Information System (CCRIS)* Literature database covering the period 1985–2011.<sup>32</sup>
- *Chemical Carcinogens: structures and experimental data from Istituto Superiore di Sanita (ISSCAN)* The ISSCAN database contains information on approximately 890 chemical compounds tested using the carcinogenicity bioassay on rodents (rat, mouse) and Ames tests (*in vitro*, *Salmonella* mutagenicity). Structures and experimental data are provided by the *Istituto Superiore di Sanita*.<sup>33</sup>

- *eChemPortal* Data provider developed by the OECD. Data can be searched by compound/property or the Globally Harmonized System of Classification and Labelling of Chemicals. Properties include physico-chemical parameters, environmental fate and pathways, as well as toxicological and ecotoxicological endpoints. Different sources are searched, e.g. CCR, J-Check, ECHA CHEM and OECD SIDS IUCLID. Each query result table is restricted to 10 000 entries. Query details can be saved and reloaded.<sup>34</sup>
- *European Chemicals Agency (ECHA) database for REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals)* The data comes from registration dossiers submitted by companies to ECHA. REACH data requirements are oriented by production tonnage band. Data can be reviewed on a case-by-case basis. A batch extraction of data per compound or per endpoint is possible *via* eChemPortal.<sup>35</sup>
- *Gene-tox* Mutagenicity test data from the United States Environmental Protection Agency (EPA).<sup>36</sup>
- *International Agency for Research on Cancer and carcinogenicity classification (IARC)*: IARC monographs including carcinogenicity classification.<sup>37</sup>
- *International Programme on Chemical Safety (IPCS INCHEM)* Open access search capabilities over a variety of summary documents.<sup>38</sup>
- *Integrated Risk Information System (IRIS)* Data from the US EPA in support of human health risk assessment, focusing on hazard identification and dose-response assessment.<sup>39</sup>
- *Japan Existing Chemical Data Base (JECDB)* Open access database containing high production volume chemicals.<sup>40</sup>
- *Leadscope Drugs Genetox Database* Genetic toxicity database from the US Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) listing product approval reviews.<sup>41</sup>
- *Leadscope Food Safety Genetox Database* Genetic toxicity database from the US FDA Center for Food Safety and Applied Nutrition (CFSAN) reviews.<sup>41</sup>
- *National Toxicology Program (NTP) database* Access to the NTP's genetic toxicity and carcinogenicity results.<sup>42,43</sup>
- *PharmaPendium* Toxicity data from FDA and European Medicines Agency approval documents.<sup>44</sup>
- *Registry of Toxic Effects of Chemical Substances (RTECS)* A collection of basic toxicity information, including prescription and non-prescription drugs, food additives, pesticides, fungicides, herbicides, solvents, diluents, chemical wastes, reaction products of chemical waste, and substances used in industrial and household situations. The RTECS database contains 46 385 RTECS mutation studies for 13 343 chemicals and 10 517 tumorigenic studies for 3724 chemicals.<sup>45</sup>
- *Vitic from Lhasa Limited* Commercial data from published and unpublished sources.<sup>46</sup>

Searching these databases individually is time-consuming and prone to error. A number of efforts have been established to consolidate and harmonize the data across many different databases to allow users to search several sources simultaneously. Harmonization also supports the generation of training and reference sets for use in building *in silico* models. Some examples are:

- *The Leadscope SAR Genetox Database*, which currently includes 11 028 compounds with 179 732 test results from FDA CDER, FDA CFSAN, NTP, CCRIS, and other primary data sources including on-going data harvesting from the literature.<sup>41</sup>
- *The Leadscope SAR Carcinogenicity Database* includes 3598 compounds with 11 538 test results from FDA CDER, FDA CFSAN, NTP, CCRIS, and other primary data sources. Available endpoints with populations include male rat (1774), female rat (1725), male mouse (1640), and female mouse (1675).<sup>41</sup>

### 9.2.4 Reproductive and Developmental Toxicity

The following databases include information on reproductive and developmental studies.

- *DART Developmental and Reproductive Toxicology Database* and environmental teratology information.<sup>47</sup>
- *National Center for Toxicological Research Endocrine Disruptor Knowledge Base (EDKB) Database* for predicting estrogen and androgen activity.<sup>48</sup>
- *ECHA REACH database* European Chemicals Agency database for REACH.<sup>35</sup>
- *FedTex* Non-commercial database on mainly one- to two- generation studies in rodents developed by Fraunhofer ITEM. Currently ~530 studies with oral application for 270 compounds are included. Studies are extracted from peer-reviewed publications. Examined targets and organs per generation are observed and the effects and dose levels are documented. Access is granted upon request to Fraunhofer ITEM.<sup>49</sup>
- *ILSI International Life Sciences Institute developmental toxicity data*.<sup>50</sup>
- *LactMed* Possible adverse effects in the nursing infant.<sup>51</sup>
- *Leadscope Drugs Repro-Developmental Database* Contains reproductive and developmental toxicity studies from US FDA CDER product approval reviews.<sup>41</sup>
- *Leadscope Food Safety Repro-Developmental Database* Contains reproductive and developmental studies from US FDA CFSAN reviews.<sup>41</sup>
- *Mario Negri* Mario Negri Institute for Pharmacological Research developmental and reproductive toxicity database.<sup>52</sup>
- *NTP* Access to the NTP's reproductive and developmental results.<sup>42</sup>
- *Reprotox* Pregnancy, reproduction, and development database.<sup>53</sup>

- *Shepard's Catalog of Teratogenic Agents* Information on teratogenic agents.<sup>54</sup>
- *TERIS (Teratogen Information System)* Data on teratogenic exposures in pregnant women.<sup>55</sup>
- *RTECS* The RTECS database covers reproductive effects and contains 25 558 reproductive studies for 3724 chemicals along with high-level endpoint data TCLO and TDLo values for select species and routes of administration.<sup>45</sup>
- *ToxRefDB* Toxicity Reference Database containing *in vivo* animal toxicity studies on hundreds of chemicals, incorporating an ontology as part of the data entry. The database primarily includes oral repeated dose toxicity studies in rodents and developmental and reproductive toxicology studies, with the last release in 2014.<sup>56</sup>

### 9.2.5 Acute and Repeated Dose Toxicity

The following databases include information on acute and repeated dose toxicity studies:

- *COSMOS* Database of repeated-dose toxicological information.<sup>57</sup>
- *ECHA REACH database* European Chemicals Agency database for REACH.<sup>35</sup>
- *Hazard Evaluation Support System (HESS)* The toxicological database contains information on repeated-dose toxicity and toxicity mechanisms. A separate metabolism knowledge database contains data on rat metabolism maps and information on absorption, distribution, metabolism and excretion in rats and humans.<sup>58,59</sup>
- *Leadscope Drugs Chronic/Sub-chronic Database* Chronic and subchronic toxicity database of US FDA CDER product approval reviews.<sup>41</sup>
- *Leadscope Food Safety Acute Toxicity Database* Acute toxicity data from the US FDA CFSAN Priority-based Assessment of Food Additives database.<sup>41</sup>
- *Leadscope Food Safety Chronic/Subchronic Database* Chronic/subchronic toxicity database of US FDA CFSAN reviews.<sup>41</sup>
- *Mario Negri*: Mario Negri Institute for Pharmacological Research repeated-dose toxicity databases.<sup>52</sup>
- *National Toxicology Program* Access to the NTP's subchronic and chronic studies.<sup>42</sup>
- *RepDose (from the Fraunhofer Institute of Toxicology and Experimental Medicine repeated-dose toxicity database)* Repeated-dose toxicity database from subacute to chronic exposure with oral, inhalation, and dermal exposure. An ontology is used for data entry. The database includes rat, mouse, and dog species information. There are currently ~3000 studies for 800 compounds and one-third of the database is online.<sup>60–65</sup>
- *RTECS* Registry of Toxic Effects of Chemical Substances, including 279 237 acute studies for 154 702 chemicals and 52 730 multiple-dose toxicity studies for 13 953 chemicals.<sup>45</sup>

- *Vitic/EtoxSys* Continuously growing commercial database on *in vivo* animal studies developed in the Innovative Medicines Initiative eTOX project. Repeated-dose toxicity data are provided by 13 pharmaceutical companies. The current eTOX database (2015.02) has about 1700 compounds and 6000 studies and is available to users either from their internal version or from a secured connection at a data broker (Lhasa Limited). Currently, the majority of studies are short term and subchronic. The ontology Hpath (see Section 9.2.2) is used to standardize and link data.<sup>46</sup>

## 9.3 *In silico* Methodologies

### 9.3.1 Overview

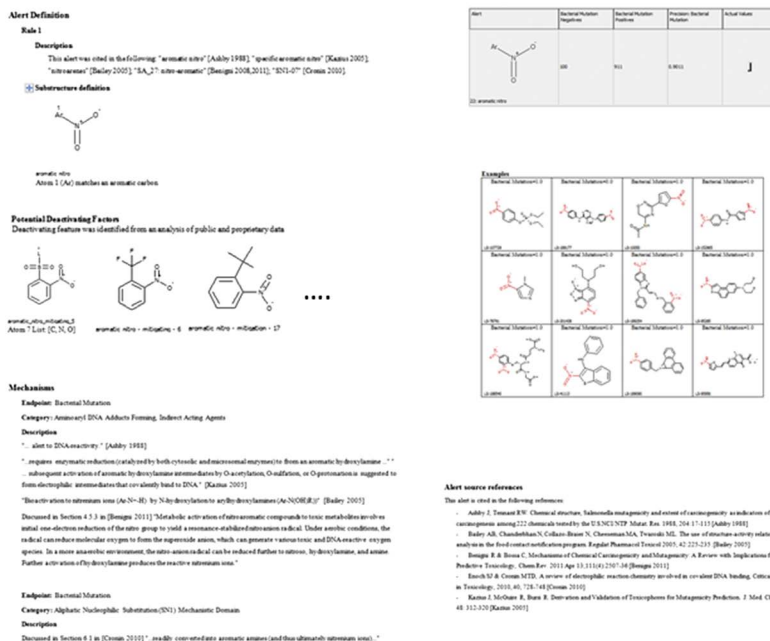
*In silico* methodologies, which includes expert alerts, QSAR models, and read-across, are being increasingly used as part of the chemical R&D process and submissions for regulatory authorities. To support their use, the OECD published five (Q)SAR model validation principles: “1. a defined endpoint; 2. an unambiguous algorithm; 3. a defined domain of applicability; 4. appropriate measures of goodness-of-fit, robustness and predictivity; 5. a mechanistic interpretation, if possible.”<sup>66</sup> These principles support the evaluation of any model used in predicting toxicity. The following sections outline these three major *in silico* methodologies.

### 9.3.2 Expert Alerts

Expert alerts (also referred to as expert rule-based or structural alerts) is a methodology for generating predictions for specific toxicity endpoints. Commonly used commercial systems include the Leadscape Genetox Expert Alerts<sup>67</sup> and Derek Nexus from Lhasa Limited.<sup>68</sup> Non-commercial expert alerts include the Benigini Bossa rulebase, which is part of the OECD toolbox<sup>69</sup> and ToxTree.<sup>70</sup> This methodology makes use of intellectually derived structural rules or alerts that are generally associated with specific toxic effects or mechanisms. These alerts are usually encoded as one or more molecular substructures that may have been reported in the public literature alongside a mechanistic justification for the structural features. Information on the precise structural definition needs to be encoded alongside rules to describe any criteria where the alert would not match a test chemical.

The aromatic nitro substructure is an example of an expert rule or alert for mutagenicity, as shown in Figure 9.1. It has been cited in a number of publications.<sup>71–75</sup> The structural definition defines both the aromatic nitro substructure along with different substituents that potentially deactivate its mutagenicity. In Figure 9.1 three deactivating fragments are shown for illustration. The alert will match any compound containing an aromatic nitro; however, if the test compound also contains one of the deactivating fragments then the alert will not match. In addition to the structural definition of the alert, information on the mechanism has also been collected from

## 22: aromatic nitro (Bacterial Mutation)



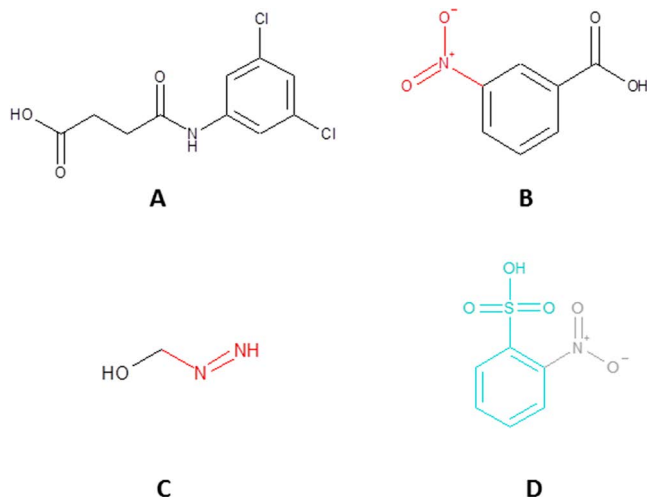
**Figure 9.1** Example alert (aromatic nitro) for mutagenicity from the Leadscape expert alerts.

different sources. The number of positive and negative known examples is reported along with data behind these tested chemicals.

Any alert reported in the literature should be critically evaluated before it is incorporated into the knowledge base of the alert's system and used to make prediction. A database of known positive and negative results (*i.e.* a reference set) can help to qualify the alerts by performance. Each prospective alert should have a significant positive association with the reference set data. This positive association results from when there is a significantly higher than expected number of positive compounds containing the alert than expected from a random sample. The number of positive and negative compounds matching the aromatic nitro alert is shown in Figure 9.1. In this example, 911 were positive and 100 were negative (or 90% positive). In comparison to the overall number of positive and negative examples in the reference set (48% positives), the association was determined to be significant and the aromatic nitro was included as an alert for mutagenicity.

If an expert alert system adheres to the OECD (Q)SAR validation principle 4, before any prediction is made, it should be established whether the test chemical is within the system's applicability domain. A positive prediction occurs when an alert is present in the test compound (based on the structural definition of the alert). If no alert occurs and the chemical is within the





**Figure 9.2** Four examples that were predicted using an expert alert system.

applicability domain, then a negative prediction is made. It is possible that there is some conflicting evidence for the alert which may result in the alert being classified as indeterminate.

Figure 9.2 shows four examples where predictions for bacterial mutagenicity were generated using the Leadscope Genetox Expert Alerts system.<sup>67</sup> Compound A was determined to be within the domain of applicability for the expert alert system, no alerts were identified and it was predicted to be negative. Compound B was predicted to be positive based on the presence of an aromatic nitro group highlighted in red. Based on the alert definition, no other group appears to deactivate the mutagenicity and hence it was predicted to be positive. An alkyl hydrazine alert was identified in compound C. There was conflicting data in the reference set matching this alert and hence it was assigned as indeterminate. Like compound B, compound D also contains an aromatic nitro group (shown in gray) but was predicted to be negative based on the mitigated fragment highlighted in light blue/green. Further information on any of the matched alerts, as illustrated in Figure 9.1, would provide additional supporting information to assess the result.

### 9.3.3 QSARs

QSAR models represent another important methodology to support the prediction of toxicity. They are based on a mathematical model that uses different descriptors generated automatically from the chemical structure. Commonly used commercial tools include the Leadscope Model Applier, covering 86 toxicity endpoints,<sup>76–83</sup> CASE Ultra from MultiCASE<sup>84–86</sup> and Sarah Nexus from Lhasa Limited for prediction of mutagenicity.<sup>87</sup> A number of non-commercial tools are available, including those available through OpenTox ToxPredict,<sup>88,89</sup> the Danish (Q)SAR Database,<sup>90</sup> and Vega.<sup>91</sup>

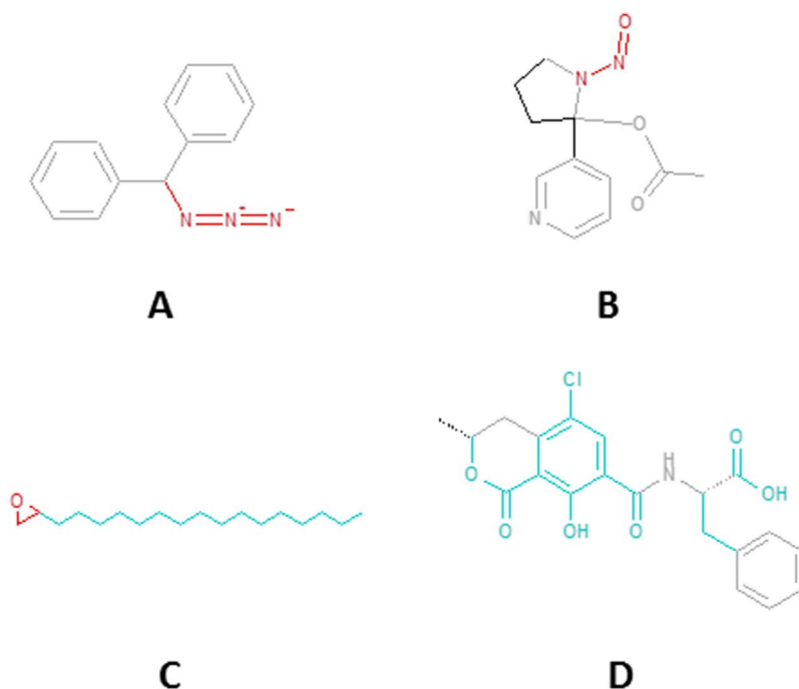
The most important part of building a QSAR model is assembling a training set. To assemble this training set, it is necessary to collect previously generated toxicity studies, potentially from a variety of sources, including on-line databases as well as from published literature. In collecting this information, establishing criteria concerning the type and quality of the studies is necessary. An overall assessment (*e.g.* mutagenic/non-mutagenic potential) should be collected alongside any supporting data on the study design or results determined to be helpful. During this process a compound may be found to have been tested more than once in different laboratories with inconsistent results that must be rectified. It is also essential to capture the chemical structure of the tested material in an electronic format.

A number of molecular descriptors can be generated and used in the QSAR model.<sup>92</sup> These descriptors are calculated directly from the chemical structure and may reflect properties of the whole molecule (such as molecular weight or logP) or the presence or absence of specific structural features such as different functional groups. A subset of these descriptors is selected and used to build and optimize a mathematical model describing the relationship between these descriptors to the overall toxicity assessment. Once the model is built, it is then necessary to validate the model, ideally using a large external test set representing the toxicity modeled in the QSAR model. The performance of the QSAR model is determined by comparing the actual and predicted values. A series of statistics, such as the Cooper statistics,<sup>93</sup> are calculated to assess overall model performance.

QSAR models that comply with OECD principles must first determine whether the chemicals being predicted are within the applicability domain of the model. Does the model adequately cover the structural features of the compound being tested? This analysis often takes into account the features used in the QSAR model as well as the chemical classes in the model's training set. As long as the chemical is within the applicability domain of the QSAR model, a prediction can be calculated. The same molecular descriptors as used in building the QSAR model are calculated for the test chemical. These descriptors are used as part of a mathematical model to calculate a prediction for the same endpoint used to train the model. When the endpoint being predicted is categorical, it may be necessary to convert a prediction (such as the probability of a positive outcome) to a categorical outcome (such as positive or negative) by applying cut-offs (such as positives are >0.6 probability, negatives are <0.4 probability, and indeterminate results are between 0.4 and 0.6).

When making a prediction, it is important to understand in detail how the QSAR model generated the prediction. For simple models, it may be possible to inspect the descriptors and equations used; however, complex and non-linear models are often used to generate QSAR models, making interpretation challenging. To ensure QSAR models are not considered a "black-box", many modelling approaches are accompanied by an explanation functionality. This will often highlight the portions of the test molecule that significantly influenced a positive or negative prediction.

In the following example, the Leadscape Model Applier software was used to create a prediction for *Salmonella* mutagenicity.<sup>81,83</sup> Predictions were made for four compounds shown in Figure 9.3. The probability of a positive outcome was predicted and a positive call made if this probability exceeded 0.6 and a negative prediction made if the probability was below 0.4. All compounds were within the applicability domain of the QSAR model. The structural diagrams in Figure 9.3 have been highlighted with red atoms and bonds supporting a positive prediction, green/blue atoms and bonds supporting a negative prediction and gray atoms and bonds when they do not appear to significantly activate or deactivate *Salmonella* mutagenicity. Any atom and bond that was not considered in the QSAR model is shown in black. Compound A was predicted to be positive with a predicted probability value of 0.991. All atoms and bonds of compound A were covered by at least one structural feature of the QSAR model, with the azide group shown to contribute towards the positive prediction. Compound B was also predicted to be positive with a predicted probability of 0.973. Most of the atoms and bonds are considered in the prediction; however, four bonds are shown in black indicating that no structural features in the model considered these atoms and bonds. The nitroso group is highlighted in red, indicating that this group is primarily responsible for the positive prediction. Compound C is predicted



**Figure 9.3** Four examples that were predicted using Leadscape's quantitative structure–activity response (QSAR) model for *Salmonella* mutagenicity.

to be negative based on a predicted probability of 0.368. All atoms and bonds were considered in this prediction. An epoxide (shown in red) provides evidence of potential *Salmonella* mutagenicity; however, the long alkyl chain (shown in light green/blue) appears to deactivate the mutagenicity. Finally, compound D is predicted to be negative, based on a predicted probability of 0.00418. Only one bond is shown in black, indicating that the majority of the compound's structural features were considered in making this prediction.

### 9.3.4 Read-Across

Read-across is an expert-based *in silico* approach, whereby experimental data from one or more chemical analogs (termed source compounds) are used to support the prediction of toxicity for a chemical with no data (termed target compound) for a toxicological endpoint of interest.<sup>94–97</sup> A one-to-one prediction is termed an analogue approach and a many-to-one prediction is termed a category approach. Qualitative predictions like hazard or quantitative predictions such as NOAELs are possible.

Any read-across approach and workflow is usually handled on a case-by-case basis. As it is time consuming and requires expert knowledge, read-across is most often applied to predict complex toxicological endpoints such as the expected toxicity after repeated exposure of low doses or reproductive toxicity, as well as endpoints for which validated QSAR models or *in vitro* testing batteries are not able to replace *in vivo* testing.

In Europe under REACH, read-across was the most frequently used alternative method to predict repeated-dose toxicity.<sup>98</sup> It was used in 3220 (33%) out of 9786 dossiers entries to predict repeated-dose toxicity (all routes, all study durations) of phase-in-substances with production volumes of 100–1000 tonnes per year.

The challenge for read-across is to identify “similar” source compounds. In principle, convincing evidence of “similarity” of source and target compounds have to be provided with regard to chemical as well as biological similarity. In cases where metabolism is an important aspect, chemical and biological similarity must also be evaluated for critical metabolites.

Chemical similarity is defined by shared structural features, *e.g.* functional groups, but also by physico-chemical properties. These features and/or properties should be carefully evaluated for their relevance to the predicted toxicological endpoint. Endpoints related to reactivity may focus on the presence or absence of reactive functional groups contained within the source compounds (or their metabolites). In other cases a consistent trend, *e.g.* the effect of different aliphatic side chain lengths, may also be appropriate. Physico-chemical properties are often used as a first indication of absorption, distribution, and elimination *in vivo*, and might accordingly be similar or follow a consistent trend. Next, the biological similarity has to be addressed. The question is “do the source compounds have a similar mode of action?” and “how can this be proven?” Based on the analysis of appropriate *in vivo* data, *e.g.* from animal or human investigations, a read-across hypothesis has to be defined. For repeated-dose toxicity, biological similarity may include the evaluation of similar

critical findings, *e.g.* neoplastic or non-neoplastic lesions in identical target organs. As an example, organo(thio)phosphates typically induce a decrease in acetylcholinesterase (AChE) which leads to neurological symptoms. Beside shared toxicodynamics, toxicokinetic information is crucial to conclude that the behavior of the source compounds *in vivo* is similar and to make a prediction such as a NOAEL value with confidence. When looking for analogs with relevant toxicological data, toxicity databases, as discussed in Section 9.2 provide a critical resource for this process.

In traditional risk assessment, which is mainly based on *in vivo* animal studies, often the underlying mechanisms of actions or adverse outcome pathways are not known. However, this mechanistic information might be a key element for further supporting the hypothesis of biological similarity. Recently, a framework has been developed in the SEURAT-1 project with the aim to predict systemic toxicity after repeated exposure for chemicals and cosmetics through the integration of mechanistic data from *in vitro* investigations.<sup>99</sup> Four different read-across scenarios were investigated, to demonstrate how evidence from *in vitro* molecular screening, '-omics' assays, and computational models can support a traditional read-across based on structural similarities between source and target substance.<sup>100</sup> These four read-across scenarios are as follows.

- (1) Chemically similar compounds that do not require/undergo metabolism to exert a potential adverse human health effect; for example perfluoroalkyl acids<sup>101</sup>
- (2) Chemically similar compounds involving metabolism (resulting in exposure to the same/similar toxicant); for example  $\beta$ -unsaturated alcohols
- (3) Chemicals with general low or no toxicity; for example primary alcohols
- (4) Chemicals in a structurally similar category with variable toxicities based on a mode of action hypothesis; for example alkylphenols.

Read-across is a prediction containing a certain uncertainty, which needs to be assessed. The selection and accuracy of the analogues, the quality of the experimental data, the relevance of data and accuracy with regard to the toxicological endpoint, as well as data gaps need to be evaluated and documented.<sup>102</sup> ECHA has recently drafted a Read-Across Assessment Framework for human endpoints to support a transparent and systematic read-across assessment and workflow.<sup>103</sup>

## 9.4 Expert Reviews

### 9.4.1 Assessing Experimental Data

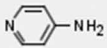
Before running any *in silico* models, the first step should be to search toxicity databases for appropriate laboratory results performed on the test material. Any results from a database search should contain sufficient information necessary to understand the adequacy of the study.<sup>104</sup>

In the following example, the mutagenicity of 4-pyridylamine is being assessed (shown in Figure 9.4). A bacterial reverse mutation assay may be used to assess mutagenicity as defined in OECD guideline 471.<sup>7</sup> When assessing the mutagenicity of 4-pyridylamine, the first step is to identify whether any bacterial reverse mutation assay data are available, if the experimental design and results are available for inspection, and whether the study was performed according to the OECD 471 guideline. In this example a five-strain GLP assay was performed in a manner consistent with the guideline. Since the study was adequately performed, no further *in silico* predictions need be generated, and it can be concluded that 4-pyridylamine is non-mutagenic from the laboratory data.

Several issues can make retrieving experimental data challenging. An analysis of the study results may reveal deviations from standard test protocols. A database search might retrieve multiple study results for the same compound (or different forms of the same compound) from different laboratories, and those results might be conflicting. A critical assessment of the experiments should be considered and conclusions drawn. Searching many individual public and commercial toxicity databases along with in-house databases can be time consuming. Utilizing an intermediate database search service that searches many different sources simultaneously would save a great deal of time. For example, a single search of the Leadscape SAR Genetox database (as discussed in Section 9.2.3) returns studies from more than 11 different sources.

### 9.4.2 Drawing Conclusions from Multiple Systems

Each *in silico* model that adheres to the OECD (Q)SAR validation principles should include information concerning its performance, often based on a cross-validation exercise or using an external test set.<sup>66</sup> It has been shown that using two or more complementary *in silico* models may improve the detection of positive compounds (resulting in a higher sensitivity). This conservative approach minimizes the chance of missing a positive outcome; however, this approach also increases the number of false positive (resulting in lower specificity). For example, it was recommended that combining the results from two complementary (Q)SAR methodologies (one expert rule based and one statistical based) should be used as part of the ICH M7 guideline to assess pharmaceutical impurities.<sup>105,106</sup> The performance statistics resulting from combining the two methodologies are illustrated here using the Leadscape expert rule-based and statistical-based models. Table 9.2 shows the results of running the two methodologies over the Hansen set<sup>107</sup> after removing any compounds contained in the training set of the statistical-based model. The results of the individual models are shown in Table 9.2, along with the overall conclusion column, by combining the results from the two models. The sensitivity of the overall conclusion increased by 8% (compare with the expert rule-based methodology); however, the specificity decreased.

<div>  <p>LS-130202</p> </div>	<b>Study call:</b>	Negative
	<b>Title:</b>	Bacterial Reverse Mutation Assay
	<b>Reference:</b>	<a href="http://www.accessdata.fda.gov/drugsatfda_docs/nda/2010/022250s000_PharmaR.pdf#page=155">http://www.accessdata.fda.gov/drugsatfda_docs/nda/2010/022250s000_PharmaR.pdf#page=155</a>
	<b>Study type:</b>	bacterial mutagenesis
	<b>Source:</b>	cdcr
	<b>Species:</b>	Salmonella typhimurium (16); Escherichia coli (4)
	<b>Strains:</b>	TA98 (4); TA100 (4); TA1535 (4); TA1537 (4); WP2uvrA (4)
	<b>Metabolic activation:</b>	Present (10); Absent (10)
	<b>Metabolic activation system:</b>	S9 Rat Liver Macrophages (2); S9 Rat Liver Hepatocytes (6); S9 Rat Liver (2)
	<b>Test calls:</b>	Negative (20)
	<b>Dose summary:</b>	1.5 micro-g/plate (10), 5.0 micro-g/plate (10), 15.0 micro-g/plate (10), 50.0 micro-g/plate (20), 150.0 micro-g/plate (20), 500.0 micro-g/plate (20), 1500.0 micro-g/plate (20), 5000.0 micro-g/plate (20)
	<b>Study Report:</b>	<a href="#">Leadscope DB Study Report.pdf</a>

**Figure 9.4** Results of a database search returning mutagenicity data for 4-pyridylamine.



**Table 9.2** Combining multiple *in silico* methodologies to increase overall sensitivity.<sup>a</sup>

	Leadscope genetox expert alerts	Leadscope statistical <i>Salmonella</i> QSAR	Overall conclusion
Concordance	82%	76%	81%
Sensitivity	90%	82%	98%
Specificity	71%	68%	59%
Positive predictivity	81%	78%	76%
Negative predictivity	83%	73%	96%
Coverage	97%	87%	99.9%

<sup>a</sup>QSAR: quantitative structure–activity relationship model.

High sensitivity is usually a desired objective when combining the results from more than one methodology and this can be achieved using the general rule that a positive from any of the methodologies results in an overall positive conclusion. A clear negative conclusion may be generated when all methodologies generate a negative response. However, it is possible that any prediction methodology may be unable to generate a classification call with adequate confidence, *i.e.* the prediction is inconclusive. In addition, one or more of the models may be unable to generate a prediction when the test compound is outside the applicability domain of the model. In these situations it may be necessary to analyze the results further to construct an expert review to reach an overall conclusion.

### 9.4.3 Reviews Accepting or Refuting An *In silico* Result

Expert reviews accepting or refuting results from any individual prediction are used to improve the overall accuracy. They may supersede a positive or negative *in silico* result based upon expert knowledge. These reviews can help resolve inconclusive predictions or out-of-domain results. The use of expert knowledge has been shown to improve predictivity.<sup>108–111</sup> Documentation of such reviews is essential to ensure that overall conclusions are transparent and defensible. Expert reviews often fall into four categories:

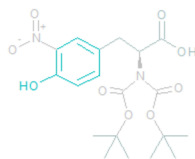
- *Reviews based on an understanding of the biological mechanism* A positive or inconclusive prediction may be based upon an alerting structural feature present in the test chemical. The alert may have been identified from a sufficiently strong mechanistic basis or statistical association with the toxicity endpoint being modeled. However, a specific test compound with an alerting fragment may not be toxic. This may be explained through an assessment of the test chemical's environment and how the mechanism does not apply to this specific chemical. Hence, a review based on the biological mechanism may be used to refute positive or inconclusive predictions.

- *Reviews based on an assessment of the relevancy of the features in the (Q)SAR model* An examination of the structural features used in a (Q)SAR model to make a prediction may reveal that the underlying basis for these features is not sufficient and a positive or inconclusive prediction should be refuted. For example, structural features are selected automatically when building a statistical-based model. Positively weighted features may be identified in a set of predominantly positive training set examples. However, a visual inspection of these compounds may reveal that other co-occurring features are most likely responsible for the toxicity. If these coincidental features significantly contributed to the positive prediction, then an expert review refuting the prediction may be made. Other possible reasons to refute a prediction include structural mitigation of the toxicity, insufficient examples used to identify the positively weighted structural feature, results based upon low-weighted structural features (*i.e.* there is no clear basis for the positive prediction), and irrelevant training set examples or examples based upon unreliable test results.
- *Reviews based on a comparison with chemical analogs* The results of a structure similarity search of toxicity databases or other collections may be used to refute a positive, negative, or inconclusive prediction. They may even be used in an expert review for out-of-domain results. One approach to refute a positive or inconclusive prediction is to identify negatively tested analogs that share the same alerting feature in the same environment as the test chemical. It is important to verify that there are no additional unrelated features that contribute to the positive prediction and that there are no fragments deactivating the toxicity of the negatively tested analogs that are missing in the test chemical. Another strategy to refute positive or inconclusive predictions or even out-of-domain results is to identify analogs whose only difference is the addition of non-toxic structural features (such as a protecting group).
- *Reviews based on a visual inspection of the results* To address situations where a lack of evidence exists, such as when one of the *in silico* methodologies does not generate a prediction, it may be possible to construct an expert review based upon a visual examination of the results. For example, an expert toxicologist or chemist may be able to examine the compound and confirm that there are no plausible structural alerts for the toxicity being assessed. In addition, it may be possible to opine based on the confidence or probability of the accuracy of a prediction.

In the following example (shown in Figure 9.5), the test chemical was predicted to be inconclusive in one of the *in silico* methodologies and negative in another for mutagenicity. An aromatic nitro group is highlighted in the test chemical. A negatively tested analog was identified that also contains the aromatic nitro in the same environment as the test chemical. The only difference is the addition of a non-reactive group; hence an expert review concluded that the test chemical was negative.

**ACT Case Study 5 (impurity) #4 - Negative result analysis - evaluation of suitable analogs**

The following chemical analogs were identified from the Leadscape Genetox SAR Database. Both alerts and model features found on the test impurity are highlighted on the analogs. (Alerts and positive model features in red, negative model features in blue-green, and indeterminate in gray). **The list of analogs has been abbreviated for this report**



ACT Case Study 5 (impurity)

Similarity=0.58 Bacterial Mutation=0.0  LS-188180				
--	--	--	--	--

The selected analogs support the negative *in silico* prediction.

**Comments:** The alerting feature (aromatic nitro) is fully contained with the same environment within LS-188180 which is a known negative. The only difference between the test compound and the known negative is the addition of a non-reactive group.

**Figure 9.5** Review refuting an inconclusive prediction.

#### 9.4.4 Documenting *In silico* Results

Any report that documents the *in silico* results and conclusions may contain the following components:

- **Materials and methods** The software system (model and/or alerts) used in the *in silico* assessment. This should also include the version number, since results can change when new version of these models and alerts are run. In addition, if any databases were used, the name and version numbers may be included.
- **Summary of results and conclusions** The results may be presented as a table, showing for each chemical structure investigated, the results from the individual *in silico* methodologies as well as an overall call. This may include experimental data if it exists for any of the chemicals. Overall conclusions as well as any supporting remarks may be helpful to include in this summary table.
- **Supporting information** Documentation of any supporting or refuting expert reviews used in the assessment of the chemicals may also be included to ensure that the results are fully transparent and defensible, including appropriate references to supporting material from the literature.
- **Appendices** The full *in silico* reports generated by the software, study reports containing experimental data and documentation detailing the models built may be included here. Two standardized reports may be used: a QSAR prediction reporting format (*i.e.* a standardized report detailing how a prediction was generated) and a QSAR model reporting format (*i.e.* a standardized report detailing how the model was built and validated).<sup>17</sup> These types of reports may be included here. A read-across report may also be documented as outlined by Schultz *et al.*<sup>112</sup>

Figures 9.6 and 9.7 illustrate some components of an *in silico* report to document the assessment of mutagenicity for four potential impurities (per ICH M7).<sup>105,106</sup>

### 9.5 Conclusions

This chapter reviews a series of *in silico* methodologies for predicting toxicity, which have been discussed in the context of current *in vivo* and *in vitro* approaches. Decisions are increasingly being made based on all three approaches—*in vivo*, *in vitro*, and *in silico*. Underpinning all *in silico* methodologies is the need to access high-quality and up-to-date toxicity study data from a variety of sources. Methods for organizing toxicity data in a harmonized manner (such as ToxML) support combining this disparate data. The process of collecting this information is time-consuming and expensive; however, the quality of any of prediction models is directly related to the quality and amount of data collected.

### Materials and methods

An assessment aligned with ICH M7 was performed on the actual and potential impurities listed in the table below using the following QSAR methodologies and systems:



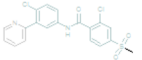
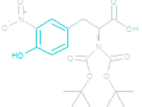
Expert rule-based methodology and parameters:	Leadscope genetox expert alerts v2 (System: Leadscope Model Applier v2.0.5.2); the domain assessment was turned on
Statistical-based methodology and parameters:	Leadscope Salmonella statistical-based QSAR model v3, Leadscope E.coli/TA102 statistical-based QSAR model v1 (System: Leadscope Model Applier v2.0.5.2); probabilities above 0.6 set to positive, probabilities below 0.4 set to negative and domain assessment was turned on
Genetic toxicity database used for searching:	Leadscope SAR genetox 2015
Rodent carcinogenicity database used for searching:	Leadscope SAR carcinogenicity 2015

These *in silico* methodologies follow the general validation principles set forth by the Organisation for Economic Co-operation and Development (OECD). The Leadscope statistical-based methodologies and the Leadscope SAR databases were developed through a research collaboration agreement with the US FDA.

**Figure 9.6** Example materials and methods component of an *in silico* report document.

## Results

The following table and notes summarize the results of this QSAR analysis for 4 impurities. Structural alerts and significant model features are highlighted. Alerts and positive model features in red, negative model features in blue-green, and indeterminate in gray.

#	Impurity Structure	Laboratory Data	Expert rule based system	Statistical based system	Overall assessment	M7 Class assignment	Additional supportive evidence and comments
1	 ACT Case Study 2	None	Negative	Negative	Non-mutagenic	5	<b>Accepted negative <i>in silico</i> result.</b> The impurity lacks obvious reactive potential.
2	 ACT Case Study 3	None	Positive	Positive	Predicted Mutagenic	3	<b>Accepted positive <i>in silico</i> result.</b> Matching alerts: 267: aromatic amine(NH2) (strong activating anilines) (0.92)
3	 ACT Case Study 4	None	Negative	Negative	Non-mutagenic	5	<b>Accepted negative <i>in silico</i> result.</b> The impurity lacks obvious reactive potential.
4	 ACT Case Study 5 (impurity)	None	Indeterminate	Negative	Non-mutagenic	5	<b>Accepted negative <i>in silico</i> result.</b> See 'ACT Case Study 5 (impurity) #4 - Appendices. [4-appendix]'

**Figure 9.7** Example of a summary table view of *in silico* results.

Three of the most commonly used methodologies in predicting toxicity are expert alerts, QSAR models and read-across. These complementary approaches provide different viewpoints concerning the structural and mechanistic basis for any prediction, alongside an analysis and rationale for supporting analog data. This information can then be assimilated within an expert review to generate a final conclusion.

*In silico* methods are being increasingly used to predict toxicity. The documentation of real-world case studies and guidelines will further support the adoption and use of *in silico* methods. New and updated models are important, particularly to predict safe dose levels and better assess prediction confidence alongside new educational programs that support the use and interpretation of these computational models.

## Acknowledgements

The authors would like to thank Catrin Hasselgren and Adam Woolley for their helpful discussion.

## References

1. [https://echa.europa.eu/documents/10162/13632/information\\_requirements\\_r6\\_en.pdf](https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf).
2. <http://www.oecd.org/chemicalsafety/>.
3. <http://www.ich.org/home.html>.
4. H. Olson, G. Betton and D. Robinson, *et al.*, Concordance of the toxicity of pharmaceuticals in humans and in animals, *Regul. Toxicol. Pharmacol.*, 2000, **32**, 56–67, DOI: 10.1006/rtph.2000.1399.
5. M. R. Fielden and K. L. Kolaja, The role of early *in vivo* toxicity testing in drug discovery toxicology, *Expert Opin. Drug Saf.*, 2008, **7**, 107–110, DOI: 10.1517/14740338.7.2.107.
6. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products.
7. OECD, *Test 471. Bacterial Reverse Mutation Test OECD Guideline for Testing of Chemicals Section 4*, 1997.
8. OECD, *Test No. 473: In Vitro Mammalian Chromosomal Aberration Test*, 2014.
9. OECD, *Test No. 487: In Vitro Mammalian Cell Micronucleus Test*, 2014.
10. OECD, *Test No. 428: Skin Absorption: In Vitro Method*, 2004.
11. OECD, 2004b. *Test No. 431: In Vitro Skin Corrosion: Human Skin Model Test*, 2004.
12. <https://federalregister.gov/a/2015-15182>.
13. <https://eurl-ecvam.jrc.ec.europa.eu/>.
14. <http://ntp.niehs.nih.gov/pubhealth/evalatm/iccvam/index.html>.
15. G. Daston, D. J. Knight, M. Schwarz, T. Gocht, R. S. Thomas, C. Mahony and M. Whelan, SEURAT: Safety Evaluation Ultimately Replacing Animal Testing—Recommendations for future research in the field of predictive toxicology, *Arch. Toxicol.*, 2015, **89**, 15–23, DOI: 10.1007/s00204-014-1421-5.



16. A. Amberg, J. Harvey, A. Czich, H. P. Spirkel, S. Robinson, A. White and D. P. Elder, Do Carboxylic/Sulfonic Acid Halides Really Present a Mutagenic and Carcinogenic Risk as Impurities in Final Drug Products?, *Org. Process Res. Dev.*, 2015, **19**, 1495–1506, DOI: 10.1021/acs.oprd.5b00106.
17. <https://eurl-ecvam.jrc.ec.europa.eu/databases/jrc-qsar-model-database-and-qsar-model-reporting-formats>.
18. M. Ali, M. Patel, D. Wilkinson, P. Judson, K. Cross and D. Bower, ToxML, a data exchange standard with content controlled vocabulary used to build better (Q)SAR models, *SAR QSAR Environ. Res.*, 2013, **24**, 429–438, DOI: 10.1080/1062936X.2013.783506.
19. <http://www.w3.org/XML/>.
20. <http://www.cdisc.org/sdtm>.
21. <http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM312964.pdf>.
22. <http://iuclid.eu/>.
23. M. Skydsgaard, International Harmonization of Nomenclature and Diagnostic Criteria (INHAND) for Lesions in the Minipig, *Toxicol. Pathol.*, 2016, **44**, 480–481, DOI: 10.1177/0192623315614119.
24. D. Rudmann and R. Cardiff, *et al.*, Proliferative and nonproliferative lesions of the rat and mouse mammary, Zymbal's, preputial, and clitoral glands, *Toxicol. Pathol.*, 2012, **40**(suppl. 6), 7S–39S, DOI: 10.1177/0192623312454242.
25. D. Creasy and A. Bube, *et al.*, Proliferative and nonproliferative lesions of the rat and mouse male reproductive system, *Toxicol. Pathol.*, 2012, **40**(suppl. 6), 40S–121S, DOI: 10.1177/0192623312454337.
26. W. Kaufmann and B. Bolon, *et al.*, Proliferative and nonproliferative lesions of the rat and mouse central and peripheral nervous systems, *Toxicol. Pathol.*, 2012, **40**(suppl. 4), 87S–157S, DOI: 10.1177/0192623312439125.
27. K. S. Frazier and J. C. Seely, *et al.*, Proliferative and nonproliferative lesions of the rat and mouse urinary system, *Toxicol. Pathol.*, 2012, **40**(suppl. 4), 14S–86S, DOI: 10.1177/0192623312438736.
28. R. Renne and A. Brix, *et al.*, Proliferative and nonproliferative lesions of the rat and mouse respiratory tract, *Toxicol. Pathol.*, 2009, **37**(suppl. 7), 5S–73S, DOI: 10.1177/0192623309353423.
29. <http://goreni.org>.
30. <http://ntp.niehs.nih.gov/nnl/>.
31. <http://toxnet.nlm.nih.gov/newtoxnet/cpdb.htm>.
32. CCRIS (Chemical Carcinogenesis Res. Information System): Developed and maintained by National Cancer Institute, <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>.
33. R. Benigni, C. Bossa, A. M. Richard and C. Yang, A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity, *Ann. Ist. Super. Sanita*, 2008, **44**, 48–56.
34. <http://www.echemportal.org>.

35. <http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances>.
36. <http://toxnet.nlm.nih.gov/newtoxnet/genetox.htm>.
37. <http://monographs.iarc.fr/ENG/Classification/index.php>.
38. <http://www.inchem.org>.
39. <http://www.epa.gov/iris/>.
40. [http://Dra4.nihs.go.jp/mhlw\\_data/jsp/SearchPagesEND.jsp](http://Dra4.nihs.go.jp/mhlw_data/jsp/SearchPagesEND.jsp).
41. [http://www.leadscope.com/toxicity\\_databases/](http://www.leadscope.com/toxicity_databases/) and [http://www.leadscope.com/toxicity\\_databases/regulatory\\_databases/](http://www.leadscope.com/toxicity_databases/regulatory_databases/).
42. <http://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm>.
43. R. W. Tennant, The genetic toxicity database of the National Toxicology Program: evaluation of the relationships between genetic toxicity and carcinogenicity, *Environ. Health Perspect.*, 1991, **96**, 47–51.
44. <http://www.pharmapendium.com>.
45. <http://www.cdc.gov/niosh/rtecs/default.html>.
46. <http://www.lhasalimited.org/products/vitic-nexus.htm>.
47. <http://toxnet.nlm.nih.gov/newtoxnet/dart.htm>.
48. <http://www.fda.gov/ScienceResearch/BioinformaticsTools/Endocrine-DisruptorKnowledgebase/default.htm>.
49. <http://www.item.fraunhofer.de/en/research-areas/chemical-risk-assessment-databases-expert-systems/testing-strategies-structure-activity-relationships.html>.
50. E. Julien, C. C. Willhite, A. M. Richard, J. M. DeSesso and I. R. W. Group, Challenges in constructing statistically based structure-activity relationship models for developmental toxicity, *Birth Defects Res., Part A*, 2004, **70**, 902, DOI: 10.1002/bdra.20087.
51. R. B. Fitzpatrick, LactMed: Drugs and Lactation database, *J. Electron. Resour. Med. Libr.*, 2007, **4**, 155.
52. <https://services.toxbank.net/>.
53. A. R. Scialli, A. Lione and G. K. B. Padgett, *Reproductive Effects of Chemical, Physical, and Biologic Agents: REPROTOX*, Johns Hopkins University Press, 1995.
54. T. H. Shepard, *Catalog of Teratogenic Agents*, The Johns Hopkins University Press, 10th edn, 2001.
55. <http://depts.washington.edu/terisdb/index.html>.
56. <http://www2.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data>.
57. <http://cosmosdb.eu/>.
58. Compatible with OECD toolbox, <http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm>, stand-alone application available upon registration, <http://www.nite.go.jp/en/chem/qsar/hess-e.html>.
59. Y. Sakuratani, H. Q. Zhang, S. Nishikawa, K. Yamazaki, T. Yamada, J. Yamada, K. Gerova, G. Chankov, O. Mekenyan and M. Hayashi, Hazard Evaluation Support System (HESS) for predicting repeated dose toxicity using toxicological categories, *SAR QSAR Environ. Res.*, 2013, **24**, 351–363, DOI: 10.1080/1062936X.2013.773375.
60. <http://fraunhofer-repdose.de/>.

61. M. Batke, S. Escher, S. Hoffmann-Doerr, C. Melber, H. Messinger and I. Mangelsdorf, Evaluation of time extrapolation factors based on the database RepDose, *Toxicol. Lett.*, 2011, **205**, 122–129, DOI: 10.1016/j.toxlet.2011.05.1030.
62. M. Batke, T. Aldenberg, S. Escher and I. Mangelsdorf, Relevance of non-guideline studies for risk assessment: the coverage model based on most frequent targets in repeated dose toxicity studies, *Toxicol. Lett.*, 2013, **218**, 293–298, DOI: 10.1016/j.toxlet.2012.09.002.
63. A. Bitsch, S. Jacobi, C. Melber, U. Wahnschaffe, N. Simetska and I. Mangelsdorf, REPDOSE: A database on repeated dose toxicity studies of commercial chemicals – a multifunctional tool, *Regul. Toxicol. Pharmacol.*, 2006, **46**, 202–210, DOI: 10.1016/j.yrtph.2006.05.013.
64. S. E. Escher, I. Tluczkiewicz, M. Batke, A. Bitsch, C. Melber, E. D. Kroese, H. E. Buist and I. Mangelsdorf, Evaluation of inhalation TTC values with the database RepDose, *Regul. Toxicol. Pharmacol.*, 2010, **58**, 259–274, DOI: 10.1016/j.yrtph.2010.06.009.
65. S. E. Escher, M. Batke, S. Hoffmann-Doerr, H. Messinger and I. Mangelsdorf, Interspecies extrapolation based on the RepDose database—A probabilistic approach, *Toxicol. Lett.*, 2013, **218**, 159–165, DOI: 10.1016/j.toxlet.2013.01.027.
66. *Guidance Document on the Validation of (Quantitative) Structure Activity Relationships [(Q)SAR] Models*, [http://ihcp.jrc.ec.europa.eu/our\\_labs/predictive\\_toxicology/background/oecd-principles](http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/background/oecd-principles).
67. [http://www.leadscope.com/white\\_papers/ICHM7-WhitePaper-0314.pdf](http://www.leadscope.com/white_papers/ICHM7-WhitePaper-0314.pdf) and [http://www.leadscope.com/genetox\\_expert\\_alerts/](http://www.leadscope.com/genetox_expert_alerts/).
68. J. C. Dearden and M. D. Barratt, *et al.*, The Development and Validation of Expert Systems for Predicting Toxicity, *ATLA*, 1997, **25**, 223–252.
69. <http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartool-box.htm>.
70. <http://toxtree.sourceforge.net/>.
71. J. Ashby and R. W. Tennant, Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested by the U.S.NCI/NTP, *Mutat. Res.*, 1988, **204**, 17–115, DOI: 10.1016/0165-1218(88)90114-0.
72. J. Kazius, R. McGuire and R. Bursi, Derivation and Validation of Toxicophores for Mutagenicity Prediction, *J. Med. Chem.*, 2005, **48**, 312–320, DOI: 10.1021/jm040835a.
73. A. B. Bailey, N. Chanderbhan, N. Collazo-Braier, M. A. Cheeseman and M. L. Twaroski, The use of structure-activity relationship analysis in the food contact notification program, *Regul. Toxicol. Pharmacol.*, 2005, **42**, 225–235, DOI: 10.1016/j.yrtph.2005.04.006.
74. R. Benigni and C. Bossa, Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology, *Chem. Rev.*, 2011, **111**, 2507–2536, DOI: 10.1021/cr100222q.
75. S. J. Enoch and M. T. D. Cronin, A review of electrophilic reaction chemistry involved in covalent DNA binding, *Crit. Rev. Toxicol.*, 2010, **40**, 728–748, DOI: 10.3109/10408444.2010.494175.

76. G. J. Myatt and K. P. Cross, *In silico* Solutions for Predicting Efficacy and Toxicity in: *Human-Based Systems for Translation Research*, ed. R. Coleman, Royal Society of Chemistry, UK, 2015, pp. 190–214.
77. C. J. Ursem, N. L. Kruhlak, J. F. Contrera, P. M. MacLaughlin, R. D. Benz and E. J. Matthews, Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans. Part A: use of FDA post-market reports to create a database of hepatobiliary and urinary tract toxicities, *Regul. Toxicol. Pharmacol.*, 2009, **54**, 1–22, DOI: 10.1016/j.yrtph.2008.12.009.
78. E. J. Matthews, C. J. Ursem, N. L. Kruhlak, R. D. Benz, D. A. Sabaté, C. Yang, G. Klopman and J. F. Contrera, Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract, *Regul. Toxicol. Pharmacol.*, 2009, **54**, 23–42, DOI: 10.1016/j.yrtph.2009.01.009.
79. E. J. Matthews, N. L. Kruhlak, R. D. Benz, D. A. Sabaté, C. A. Marchant and J. F. Contrera, Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part C: Use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities, *Regul. Toxicol. Pharmacol.*, 2009, **54**, 43–65, DOI: 10.1016/j.yrtph.2009.01.007.
80. E. J. Matthews and A. A. Frid, Prediction of drug-related cardiac adverse effects in humans—A: creation of a database of effects and identification of factors affecting their occurrence, *Regul. Toxicol. Pharmacol.*, 2010, **56**, 247–275, DOI: 10.1016/j.yrtph.2009.11.006.
81. L. Stavitskaya, *et al.*, *Development of Improved Salmonella Mutagenicity QSAR Models Using Structural Fingerprints of Known Toxicophores*, Soc. Of Tox, 2013, <http://www.leadscope.com/media/SOT%202013%20Stavitskaya%20CDER%20poster.pdf>.
82. L. Stavitskaya, *et al.*, *Development of Improved QSAR Models for Predicting A-T Base Pair Mutations*, *Genetic Toxicity Association Poster Presentation*, 2013, [http://www.leadscope.com/media/GTA\\_LStavitskaya.pdf](http://www.leadscope.com/media/GTA_LStavitskaya.pdf).
83. [http://www.leadscope.com/model\\_appliers/](http://www.leadscope.com/model_appliers/).
84. M. T. Kim, A. Sedykh, S. K. Chakravarti, R. D. Saiakhov and H. Zhu, Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches, *Pharm. Res.*, 2014, **31**, 1002–1014, DOI: 10.1007/s11095-013-1222-1.
85. R. D. Saiakhov, S. K. Chakravarti and G. Klopman, Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs, *Mol. Inf.*, 2012, **32**, 87–97, DOI: 10.1002/minf.201200081.
86. S. K. Chakravarti, R. D. Saiakhov and G. Klopman, Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts, *J. Chem. Inf. Model.*, 2012, **52**, 2609–2618, DOI: 10.1021/ci300111r.
87. <http://www.lhasalimited.org/products/sarah-nexus.htm>.

88. <http://www.opentox.org/>.
89. <https://apps.ideaconsult.net/ambit2/ui>.
90. <http://qsar.food.dtu.dk/>.
91. <http://www.vega-qsar.eu/>.
92. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, ed. R. Mannhold, H. Kubinyi and G. Folkers, Wiley-VCH, Weinheim, Germany, 2009.
93. J. A. Cooper, R. Saracci and P. Cole, Describing the Validity of Carcinogen Screening Tests, *Br. J. Cancer*, 1979, **39**, 87–89.
94. S. Dimitrov and O. Mekenyan, An introduction to read-across for the prediction of effects of chemicals, in *In Silico Toxicology: Principles and Applications*, ed. M. T. D. Cronin and J. C. Madden, The Royal Society of Chemistry, Cambridge, 2010, pp. 372–384.
95. K. van Leeuwen and T. Schultz, *et al.*, Using chemical categories to fill data gaps in hazard assessment, *SAR QSAR Environ. Res.*, 2009, **20**, 207–220, DOI: 10.1080/10629360902949179.
96. M. T. D. Cronin, An introduction to chemical grouping, categories and read across to predict toxicity, in *Chemical Toxicity Prediction: Category Formation and Read-Across Applications*, ed. M. T. D. Cronin, J. C. Madden, S. J. Enoch and D. W. Roberts, The Royal Society of Chemistry, Cambridge, 2013a, pp. 1–29.
97. M. T. D. Cronin, Evaluation of categories and read-across for toxicity prediction allowing for regulatory acceptance, in *Chemical Toxicity Prediction: Category Formation and Read-Across Applications*, ed. M. T. D. Cronin, J. C. Madden, S. J. Enoch and D. W. Roberts, The Royal Society of Chemistry, Cambridge, 2013b, pp. 155–167.
98. [http://echa.europa.eu/documents/10162/13628/raaf\\_en.pdf](http://echa.europa.eu/documents/10162/13628/raaf_en.pdf).
99. T. Gocht and E. Berggren, *et al.*, The SEURAT-1 Approach towards Animal Free Human Safety Assessment, *Altex-Altern. Anim. Ex.*, 2015, **32**, 9–24, DOI: 10.14573/altex.1408041.
100. E. Berggren, P. Amcoff, R. Benigni, K. Blackburn, E. Carney, M. Cronin, H. Deluyker, F. Gautier, R. S. Judson, G. E. N. Kass, D. Keller, D. Knight, W. Lilienblum, C. Mahony, I. Rusyn, T. Schultz, M. Schwarz, G. Schüürmann, A. White, J. Burton, A. M. Lostia, S. Munn and A. Worth, Chemical Safety Assessment Using Read-Across: Assessing the Use of Novel Testing Methods to Strengthen the Evidence Base for Decision Making, *Environ. Health Perspect.*, 2015, **123**, 1232–1240, DOI: 10.1289/ehp.1409342.
101. T. W. Schultz, C. L. Mellor, K. R. Przybylak, S. Escher, R. Judson, I. Tsakovka, M. T. D. Cronin, A. N. Richarz, *Regul. Toxicol. Pharmacol.*, Summary and Reporting of the Grouping of Perfluoroalkyl Acids to Read Across 90-Day Oral Repeated-dose Toxicity, submitted for publication.
102. K. Blackburn and S. B. Stuard, A framework to facilitate consistent characterization of read across uncertainty, *Regul. Toxicol. Pharmacol.*, 2014, **68**, 353–362, DOI: 10.1016/j.yrtph.2014.01.004.

103. [http://echa.europa.eu/documents/10162/13628/raaf\\_en.pdf](http://echa.europa.eu/documents/10162/13628/raaf_en.pdf).
104. H. J. Klimisch, M. Andreae and U. Tillmann, A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data, *Regul. Toxicol. Pharmacol.*, 1997, **25**, 1–5, DOI: 10.1006/rtp.1996.1076.
105. ICH M7 – assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk, <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM347725>.
106. Addendum to ICH M7: Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk: application of the principles of the ICH M7 guideline to calculation of compound-specific acceptable intakes M7 (R1).
107. K. Hansen, S. Mika, T. Schroeter, A. Sutter, A. Laak, S. T. Hartmann, N. Heinrich and K. R. Muller, Benchmark data set for *in-silico* prediction of Ames mutagenicity, *J. Chem. Inf. Model.*, 2009, **49**, 2077–2081, DOI: 10.1021/ci900161g.
108. K. L. Dobo, N. Greene, C. Fred, S. Glowienke, J. S. Harvey, C. Hasselgren, R. Jolly, M. O. Kenyon, J. B. Munzner, W. Muster, R. Neft, M. V. Reddy, A. T. White and S. Weiner, *In silico* methods combined with expert knowledge rule out mutagenic potential of pharmaceutical impurities: an industry survey, *Regul. Toxicol. Pharmacol.*, 2012, **62**, 449–455, DOI: 10.1016/j.yrtph.2012.01.007.
109. N. L. Kruhlak, R. D. Benz, H. Zhou and T. J. Colatsky, (Q)SAR modeling and safety assessment in regulatory review, *Clin. Pharmacol. Ther.*, 2012, **91**, 529–534, DOI: 10.1038/clpt.2011.300.
110. R. T. Naven, N. Greene and R. V. Williams, Latest advances in computational genotoxicity prediction, *Expert Opin. Drug Metab. Toxicol.*, 2012, **8**, 1579–1587, DOI: 10.1517/17425255.2012.724059.
111. M. W. Powley, (Q)SAR assessments of potentially mutagenic impurities: A regulatory perspective on the utility of expert knowledge and data submission, *Regul. Toxicol. Pharmacol.*, 2015, **71**, 295–300, DOI: 10.1016/j.yrtph.2014.12.012.
112. T. W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D. J. Knight, C. Mahony, M. Schwarz, A. White and M. T. D. Cronin, A strategy for structuring and reporting a read-across prediction of toxicity, *Regul. Toxicol. Pharmacol.*, 2015, **72**, 586–601, DOI: 10.1016/j.yrtph.2015.05.016.

## CHAPTER 10

# *Data Sources for Herbal and Traditional Medicines*

HSUEH-FEN JUAN<sup>a</sup>

<sup>a</sup>Department of Life Science, Institute of Molecular & Cellular Biology, Institute of Biomedical Electronics & Bioinformatics, Genome and Systems Biology Degree Program, Center for Systems Biology, National Taiwan University, No. 1, Section 4, Roosevelt Rd., Da'an District, Taipei 10617, Taiwan

\*E-mail: [yukijuan@ntu.edu.tw](mailto:yukijuan@ntu.edu.tw)

## 10.1 Introduction

Traditional herbal medicines are naturally occurring, plant-derived substances with minimal or no industrial processing that have been used to treat illness within local or regional healing practices.<sup>1</sup> Chinese herbs are important components of traditional Chinese medicine (TCM), which has been used for thousands of years as a major preventive and therapeutic strategy against disease.<sup>2</sup> Currently, more than 3200 species of medicinal plants are used as TCM.<sup>2</sup> A fundamental feature of TCM is compound formulas, which are composed of many kinds of herbs and sometimes minerals or animal components, similar to a cocktail therapy.<sup>3</sup> Each TCM compound formula is usually designed to combat specific symptoms and combined with other herbs or prescriptions tailored to individual needs. Herbal extracts have been investigated for use in treating various diseases and have been used as a complementary or alternative form of medical therapy for cancer patients.<sup>4–6</sup>

---

Issues in Toxicology No. 31

Computational Systems Pharmacology and Toxicology

Edited by Dale E. Johnson and Rudy J. Richardson

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)



In other studies that focus on chronic kidney disease,<sup>7</sup> neurodegenerative disease,<sup>8</sup> and diabetes mellitus,<sup>9</sup> Chinese herbs have been reported to alleviate symptoms and mediate signal transduction. The Nobel Prize in Physiology or Medicine 2015 was awarded to Tu Youyou "for her discoveries concerning a novel therapy against Malaria". Her Nobel lecture was entitled "Artemisinin—a Gift from Traditional Chinese Medicine to the World".<sup>10</sup> This indicates that TCM has become significant, receiving attention in health debates around the world.

Systems biology, which combines computational and experimental approaches to analyze complex biological systems, focuses on understanding functional activities from a systems-wide perspective.<sup>11</sup> With the advent of sequencing and computational methods, the application of omics data such as genomics, proteomics, and metabolomics has become a viable approach for studying the molecular mechanism and components of TCM as well as improving our knowledge of health and disease. In this chapter, I focus on introducing the TCM databases, as shown in Table 10.1. Additionally, I introduce the omics data in TCM.

## 10.2 TCM Databases

Here, I introduce the following six TCM databases: (1) Chem-TCM (Chemical Database of Traditional Chinese Medicine);<sup>12</sup> (2) HIT (Linking Herbal Active Ingredients to Targets);<sup>13</sup> (3) TCMSP (Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform);<sup>14</sup> (4) TCMGeneDIT (a database for associated TCM, gene and disease information using text mining);<sup>15</sup> (5) TCMID (Traditional Chinese Medicine Integrative Database for herb molecular mechanism analysis);<sup>16</sup> and (6) TTD (Therapeutic Target Database).<sup>17</sup>

### 10.2.1 Chem-TCM (Chemical Database of Traditional Chinese Medicine)

Chem-TCM<sup>12</sup> was launched in October 2011 by researchers at King's College London in collaboration with the Shanghai Institute of Materia Medica. This database lists 12 070 chemical records, constituents of approximately 350 herbs used in TCM. There are >9500 unique molecular records adjusted for overlapping presence in multiple plants and different stereochemistry.<sup>12</sup> Chem-TCM contains four parts: chemical information, botanical information, Western therapeutic targets, and assessment of molecular activity according to 26 TCM categories (Table 10.2).

### 10.2.2 HIT (Linking Herbal Active Ingredients to Targets)

HIT<sup>13</sup> is a curated database to complement available resources on protein targets for United States Food and Drug Administration-approved drugs and promising precursors. It currently contains ~1301 known protein targets

**Table 10.1** Databases for herbal and traditional medicines.

Database	Website	Description <sup>a</sup>	New release
Chem-TCM: Chemical Database of Traditional Chinese Medicine	<a href="http://www.chemtcm.com/">www.chemtcm.com/</a>	The Chem-TCM database gathers 12 070 chemical records, constituents of ~350 herbs used in traditional Chinese medicine. There are >9500 unique molecular records adjusted for overlapping presence in multiple plants and different stereochemistry	2011
HIT: linking herbal active ingredients to targets	<a href="http://lifecenter.sgst.cn/hit/">http://lifecenter.sgst.cn/hit/</a>	HIT is a comprehensive and fully curated database to complement available resources on protein targets for FDA-approved drugs, as well as promising precursors	2011
TCMSP: Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform	<a href="http://lsp.nwsuaf.edu.cn/tcmsp.php">http://lsp.nwsuaf.edu.cn/tcmsp.php</a>	TCMSP is a unique systems pharmacology platform of Chinese herbal medicines that captures the relationships between drugs, targets and diseases	2014
TCMGeneDIT: a database for associated TCM, gene and disease information using text mining	<a href="http://tcm.lifescience.ntu.edu.tw/index.html">http://tcm.lifescience.ntu.edu.tw/index.html</a>	TCMGeneDIT is a database system providing association information about TCMs, genes, diseases, TCM effects, and TCM ingredients automatically mined from vast amount of biomedical literature	2008
TCMID: Traditional Chinese Medicine Integrated Database for herb molecular mechanism analysis	<a href="http://www.megabionet.org/tcmid/">www.megabionet.org/tcmid/</a>	TCMID is a comprehensive database to provide information and bridge the gap between TCM and modern life sciences	2012
TTD: Therapeutic Targets Database	<a href="http://bidd.nus.edu.sg/group/ttd/ttd.asp">http://bidd.nus.edu.sg/group/ttd/ttd.asp</a>	TTD Version 5.1.01 provides 1008, 369, and 119 nature-derived approved, clinical trial, and preclinical drugs, respectively, together with their species origin information. All data are available for users to download	2016

<sup>a</sup>Information is taken from the homepage of each database. FDA: Food and Drug Administration.

**Table 10.2** Chem-TCM built the database according to 26 traditional Chinese medicine categories.

1	Release the exterior (warm, acrid herbs)
2	Release the exterior (cool, acrid herbs)
3	Drain fire
4	Cool the blood
5	Clear heat and relieve toxicity
6	Purgatives
7	Cathartics
8	Drain dampness
9	Dispel wind-damp
10	Cool and transform phlegm-heat
11	Warm and transform phlegm-cold
12	Treat cough and wheezing
13	Aromatically transform damp
14	Regulate the Qi
15	Stop bleeding
16	Invigorate the blood
17	Warm the interior and expel cold
18	Tonify the Qi
19	Tonify the blood
20	Tonify the Yang
21	Tonify the Yin
22	Stabilize and bind (astringents)
23	Nourish the heart and calm the spirit
24	Aromatically clear the orifices of the heart
25	Extinguish wind and stop tremors
26	Expel parasites

derived from 3258 items of literature, which covers ~586 active compounds from 1305 Chinese herbs. The molecular target information involves those proteins being directly/indirectly activated or inhibited, protein binders, and enzymes whose substrates or products are those compounds. Detailed interaction values such as  $IC_{50}$  and  $K_d/K_i$  are also provided. HIT also includes genes that are up- or down-regulated under the treatment of individual ingredients.

The homepage (<http://lifecenter.sgst.cn/hit/>) of HIT is shown in Figure 10.1A. It may be searched by compound, herb, and protein target in Chinese or English. For example, keying in “甘草” or “Radix Glycyrrhizae” as the key word in “HERB”, will produce the search result page as shown in Figure 10.1B. To discover what kind of compounds or herbs target a protein of interest, keying the protein name in “PROTEIN” will list which compound will target the protein directly or indirectly, and which herbs contain the compound.

### 10.2.3 TCMSP (Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform)

The traditional Chinese medicine systems pharmacology database and analysis platform (TCMSP) was built based on the framework of systems pharmacology for herbal medicines.<sup>14</sup> It consists of all the 499 Chinese herbs

A

HomeBrowseSimilarityStatisticsContact UsHelp

# HIT

## Herbal Ingredients' Targets Database

### Introduction

HIT is a comprehensive and fully curated database to complement available resources on protein targets for FDA-approved drugs as well as the promising precursors. It currently contains about 1,301 known protein targets(221 proteins are described as direct targets) derived from more than 3,250 literatures, which covers about 586 active compounds from more than 1,300 reputable Chinese herbs. The molecular target information involves those proteins being directly/indirectly activated or inhibited, protein binders, and enzymes whose substrates or products are those compounds. Detailed interaction values such as IC50 and Kd/Ki are collected if possible. Those up or down regulated genes are also included under the treatment of individual ingredients.



### Keyword Search

COMPOUND  
eg: EGCG; CAS:989-51-5; CID:5280489; flavone

HERB  
eg: Radix Glycyrrhizae; gan cao; 甘草

PROTEIN  
eg: Caspase-3; P42574; hsa:2534; GO:0070330; PF07714

FULL TEXT

SearchReset

**Claim:** HIT is strictly for academic use only. Any commercial application should contact us at [zwcao@tongji.edu.cn](mailto:zwcao@tongji.edu.cn)

**Citation:** Hao Ye, Li Ye, Hong Kang, Duanfeng Zhang, Lin Tao, Kailin Tang, Xueping Liu, Ruixin Zhu, Qi Liu, Y. Z. Chen, Yixue Li and Zhiwei Cao. (2011) HIT: linking herbal active ingredients to targets. *Nucleic Acids Res.*, 39, D1055–D1059. [PDF]

26230

B

HomeBrowseCompound SimilarityTarget SimilarityStatisticsContact UsHelp

# Herbal Ingredients' Targets Database

## Herb Search Result

HERB DETAILS	
Herb	Radix Glycyrrhizae(gan cao'甘草) <a href="#">Link to TCM-ID</a>
Function	To reinforce the function of the spleen and replenish qi, to remove heat and counteract toxicity, to dispel phlegm and relieve cough, to alleviate spasmodic pain, and to moderate drug actions.
Ingredients	<div>nicotiflorin [C0113]</div> <div>monosodium glycyrrhizinate [C0120]</div> <div>liquiritin [C0131]</div> <div>glycyrrhizic acid [C0267]</div> <div>glycyrrhizin [C0270]</div> <div>sinapic acid [C0320]</div> <div>formononetin [C0340]</div> <div>18alpha-glycyrrhetic acid [C0472]</div> <div>glycyrrhetic acid [C0473]</div> <div>18beta-glycyrrhetic acid [C0474]</div> <div>rutin [C0514]</div>
Herb	Glycyrrhiza uralensis Fisch(gan cao jie'甘草节)
Function	To reinforce the function of the spleen and replenish qi, to remove heat and counteract toxicity, to dispel phlegm and relieve cough, to alleviate spasmodic pain, and to moderate drug actions.
Ingredients	<div>nicotiflorin [C0113]</div> <div>monosodium glycyrrhizinate [C0120]</div> <div>liquiritin [C0131]</div> <div>glycyrrhizic acid [C0267]</div> <div>glycyrrhizin [C0270]</div> <div>sinapic acid [C0320]</div> <div>formononetin [C0340]</div> <div>18alpha-glycyrrhetic acid [C0472]</div> <div>glycyrrhetic acid [C0473]</div> <div>18beta-glycyrrhetic acid [C0474]</div> <div>rutin [C0514]</div>
Herb	Radix Glycyrrhizae uralensis(gan cao shao'甘草梢) <a href="#">Link to TCM-ID</a>
Function	To reinforce the function of the spleen and replenish qi, to remove heat and counteract toxicity, to dispel phlegm and relieve

Figure 10.1 (A) The homepage of HIT and (B) search results.

Published on 01 March 2017 on http://pubs.rsc.org | doi:10.1039/9781782623731-00243

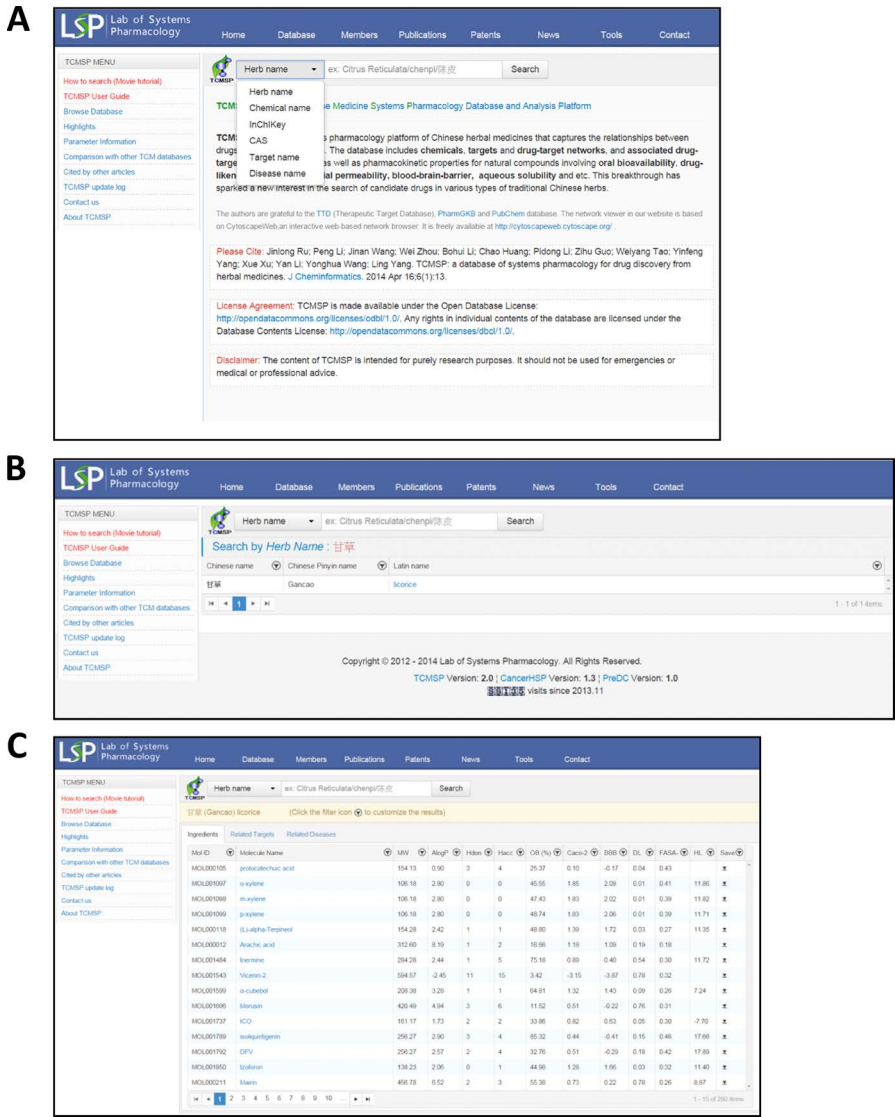
registered in the Chinese pharmacopoeia, with 29384 ingredients, 3311 targets and 837 associated diseases.<sup>14</sup> It is divided into three major categories: (1) compounds, targets, and diseases information; (2) herbal ingredients with their ADME (absorption, distribution, metabolism, and excretion)-related properties; and (3) compound–target relationships and target–disease relationships. TCMSP not only provides 12 important ADME-related properties such as human oral bioavailability, half-life, drug-likeness, Caco-2 permeability, blood–brain barrier status, and Lipinski’s rule of five for drug screening and evaluation, but also physicochemical parameters such as hydrogen bond donor, hydrogen bond acceptor, molecular weight, fractional negative accessible surface area, topological polar surface area, and number of rotatable bonds.

The homepage (<http://lsp.nwsuaf.edu.cn/tcmsp.php>) of TCMSP is shown in Figure 10.2A. It shows a choice of six categories including herb name, chemical name, InChiKey, Chemical Abstracts Service (CAS) registry number, target name, and disease name from the manual bar. For example, choosing “herb name” and inputting “甘草” or “Radix Glycyrrhizae” as the query, will obtain the search result as shown in Figure 10.2B. Detailed information includes the ingredients, related targets, and related diseases of Radix Glycyrrhizae (Figure 10.2C). Additionally, the drug–target or drug–disease networks can be gathered (Figure 10.3A and B), which will help to reveal the mechanisms of action of Chinese herbs, uncover the nature of TCM theory and develop new herb-oriented drugs.

#### 10.2.4 TCMGeneDIT (A Database for Associated TCM, Gene and Disease Information Using Text Mining)

TCMGeneDIT is a database system providing association information about TCM, genes, diseases, TCM effects, and TCM ingredients automatically mined from vast amount of biomedical literature.<sup>15</sup> Integrated protein–protein interaction and biological pathways information collected from public databases are also available.<sup>15</sup> In addition, the transitive relationships among genes, TCMS, and diseases could be inferred through the shared intermediates. Furthermore, TCMGeneDIT is useful in deducing possible synergistic or antagonistic contributions of the prescription components to the overall therapeutic effects. TCMGeneDIT is a unique database of various association information about TCMS. The database integrates TCMS with life sciences and biomedical studies facilitates modern clinical research and the understanding of therapeutic mechanisms of TCMS and gene regulations.

The homepage (<http://tcm.lifescience.ntu.edu.tw/index.html>) of TCMGeneDIT is shown in Figure 10.4A. Information can be searched by TCMS and genes, diseases or TCM effects. For example, inputting “Radix Glycyrrhizae” as the query in TCM, produces the search result shown in Figure 10.4B. Further information may be obtained about the ingredients, and putative relationships between TCM effects and sources extracted from the literature.

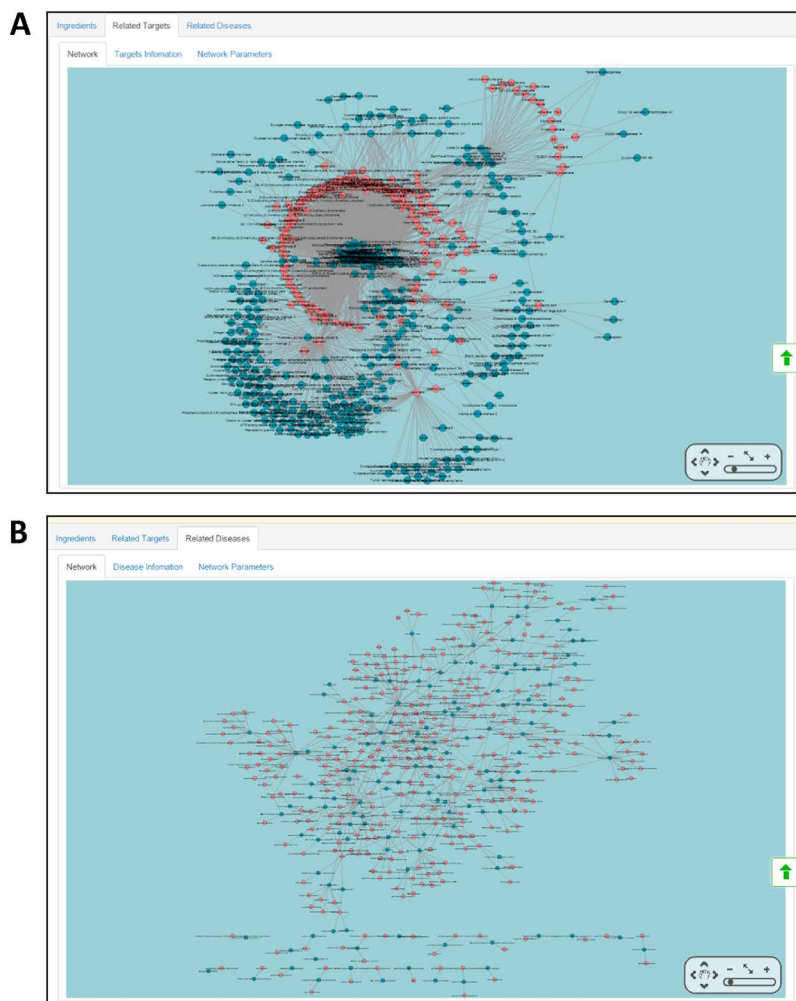


**Figure 10.2** (A) The homepage of TCMSP, (B) search results and (C) the ingredients, related targets and related diseases of Radix Glycyrrhizae.

### 10.2.5 TCMID (Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis)

TCMID is composed of six data fields, namely prescriptions, herbs, ingredients, targets, drugs and diseases.<sup>16</sup> The information and data in those fields were integrated from related web-based databases and text mining of books and published articles.<sup>16</sup> Currently, TCMID contains 8159 herbs, 25 210



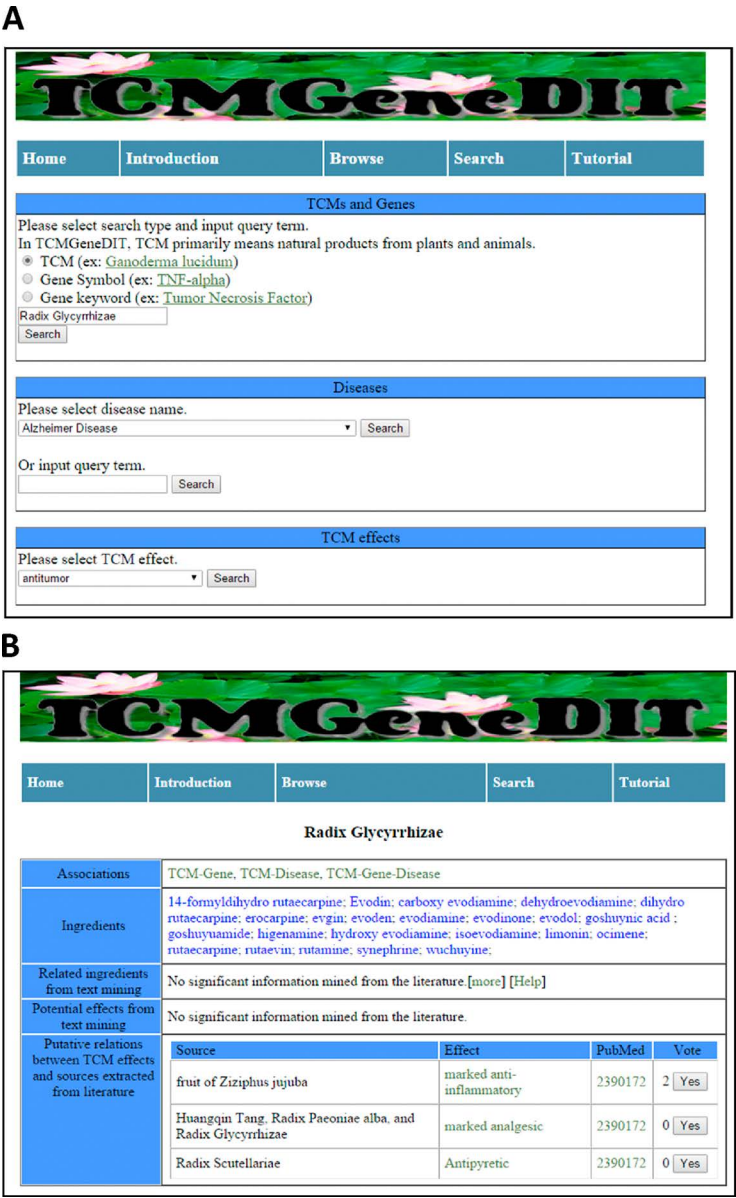


**Figure 10.3** (A) Drug–target and (B) drug–disease networks in TCMSP.

ingredients, 6826 drugs, 17 521 targets, 46 914 prescriptions, and 3791 diseases. Furthermore, TCMID provides network displays such as herb–disease network, herbal ingredient–target interaction network, and herbal ingredient–target–disease–drug network, which will facilitate the study of combination therapy and understanding of the underlying mechanisms for TCM at molecular level.<sup>16</sup>

The homepage (<http://www.megabionet.org/tcmid/>) of TCMID is shown in Figure 10.5A. The information may be searched by prescriptions, herbs, ingredients, targets, drugs or diseases. Inputting “Radix Glycyrrhizae” as the query in TCM will not generate a result. The Pinyin name “gan cao” will generate the search result as shown in Figure 10.5B. Herb–disease network data may be downloaded from TCMID.



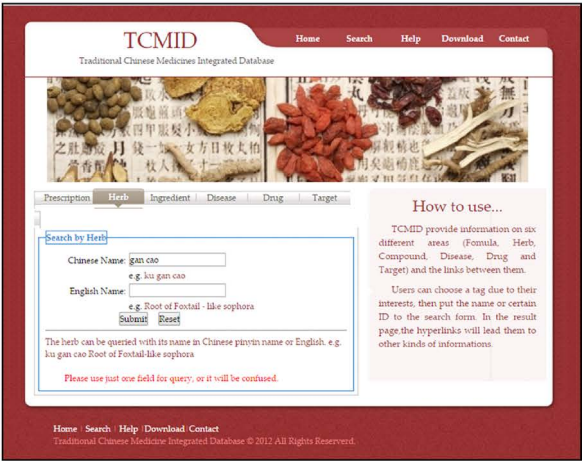


**Figure 10.4** (A) The homepage of TCMGeneDIT and (B) search results.

### 10.2.6 TTD (Therapeutic Target Database)

TTD Version 5.1.01 provides 397 successful and 1469 research targets, and 2071 approved and 17 803 investigative drugs.<sup>17</sup> Among these, it contains 1008 approved drugs derived from nature, 369 clinical trial drugs and 119 pre-clinical drugs, together with their species origin information. Furthermore,

A



B



Figure 10.5 (A) The homepage of TCMID (B) search results.

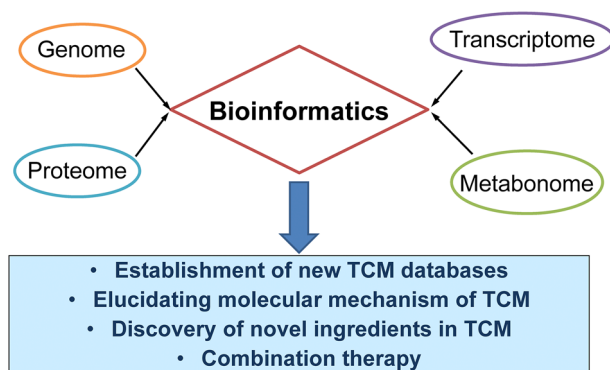
it includes cross-links of most TTD target and drug entries to the corresponding pathway entries of the Kyoto Encyclopedia of Genes and Genomes, MetaCyc/BioCyc, NetPath, PANTHER (Protein Analysis Through Evolutionary Relationships) pathway, the Pathway Interaction Database (PID), PathWhiz, Reactome and WikiPathways, as well as convenient access to multiple targets and drugs cross-linked to each of these pathway entries.<sup>17</sup> TTD is accessible at <http://bidd.nus.edu.sg/group/ttd/ttd.asp>. All data are available for users to download.

### 10.3 Omics Data in TCM

High-throughput data can be used to provide a comprehensive inventory of all the biological processes of cells, display their complexity and increase data accuracy.<sup>18</sup> An accurate picture of the differential expression of experimental samples is important for defining precise targets and networks.<sup>18</sup> An increasing number of “-omics” methods, such as genomics, transcriptomics, proteomics and metabonomics, are gradually being adopted by TCM researchers.<sup>16</sup> Integrating various omics data and analyzing the data using bioinformatics will assist in establishing new TCM databases, which will facilitate searching prescriptions, herbs, ingredients, targets, drugs and diseases described above, as well as for elucidating the molecular mechanism of TCM, discovery of novel ingredients in TCM and combination therapy (Figure 10.6). Here, I review the data resources from omics studies.

#### 10.3.1 Genomics in TCM

With rapid advances in high-throughput sequencing technologies and greatly reduced costs, a new discipline called “herbal genomics” or TCM genomics has emerged.<sup>19,20</sup> Researchers are now systematically categorizing medicinal herbs by sequencing, assembling, and annotating their



**Figure 10.6** The benefits to TCM of integrating various omics data and analyzing the data by bioinformatics.

genomes, and by analyzing their genes' functions.<sup>19</sup> Among these, *Ganoderma lucidum* (ChiZhi) is one of the best known medical macrofungi. *G. lucidum* is more than just an ordinary fungus; it has long been used in traditional Chinese medicinal remedies and for the promotion of health and longevity in many Asian countries.<sup>21,22</sup> The *G. lucidum* resource includes genome assembly, 16 113 gene models and their corresponding gene function annotations, and transcriptomic data from three different developmental stages, namely, mycelia, primordial, and fruiting bodies.<sup>23</sup> It also presents predicted gene clusters and chemical compounds identified in *G. lucidum*.<sup>23</sup>

The purpose of TCM genomics is to study the genome of prescriptions and to discover the molecular mechanism of TCM acting on the human genome. It prescriptions precise and simple, and pushes TCM to the molecular level or finds the gene for active ingredients.<sup>20</sup> One TCM genomics database, the Medicinal Plant Metabolomics Resource, includes 14 taxonomically diverse medicinal plant species. This information is summarized in Table 10.3. The database contains not only the genome information of 14 taxonomically diverse medicinal plant species, but also transcriptome and metabolome data. Users can download the data directly from the website ([http://met-netweb.gdcb.iastate.edu/mpmr\\_public/](http://met-netweb.gdcb.iastate.edu/mpmr_public/)).

### 10.3.2 Transcriptomics in TCM

The transcriptome can be measured either by microarrays or next-generation sequencing. Many transcriptional profiles have been mined to search for target molecules of TCM extract treatments. These results were used to evaluate whether TCM extracts could be used as complementary drugs to treat specific symptoms or diseases. Here, *G. lucidum*-treated transcriptomes in immunomodulatory and anti-cancer therapy are described.

*G. lucidum*, a well-known herb, contains an abundance of polysaccharides with immunostimulatory properties.<sup>24</sup> Transcriptional profiles of polysaccharides from *G. lucidum*-treated dendritic cells showed a decrease in the expression of some phagocytosis-related genes.<sup>25</sup> F3 (a polysaccharide fraction extracted from lingzhi)-treated human leukemia THP-1 cells have been found to undergo apoptosis through death receptor pathways as determined through microarray analysis, and induces macrophage-like differentiation by caspase cleavage and p53 activation in THP-1 cells.<sup>21,22</sup>

Transcriptome data should be deposited at the Gene Expression Omnibus (GEO), which is a public functional genomics data repository supporting MIAME (*minimum information about a microarray experiment*)-compliant data submissions.<sup>26</sup> GEO accepts array- and sequence-based data and provides tools to help users query and download experiments and curated gene expression profiles.<sup>26</sup> For example, the array data from F3-treated human leukemia THP-1 cells have been deposited to the GEO database and the series record is GSE16014. The data can be download freely and used for further analysis.

**Table 10.3** 14 taxonomically diverse medicinal plant species in the Medicinal Plant Metabolomics Resource.

Species	Common name	Description of drug uses <sup>a</sup>
<i>Atropa belladonna</i>	Deadly nightshade; belladonna	One of the most poisonous plants known, with all parts of the plant containing toxic tropane alkaloids
<i>Camptotheca acuminata</i>	Xi Shu; happy tree	A Chinese tree that produces the pentacyclic quinolines camptothecin and 10-hydroxycamptothecin through the monoterpene indole alkaloid pathway
<i>Cannabis sativa</i>	Marijuana; hemp; pot	Has acquired considerable importance as a medicinal plant all over the world
<i>Catharanthus roseus</i>	Rosy periwinkle	Used as a folk medicine, but was recognized by western physicians in the 1950s to produce compounds with anti-cancer activity
<i>Digitalis purpurea</i>	Common foxglove	Used as a folk medicine, but was reported to contain a cardio-active compound in 1785
<i>Dioscorea villosa</i>	Wild yam	Phyto-estrogenic properties for post-menopausal women
<i>Echinacea pupurea</i>	Eastern purple coneflower	Used as a folk medicine, but recognized to produce compounds that stimulate the immune response
<i>Ginkgo biloba</i>	Ginkgo; maidenhair tree	Used as a vasodilator to help improve cognitive function/Alzheimer's disease, reduction of cerebral ischemia during stroke, reduce varicose conditions, as a treatment for Raynaud's syndrome and as an aid to peripheral blood flow
<i>Hoodia gordonii</i>	Hoodia	Used to suppress appetite and thirst; as a dietary supplement weight-loss product
<i>Hypericum perforatum</i>	St John's wort	Antiretroviral, anti-cancer, anti-depressive, anti-autoimmune
<i>Panax quinquefolius</i>	American ginseng	Stimulant, stress reducer, enhances cognitive function and mental awareness
<i>Rauvolfia serpentina</i>	Serpentwood; Indian snakeroot	E.g. The treatment of snake bites, fever and insanity
<i>Rosmarinus officinalis</i>	Rosemary	Rosemary leaves contains contain 20–30 different monoterpenes with antifungal and antimicrobial activity
<i>Valeriana officinalis</i>	Common valerian; garden heliotrope	Anxiolytic, sedative, central nervous system depressant and sleep aid

<sup>a</sup>Descriptions have been extracted from the Medicinal Plant Metabolomics Resource website. Greater detail is available therein.

### 10.3.3 Proteomics in TCM

The term proteome—analogue to genome—was first introduced in 1996, defined as the entire complement of proteins expressed in a specific state of a cell, tissue, or an organism.<sup>27</sup> Despite the similarities between the genome, transcriptome, and proteome, the proteomic profiling of a subject cannot be directly transferred from its genome or transcriptome. High-throughput protein identification and quantification analysis based on mass spectrometry are evolving at a rapid pace. State-of-the-art mass spectrometry provides a platform to identify complicated proteome with high sensitivity at a relatively low cost and high reproducibility.<sup>28,29</sup>

The ProteomeXchange (PX) consortium ([www.proteomexchange.org](http://www.proteomexchange.org)) was formed in 2006 to store proteomic raw data, comprising primary (PRIDE, PASSSEL) and secondary resources (PeptideAtlas, UniProt) and represent the journals regularly publishing proteomics data.<sup>30,31</sup> It was established to provide a coordinated submission of mass spectrometry-based proteomics data to the main existing proteomics repositories, and to encourage optimal data dissemination.<sup>32</sup>

Tanshinone IIA (TIIA) is an example: TIIA, a diterpene quinone extracted from the plant Danshen (*Salvia miltiorrhiza*), is used for the prevention of cardiac disease,<sup>33</sup> protection of the nervous system<sup>34</sup> and hepatocytes,<sup>35</sup> and inhibition of osteoporosis.<sup>36</sup> Proteomic expression profiling of TIIA-treated gastric cells was performed using isobaric tags for relative and absolute quantification (iTRAQ)-based quantitative proteomics analysis.<sup>18</sup> The mass spectrometry proteomics data were uploaded to the ProteomeXchange Consortium *via* the PRIDE partner repository, with the data set identifier PXD000998 and DOI 10.6019/PXD000998. Researchers can freely download and re-analyze the data.

### 10.3.4 Metabonomics in TCM

Metabolomics involves the study of targeted small molecule metabolites (<1500 Da). In 1998, the metabolome was first introduced in the elucidation of yeast gene function.<sup>37</sup> However, the application of this concept can be traced back to the development of traditional medicine; while metabonomics chemically tracks metabolites in urine, feces, and so forth, traditional medicine used color, smell, and taste to facilitate diagnoses.<sup>2</sup> Over genomics and proteomics, metabolomics can provide a more solid link between genotype and phenotype.<sup>2</sup>

As described earlier, the Medicinal Plant Metabolomics Resource contains not only the genome information of 14 taxonomically diverse medicinal plant species, but also transcriptome and metabolome data. Researchers interested in these plant metabolomes can access the Medicinal Plant Metabolomics Resource website. Researchers interested in human metabolomes after treatment with TCM can access the Human Metabolome Database (HMDB).<sup>38</sup> HMDB is a freely available electronic database containing

detailed information about small molecule metabolites found in the human body.<sup>38</sup> The database contains 41 993 metabolite entries, including both water-soluble and lipid-soluble metabolites as well as metabolites that would be regarded as either abundant ( $>1\ \mu\text{M}$ ) or relatively rare ( $<1\ \text{nM}$ ).<sup>39</sup>

## 10.4 Summary

With the development of integrative medicine, the integration of TCM and modern technologies is increasing rapidly. Currently, the major focus of research is divided into two: understanding the mechanisms of TCM from the systems biology perspective and facilitating novel drug design based on the analysis of TCM herbal medicine.<sup>40</sup> Although massive amounts of research data all over the world are private and isolated, which will restrict the communication of the community and weaken the power of integrative medicine, research communities such as GEO and ProteomeXchange provide excellent platforms for researchers to have a stable repository for raw genomic, transcriptomic and mass spectrometry proteomics data.

This chapter introduces data sources for herbal and traditional medicines, especially TCM. The TCM databases provide readers with the prescriptions, ingredients, and targets of these TCM substances. Additionally, through analyzing omics data such as the genome, transcriptome, and proteome as well as the metabolome, the molecular mechanisms of TCM in human diseases have been revealed rapidly. However, the effects or molecular functions of a lot of TCM therapies are still unknown. Further research effort is needed into TCM.

## Acknowledgements

This work was supported by the National Taiwan University (NTU) Cutting-Edge Steering Research Project (NTU-CDP-105R7717), Ministry of Science and Technology (NSC 102-2628-B-002-041-MY3) and 2015 Collaboration Research Project between NTU Hospital and NTU Colleges (UN105-016).

## References

1. WHO, [www.who.int/bulletin/volumes/86/8/07-042820/en/](http://www.who.int/bulletin/volumes/86/8/07-042820/en/).
2. L. L. Lin, Y. H. Wang, C. Y. Lai, C. L. Chau, G. C. Su, C. Y. Yang, S. Y. Lou, S. K. Chen, K. H. Hsu, Y. L. Lai, W. M. Wu, J. L. Huang, C. H. Liao and H. F. Juan, Systems biology of meridians, acupoints, and Chinese herbs in disease, *J. Evidence-Based Complementary Altern. Med.*, 2012, **2012**, 372670.
3. X. Xu, K. You, X. Bao and K. Kang, *Essentials of Traditional Chinese Medicine English-Chinese Encyclopedia of Practical TCM*, Higher Education Press, Beijing, China, 1991.
4. M. A. Richardson, T. Sanders, J. L. Palmer, A. Greisinger and S. E. Singletary, Complementary/alternative medicine use in a comprehensive cancer center and the implications for oncology, *J. Clin. Oncol.*, 2000, **18**, 2505–2514.



5. S. Gözümlü, A. Tezel and M. Koc, Complementary alternative treatments used by patients with cancer in eastern Turkey, *Cancer Nurs.*, 2003, **26**, 230–236.
6. O. DePeyer, Herbal extracts help fight cancer, *Nat. News*, 2000, DOI: 10.1038/news000831-1.
7. K. Wojcikowski, D. W. Johnson and G. Gobe, Herbs or natural substances as complementary therapies for chronic kidney disease: ideas for future studies, *J. Lab. Clin. Med.*, 2006, **147**, 160–166.
8. S. Y. Su, C. Y. Cheng, T. H. Tsai, C. Y. Hsiang, T. Y. Ho and C. L. Hsieh, Paeonol attenuates H<sub>2</sub>O<sub>2</sub>-induced NF- $\kappa$ B-associated amyloid precursor protein expression, *Am. J. Chin. Med.*, 2010, **38**, 1171–1192.
9. G. C. Han, S. K. Ko, J. H. Sung and S. H. Chung, Compound K enhances insulin secretion with beneficial metabolic effects in db/db mice, *J. Agric. Food Chem.*, 2007, **55**, 10641–10648.
10. [www.nobelprize.org/nobel\\_prizes/medicine/laureates/2015/tu-lecture-slides.pdf](http://www.nobelprize.org/nobel_prizes/medicine/laureates/2015/tu-lecture-slides.pdf).
11. H. F. Juan and H. C. Huang, *System Biology: Applications in Cancer-Related Research*, World Scientific, Singapore, 2012.
12. <http://www.chemtem.com/database.html>.
13. H. Ye, L. Ye, H. Kang, D. Zhang, L. Tao, K. Tang, X. Liu, R. Zhu, Q. Liu, Y. Z. Chen, Y. Li and Z. Cao, HIT: linking herbal active ingredients to targets, *Nucleic Acids Res.*, 2011, **39**, D1055–D1059.
14. J. Ru, P. Li, J. Wang, W. Zhou, B. Li, C. Huang, P. Li, Z. Guo, W. Tao, Y. Yang, X. Xu, Y. Li, Y. Wang and L. Yang, TCMSP: a database of systems pharmacology for drug discovery from herbal medicines, *J. Cheminf.*, 2014, **6**, 13.
15. Y. C. Fang, H. C. Huang, C. C. Chen and H. F. Juan, TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining, *BMC Complementary Altern. Med.*, 2008, **8**, 58.
16. R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen and T. Shi, TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis, *Nucleic Acids Res.*, 2013, **41**, D1089–D1095.
17. H. Yang, C. Qin, Y. H. Li, L. Tao, J. Zhou, C. Y. Yu, F. Xu, Z. Chen, F. Zhu and Y. Z. Chen, Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information, *Nucleic Acids Res.*, 2016, **44**, D1069–D1074.
18. L. L. Lin, C. R. Hsia, C. L. Hsu, H. C. Huang and H. F. Juan, Integrating transcriptomics and proteomics to show that tanshinone IIA suppresses cell growth by blocking glucose metabolism in gastric cancer cells, *BMC Genomics*, 2015, **16**, 41.
19. S. Chen, J. Song, C. Sun, J. Xu, Y. Zhu, R. Verpoorte and T. P. Fan, Herbal genomics: examining the biology of traditional medicines, *Science*, 2015, **347**(suppl. 6219), S27–S29.
20. M. Yuan, R. Wang, X. Wu, Y. An and Z. Sun, TCM genomics—a key factor to modernization of traditional Chinese medicine, *J. Single Cell Genomics Proteomics*, 2012, **1**, e102.

21. J. W. Hsu, H. C. Huang, S. T. Chen, C. H. Wong and H. F. Juan, *Ganoderma lucidum* polysaccharides induce macrophage-like differentiation in human leukemia THP-1 cells *via* caspase and p53 activation, *J. Evidence-Based Complementary Altern. Med.*, 2011, **2011**, 358717.
22. K. C. Cheng, H. C. Huang, J. H. Chen, J. W. Hsu, H. C. Cheng, C. H. Ou, W. B. Yang, S. T. Chen, C. H. Wong and H. F. Juan, *Ganoderma lucidum* polysaccharides in human monocytic leukemia cells: from gene expression to network construction, *BMC Genomics*, 2007, **8**, 411.
23. S. Chen, J. Xu, C. Liu, Y. Zhu, D. R. Nelson, S. Zhou, C. Li, L. Wang, X. Guo, Y. Sun, H. Luo, Y. Li, J. Song, B. Henrissat, A. Levasseur, J. Qian, J. Li, X. Luo, L. Shi, L. He, L. Xiang, X. Xu, Y. Niu, Q. Li, M. V. Han, H. Yan, J. Zhang, H. Chen, A. Lv, Z. Wang, M. Liu, D. C. Schwartz and C. Sun, Genome sequence of the model medicinal mushroom *Ganoderma lucidum*, *Nat. Commun.*, 2012, **3**, 913.
24. D. Sliva, Cellular and physiological effects of *Ganoderma lucidum* (Reishi), *Mini-Rev. Med. Chem.*, 2004, **4**, 873–879.
25. Y. L. Lin, S. S. Lee, S. M. Hou and B. L. Chiang, Polysaccharide purified from *Ganoderma lucidum* induces gene expression changes in human dendritic cells and promotes T helper 1 immune response in BALB/c mice, *Mol. Pharm.*, 2006, **70**, 637–644.
26. [www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/).
27. M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams and D. F. Hochstrasser, From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis, *Bio/Technology*, 1996, **14**, 61–65.
28. T. Nilsson, M. Mann, R. Aebersold, J. R. Yates 3rd, A. Bairoch and J. J. Bergeron, Mass spectrometry in high-throughput proteomics: ready for the big time, *Nat. Methods*, 2010, **7**, 681–685.
29. M. Mann and N. L. Kelleher, Precision proteomics: the case for high resolution and high mass accuracy, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 18132–18138.
30. H. Hermjakob and R. Apweiler, The Proteomics Identifications Database (PRIDE) and the ProteomeXchange Consortium: making proteomics data accessible, *Expert Rev. Proteomics*, 2006, **3**, 1–3.
31. J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dienes, Z. Sun, T. Farrar, N. Bandeira, P. A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H. J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones and H. Hermjakob, ProteomeXchange provides globally coordinated proteomics data submission and dissemination, *Nat. Biotechnol.*, 2014, **32**, 223–226.
32. [www.proteomexchange.org/](http://www.proteomexchange.org/).
33. M. Ding, G. R. Zhao, Y. J. Yuan and Z. X. Guo, Aqueous extract of *Salvia miltiorrhiza* regulates adhesion molecule expression of tumor necrosis factor alpha-induced endothelial cells by blocking activation of nuclear factor kappaB, *J. Cardiovasc. Pharmacol.*, 2005, **45**, 516–524.

34. Y. Ren, P. J. Houghton, R. C. Hider and M. J. Howes, Novel diterpenoid acetylcholinesterase inhibitors from *Salvia miltiorhiza*, *Planta Med.*, 2004, **70**, 201–204.
35. S. Wasser, J. M. Ho, H. K. Ang and C. E. Tan, *Salvia miltiorrhiza* reduces experimentally-induced hepatic fibrosis in rats, *J. Hepatol.*, 1998, **29**, 760–771.
36. H. H. Kim, J. H. Kim, H. B. Kwak, H. Huang, S. H. Han, H. Ha, S. W. Lee, E. R. Woo and Z. H. Lee, Inhibition of osteoclast differentiation and bone resorption by tanshinone IIA isolated from *Salvia miltiorrhiza* Bunge, *Biochem. Pharmacol.*, 2004, **67**, 1647–1656.
37. S. G. Oliver, M. K. Winson, D. B. Kell and F. Baganz, Systematic functional analysis of the yeast genome, *Trends Biotechnol.*, 1998, **16**, 373–378.
38. D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert, HMDB3.0–The Human Metabolome Database in 2013, *Nucleic Acids Res.*, 2013, **41**, D801–D807.
39. [www.hmdb.ca/](http://www.hmdb.ca/).
40. H. S. Parekh, G. Liu and M. Q. Wei, A new dawn for the use of traditional Chinese medicine in cancer therapy, *Mol. Cancer*, 2009, **8**, 21.

## CHAPTER 11

# *Network Pharmacology Research Approaches for Chinese Herbal Medicines*

DALE E. JOHNSON<sup>a,b</sup>

<sup>a</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA;

<sup>b</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Morgan Hall, Berkeley, CA 94720-3104, USA

\*E-mail: [daleej@umich.edu](mailto:daleej@umich.edu)

## 11.1 Introduction

Network pharmacology represents the integration of several multidisciplinary concepts including biochemical, bioinformatics, and systems biology.<sup>1–8</sup> Zhang *et al.*<sup>2</sup> have shown the advantages of this concept in traditional Chinese medicine (TCM) where the concept of a single target is replaced by a holistic view of multi-target effects which can be used to uncover combination effects of several active components of a formulation. From a drug discovery standpoint, this also includes the understanding of the regulation of signaling pathways with multiple channels, a potential efficacy increase in drug and drug–herb combinations, and the potential reduction in side effects.<sup>2,7,8</sup> Two main network pharmacology approaches are currently being used:

---

Issues in Toxicology No. 31

Computational Systems Pharmacology and Toxicology

Edited by Dale E. Johnson and Rudy J. Richardson

© The Royal Society of Chemistry 2017

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

- by establishing a network model through the utilization of online databases and tools, which would include both disease and chemical interaction nodes. This type of network model defines predictions of mechanism of action and potential synergy between different chemical entities and provides the opportunity to revise components within a multi-chemical formulation;<sup>2,8,9</sup>
- by combining the disease-target(s) model with high-throughput screening, targeted assays, and directed modeling to provide validation of the predictions or forecasts and also to probe different modified structures of phytochemical analogues.<sup>2,7,10,11</sup> Key compound components of these methods include the prediction of oral bioavailability and drug-likeness, along with projected target screening. Important rules relating to physicochemical properties and chemical structure include lipophilicity (octanol–water partition), molecular weight, molecular flexibility estimates calculated by the number of hydrogen bond donors and acceptors, number of rotatable bonds, and polar surface area.<sup>12</sup> As an example, Kumar *et al.*<sup>13</sup> have described a prediction model for oral bioavailability by a support vector machine-based kernel learning approach using physicochemical properties. Possible gastrointestinal absorption problems are alerted if any two of the following conditions are satisfied: molecular weight >500; number of hydrogen-bond acceptors >10; number of hydrogen-bond donors >5; calculated logP >5.0 (if ClogP is used) or >4.15 (if MlogP is used).<sup>14</sup>

One of the difficulties in predictive drug or phytochemical research is the estimated potency of the chemical at the biological target or at multiple targets proposed. Desirable potency at the target would be a high value for  $pIC_{50}$ , which reduces the risk of non-specific, off-target pharmacological effects and could allow for lower total doses. Yu *et al.*<sup>15</sup> designed a set of *in silico* tools incorporating chemical, genomic, and pharmacological information. In the framework, molecular descriptors were combined with structural and physicochemical property descriptors. Importantly for drug research, the typical process which follows in sequence after *in silico* prediction includes initial target identification and validation, assay development, high-throughput screening, hit identification, lead optimization, and selection of candidate clinical molecule(s). With TCM therapeutics, the starting point typically follows a physiology-based concept where herbs and/or concoctions have been used for certain diseases or conditions, and the disease phenotype and associated biological pathways are then explored with interactions with active phytochemicals. Zhang *et al.*<sup>2</sup> discuss the importance of database development in these processes, including the incorporation of several TCM databases available online (discussed below). Another aspect of TCM systems pharmacology developing currently is the definition of susceptible individuals: those most likely to respond positively or potentially negatively to a TCM concoction and/or TCM plus Western therapeutic(s). Some of these projections into personalized therapies will require an increased knowledge and

incorporation of pharmacogenomics and metabolic enzyme variances due to specific polymorphisms, particularly as it relates to difference seen in various ethnic groups.

### 11.1.1 Modernization of TCM

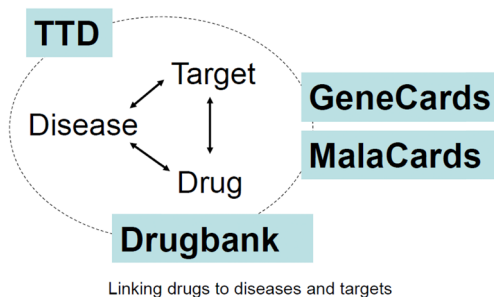
Xu *et al.*<sup>16</sup> reviewed a new era for the modernization of TCM, highlighting the interest and importance of TCM internationally. The authors discuss advances in TCM in three recent phases: phase I (1950s–1970s), focused on higher education, both scientific and economic foundations, and the establishment of hospital networks in China; phase II (1980s–2000s) was important for the development of legal, economic and scientific foundations for TCM and for the development of international networks; and phase III (2011 onwards) is focused on consolidating the scientific basis and international networks through multiple collaborations. This new era also has seen the launch of the Good Practice in Traditional Chinese Medicine Research in the Post-genomic Era (GP-TCM) project and the European Union's Seventh Framework Programme (FP7) coordination action.<sup>17</sup> The main objectives are:

- develop a European–Chinese network which would collaborate on functional genomics research in TCM;
- review the current practice of TCM research, identifying problems and practical solutions;
- propose prioritized areas for future TCM research; and
- develop online resources for the enhancement and support of pan-European studies of TCM research.

### 11.1.2 Concept of Network Pharmacology

As part of the modernization of TCM, the concept of network pharmacology has come to the forefront in TCM research.<sup>1–8</sup> Network pharmacology is a field based within systems biology, where the focus is on complex interactions in biological systems from a holistic basis, rather than on single drug targets.<sup>18</sup> Advancements in the basic understanding that drugs in several therapeutic classes act on multiple targets—polypharmacology—and in this aspect, the network-based view of drug action is constructed by mapping polypharmacology networks onto biological networks.<sup>19</sup> This concept is now fundamental in TCM, because various herbs consist of several phytochemicals and TCM concoctions consist of several plant and animal sources, that is, multiple chemical compounds. This concept is now widely accepted as essential in understanding Chinese herbal medicines from systems biology and bioinformatics standpoints.<sup>20</sup> A brief example as a starting point for this type of approach is illustrated in Figure 11.1, where four different freely available online sources are identified. These sources are also key information links present in several other online resources.

Therapeutics Targets Database:  
linking diseases to targets



**Figure 11.1** Target–drug–disease–gene connectivity.

- GeneCards<sup>21</sup> is a searchable, integrative database providing comprehensive, user-friendly information on all known and predicted human genes. It automatically integrates gene-centric data from more than 100 web sources, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information.
- MalaCards<sup>22</sup> is an integrated database of human diseases (maladies) and their annotations, modeled on the architecture of the GeneCards database. The MalaCards database is organized into "disease cards", each integrating prioritized information, including known aliases for each disease, inter-disease connections, symptoms, drugs, publication links, genes, clinical trials, and related diseases/disorders. As of April 2016, the database contained 19915 disease entries consolidated from 68 sources.
- The Therapeutic Target Database (TTD)<sup>23</sup> provides information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and the corresponding drugs directed at each target. The database also includes links to relevant databases that contain information about target function, sequence, 3D structure, ligand binding properties, enzyme nomenclature and drug structure, therapeutic class, and the clinical development status of each molecule. All information is fully annotated.
- The DrugBank<sup>24</sup> database is a bioinformatics and cheminformatics resource that combines detailed drug (*i.e.* chemical, pharmacological, and pharmaceutical) data with comprehensive drug target (*i.e.* sequence, structure, and pathway) information. As of April 2016, the database contained 8206 drug entries including 1991 United States Food and Drug Administration (FDA)-approved small molecule drugs, 207 FDA-approved biologic drugs, 93 nutraceuticals and >6000 experimental drugs. Additionally, >4000 non-redundant protein (*i.e.* drug target/enzyme/transporter/carrier) sequences are linked to the drug entries. DrugBank is the standard source of drug information for a majority of online resources in systems pharmacology.



As discussed in following sections, the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP)<sup>25</sup> is a platform that captures and predicts the relationships between drugs, targets, and diseases and uses the DrugBank<sup>24</sup> as a key resource. Essential parts of the database include chemicals, targets, drug–target networks, and associated drug–target–disease networks. Also included are pharmacokinetic-related properties of natural compounds including oral bioavailability, drug-likeness, intestinal permeability, blood–brain barrier permeability, and other physicochemical properties essential for the prediction of pharmacologic activity.

## 11.2 Network Pharmacology in TCM Research

There is an increasing interest in clarifying and understanding the mechanisms of therapeutic actions of compounds in TCM formulations and correlating these with *Zheng* disease subtypes and disease categories used worldwide for therapeutic interventions.<sup>8,10</sup>

Shi *et al.*<sup>26</sup> used the systems pharmacology approach to study *Bushenhuoxue* formula (BSHX), a traditional medicine used for the treatment of chronic kidney disease. The authors describe a stepwise approach to elucidate the network pharmacology process. This includes:

- (1) determine known targets and candidate genes for the specific disease;
- (2) determine phytochemical ingredients of TCM formulations or individual medicinal herbs;
- (3) construct a natural product–target network; confirm certain interactions with molecular docking;
- (4) construct protein–protein interaction networks and elucidate biological function analysis;
- (5) construct molecule–target networks, overlapping on the disease network.

In this publication, the authors showed that BSHX would have therapeutic effects through multi-channel network regulation, which includes regulating the coagulation and fibrinolytic balance, expression of inflammatory factors, and inhibiting abnormal extracellular matrix accumulation. They also indicate that through this process, the ingredients of BSHX—tanshinone IIA, rhein, curcumin, calycosin, and quercetin—are identified as the potential pharmacologically active entities. In a continuation of this work, Liang *et al.*<sup>27</sup> introduced new methods and employed them in analyzing *Liu Wei Di Huang*, a classic herbal medicine used to tonify the “Yin deficiency pattern” in TCM. Further clinical work with the formula established potential disease targets within hypertension and esophageal cancer. The authors used a discovery algorithm consisting of: herb formulation → herb chemical compound collection → chemical characters analysis → drug-likeness screening → target prediction → target selection and compound ranking → prediction of therapy

effects of herb formulation → understanding different effects of formula on same disease and understanding common effects of the formula on different diseases. Using a series of ranking efficacy scores, networks were constructed. The seven different compounds of interest were purchased commercially and a series of assays were conducted with targeted proteins where expression levels could be determined. The results suggested that the key pharmacological effects of the formulation may be associated with maintaining homeostasis in the endocrine system, the immune system, and in metabolism.

Additionally, several researchers have developed and use network pharmacology approaches to attempt to understand detailed mechanisms of action of individual phytochemicals in herbal preparations and to suggest ways to enhance the synergy of herb combinations when used for specific diseases or conditions. Li *et al.*<sup>28</sup> used an integrated systems pharmacology platform to dissect the underlying mechanism of TCM prescriptions, and to form the basis of the herbal addition and subtraction theory (AST). In this approach, the researchers used a systems approach to add or remove herbal medicines from the original formulations *Xiao Chaihu* and *Da Chaihu*. This approach could develop into a practice that could play an important role in individualizing TCMs to account for certain measurable traits such as susceptibility to both adverse and effective drug response. As a starting point, all herbal ingredients were identified using the TCMSP database,<sup>25</sup> and information was collated in a database. Calculations for oral bioavailability were made using an in-house OBiavail1.1 tool; half-life was predicted through an *in silico* model created and supported by information from DrugBank.<sup>24</sup> These predictions were accomplished for all compounds to identify the most likely active compounds, and targets were predicted using the SySDT tool developed in their own laboratories. Targets were searched for corresponding diseases using TTD.<sup>23</sup> The main findings included 63 bioactive ingredients with 65 potential targets, confirming the anti-inflammatory, antioxidant, and anti-cytotoxicity activities of the concoctions. The authors approach highlights the importance of exploring the importance of complementary synergistic effects and eventually for personalizing and optimizing TCM recipes.

Wang *et al.*<sup>29</sup> used a systems biology model that integrated oral bioavailability, drug-likeness screening, target identification, and network pharmacology, and applied it to examine four widely used herbal medicines used in treating cardiovascular disease. These included *Radix Astragali Mongolici*, *Radix Puerariae Lobatae*, *Radix Ophiopogonis Japonici*, and *Radix Salviae Miltiorrhiza*. Their analysis explained the enhancement of pharmacological synergy with a combination (within a concoction) of herbal medicines and provides methods to explore drug-herb combinations that would enhance pharmacological action.

Zhang *et al.*<sup>30</sup> described a multi-component TCM network pharmacology platform developed in their own laboratories and used to create a network regulation mechanism which identifies active ingredients within a formulation as well as synergistic combinations of herbs. They illustrate the platform utility with an analysis of *Qing Luo Yin* in rheumatoid arthritis. The authors

propose their work supports the translation of TCM from experience-based to evidence-based medicine using network target multicomponent therapeutics.

Wang *et al.*<sup>31</sup> used a network pharmacology approach to study 721 compounds in the *Erxian* decoction. The compounds were determined from the TCM Database@Taiwan<sup>32</sup> and the TCMSP database.<sup>25</sup> Compound-protein interactions were obtained from Search Tool for Interactions of Chemicals (STITCH) 4.0<sup>33</sup> and compound-gene interactions were obtained from the Comparative Toxicogenomics Database (CTD).<sup>34</sup> Lipinski's rule (TCMSP)<sup>25</sup> and enrichment analyses (JEPETTO and DAVID 6.7)<sup>31</sup> of interactions were used to identify 20 potentially effective ingredients with acceptable oral bioavailability for relieving menopausal symptoms.

Liu *et al.*<sup>19</sup> used a drug-target-disease systems pharmacology model integrating oral bioavailability screening, drug-likeness evaluation, blood-brain barrier permeation, target identification with docking technologies, and network analysis. They identified 73 bioactive components (out of 287 ingredients) and 91 potential targets for licorice, the root of three *Glycyrrhiza* species. Licorice has been reported to be a constituent of ~60% of all TCM prescriptions and has been used for cough relief and as a detoxifying agent. In addition, it is widely used in TCM as an anti-inflammatory, anti-anabrosis, immunomodulatory, anti-platelet, and antiviral agent. The authors' detailed analysis of diseases associated with licorice components include diseases of the respiratory, gastrointestinal, and cardiovascular systems. As an example in the study, licorice flavonoids targeted ischemia-related proteins HTR1A, OPRD1, GSK3B, HRH1, MAPK10, F2, ADRA2A, and AChE.

Fu *et al.*<sup>35</sup> discuss the known and proposed pharmacodynamics activities of TCM components and the advantages of using the TCMSP framework.<sup>25</sup> This includes large-scale datamining to collect chemical, genomic, and pharmacologic data, and statistical analysis of the collected information. The framework includes *in silico* models to predict absorption, distribution, metabolism, and excretion (ADME) and toxicity properties of the constituent chemicals. Target predictions were made *via* methods from Yu<sup>36</sup> and Ru.<sup>25</sup> These methodologies were used to study synergistic mechanisms of herbs used in cardiovascular disease treatment including astragalus, kudzu, ophiopogon tuber, and salvia. A similar analysis of Folium Eriobotryae and the mechanisms underlying the use as an anti-inflammatory by Zhang *et al.*<sup>37</sup> 11 anti-inflammatory ingredients, mainly flavonoids and triterpene acids, were identified along with 43 proteins targeted by the compounds. Six of the targets including COX2, ALOX5, PPARG, ICAM1, COX1, and PPAED are also anti-inflammatory targets of Western therapeutics.

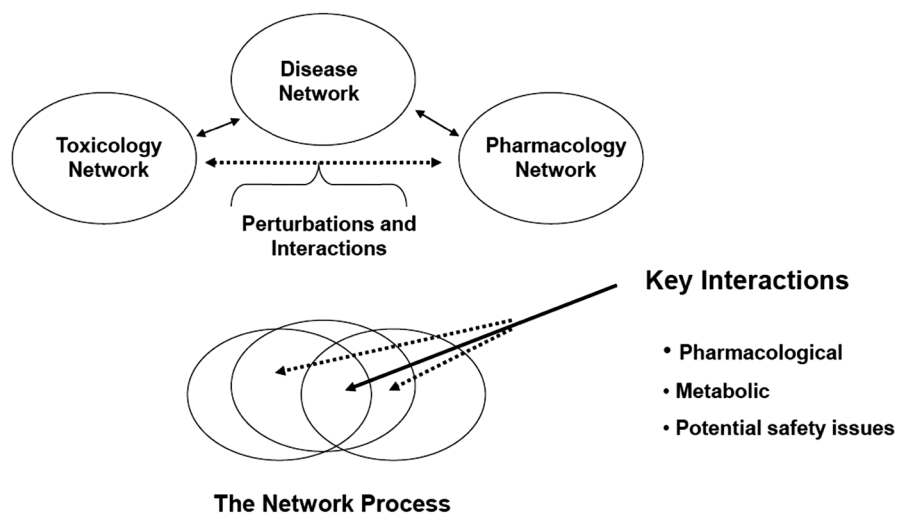
Liang *et al.*<sup>38</sup> studied TCM in relation to osteoarthritis, where intracellular signals lead to the overexpression of inflammatory mediators such as prostaglandins (PTGS1 and PTGS2); the NF- $\kappa$ B pathway is critical for translating the mechanical stress signaling into synthesis of inflammatory mediators to induce an inflammatory response. *Hong Teng* (*Sargentodoxa cuneata*) is a TCM that is often prescribed for "heat"-related inflammation and it is frequently used for the treatment of osteoarthritis. Phytochemicals were

determined using the TCM Database@Taiwan<sup>32</sup> and the .mol2 file detailing the structure of each molecule was obtained. The .mol2 files were entered into PharmMapper,<sup>39</sup> a web-based docking program that has the capacity to predict associations with protein targets in the human body. PharmMapper<sup>39</sup> was configured to generate 300 conformations per phytochemical and only analyze those conformers against human protein targets. The list of protein targets generated was then translated into a standard form by manual searching the protein names through Genecards.<sup>21</sup> The standardized list of putative protein targets was then entered into the CTD MyVenn tool<sup>34</sup> and compared to the human proteome affiliated with inflammatory processes to eliminate any protein targets that are not part of an inflammatory pathway. The resulting data set was then compared to the protein targets of industry-standard osteoarthritis medications. The primary overlaps were PTGS1, AKT1, BIRC, and ALB, all targets of Western anti-inflammatory medications. This study confirms from a computational basis, the pharmacological activity of components of the TCM *Hong Teng* in osteoarthritis.

### 11.3 Network Pharmacology in the Understanding of Herb–Drug Interactions

The systems pharmacology approach can also be used to establish known toxicities (adverse events) and potential herb–herb or drug–herb interactions by overlapping the disease, pharmacological action, and potential toxicity networks (Figure 11.2).

Chan *et al.*<sup>40</sup> used the approach to study potential interactions between TCM and Western therapeutics, using *Kang Ai Pian*, a TCM formulation used



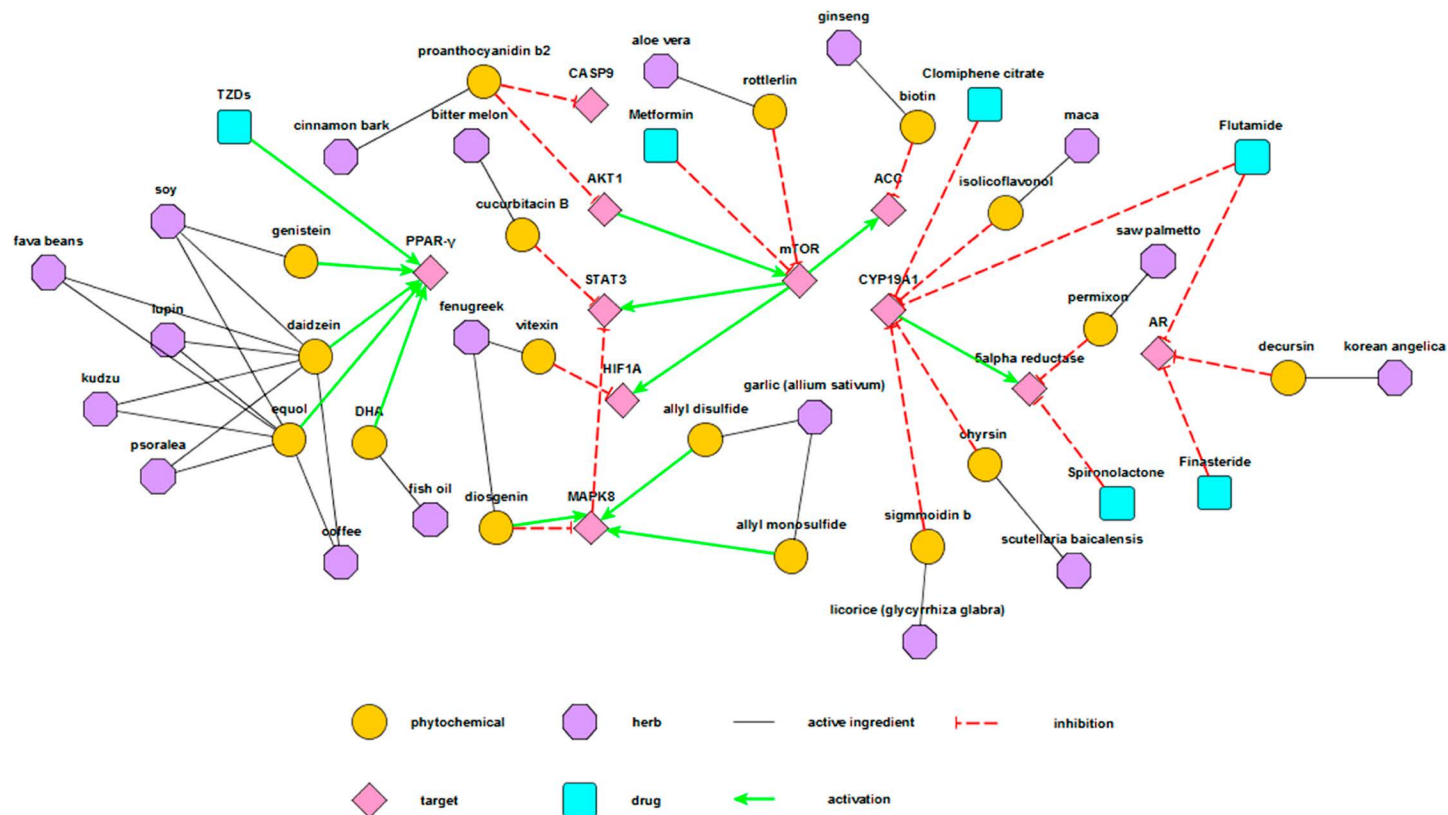
**Figure 11.2** The network process: linking diseases, phytochemicals, and mechanisms of action.

in various cancers such as cervical, ovarian, breast, nasopharyngeal, lung, liver, and gastrointestinal cancers, along with drugs used in the same cancers. *Kang Ai Pian* is comprised primarily of *chen pi* (*Citrus reticulata*; Citri Reticulatae Pericarpium), *huang bo* (*Phellodendron amurense*; Phellodendri Cortex), *huang lian* (*Coptis chinensis*, *Coptis deltoidea* and *Coptis teeta*; Coptidis Rhizoma), *huang qin* (*Scutellaria baicalensis*; Scutellariae Radix), *hu po* (amber; Succinum), *niu huang* (*Bos taurus domesticus*; Bovis Calculus), and *san qi* (*Panax notoginseng*; Notoginseng Radix). Using a network approach, a scheme was generated to include 15 different Western chemotherapy treatments that may be used concurrently. The primary tool used in this research was MetaDrug (Thomson Reuters).<sup>41</sup> This approach illustrates the potential herb–drug interactions (positive and negative) of *Kang Ai Pian* when used concomitantly with 15 different Western chemotherapy treatments.

## 11.4 Pharmacogenomics in TCM

An area of research in the TCM field, expected to increase in future, is the determination of individual susceptibilities in relation to polymorphic metabolizing cytochrome P450 enzymes. These enzymes represent the major phase I metabolic routes for a majority of marketed drugs and for individual phytochemicals within medicinal herbs. As an example of the necessity of this, Fricke-Galindo *et al.*<sup>42</sup> have shown in a world population comparative analysis that CYP2C19 shows important interethnic variations of alleles which can affect the metabolic phenotype in individuals, *e.g.* poor metabolizers were more frequent among Asians than in Europeans, which was contrary to the phenomenon reported for Europeans for CYP2D6. Poor metabolizers may lack functional CYP2C19 functional enzymes; therefore, the expected pharmacokinetics and pharmacodynamics could be significantly altered for a drug or phytochemical which could lead to excessive pharmacologic action, no pharmacologic activity, or possible toxicity. These findings represent important considerations when applying individualized TCMs and TCM–drug combinations. As of early 2016, there were 204 FDA-approved drug labels that contained pharmacogenomics information; in some cases these represent warnings and/or pharmacogenomics screening requirements. Approximately 40% of the labels involve drug metabolism [as in the Pharmacogenomics Knowledge Base (PharmGKB)].<sup>43</sup> Several publications report metabolism-based TCM and Western drug interactions and the online source Medline Plus<sup>44</sup> details known interactions of several herbal medicines. In addition, as of early 2016, there were 387 clinical studies on pharmacogenomics in [www.ClinicalTrials.gov](http://www.ClinicalTrials.gov);<sup>45</sup> however, none of these involved TCM formulations.

Ng *et al.*<sup>46</sup> used a series of online sources, including the Traditional Chinese Medicines Integrated Database (TCMID),<sup>47</sup> STITCH,<sup>33</sup> DrugBank,<sup>24</sup> PharmGKB,<sup>43</sup> the Kyoto Encyclopedia of Genes and Genomes (KEGG),<sup>48,49</sup> and the Side Effect Resource database (SIDER)<sup>50</sup> to study different Western medications and TCMs used in polycystic ovarian syndrome. A network diagram of phytochemicals, drugs, and targets is shown in Figure 11.3.



**Figure 11.3** Possible interactions between different phytochemicals and Western medications and their relative molecular targets.

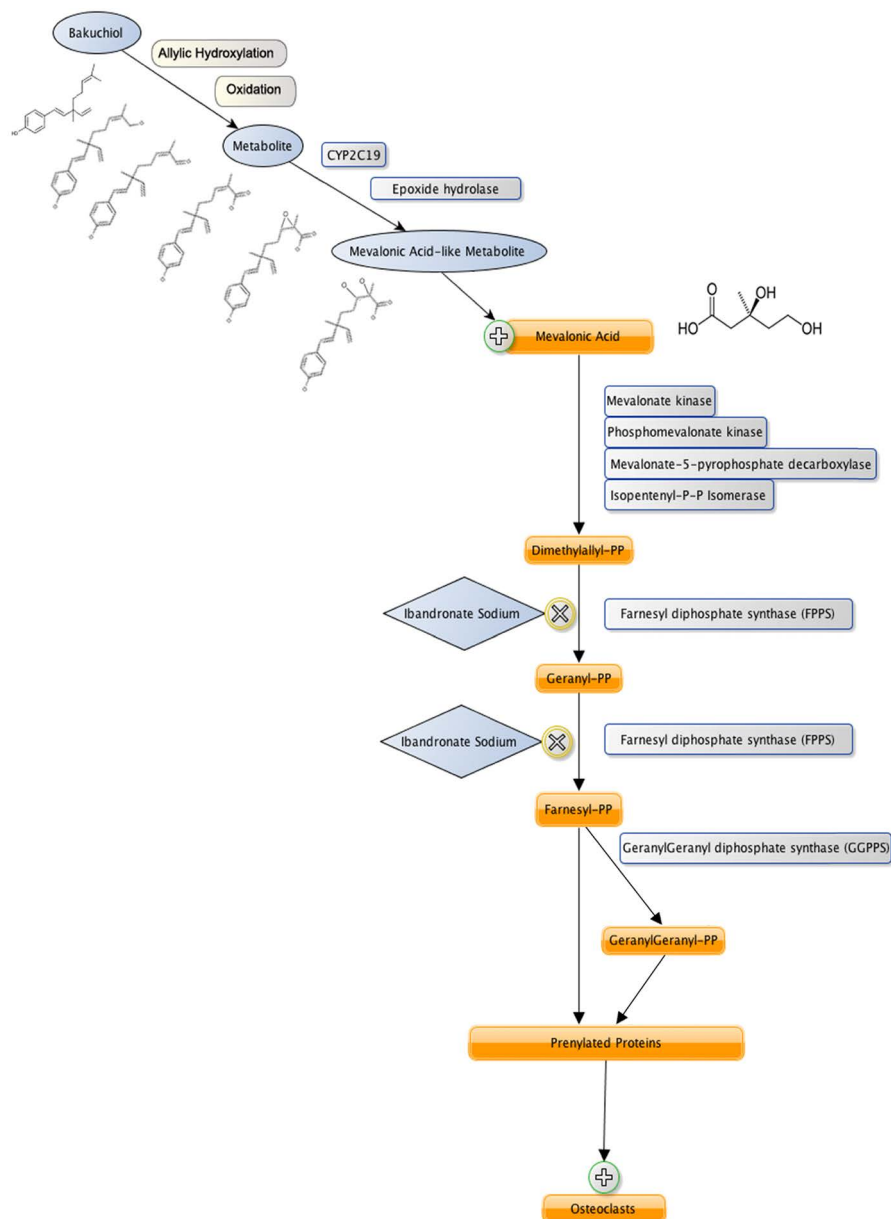
The overall goal of this study was to identify ethnicity-related differences in susceptibility to different therapeutic approaches. The study highlights the importance of understanding the metabolizing enzymes of drugs used for specific diseases and the same information for key active ingredients of TCM formulations. This could become important if certain individuals fall in the ultra- or poor metabolizer categories based on polymorphisms in metabolizing enzymes such as CYP2D6 and/or CYP2C19.<sup>51</sup> Potential pharmacogenomics of active metabolites of phytochemicals in TCM formulations can also play an important role in therapeutic outcome. Chang and Johnson<sup>52</sup> studied this in relation to Western and TCM treatments for postmenopausal osteoporosis. A comparative analysis of ibandronate sodium (Boniva®), which acts on the mevalonate (HMG-CoA reductase) pathway and the herb *Bu Gu Zi* (*Psoralea corylifolia*), a TCM used in osteoporosis, was conducted. A common TCM called BoneVigor® contains *P. corylifolia*, and one of its main components is a meroterpene called bakuchiol. Several databases and tools were used to compare genes associated with osteoporosis and how the two compounds interacted with key disease pathways. These included CTD,<sup>34</sup> KEGG,<sup>48,49</sup> PharmGKB,<sup>43</sup> Meteor Nexus,<sup>53</sup> and SMARTCyp,<sup>54</sup> which highlighted the possible overlaps between the major bakuchiol metabolite and ibandronate sodium pathways involved in osteoporosis. CTD,<sup>34</sup> KEGG,<sup>48,49</sup> and PharmGKB<sup>43</sup> were used to provide detailed biological information involved in the mevalonate and osteoporotic pathways. Ibandronate sodium was shown to inhibit farnesyl diphosphate synthase in the mevalonate pathway and is not metabolized by CYP450 enzymes, while bakuchiol is an estrogen receptor- $\alpha$  agonist and is subject to CYP450 metabolism. Using data from Meteor Nexus<sup>53</sup> and SMARTCyp,<sup>54</sup> bakuchiol was predicted to undergo allylic hydroxylation and then oxidation in mammals to yield a metabolite susceptible to further phase I biotransformation (Figure 11.4).

This metabolite contains a trisubstituted alkene that is susceptible to epoxidation by CYP2C19 and the opening of its hydrolytic ring by epoxide hydrolase. This bakuchiol metabolite resembles mevalonic acid, and presumably could enter the mevalonate pathway, which may account for its biological effects. Accordingly, polymorphisms in CYP2C19 may affect bakuchiol metabolism in osteoporotic patients, thereby affecting the way bakuchiol would interact with ibandronate sodium. This study provides a better understanding of how bakuchiol may be metabolized and interact with ibandronate sodium to help develop more personalized medicine regimens for osteoporotic patients with different CYP2C19 polymorphisms.

## 11.5 TCMs in Clinical Trials

An increasing number of TCM formulations are currently being tested in clinical trials and reported on [www.ClinicalTrials.gov](http://www.ClinicalTrials.gov),<sup>45</sup> a service of the US National Institutes of Health. [www.ClinicalTrials.gov](http://www.ClinicalTrials.gov) is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. As of April 2016, there were >600 clinical





**Figure 11.4** Diagram of proposed bakuchiol and ibandronate sodium in relation to the mevalonate pathway. Ibandronate sodium is a bisphosphonate drug that inhibits FPPS in the mevalonate pathway. From Meteor Nexus, bakuchiol can first undergo allylic hydroxylation and oxidation to yield a trisubstituted alkene metabolite. CYP2C19 and epoxide hydrolase can metabolize it to yield a mevalonic acid-like metabolite. This metabolite can then possibly act as a competitive inhibitor of mevalonic acid and stimulate the mevalonate pathway.

trials reported using different TCM formulations alone or in combination with Western therapeutics, involving >700 conditions.

## 11.6 The Future of Network Pharmacology in Traditional Medicine

The network pharmacology approach for traditional medicines can provide a new vision for drug discovery, starting with well documented TCM remedies, creating a detailed understanding of mechanisms of action, and proposing new multi-target agents and therapeutically relevant drug and herbal combinations.<sup>11,15,18,20,55,56</sup> The importance of bioinformatics methods and tools in these efforts has been apparent,<sup>56–58</sup> particularly with the introduction of several online databases and tools, as described previously, used to parse out the “active” ingredients from herbs and herb concoctions and to aid in the creation of network maps and follow-on research.

Essays on the future of traditional medicines worldwide has been published in a three-part series entitled *The Art and Science of Traditional Medicine*, published by Science/AAAS Custom Publishing Office.<sup>59–61</sup> This series makes the case for the integration of TCM into modern medical practice. The three series are: *Part 1: TCM Today—A Case for Integration*;<sup>59</sup> *Part 2: Multidisciplinary Approaches for Studying Traditional Medicine*;<sup>60</sup> and *Part 3: The Global Impact of Traditional Medicine*.<sup>61</sup> In Part 1, Qi and Kelly<sup>62</sup> discuss how traditional and complementary medicine (T&CM) continue to be popular and in demand worldwide. Several countries have integrated certain types of traditional medicine into their healthcare systems. These include China, the Republic of Korea (South Korea), India, and Vietnam. In China, conventional and traditional Chinese medicine are integrated at every level of the healthcare service, and public and private insurance cover both forms of treatment.<sup>62</sup> Based on both needs and challenges, the current World Health Organization strategy has established the following objectives:

- (1) To build the knowledge base for active management of T&CM through appropriate national policies.
- (2) To strengthen quality assurance, safety, proper use, and effectiveness of T&CM by regulating T&CM products, practices, and practitioners.
- (3) To promote universal health coverage by integrating T&CM services into health care service delivery and self-health care.<sup>62</sup>

Wang and Xu<sup>63</sup> describe TCM as an ancient medical practice system that is based on regulating the integrity of the human body along with its inter-relationship with natural environments. *Zheng*, defined as syndrome or pattern, represents the overall physiological and/or pathological responses of the body to any internal or external influence, and in TCM, this is fundamental in both diagnosis and treatment of diseases. Wang and Xu<sup>63</sup> propose a comprehensive *Zheng* map that connects all *Zheng* based on molecular and

cellular relationships. And, as a new ‘omics’ endeavor, they propose creating a “Zhengome” used to investigate the molecular and systems hierarchies, environmental factors, common treatments, and relationships between different *Zhenges*. In relation to drug discovery, Wang and Xu<sup>63</sup> propose a *Zheng*-to-TCM reverse target and screening approach to identify active components in herbal medicines and to uncover multi-ingredient synergism within formulations and as used with Western drug combinations. In a TCM-to-*Zheng* strategy, they propose screening herbal ingredients for ADME and toxicity properties, creating databases, constructing and analyzing networks, and connecting all data into *Zhenges* and diseases.

## 11.7 Conclusion

Network pharmacology has been defined as a process that encompasses a network approach to represent and analyze the complex biological systems underlying diseases and drug actions. It has become a critical part of drug discovery, drug design, and drug development, thereby sharing a holistic perspective characteristic of the science and practice of TCM. The basic principle is to construct a biological network that deconstructs a disease or condition into genes, gene products, and related associations such as connectivity and feedback, which when interrogated will provide intuitive information on therapeutic interventions from herbal ingredients in TCM formulations and TCM in combination with Western drugs.<sup>64</sup> This concept is now fundamental in TCM because various herbs consist of several phytochemicals and TCM concoctions consist of several plant and animal sources, therefore, multiple chemical compounds. Accordingly, the “polypharmacology” concept is now widely accepted as essential in understanding Chinese herbal medicines from systems biology and bioinformatics standpoints.<sup>20</sup>

## References

1. W. Zhang, Network pharmacology: A further description, *Network Pharmacol.*, 2016, **1**(1), 1–14.
2. G. Zhang, Q. Li, Q. Chen and S. Su, Network Pharmacology: A New Approach for Chinese Herbal Medicine Research, *J. Evidence-Based Complementary Altern. Med.*, 2013, **2013**, 621423, DOI: 10.1155/2013.
3. A. Hopkins, Network biology illuminates our understanding of drug action, *Nat. Biotechnol.*, 2007, **25**(10), 1110–1111.
4. R. Verpoorte, D. Crommelin, M. Danhof, L. Gilissen, H. Schuitmaker, J. van der Greef and R. Witkamp, A systems view on the future of medicine: Inspiration from Chinese medicine?, *J. Ethnopharmacol.*, 2009, **121**, 479–481.
5. J. Qiu, ‘Back to the future’ for Chinese herbal medicines, *Nat. Rev. Drug Discovery*, 2007, **6**, 506–507.
6. R. Verpoorte, Y. Choi and H. Kim, Ethnopharmacology and systems biology: a perfect holistic match, *J. Ethnopharmacol.*, 2005, **100**, 53–56.

7. J. Gertsch, Botanical drugs, synergy, and network pharmacology: forth and back to intelligent mixtures, *Planta Med.*, 2011, **77**, 1086–1098.
8. J. Zhao, P. Jiang and W. Zhang, Molecular networks for the study of TCM pharmacology, *Briefings Bioinf.*, 2009, **11**(4), 417–430.
9. X. Wu, M. Jiang, M. Zhang and S. Li, Network-based global inference of human disease genes, *Mol. Syst. Biol.*, 2008, 189, DOI: 10.1038/msb.2008.27.
10. X. Liu, W.-Y. Wu, B.-H. Jiang, M. Yang and D.-A. Guo, Pharmacological tools for the development of traditional Chinese medicine, *Trends Pharmacol. Sci.*, 2013, **34**(11), 620–628.
11. R. Graziose, M. Lila and I. Raskin, Merging traditional Chinese medicine with modern drug discovery technologies to find novel drugs and functional foods, *Curr. Drug Discovery Technol.*, 2010, **7**, 2–12.
12. D. Veber, S. Johnson, H.-Y. Cheng, B. Smith, K. Ward and K. Kopple, Molecular properties that influence oral bioavailability of drug candidates, *J. Med. Chem.*, 2002, **45**, 2615–2623.
13. R. Kumar, A. Sharma and P. Varadivai, A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine, *J. Nat. Sci., Biol. Med.*, 2011, **2**(2), 168–173.
14. D. Clark and S. Pickett, Computational methods for the prediction of “drug-likeness”, *Drug Discovery Today*, 2000, **5**(2), 49–58.
15. H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang and Y. Wang, A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data, *PLoS One*, 2012, **7**, e37608.
16. Q. Xu, R. Bauer, B. Hendry, T.-P. Fan, Z. Zhao, P. Duez, M. Simmonds, C. Witt, A. Lu, N. Robinson, D. Guo and P. Hylands, The quest for modernization of traditional Chinese medicine, *BMC Complementary Altern. Med.*, 2013, **13**, 132–142.
17. Good Practice in Traditional Chinese Medicine Research, [www.gp-tcm.org/the-project/](http://www.gp-tcm.org/the-project/), H. Uzner, T.-P. Fan, A. Dias, D. Guo, H. El-Nezami and Q. Xu, Establishing an EU-China consortium on traditional Chinese medicine research, *Chin. Med.*, 2010, **5**, 42, <http://www.cmjournal.org/content/5/42>.
18. D.-X. Kong, X.-J. Li, G.-Y. Tang and H.-Y. Zhang, How many traditional Chinese medicine components have been recognized by modern western medicine? A chemoinformatic analysis and implications for finding multicomponent drugs, *ChemMedChem*, 2008, **3**, 233–236.
19. H. Liu, J. Wang, W. Zhou, Y. Wang and L. Yang, Systems approaches and polypharmacology for drug discovery from herbal medicines: An example using licorice, *J. Ethnopharmacol.*, 2013, **146**, 773–793.
20. M. Kibble, N. Saarinen, J. Tang, K. Wennerberg, S. Mäkelä and T. Aittokallio, Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products, *Nat. Prod. Rep.*, 2015, **32**, 1249–1266.
21. *GeneCards – Human Genes, Gene Database, Gene Search*, <http://www.genecards.org/>.

22. *MalaCards – Human Disease Database*, [www.malacards.org](http://www.malacards.org).
23. *TTD: Therapeutic Target Database. BIDD*, <http://bidd.nus.edu.sg/group/cjttd/>.
24. *DrugBank*, [www.drugbank.ca/](http://www.drugbank.ca/).
25. J. Ru, P. Li, J. Wang, W. Zhou, B. Li, C. Huang, P. Li, Z. Guo, W. Tao, Y. Yang, X. Xu, Y. Li, Y. Wang, L. Wang and L. Yang, TCMSP: a database of systems pharmacology for drug discovery from herbal medicines, *J. Cheminf.*, 2014, **6**, 13.
26. S. Shi, Y. Cai, X. Cai, X. Zheng, D. Cao and F. Ye, A Network Pharmacology Approach to Understanding the Mechanisms of Action of Traditional Medicines: Bushenhuoxue Formula for the treatment of chronic kidney disease, *PLoS One*, 2014, **9**(3), e89123, DOI: 10.1371/journal.pone.0089123.
27. X. Liang, H. Li and S. Li, A novel network pharmacology approach to analyse traditional herbal formulae: the Liu-Wei-Di-Huang pill as a case study, *Mol. BioSyst.*, 2014, **10**, 1014–1022.
28. B. Li, W. Tao, C. Zheng, P. Shar, C. Huang, Y. Fu and Y. Wang, Systems pharmacology-based approach for dissecting the addition and subtraction theory of traditional Chinese medicine: An example using Xiao-Chaihu-Decoction and Da-Chaihu-Decoction, *Comput. Biol. Med.*, 2014, **53**, 19–29.
29. X. Wang, X. Xu, W. Tao, Y. Li, Y. Wang and L. Yand, A Systems Biology Approach to Uncovering Pharmacological Synergy in Herbal Medicines with Application to Cardiovascular Disease, *J. Evidence-Based Complementary Altern. Med.*, 2012, **2012**, 519031, DOI: 10.1155/2012/51931.
30. B. Zhang, X. Wang and S. Li, An Integrative Platform of TCM Network Pharmacology and its application on a Herbal Formula, Qing-Luo-Yin, *J. Evidence-Based Complementary Altern. Med.*, 2013, **2013**, 456747, DOI: 10.1155/2013/456747.
31. S. Wang, Y. Tong, T. Ng, L. Lao, J. Lam, K. Zhang, Z. Zhang and S. Sze, Network pharmacological identification of active compounds and potential actions of Erxian decoction in alleviating menopause-related symptoms, *Chin. Med.*, 2015, **10**, 19, DOI: 10.1186/s13020-015-0051-z.
32. *Traditional Chinese Medicine Database @ Taiwan*, <http://tcm.cmu.edu.tw/>.
33. *STITCH 4.0*, <http://stitch.embl.de/>.
34. *Comparative Toxicogenomics Database (CTD)*, <http://ctdbase.org/>.
35. Y. Fu, Y. Wang and B. Zhang, Systems pharmacology for traditional Chinese medicine with application to cardio-cerebrovascular diseases, *J. Tradit. Chin. Med. Sci.*, 2014, **1**, 84–91.
36. H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang and Y. Wang, A Systematic Prediction of Multiple Drug-Target Interactions from Chemical, Genomic, and Pharmacological Data, *PLoS One*, 2012, **7**(5), e37608, DOI: 10.1371/Journal.pone.0037608.
37. J. Zhang, Y. Li, S.-S. Chen, L. Zhang, J. Wang, Y. Yang, S. Zhang, Y. Pan, Y. Wang and L. Yang, Systems Pharmacology Dissection of the Anti-Inflammatory Mechanism for the Medicinal Herb Folium Eriobotryae, *Int. J. Mol. Sci.*, 2015, **16**, 2913–2941.

38. G. Liang, X. Huang, R. Trimble, C. Park and D. Johnson, Computational Comparison of the Anti-Inflammatory Targets of the Traditional Chinese Medicine *Sargentodoxa cuneate* (Hong Teng) and Western Therapeutics for the Treatment of Osteoarthritis, *Toxicologist*, 2015, **144**(1), 279, Poster: ResearchGate.
39. *PharmMapper*, <http://59.78.96.61/pharmmapper/index.php>.
40. E. Chan, M. Tan, J. Xin, S. Sudarsanam and D. Johnson, Interactions between Traditional Chinese Medicines and Western Therapeutics, *Curr. Opin. Drug Discovery Dev.*, 2010, **13**(1), 58–73.
41. *MetaDrug. Manually Curated Systems Pharmacology Solution*, [www.ThomsonReuters.com](http://www.ThomsonReuters.com).
42. I. Fricke-Galindo, C. Céspedes-Garro, F. Rodrigues-Soares, M. Naranjo, Á. Delgado, F. de Andrés, M. López-López, E. Peñas-Lledó and A. Llerena, Interethnic variation of *CYP2C19* alleles, 'Predicted' phenotypes and 'measured' metabolic phenotypes across world populations, *Pharmacogenomics J.*, 2016, **16**, 113–123.
43. *PharmGKB. The Pharmacogenomics Knowledgebase*, [www.pharmgkb.org](http://www.pharmgkb.org).
44. *Herbs and Supplements: MedlinePlus*, [www.nlm.nih.gov/medlineplus/druginfo/herb\\_All.html](http://www.nlm.nih.gov/medlineplus/druginfo/herb_All.html).
45. *ClinicalTrials.gov. A service of the U.S. National Institutes of Health*, <https://clinicaltrials.gov>.
46. S. Ng, E. Qui, A. Chang, L. Phan and D. Johnson, A Computational Analysis of Ethnicity-Specific Polycystic Ovarian Syndrome Treatment using Western and Traditional Medicines, *Toxicologist*, 2015, **144**(1), 280, Poster: ResearchGate.
47. *Traditional Chinese Medicine Integrated Database (TCMID)*, [www.megabionet.org/tcmid](http://www.megabionet.org/tcmid).
48. *KEGG Pathway Database*, [www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html).
49. *KEGG ENVIRON*, [www.genome.jp/kegg/drug/enviro.html](http://www.genome.jp/kegg/drug/enviro.html).
50. *SIDER. Side Effect Resource*, <http://sideeffects.embl.de/>.
51. M. Liu, Y. Zhang, M. Zeng, F. He, Z. Luo, J. Luo, J. Wen, X. Chen, H. Zhou and W. Zhang, Pharmacogenomics and Herb-Drug Interactions: Merge of Future and Tradition, *J. Evidence-Based Complementary Altern. Med.*, 2015, **2015**, 321091, DOI: 10.1155/2015/321091.
52. R. Chang and D. Johnson, Computational Analysis of Potential Interactions from the Combined use of Western Therapeutics and Traditional Chinese Medicines in Postmenopausal Osteoporosis, *Toxicologist*, 2014, **138**(1), 112, Poster: ResearchGate.
53. *Meteor Nexus. Metabolism Expert System*, [www.lhasalimited.org](http://www.lhasalimited.org).
54. *SMARTCyp Web Service*, [www.farma.ku.dk/smartcyp/](http://www.farma.ku.dk/smartcyp/).
55. A. Harvey, R. Edrada-Ebel and R. Quinn, The re-emergence of natural products for drug discovery in the genomics era, *Nat. Rev. Drug Discovery*, 2015, **14**, 111–129.
56. J. Hughes, S. Rees, S. Kalindjian and K. Philpott, Principles of Early Drug Discovery, *Br. J. Pharmacol.*, 2011, **162**, 1239–1249.

57. D. Barlow, A. Buriani, T. Ehrman, E. Bosisio, I. Eberini and P. Hylands, *In-silico* studies in Chinese herbal medicines' research: Evaluation of *in-silico* methodologies and phytochemical data sources, and a review of research to date, *J. Ethnopharmacol.*, 2012, **140**, 526–534.
58. P. Gu and H. Chen, Modern bioinformatics meets traditional Chinese medicine, *Briefings Bioinf.*, 2015, **2015**, 321091, DOI: 10.1093/bib/bbt063.
59. T.-P. Fan, J. Briggs, L. Liu, A. Lu, J. van der Greef and A. Xu, The Art and Science of Traditional Medicine. Part 1: TCM Today – A Case for Integration, *Science*, 2014, **346**(suppl. 6216), S1–S24.
60. T.-P. Fan, J. Briggs, L. Liu, A. Lu, J. van der Greef and A. Xu, The Art and Science of Traditional Medicine. Part 2: Multidisciplinary Approaches for Studying Traditional Medicine, *Science*, 2015, **347**(suppl. 6219), S26–S51.
61. T.-P. Fan, J. Briggs, L. Liu, A. Lu, J. van der Greef and A. Xu, The Art and Science of Traditional Medicine. Part 3: The Global Impact of Traditional Medicine, *Science*, 2015, **350**(suppl. 6259), S53–S85.
62. Z. Qi and E. Kelly, The WHO Traditional Medicine Strategy 2014-2023: A Perspective, *Science*, 2014, **346**(suppl. 6216), S5–S6.
63. Y. Wang and A. Xu, *Zheng*: A systems biology approach to diagnosis and treatments, *Science*, 2014, **346**(suppl. 6216), S13–S15.
64. S. Li, Mapping ancient remedies: Applying a network approach to traditional Chinese medicine, *Science*, 2015, **350**(suppl. 6259), S72–S74.



## CHAPTER 12

# *Chemical–Disease Category Linkage (CDCL): Computational Methods Linking Traditional Chinese Medicines and Western Therapeutics*

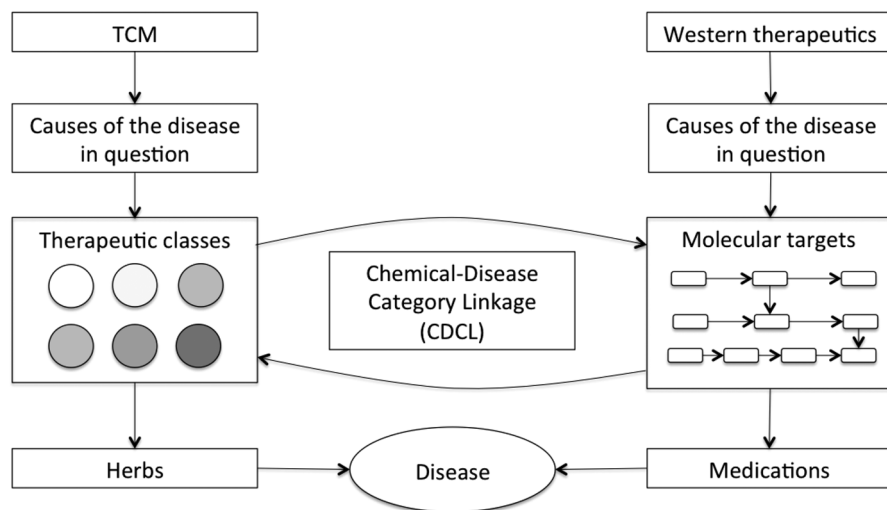
DALE E. JOHNSON<sup>\*a,b</sup> AND KIT WUN KATHY CHEUNG<sup>c</sup>

<sup>a</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA;

<sup>b</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Morgan Hall, Berkeley, CA 94720-3104, USA; <sup>c</sup>University of California, San Francisco, School of Pharmacy, San Francisco, CA 94143, USA  
\*E-mail: [daleej@umich.edu](mailto:daleej@umich.edu)

## 12.1 Introduction

Chemical–Disease Category Linkage (CDCL) is a term used to encompass various methods and approaches to link traditional Chinese medicine (TCM) and/or dietary supplements with TCM classifications or categories, diseases, the molecular basis of both diseases and therapeutic interventions, and the synergistic and/or antagonistic interactions of potential combined use with Western therapeutic treatments<sup>1</sup> (Figure 12.1).



**Figure 12.1** In Western therapeutics, clinicians treat patients with medications that can exert their effects by modulating the known molecular targets associated with the disease in question. Due to the differences in philosophy, the causes of the disease in question can be different in traditional Chinese medicine (TCM). To target those causes, TCM practitioners often use TCM herbal recipes that are comprised of herbs from various therapeutic classes. Like Western therapeutics, the active constituents in these herbal ingredients (phytochemicals) must reach a molecular target in order to exert their effects. The idea of chemical-disease category linkage (CDCL) is an effort to link TCM and/or dietary supplements with their own classifications, diseases, molecular basis of the diseases as well as therapeutic interventions, and the synergistic and/or antagonistic effects with Western therapeutics when used concomitantly.

Traditional medicine, which incorporates the therapeutic use of herbs and other natural products particularly in recipes or other combinations, has been embedded in many cultures for thousands of years. There is an extensive foundation for the therapeutic effects of herbal medicines primarily derived from their extensive use, but which also includes supportive research data including clinical trial results.<sup>2</sup> The World Health Organization (WHO) has estimated that ~80% of the global population relies on traditional herbal medicines as part of standard healthcare.<sup>3</sup> In the United States, where herbal remedies are classified as dietary supplements, an estimated one in five adults regularly consume herbal products.<sup>4</sup> Consumers are increasingly using web-based sources for diagnosis and suggestions of treatments, which presents an issue within the healthcare system, as most herbal products are not included in patient records as “other medications”, thereby limiting the understanding of potential herb–drug interactions to both the patient and healthcare professional.<sup>5</sup> China has long recognized the value of the combination of traditional and conventional (Western) medicine and originated

both the study and practice of integrated medicine. The National Institutes of Health (NIH) has defined complementary and alternative medicine (CAM) as a group of diverse medical and healthcare systems, practices, and products that are not generally considered to be part of conventional medicine.<sup>2</sup> While alternative medicine is used instead of conventional medicine, complementary medicine is used in conjunction with conventional therapeutics. Therefore, according to the NIH definition, integrated medicine encompasses the use of both conventional and alternative therapies for which there is evidence of safety and effectiveness.<sup>2</sup> In a report in 2002, it was estimated that more than one-third of US citizens had used CAM within the previous 12 months, and more than half of patients aged 18 years or older used CAM therapies in addition to conventional therapies because of the belief that CAM therapies increased the beneficial effects or reduced side effects of conventional drugs.<sup>2</sup>

Chan *et al.*<sup>2</sup> reviewed the history and principles of TCM and discussed a method to deconvolute TCM formulations into constituent phytochemicals, along with identification of potential molecular targets and the proposed relevant biological activity. The review also details several synergistic and antagonistic examples of herbs and phytochemicals within a formulation, and methods to explore interactions between TCM and Western therapeutics.

Recent Western-related milestones for TMC include:

- first TCM approved for the European Union market in 2012: *Di'ao Xin Xue Kang*, an extract of *Dioscorea nipponica* approved in the Netherlands for pain relief;<sup>5</sup>
- second ever botanical drug approved by the US Food and Drug Administration (FDA) in 2012. Fulyzaq (crofelemer); derived from the red sap of *Croton lechleri* for HIV-associated diarrhea. In 2006, the US FDA approved the first botanical prescription drug, Veregen (sinecatechins), a treatment for external genital and perianal warts<sup>6</sup>
- Youyou Tu was awarded the 2011 Lasker–DeBakey Clinical Medical Research Award for discovering artemisinin as a treatment for malaria,<sup>7</sup> and shared the 2015 Nobel Prize in Physiology or Medicine for discovering the novel treatment for malaria<sup>8</sup>

In the discovery process for artemisinin, Tu started with a large-scale screen of herbal remedies in malaria-infected animals, and revisited the ancient literature and discovered clues that guided her in her quest to successfully extract the active component from *Artemisia annua*.

### 12.1.1 Databases for CDCL Information and Study

Several databases are used to link TCM with Western therapeutics following methods that include identification of TCM herb and formulae usage in certain diseases and conditions, identifying the most important active constituents (phytochemicals), Western therapeutics used in the same diseases,

disease pathway analyses, drug and/or herb target identification, and potential interactions. Several of these information sources are highlighted below.

### 12.1.1.1 TCM and Chemical Constituents

- Traditional Chinese Medicine Database@Taiwan<sup>9</sup> contains information on medicinal herbs and chemical constituents, including 3D-structures of chemicals which can be used for structural docking.
- TCMID (Traditional Chinese Medicine Integrated Database)<sup>10</sup>—contains information on herbs, formulae, chemical ingredients, drug targets and/or disease genes/proteins.
- TCMgeneDIT<sup>11</sup>—a database system about TCM, genes, diseases, TCM effects, and TCM ingredients.
- HerbMed<sup>12</sup>—a database provided by Alternative Medicine Foundation, Inc. The herbal information is categorized and is evidence-based. Open access is limited.
- Dr Duke's Phytochemical and Ethnobotanical Databases<sup>13</sup>—information on chemical constituents.

### 12.1.1.2 TCM Classification and Systems Approach

- Chem-TCM<sup>14</sup>—Based on the publication by Ehrman *et al.*,<sup>15</sup> Chem-TCM is a database of chemical constituents of plants used in TCM with the following parts: chemical identification, botanical identification, predicted activity against common Western therapeutic targets, and estimated molecular activity according to TCM categories. The goal is to connect Chinese and Western medicines on a molecular level. The database lists predicted activities in 41 therapeutically significant targets in Western medicine.
- TCMSP (Traditional Chinese Medicine Systems Pharmacology) Database and Analysis Platform.<sup>16</sup> The TCMSP database includes herbs, chemical ingredients with structure files and absorption, distribution, metabolism and excretion (ADME) properties, and compound–target–disease network generation. TCMSP is a platform for combining TCM theory, new TCM combination development, active component identification and screening, and systems pharmacology analyses.
- TCM-ID The Traditional Chinese Medicine Information Database is referenced in several publications, however it's use is currently unavailable. It provided information on TCM formulations, herbal compositions, chemical compositions, molecular structures and functional properties, therapeutic and side effects, clinical indications and applications, and links to relevant information.

Evidence for pharmacological activity in these databases includes predictive, *in vitro*, animal based results, publications with human subjects, and in some cases, results from controlled clinical trials.

### 12.1.1.3 Western Therapeutics

- DrugBank<sup>18,19</sup>—a unique bioinformatics and chemoinformatics resource that combines detailed drug data with comprehensive target information.

### 12.1.1.4 Therapeutic Targets and Protein Interactions

- SuperTarget<sup>20,21</sup>—an extensive web resource for analyzing >300 000 drug–target interactions.
- TTD (Therapeutic Target Database)<sup>22,23</sup>—provides information about known and explored therapeutic protein and nucleic acid targets, the target-related disease, pathway information, and corresponding drugs targeted at each target.
- Matador<sup>24</sup>—resource for protein–chemical interactions, including both direct and indirect interactions.
- PDTD (Potential Drug Target Database)<sup>25,26</sup>—associates an informatics database to a structural database of known and potential drug targets, focusing on drug targets with known 3D structures.
- TRMP (Therapeutically Relevant Multiple Pathways)<sup>27,28</sup>—a database that integrates information on therapeutic targets and disease-associated signaling pathways.
- STITCH (Search Tool for Interactions of Chemicals)<sup>29,30</sup>—a tool that integrates interactions of proteins and chemicals from different databases.
- HIT (Herbal Ingredients' Targets) Database<sup>31,32</sup>—a comprehensive and fully curated database linking protein targets and active herbal ingredients.

### 12.1.1.5 Pathway Analysis

- KEGG (Kyoto Encyclopedia of Genes and Genomes) and KEGG Pathway Database—a collection of manually drawn pathway maps on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development. There is also a direct link and searching capability for medicinal herbs.<sup>33</sup>

## 12.1.2 TCM Classifications

In TCM, herbs are categorized into therapeutic classes, and TCM practitioners use a mixture of herbs from different classes to exert the therapeutic effects and modulate different bodily functions. Ehrman *et al.*<sup>15</sup> published a comprehensive classification of TCM which included TCM category, an approximate Western equivalent, signs and symptoms/conditions, and representative herbs. The information sets also include the number of herbs and number of compounds

(phytochemicals) associated with each category. As mentioned, these data are the source of information used to construct the Chem-TCM database. An extract of these data are presented below by TCM category, the approximate Western equivalent (WE), and selected signs and symptoms (SS).<sup>15</sup>

- *Wind cold* (WE) diaphoretic, antiviral, antibacterial; (SS) chills, headache, body and neck pain, no to mild fever
- *Wind heat* (WE) diaphoretic, antiviral, antibacterial; (SS) fever, sore throat, mild chills, deep-seated infections, rashes, eye problems
- *Heat (Qi)* (WE) refrigerant, antipyretic, anti-inflammatory, antimicrobial; (SS) high fever, irritability, thirst, delirium, skin diseases
- *Heat (blood)* (WE) refrigerant, styptic, coagulant; (SS) fever, rash, nose-bleed, vomiting blood, blood in stool/urine, skin diseases
- *Damp heat* (WE) antimicrobial, antipyretic, anti-inflammatory; (SS) dysentery, urinary difficulty, jaundice, eczema-type skin disease
- *Toxic heat* (WE) detoxicant, anti-inflammatory, antimicrobial, antiviral, diuretic; (SS) painful swellings, purulent infections, abscesses, dysentery, mumps, encephalitis, skin disease
- *Heat (deficiency)* (WE) antipyretic, anti-inflammatory, antimicrobial, antiviral; (SS) fever, night fever
- *Laxative* (WE) laxative, purgative; (SS) constipation
- *Cathartic* (WE) cathartic; (SS) edema, ascites, pleurisy
- *Drain dampness* (WE) diuretic; (SS) edema, urinary dysfunction, jaundice, heart disease
- *Wind damp* (WE) antirheumatic conditions, analgesic, antipyretic, anti-inflammatory, anticoagulant; (SS) joint and muscle pain and numbness, rheumatic conditions
- *Phlegm heat* (WE) expectorant, antitussive, anti-inflammatory, sedative; (SS) dry cough, convulsions, some psychiatric conditions
- *Phlegm cold* (WE) expectorant, decongestant; (SS) productive cough, phlegm
- *Coughing and wheezing* (WE) antitussive, expectorant, antibiotic, diuretic, laxative, antiasthmatic; (SS) persistent cough, wheezing, asthma
- *Emetic* (WE) emetic; (SS) severe phlegm, retained food, certain types of jaundice
- *Aromatic (damp)* (WE) digestive stimulant; (SS) distention, nausea, vomiting, poor appetite, greasy tongue coating
- *Regulate Qi* (WE) digestive stimulant, circulatory stimulant, analgesic; (SS) pain, diarrhea/constipation, irregular menstruation, stifling sensation in chest
- *Stop bleeding* (WE) styptic; (SS) bleeding, vomiting and coughing blood, hematuria, excessive menstruation, trauma
- *Invigorate blood* (WE) anticoagulant, circulatory stimulant; (SS) pain, abscesses, ulcers, abdominal masses, thrombosis, ischemia
- *Interior cold* (WE) circulatory stimulant, cardiogenic; (SS) cold extremities, lack of thirst, loose stool, diarrhea, nausea, chest and abdominal pain, slow pulse

- *Tonify Qi* (WE) endocrine agent, immunostimulant; (SS) lethargy, weakness, poor appetite, weak voice, pale complexion, breathlessness, immunodeficiency
- *Tonify blood* (WE) anti-anemic; (SS) pallor, dizziness, vertigo, poor vision, lethargy, palpitations, amenorrhea, insomnia, pale tongue, fine pulse
- *Tonify Yang* (WE) endocrine agent, stimulant; (SS) systemic exhaustion, fear of cold, cold extremities, withdrawal, sore and weak lower back, slow and deep pulse
- *Tonify Yin* (WE) endocrine agent, antidiuretic, antihypertensive, anti-cholesterolemic; (SS) dizziness, tinnitus, weak lower back and knees, low-grade fever, menopausal symptoms, scant dark urine, red dry tongue, thin pulse
- *Astringent* (WE) astringent, endocrine agent; (SS) diarrhea, polyuria, sweating, prolapse, discharge
- *Shen* (WE) tranquillizer, sedative, nerve tonic; (SS) palpitations, anxiety, insomnia
- *Phlegm (heart)* (WE) resuscitant, tranquillizer, stimulant, nerve agent; (SS) delirium, seizure, coma, psychiatric conditions, bipolar disease
- *Internal wind* (WE) antihypertensive, sedative, nerve tonic; (SS) tremor, spasm, hemiplegia, aphasia, paralysis, blurred vision, dizziness, paresthesia, stroke

The TCMSP database<sup>16</sup> uses a similar classification, referred to as therapeutic classes, as listed below:

- Spirit calming
- Eliminating toxic materials, dissolving rottenness, and growing new muscles
- Tonifying weakness
- Dissolving dampness by flavors
- Dissolving phlegm, stopping cough, and soothing breathing
- Blood activation and stasis removal
- Relieving exterior syndrome
- Toxicification reduction, anthelmintic, dampness removal, and itching control
- Mind opening
- Antitumor
- Regulation of Qi
- For promoting diuresis and penetrating dampness
- Anesthesia
- Calming liver and containing wind
- Heat clearance
- Anthelmintic
- Dispelling wind dampness
- Astringency
- Warming exterior
- Warming interior



- Food digestion
- Purging
- Promotion of vomiting
- Bleeding control
- Others

The TCM-ID database<sup>17</sup> uses traditional classifications according to specific functions within therapeutic classes. Currently, there are 146 different functions, such as tonify Qi, tonify the blood, tonify the kidney, and tonify the spleen.

### 12.1.3 Active Ingredients in Herbs

As indicated earlier, chemical ingredients/phytochemicals of herbal medicines can be accessed through several programs including TCMSP,<sup>16</sup> Chem-TCM,<sup>14</sup> TCMID,<sup>10</sup> TCM Database@Taiwan,<sup>9</sup> KEGG Genome, Drug-Environ,<sup>17</sup> and HIT.<sup>31</sup> In order to combine herbs, phytochemicals, classifications, and potential therapeutic effects, the WHO monographs on selected medicinal plants, volumes 1–4 were used where the evidence for mechanism of action as well as the level of human evidence can be ascertained.<sup>34</sup> Four of the Chem-TCM<sup>14</sup> categories (Section 12.1.2) were selected based on the level of clinical evidence.

#### 12.1.3.1 Wind Cold

- (WE) diaphoretic, antiviral, antibacterial
- (SS) chills, headache, body and neck pain, no to mild fever
- Representative herb: *Ephedra sinica*

Herba Ephedrae consists of the dried stem or aerial part of *Ephedra sinica* or other ephedrine-containing *Ephedra* species. Medical uses supported by clinical data include the treatment of nasal congestion due to hay fever, allergic rhinitis, acute coryza, common cold, and sinusitis. It is also used as a bronchodilator in the treatment of bronchial asthma. The mechanism for the pharmacological action is the sympathomimetic effects and the release of norepinephrine. Effects include bronchial muscle relaxation due to the activation of  $\beta$ -adrenoceptors. Primary active ingredients include ephedrine and pseudoephedrine.

#### 12.1.3.2 Heat (Blood)

- (WE) refrigerant, styptic, coagulant
- (SS) fever, rash, nosebleed, vomiting blood, blood in stool/urine, skin diseases
- Representative herb: *Rehmannia glutinosa*

Radix Rehmanniae consists of the dried roots of *Rehmannia glutinosa*. Published case reports indicate that the herb is used for the treatment of rheumatoid arthritis and hypertension; however, data from controlled clinical trials are lacking. Information related to the proposed pharmacological actions noted above are taken from traditional medicine pharmacopeia. Primary active ingredients include rehmanniosides A, B, C, and D.

### 12.1.3.3 Tonify Qi

- (WE) endocrine agent, immunostimulant
- (SS) lethargy, weakness, poor appetite, weak voice, pale complexion, breathlessness, immunodeficiency
- Panax ginseng

Radix Ginseng is the dried root of *Panax ginseng*. Medical uses supported by clinical data include as a prophylactic and restorative agent for the enhancement of mental and physical capacities, in cases of weakness, exhaustion, tiredness and loss of concentration, and during convalescence. The mechanism of action is most likely through the hypothalamus–pituitary–adrenal axis and through its immune-stimulant effect. Active ingredients include ginsenosides A1, Rb1, Rb2, Rc, Rd, Re, Rf, Rg1, Rg2.

### 12.1.3.4 Tonify Blood

- (WE) anti-anemic
- (SS) pallor, dizziness, vertigo, poor vision, lethargy, palpitations, amenorrhea, insomnia, pale tongue, fine pulse
- Angelica sinensis

Radix Angelicae Sinensis consists of the dried roots of *Angelica sinensis*. There are no known medicinal uses supported by clinical data, although multiple animal pharmacology studies have been conducted. Active ingredients include angelicide, (Z)-butylidenephthalide,  $\beta$ -cadinene, carvacrol, (E)-ferulic acid, (Z)-ligusticum lactone, (Z)-ligustilide, umbelliferone, valerophenone-*o*-carboxylic acid, and vanillic acid.

## 12.2 Open Access Tools for CDCL Informatics

Several publications have reported the value of a systems pharmacological approach to studying TCM recipes and formulations starting with a detailed network or pathway analysis.<sup>35–42</sup> Wang *et al.*<sup>42</sup> reviewed TCM network pharmacological research highlighting a process involving (1) retrieval of interaction information from various databases, (2) network construction, and (3) knowledge discovery based on network models. The authors present a

summary of all relevant databases for systems/network pharmacology and present detailed network frameworks. Zhang *et al.*<sup>35</sup> used a TCM network pharmacology platform to analyze *Qing Luo Yin* and its use in rheumatoid arthritis, while Ding *et al.*<sup>40</sup> used detailed network pharmacology approaches to study herbal formulations used for cardiovascular diseases. Li *et al.*<sup>36</sup> applied an addition and subtraction theory to herb combinations (as in recipes) to create new formulations based on systems pharmacology analyses using several tools including the TCMSP platform. They applied these methods to two classical prescriptions, *Xiao Chaihu* and *Da Chaihu* decoctions. The findings included 63 bioactive compounds with 65 potential targets and a proposal of dividing the formulations into two aspects: the fundamental formula that is primarily responsible for the basic pharmacological effect, and the additive herbs that exhibit reinforcement functions, thereby achieving a complementary synergistic effect. This can be envisioned as a process to establish new formulations with enhanced synergistic effects. Liu *et al.*<sup>41</sup> used a pathway-based strategy by combining pathway integration, target selection, reverse drug targeting, and network analysis to analyze Reduning injection, a widely used herbal medicine for combating inflammation. Ten key chemical constituents were identified and used for *in silico* assessment and the predicted results were experimentally validated in lipopolysaccharide-stimulated RAW 265.7 cells. The results provide a systemic understanding of the mechanisms of herbal medicines that act on specific disease-related pathways. In a study that illustrates the wide use of several databases and tools, Wang *et al.*<sup>42</sup> used a network-based pharmacological identification of active compounds of *Erxian* decoction used for the alleviation of menopause-related symptoms. Of the total 721 compounds in the medicine, 20 compounds related to 34 significant pathways or genes associated with menopause. The researchers used TCMSP,<sup>16</sup> TCM Database@Taiwan,<sup>9</sup> the Comparative Toxicogenomics Database (CTD),<sup>43</sup> STITCH,<sup>29</sup> and DAVID 6.7<sup>44</sup> for the analyses. This study represents a good example of the use of open access tools for TCM analyses.

Chu *et al.*<sup>45</sup> studied the potential of alleviating side effects of chemotherapy drugs (taxol as an example) with herbal medications in the treatment of breast cancer. Several resources were used including the TCM Database@Taiwan<sup>9</sup> and DrugBank.<sup>18</sup> Five herbs were identified as being used in breast cancer and these are listed with therapeutic functions, primary active phytochemical(s), and potential interactive effects:

- *Dian Huang Qin*: clears heat; baicalein, baicalin; (inhibition of VEGF, cMyc)
- *Dan Shen*: improves blood circulation; baiclin, tanshinones; (inhibition of cyclin D)
- *Shan Ci Gu*: resolves abscesses and masses; colchicine; (tubulin)
- *Bai Zhu*: improves spleen function and appetite; atracydolin; (induction of ghrelin)
- *Dang Gui*: improves blood circulation; ferulic acid; (possible antagonistic effects with taxol involving VEGFA, ER $\alpha$ )

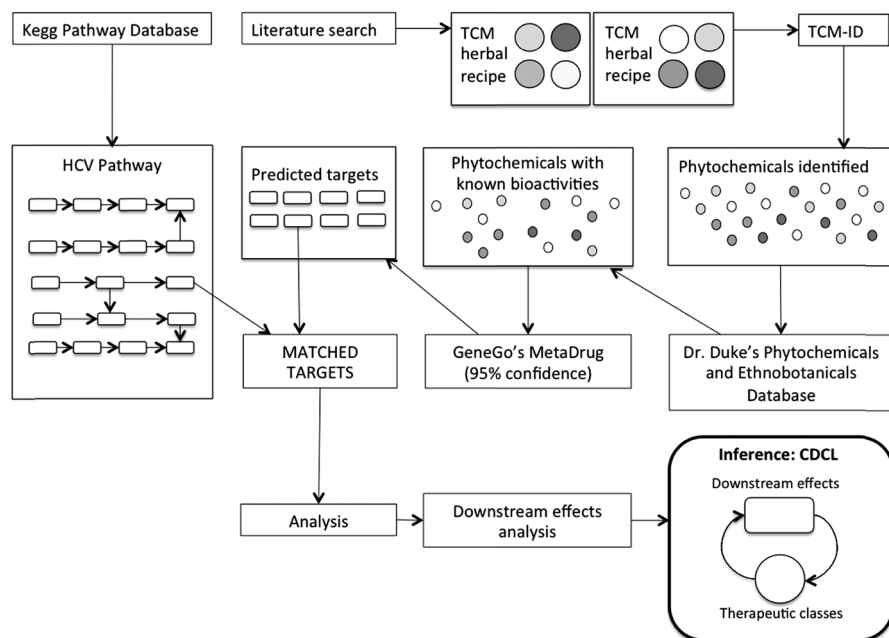
## 12.3 Computational CDCL Studies with Commercial Tools

The following two studies were conducted using a commercially available systems pharmacology platform, MetaDrug (Genego, Thomson Reuters).<sup>46</sup> MetaDrug incorporates extensive manually curated information about chemical compound targets, metabolic fate, ADME properties, and both therapeutic and side effects. Targets come with protein interactions with an exploration of biological pathways including network-neighborhood interactions. In addition, there are more than 70 QSAR models to predict therapeutic activity, ADME properties, and potential toxicity of each compound.

The MetaDrug platform was used by Chan *et al.*<sup>2</sup> to study potential effects of *Kang Ai Pian*, a TCM formulation used to treat cervical, ovarian, breast, nasopharyngeal, lung, liver, and gastrointestinal cancer. Using a systems pharmacology approach, the authors looked at potential positive and negative interactions using the TCM formulation along with representative Western therapeutics. *Kang Ai Pian* is comprised primarily of *chen pi* (*Citrus reticulata*; Citri reticulatae Pericarpium), *huang bo* (*Phellodendron amurense*; Phellodendri Cortex), *huang lian* (*Coptis chinensis*, *Coptis deltoidea* and *Coptis teeta*; Coptidis Rhizoma), *huang qin* (*Scutellaria baicalensis*; Scutellariae Radix), *hu po* (*amber*; Succinum), *niu huang* (*Bos taurus domesticus*; Bovis Calculus) and *san qi* (*Panax notoginseng*; Notoginseng Radix). The active phytochemicals were mapped according to activation and/or inhibition of predicted therapeutic targets and co-mapped with commonly used Western therapeutics used in these diseases. This provided a view of both synergistic and antagonistic interactions and the method can be used to design new formulations as well as decide the most appropriate combinations to be used.

In a hepatitis C (hepC) study by Cheung and Johnson<sup>1</sup> looking at possible CDCL connections, the combination of various herbs and recipes in relation to the known hepC pathway from KEGG Pathways<sup>33</sup> were studied, and as a comparator, peg-interferon was used as a typical Western therapeutic for hepC. Herbs and TCM recipes were selected from Zhao *et al.*<sup>47</sup> who showed that the combination of interferon and various TCM herbs in randomized clinical trials gave a higher sustained virological response, compared to interferon alone (for methodology see Figure 12.2).

These herbs included *Ku Shen* (root of light yellow sophora; Radix Sophorae Flavescentis), *Jianpi Bushen* formula, *Yiganyin*, *Jianpi Huoxue* formula, *Jianpi Qinghua* formula, traditional Chinese drugs I and traditional Chinese drugs II. Herbal ingredients from these selected recipes belong to a total of 11 therapeutic classes from TCM-ID: heat clearance, tonifying weakness, blood activation, promoting diuresis and penetration, bleeding control, blood activation and stasis control, regulation of Qi, dispelling wind-dampness, antitumor, relieving exterior syndrome and astriction.<sup>17</sup> Phytochemical constituents of these herbs were identified with TCM-ID; 210 different phytochemicals from 37 herbs in seven of the 11 therapeutic classes: heat clearance, tonifying weakness, promoting diuresis and penetrating dampness, blood activation and stasis removal, regulation of Qi, dispelling wind-dampness and



**Figure 12.2** Methodology for hepatitis C (hepC) case study. In this study, possible chemical-disease category linkage (CDCL) connections between various herbs and traditional Chinese medicine (TCM) recipes, and the known hepatitis C pathway were scrutinized. Peg-interferon, a typical Western therapeutic for hepC, was used as a comparator. Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, molecular pathways in human cells triggered by the invasion and replication of hepC virus were identified. Several herbs and TCM herbal recipes that were used for treatment of hepC were selected from literature. Specifically, these herbs were shown to provide higher sustained virological response when used in combination with interferon *versus* interferon alone. By inputting the herbal ingredients into TCM-ID and Dr Duke's Phytochemicals and Ethnobotanicals Database, the TCM therapeutic classes and the phytochemicals with bioactivities were identified. The molecular targets of these phytochemicals were then predicted using GeneGo's MetaDrug. Phytochemicals with agonistic or antagonistic effects and the corresponding targets were chosen for further analysis because such binding could lead to downstream events that might affect translation, regulation, and cell proliferation. This analysis not only characterized the molecular interactions between the Western therapeutic and herbal ingredients, but also provided a hypothesis for the possible molecular connections between the TCM herbal therapeutic classes and the corresponding Western therapeutics.

relieving exterior syndromes, were identified to have known bioactivities in relation to hepC therapeutics (Table 12.1).

23 out of these 210 phytochemicals were predicted, using GeneGo's MetaDrug, to interact with 19 molecular targets involved in the hepC pathway, including NF- $\kappa$ B, PKR, p53, EGFR and GSK3 $\beta$  (Table 12.1). The analysis

**Table 12.1** A list of phytochemicals and herbal ingredients found in *Ku Shen* and one of the six herbal recipes selected in the hepatitis C (hepC) study. The corresponding therapeutic class, phytochemicals with bioactivities and those that interact with targets in the hepC pathway for all but one ingredient were identified using various databases and computational resources. In this recipe, available data shows that the seven herbs belong to five traditional Chinese medicine (TCM) therapeutic classes—an example that demonstrates the TCM treatment modality where the TCM practitioners use a mixture of herbs to exert different therapeutic effects and modulate different bodily functions. In the study, 210 different phytochemicals from 37 herbs in seven therapeutic classes were found to have known bioactivities, and 23 of these 210 phytochemicals were predicted to interact with 19 molecular targets in the hepC pathway.

Formula	Herb [common Chinese name (English name; Chinese pharmaceutical name)]	Therapeutic class	Number of phytochemicals identified	Number of phytochemicals with known bioactivities	Phytochemicals with targets found in the HCV pathway
<i>Ku shen</i>	<i>Ku shen</i> (root of light yellow sophora; Radix Sophorase Flavescentis)	Heat clearance	39	17	Lauric acid, 1,8-cineole
<i>Jianpi Bushen formula</i>	<i>Tai zi shen</i> (heterophylly Falsestarwort root; Radix Pseudostellariae)	Tonifying weakness	16	9	β-sitosterol, palmitic acid, fructose
	<i>Fu ling</i> (Indian bread; Porla)	Promoting diuresis and penetrating dampness	2	1	Adenine
	<i>San qi</i> (Sanchi; Radix Notoginseng)	Bleeding control	6	1	
	<i>Dan shen</i> (root of Ligulilobe sage; Radix Salviae Liguliobae)	Blood activation and stasis control	17	9	
	<i>Chi shao</i> (red Peony root; Radix Paeoniae Rubra)	Heat clearance	16	9	β-sitosterol, palmitic acid
	<i>Du zhong</i> (Eucommia bark; Cortex eucommiae)	Tonifying weakness	2	1	
	<i>Tian ji Huang</i> (all-grass of Japanese St John's wort; Herba Hyperici Japonici)	Not available	Not available	Not available	Not available
	<i>Tu si zi</i> (Dodder seed; Semen cuscutae)	Tonifying weakness	1	1	Campesterol

demonstrated that the phytochemicals were not predicted to interact with interferon or targets involved in the interferon stimulating pathway, but rather they bind to other protein targets in the hepC pathway and exert their agonistic or antagonistic effects, which may lead to different downstream events, such as regulation of translation, cell proliferation, and apoptosis. Phytochemicals in herbs that belong to “heat clearance” were shown to interact with EGFR, p53, JNK, RXR $\alpha$ , PPAR $\alpha$  and LXR $\alpha$ , while those in herbs that belong to “tonifying weakness” were shown to interact with EGFR, p53, JNK, IKK $\beta$ , GSK3 $\beta$ , RXR $\alpha$ , PPAR $\alpha$  and LXR $\alpha$ . This examination showed that phytochemicals in a specific therapeutic class may affect multiple target proteins in specific pathways, such as cell survival and lipid metabolism. Phytochemicals in the “tonifying weakness” class were shown to interact mainly with targets responsible for cell survival; however, they were also shown to interact with targets that caused immune responses, such as IKK $\beta$  and GSK3 $\beta$ . For “promoting diuresis and penetrating dampness”, phytochemicals seemed to interact with targets that are responsible for metabolism, in addition to cell survival and immune responses (Table 12.2).

This study provided a possible molecular connection between the TCM herbal classification systems and corresponding Western medicines used to treat the disease in question.

## 12.4 Herb–Drug Interactions

### 12.4.1 Pharmacokinetic Interactions

Several potential interactions between TCM herbs and formulations and Western therapeutics have been reported, and these are typically pharmacokinetic-based, where the phytochemicals in herbs can be substrates, inhibitors, or inducers of various CYP enzymes, thereby changing the pharmacokinetic profiles of the concurrently administered drug or metabolites. Chan<sup>2</sup> described the continual need for and the difficulties in obtaining accurate information on interactions in the practice of integrative medicine. Some of the issues include the evaluation of *in vitro* data, quality of chemicals used in the various assays, and comparison to human case studies. Approaches to meet those needs include the development of the Herbal Ingredient Metabolism database (HIM) by Kang *et al.*<sup>48</sup> The database collects almost all the available *in vivo* metabolism information for herbal active ingredients, as well as their corresponding bioactivity, organs and/or tissues distribution, toxicity, ADME data, and the clinical research profile when possible. Currently, HIM<sup>49</sup> contains 361 ingredients and 1104 corresponding *in vivo* metabolites from 673 reputable herbs. Tools to investigate structural similarity, substructure search and Lipinski’s rule of five are also provided. Various links are also available to PubChem, PubMed, TCM-ID and HIT.

Wu *et al.*<sup>50</sup> reviewed the most common herbs used in TCM formulations and highlighted the current understanding of phytochemicals acting as substrates, inhibitors, or inducers of human CYP enzymes. Cho and Yoon<sup>51</sup>



**Table 12.2** Analysis of selected therapeutic classes and their respective predicted targets.<sup>a</sup>

	Cell survival					Cell survival and immune response			Metabolism, apoptosis and cell cycle		
	EGFR	p53	JNK	p38	ERK1/2	IKK beta	NF-kappa B	PKR	GSK3 beta	PI3K	Akt
Heat clearance	1	1									
Tonifying weakness	7	2	3			3			3		
Promoting diuresis and penetrating dampness	4	1	3	1	2	1	2	1	2	1	1

<sup>a</sup>In this table, the predicted targets were grouped based on the main downstream effects. It shows clearly that herbs in the therapeutic class of “heat clearance” only interact with targets that are responsible for cell survival, while those in the class of “promoting diuresis and penetrating dampness” interact with those responsible for immune response and metabolism, in addition to cell survival.

reviewed the modulation of CYP enzymes and P-glycoprotein by 10 popular medicinal and/or dietary herbs and their phytochemicals in relationship pharmacokinetic interactions. Evidence included *in vitro*, *in vivo*, and human-based assays. Tsai *et al.*<sup>52</sup> reviewed the interactions between anticoagulant/antiplatelet therapeutics and TCM. They discuss 306 documented interactions; 155 were attributable to pharmacodynamic factors; almost all were rated as moderate to severe. In another study, Tsai *et al.*<sup>53</sup> compiled an extensive search and documented herbs and dietary supplement interactions with the concomitant use of drugs. Their data show that formulations and products containing St John's wort, magnesium, calcium, iron, and ginkgo had the highest number of documented interactions, and flaxseed, Echinacea, and yohimbe had the highest number of documented interactions. Western therapeutics affecting the central nervous and cardiovascular systems had the highest documented interactions with herbal medicines. Wanwimolruk *et al.*<sup>54,55</sup> in two publications, review several herb drug interactions, observing that most information derives from animal and *in vitro* studies. The authors highlight the importance of confirming current evidence with clinical studies.

Detailed information on herbal medicines can be found at Medline Plus.<sup>56</sup> Summaries of information include answers to the questions What is it? How effective is it? How does it work? Are there safety concerns? Are there interactions with medications? Are there interactions with herbs and supplements? Are there interactions with foods? What dosage is used? Plus other names, methodology, and references.

MediHerb provides a chart of potential herb–drug interactions for commonly used herbs.<sup>57</sup>

### 12.4.2 Pharmacogenomic-Related Interactions

Although multiple reports have been published on the interactions of herbs and drugs, and most of the interactions involve metabolizing enzymes and transporters, very few reports exist on pharmacogenomic-related interactions. For each of the different enzymes and transporters, there are known polymorphisms that have been shown to affect the intended therapeutic outcomes with Western therapeutics and several of these appear in the approved labeling of these drugs. Liu *et al.*<sup>41</sup> highlighted these polymorphisms and drew correlations with various herbal medications based on the known information based on substrate, inducer, or inhibitor molecular interactions. This research, while very important, still remains at an early stage.

## 12.5 Combination Therapies and Future Directions

Liu *et al.*<sup>58</sup> discuss the history of TCM including the modernization of TCM based on a strong connection with the field of pharmacology. This connection has opened the door for new approaches for TCM-based therapeutics and these are classified by (1) chemistry-focused pharmacological studies,

(2) target-directed pharmacological studies, and (3) systems biology-based pharmacological studies. Whereas (1) and (2) occupy important aspects of current TCM research, (3) has taken on an equally important role. The authors highlight the concern that (3) requires the continual development and update of important technologies and databases, which, if incomplete or insufficiently updated, could reduce the credibility of certain results and conclusions.

Che *et al.*<sup>59</sup> provide a detailed overview of herb–herb combinations used in Chinese medicine practice, highlighting available scientific and clinical evidence supporting combinations. An example of a complementary combination is the decoction Ephedra (*Mahung Tang*), which contains ephedra, cinnamon twig, bitter apricot seed, and licorice root. The combination prescription is used for symptoms of excessive “coldness” and “wind” in the body. Therefore, it is used for its diaphoretic effect and for relief of coughing and asthma. The authors compare this to Western combination medicines with antipyretic, cough suppressant, and nasal decongestant activities.

Yao *et al.*<sup>60</sup> discuss the history and TCM philosophy of combinations of herbs in formulations, noting that TCM mixtures normally contain many active constituents which act on multiple targets. They deciphered the TCM combination principle of “*Jun-Chen-Zuo-Shi*” where the *Jun* herb acts on the main disease targets; the *Chen* herb enhances the *Jun* mechanism of action effect by interacting with common targets of *Jun*, possibly leading to lower doses of the *Jun* herb; and the *Zuo-Shi* herbs improve the bioavailability and decrease the toxicity of the *Jun-Chen* herbs. In their research, they used a common formulation, *Ma-huang*, to develop a protocol to elucidate the mechanisms of action and the systems and poly-pharmacology of the formulation ingredients. They proposed and used a multi-step process, which included (1) molecular database building of chemicals of the four herbs; (2) herb ingredient comparison through measurement of physiochemical properties and chemical structural features; (3) oral bioavailability screening and drug-likeness evaluation; (4) evaluation of interactions with drug metabolizing enzymes; (5) drug targeting through various *in silico* techniques; and (6) network construction and analysis to obtain drug–target interactions and mechanisms of action. This work provides a detailed systems pharmacology approach to connect TCM theory to molecular-based pharmacology.

Kibble *et al.*<sup>61</sup> provided an overview of key network pharmacology-based concepts successfully applied to TCM research and the process of developing effective therapeutic combinations. The authors also discuss the concepts of extending the druggable space of proteins implicated in various complex diseases.

## 12.6 Conclusions

CDCL is a term used to encompass various methods and approaches to link TCM with TCM classifications or categories, diseases, and the molecular basis of both diseases and Western therapeutic interventions.

The development of various systems-pharmacology tools and databases has made this concept a reality within the past 10 years, with more publications describing the systems approaches currently in use. These connections are important particularly from the standpoint of integrative medicine practice, but also from the standpoint of research and development, so as to create a clearer understanding of the synergistic and/or antagonistic interactions of components of formulations and of potential combined use with Western therapeutic treatments.

## References

1. K. Cheung and D. Johnson, Chemical-Disease Category Linkage (CDCL): Methods linking Traditional Chinese Medicine and Western Therapeutics, *Toxicologist*, 2012, **126**(1), 518.
2. E. Chan, M. Tan, J. Xin, S. Sudarsanam and D. Johnson, Interactions between Traditional Chinese Medicines and Western Therapeutics, *Curr Opin Drug Disc Devel*, 2010, **13**(1), 58–73.
3. B. Foster, J. Arnason and C. Briggs, Natural Health Products and Drug Disposition, *Annu. Rev. Pharmacol. Toxicol.*, 2005, **45**, 203–226.
4. S. Bent, Herbal Medicine in the United States: Review of Efficacy, Safety, and Regulation, *J. Gen. Intern. Med.*, 2008, **23**(6), 854–859.
5. www.Europe.chinadaily.com.cn/culture/2012-04/19/content\_15091891.htm.
6. Fulyzaq: <http://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&varApplNo=202292>, Veregan: <http://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=021902>.
7. U. Neill, From branch to bedside: Youyou Tu is awarded the 2011 Lasker-DeBakey Clinical Medical Research Award for discovering artemisinin as a treatment for malaria, *J. Clin. Invest.*, 2011, **121**(10), 3768–3773.
8. D. Normile, Drug Discovery: Nobel for antimalarial drug highlights East-West divide, *Science*, 2015, **350**(6258), 265, DOI: 10.1126/science.350.6258.265.
9. *Traditional Chinese Medicine Database @ Taiwan*, <http://TCM.cmu.edu.tw>.
10. *Traditional Chinese Medicine Integrated Database (TCMID)*, [www.megabionet.org/tcmid/](http://www.megabionet.org/tcmid/).
11. *TCMgeneDIT*, <http://tcm.lifescience.ntu.edu.tw/>.
12. [www.herbmed.org](http://www.herbmed.org).
13. *Dr Duke's Phytochemical and Ethnobotanical Databases*, [www.ars-grin.gov/duke/](http://www.ars-grin.gov/duke/).
14. *Chem TCM*, [www.chemtcm.com](http://www.chemtcm.com).
15. T. Ehrman, D. Barlow and P. Hylands, Phytochemical informatics of Traditional Chinese Medicine and Therapeutic Relevance, *J. Chem. Inf. Model.*, 2007, **47**, 2316–2334.
16. *TCMSP*, <http://lsp.nwsuaf.edu.cn/tcmsp.php>.

17. KEGG, [www.genome.jp/kegg/drug/envron.html](http://www.genome.jp/kegg/drug/envron.html). Search under Medicinal Herbs.
18. *DrugBank*, <http://www.drugbank.ca/>.
19. D. Wishart, C. Knox, A. Guo, S. Shrivastara, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *DrugBank: a comprehensive resource for in silico drug discovery and exploration*, *Nucleic Acids Res.*, 2006, **34**, D668–D672, 16381955.
20. *SuperTarget*, <http://insilico.charite.de/supertarget/>.
21. S. Gunther, M. Kuhn, D. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. Urdiales, A. Gewiess, L. Jensen, R. Schneider, R. Skoblo, R. Russell, P. Bourne, P. Bork and R. Preissner, *SuperTarget and Mator: resources for exploring drug-target relationships*, *Nucleic Acids Res.*, 2007, **36**, D919–D922.
22. *TTD*, <http://bidd.nus.edu.sg/group/cjttd/ttd.asp>.
23. X. Chen, Z. Ji and Y. Chen, *TTD: Therapeutic Target Database*, *Nucleic Acids Res.*, 2002, **30**, 412–415.
24. *MATADOR*, <http://matador.embl.de/>.
25. *PDTD*, [www.dddc.ac.cn/pdtd/](http://www.dddc.ac.cn/pdtd/).
26. Z. Gao, H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, W. Zhu, K. Chen, X. Wang and H. Jiang, *PDTD: a web-accessible protein database for drug target identification*, *BMC Bioinf.*, 2008, **9**, 104, DOI: 10.1186/1471-2105-9-104.
27. *TRMP*, <http://bidd.nus.edu.sg/group/trmp/trmp.asp>.
28. C. Zheng, H. Zhou, B. Xie, L. Han, C. Yap and Y. Chen, *TRMP: a database of therapeutically relevant multiple pathways*, *Bioinformatics*, 2004, **20**, 2236–2241.
29. *STITCH*, <http://stitch.embl.de>.
30. M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. Blicher, C. von Mering, L. Jensen and P. Bork, *STITCH4: integration of protein-chemical interactions with user data*, *Nucleic Acids Res.*, 2014, **42**, D401–D407, DOI: 10.1093/nar/gkt1207.
31. *HIT: Herbal Ingredients' Target Database*, <http://lifecenter.sgst.cn/hit/>.
32. H. Ye, L. Ye, H. Kang, D. Zhang, L. Tao, X. Liu, R. Zhu, Q. Liu, Y. Chen, X. Li and Z. Cao, *HIT: linking herbal active ingredients to targets*, *Nucleic Acids Res.*, 2011, **39**, D1055–D1059.
33. KEGG: Kyoto Encyclopedia of Genes and Genomes, [www.kegg.jp](http://www.kegg.jp), *KEGG Pathway Database*, [www.kegg.jp/kegg/pathway.html](http://www.kegg.jp/kegg/pathway.html).
34. WHO Monographs on selected medicinal plants: World Health Organization, vol I 1999 (28 monographs), vol II 2003 (30 monographs), vol III 2007 (31 monographs), vol IV 2009 (28 monographs); Medicinal Plants Commonly used in New Independent States 2010 (13 new monographs, 17 from previous volumes).
35. B. Zhang, X. Wang and S. Li, *An Integrative Platform of TCM Network Pharmacology and its Application on a Herbal Formula*, Qing-Luo-Yin, *J. Evidence-Based Complementary Altern. Med.*, 2013, **2013**, 456747, DOI: 10.1155/2013/456747.

36. B. Li, W. Tao, C. Zheng, P. Shar, C. Huang, Y. Fu and Y. Wang, Systems Pharmacology-based approach for dissecting the addition and subtraction theory of traditional Chinese medicine: and example using xiao-chaihu-decoction and da-chaihu-decoction, *Comput. Biol. Med.*, 2014, **53**, 19–29.
37. S. Wang, Y. Tong, T.-B. Ng, L. Lao, J. Lam, K. Zhang, Z.-J. Zhang and S. Sze, Network Pharmacological identification of active compounds and potential actions of Erxian decoction in alleviating menopause-related symptoms, *China's Med.*, 2015, **10**, 19, DOI: 10.1186/s13020-015-0051-2.
38. J. Zhao, P. Jiang and W. Zhang, Molecular networks for the study of TCM pharmacology, *Briefings Bioinf.*, 2010, **11**(4), 417–430.
39. T. Ehrman, D. Barlow and P. Hylands, *In silico* search for multi-target anti-inflammatories in Chinese herbs and medicines, *Bioorg. Med. Chem.*, 2010, **18**, 2204–2218.
40. F. Ding, Q. Zhang, C. Ung, Y. Wang, Y. Han, Y. Hu and J. Qi, An analysis of Cemical Ingredients Network of Chinese Herbal Formulations for the Treatment of Coronary Heart Disease, *PLoS One*, 2015, **10**(2), e0116441, DOI: 10.1371/journal.pone.0116441.
41. M.-Z. Liu, Y.-L. Zhang, M.-Z. Zeng, F.-Z. He, Z.-Y. Luo, J.-Q. Luo, J.-G. Wen, X.-P. Chen, H.-H. Zhou and W. Zhang, Pharmacogenomics and Herb-Drug Interactions. Merge of Future and Tradition, *J. Evidence-Based Complementary Altern. Med.*, 2015, **2015**, 456747, DOI: 10.1155/2015/321091.
42. X. Wang, X. Xu, W. Tao, Y. Li, Y. Wang and L. Wang, A systems biology approach to uncovering pharmacological synergy in herbal medicines with applications to cardiovascular diseases, *J. Evidence-Based Complementary Altern. Med.*, 2012, **2012**, 519031, DOI: 10.1155/2012/519031.
43. *Comparative Toxicogenomics Database*, <http://ctdbase.org/>.
44. *DAVID 6.7*, <https://david.ncifcrf.gov/>.
45. A. Chu, M. Lin, D. Loomba, J. Ma and D. Johnson, Computational analysis of the combined therapeutic effects of Traditional Chinese Medicines and Western Therapeutics in breast cancer, *Toxicologist*, 2014, **138**(1), 112.
46. *Genego Metadrag Thomson Reuters*, <http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metacore.html>.
47. S. Zhao, E. Lin, K. Wei, S. Lu, Y. Chu, Y. Li, Y. Wang, B. Huang, Y. Chen and P. Yang, Interferon plus Chinese herbs are associated with higher sustained virological response than interferon alone in chronic Hepatitis C: A meta-analysis of randomized trials, *Antiviral Res.*, 2011, **89**, 156–164.
48. H. Kang, K. Tang, Q. Liu, Y. Sun, Q. Huang, R. Zhu, J. Gao, D. Zhang, C. Huang and Z. Cao, HIM-herbal ingredients in-vivo metabolism database, *J. Cheminf.*, 2013, **5**, 28, <http://www.jcheminf.com/content/5/1/28>.
49. *HIM: Herbal Ingredients In vivo Metabolism Database*, <http://58.40.126.120:8080/him/>.

50. J.-J. Wu, C.-Z. Ai, Y. Liu, Y.-Y. Zhang, M. Jiang, X.-R. Fan, A.-P. Lv and L. Yang, Interactions between phytochemicals from traditional Chinese medicines and human cytochrome P450 enzymes, *Curr. Drug Metab.*, 2012, **13**, 599–614.
51. H.-J. Cho and I.-S. Yoon, Pharmacokinetic interactions of herbs with cytochrome P450 and glycoprotein, *J. Evidence-Based Complementary Altern. Med.*, 2015, 736431, DOI: 10.1155/2015/736431.
52. H.-H. Tsai, H.-W. Liu, Y.-H. Lu, Y.-L. Chen and G. Mahady, A review of potential harmful interactions between anticoagulant/antiplatelet agents and Chinese herbal medicines, *PLoS One*, 2013, **8**(5), e64255, DOI: 10.1371/journal.pone.0064255.
53. H.-H. Tsai, H.-W. Lin, A. Pickard, H.-Y. Tsai and G. Mahady, Evaluation of Documented Drug Interactions and Contraindications associated with Herbs and Dietary Supplements, *Int. J. Clin. Pract.*, 2012, **66**(11), 1056–1078.
54. S. Wanwimolruk and V. Prachayasittikul, Cytochrome P450 enzyme mediated herbal drug interactions (Part1), *EXCLI J.*, 2014, **13**, 347–391.
55. S. Wanwimolruk, K. Phopin and V. Prachayasittikul, Cytochrome P450 enzyme mediated herbal drug interactions (Part 2), *EXCLI J.*, 2014, **13**, 869–896.
56. [www.nlm.nih.gov/medlineplus/](http://www.nlm.nih.gov/medlineplus/).
57. [www.standardprocess.com/MediHerb-Documnet-Library/Catalog-Files/herb-drug-interaction-chart.pdf](http://www.standardprocess.com/MediHerb-Documnet-Library/Catalog-Files/herb-drug-interaction-chart.pdf).
58. X. Liu, W.-Y. Wu, B.-H. Jiang, M. Yang and D.-A. Guo, Pharmacological tools for the development of traditional Chinese medicines, *Trends Pharmacol. Sci.*, 2013, **34**(11), 620–628.
59. C. Che, Z. Wang, M. Chow and C. Lam, Herb-herb Combination for Therapeutic Enhancement and Advancement: Theory, Practice, and Future Perspectives, *Molecules*, 2013, **18**, 5125–5141.
60. Y. Yao, X. Zhang, Z. Wang, C. Zhen, P. Li, C. Huang, W. Tao, W. Xiao, Y. Wang, L. Huang and L. Yang, Deciphering the combination principles of Traditional Chinese Medicine from a systems pharmacology perspective based on Ma-huang decoction, *J. Ethnopharmacol.*, 2013, **150**, 619–638.
61. M. Kibble, N. Saarinen, J. Tang, K. Wennerberg, S. Makela and T. Aittokallio, Network pharmacology applications to map unexplored target space and therapeutic potential of natural products, *Nat. Prod. Rep.*, 2015, **32**, 1249–1266.



## CHAPTER 13

# *Educational Programs for Computational Toxicology and Pharmacology*

DALE E. JOHNSON<sup>\*a,b</sup> AND RUDY J. RICHARDSON<sup>a</sup>

<sup>a</sup>University of Michigan, School of Public Health, Department of Environmental Health Sciences, Ann Arbor, MI 48109-2029, USA;

<sup>b</sup>University of California, Berkeley, Department of Nutritional Sciences and Toxicology, Morgan Hall, Berkeley, CA 94720-3104, USA

\*E-mail: [daleej@umich.edu](mailto:daleej@umich.edu)

## 13.1 Introduction

Computational toxicology is an expanding research endeavor that attempts to meld advances in molecular biology and chemistry with modeling and computational science, in order to increase the predictive power of toxicological assessments. The goals of computational toxicology are to create a more rapid approach to access the potential toxicity of potential new drugs in early stages of research and development, and to create greater efficiency and effectiveness in determining and understanding the hazards and potential risk of the many chemical-related stressors that exist or that will potentially enter the environment. An overall goal is to decrease uncertainties in information (or lack of information) necessary to protect human health and sustainability of the environment. The integration of information from

large-scale datasets utilizing many levels of biological organization and systems biology pathways requires several types of computer tools and modeling techniques, and an intuitive approach to solve problems in new and different ways.<sup>1-4</sup>

In the past several years, we have witnessed rapid advances in computer science, systems biology, chemistry, and other disciplines that continue to enable powerful new computational tools and models highly applicable for toxicology and pharmacology.<sup>4-8</sup> These tools and models hold tremendous promise for advancing applied and basic science, accelerating and streamlining drug efficacy and safety testing,<sup>9-11</sup> and to increase the efficiency and effectiveness of hazard identification and risk assessment for the exposure to environmental chemicals.<sup>3,5,8</sup> These approaches also offer the potential to improve toxicological experimental design, reduce the overall number of experimental toxicology studies needed, and reduce the number of animals used in experimentation. The principles of drug action, covering both efficacy and safety aspects, involve complex interactions, both predicted and experimentally determined between a chemical or biological therapeutic targets and multiple pathways and networks within the body. These *in vivo* phenomena are influenced by disease, overall health status, and the influences of co-exposures to dietary and environmental chemicals and other therapeutics.<sup>9-14</sup>

Regulatory agencies world-wide have established computational toxicology programs internally; there is a National Center for Computational Toxicology established by the United States Environmental Protection Agency (EPA) and computational toxicology groups at the US Food and Drug Administration (FDA) in both the drugs and foods divisions.<sup>3,8,11,15</sup> Computational approaches are ideally suited to organize, process, and analyze the vast libraries and databases of scientific information and to simulate complex biological phenomena.<sup>6,7</sup>

Because of these continuing advances, today's educational programs in toxicology must be adapted to incorporate a wide range of computational tools to give students the ability to innovate and see a glimpse of the future while learning the basics of the science of toxicology.

## 13.2 Historical Context: Computational Toxicology

### 13.2.1 Background

Computational toxicology had its early roots in the 1980s in the process known as combinatorial chemistry, where rapid synthesis or computer simulation of a large number of different but structurally related molecules or materials (by building blocks) were generating large libraries of compounds for initial screening of potential hits against molecular targets. This was accomplished by highly parallel or split-pool chemical synthesis, resulting in the generation of thousands to millions of compounds. Initially, thousands of compounds were present in mixtures (liquid state or solid state)

and de-convolution of the mixtures was accomplished by structural similarity categories and rank order elimination algorithms based on targeted screening of structural analogs. The key lessons in attempting to decipher potential safety concerns in large sets of structure data were that analog identification and categorization were crucial for unknowns and that structural features were related to chemical-biological effects. Early on, structure–activity relationships (SAR) and quantitative structure–activity relationships (QSAR) were found to be useful to fill data gaps, and were particularly useful to rank order individual compounds in a series so as to start the selection process for potential development leads. It was also generally recognized that there was a huge difference in rank ordering compounds and actually predicting endpoints, because of the limitations of chemical space within computational models. It was also determined that proper weighting of endpoint criteria was essential; these weighting criteria for toxicity later became the “filters” used to make decisions on potential drug candidates from analog series, for instance projected electrophilic metabolites as an indicator of potential toxicity and key physicochemical properties that predicted key absorption, distribution, metabolism, and excretion (ADME) properties of chemicals.

### 13.2.2 Programs at University of California Berkeley and University of Michigan

At University of California, Berkeley, it was recognized that a new undergraduate major in molecular toxicology must include computational instruction, and as such computational toxicology lectures were included in basic toxicology courses starting in 2001. Specific computational toxicology courses were started for undergraduates at Berkeley in 2006, and since that time the Berkeley course has been a requirement for molecular toxicology majors in the Department of Nutritional Sciences and Toxicology. In addition, starting in 2015, it is a requirement for majors in nutritional sciences and toxicology with a specialization in toxicology. At University of Michigan, lectures and exercises were included in various toxicology courses since the early 2000s and a specific course was created for graduate students in 2012. At Michigan, the course is taught in the Department of Environmental Health Sciences in the School of Public Health, where students taking the course are typically Masters in Public Health (MPH) candidates. Computational pharmacology approaches are included in a course entitled “Principles of drug action”, started at UC Berkeley in 2011, a course that is open to all undergraduate majors.

In both disciplines, independent study and honors research courses are also offered, which allow students to continue projects after the regular course semester has ended, or pursue projects based on personal interest. At UC Berkeley, starting in 2005, an internship association was started with US FDA, where approximately two students per year work with scientists at the FDA for one or more years on computational toxicology projects. These internships have been in both the drug and foods divisions.

Initially at Berkeley, computational toxicology tools were acquired through donated licenses (short term) by commercial vendors. These included MDL QSAR (MDL Information Systems, Inc.), ADMET Predictor (Simulations Plus, Inc.), Multicase MCWeb (Multicase, Inc.), Leadscope (Leadscope Inc.), and Genego (Thompson Reuters).

Course work and student projects in the first years at Berkeley were centered on QSAR modeling, carcinogenicity and mutagenicity predictions, cardiovascular side effects of drugs, structural alerts for toxicity endpoints, and predictions of chemical persistence and bioaccumulation in the environment.

While these commercial tools were and continue to be highly efficient, feedback from students indicated a preference for programs and tools that could be used after graduation. Accordingly, there was a strong preference for use of open-access tools. In addition, a more systems-based approach was developing, which included the adverse outcome pathway (AOP) and pathway and systems approaches in drug research and development. This included the concepts of target modulation and biomarkers of the induction of desired biological effect(s) and proof of concept on how the biological effect alters disease.<sup>16</sup> At Berkeley, there was also a strong desire to explore and understand herbal or traditional medicines, and particularly their interactions with Western therapeutics and environmental chemical exposures. Currently, the program at Berkeley utilizes Derek and Meteor Nexus (Lhasa Ltd) under an academic license.

### 13.3 Inquiry-Based Science Courses

At Berkeley and Michigan, computational toxicology courses are structured as an inquiry-based experience for hands-on problem solving in toxicology using computational approaches. Students work in small cooperative groups and are given tools, data, and basic concepts to solve toxicity-related environmental, public health, and/or disease-oriented problems in novel ways. The courses are designed to provide students an understanding of the basic principles of computational toxicology and the current methods of predictive toxicology using chemical structures, toxicity-related databases, and biological systems and pathway tools. In addition, students learn how to connect diseases with chemicals and associated genes and to create network pharmacology/toxicology connectivity maps. Most importantly, students gain experience in solving complex problems in new ways.

Students use a case-study approach to learn the application of computer technology and mathematical/computational models used to analyze, model and/or predict potential toxicological effects from chemical structure (parent compound or metabolites); inference from similar compounds; exposure, bioaccumulation, and persistence in the environment; differential indicators or patterns related to exposure (biomarkers); and networks of biological pathways affected by a chemical. During the semester students work to further understand mechanisms of toxicity, particularly in relation to organism,

organ, and disease specificity. Students also probe why certain individuals, ethnic groups, or populations are more susceptible to chemical exposures, draw associations between chemical exposure and increased risk for certain diseases, and gain an understanding of naturally occurring chemicals and particularly interactions between phytochemicals and Western therapeutics. By the end of the semester, students have identified a problem to solve and completed a project worthy of presentation at local and national scientific meetings.

Computational pharmacology instruction at UC Berkeley is designed to provide students with a basic understanding of the principles of pharmacology, pharmacokinetics, and drug safety. In addition, students learn basic principles of drug discovery and development from concept to reality and gain an appreciation of current research issues important to discovering new therapeutics. The course is taught in a problem-based learning structure where students take on real-world challenges similar if they were practitioners of the discipline. Collaborative groups are formed, with the view that different students from a variety of majors bring different intellectual currency to solving problems. The ability to fuse ideas—especially ideas that cross disciplines and expertise—is a crucial aspect of accelerating innovation. The course is structured as an entrepreneurial-style model where the instructor and students create a mythical biotechnology start-up company and form drug discovery groups studying different disease categories. Groups learn from lectures, discussions, and supplied materials and present a proposal of what disease(s) and drug candidate(s) the biotech company should focus on. Each proposal includes filing a mythical investigational new drug application with the FDA, which includes the design of the first clinical trial. Students experience and learn target-based, physiology-based, and clinical precedence-based drug discovery in a collaborative setting. In addition, they gain an understanding of disease mechanisms, disease genes, and important polymorphisms potentially associated with efficacy and toxicity. They discover the design characteristics that lead to drug-like molecules and explore critical drug interactions. They also gain a background in pharmacokinetics and dive into pharmacogenomics and personalized medicine. Students also create a “toolbox” of publicly available websites where chemical structures can be matched to potential protein targets and disease pathways, and this information can be used to suggest different delivery techniques. The course gives students a “life-cycle” approach in understanding and experiencing the concepts of health care and therapeutic interventions. Individual groups make presentations at the end of the year on their proposed new “drug” to instructors and a panel of outside biotech scientists.

### 13.4 Current Computational Toxicology Courses

An assumption is made that students will have a basic understanding of toxicology.

### 13.4.1 Toxicology Tutorials

As a review, or for those students without a didactic background in toxicology, the following tutorials are useful. The following tutorials are provided as open-access websites as a public service by the US Department of Human Health Services, National Library of Medicine, National Institutes of Health, Environmental Health and Toxicology Specialized Information Services:<sup>17</sup>

- Toxicology Tutor I: Basic Principles<sup>18</sup>
- Toxicology Tutor II: Toxicokinetics<sup>19</sup>
- Toxicology Tutor III: Cellular Toxicology<sup>20</sup>

ToxLearn: A Gateway to Toxicology is an open-access joint project of the US National Library of Medicine Toxicology and Environmental Health Information Program and the US Society of Toxicology:<sup>21</sup>

- ToxLearn Module I: Introduction to Toxicology and Dose-Response<sup>22</sup>
- ToxLearn Module II: Cells and Tissue: Injury and Repair<sup>23</sup>

### 13.4.2 Course Concepts

In both courses the case-study approach allows students to explore and use different techniques as well as start to a collaborative approach within study and project groups. During this stage when groups are formed, each group creates a communication network such as a Google group or through networks through university course websites, and students are also given group evaluation criteria and forms to help expedite communication and workflow. Case studies are tailored to topics and potential projects of interest to students during the semester or as indicated by students in instructor-student meetings.

A typical lecture series combined with relevant case studies will explore these concepts:

- The application of computer technology and mathematical/computational models to analyze, model, and/or predict potential toxicological effects from:
  - chemical structure (parent compound or metabolites)
  - inference from similar compounds
  - published data (typically not human data)
  - exposure, bioaccumulation, persistence
  - biomonitoring data
  - plasma or tissue concentrations
  - differential indicators (biomarkers) or patterns of outcome related to exposure
  - networks of biological pathways affected by the chemical
  - correlations to known human diseases

- To further understand mechanisms of toxicity
  - organ specific
  - organism specific
  - disease specific
- To explain why certain individuals are more susceptible
- Key methods
  - chemical fragment or structural similarities (structural alerts)
  - categorization or grouping for inference (*i.e.* similar compounds causing similar effects): analogs, categories based on mechanism, mode of action
  - structure–activity relationships (SARs, QSARs)
  - biological pathway perturbations
  - biomarkers of exposure, susceptibility, outcome

### 13.4.3 Case Studies

Examples of case studies associated with relevant topics are presented.

#### 13.4.3.1 *Chemical Structural Features Determine Biological Effects*

Underlying theory: there is a connection between a chemical structure and its interaction with biological systems. Students learn the sources of information and the different identification systems and apply this in an environmental assessment.

##### 13.4.3.1.1 Case Study Example 1.

- Start with these compounds:
  - bisphenol A
  - dibutylphthalate
  - 1,3-butadiene
  - benzene
- Chemical information queries require at least three notations:
  - CAS#, SMILES code, SDF. Obtain each for each chemical using PubChem,<sup>24</sup> ChemIDPlus,<sup>25</sup> or ChemSpider<sup>26</sup>
  - search toxicology information using PubChem<sup>24</sup> compound summary and various categories of ToxNet<sup>27</sup>
- Classify each compound by hazard traits such as carcinogenicity, reproductive hazard, persistence in environment, hepatotoxicity, neurotoxicity, *etc.*



– give brief evidence

- Discuss the physicochemical properties that cause it to be persistent or not, bioaccumulative, and/or toxic to the environment using PBT Profiler<sup>28</sup>
- Note the household products that contain the compound—Household Products Database<sup>29</sup>
- In particular, note how you can move from one information source to another

**13.4.3.1.2 Case Study Example 2.** There is wide agreement that chemicals of great concern are those that: persist (P), bioaccumulate (B), and present toxicity (T) concerns, *i.e.* PBTs. Only a fraction of all chemicals have been tested to determine if they are PBTs. PBT profiling is aimed at all raw materials, as an important predictive screen. Students are given links to global PBT concerns including those from the US, European Union, and Canada and use the public access to the PBT Profiler<sup>28</sup> and links and tutorials to ChemIDPlus,<sup>25</sup> ToxNet,<sup>27</sup> Cactus,<sup>30</sup> Pubchem,<sup>24</sup> and ChemSpider.<sup>26</sup>

- Exercise
- use naphthalene as the core structure
- substitute various functional groups in the 1 and 2 positions
- see how lipophilicity changes with structural substitutions
- create a spreadsheet with chemical name, SMILES code, CAS number, and SDF file
- use the name, SMILES code, and CAS number to compare the environmental impact of each compound in PBT Profiler (EPA)
- in ChemID Plus<sup>25</sup> and ToxNet<sup>27</sup> find the carcinogenicity and mutagenicity results for each compound if present in database
- use the SDF file to enter the compound (or multiple compounds) in a commercial software package such as Genego<sup>31</sup>
- compare a few chemical properties that may be involved in the difference in environmental impact, and/or carcinogenicity or mutagenicity findings in a program such as ToxTree<sup>32</sup>
- compare the “predicted” Ames test in the predictive models with the actual results in ChemID Plus<sup>25</sup>
- if there are no carcinogenicity results for any individual compound, predict the findings using real or predicted mutagenicity findings

#### 13.4.3.1.3 Case Study Example 3.

- Select two out of the three chemicals listed below and add a third chemical of your choice: butylbenzyl phthalate (BBP), diethyl hexyl phthalate (DEHP), and perfluorooctane sulfonic acid (PFOS)
- Use the CTD,<sup>33</sup> T3DB,<sup>34</sup> and Chemspider<sup>26</sup> and develop a chemical–disease–gene linkage for a relevant toxicological endpoint for each chemical

- Use STITCH<sup>35</sup> to display interactions
- For a relevant zip code, city, or county (EPA Tri,<sup>36</sup> Scorecard<sup>37</sup>) determine the most predominant environmental contaminant and using T3DB<sup>34</sup> and CTD,<sup>33</sup> suggest a linkage to toxicity and/or disease

#### 13.4.3.1.4 Case Study Example 4.

- Expand case study 3 into chemical–disease pathway analyses
- Find a relevant disease pathway using tools below and form a network toxicology diagram and discuss the feasibility of the chemical disease linkage:
- KEGG pathway database: wiring diagrams of molecular interactions, reactions, and relations<sup>38</sup>
- Pathway Commons<sup>39</sup>
- Wiki Pathways<sup>40</sup>
- Pathway Maps. Life Science Research<sup>41</sup>
- Quick search for pathways using Google image: search using the chemical name, disease, and the word pathway. Important pathway images will appear and these will be linked directly to the information source.
  - Draw the new pathway using yED graph editor from the yWorks Diagraming Company<sup>42</sup>

#### 13.4.3.1.5 Typical Report for Case Studies (One Per Team if Applicable).

Introduction (explain what you are attempting to do)

Methods (explain all compound substitutions and the programs used to evaluate)

Results (show results of all compounds in table format)

Show figures of pathways if applicable

Discussion (compare results from different compounds and suggest why there are differences between compounds)

Conclusion

### 13.4.3.2 Herbal Traditional Medicines

#### 13.4.3.2.1 Case Study Example 5.

- Using herbal and traditional medicine resources (listed below), select either a single herb or combination of herbs (as in a recipe)
- discuss the potential treatment of a condition and/or disease
- list the relevant phytochemicals
- determine if the targeted phytochemicals have potential positive effect on the condition
- show a gene–target interaction diagram

- Resources

- KEGG Environ<sup>43</sup>
- Traditional Chinese Medicine Integrated Database: TCMID<sup>44</sup>
- TCMgeneDIT:<sup>45</sup> TCMs, genes, diseases, effects, and ingredients
- Traditional Chinese Medicine Database@Taiwan<sup>46</sup>
- Natural Center for Complementary and Alternative Medicines (NCCAM): herbs at a glance<sup>47</sup>
- Medline Plus: herbs and supplements<sup>48</sup>
- Traditional Chinese Medicine Systems Pharmacology Database (TCMSP) and analysis platform<sup>49</sup>

### 13.4.3.2.2 Case Study Example 6.

- Using an example from case study 5 suggest a drug–herb interaction that may involve drug metabolism or individual differences in response
- Show the potential interaction in pathway form and discuss whether this is a known interaction or hypothesized from your work

### 13.4.3.3 Environmental Chemicals and Health Relationships

**13.4.3.3.1 Case Study Example 7.** A current topic is selected that recently appeared in the scientific literature or the news media. This topic changes each year. The following is an example.

NBC News recently did an investigative report on the safety of our children's playgrounds and sports fields. For more than 5 years now, there have been a slew of studies looking to see if there is a link between the rise of cancer cases among athletes and the surfaces they play on.<sup>50</sup> Artificial turf fields can be found all over the United States as the popular alternative to natural grass fields which require a lot more upkeep; however, the chemicals involved in making the turf has been drawing some concern.<sup>51,52</sup>

Of the most concern is artificial grass that uses infill called crumb rubber made of pieces of old tires. This is what makes these fields more bouncy, protecting players better from more serious injuries like concussions. The downside is that when players' bodies connect with the field, little pieces of the infill fly up and scatter everywhere. After a game or practice, players will find these crumbs not only in their uniforms, but in their hair and inside cuts and abrasions they received on the field. For soccer goalies however, with their full body dives to the ground, these crumbs can also get in their mouths during play.

Because different types of tires made from a large variety of materials are used to make the turf, it is hard to pinpoint the particular chemicals that could be causing the increase in cancer rates. The EPA lists mercury, lead, benzene and arsenic as well as other chemicals and carcinogens as being the ingredients in tires.

- You are asked to comment scientifically about this issue offering an informed opinion
- Discuss what you know about this issue
- Discuss what you need to know to form the informed opinion

## 13.5 Course Projects

### 13.5.1 Starting Projects

- Students work in small groups (three students optimal)
- Identify a problem (relevance to the group is optimal)
- Organize ideas and prior knowledge
  - what do we know?
- Pose questions
  - what do we need to know?
- Assign responsibility for investigating aspects of question
  - discuss resources and approach
  - defining project question is the midterm assignment
- Research the question: analyze and summarize findings
- Innovate: try new approaches
- Integrate new information and keep refining the question
- Present solution to the problem
  - project presentation; final report

### 13.5.2 Typical Project Categories

- Chemical exposure focus (environmental risk)
  - chemicals, consumer products, hazard traits associated with chemicals, diseases
- Disease focus (disease–chemical–exposure linkage)
  - linkage to chemical exposure(s)
  - geographic clusters: disease and/or chemical exposure
- Therapeutics focus (drug safety/side effects)
  - adverse drug reactions
  - susceptible populations

- Phytochemical, herbal medicine focus (systems pharmacology/toxicology)
  - traditional Chinese medicines and Western therapeutics
- Other

### 13.5.3 Therapeutics vs. Environmental Chemicals

- Therapeutics
  - extensive biochemical and cell-based screening
  - full battery of animal testing
  - extensive clinical data
  - detailed dose/exposure data exists
  - risk is always balanced by benefit
- Environmental chemicals
  - major data gaps in animal testing
  - human data sporadic, primarily epidemiological, based on unintentional exposure
  - risk is not balanced by benefit
- Both deal with toxicity and adverse outcome pathways

### 13.5.4 Challenges in Computational Toxicology

- Delineating initiating events in toxicity pathways
- Determining minimal concentrations at which biological events occur
- Understanding relevant metabolic transformations of compounds and the proximal toxicant
- Discovering biomarkers of key events
- Determining susceptibility indicators

#### 13.5.4.1 Hazard-Based Information Gathering

- Identify molecular targets or biological pathways linked to toxicity
  - mode of action (MOA)/AOP anchor 1
  - chemicals perturbing these can lead to an adverse event
  - find information on assays used to probe molecular initiating events for these targets or pathways
  - information linking a chemical to key assays can lead to a prediction of MOA/anchor 1

- Look for predictive *in silico* models
  - prioritize chemicals based on the hazard of interest
  - suggest/distinguish possible AOP/MOA for chemical(s)

## 13.6 Sample Project: The Chemical of Concern Question

Question: Do certain chemicals reach a level of concern for a specific disease, and would this trigger further testing and/or action?

This type of project offers students a unique opportunity to apply course methodology to a defined human health issue; in this example environmental risk factors for human breast cancer. Within the breast cancer “knowledgebase” are compounds known to increase the incidence or risk of breast cancer in humans and biological pathways identified as important for this chemically induced disease. Ideally, a screening process to develop information useful for making regulatory decisions is essential to make an impact on the disease. However, there are >80 000 compounds identified as environmental contaminants. Thousands of “new” compounds enter the environment (including consumer products, drugs, and foods) each year, so identifying compounds of concern that may increase the risk of human breast cancer to enter the screening program is a critical issue.

The goal of this project is to evaluate *in silico* methods for identifying and prioritizing compounds of concern. Several categorization schemes can be suggested for compounds, such as (1) hormonal activity, (2) reactivity to form toxicophores, and (3) potential for localization in breast tissue and milk. In addition, metabolism and transport to breast tissue should be characterized, as well as the predicted ability to interact with known breast cancer targets and key metabolic pathways. Realistically, several of these compounds will not have the data to be categorized effectively, and methods to “fill the gaps” need to be proposed.

Students use certain tools and data sources such as the National Health and Nutrition Examination Survey (NHANES)<sup>53</sup> biomonitoring data, presence in breast tissue/milk, hormonal activity, compound reactivity, modulation of critical metabolic enzymes or transporters in breast tissue, and interaction with known human breast cancer targets. Candidate compounds identified would reflect possible mechanistic causality, and those identified could then enter a pipeline for subsequent rigorous testing through the experimental methods-based algorithm.

A systematic process for filling information gaps has to be established using a number of software programs including the ability to predict key molecular targets and activating pathways for each compound from large sets of compounds. In addition, the potential for each compound to activate key metabolic enzymes expressed in breast tissue needs to be assessed. This complex algorithm should provide valuable assistance in the development of a roadmap to guide chemical screening for agents likely to be involved in

breast cancer development and progression. Application with several data-sets will refine the algorithm and it is projected that this scheme will form the basis of future screens for other human diseases and/or location specific effects of environmental contaminants.

The project would develop in the following way.

### 13.6.1 Health Effects Inquiry

- Requirements for inquiry
  - establish disease-specific endpoints
  - correlate with molecular network system
- Collation and organization of data available for the compound
  - assessment of quality of data
  - define data gaps in core endpoints for each chemical
  - support endpoints with data and/or risk evaluations
- Define strategy to predict toxicity
- Data gaps “filled” by use of *in silico* methods
- Does level of concern require further testing and/or action?

### 13.6.2 Endpoints for Breast Cancer

- Core endpoints
  - classification as a human carcinogen
  - positive findings in animal carcinogenicity studies
  - site-specific findings
  - positive in mutagenicity studies
  - known endocrine disruptor
  - structural alerts for reactivity
  - persistence in the environment
- Secondary endpoints
  - biomonitoring evidence of human exposure
  - interaction or modulation of key biological pathways
  - biotransformation effects that increase hazard
  - presence in breast tissue or milk
- Risk endpoints
  - chemical use/manufacturing data
  - exposure to susceptible populations



### 13.6.3 Project Question

Can a breast cancer MOA be postulated for an unknown compound?

- Related to chemical structure?
- Hormonal (*i.e.* xenoestrogen)
- Forms reactive metabolite in breast tissue?
- Decreases detoxification pathways?
- Does projected human exposure suggest a risk?
- Is there a susceptible population particularly at risk?
- Is there supporting information?
- Potential localization in breast tissue/milk
  - physio-chemical properties
- Potential hormonal activity
  - contains active pharmacophore? – visual inspection/QSAR
  - QSAR estrogen receptor
  - docking to active forms of receptors
- Reactivity (toxicophore formation)
  - metabolite modeling
  - mutagenicity – QSAR
  - carcinogenicity – QSAR
  - activate or inhibit key metabolic enzymes
  - pathway mapping to breast cancer molecular network

## 13.7 Course Projects Presented and Published

### 13.7.1 Projects Presented at National Scientific Meetings

#### 13.7.1.1 *Society of Environmental Toxicology and Chemistry (SETAC)*

- Analysis of persistent polycyclic aromatic hydrocarbons in the Exxon-Valdez oil spill using computational systems biology<sup>54</sup>
- *In silico* modeling of manganese- and iron-binding properties of a designed peptide inspired by the radioresistant bacterium *Deinococcus radiodurans*<sup>55</sup>

#### 13.7.1.2 *National Society of Toxicology Through 2015*

- Decision trees for the Organisation for Economic Co-operation and Development (OECD) QSAR Toolbox, a primary local QSAR model builder for green chemistry initiatives<sup>56</sup>

- Environmental chemicals and breast cancer: geographic association between livestock waste, contaminated surface water, and breast cancer incidence in the San Francisco Bay Area<sup>57</sup>
- Chemical–disease category linkage methods linking traditional Chinese medicine and Western therapeutics<sup>58</sup>
- A computational analysis of the mechanistic relationship between cigarette smoking and risk of rheumatoid arthritis<sup>59</sup>
- An examination of the combined effects of environmental factors and susceptible genes on increasing the risk of Alzheimer's disease<sup>60</sup>
- Lead interaction with three key haplotypes of delta-aminolevulinic acid dehydratase, vitamin D receptor, and human leukocyte antigen may affect susceptibility to multiple sclerosis<sup>61</sup>
- Role of beta-methylamino-L-alanine in GRIK1-mediated amyotrophic lateral sclerosis disease pathway<sup>62</sup>
- Developing a gene–gene interaction network for nonsyndromic orofacial clefts: connecting vitamin A and folic acid metabolic pathways with orofacial cleft gene candidates<sup>63</sup>
- Depleted uranium induction of lowered levels of folate in mothers increases the risk of neural tube defects in infants<sup>64</sup>
- Tea and atherosclerosis: an examination of green tea, oolong tea, and black tea and their effects on cholesterol biosynthesis and atherosclerosis<sup>65</sup>
- Arsenic and pancreatic cancer: a computational systems toxicology analysis of risk<sup>66</sup>
- Computational analysis of the combined therapeutic effects of traditional Chinese medicines and Western therapeutics in breast cancer<sup>67</sup>
- Computational analysis of potential interactions from the combined use of Western therapeutics and traditional Chinese medicines in postmenopausal osteoporosis<sup>68</sup>
- LocaTox: an interactive tool linking known environmental contaminants, associated chemical–human gene interactions, and birth defect statistics in specific locales<sup>69</sup>
- Computational identification of active phytochemicals and potential mechanisms for the alleviation of bleomycin-induced pulmonary fibrosis<sup>70</sup>
- Computational alternative analysis: network relationship of structural motifs for chemicals of concern correlated with health, ecological, and lifecycle impacts<sup>71</sup>
- Computational investigation of the combination of traditional Chinese medicine and Western therapeutics for the treatment of non-small cell lung cancer<sup>72</sup>
- A computational analysis of the potential environmental and lifestyle risk factors associated with breast cancer incidence in South Napa, California<sup>73</sup>
- A computational analysis of ethnicity-specific polycystic ovarian syndrome treatments using Western and traditional medications<sup>74</sup>

- Computational analysis of active phytochemicals and potential synergism with Western therapeutics in the treatment of Parkinson's disease<sup>75</sup>
- Computational analysis of coffee constituents and their potential neuroprotection action in Alzheimer's disease<sup>76</sup>
- Computational comparison of the anti-inflammatory targets of the traditional Chinese medicine *Sargentodoxa cuneata* (Hong Teng) and Western therapeutics for the treatment of osteoarthritis<sup>77</sup>

### 13.7.2 Projects from the Courses Published in Journals

- Computational toxicology: heading toward more relevance in drug discovery and development<sup>78</sup>
- Clustering and its application in multi-target prediction<sup>79</sup>
- CYP2C9 polymorphisms: considerations in NSAID therapy<sup>80</sup>
- Current HIV therapeutics: mechanistic and chemical determinants of toxicity<sup>81</sup>
- Interactions between traditional Chinese medicines and Western therapeutics<sup>82</sup>
- Ethanol toxicity in breast cancer<sup>83</sup>
- Evidence for a potential protective effect of carnitine-pantothenic acid co-treatment on valproic acid-induced hepatotoxicity<sup>84</sup>

## 13.8 Computational Pharmacology as Part of the Principles of Drug Action

As mentioned earlier, computational pharmacology instruction at UC Berkeley is designed to provide students with a basic understanding of the principles of pharmacology, pharmacokinetics, and drug safety. In addition, students learn basic principles of drug discovery and development from concept to reality and gain an appreciation of current research issues important to discovering new therapeutics. The first classroom project is to establish research and development divisions and form group discovery teams. The following divisions have been selected by students since 2011.

- Standard research and development divisions
  - cardiovascular disease
  - respiratory disease
  - oncology
  - metabolic disease
  - infectious disease
  - neurological disease
  - autoimmune disease

As an example of student innovation, the following are selected previous proposals developed during the class.

- Cardiovascular
  - a novel treatment for hypercholesterolemia: ApoE replacement + LCAT activator
  - an anti-antisense therapy for hypercholesterolemia
- Respiratory
  - a novel cationic shell-cross linked lipid nanoparticle encased IL-4R $\alpha$  siRNA asthma therapeutic targeting the IL-4/IL-13/STAT6 pathway
  - blocking IgE responses in allergic rhinitis
- Oncology
  - small molecule inhibitor of BRFZ-RNA Pol III binding as a novel therapeutic for squamous cell lung carcinoma
  - a small molecule targeting TRPM-7 for pancreatic cancer with delivery *via* liposomal nanoparticles
  - a peptide ligand (Pep42) conjugated to ixabepilone that binds to GRP78 on cancer cells, internalizes, and is cleaved by cathepsin B to deliver the cytotoxin inside cancer cells
  - a nanoparticle-tethered peptide targeting the NF- $\kappa$ B pathway in late-stage cancer patients with cachexia
  - ovarian cancer treatment and diagnosis based on nanotube technology
  - anti-pancreatic cancer therapy *via* a CDK-inhibitor combination
  - human mammary tumor virus therapy through a novel gene editing tool
  - jemsgydumab: a proposed novel therapeutic agent for melanoma
  - BlockX: an oral medication that treats acute myeloid leukemia through the inhibition of a tyrosine kinase receptor
- Metabolic disease
  - a small molecule targeting the NR2C subunit of the NMDA receptor as a treatment of type 2 diabetes through the CNS
  - inverse vaccination to inhibit an auto-immune response in type I diabetes: novel adjuvant linked to GAD65
  - development of a  $\beta$ 3 adrenergic receptor agonist targeting perilipin A as a treatment for obesity through the upregulation of lipolysis
  - a bi-functional small molecule targeting ACC2 and SOCS3 for the treatment of type 2 diabetes

- an inhibitory peptide therapeutic for celiac disease targeting the HLA-DQ2 receptor on intestinal antigen presenting cells
- controlling cholesterol synthesis *via* mechanisms involving lanosterol synthase
- lowering cholesterol with a combination LXR agonist and FXR antagonist
- type II diabetes therapy through a novel GSK-3 inhibitor
- Infectious disease
  - a di-thiol compound-based drug to combat lipodystrophy, hepatic steatosis, and mitochondrial toxicity among HIV patients taking anti-retroviral drugs
  - aptamer-based targeting of NS4B protein for the treatment of hepatitis C
  - a sepsis pro-drug: maltodextrin-linked antibiotic delivered to bacterial maltodextrin receptor with cleavage by bacterial protease DegP
  - a meningitis B therapeutic targeting and disabling the CNS factor H binding protein *via* a nanoparticle delivery system
  - targeting the HBx antigen of HBV patients by utilizing shRNA and pegylated interferons
  - novel treatment of tuberculosis using a UDP-galactofuranose inhibitor
- Neurological disease
  - targeting beta-amyloid plaques in Alzheimer's disease *via* activation of CR-1 with transcranial magnetic stimulation and cannabinoid compounds
  - a novel treatment of Alzheimer's disease targeting beta-secretase 1 with a pro-drug in a time-release dermal patch. BBB passage facilitated by attachment to a camelid VHH antibody
  - targeting the PSEN-1 subunit of  $\gamma$ -secretase for Alzheimer's disease
  - a novel approach to the treatment of Parkinson's disease with a preventative analog of curli chaperone
- Autoimmune disease
  - A novel drug administered *via* an injectable implant for psoriasis
  - development of a dual Toll-like receptor 7 and 9 antagonist for the treatment of systemic lupus erythematosus

## 13.9 Conclusion

The value of inquiry-based learning experiences in computational toxicology and pharmacology for both undergraduate and graduate students is highlighted and confirmed by the high level of classroom innovation

at two universities. By definition, the methods cannot remain static year-to-year, but must match the rapid advances in computer science and systems biology, and the emergence of more advanced open-access tools each year. Solving relevant, real-life problems as uncovered and developed by the students themselves in the class provides a rewarding learning experience.

## References

1. B. A. Fowler, *Computational Toxicology. Methods and Applications for Risk Assessment*, ed. B. A. Fowler, Academic Press-Elsevier, London, 2013, pp. ix–xiv.
2. B. Reisfeld and A. Mayeno, *Computational Toxicology, Methods in Molecular Biology*, ed. B. Reisfeld and A. Mayeno, Humana Press, Springer Science, NY, 2013, pp. 3–7.
3. R. J. Kavlock, G. Aukley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R. W. Setzer, I. Shah, D. Villeneuve and E. Weber, Computational Toxicology – A State of the Science Mini Review, *Toxicol. Sci.*, 2008, **103**, 14–27.
4. S. Ekins, Progress in Computational Toxicology, *J. Pharmacol. Toxicol. Methods*, 2014, **69**, 115–140.
5. I. Rusyn and G. P. Daston, Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century, *Environ. Health Perspect.*, 2010, **118**, 1047–1050.
6. J. Zhang, J.-H. Hsieh and H. Zhu, Profiling Animal Toxicants by Automatically mining Public Bioassay Data: The Big Data Approach for Computational Toxicology, *PLoS One*, 2014, **9**, 1–11.
7. M. Bouhifd, H. T. Hogberg, A. Kleensang, A. Maertens, L. Zhao and T. Hartung, Mapping the Human Toxome by Systems Toxicology, *Basic Clin. Pharmacol. Toxicol.*, 2014, **115**, 24–31.
8. R. Judson, Public Databases Supporting Computational Toxicology, *J. Toxicol. Environ. Health*, 2010, **13**, 218–231.
9. W. Munster, A. Breidenbach, H. Fischer, S. Kirchner, L. Muller and A. Pahler, Computational Toxicology in Drug Development, *Drug Discovery Today*, 2008, **7/8**, 303–310.
10. C. Merlot, Computational Toxicology – A Tool for Early Safety Evaluation, *Drug Discovery Today*, 2010, **15**, 16–22.
11. L. G. Valerio, *Toxicol. Appl. Pharmacol.*, 2009, **241**, 356–370.
12. A. Mostrag-Szlichtyng, J.-M. Z. Comenges and A. P. Worth, Computational Toxicology at the European Commission's Joint Research Centre, *Expert Opin. Drug Metab. Toxicol.*, 2010, **6**, 785–792.
13. A. Roncaglioni, A. A. Toropov, A. P. Toropova and E. Benfenati, *In Silico Methods to Predict Drug Toxicity*, *Curr. Opin. Pharmacol.*, 2013, **13**, 802–806.
14. A. A. Toropov, A. P. Toropova, I. Raska, D. Leszczynska and J. Leszczynski, Comprehension of Drug Toxicity: Software and Databases, *Comput. Biol. Med.*, 2014, **45**, 20–25.

15. L. Valerio, *In silico Toxicology Models and Databases as FDA Critical Path Initiative Tools*, *Hum. Genomics*, 2011, 5, 200–207.
16. H. Larson, E. Chan, S. Sudarsanam and D. E. Johnson, in *Biomarkers in Computational Toxicology, Methods in Molecular Biology*, ed. B. Reisfeld and A. Mayeno, Humana Press, Springer Science, NY, 2013, Chapter 11, pp. 253–273.
17. U.S. Department of Health and Human Services, *Specialized Information Services*, <https://sis.nlm.nih.gov/enviro.html>.
18. *Toxicology Tutor I*, <http://sis.nlm.nih.gov/enviro/toxtutor/Tox1/index.html>.
19. *Toxicology Tutor II*, <http://sis.nlm.nih.gov/enviro/toxtutor/Tox2/index.html>.
20. *Toxicology Tutor III*, <http://sis.nlm.nih.gov/enviro/toxtutor/Tox3/index.html>.
21. *ToxLearn*, <https://toxlearn.nlm.nih.gov>.
22. *ToxLearn Module I*, <https://toxlearn.nlm.nih.gov/Module1.htm>.
23. *ToxLearn Module II*, <https://toxlearn.nlm.nih.gov/Module2.htm>.
24. *PubChem Project*, [Pubchem.ncbi.nlm.nih.gov](http://pubchem.ncbi.nlm.nih.gov).
25. *ChemID Plus*, [chem.sis.nlm.nih.gov/chemidplus](http://chem.sis.nlm.nih.gov/chemidplus), Lite version: <https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp>, Advanced version: <https://chem.nlm.nih.gov/chemidplus/>.
26. *ChemSpider*, [www.chemspider.com](http://www.chemspider.com).
27. *ToxNet Toxicology Data Network*, [toxnet.nlm.nih.gov](http://toxnet.nlm.nih.gov).
28. *PBT Profiler*, [www.pbtprofiler.net](http://www.pbtprofiler.net).
29. *Household Products Database*, <https://householdproducts.nlm.nih.gov>.
30. *Cactus*, <https://cactus.nci.nih.gov/ncidb2.2/>.
31. *Thomson Reuters Systems Biology Solutions*, <https://portal.genego.com>.
32. *ToxTree*, [Toxtree.sourceforge.net](http://Toxtree.sourceforge.net).
33. *Comparative Toxicogenomics Database*, [ctdbase.org](http://ctdbase.org).
34. *Toxin and Toxin – Target Database; Toxic Exposome Database*, [www.t3db.ca](http://www.t3db.ca).
35. *STITCH: Chemical Protein Interactions*, [stitch.embl.de](http://stitch.embl.de).
36. *Toxics Release Inventory (TRI) Program US EPA*, <https://www.epa.gov/toxics-release-inventory-tri-program>.
37. *Scorecard*, The Pollution Information site, [www.Scorecard.goodguide.com](http://www.Scorecard.goodguide.com).
38. *KEGG Pathways*, <http://www.genome.jp/kegg/pathway.html>.
39. *Pathway Commons*, [www.pathwaycommons.org](http://www.pathwaycommons.org).
40. *WikiPathways*, [www.wikipathways.org](http://www.wikipathways.org).
41. *Pathway Maps*, <http://pathwaymaps.com>.
42. *yED graph editor*, [www.yworks.com](http://www.yworks.com).
43. *KEGG Environ*, [www.genome.jp/kegg/drug/enviro.html](http://www.genome.jp/kegg/drug/enviro.html).
44. *TCMID*, <http://www.megabionet.org/tcmid/>.
45. *TCMGeneDIT*, <http://tcm.lifescience.ntu.edu.tw/>.
46. *Traditional Chinese Medicine Database@Taiwan*, <http://tcm.cmu.edu.tw/>.
47. *NCCAM.NIH.gov/health/herbsataglance.htm*.
48. *Medline Plus, Herbs and Supplements*, [https://www.nlm.nih.gov/medline-plus/druginfo/herb\\_All.html](https://www.nlm.nih.gov/medline-plus/druginfo/herb_All.html).
49. *TCMSP*, <http://lsp.nwsuaf.edu.cn/tcmsp.php>.



50. *Fact Sheet. Crumbs – Rubber Infilled Synthetic Turf Athletic Fields*, New York State Department of Health, 2008, [https://www.health.ny.gov/environmental/outdoors/synthetic\\_turf/crumb-rubber\\_infilled/docs/fact\\_sheet.pdf](https://www.health.ny.gov/environmental/outdoors/synthetic_turf/crumb-rubber_infilled/docs/fact_sheet.pdf).
51. D. Brown, eEnvironment and Human Health, Inc, Project Report, 2007, [www.ehhi.org](http://www.ehhi.org).
52. M. Llompart, L. Sanchez-Prado, J. Lamas, C. Garcia-Jares, E. Roca and T. Dagnac, Hazardous Organic Chemicals in Rubber Recycled Tire Playgrounds and Pavers, *Chemosphere*, 2013, **90**, 423–431.
53. *Fourth National Report on Human Exposure to Environmental Chemicals*, 2015, <http://www.cdc.gov/exposurereport>.
54. R. A. Brown and D. E. Johnson, Analysis of persistent polycyclic aromatic hydrocarbons in the Exxon-Valdez oil spill using computational systems biology, *SETAC Annu. Meet.*, 2009.
55. D. M. Faulkner, M. A. Bradley, S. J. Wijeyesakere, D. E. Johnson and R. J. Richardson, *In silico* modeling of manganese- and iron-binding properties of a designed peptide inspired by the radioresistant bacterium *Deinococcus radiodurans*, *SETAC Annu. Meet.*, 2013.
56. S. Maxey, G. Mao, J. Chan, H. W. Kim and D. E. Johnson, Decision trees for the OECD Toolbox, a primary local QSAR model builder for green chemistry initiatives, *Toxicologist*, 2010, **114**(1), 44.
57. H. Larson, M. Mun, W. Onyenwe and D. E. Johnson, Environmental Chemicals and Breast Cancer: Geographic association between livestock waste, contaminated surface water, and breast cancer in the San Francisco Bay area, *Toxicologist*, 2011, **120**(2), 49.
58. K. W. K. Cheung and D. E. Johnson, Chemical-Disease category linkage (CDCL): methods linking Traditional Chinese Medicine and western therapeutics, *Toxicologist*, 2012, **126**(1), 518.
59. L. Cai, J. Lu, S. Ng and D. E. Johnson, A computational analysis of the mechanistic relationship between cigarette smoking and risk of rheumatoid arthritis, *Toxicologist*, 2012, **126**(1), 169.
60. C. Ho, L. Feng and D. E. Johnson, An examination of the combined effects of environmental factors and susceptible genes on increasing the risk of Alzheimer's Disease, *Toxicologist*, 2012, **126**(1), 170.
61. F. Su, M. Gooding, G. Anderson and D. Johnson, A computational analysis of environmental factors potentially associated with multiple sclerosis susceptibility, *Toxicologist*, 2013, **132**(1), 25.
62. S. Firouzbakht, Y. Iizuka and D. Johnson, Role of beta-methylamino-L-alanine in GRIK1-mediated amyotrophic lateral sclerosis disease pathway, *Toxicologist*, 2013, **132**(1), 25.
63. J. McKone, A.-A. N. Van, E. Lachenauer and D. Johnson, Developing a gene-gene interaction network for nonsyndromic orofacial clefts using computational analysis, *Toxicologist*, 2013, **132**(1), 25.
64. S. Aghaee, H. Ngo and D. Johnson, Computational analysis of a potential mechanistic relationship between depleted uranium (DU) exposure and risk of Spina Bifida Cystica, *Toxicologist*, 2013, **132**(1), 25.

65. C. Lei, M. Nhan, C. Ha and D. Johnson, Computational analysis of catechins in teas and potential relationship to cardiovascular health, *Toxicologist*, 2013, **132**(1), 25.
66. R. Coronel, C. Han, K. Kan, P. Nguyen and D. Johnson, Arsenic and Pancreatic Cancer: A Computational Systems Toxicology Analysis of Risk, *Toxicologist*, 2014, **138**(1), 603.
67. A. Chu, M. Lin, D. Loomba, J. Ma and D. Johnson, Computational Analysis of the Combined Therapeutic Effects of Traditional Chinese Medicines and Western Therapeutics in Breast Cancer, *Toxicologist*, 2014, **138**(1), 112.
68. R. Chang and D. Johnson, Computational Analysis of Potential Interactions from the Combined Use of Western Therapeutics and Traditional Chinese Medicines in Postmenopausal Osteoporosis, *Toxicologist*, 2014, **138**(1), 112.
69. K. Aquino, K. Hald, M. Reyna, M. Zamora and D. Johnson, LocaTox: An Interactive Tool Linking Known Environmental Contaminants, Associated Chemical-Human Gene Interactions, and Birth Defect Statistics in Specific Locales, *Toxicologist*, 2014, **138**(1), 109.
70. N. Chan, E. Perez, S. Li and D. E. Johnson, Computational Identification of Active Phytochemicals and Potential Mechanisms for the Alleviation of Bleomycin-Induced Pulmonary Fibrosis, *Toxicologist*, 2014, **138**(1), 605.
71. Y. He, B. Ford, N. Winje and D. Johnson, Computational Alternative Analysis: Network Relationship of Structural Motifs for Chemicals of Concern Correlated with health, Ecological, and Lifecycle Impacts, *Toxicologist*, 2015, **144**(1), 286.
72. J. Dovedy, A. Chen, A. Ly, M. Lucia and D. Johnson, A Computational Investigation of the Combination of Traditional Chinese Medicine (TCM) and Western Therapeutics for the Treatment on Non-Small Cell Lung Cancer (NSCLC), *Toxicologist*, 2015, **144**(1), 281.
73. A. Goldbaum, R. Sood, R. Sail and D. Johnson, A Computational Analysis of the Potential Environmental and Lifestyle Risk Factors Associated and Breast Cancer Incidence in South Napa, California, *Toxicologist*, 2015, **144**(1), 280.
74. S. Ng, E. Qiu, A. Chang, L. Phan and D. Johnson, A Computational Analysis of Ethnicity-Specific Polycystic Ovarian Syndrome Treatments Using Western and Traditional Medicines, *Toxicologist*, 2015, **144**(1), 280.
75. V. Ho, Y. Han, Y. Aldaher and D. Johnson, Computational Analysis of Active Phytochemicals and Potential Synergism with Western Therapeutics in the Treatment of Parkinson's Disease, *Toxicologist*, 2015, **144**(1), 280.
76. T. Bradford, V. Man, M. Garg, N. Wong and D. Johnson, Computational Analysis of Coffee Constituents and Their Potential Neuroprotection Action in Alzheimer's Disease, *Toxicologist*, 2015, **144**(1), 280.
77. G. Liang, X. Huang, R. Trimble, C. Park and D. Johnson, Computational Comparison of the Anti-Inflammatory Targets of the Traditional Chinese Medicine *Sargentodoxa Cuneata* (Hong Teng) and Western Therapeutics for the Treatment of Osteoarthritis, *Toxicologist*, 2015, **144**(1), 279.

78. D. E. Johnson and A. D. Rodgers, Computational toxicology: Heading toward more relevance in drug discovery and development, *Curr. Opin. Drug Discovery Dev.*, 2006, **9**(1), 29–37.
79. W. Liu and D. E. Johnson, Clustering and its application in multi-target prediction, *Curr. Opin. Drug Discovery Dev.*, 2009, **12**(1), 98–107.
80. Z. K. Ali, R. J. Kim and F. M. Ysla, CYP2C9 polymorphisms: considerations in NSAID therapy, *Curr. Opin. Drug Discovery Dev.*, 2009, **12**(1), 108–114.
81. S. Chiao, D. Romero and D. E. Johnson, Current HIV Therapeutics: mechanistic and chemical determinants of toxicity, *Curr. Opin. Drug Discovery Dev.*, 2009, **12**(1), 53–60.
82. E. Chan, M. Tan, J. Xin, S. Sudarsanam and D. E. Johnson, Interactions between Traditional Chinese Medicines and Western Therapeutics, *Curr. Opin. Drug Discovery Dev.*, 2010, **13**(1), 58–73.
83. R. Chan, Ethanol Toxicity in Breast Cancer, *Berkeley McNair Res. J.*, 2011, **18**, 45–61.
84. D. Felker, A. Lynn, S. Wang and D. Johnson, Evidence for a Potential Protective Effect of Carnitine-Pantothenic Acid Co-treatment on Valproic Acid-Induced Hepatotoxicity, *Expert Rev. Clin. Pharmacol.*, 2014, **7**(2), 211–218.

# Subject Index

- absorption, metabolism,  
distribution, excretion, and  
toxicity (ADMET), 38
- ACD Percepta  
([www.acdlabs.com](http://www.acdlabs.com)), 45–46
- acetylcholinesterase (AChE)  
inhibition, 79
- activity landscapes, 131–137
- adenosine deaminase (ADA), 179,  
180
- ADMET. *See* absorption, meta-  
bolism, distribution, excretion,  
and toxicity (ADMET)
- ADMET Predictor  
([www.simulations-plus.com](http://www.simulations-plus.com)), 46
- adverse outcome pathway (AOP)  
vs. aggregate exposure  
pathway (AEP), 80–81
- approach, 2, 10
- concept, 193–194
- environmental exposure,  
73–75
- key event relationships  
(KERs), 64
- organizing principle, 64–67  
data resources for,  
70–73
- knowledge  
management, 67–68
- phases of, 68–70
- pharmacokinetic  
considerations, 73–75
- Aggregated Computational  
Toxicology Resource  
(ACToR), 24
- aggregate exposure pathway (AEP)  
vs. adverse outcome  
pathway (AOP), 80–81
- data resources for, 79–80
- organizing principle, 75–79
- ALTBIB, 24
- Analog Identification  
Methodology (AIM), 49
- ArrayExpress, 72
- benchmark dose (BMD), 195
- BindingDB Binding Database, 30
- Biological General Repository  
for Interaction Datasets  
(BioGrid), 27
- blood–brain barrier (BBB), 38–39
- Caco-2 cell lines, 39
- CACTVS/PubChem keys, 123
- CAESAR, 29
- CAM. *See* Complementary and  
alternative medicine (CAM)
- Carcinogenic Potency Database  
(CPDB), 24, 217
- CDCL. *See* chemical–disease  
category linkage (CDCL)
- Chemaxon Suite (Marvin Sketch  
and Metabolizer)  
([www.chemaxon.com/](http://www.chemaxon.com/)), 49

- Chemical Abstracts Service (CAS), 22
- Chemical and Product Categories database (CPCat), 24
- Chemical Carcinogenesis Research Information System (CCRIS), 24
- chemical carcinogens, 217
- chemical database of traditional chinese medicine (Chem-TCM), 244, 246
- chemical-disease category linkage (CDCL)
- classifications, 283–286
  - with commercial tools, 289–292
  - databases for, 281–283
  - herbs, active ingredients in, 286–287
  - open access tools for, 287–288
- chemical-disease inference system (ChemDIS), 27
- Chemical Effects in Biological Systems (CEBS), 24–25
- chemical-gene-protein databases, 26–27
- Chemicalize (www.chemicalize.com), 49
- Chemical Similarity Network Analysis Pulldown (CSNAP), 140
- Chemical Space Networks (CSN), 141
- chemical spaces, 131–137
- chemical toxicities
- biomarkers, 8
  - living system, 37
  - potential adverse reactions, 5–6
  - safer chemical design, 6–7
  - single-target toxicity concepts, 3–4
- ChemIDPlus, 21–22
- Chemmapper, 30
- ChemProp, 28
- Chemprot, 27
- ChemSpider
- (www.chemspider.com), 28, 49
- Chem-TCM, 282
- Comparative Molecular Field Analysis (CoMFA), 145
- Comparative Molecular Similarity Indices Analysis (CoMSIA) method, 147
- Comparative Toxicogenomics Database (CTD), 72
- Complementary and alternative medicine (CAM), 281
- computational toxicology and pharmacology
- chemical structural features determine biological effects, 306–308
  - course concepts, 305–306
  - educational programs, 302–303
    - course projects, 310–312
    - drug action, 316–318
    - sample project, 312–314
  - environmental chemicals and health relationships, 309–310
  - herbal traditional medicines, 308–309
  - historical context, 301–302
  - inquiry-based science courses, 303–304
  - toxicology tutorials, 305
- University of California Berkeley, 302–303
- University of Michigan, 302–303

- Consolidated Human Activity Database (CHAD), 24  
 CORAL, 28  
 CPDB. *See* Carcinogenic Potency Database (CPDB)  
 CSN. *See* Chemical Space Networks (CSN)  
 CTD. *See* Comparative Toxicogenomics Database (CTD)  
 cytidine, 132, 133  
 Cytoscape, 28
- Danish (Q)SAR database, 28  
 DART, 22  
 database for pharmacokinetic properties (PK/DB), 30  
 databases and tools  
   systems toxicology  
     additional toxicity databases, 25  
     chemical-gene-protein databases, 26–27  
     chemistry, 28–29  
     drug and drug target databases, 30  
     national toxicology program databases, 24–25  
     pathway-network databases, 27–28  
     QSAR databases and tools, 28–29  
     structural alert, 28–29  
     TOXicology Data Network (TOXNET), 21–24  
     US EPA chemical toxicity databases, 24  
     websites with extensive links to, 31
- Derek and Meteor Nexus from Lhasa Limited ([www.lhasalimited.org](http://www.lhasalimited.org)), 47  
 Directory of Useful Decoys (DUD), 141  
 Directory of Useful Decoys, Enhanced (DUD-E), 141  
 DITOP, 26  
 drug and drug target databases, 30  
 DrugBank database, 30  
 drug interactions with phytochemicals  
   enhancement of absorption, 114–115  
   induction of metabolism, 113–114  
   inhibition of absorption, 115  
   inhibition of metabolism, 112–113  
   pharmacodynamic interactions, 115–116  
 DrugMatrix, 25  
 DSSTox, 24  
 dynamic flexibility, 144
- eChemPortal, 218  
 Economic Co-operation and Development (OECD), 66  
 educational programs, 302–303.  
   *See also* computational toxicology and pharmacology  
     course projects, 310–312  
     drug action, 316–318  
     sample project, 312–314  
 Environmental Protection Agency (EPA), 7, 22, 37, 62  
 environmental toxicology  
   adverse outcome pathway, 9–10  
   expanding exposure concepts, 10–11  
   exposome, 11

- European Chemicals Agency (ECHA) database for REACH, 218
- European Inventory of Existing Commercial Chemical Substances (EINECS), 22
- Extended Connectivity Fingerprints (ECFP), 123
- FAERS, 25
- fluorine substitution, 135
- Gene Expression Omnibus (GEO), 72, 254
- Genetic Toxicology Data Bank (GENE-TOX), 24
- Gene-tox, 218
- genotoxicity, 38–39
- green molecular design (GMD)
  - case study, 50–51
  - decision trees, 44
  - design of, 51–54
  - expert systems, 39–44
  - QSAR tools, 44–45
  - representative tools, 45–50
- Hazard Evaluation Support System (HESS), 220
- Hazardous Substances Data Bank (HSDB), 21
- Haz-Map, 23
- herbal and traditional medicines. *See* traditional Chinese medicine (TCM)
- Herbal Ingredients' Targets (HIT) database, 283
- herb–drug interactions
  - pharmacogenomic-related interactions, 294
  - pharmacokinetic interactions, 292–294
- high-throughput toxicokinetics (HTTK) models, 10
- Historical Control Databases, 25
- human intestinal absorption (HIA), 150
- Human Metabolome Database (HMDB), 72
- Human Protein Atlas, 27
- iCSS Dashboard, 24
- indoline, 52
- in silico* chemical–protein docking
  - bisphenol-A to estrogen-related receptor- $\gamma$ , 180–182
  - 1,3-dinitrobenzene (1,3-DNB), 179–180
  - genetic algorithms, 177
  - inverse docking, 179
  - matching algorithms, 177–178
  - molecular dynamics
    - organophosphorous compounds, 185–187
    - simulations, 182–183
    - trajectories, 183–185
  - Monte Carlo procedure, 177
  - scoring ligand poses, 178
- in silico* approaches, 210–211
  - assessing experimental data, 227–228
  - documenting, 233
  - expert alerts, 221–223
  - multiple systems, 228–230
  - QSARs, 223–226
  - read-across, 226–227
  - result, 230–232
- integrated approaches to testing and assessment (IATA), 81
- Integrated Risk Information System (IRIS), 23, 218
- International Agency for Research on Cancer (IARC), 22, 218



- International Conference on  
Harmonisation (ICH), 43
- International Consortium for  
Innovation and Quality  
in Pharmaceutical  
Development (IQ), 3
- International Programme on  
Chemical Safety (IPCS), 2, 62
- International Toxicity Estimates  
for Risk (ITER), 23
- International Uniform  
Chemical Information  
Database (IUCSID), 216
- in vitro* approaches, 210–211
- in vivo* approaches, 210–211
- iPathwayGuide, 28
- Japan Existing Chemical Data  
Base (JECDB), 26, 218
- key event relationships (KERs),  
194
- Kyoto Encyclopedia of Genes and  
Genomes (KEGG), 26, 27
- LactMed, 22
- lamivudine, 132, 133
- Lazy Structure–Activity  
Relationships (LAZAR), 28
- Leadscope Drugs Chronic/  
Sub-chronic Database, 220
- Leadscope Food Safety Chronic/  
Subchronic Database, 220
- Leadscope Food Safety Genetox  
Database, 218
- local lymph node assay  
(LLNA), 81
- lowest observed adverse effect  
levels (LOAELs), 195
- MACCS keys, 123
- Madin-Darby Canine Kidney  
Epithelial cells (MDCK), 39
- Matador, 283
- Matched Molecular Pair (MMP)  
analysis, 149
- Medchem Designer, 46–47
- Medical Subject Headings  
(MeSH®), 26
- methyl substitution, 134
- Mobyle@RPBS (<http://mobyle.rpbs.univ-paris-diderot.fr>),  
49–50
- molecular docking  
genetic algorithms, 177  
matching algorithms,  
177–178  
Monte Carlo procedure, 177
- Molecular Field Topology  
Analysis (MFTA), 124, 145
- molecular initiating events  
(MIEs), 42, 193
- molecular similarity  
activity prediction, 139–141  
applicability of, 122  
applications of, 138–142  
clustering, 141–142  
concept, 121–122  
diversity, 141–142  
functions, 125–131  
graph-based  
representations, 127  
networks, 141–142  
set-based representations,  
126  
structure representations  
for, 122–124  
virtual screening, 138–139
- National Institute of Allergy and  
Infectious Diseases  
(NIAID), 22
- National Institute of Standards  
and Technology (NIST), 22
- National Institutes of Health  
(NIH), 7, 281

- National Research Council (NRC), 2, 61, 198
- National Toxicology Program (NTP) database, 218, 220
- national toxicology program databases, 24–25
- NDEx, 28
- no observed adverse effect levels (NOAELs), 195
- NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM LLMA Database), 25
- OECD QSAR Toolbox, 48
- OECD Toolbox, 28
- Offsides, 26
- Online Chemical Database, 29
- ordinary differential equations (ODE), 91
- Pathway Commons, 28
- pathway-network databases, 27–28
- PBPK modeling, 196–197
- pharmacodynamic effects, mathematical description of direct effects and effect compartments, 106–107 indirect effects, 107–109
- pharmacokinetic and toxicokinetic models compartmental models, 92–96 K-PD models, 110–111 mathematical description of, 105–109 physiological models, 96–98 PK/PD and TK/TD modeling, 109–110 PK/PD relationships, 102–104 structural models, 92–98 variance models, 98–102
- pharmacokinetic properties (PK/DB), database for, 30
- pharmacostatistical modeling, 99
- Pharmmapper, 30
- phytochemicals, drug interactions with enhancement of absorption, 114–115 induction of metabolism, 113–114 inhibition of absorption, 115 inhibition of metabolism, 112–113 pharmacodynamic interactions, 115–116
- Potential Drug Target Database (PDTD), 30, 283
- proton-exchange membrane (PEM) fuel cells, 50
- PubChem, 28
- Qikprop ([www.schrodinger.com/QikProp](http://www.schrodinger.com/QikProp)), 48
- QSAR databases and tools, 28–29
- quantitative structure–activity/property relationships (QSAR/QSPR), 142–143 activity towards individual targets, 152–153 ADMET properties, 150–152 congeneric series, 143–148 consistent mechanisms, 143–148 diverse series and big data, 148–153 physico-chemical properties, 153 potential drug targets, 152 web-based QSAR/QSPR services, 153

- quantitative structure–activity relationship (QSAR), 74
- quantitative structure–toxicity relationship (QSTR), 53, 197–198
- quantitative structure–transcription–assay relationships (QSTAR), 6
- Quantitative Systems Pharmacology (QSP), 12, 13
- Rapid Overlay of Chemical Structures (ROCS), 144
- Reactome, 27
- Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), 82
- Registration, Evaluation, Authorisation, and Restriction of Chemical Substances (REACH), 2, 192
- Registry of Toxic Effects of Chemical Substances (RTECS), 218, 220–221
- RepDose, 220
- SAR modeling, 197–198
- Search Tool for Interacting Chemicals (STITCH), 27
- Search Tool for Interactions of Chemicals (STITCH), 283
- Side Effect Resource (SIDER), 25
- Similarity Ensemble Approach (SEA), 140
- source structure–activity data, 149
- Structure–Activity Landscape Index (SALI), 135
- structure–activity relationships (SAR), 194
- SuperTarget, 283
- systems and network pharmacology identification and pursuance, 12
- potential adverse effects, 14
- secondary pharmacology and off-target effects, 13
- systems toxicology, 2–3
- additional toxicity databases, 25
- chemical–gene–protein databases, 26–27
- chemistry, 28–29
- drug and drug target databases, 30
- national toxicology program databases, 24–25
- pathway–network databases, 27–28
- QSAR databases and tools, 28–29
- structural alert, 28–29
- TOXicology Data NETwork (TOXNET), 21–24
- US EPA chemical toxicity databases, 24
- tanshinone IIA (TIIA), 256
- TCM. *See* traditional Chinese medicine (TCM)
- Therapeutically Relevant Multiple Pathways (TRMP), 283
- Therapeutic Target Database (TTD), 30, 283
- ToxAlerts, 29
- ToxFx, 25
- toxicity databases
- acute and repeated dose toxicity, 220–221
- carcinogenicity, 217–219
- database organization, 216–217
- developmental toxicity, 219–220
- genetic toxicity, 217–219

- reproductive toxicity, 219–220
- Toxicity Estimation Software Tool (TEST), 28
- toxicity pathways, computational modeling of
  - case studies, 203–205
  - concept of, 198–199
  - education on, 205
  - model toxicity pathways, 200–203
  - pathway modeling software tools, 205–206
  - purpose of, 199–200
- Toxicity Reference Database (ToxRefDB), 24
- Toxicogenomics Project Genomics Assisted Toxicity Evaluation System (TG-GATES), 72
- TOXicology Data Network (TOXNET), 21–24
- Toxic Substances Control Act (TSCA), 22
- Toxin and Toxin Target Database (T3DB), 25
- TOXMAP, 22–23
- ToxRead, 29
- Toxtree (<http://toxtree.sourceforge.net/>), 29, 48–49
- traditional apical endpoints testing, quantitative methods in, 194–196
- traditional Chinese medicine (TCM)
  - and chemical constituents, 282
  - Chem-TCM (chemical database of traditional chinese medicine), 244, 246
  - classification and systems approach, 282
  - in clinical trials, 271–273
  - combination therapies, 294–295
  - herb–drug interactions, 268–269
  - HIT (linking herbal active ingredients to targets), 245–246
  - modernization of, 263
  - network pharmacology, 264–268, 273–274
  - omics data in
    - genomics in, 253–254
    - metabonomics in, 256–257
    - proteomics in, 256
    - transcriptomics in, 254–255
  - pathway analysis, 283
  - pharmacogenomics in, 269–271
- TCMGeneDIT, 248–249
- TCMID, 249–251
- TCMSP (traditional chinese medicine systems pharmacology database and analysis platform), 246–248
- therapeutic targets and protein interactions, 283
- TTD (therapeutic target database), 251–253
- Western therapeutics, 283
- Traditional Chinese Medicine Information Database (TCM-ID), 282
- Traditional Chinese Medicine Systems Pharmacology (TCMSP), 282
- tramadol pharmacokinetics, 101
- TRI, 23
- Twosides, 26

US EPA chemical toxicity  
databases, 24

Vaccine Adverse Event Reporting  
System (VAERS), 26  
ventricular tachyarrhythmia (VT), 6

Virtual Computational Chemistry  
Laboratory, 28  
Vitic from Lhasa Limited, 218

World Health Organization  
(WHO), 280