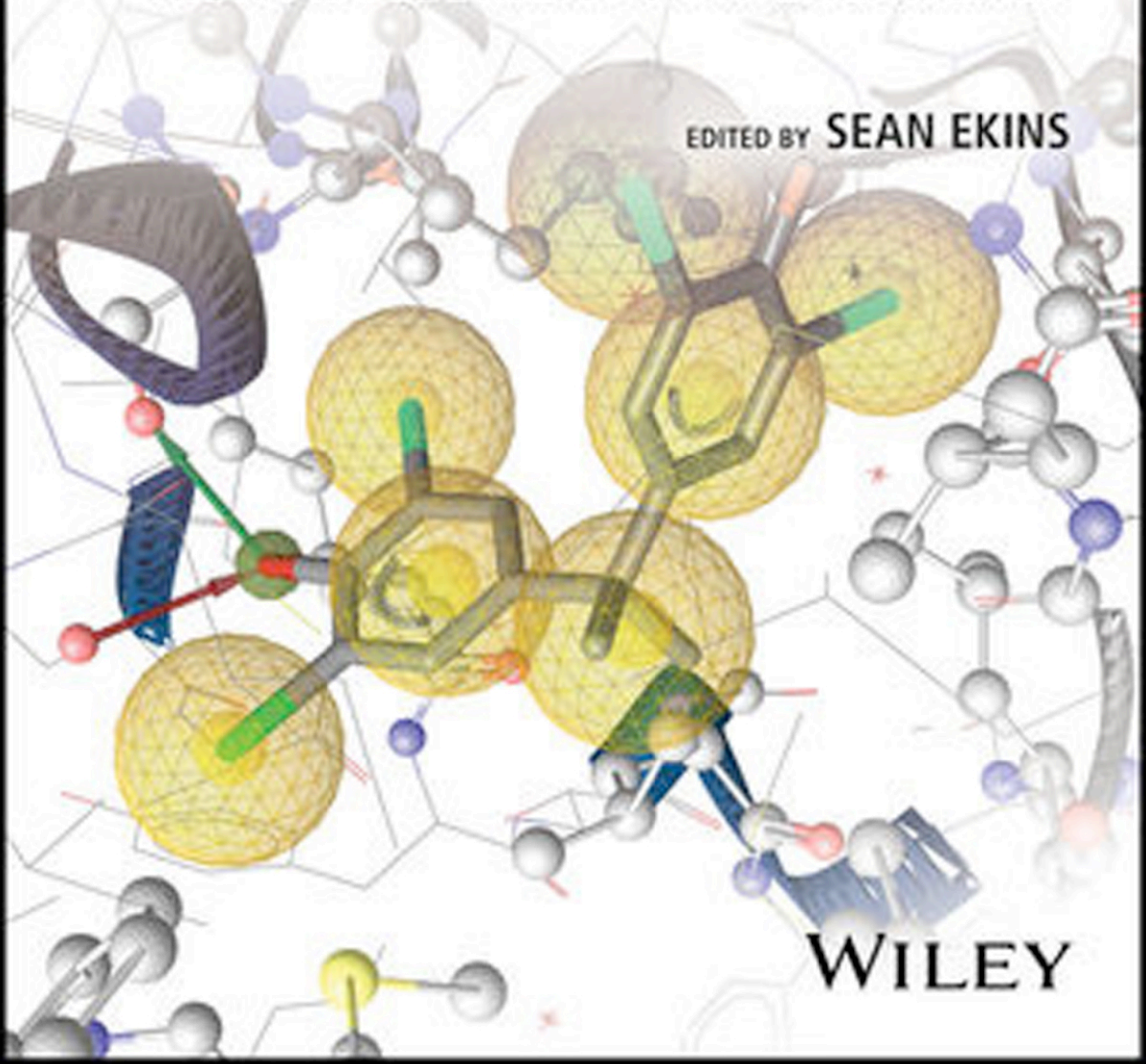
Wiley Series on Technologies for the Pharmaceutical Industry

Sean Ekins, Series Editor

COMPUTATIONAL TOXICOLOGY

RISK ASSESSMENT FOR CHEMICALS

EDITED BY SEAN EKINS



WILEY

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 Šašić

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

Drug Efficacy, Safety, and Biologics Discovery: Enmerging 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

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

10 9 8 7 6 5 4 3 2 1

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

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.

Sean Ekins
November 12, 2017 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 Ekins[1], Alex M. Clark[2], Alexander L. Perryman[3], Joel S. Freundlich[3,4], Alexandru Korotcov[5], and Valery Tkachenko[6]

[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

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

Computational Toxicology: Risk Assessment for Chemicals, First Edition. Edited by Sean Ekins.

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 used 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 pharmaceuticals 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