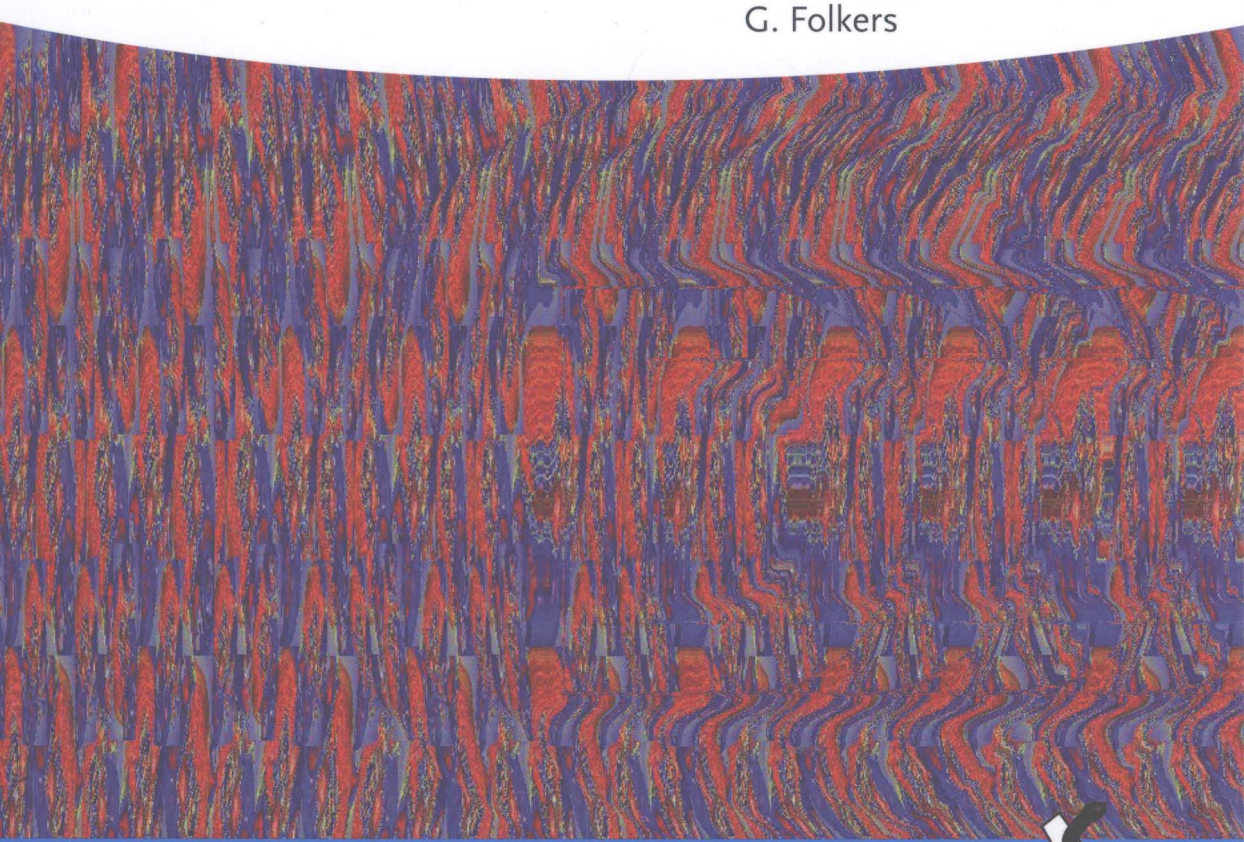
WILEY-VCH

Ed[redacted]
Rémy D. Hoffmann, Arnaud Gohier, Pavel Pospisil

Data Mining in Drug Discovery

Volume 57
Series Editors:
R. Mannhold, H. Kubinyi,
G. Folkers



Methods and Principles in Medicinal Chemistry

Edited by Rémy D. Hoffmann, Arnaud Gohier, and Pavel Pospisil

Data Mining in Drug Discovery

WILEY-VCH
Verlag GmbH & Co. KGaA

Series Editors

Prof. Dr. Raimund Mannhold
Rosenweg
740489 Düsseldorf
Germany
mannhold@uni-duesseldorf.de

Prof. Dr. Hugo Kubinyi
Donnersbergstrasse 9
67256 Weisenheim am Sand
Germany
kubinyi@t-online.de

Prof. Dr. Gerd Folkers
Collegium Helveticum
STW/ETH Zurich
8092 Zurich
Switzerland
folkers@collegium.ethz.ch

Volume Editors

Dr. Rémy D. Hoffmann
Prestwick Chemical
Bld. Gonthier d'Andernach
67400 Strasbourg-Illkirch
France

Dr. Arnaud Gohier
Institut de Recherches
Servier
125 Chemin de Ronde
78290 Croissy-sur-Seine
France

Dr. Pavel Pospisil
Philip Morris Int. R&D
Quai Jeanrenaud 5
Biological Systems Res.
2000 NEUCHÂTEL
Switzerland

Cover Description

The cover picture is a 3D stereogram. The pattern is built from a mix of pictures showing complex molecular networks and structures.
The aim of this stereogram is to symbolize the complexity of data to data mine: when looking at them "differently," a shape of a drug pill with a letter D appears!
In order to see it, try parallel or cross-eyed viewing (either you focus your eyes somewhere behind the image or you cross your eyes).

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at ⟨http://dnb.d-nb.de⟩.

Typesetting Thomson Digital, Noida, India
Printing and Binding Markono Print Media Pte Ltd, Singapore
Cover Design Grafik-Design Schulz, Fußgönheim

Print ISBN: 978-3-527-32984-7
ePDF ISBN: 978-3-527-65601-1
ePub ISBN: 978-3-527-65600-4
mobi ISBN: 978-3-527-65599-1
oBook ISBN: 978-3-527-65598-4

Printed on acid-free paper
Printed in Singapore

Edited by
Rémy D. Hoffmann
Arnaud Gohier
Pavel Pospisil

Data Mining in Drug Discovery

Methods and Principles in Medicinal Chemistry

Previous Volumes of this Series:

List of Contributors

Mohammad Afshar
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Kamal Azzaoui
Novartis Institutes for Biomedical Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Igor I. Baskin
Strasbourg University
Faculty of Chemistry
UMR 7177 CNRS
1 rue Blaise Pascal
67000 Strasbourg
France

and

MV Lomonosov Moscow State University
Leninsky Gory
119992 Moscow
Russia

James N.D. Battey
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Sven Bergmann
Université de Lausanne
Department of Medical Genetics
Rue du Bugnon 27
1005 Lausanne
Switzerland

Sharon D. Bryant
Inte:Ligand GmbH
Clemens Maria Hofbauer-Gasse 6
2344 Maria Enzersdorf
Austria

Allen Cornett
Novartis Institutes for Biomedical Research (NIBR/DMP)
220 Massachusetts Avenue
Cambridge, MA 02139
USA

Renée Deehan
Selventa
One Alewife Center
Cambridge, MA 02140
USA

David A. Drubin
Selventa
One Alewife Center
Cambridge, MA 02140
USA

Christof Gaenzler
TIBCO Software Inc.
1235 Westlake Drive, Suite 210
Berwyn, PA 19132
USA

Michael Gilson
University of California San Diego
Skaggs School of Pharmacy and
Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Janna Hastings
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD
UK

Julia Hoeng
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Nikolai V. Ivanov
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Edgar Jacoby
Janssen Research & Development
Turnhoutseweg 30
2340 Beerse
Belgium

Jeremy L. Jenkins
Novartis Institutes for Biomedical
Research (NIBR/DMP)
220 Massachusetts Avenue
Cambridge, MA 02139
USA

Nathalie Jullian
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Esther Kellenberger
UMR 7200 CNRS-UdS
Structural Chemogenomics
74 route du Rhin
67400 Illkirch
France

Thierry Langer
Prestwick Chemical SAS
220, Blvd. Gonthier d'Andernach
67400 Illkirch-Strasbourg
France

Tiging Liu
University of California
San Diego
Skaggs School of Pharmacy and
Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Gilles Marcou
Strasbourg University
Faculty of Chemistry
UMR 7177 CNRS
1 rue Blaise Pascal
67000 Strasbourg
France

and

MV Lomonosov Moscow State University
Leninsky Gory
119992 Moscow
Russia

Elyette Martin
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Florian Martin
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Aurélien Monge
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

David Mosenkis
TIBCO Software Inc.
1235 Westlake Drive, Suite 210
Berwyn, PA 19312
USA

George Nicola
University of California San Diego
Skaggs School of Pharmacy and Pharmaceutical Sciences
9500 Gilman Drive
La Jolla, CA 92093
USA

Florian Nigsch
Novartis Institutes for Biomedical Research (NIBR)
CPC/LFP/MLI
4002 Basel
Switzerland

Manuel C. Peitsch
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Maxim Popov
Novartis Institutes for Biomedical Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Pavel Pospisil
Philip Morris International R&D
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

John P. Priestle
Novartis Institutes for Biomedical Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Josep Prous Jr.
Prous Institute for Biomedical Research
Research and Development
Rambla Catalunya 135
08008 Barcelona
Spain

Jordi Quintana
Parc Científic Barcelona (PCB)
Drug Discovery Platform
Baldiri Reixac 4
08028 Barcelona
Spain

Didier Rognan
UMR 7200 CNRS-UdS
Structural Chemogenomics
74 route du Rhin
67400 Illkirch
France

Ansgar Schuffenhauer
Novartis Institutes for Biomedical Research (NIBR/CPC/iSLD)
Forum 1 Novartis Campus
4056 Basel
Switzerland

Alain Sewer
Philip Morris International R&D
Biological Systems Research
Quai Jeanrenaud 5
2000 Neuchâtel
Switzerland

Christoph Steinbeck
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
UK

Ty M. Thomson
Selventa
Cambridge, MA 02140
USA

Yannic Tognetti
Ariana Pharma
28 rue Docteur Finlay
75015 Paris
France

Antoni Valencia
Prous Institute for Biomedical Research, SA
Computational Modeling
Rambla Catalunya 135
08008 Barcelona
Spain

Thibault Varin
Eli Lilly and Company
Lilly Research Laboratories
Lilly Corporate Center
Indianapolis, IN 46285
USA

Jurjen W. Westra
Selventa
Cambridge, MA 02140
USA

Preface

In general, the extraction of information from databases is called data mining. A database is a data collection that is organized in a way that allows easy accessing, managing, and updating its contents. Data mining comprises numerical and statistical techniques that can be applied to data in many fields, including drug discovery. A functional definition of data mining is the use of numerical analysis, visualization, or statistical techniques to identify nontrivial numerical relationships within a data set to derive a better understanding of the data and to predict future results. Through data mining, one derives a model that relates a set of molecular descriptors to biological key attributes such as efficacy or ADMET properties. The resulting model can be used to predict key property values of new compounds, to prioritize them for follow-up screening, and to gain insight into the compounds' structure–activity relationship. Data mining models range from simple, parametric equations derived from linear techniques to complex, nonlinear models derived from nonlinear techniques. More detailed information is available in literature [1–7].

This book is organized into four parts. Part One deals with different sources of data used in drug discovery, for example, protein structural databases and the main small-molecule bioactivity databases.

Part Two focuses on different ways for data analysis and data enrichment. Here, an industrial insight into mining HTS data and identifying hits for different targets is presented. Another chapter demonstrates the strength of powerful data visualization tools for simplification of these data, which in turn facilitates their interpretation.

Part Three comprises some applications to polypharmacology. For instance, the positive outcomes are described that data mining can produce for ligand profiling and target fishing in the chemogenomics era.

Finally, in Part Four, systems biology approaches are considered. For example, the reader is introduced to integrative and modular analysis approaches to mine large molecular and phenotypical data. It is shown how the presented approaches can reduce the complexity of the rising amount of high-dimensional data and provide a means for integrating different types of omics data. In another chapter, a set of novel methods are established that quantitatively measure the biological impact of chemicals on biological systems.

The series editors are grateful to Remy Hoffmann, Arnaud Gohier, and Pavel Pospisil for organizing this book and to work with such excellent authors. Last but not least, we thank Frank Weinreich and Heike Nöthe from Wiley-VCH for their valuable contributions to this project and to the entire book series.

Düsseldorf
Weisenheim am Sand
Zürich
May 2013

Raimund Mannhold
Hugo Kubinyi
Gerd Folkers

References

1 Cruciani, G., Pastor, M., and Mannhold, R. (2002) Suitability of molecular descriptors for database mining: a comparative analysis. *Journal of Medicinal Chemistry*, **45**, 2685–2694.
2 Obenshain, M.K. (2004) Application of data mining techniques to healthcare data. *Infection Control and Hospital Epidemiology*, **25**, 690–695.
3 Weaver, D.C. (2004) Applying data mining techniques to library design, lead generation and lead optimization. *Current Opinion in Chemical Biology*, **8**, 264–270.
4 Yang, Y., Adelstein, S.J., and Kassis, A.I. (2009) Target discovery from data mining approaches. *Drug Discovery Today*, **14**, 147–154.
5 Campbell, S.J., Gaulton, A., Marshall, J., Bichko, D., Martin, S., Brouwer, C., and Harland, L. (2010) Visualizing the drug target landscape. *Drug Discovery Today*, **15**, 3–15.
6 Geppert, H., Vogt, M., and Bajorath, J. (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, **50**, 205–216.
7 Hasan, S., Bonde, B.K., Buchan, N.S., and Hall, M.D. (2012) Network analysis has diverse roles in drug discovery. *Drug Discovery Today*, **17**, 869–874.

A Personal Foreword

The term data mining is well recognized by many scientists and is often used when referring to techniques for advanced data retrieval and analysis. However, since there have been recent advances in techniques for data mining applied to the discovery of drugs and bioactive molecules, assembling these chapters from experts in the field has led to a realization that depending upon the field of interest (biochemistry, computational chemistry, and biology), data mining has a variety of aspects and objectives.

Coming from the ligand molecule world, one can state that the understanding of chemical data is more complete because, in principle, chemistry is governed by physicochemical properties of small molecules and our "microscopic" knowledge in this domain has advanced considerably over the past decades. Moreover, chemical data management has become relatively well established and is now widely used. In this respect, data mining consists in a thorough retrieval and analysis of data coming from different sources (but mainly from literature), followed by a thorough cleaning of data and its organization into compound databases. These methods have helped the scientific community for several decades to address pathological effects related to simple (single target) biological problems. Today, however, it is widely accepted that many diseases can only be tackled by modulating the ligand biological/pharmacological profile, that is, its "molecular phenotype." These approaches require novel methodologies and, due to increased accessibility to high computational power, data mining is definitely one of them.

Coming from the biology world, the perception of data mining differs slightly. It is not just a matter of literature text mining anymore, since the disease itself, as well as the clinical or phenotypical observations, may be used as a starting point. Due to the complexity of human biology, biologists start with hypotheses based upon empirical observations, create plausible disease models, and search for possible biological targets. For successful drug discovery, these targets need to be druggable. Moreover, modern systems biology approaches take into account the full set of genes and proteins expressed in the drug environment (omics), which can be used to generate biological network information. Data mining these data, when structured into such networks, will provide interpretable information that

leads to an increased knowledge of the biological phenomenon. Logically, such novel data mining methods require new and more sophisticated algorithms.

This book aims to cover (in a nonexhaustive manner) the data mining aspects for these two parallel but meant-to-be-convergent fields, which should not only give the reader an idea of the existence of different data mining approaches, algorithms, and methods used but also highlight some elements to assess the importance of linking ligand molecules to diseases. However, there is awareness that there is still a long way to go in terms of gathering, normalizing, and integrating relevant biological and pharmacological data, which is an essential prerequisite for making more accurate simulations of compound therapeutic effects.

This book is structured into four parts: Part One, Data Sources, introduces the reader to the different sources of data used in drug discovery. In Chapter 1, Kellenberger *et al.* present the Protein Data Bank and related databases for exploring ligand–protein recognition and its application in drug design. Chapter 2 by Nicola *et al.* is a reprint of a recently published article in *Journal of Medicinal Chemistry* (2012, 55 (16): 6987–7002) that nicely presents the main small-molecule bioactivity databases currently used in medicinal chemistry and the modern trends for their exploitation. In Chapter 3, Hastings *et al.* point out the importance of chemical ontologies for the standardization of chemical libraries in order to extract and organize chemical knowledge in a way similar to biological ontologies. Chapter 4 by Martin *et al.* presents the importance of a corporate chemical registry system as a central repository for uniform chemical entities (including their spectrometric data) and as an important point of entry for exploring public compound activity databases for systems biology data.

Part Two, Analysis and Enrichment, describes different ways for data analysis and data enrichment. In Chapter 5, Battey *et al.* didactically present the basics of plant pathway construction, the potential for their use in data mining, and the prediction of pathways using information from an enzymatic structure. Even though this chapter deals with plant pathways, the information can be readily interpreted and applied directly to metabolic pathways in humans. In Chapter 6, Azzaoui *et al.* present an industrial insight into mining HTS data and identifying hits for different targets and the associated challenges and pitfalls. In Chapter 7, Mosenkis *et al.* clearly demonstrate, using different examples, how powerful data visualization tools are key to the simplification of complex results, making them readily intelligible to the human brain and eye. We also welcome Chapter 8 by Marcou *et al.* that provides a concrete example of the increasingly frequent need for powerful statistical processing tools. This is exemplified by the use of R in the chemoinformatics process. Readers will note that this chapter is built like a tutorial for the R language in order to process, cluster, and visualize molecules, which is demonstrated by its application to a concrete example. For programmers, this may serve as an initiation to the use of this well-known bioinformatics tool for processing chemical information.

Part Three, Applications to Polypharmacology, contains chapters detailing tools and methods to mine data with the aim to elucidate preclinical profiles of small

molecules and select potential new drug targets. In Chapter 9, Prous *et al.* nicely present three examples of knowledge bases that attempt to relate, in a comprehensive manner, the interactions between chemical compounds, biological entities (molecules and pathways), and their assays. The second part of this chapter presents the challenges that these knowledge-based data mining methodologies face when searching for potential mechanisms of action of compounds. In Chapter 10, Jullian *et al.* introduce the reader to the advantages of using rule-based methods when exploring polypharmacological data sets, compared to standard numerical approaches, and their application in the development of novel ligands. Finally, in Chapter 11, Bryant *et al.* familiarize us with the positive outcomes that data mining can produce for ligand profiling and target fishing in the chemogenomics era. The authors expose how searching through ligand and target pharmacophoric structural and descriptor spaces can help to design or extend libraries of ligands with desired pharmacological, yet lowered toxicological, properties.

In Part Four, Systems Biology Approaches, we are pleased to include two exciting chapters coming from the biological world. In Chapter 12, Bergmann introduces us to integrative and modular analysis approaches to mine large molecular and phenotypical data. The author argues how the presented approaches can reduce the complexity of the rising amount of high-dimensional data and provide a means to integrating different types of omics data. Moreover, astute integration is required for the understanding of causative links and the generation of more predictive models. Finally, in the very robust Chapter 13, Sewer *et al.* present systems biology-based approaches and establish a set of novel methods that quantitatively measure the biological impact of the chemicals on biological systems. These approaches incorporate methods that use mechanistic causal biological network models, built on systems-wide omics data, to identify any compound's mechanism of action and assess its biological impact at the pharmacological and toxicological level. Using a five-step strategy, the authors clearly provide a framework for the identification of biological networks that are perturbed by short-term exposure to chemicals. The quantification of such perturbation using their newly introduced impact factor "BIF" then provides an immediately interpretable assessment of such impact and enables observations of early effects to be linked with long-term health impacts.

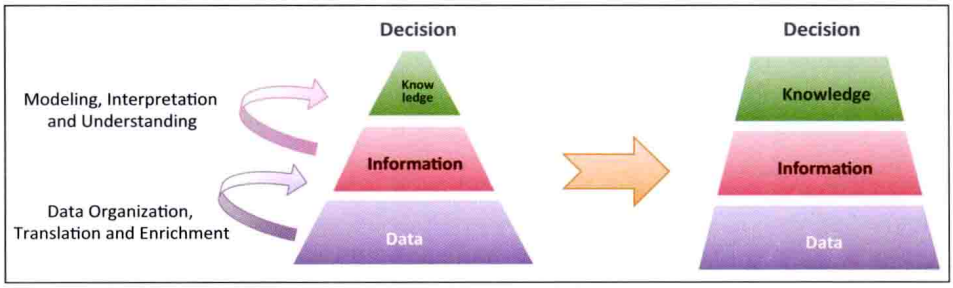
We are pleased that you have selected this book and hope that you find the content both enjoyable and educational. As many authors have accompanied their chapters with clear concise pictures, and as someone once said "one figure can bear thousand words," this Personal Foreword also contains a figure (see below). We believe that the novel applications of data mining presented in these pages by authors coming from both chemical and biological communities will provide the reader with more insight into how to reshape this pyramid into a trapezoidal form, with the enlarged knowledge area. Thus, improved data processing techniques leading to the generation of readily interpretable information, together with an increased understanding of the therapeutical processes, will enable scientists

to take wiser decisions regarding what to do next in their efforts to develop new drugs.

We wish you a happy and inspiring reading.

Strasbourg, March 14, 2013

Remy Hoffmann, Arnaud Gohier, and Pavel Pospisil



Contents