

Wiley Series on Technologies for the
Pharmaceutical Industry

Sean Ekins, Series Editor

COMPUTATIONAL TOXICOLOGY

RISK ASSESSMENT FOR CHEMICALS

EDITED BY SEAN EKINS

WILEY

The background of the cover is a complex collage of molecular models. It features various representations of chemical structures: ball-and-stick models with grey, white, red, and blue spheres; wireframe skeletal structures; and semi-transparent yellow and purple mesh surfaces. The models are scattered across the white background, creating a sense of depth and scientific complexity.

Computational Toxicology

Wiley Series on Technologies for the Pharmaceutical Industry **Sean Ekins, Series Editor**

Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals
Edited by Sean Ekins

Pharmaceutical Applications of Raman Spectroscopy
Edited by Slobodan Sašić

Pathway Analysis for Drug Discovery: Computational Infrastructure and Applications
Edited by Anton Yuryev

Drug Efficacy, Safety, and Biologics Discovery: Emerging Technologies and Tools
Edited by Sean Ekins and Jinghai J. Xu

The Engines of Hippocrates: From the Dawn of Medicine to Medical and Pharmaceutical Informatics
Barry Robson and O.K. Baek

Pharmaceutical Data Mining: Applications for Drug Discovery
Edited by Konstantin V. Balakin

The Agile Approach to Adaptive research: Optimizing Efficiency in Clinical Development
Michael J. Rosenberg

Pharmaceutical and Biomedical Project Management in a Changing Global Environment,
Edited by Scott D. Babler

Systems Biology in Drug Discovery and Development
Edited by Daniel L. Young and Seth Michelson

Collaborative Computational Technologies for Biomedical Research
Edited by Sean Ekins, Maggie A.Z. Hupcey and Antony J. William

Predictive Approaches in Drug Discovery and Development: Biomarkers and In Vitro/ In Vivo correlations
Edited by J. Andrew Williams, Richard Lalonde, Jeffrey Koup and David D. Christ

Collaborative Innovation in Drug Discovery, Strategies for Public and Private Partnerships
Edited by Rathnam Chaguturu

Computational Toxicology: Risk Assessment for Chemicals
Edited by Sean Ekins

Computational Toxicology

Risk Assessment for Chemicals

Edited by

Sean Ekins

Collaborations Pharmaceuticals, Inc. Raleigh, USA

WILEY

This edition first published 2018
© 2018 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Sean Ekins to be identified as the Editor in this work has been asserted in accordance with law.

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data:

Names: Ekins, Sean, editor.

Title: Computational toxicology : risk assessment for chemicals / edited by Sean Ekins.

Description: First edition. | Hoboken, NJ : John Wiley & Sons, 2018. |

Series: Wiley series on technologies for the pharmaceutical industry |

Includes bibliographical references and index. |

Identifiers: LCCN 2017037714 (print) | LCCN 2017046458 (ebook) | ISBN 9781119282570 (pdf) | ISBN 9781119282587 (epub) | ISBN 9781119282563 (cloth)

Subjects: LCSH: Toxicology--Mathematical models. | Toxicology--Computer simulation. | QSAR (Biochemistry)

Classification: LCC RA1199.4.M37 (ebook) | LCC RA1199.4.M37 C66 2018 (print) | DDC 615.90285--dc23

LC record available at <https://lccn.loc.gov/2017037714>

Cover Design: Wiley

Cover Images: (Front cover) Courtesy of Daniela Schuster; (Author photo) Courtesy of Sean Ekins

Set in 10/12pt WarnockPro by SPi Global, Chennai, India

*I should have no objection to go over the same life from its beginning to the end:
requesting only the advantage authors have, of correcting in a second edition
the faults of the first.*

Benjamin Franklin

To my family and collaborators.

Contents

List of Contributors *xvii*

Preface *xxi*

Acknowledgments *xxiii*

Part I Computational Methods 1

- 1 Accessible Machine Learning Approaches for Toxicology** 3
Sean Ekins, Alex M. Clark, Alexander L. Perryman, Joel S. Freundlich, Alexandru Korotcov, and Valery Tkachenko
- 1.1 Introduction 3
- 1.2 Bayesian Models 5
- 1.2.1 CDD Models 7
- 1.3 Deep Learning Models 13
- 1.4 Comparison of Different Machine Learning Methods 16
- 1.4.1 Classic Machine Learning Methods 17
- 1.4.1.1 Bernoulli Naive Bayes 17
- 1.4.1.2 Linear Logistic Regression with Regularization 18
- 1.4.1.3 AdaBoost Decision Tree 18
- 1.4.1.4 Random Forest 18
- 1.4.1.5 Support Vector Machine 19
- 1.4.2 Deep Neural Networks 19
- 1.4.3 Comparing Models 20
- 1.5 Future Work 21
- Acknowledgments 21
- References 21
- 2 Quantum Mechanics Approaches in Computational Toxicology** 31
Jakub Kostal
- 2.1 Translating Computational Chemistry to Predictive Toxicology 31

2.2	Levels of Theory in Quantum Mechanical Calculations	33
2.3	Representing Molecular Orbitals	38
2.4	Hybrid Quantum and Molecular Mechanical Calculations	39
2.5	Representing System Dynamics	40
2.6	Developing QM Descriptors	42
2.6.1	Global Electronic Parameters	42
2.6.1.1	Electrostatic Potential, Dipole, and Polarizability	43
2.6.1.2	Global Electronic Parameters Derived from Frontier Molecular Orbitals (FMOs)	45
2.6.2	Local (Atom-Based) Electronic Parameters	47
2.6.2.1	Parameters Derived from Frontier Molecular Orbitals (FMOs)	48
2.6.2.2	Partial Atomic Charges	51
2.6.2.3	Hydrogen-Bonding Interactions	51
2.6.2.4	Bond Enthalpies	53
2.6.3	Modeling Chemical Reactions	53
2.6.4	QM/MM Calculations of Covalent Host-Guest Interactions	56
2.6.5	Medium Effects and Hydration Models	59
2.7	Rational Design of Safer Chemicals	61
	References	64

Part II Applying Computers to Toxicology Assessment: Pharmaceutical, Industrial and Clinical 69

3	Computational Approaches for Predicting hERG Activity	71
	<i>Vinicius M. Alves, Rodolpho C. Braga, and Carolina Horta Andrade</i>	
3.1	Introduction	71
3.2	Computational Approaches	73
3.3	Ligand-Based Approaches	73
3.4	Structure-Based Approaches	77
3.5	Applications to Predict hERG Blockage	77
3.5.1	Pred-hERG Web App	79
3.6	Other Computational Approaches Related to hERG Liability	82
3.7	Final Remarks	83
	References	83
4	Computational Toxicology for Traditional Chinese Medicine	93
	<i>Ni Ai and Xiaohui Fan</i>	
4.1	Background, Current Status, and Challenges	93
4.2	Case Study: Large-Scale Prediction on Involvement of Organic Anion Transporter 1 in Traditional Chinese Medicine-Drug Interactions	99

4.2.1	Introduction to OAT1 and TCM	99
4.2.2	Construction of TCM Compound Database	101
4.2.3	OAT1 Inhibitor Pharmacophore Development	101
4.2.4	External Test Set Evaluation	102
4.2.5	Database Searching	102
4.2.6	Results: OAT1 Inhibitor Pharmacophore	103
4.2.7	Results: OAT1 Inhibitor Pharmacophore Evaluation	104
4.2.8	Results: TCM Compound Database Searching Using OAT1 Inhibitor Pharmacophore	104
4.2.9	Discussion	110
4.3	Conclusion	114
	Acknowledgment	114
	References	114
5	Pharmacophore Models for Toxicology Prediction	121
	<i>Daniela Schuster</i>	
5.1	Introduction	121
5.2	Antitarget Screening	125
5.3	Prediction of Liver Toxicity	125
5.4	Prediction of Cardiovascular Toxicity	127
5.5	Prediction of Central Nervous System (CNS) Toxicity	128
5.6	Prediction of Endocrine Disruption	130
5.7	Prediction of ADME	135
5.8	General Remarks on the Limits and Future Perspectives for Employing Pharmacophore Models in Toxicological Studies	136
	References	137
6	Transporters in Hepatotoxicity	145
	<i>Eleni Kotsampasakou, Sankalp Jain, Daniela Digles, and Gerhard F. Ecker</i>	
6.1	Introduction	145
6.2	Basolateral Transporters	146
6.3	Canalicular Transporters	148
6.4	Data Sources for Transporters in Hepatotoxicity	148
6.5	<i>In Silico</i> Transporters Models	150
6.6	Ligand-Based Approaches	150
6.7	OATP1B1 and OATP1B3	150
6.8	NTCP	154
6.9	OCT1	154
6.10	OCT2	154
6.11	MRP1, MRP3, and MRP4	155
6.12	BSEP	155
6.13	MRP2	156
6.14	MDR1/P-gp	156

- 6.15 MDR3 157
- 6.16 BCRP 157
- 6.17 MATE1 158
- 6.18 ASBT 159
- 6.19 Structure-Based Approaches 159
- 6.20 Complex Models Incorporating Transporter Information 160
- 6.21 *In Vitro* Models 160
- 6.22 Multiscale Models 161
- 6.23 Outlook 162
- Acknowledgments 164
- References 164

7 Cheminformatics in a Clinical Setting 175

Matthew D. Krasowski and Sean Ekins

- 7.1 Introduction 175
- 7.2 Similarity Analysis Applied to Drug of Abuse/Toxicology Immunoassays 177
- 7.3 Similarity Analysis Applied to Therapeutic Drug Monitoring Immunoassays 187
- 7.4 Similarity Analysis Applied to Steroid Hormone Immunoassays 191
- 7.5 Cheminformatics Applied to "Designer Drugs" 195
- 7.6 Relevance to Antibody-Ligand Interactions 202
- 7.7 Conclusions and Future Directions 202
- Acknowledgment 203
- References 204

Part III Applying Computers to Toxicology Assessment: Environmental and Regulatory Perspectives 211

8 Computational Tools for ADMET Profiling 213

*Denis Fourches, Antony J. Williams, Grace Patlewicz, Imran Shah,
Chris Grulke, John Wambaugh, Ann Richard, and Alexander Tropsha*

- 8.1 Introduction 213
- 8.2 Cheminformatics Approaches for ADMET Profiling 214
- 8.2.1 Chemical Data Curation Prior to ADMET Modeling 215
- 8.2.2 QSAR Modelability Index 217
- 8.2.3 Predictive QSAR Model Development Workflow 218
- 8.2.4 Hybrid QSAR Modeling 220
- 8.2.4.1 Simple Consensus 223
- 8.2.4.2 Mixed Chemical and Biological Features 223

8.2.4.3	Two-Step Hierarchical Workflow	224
8.2.5	Chemical Biological Read-Across	226
8.2.6	Public Chemotype Approach to Data-Mining	229
8.3	Unsolved Challenges in Structure Based Profiling	230
8.3.1	Biological Data Curation	231
8.3.2	Identification and Treatment of Activity and Toxicity Cliffs	233
8.3.3	<i>In Vitro</i> to <i>In Vivo</i> Continuum in the Context of AOP	233
8.4	Perspectives	234
8.4.1	Profilers on the Go with Mobile Devices	235
8.4.2	Structure–Exposure–Activity Relationships	236
8.5	Conclusions	237
	Acknowledgments	237
	Disclaimer	237
	References	238
9	Computational Toxicology and Reach	245
	<i>Emilio Benfenati, Anna Lombardo, and Alessandra Roncaglioni</i>	
9.1	A Theoretical and Historical Introduction to the Evolution Toward Predictive Models	245
9.2	Reach and the Other Legislations	247
9.3	Annex XI of Reach for QSAR Models	248
9.3.1	The First Condition of Annex XI and QMRF	249
9.3.2	The Second Condition and the Applicability Domain	251
9.3.3	The Third Condition of Annex XI, and the Use of the QSAR Models	252
9.3.4	Adequate and Reliable Documentation of the Applied Method	254
9.4	The ECHA Guidelines and the Use of QSAR Models within ECHA	255
9.4.1	Example of Bioconcentration Factor (BCF)	255
9.4.2	Example of Mutagenicity (Reverse-Mutation Assay) Prediction	260
9.5	Conclusions	266
	References	266
10	Computational Approaches to Predicting Dermal Absorption of Complex Topical Mixtures	269
	<i>Jim E. Riviere and Jason Chittenden</i>	
10.1	Introduction	269
10.2	Principles of Dermal Absorption	270
10.3	Dermal Mixtures	274
10.4	Model Systems	275
10.5	Local Skin Versus Systemic Endpoints	277
10.6	QSAR Approaches to Model Dermal Absorption	278

- 10.7 Pharmacokinetic Models 281
- 10.8 Conclusions 284
- References 285

Part IV New Technologies for Toxicology, Future Perspectives 291

- 11 Big Data in Computational Toxicology: Challenges and Opportunities 293**
Linlin Zhao and Hao Zhu
 - 11.1 Big Data Scenario of Computational Toxicology 293
 - 11.2 Fast-Growing Chemical Toxicity Data 295
 - 11.3 The Use of Big Data Approaches in Modern Computational Toxicology 299
 - 11.3.1 Profiling the Toxicants with Massive Biological Data 299
 - 11.3.2 Read-Across Study to Fill Data Gap 301
 - 11.3.3 Unstructured Data Curation 302
 - 11.4 Challenges of Big Data Research in Computational Toxicology and Relevant Forecasts 303
 - References 304

- 12 HLA-Mediated Adverse Drug Reactions: Challenges and Opportunities for Predictive Molecular Modeling 313**
George van Den Driessche and Denis Fourches
 - 12.1 Introduction 313
 - 12.2 Human Leukocyte Antigens 314
 - 12.2.1 HLA Proteins 314
 - 12.2.2 ADR–HLA Associations 316
 - 12.2.3 HLA-Drug–Peptide Proposed T-Cell Signaling Mechanisms 321
 - 12.3 Structure-Based Molecular Docking to Study HLA-Mediated ADRs 322
 - 12.3.1 Structure-Based Docking 324
 - 12.3.2 Case Study: Abacavir with B*57:01 326
 - 12.3.3 Limitations 332
 - 12.4 Perspectives 334
 - References 335

- 13 Open Science Data Repository for Toxicology 341**
Valery Tkachenko, Richard Zakharov, and Sean Ekins
 - 13.1 Introduction 341
 - 13.2 Open Science Data Repository 342
 - 13.3 Benefits of OSDR 344

13.3.1	Chemically and Semantically Enabled Scientific Data Repository	344
13.3.2	Chemical Validation and Standardization Platform	346
13.3.3	Format Adapters	347
13.3.4	Open Platform for Data Acquisition, Curation, and Dissemination	350
13.3.5	Dataledger	350
13.4	Technical Details	351
13.5	Future Work	353
13.5.1	Implementation of Ontology-Based Properties	356
13.5.2	Implementation of an Advanced Search System	357
13.5.3	Implementation of a Scientist Profile, Advanced Security, Data Sharing Capabilities and Notifications Framework	357
	References	358
14	Developing Next Generation Tools for Computational Toxicology	363
	<i>Alex M. Clark, Kimberley M. Zorn, Mary A. Lingerfelt, and Sean Ekins</i>	
14.1	Introduction	363
14.2	Developing Apps for Chemistry	364
14.3	Green Chemistry	364
14.3.1	Green Solvents and Lab Solvents	367
14.3.2	Green Lab Notebook	370
14.4	Polypharma and Assay Central	374
14.4.1	Future Efforts with Assay Central	380
14.5	Conclusion	382
	Acknowledgments	383
	References	383
	Index	389

List of Contributors

Ni Ai

Pharmaceutical Informatics Institute
College of Pharmaceutical Sciences
Zhejiang University
Hangzhou
Zhejiang, PR
China

Vinicius M. Alves

LabMol – Laboratory for Molecular
Modeling and Design, Faculty of
Pharmacy
Federal University of Goias
Goiania, GO
Brazil

Carolina Horta Andrade

LabMol – Laboratory for Molecular
Modeling and Design, Faculty of
Pharmacy
Federal University of Goias
Goiania, GO
Brazil

Rodolpho C. Braga

LabMol – Laboratory for Molecular
Modeling and Design, Faculty of
Pharmacy
Federal University of Goias
Goiania, GO
Brazil

Jason Chittenden

Center for Chemical Toxicology
Research and Pharmacokinetics
Biomathematics Program
North Carolina State University
Raleigh, NC
USA

Alex M. Clark

Molecular Materials Informatics, Inc.
Montreal, Quebec
Canada

Daniela Digles

Department of Pharmaceutical
Chemistry
University of Vienna
Wien
Austria

George van Den Driessche

Department of Chemistry
Bioinformatics Research Center
North Carolina State University
Raleigh, NC
USA

Gerhard F. Ecker

Department of Pharmaceutical
Chemistry
University of Vienna
Wien
Austria

Sean Ekins

Collaborations Pharmaceuticals, Inc.
Raleigh, NC
USA

Emilio Benfenati

IRCCS – Istituto di Ricerche
Farmacologiche “Mario Negri”
Laboratory of Environmental
Chemistry and Toxicology
Milan
Italy

Xiaohui Fan

Pharmaceutical Informatics Institute
College of Pharmaceutical Sciences
Zhejiang University
Hangzhou
Zhejiang, PR
China

Denis Fourches

Department of Chemistry
Bioinformatics Research Center
North Carolina State University
Raleigh, NC
USA

Joel S. Freundlich

Department of Pharmacology &
Physiology
New Jersey Medical School
Rutgers University
Newark, NJ
USA

and

Division of Infectious Disease
Department of Medicine and the Ruy
V. Lourenço Center for the Study of
Emerging and Re-emerging
Pathogens
New Jersey Medical School, Rutgers
University
Newark, NJ
USA

Chris Grulke

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Sankalp Jain

Department of Pharmaceutical
Chemistry
University of Vienna
Wien
Austria

Alexandru Korotcov

Gaithersburg, MD
USA

Jakub Kostal

Chemistry Department
The George Washington University
Washington DC
USA

Eleni Kotsampasakou

Department of Pharmaceutical
Chemistry
University of Vienna
Wien
Austria

Matthew D. Krasowski

Department of Pathology
University of Iowa Hospitals and
Clinics
Iowa City, IA
USA

Mary A. Lingerfelt

Collaborations Pharmaceuticals, Inc.
Raleigh, NC
USA

Anna Lombardo

IRCCS – Istituto di Ricerche
Farmacologiche “Mario Negri”
Laboratory of Environmental
Chemistry and Toxicology
Milan
Italy

Grace Patlewicz

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Alexander L. Perryman

Department of Pharmacology &
Physiology
New Jersey Medical School
Rutgers University
Newark, NJ
USA

Ann Richard

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Jim E. Riviere

Center for Chemical Toxicology
Research and Pharmacokinetics
Biomathematics Program
North Carolina State University
Raleigh, NC
USA

Alessandra Roncaglioni

IRCCS – Istituto di Ricerche
Farmacologiche “Mario Negri”
Laboratory of Environmental
Chemistry and Toxicology
Milan
Italy

Daniela Schuster

Institute of
Pharmacy/Pharmaceutical
Chemistry
University of Innsbruck
Innsbruck
Austria

Imran Shah

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Valery Tkachenko

Rockville, MD
USA

Alexander Tropsha

UNC Eshelman School of Pharmacy
University of North Carolina at
Chapel Hill
Chapel Hill, NC
USA

John Wambaugh

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Antony J. Williams

National Center for Computational
Toxicology, Office of Research and
Development
U.S. Environmental Protection
Agency
Research Triangle Park
Durham, NC
USA

Richard Zakharov

Rockville, MD
USA

Linlin Zhao

Center for Computational and
Integrative Biology
Rutgers University
Camden, NJ
USA

Hao Zhu

Center for Computational and
Integrative Biology
Rutgers University
Camden, NJ
USA

and

Department of Chemistry
Rutgers University
Camden, NJ
USA

Kimberley M. Zorn

Collaborations Pharmaceuticals, Inc.
Raleigh, NC
USA

Preface

Since the publication of *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals* in 2007 a lot has happened both in the career of the editor and in science in general. For one, my focus has expanded towards many computational applications to drug discovery rather than solely focused on ADME/Tox. I have also garnered new collaborators some of whom have very graciously agreed to contribute to this volume. Science is changing. Publishing may be adjusting slowly too. This book will likely be read as much on mobile devices or computers as in physical hard copies. Computational toxicology has also evolved in the past decade with the dramatic increase in public data availability. There have also been a number of more collaborative projects in Europe around toxicology (e.g. e-Tox and OpenTox), in addition we have seen a growth in open computational tools and model sharing (QSAR toolbox, Chembench, CDD, Bioclipse etc.). Groups like the EPA have developed and expanded ToxCast which represents a valuable resource for toxicology modeling. We are now therefore in the age of truly Big Data compared with a decade ago and there have been several efforts to combine different types of data for toxicology. To round this off, the growth in nanotechnology has seen the emergence of computational nanotoxicology which would not have been predicted my earlier book.

This book is therefore aimed at this next generation of computational toxicology scientist, comprehensively discussing the state-of-the-art of currently available molecular-modelling tools and the role of these in testing strategies for different types of toxicity. The overall role of these computational approaches in addressing environmental and occupational toxicity is also covered. These chapters before you aim to describe topics in an accessible manner especially for those who are not experts in the field. My goal with this book was to not cover too much of the same ground as the earlier book because much of what we published then is still generally valid, but to make the book focused on newer topics. I hope this book also serves to introduce some of the younger scientists from around the world who will likely drive this next generation of computational toxicology for many years to come. Finally, I hope this book inspires

scientists to pursue computational toxicology so that it continues to expand across different industries from pharmaceutical to consumer products and its importance increases, as it has over the past decade.

November 12, 2017

Sean Ekins
Fuquay Varina, NC, USA

Acknowledgments

I am extremely grateful to Jonathan Rose and colleagues at Wiley for their assistance and considerable patience. My proposal reviewers are gratefully acknowledged for their many suggestions which helped shape this.

I would like to acknowledge my many collaborators over the years whose work in some cases has been mentioned here. In particular, Dr Joel S. Freundlich, Dr Antony J. Williams, Dr Alex M. Clark, Dr Matthew D. Krasowski, Dr Carolina H. Andrade, and many others. I am also grateful for the support of SC Johnson who have kept me challenged and engaged with new applications for computational toxicology over the years. I would also like to acknowledge Dr Daniela Schuster for the kind use of her graphic for the book cover.

This book would not have been possible without the support of Dr Maggie A.Z. Hupcey and my family who have tolerated late nights, and frequent disappearances to the library to write over the holidays.

Part I

Computational Methods

1

Accessible Machine Learning Approaches for Toxicology

Sean Ekins¹, Alex M. Clark², Alexander L. Perryman³, Joel S. Freundlich^{3,4}, Alexandru Korotcov⁵, and Valery Tkachenko⁶

¹ Collaborations Pharmaceuticals, Inc., Raleigh, NC, USA

² Molecular Materials Informatics, Inc., Montreal, Quebec, Canada

³ Department of Pharmacology & Physiology, New Jersey Medical School, Rutgers University, Newark, NJ, USA

⁴ Division of Infectious Disease, Department of Medicine and the Ruy V. Lourenço Center for the Study of Emerging and Re-emerging Pathogens, New Jersey Medical School, Rutgers University, Newark, NJ, USA

⁵ Gaithersburg, MD, USA

⁶ Rockville, MD, USA

CHAPTER MENU

Introduction, 3

Bayesian Models, 5

Deep Learning Models, 13

Comparison of Different Machine Learning Methods, 16

Future Work, 21

1.1 Introduction

Computational approaches have in recent years played an increasingly important role in the drug discovery process within large pharmaceutical firms. Virtual screening of compounds using ligand-based and structure-based methods to predict potency enables more efficient utilization of high throughput screening (HTS) resources, by enriching the set of compounds physically screened with those more likely to yield hits [1–4]. Computation of absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) properties exploiting statistical techniques greatly reduces the number of expensive assays that must be performed, now making it practical to consider these factors very early in the discovery process to minimize late-stage failures of potent lead compounds that are not drug-like [5–11]. Large pharma have successfully

integrated these *in silico* methods into operational practice, validated them, and then realized their benefits, because these firms have (i) expensive commercial software to build models, (ii) large, diverse proprietary datasets based on consistent experimental protocols to train and test the models, and (iii) staff with extensive computational and medicinal chemistry expertise to run the models and interpret the results. Drug discovery efforts centered in universities, foundations, government laboratories, and small biotechnology companies, however, generally lack these three critical resources and, as a result, have yet to exploit the full benefits of *in silico* methods. For close to a decade, we have aimed to use machine learning approaches and have evaluated how we could circumvent these limitations so that others can benefit from current and emerging best industry practices.

The current practice in pharma is to integrate *in silico* predictions into a combined workflow together with *in vitro* assays to find “hits” that can then be reconfirmed and optimized [12]. The incremental cost of a virtual screen is minimal, and the savings compared with a physical screen are magnified if the compound would also need to be synthesized rather than purchased from a vendor. Imagine if the blind hit rate against some library is 1%, and the *in silico* model can pre-filter the library to give an experimental hit rate of 2%, then significant resources are freed up to focus on other promising regions of chemical property space [13]. Our past pharmaceutical collaborations [14, 15] have suggested that computational approaches are critical to making drug discovery more efficient.

The relatively high cost of *in vivo* and *in vitro* screening of ADME and toxicity properties of molecules has motivated our efforts to develop *in silico* methods to filter and select a subset of compounds for testing. By relying on very large, internally consistent datasets, large pharma has succeeded in developing highly predictive proprietary models [5–8]. At Pfizer (and probably other companies), for example, many of these models (e.g., those that predict the volume of distribution, aqueous kinetic solubility, acid dissociation constant, and distribution coefficient) [5–8, 16] are believed (according to discussions with scientists) to be so accurate that they have essentially put experimental assays out of business. In most other cases, large pharma perform experimental assays for a small fraction of compounds of interest to augment or validate their computational models. Efforts by smaller pharma and academia have not been as successful, largely because they have, by necessity, drawn upon much smaller datasets and, in a few cases, tried to combine them [11, 17–22]. However, this is changing rapidly, and public datasets in PubChem, ChEMBL, Collaborative Drug Discovery (CDD) and elsewhere are becoming available for ADME/Tox properties. For example, the CDD public database has >100 public datasets that can be used to generate community-based models, including extensive neglected infectious disease structure–activity relationship (SAR) datasets (malaria, tuberculosis, Chagas disease, etc.), and ADMEdata.com

datasets that are broadly applicable to many projects. Recent efforts with them have led to a platform that enables drug discovery projects to benefit from open source machine learning algorithms and descriptors in a secure environment, which allows models to be shared with collaborators or made accessible to the community.

In the area of pharmaceutical research and development and specifically that of cheminformatics, there are many machine learning methods, such as support vector machines (SVM), k -nearest neighbors, naïve Bayesian, and decision trees, [23] which have seen increasing use as our datasets, have grown to become “big data” [24–27]. These methods [23] can be used for binary classification, multiple classes, or continuous data. In more recent years, the biological data amassed from HTS and high content screens has called for different tools to be used that can account for some of the issues with this bigger data [26]. Many of these resulting machine learning models can also be implemented on a mobile phone [28, 29].

1.2 Bayesian Models

Our machine learning experience over a decade [14, 30–46] has focused on Bayesian approaches (Figure 1.1). Bayesian models classify data as active or inactive on the basis of user-defined thresholds using a simple probabilistic classification model based on Bayes’ theorem. We initially used the Bayesian modeling software within the Pipeline Pilot and Discovery Studio (BIOVIA) with many ADME/Tox and drug discovery datasets. Most of these models have used molecular function class fingerprints of maximum diameter 6 and several other simple descriptors [47, 48]. The models were internally validated through the generation of receiver operator characteristic (ROC) plots. We have also compared single- and dual-event Bayesian models utilizing published screening data [49, 50]. As an example, the single-event models use only whole-cell antitubercular activity, either at a single compound concentration or as a dose–response IC_{50} or IC_{90} (amount of compound inhibiting 50% or 90% of growth, respectively), while the dual-event models also use a selectivity index ($SI = CC_{50}/IC_{90}$, where CC_{50} is the compound concentration that is cytotoxic and inhibits 50% of the growth of Vero cells). While single-event models [13, 51, 52] are widely published, dual-event models [53] attempt to predict active compounds with acceptable relative activity against the pathogen (in this case, *Mtb*), versus the model mammalian cell line (e.g., Vero cells). Our models identified 4–10 times more active compounds than random screening did and the models also had relatively high hit rates, for example, 14% [54], 71% (Figure 1.1) [53], or intermediate [55] for *Mtb*. Recent machine learning work on Chagas disease has identified *in vivo* active compounds [56], one of which is an approved antimalarial in Europe. Most recently, we