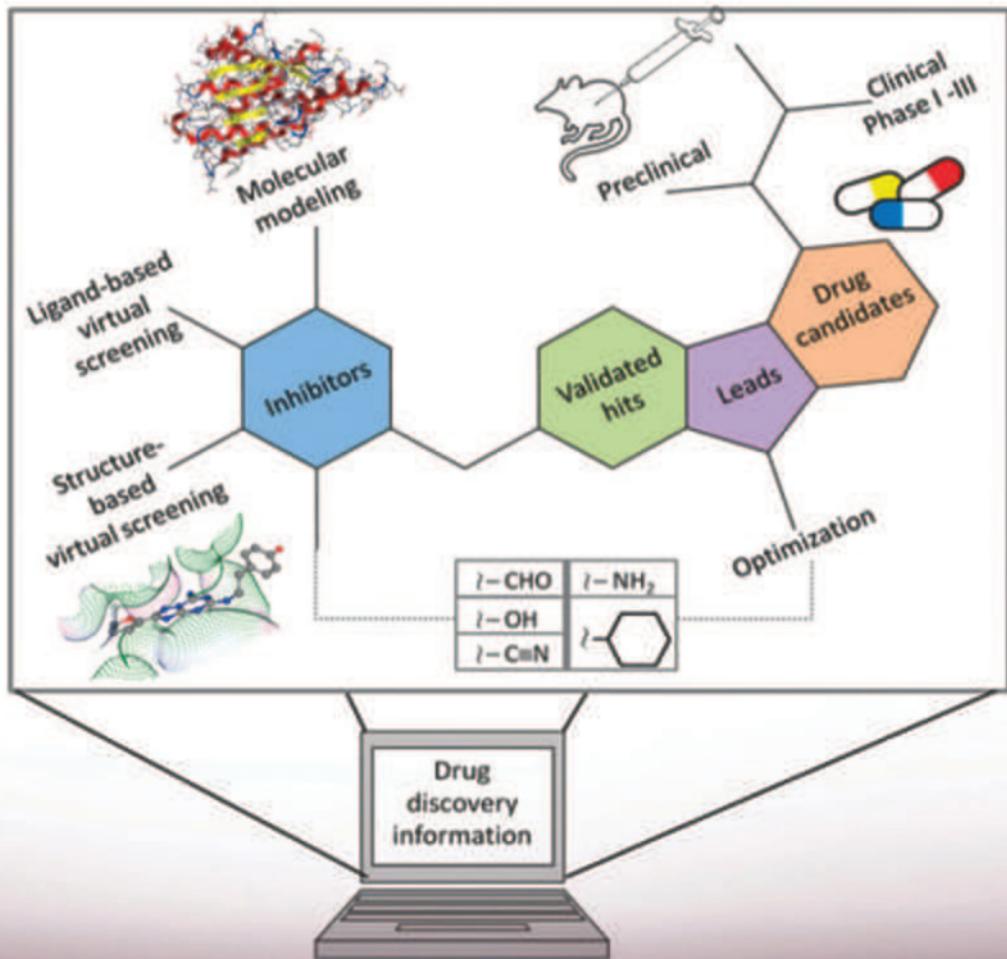


Chemoinformatics for Drug Discovery

Edited by Jürgen Bajorath



WILEY

CHEMOINFORMATICS FOR DRUG DISCOVERY

CHEMOINFORMATICS FOR DRUG DISCOVERY

Edited by

JÜRGEN BAJORATH

WILEY

Copyright © 2014 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Chemoinformatics for drug discovery / edited by Jürgen Bajorath.

pages cm

Includes index.

ISBN 978-1-118-13910-3 (cloth)

1. Chemoinformatics. 2. Drug development—Data processing. 3. Pharmacy informatics.

I. Bajorath, Jürgen, editor of compilation.

RS418.C482 2014

615.1'9—dc23

2013018927

Printed in the United States of America

CONTENTS

PREFACE	vii
CONTRIBUTORS	xiii
1 WHAT ARE OUR MODELS REALLY TELLING US? A PRACTICAL TUTORIAL ON AVOIDING COMMON MISTAKES WHEN BUILDING PREDICTIVE MODELS	1
<i>W. Patrick Walters</i>	
2 THE CHALLENGE OF CREATIVITY IN DRUG DESIGN	33
<i>Ajay N. Jain</i>	
3 A ROUGH SET THEORY APPROACH TO THE ANALYSIS OF GENE EXPRESSION PROFILES	51
<i>Joachim Petit, Nathalie Meurice, José Luis Medina-Franco, and Gerald M. Maggiora</i>	
4 BIMODAL PARTIAL LEAST-SQUARES APPROACH AND ITS APPLICATION TO CHEMOGENOMICS STUDIES FOR MOLECULAR DESIGN	85
<i>Kiyoshi Hasegawa and Kimito Funatsu</i>	
5 STABILITY IN MOLECULAR FINGERPRINT COMPARISON	97
<i>Anthony Nicholls and Brian Kelley</i>	
6 CRITICAL ASSESSMENT OF VIRTUAL SCREENING FOR HIT IDENTIFICATION	113
<i>Dagmar Stumpfe and Jürgen Bajorath</i>	
7 CHEMOMETRIC APPLICATIONS OF NAÏVE BAYESIAN MODELS IN DRUG DISCOVERY: BEYOND COMPOUND RANKING	131
<i>Eugen Lounkine, Peter S. Kutchukian, and Meir Glick</i>	

8 CHEMOINFORMATICS IN LEAD OPTIMIZATION	149
<i>Darren V. S. Green and Matthew Segall</i>	
9 USING CHEMOINFORMATICS TOOLS TO ANALYZE CHEMICAL ARRAYS IN LEAD OPTIMIZATION	179
<i>George Papadatos, Valerie J. Gillet, Christopher N. Luscombe, Iain M. McLay, Stephen D. Pickett, and Peter Willett</i>	
10 EXPLORATION OF STRUCTURE–ACTIVITY RELATIONSHIPS (SARs) AND TRANSFER OF KEY ELEMENTS IN LEAD OPTIMIZATION	205
<i>Hans Matter, Stefan Güssregen, Friedemann Schmidt, Gerhard Hessler, Thorsten Naumann, and Karl-Heinz Beringhaus</i>	
11 DEVELOPMENT AND APPLICATIONS OF GLOBAL ADMET MODELS: IN SILICO PREDICTION OF HUMAN MICROSOMAL LABILITY	245
<i>Karl-Heinz Beringhaus, Gerhard Hessler, Hans Matter, and Friedemann Schmidt</i>	
12 CHEMOINFORMATICS AND BEYOND: MOVING FROM SIMPLE MODELS TO COMPLEX RELATIONSHIPS IN PHARMACEUTICAL COMPUTATIONAL TOXICOLOGY	267
<i>Catrin Hasselgren, Daniel Muthas, Ernst Ahlberg, Samuel Andersson, Lars Carlsson, Tobias Noeske, Jonna Stålring, and Scott Boyer</i>	
13 APPLICATIONS OF CHEMINFORMATICS IN PHARMACEUTICAL RESEARCH: EXPERIENCES AT BOEHRINGER INGELHEIM IN GERMANY	291
<i>Bernd Beck, Michael Bieler, Peter Haebel, Andreas Teckentrup, Alexander Weber, and Nils Weskamp</i>	
14 LESSONS LEARNED FROM 30 YEARS OF DEVELOPING SUCCESSFUL INTEGRATED CHEMINFORMATIC SYSTEMS	321
<i>Michael S. Lajiness and Thomas R. Hagadone</i>	
15 MOLECULAR SIMILARITY ANALYSIS	343
<i>José L. Medina-Franco and Gerald M. Maggiora</i>	
INDEX	401

PREFACE

Chemoinformatics: From methods and models to pharmaceutical applications

Chem(o)informatics is a relatively young and still evolving discipline, although some of its scientific origins can be traced back at least five decades. It continues to be challenging to clearly define chemoinformatics as a scientific field. Essentially, chemoinformatics uses algorithms and computational methods, often adapted from computer science, to organize and process chemical data, analyze and predict structure–property relationships of small molecules, and design compounds. Although chemoinformatics is not confined to questions and tasks that are relevant for pharmaceutical research, this field has firm roots in drug discovery. In fact, when the term chemoinformatics was first introduced in the literature in 1998 (Brown FK. Chemoinformatics: What is it and how does it impact drug discovery. *Ann. Rep. Med. Chem.* 1998;33:375–384), there was a strong focus on drug discovery research—and this has been a characteristic of this field ever since. Accordingly, the study of biological activities of chemical compounds and analysis of their structure–activity relationships (SARs) are hallmarks of chemoinformatics as we understand it today. As a consequence, methodological boundaries between chemoinformatics, computational chemistry, and drug design are rather fluid. In more specific terms, chemoinformatics has been defined to cover a wide range of scientific topics, from chemical data collection, management, and analysis to the exploration of SARs and prediction of compound activity or *in vivo* properties (Bajorath J. Understanding chemoinformatics: A unifying approach. *Drug Discov. Today* 2004;9:13–14). The scientific diversity of the field is high (Warr WA. Some trends in chem(o)informatics. *Meth. Mol. Biol.* 2011;672:1–37) and likely to even further increase, given the advent of research disciplines such as chemical biology or nanoscience, for which concepts from chemoinformatics are also relevant. Despite the presence of fluid scientific boundaries, characteristic features of chemoinformatics include its large-scale character (i.e., very large numbers of compounds and activity data are processed and analyzed) and its dual purpose of generating computational infrastructures and predictive models or data mining methods. Given its roots, another characteristic feature of chemoinformatics is that many important developments have originated from

pharmaceutical environments, in addition to research carried out in academia. It is evident that the pharmaceutical industry is the place where the need for chemoinformatics technologies and experts has been and continues to be the greatest. One should also note that the chemoinformatics literature is dominated by reports of computational methods and benchmark investigations, rather than practical applications. This is not very surprising, given that the majority of pharmaceutical applications are a part of drug discovery campaigns and hence proprietary (at the least for the duration of a discovery project). However, there clearly is a need to evaluate and better understand what chemoinformatics can actually accomplish in practical drug discovery situations. This need is not sufficiently met by the current scientific literature.

Having briefly introduced chemoinformatics as a scientific discipline, I should address the question why this book was originally planned and ultimately written. What was the prime motivation? Different from other currently available textbooks on chemoinformatics (there are not many), this book was envisioned to mostly (but not exclusively) focus on practical applications of chemoinformatics approaches in pharmaceutical research, hence addressing the need referred to earlier. It was intended to bring together leading experts from the pharmaceutical industry and selected academic institutions to describe the practice of chemoinformatics, illustrate the interplay between academic and pharmaceutical research, and showcase collaborations. Among others, key questions for authors included: What does chemoinformatics mean to you? How is it applied in your specific research environment? How does chemoinformatics contribute to pharmaceutical research? What works? What does not? Hence, special emphasis was put on expert views and experience values that might reflect the “true” impact of chemoinformatics approaches in drug discovery. In addition, a few selected methodological concepts were considered to further expand the spectrum of the presentations.

The 15 chapters presented herein include contributions from major pharmaceutical companies, a leading software firm, and several academic groups. They also cover collaborative efforts between academia and pharma. The chapters are arranged to follow a conceptual path from the description of methods and models to drug discovery applications and the design of chemoinformatics infrastructures. Hence, they span a wide range of topics.

Chapter 1 by W. Patrick Walters from Vertex presents a practical guide to the generation and evaluation of predictive models. It emphasizes common pitfalls in model building and assessment and shows how to avoid them. Many practical examples are provided including source code, which results in an instructive and much needed contribution. In Chapter 2 by Ajay Jain of the University of California at San Francisco, computational methods and models are considered from a principal point of view. The argument is made—and well supported—that the success of computational models often depends on the incorporation of sound physical principles (termed physical reality), although their consideration inevitably also introduces approximations. A number of well-selected methodological examples are presented.

Chapter 3 by Gerald M. Maggiore of the University of Arizona and collaborators of the Mayo Clinic and the Torrey Pines Institute for Molecular Studies reports the

adaptation of a new approach for chemoinformatics, that is, rough set theory, and discusses opportunities of this approach for drug discovery applications. In Chapter 4, Kiyoshi Hasegawa of the Chugai Pharmaceutical Company and Kimito Funatsu of the University of Tokyo also introduce new methodology. Their collaborative effort describes the application of the bimodal partial least-squares regression technique to analyze compound activity data by taking both ligand and target representations into account. Furthermore, in Chapter 5, Anthony Nicholls and Brian Kelley of OpenEye Scientific Software investigate search characteristics of different types of two-dimensional fingerprints, which are among the most popular molecular representations for chemical similarity searching and ligand-based virtual screening. Nicholls and Kelley pay particular attention to the way molecular similarity relationships are accounted for by different fingerprint representations and analyze how similarity assessment might be biased by fingerprints having high or low chemical resolution. On the basis of their findings, differences in search characteristic between fingerprints of alternative design can be rationalized. Practical implications of these results and possible methodological extensions are also discussed. Chapter 6, a contribution from our research group, further expands on ligand-based virtual screening, puts the approach into scientific context, and presents a critical assessment of practical virtual screening applications. Then, Meir Glick and colleagues of the Novartis Institutes for Biomedical Research, the authors of Chapter 7, describe a variety of applications of Bayesian modeling methods in drug discovery. Bayesian methods currently are among the most popular chemoinformatics approaches for compound classification, activity prediction, and target assignment. The topics discussed in this contribution include the analysis of phenotypic screening data and the prediction of off-target effects of drugs.

The contributions described thus far largely focus on approaches for the identification and characterization of active compounds. Once new active molecules have been identified, early-phase drug discovery projects transition into the hit-to-lead and lead optimization phases. Chapter 8 by Darren Green of GlaxoSmithKline and Matthew Segall of Optibrium Ltd. presents a thoughtful account of the evolution of lead optimization strategies and illustrates how different chemoinformatics concepts are adapted to aid in the optimization process. This contribution is very well complemented by Chapter 9 that reports on lead optimization collaborations between academia and the pharmaceutical industry. This work involved Valerie Gillet and Peter Willett of the University of Sheffield and George Papadatos et al. of GlaxoSmithKline. Here, the use of compound arrays for lead optimization is the major topic. A variety of chemoinformatics approaches have been employed to aid in the design of compound arrays and analyze progress made over time in lead optimization projects. This contribution also illustrates practical constraints involved in data assembly that affect medicinal chemistry projects and often work against a systematic and timely application of computational methods during lead optimization. In Chapter 10, Hans Matter and colleagues of Sanofi-Aventis further extend the lead optimization theme. They present a thorough and extensively referenced review of chemoinformatics methodologies for the analysis and prediction of SARs and demonstrate how such approaches have specifically been adapted for in-house applications.

The chapter also contains a discussion of methods to transfer SARs from one chemical series to another, which is a topic of high interest in medicinal chemistry.

The optimization of leads and generation of clinical candidates is a complex multi-parametric process in which *in vivo* compound characteristics such as absorption, distribution, metabolism, extraction, and toxicology (ADMET) properties are as important as compound potency and specificity. The following two contributions address these issues. In Chapter 11, Karl-Heinz Baringhaus et al., also of Sanofi-Aventis, discuss how different types of computational ADMET models are generated and present a case study in which a model of human liver microsomal lability (a measure of metabolic instability of compounds) was derived for in-house use. Then, in Chapter 12, Scott Boyer and colleagues of AstraZeneca further expand the discussion of ADMET models with a focus on toxicology assessment. Their contribution also highlights the critically important role primary *in vivo* data play for predictive model building, given their sparseness and expected error margins. Both contributions cover a wide range of chemoinformatics methodologies for the derivation of ADMET models. With a concluding discussion of data delivery and communication issues, Chapter 12 also represents a transition point to another important thematic section of the book.

The contributions described thus far introduce scientific concepts, derive increasingly complex prediction models, and illustrate how such models are practically applied in drug discovery. As such, they represent a major category of chemoinformatics approaches in pharmaceutical research, that is, modeling and prediction of various compound properties. Another major category includes the design and implementation of computational infrastructures and information systems that is equally important for drug discovery as data mining and predictive modeling. In fact, pharmaceutical research environments heavily rely on the availability of specialized database structures and information systems to enable data warehousing with consistent deposition, distribution, access, and use across an organization. For large pharmaceutical companies, these requirements represent challenging tasks. The last two contributions in this book address these challenges. In Chapter 13, Nils Weskamp et al. describe how comprehensive chemoinformatics and database structures have been designed and implemented at Boehringer-Ingelheim. Here, it becomes clear that data archiving and handling is only a part of the equation—it is equally important to provide general access to modeling tools to, for example, analyze high-throughput screening data or characterize SARs. This presents considerable challenges for chemoinformaticians because such computational tools must not only be generated or adopted but also be made accessible to nonexpert users in the form of automated and easy-to-use workflows. In addition, results must be communicated in an intuitive and interpretable manner. Furthermore, in Chapter 14, Michael S. Lajiness and Thomas R. Hagadone of Eli Lilly and Company discuss lessons learned from over three decades of design and implementation of different generations of chemoinformatics systems for pharmaceutical research. These investigators are among the pioneers in building and maintaining such computational infrastructures in different company-specific environments. Their contribution illustrates how such systems have evolved, and continue to evolve, as computational resources and requirements rapidly change and data volumes and drug discovery demands further increase. On the basis of their

long experience, Lajiness and Hagadone comment on a number of practical aspects associated with system design that should be taken into consideration to ensure quality, accessibility, and utility of chemoinformatics infrastructures in drug discovery settings.

The book begins with chemoinformatics methodology and so it ends. To close the circle, in the final chapter (Chapter 15), José L. Medina-Franco of the Torrey Pines Institute for Molecular Studies and Gerald M. Maggiora of the University of Arizona describe foundations of molecular similarity analysis, one of the central themes in chemoinformatics. The evaluation and quantification of molecular similarity as an indicator of activity similarity is at the core of many chemoinformatics methods and an intensely investigated research topic to this date, conceptually linked to the design and navigation of chemical feature spaces.

Taken together, the contributions in this book highlight—from different points of view—key issues for the practice of chemoinformatics. The initial goals of this book project were quite ambitious and potential complications were expected. On the one hand, it was anticipated that it might be difficult for researchers in academia to present studies that are of high practical relevance for drug discovery; on the other hand, that it might be even more difficult for many investigators in the pharmaceutical industry to elaborate on details of their chemoinformatics work, given the proprietary nature of most of their projects. However, the chapters in this book have clearly exceeded initial expectations. Hence, I am very grateful to all authors who have spent their time and efforts to put together these excellent contributions! Without their early commitment and dedication, this project would not have been possible.

The contents of the book should be of interest to experts and practitioners in this field as well as to newcomers; there will be interesting materials for individuals with different motivations and levels of experience. Many of the questions that were initially asked have been answered in different ways and from different perspectives, which is highly desirable—after all, authors should have the last word.

Last but not least, given the critical expert views presented in this book and its practical drug discovery orientation, it is hoped that this publication will represent another important step forward in further defining and supporting chem(o)informatics as a scientific discipline at the interface between chemistry, computer science, and drug discovery.

JÜRGEN BAJORATH

CONTRIBUTORS

ERNST AHLBERG, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

SAMUEL ANDERSSON, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

JÜRGEN BAJORATH, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

KARL-HEINZ BARINGHAUS, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

BERND BECK, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

MICHAEL BIELER, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

SCOTT BOYER, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

LARS CARLSSON, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

KIMITO FUNATSU, Department of Chemical System Engineering, University of Tokyo, Tokyo, Japan

VALERIE J. GILLET, Information School, University of Sheffield, Sheffield, UK

MEIR GLICK, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

DARREN V. S. GREEN, GlaxoSmithKline Medicines Research Centre, Stevenage, Herts, UK

STEFAN GÜSSREGEN, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

PETER HAEBEL, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

THOMAS R. HAGADONE, Eli Lilly and Company, Indianapolis, IN, USA

KIYOSHI HASEGAWA, Chugai Pharmaceutical Company, Kamakura Research Laboratories, Kamakura, Kanagawa, Japan

CATRIN HASSELGREN, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

GERHARD HESSLER, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

AJAY N. JAIN, Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA

BRIAN KELLEY, OpenEye Scientific Software, Inc., Santa Fe, NM, USA

PETER S. KUTCHUKIAN, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

MICHAEL S. LAJINESS, Eli Lilly and Company, Indianapolis, IN, USA

EUGEN LOUNKINE, Novartis Institutes for BioMedical Research, Cambridge, MA, USA

CHRISTOPHER N. LUSCOMBE, GlaxoSmithKline, Medicines Research Centre, Stevenage, UK

GERALD M. MAGGIORA, College of Pharmacy and BIO5 Institute, University of Arizona, Tucson, AZ, USA; Translational Genomics Research Institute, Phoenix, AZ, USA

HANS MATTER, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

IAIN M. McLAY, The Open University, Cardiff, UK

JOSÉ LUIS MEDINA-FRANCO, Torrey Pines Institute for Molecular Studies, Port St. Lucie, FL, USA

NATHALIE MEURICE, Mayo Clinic, Scottsdale, AZ, USA

DANIEL MUTHAS, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

THORSTEN NAUMANN, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

ANTHONY NICHOLLS, OpenEye Scientific Software, Inc., Santa Fe, NM, USA

TOBIAS NOESKE, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

GEORGE PAPADATOS, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, UK

JOACHIM PETIT, Mayo Clinic, Scottsdale, AZ, USA

STEPHEN D. PICKETT, GlaxoSmithKline, Medicines Research Centre, Stevenage, UK

FRIEDEMANN SCHMIDT, R&D, LGCR, Structure, Design and Informatics, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

MATTHEW SEGALL, Optibrium Ltd., Cambridge, UK

JONNA STÅLRING, Global Safety Assessment, AstraZeneca R&D Mölndal, Mölndal, Sweden

DAGMAR STUMPFE, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

ANDREAS TECKENTRUP, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

W. PATRICK WALTERS, Vertex Pharmaceuticals Incorporated, Cambridge, MA, USA

ALEXANDER WEBER, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

NILS WESKAMP, Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany

PETER WILLETT, Information School, University of Sheffield, Sheffield, UK

CHAPTER 1

WHAT ARE OUR MODELS REALLY TELLING US? A PRACTICAL TUTORIAL ON AVOIDING COMMON MISTAKES WHEN BUILDING PREDICTIVE MODELS

W. PATRICK WALTERS

1.1 INTRODUCTION

Predictive models have become a common part of modern day drug discovery [1]. Models are used to predict a range of key parameters including:

- Physical properties such as aqueous solubility or octanol/water partition coefficients [2–4]
- Off-target activities such as CYP or hERG inhibition [5–7]
- Binding geometry and affinity of small molecules in protein targets [8].

When building these models, it is essential that the cheminformatics practitioner be aware of factors that could potentially mislead and confuse those using the models. In this chapter, we will focus on some common traps and pitfalls and discuss strategies for realistic evaluation of models.

We will consider a few important, and often overlooked, issues in the model-building process.

- How does the dynamic range of the data being modeled impact the apparent performance of the model?
- How does experimental error impact the apparent predictivity of a model?

2 WHAT ARE OUR MODELS REALLY TELLING US?

- How can we determine whether a model is applicable to a new dataset?
- How should we compare the performance of regression models?

The chapter will take a tutorial format. We will analyze some commonly used datasets and use this analysis to make a few points about the process of building and evaluating predictive models. One of the most important aspects of scientific investigation is reproducibility. As such, all of the analyses discussed in this chapter were performed using readily available, open source software. This makes it possible for the reader to follow along, carry out the analyses, and experiment with the datasets. All of the code used to perform the analyses is available in the listings section at the end of the chapter. The datasets and scripts used in this chapter can also be downloaded from the author's website <https://github.com/PatWalters/cheminformaticsbook>. It is hoped that these scripts will kindle an appreciation for aspects of the model-building process and will provide the basis for further exploration.

The software tools required for the analyses are

The Python programming language – <http://www.python.org>

The RDKit cheminformatics programming library – <http://www.rdkit.org>

The R statistics program – <http://www.r-project.org>

Python scripts can be run by executing the command

```
python script_name.py (Unix and OS-X)  
python.exe script_name.py (Windows)
```

where `script_name.py` is the name of the script to run.

R scripts can be run by executing the following two commands within the R console.

```
setwd("directory_path")  
source("script.R")
```

In the aforementioned commands, “`directory_path`” is the full path to the directory (folder) containing the scripts and data, and “`script.R`” is the name of the script to execute. The R scripts used in this chapter utilize a number of libraries that are not included as part of the base R distribution. These libraries can be easily installed by typing the command

```
source("install_libraries.R")
```

in the R console. Since these libraries are being downloaded from the Internet, it is necessary for your computer to be connected to the Internet when executing the aforementioned command.

Those unfamiliar with Python or R are urged to consult references associated with those languages [9–12]. We now live in a data rich world where every cheminformatics practitioner should possess at least rudimentary programming skills.

1.2 PRELIMINARIES

In order to better understand some of the nuances associated with the construction and evaluation of predictive models, it is useful to consider actual examples. In this chapter, we will examine a number of datasets containing measured values for aqueous solubility and use these datasets to build and evaluate predictive models. Solubility in water or buffer is an important parameter in drug discovery [13]. Poorly soluble compounds tend to have poor pharmacokinetics and can precipitate or cause other problems in assays. As such, the prediction of aqueous solubility has been an area of high interest in the pharmaceutical industry. Over the last 15 years, numerous papers have been published on methods for predicting aqueous solubility [2, 3, 14]. Although many papers have been published and commercial software for predicting aqueous solubility has been released, reliable solubility prediction remains a challenge.

The challenges in developing models for predicting solubility can arise from a number of experimental factors. The aqueous solubility of a compound can vary depending on a number of factors including:

- Temperature at which the solubility measurement is performed
- Purity of the compound
- Crystal form—different polymorphs of the same compound can have vastly different solubilities.

In addition to confounding experimental factors, a number of published solubility models are somewhat misleading due to a lack of proper computational controls. While we sometimes have limited control over the experimental data used to build models, we have complete control over the way models are evaluated and should always employ appropriate means of evaluating our models. In subsequent sections, we will use solubility datasets to examine some of these control strategies.

1.3 DATASETS

In this chapter, we will consider three different, publicly available, solubility datasets.

The Huuskonen Dataset This set of 1274 experimental solubility values ($\log S$) was one of the first large solubility datasets published [15, 16] and has subsequently been used in a number of other publications [14, 17]. The data in this set was extracted from the AQUASOL [18, 19] database, compiled by the Yalkowsky group at the

University of Arizona and the PHYSPROP [20] database, compiled by the Syracuse Research Corporation.

The JCIM Dataset This is a set of 94 experimental solubility values that were published as the training set for a “blind challenge” published in 2008 [21]. All of the solubility values reported in this paper were measured by a single group under a consistent set of conditions. The objective of this challenge was for groups to use a consistently measured set of solubility values to build a model that could subsequently be used to predict the solubility of a set of test compounds. Results of the challenge were reported in a subsequent paper in 2009 [22].

The PubChem Dataset A randomly selected subset of 1000 measured solubility values selected from a set of 58,000 values that were experimentally determined using chemiluminescent nitrogen detection (CLND) by the Sanford-Burnham Medical Research Institute and deposited in the PubChem database (AID 1996) [23]. This dataset is composed primarily of screening compounds from the NIH Molecular Libraries initiative and can be considered representative of the types of compounds typically found in early stage drug discovery programs. Values in this dataset were reported with a qualifier “>”, “=”, “<” to indicate whether the values were below, within, or above the limit of detection for the assay. Only values within the limit of detection (designated by “=”) were selected for the subset used in this analysis.

In order to compare predictions with these three datasets, we first need to format the data in a consistent fashion. We begin by formatting all of the data as Log S, the log of the molar solubility of the compounds. Data in the PubChem and JCIM datasets were originally reported in $\mu\text{g}/\text{ml}$, so the data was transformed to Log S using the formula

$$\text{LogS} = \log_{10}(\text{(solubility in } \mu\text{g/ml}) / (1000.0 * \text{MW}))$$

Where \log_{10} is the base 10 logarithm and MW is the molecular weight.

1.3.1 Exploring Datasets

One of the first things to consider in evaluating a new dataset is the range and distribution of values reported. An excellent tool for visualizing data distributions is the boxplot [24, 25]. The “box” at the center of the boxplot shows the range covered by the middle 50% of the data, while the “whiskers” show the maximum and minimum values (discounting the presence of outliers). Outliers in the boxplot are drawn as circles. More information on boxplots can be found on the Wikipedia page [26] and references therein. The anatomy of a boxplot is detailed in Figure 1.1.

Figure 1.2 shows a boxplot of the data distributions for the three solubility datasets mentioned earlier. Numeric summaries of the same data are shown in Table 1.1. Listing 1 provides the R code for loading and annotating the data, as well as generating Figure 1.2 and Table 1.1. The lower whisker and lower hinge in Table 1.1 define the lower extents of the boxplot, while the upper hinge and upper whisker define the upper extents. The interquartile range (IQR) defines the distance between the upper and lower hinges, while the range defines the distance between the upper and lower whiskers.

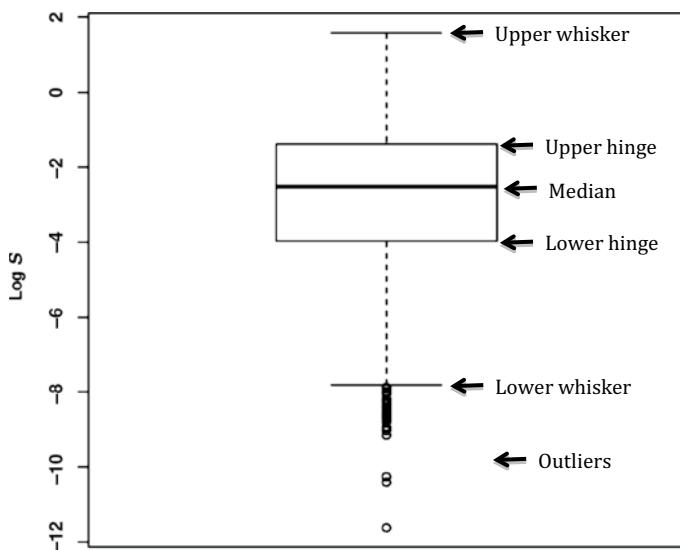


FIGURE 1.1 The anatomy of a boxplot.

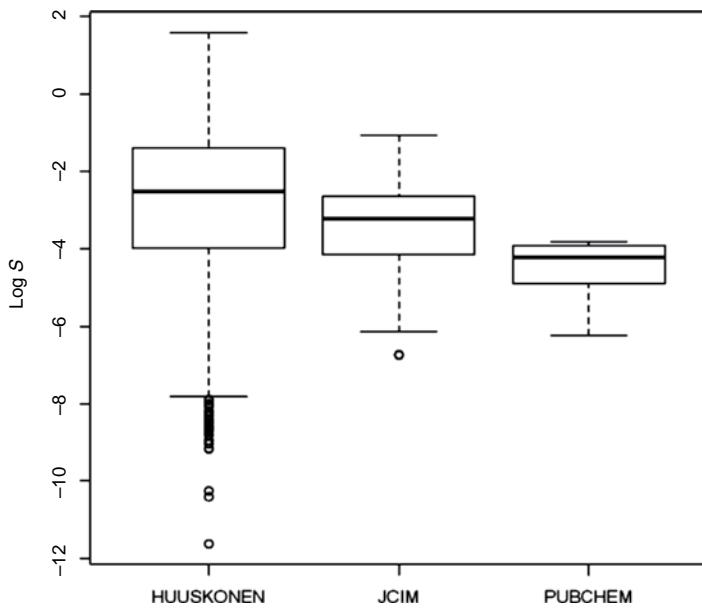


FIGURE 1.2 A boxplot comparison of Log S for the three datasets studied in this chapter.

In examining the datasets, we can see that the Huuskonen dataset spans more than 9 logs, while the JCIM dataset set spans 5 logs, and the PubChem dataset spans a much smaller 2.4 logs. Note that the IQR for the PubChem dataset is only about one log. Measured solubilities in drug discovery programs typically range between 1 and

TABLE 1.1 Boxplot Statistics for the Three Datasets Studied in This Chapter

Dataset	Lower Whisker	Lower Hinge	Median	Upper Hinge	Upper Whisker	Interquartile Range	Range
Huuskonen	-7.82	-3.98	-2.53	-1.39	1.58	2.59	9.4
JCIM	-6.14	-4.15	-3.23	-2.65	-1.06	1.5	5.08
PubChem	-6.23	-4.89	-4.22	-3.91	-3.82	0.98	2.4

100 μM ($\log S$ –6 to –4), so the PubChem dataset can be considered more representative of data that is commonly encountered in drug discovery than the other two datasets. As we will see, the range of data covered by a dataset can have a significant impact on the perceived performance of a model.

1.4 BUILDING PREDICTIVE MODELS

In order to build a predictive model, we need three things:

- Reliable experimental data (see earlier)
- Sets of molecular descriptors [27, 28] that can be derived from chemical structure
- Statistical or machine-learning [29] methods that can be used to model the association between the molecular descriptors and the experimental data.

Models can take a variety of forms, but are typically divided into two categories.

- *Classification models* attempt to determine the membership of an object in a group or category. In the case of aqueous solubility, we might attempt to classify molecules as soluble or insoluble based on a predetermined cutoff like 100 μM . We could classify all molecules with solubility greater than or equal to 100 μM as “soluble,” and all molecules with solubility less than 100 μM as “insoluble.” Of course, this cutoff can vary based on a particular situation or the needs of a project. Classification models can also contain multiple categories. For instance, molecules could be classified as “soluble,” “moderately soluble,” or “insoluble” based on two different cutoffs. There are a few problems with classification models. The first relates to what are called “edge effects.” In setting an arbitrary cutoff, we will invariably run into cases where similar values on either side of the cutoff will be placed into different classes. Let us consider a case where we have a two-class system with a cutoff of 100 μM . A value of 99 μM will be considered insoluble while a value of 101 μM will be considered soluble. Given the experimental error in the measurements, the values are the same, but are put into different categories. The other difficulty with classification models is that they provide limited direction for improving the properties of a compound. Ideally, a scientist on a drug discovery program would like to use a predictive model to

design new compounds with improved properties. While a classification model might be useful for designing a “soluble” compound from an “insoluble” compound, it will not allow one to improve the solubility of an already soluble compound.

- *Regression models* attempt to predict a real value (e.g. Log S) based on existing data and molecular descriptors. While regression models can be quite useful as part of an optimization effort. It is often difficult to create a regression model given data with a limited dynamic range. Many of the ADME assays used in drug discovery programs (e.g. solubility, hERG inhibition, CYP inhibition, metabolic stability, permeation) typically report data within a 2-log range. As we will see in subsequent sections of this chapter, a limited dynamic range makes it extremely difficult to produce reliable regression models.

1.5 EVALUATING THE PERFORMANCE OF PREDICTIVE MODELS

1.5.1 Pearson's *r*

By far, the most common method for evaluating regression models in the cheminformatics literature is Pearson's product-moment correlation [30], more commonly referred to as Pearson's *r*, or its square *r*². Pearson's *r* can be calculated in a number of ways; one of the most straightforward is shown next.

If we have paired values *X* and *Y* (e.g. predicted and corresponding experimental values), then we can calculate Pearson's *r* as:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_x} \right) \left(\frac{Y_i - \bar{Y}}{S_y} \right)$$

where \bar{X} and \bar{Y} are the means for *X* and *Y*, S_x and S_y are the corresponding standard deviations, and *n* is the number of data points.

Values of *r* can vary between -1 and 1, with 1 indicating a perfect linear correlation, -1 being a perfect inverse linear correlation, and 0 indicating an absence of correlation. The definition of what constitutes a “good” value for *r* is somewhat subjective and situation dependent, and could easily fill an entire chapter or even a book. As we will see in subsequent sections, the dynamic range of the data being considered can have a dramatic effect on Pearson's *r*. We will also see that when comparing values of Pearson's *r* for different models, we must consider the confidence intervals around *r*.

1.5.2 Kendall's Tau

One of the drawbacks of Pearson's *r* is that it is sensitive to outliers and to the distribution of the underlying data [31]. More recently, many in the cheminformatics community have begun to follow an example set many years ago by statisticians, and

reporting nonparametric measures of correlation like Kendall's tau [32]. Because these nonparametric methods employ the rank orders of values rather than the values themselves, they are less sensitive to data distribution or outliers. If we have a paired set of values X and Y , we can define Kendall's tau by counting the number of concordant and discordant pairs in the data. Pairs are considered concordant if their rank orders agree

$$x_i > x_j \text{ and } y_i > y_j \quad \text{or} \quad x_i < x_j \text{ and } y_i < y_j.$$

Pairs are considered discordant if their rank orders disagree

$$x_i > x_j \text{ and } y_i < y_j \quad \text{or} \quad x_i < x_j \text{ and } y_i > y_j.$$

Kendall's tau is then evaluated by considering all pairs,

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(1/2)n(n-1)}$$

where n is the number of pairs.

As we will see in subsequent code listings, Kendall's tau can be easily calculated by using the “Kendall” function in the “Kendall” library in R.

1.5.3 Root-Mean-Square Deviation (RMSD)

In addition to defining the correlation between predicted and experimental values, we need a means of measuring the magnitude of the error in the prediction. The most commonly used error measure in the cheminformatics literature is the root-mean-square deviation (RMSD), which is also known as the RMS error [33]. If we consider paired values X and Y , RMSD can be calculated using the following equation

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}}$$

where x_i and y_i are paired values (predicted and experimental) and n is the number of data points.

1.6 MOLECULAR DESCRIPTORS

Over the last 20 years, a vast array of molecular descriptors and machine-learning methods have been applied to the problem of building predictive models [27, 28]. A complete treatment of molecular descriptors and machine-learning methods is

beyond the scope of this chapter. Rather than explore the pros and cons of various descriptors and machine-learning methods, we will focus on model-building approaches that can be applied with any descriptors or machine-learning method. In the interest of reproducibility, we will employ a set of molecular descriptors and a machine-learning method that have been widely used and are readily available. The molecular descriptors we will use are the default set calculated by the RDKit cheminformatics programming library [34, 35]. This descriptor set contains a variety of topological indices and encodings for atom environments. As mentioned earlier, our focus here is on the analysis of the results rather than the specifics of the descriptors. The RDKit user manual contains a complete listing of the descriptors as well as the corresponding literature references. Listing 2 provides the Python code for the descriptor calculations used in this chapter.

1.7 BUILDING AND TESTING A RANDOM FOREST MODEL

A variety of machine-learning approaches have been applied to develop models relating chemical structure to physical properties and biological activity. These methods range from relatively simple approaches like linear regression to more sophisticated methods like support-vector machines. In this example, we will utilize the random forest machine-learning method that was originally published by Breiman [36]. The random forest method works by constructing an ensemble of decision trees. In practice, the method typically employs hundreds of decision trees and the consensus of multiple trees is used to generate a prediction. The random forest method as implemented in the “randomForest” library for the R statistical software can be used to perform either classification or regression [37, 38]. The choice of classification versus regression is made based on the type of the variable being predicted. If a categorical variable (e.g. “soluble” or “insoluble”) is being predicted, the method will perform classification. If a real valued variable is being predicted, the method will perform regression. In this example, we will generate a regression model. Listing 3 provides R code that demonstrates the construction and testing of a regression model using the random forest method. The steps involved in training and testing the model are listed in the following.

1. Integrate the experimental data and molecular descriptors
2. Divide the data into training and test sets
3. Build a model from the training set
4. Use this model to predict the test set

Note that training and test sets are produced by randomly sampling the dataset. In practice, we would perform this sampling multiple times (typically 50–100) to get a better idea of the overall performance of the model. Some authors have proposed other strategies for constructing training and test sets. There are publications and commercial software packages that advocate clustering the data and selecting one

TABLE 1.2 Statistics for Models Build Based on the Full Huuskonen Dataset and Subset with Log S Between –6 and –3

	Full Huuskonen Set	Realistic Subset
Training set size	891	303
Test set size	383	130
Pearson r	0.95	0.75
Kendall tau	0.81	0.56
RMS error	0.64	0.52

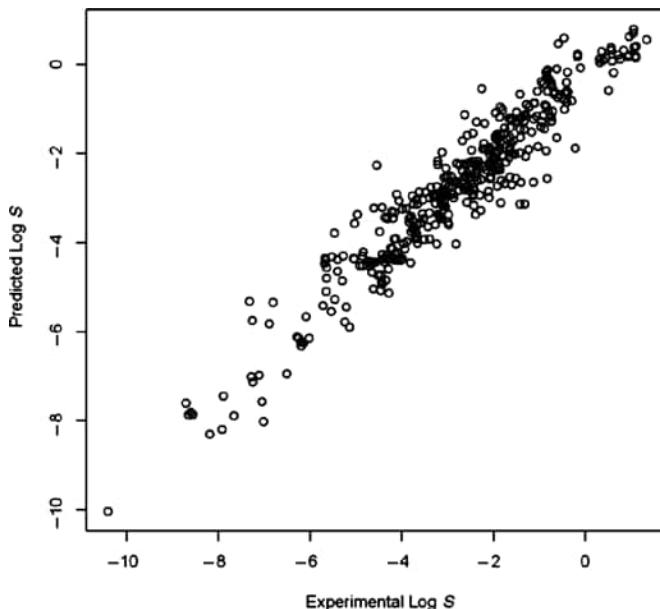


FIGURE 1.3 A plot of experimental versus predicted Log S for the Huuskonen test set.

training and one test compound from each cluster [39, 40]. In the opinion of this chapter’s author, this is a terrible idea. By sampling training and test sets in this fashion, one artificially inflates the performance of the method and creates unrealistic expectations from those using the model.

Now that we have built the random forest model, we can evaluate its performance on training and test sets generated from the Huuskonen dataset. Statistics for the test set performance are shown in the first column of Table 1.2. Figure 1.3 shows a plot of the predicted versus experimental Log S for the Huuskonen test set. The plot shows the performance of a model generated using the code in Listing 3. The model was trained using 70% of the compounds in the Huuskonen dataset and tested on the remaining 30%. At first glance, the performance appears quite good. The r for the test set is 0.95 and the RMS error is 0.62. These results are similar to the performance of literature methods on this dataset.

As an aside to those following along with the code in the listings, please note that your results will not exactly match those listed here. There is a stochastic component to the selection of training and test sets, as well as the construction of the random forest models. While your results probably won't match exactly what is reported here, they should be similar.

As we mentioned earlier, the Huuskonen dataset spans a wide range of solubility values. The range observed in this dataset is much larger than what is typically seen in drug discovery projects. Let's take a look at what would happen if we reduced this dataset to only those compounds with $\text{Log } S$ between -6 and -3 , a more typical range. Listing 4 demonstrates the selection of only those values within this range, and the subsequent construction of a model using a procedure identical to that described earlier for the full Huuskonen dataset. Statistics for model performance are shown in the second column of Table 1.2. In this case, the RMS error of the prediction for the subset is lower at 0.53 than that obtained with the larger set. However, the r for the subset is now 0.76 , lower than the value for the full Huuskonen set that spans a larger dynamic range.

As we can see, the dynamic range in a dataset can have a large impact on the apparent correlation between experimental and predicted activity. The literature is replete with examples of what appear to be impressive correlations on datasets that span an unrealistically high range. Authors who generate predictive models for protein–ligand binding affinity often use datasets that span up to 12 orders of magnitude [8]. The reality is that the binding affinity of compounds typically encountered in drug discovery programs may span 5–6 orders of magnitude. When data within this typical range is considered, these apparent correlations decrease dramatically. In practice, the utility of models for predicting–binding affinity is extremely limited.

Unfortunately, many cheminformatics practitioners have become enamored with r values, and linear plots of model performance. This has caused them to, consciously or unconsciously, choose datasets that often provide an unrealistic view of model performance. When building a predictive model, one should consider the dynamic range of the data being used to build the model and how this range compares with the range of the data to be predicted.

1.8 EXPERIMENTAL ERROR AND MODEL PERFORMANCE

One critical factor to keep in mind when building and evaluating predictive models is that every experimental data point has an error associated with it. For example, if we measure the $\text{Log } S$ of a compound as -6 and that data point has an error of 0.3 log units, the actual value could be anywhere between -6.3 and -5.7 . In a 2009 paper, Brown and coworkers [41] examined the relationship between experimental error and model performance. They carried out a series of theoretical experiments where Gaussian distributed random values were added to data to simulate experimental errors. The authors then calculated the correlation between the measured values and the same values with this simulated error. This correlation can be thought of as the maximum correlation possible given the error in the measurement. As we saw earlier,

TABLE 1.3 Maximum Possible Correlations for the Three Datasets Studied When Experimental Error Is 0.3, 0.5, and 1.0 Log

Dataset	Max@0.3	Max@0.5	Max@1.0
Huuskonen	0.98	0.94	0.81
JCIM	0.95	0.86	0.60
PubChem	0.80	0.60	0.27

this maximum correlation is a function not only of the error but also of the number of datapoints and the dynamic range of the data. The authors did a survey of predictive models in the molecular modeling and cheminformatics literature and found that correlations reported in 8 of 16 papers examined exceeded what could be expected based on experimental error alone.

Listing 5 shows how we can use R to add a random error to each data point and examine the correlation between experimental data and the data with added error. In Listing 5, we calculate the mean error for 1000 such simulations. Table 1.3 shows the maximum possible correlation for each of the three solubility datasets we have been examining when experimental errors of 0.3, 0.5, and 1.0 log are considered. Note that the error has much more of an impact on a dataset like the PubChem set, which has a limited dynamic range. If our experimental error is 0.5 log, then the best r^2 we can hope to achieve is 0.60. Also note that when the error is 1.0 log, almost half the dynamic range of the dataset, it is impossible to obtain any sort of meaningful correlation. Because the dynamic range of the Huuskonen dataset is (some might say unrealistically) large, the impact of error on the correlation is much less significant.

When building any sort of predictive model, it is important to carry out the type of analysis outlined earlier in order to account for experimental error. Any predictive models that perform better than the theoretical limit should be viewed with extreme skepticism.

1.9 MODEL APPLICABILITY

Although we would like them to be, predictive models are not universal. Regardless of the statistical or machine-learning method used, models built from molecular descriptors tend to do a reasonable job of predicting the activity of molecules similar to those in the training set. However, these models often perform poorly on molecules that bear little resemblance to those in the training set. This relationship between chemical similarity and model performance is detailed in a 2004 paper by Sheridan and coworkers [42]. In this paper, the authors examined 20 different predictive models and found that the best predictors of model performance were the similarity to the nearest neighbor in the training set and the number of neighbors in the training set. Neighbors are defined as molecules that are similar within a particular threshold (typically a Tanimoto similarity of 0.7–0.85). Molecules that were similar to those in the training set were predicted well, while those that were not similar were predicted poorly.

We can use the datasets employed so far to examine the relationship between similarity to the training set molecules and model performance. As before, we will

build a random forest model using 70% of the Huuskonen dataset (we will refer to this as our training set) and use this model to predict the solubilities of

- Remaining 30% of the Huuskonen dataset
- JCIM dataset
- PubChem dataset

For each of these datasets, we will also calculate the similarity of each molecule to every molecule in the training set. We will report the maximum similarity to any training set molecule as an indicator the molecule's similarity to the training set.

There are many ways to calculate molecular similarity, and a complete discussion of the topic is beyond the scope of this chapter. The interested reader is urged to consult some of the reviews referenced here [43–48] For purposes of illustration, we will calculate molecular similarity using 2D pharmacophore fingerprints as implemented in the RDKit library. Listing 6 provides the code for performing the similarity comparisons.

Once we have calculated the maximum similarity of each training set structure to each test set structure, we can use a boxplot to compare the similarities of our test sets. Listing 7 provides the R code for reading the similarity data, assigning labels to the datasets and plotting boxplots.

An examination of Figure 1.4 and Table 1.4 can give us some idea of the expected performance of a model built from the Huuskonen training set on our three test sets.

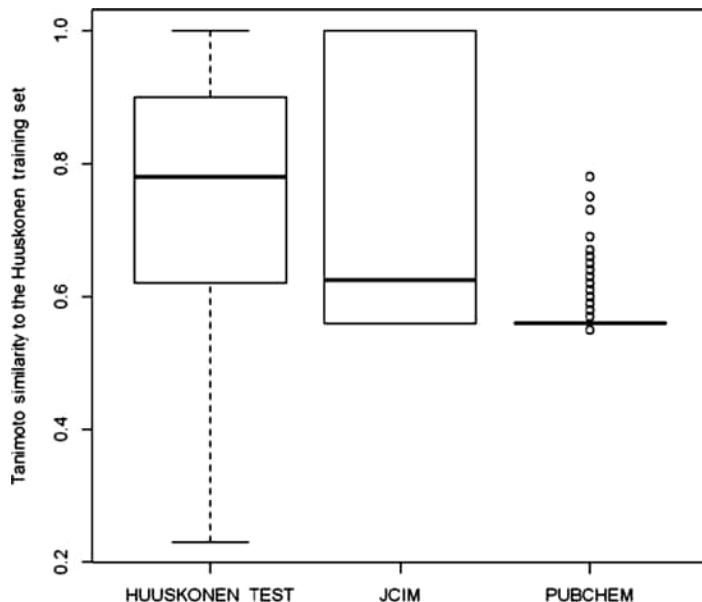


FIGURE 1.4 Tanimoto similarity of each of the three test sets to the Huuskonen training set.

TABLE 1.4 Similarity of Each Test Set to the Huuskonen Training Set

Dataset	Mean	Median
Huuskonen_Test	0.76	0.78
JCIM	0.74	0.62
PubChem	0.56	0.56

TABLE 1.5 Model Performance for the Three Test Sets

Dataset	R ²	Kendall	RMS Error
Huuskonen_Test	0.92	0.82	0.58
JCIM	0.58	0.59	0.83
PubChem	0.11	0.22	1.12

- The Huuskonen test set is similar to the training set. The median similarity for the training set is 0.78 and the mean is 0.76. By most standards, one would assume that approximately half of the molecules in the Huuskonen test set are similar to molecules in the training set. We would expect the performance of the model on this dataset to be reasonably good.
- The JCIM set is less similar to the training set. The median is 0.62 and the mean is 0.74. The boxplot and higher mean indicate that there are a number of compounds in this set that are similar to the training set. The absence of whiskers in the boxplot indicates that the similarities are more narrowly distributed. We would expect moderate performance from this dataset.
- The PubChem dataset appears to be very different from the training set. Note that the boxplot is almost flat, with a few outliers drawn as circles. The mean and median similarities to the training set are both 0.56. Similarity values in this range are what we would tend to expect from pairs of random compounds. We would expect extremely poor performance from this dataset.

Now that we have these predictions of performance in hand, we can generate a model based on our training set, use it to predict the test sets, and see if the model performance meets our expectations. Listing 8 shows how we can use R to build a model from our training set and test on each of the three test sets. Table 1.5 shows the r^2 , Kendall tau, and RMS error for the prediction of each of the three datasets earlier. The results are in line with our expectations based on the similarity of test set compounds to the training set. As was predicted earlier, the Huuskonen test set gives a reasonably good prediction with an r^2 of 0.92 and an RMS error of 0.58. Again, as expected based on similarity to the training set, the performance of the model on the JCIM dataset can be considered moderate. Finally, as expected based on similarity to the training set, we are unable to obtain reasonable predictions for the PubChem dataset.

When evaluating the potential performance of a model on a new dataset, it is important to perform this type of analysis to gauge whether one can expect useful

predictions to come from a model. This is particularly important when working with collaborators who may have a limited understanding of computational methods. Many pharmaceutical companies have set up automated prediction methods that allow scientists to simply sketch a molecule and obtain a prediction. While tools of this sort can make computation available to a wider audience, they can also mislead a naïve user. Any predictive tool, whether designed for an expert user or a novice, should provide some indication of whether the molecule being predicted falls within the applicability domain of the model.

1.10 COMPARING PREDICTIVE MODELS

In this final section, we will consider the comparison of two predictive models. The cheminformatics literature is replete with papers comparing predictive models. When developing a new method, it is always important to examine how the method compares with the current state of the art. However, when making comparisons, one must remember that correlations have an associated error. This error is a function of both the correlation coefficient and the number of data points used to obtain the correlation coefficient. When comparing correlation coefficients, we must not only consider the value of the correlation coefficient, but also the confidence intervals around the correlation coefficient. When we have a larger number of data points or a higher correlation coefficient, we are more confident in the correlation and our confidence interval is relatively narrow. When we have a smaller number of data points or our correlation coefficient is lower, the confidence interval around the correlation is larger. If the confidence intervals of two correlations overlap, we cannot claim that one predictive model is superior to another.

While the calculation of confidence intervals for a correlation is straightforward, it is rarely used in the cheminformatics literature. As such, we will provide a brief review of the method for calculating a confidence interval on a Pearson r . Since values of Pearson's r cannot exceed 1, its distribution is not normal. The distribution is closer to normal for lower values of r and becomes more skewed as r approaches 1. In order to calculate a confidence interval, values of r must be converted to Fisher's z' distribution using Equation 1.10.1.

$$z' = .5[\ln(1+r) - \ln(1-r)] \quad (1.10.1)$$

An r of 0.7 would then equate to a z' of 0.87. The z' distribution is normal and its standard deviation can be calculated as

$$\sigma_{z'} = \frac{1}{\sqrt{N-3}} \quad (1.10.2)$$

where N is the number of pairs used to calculate r . Confidence intervals can then be calculated as

$$z' \pm z\sigma_{z'} \quad (1.10.3)$$

For this example, let us assume we are dealing with a dataset where we are predicting the solubility of 23 molecules. According to standard tables of z values, we should use a value of 2.08 for a t test at 95% confidence for a two-tailed distribution with a sample size of 23. The value of $\sigma_{z'}$ would then be 0.23. For a Pearson r of 0.7, the confidence intervals on z' would be

$$\text{Lower limit} = 0.87 - (2.08 \times 0.23) = 0.39$$

$$\text{Upper limit} = 0.87 + (2.08 \times 0.23) = 1.35$$

Transforming the values from a z' distribution back to an r distribution, we end up with a confidence limit on r of

$$0.37 \leq r \leq 0.87$$

The confidence intervals for a Person's r , Kendall's tau, or Spearman's rho can be easily calculated using the function `cor.test` in R.

We will now consider the case of two predictive models, Model A and Model B, for aqueous solubility. We will use R to compare the performance of these models when tested on 25, 50, and 100 compounds. Listing 9 provides an example of how this comparison can be performed in R. In this listing, we first calculate the Pearson r and the upper and lower 95% confidence intervals for the Pearson r . Table 1.6 and Figure 1.5 show the correlations and associated bar plots. The bar plots show the value of Pearson r for each subset and the associated "whiskers" show the upper and lower limits of the 95% confidence interval.

From Figure 1.5 we can see that, for the subset of 25 compounds, the 95% confidence intervals overlap. When the confidence intervals overlap like this, we cannot say that one correlation is superior to the other. For the subset of 50 compounds, there is a very small difference between the upper bound of the 95% confidence interval for Method A (0.91) and the lower bound of the confidence interval for Method B (0.92). While there is some separation, we may still have doubts as to whether one method outperforms the other. The bars on the far right in Figure 1.5 show the correlations for the subset of 100 compounds and the associated

TABLE 1.6 Pearson r , Upper and Lower Bound for 95% Confidence Intervals for Regression Models

Dataset	Method	r	Confidence Interval	
			Lower Bound	Upper Bound
subset25	Method_A	0.80	0.60	0.91
subset25	Method_B	0.93	0.84	0.97
subset50	Method_A	0.85	0.75	0.91
subset50	Method_B	0.96	0.92	0.97
subset100	Method_A	0.81	0.73	0.87
subset100	Method_B	0.94	0.91	0.96

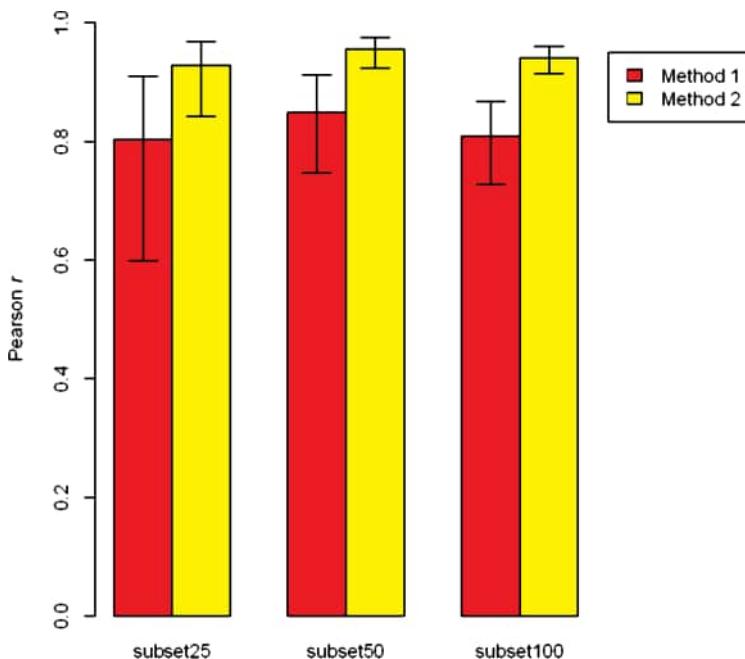


FIGURE 1.5 Comparing the errors associated with regression for two predictive models evaluated on three different datasets.

confidence intervals. In this case, there is a clear separation between the confidence intervals. As such, we can have some confidence that there is a difference between the correlation coefficients.

1.11 CONCLUSION

As mentioned in the introduction, cheminformatics and predictive modeling have become an integral part of modern drug discovery. The advent of “easy to use” software has opened the world of predictive modeling to a much wider audience. While it has become easier to generate and apply predictive models, it is still essential that we remain vigilant in evaluating both the statistical and chemical validity of the models. This chapter has presented a few factors that must be considered when evaluating the performance of predictive models:

- Dynamic range of the data
- Experimental error
- Applicability domain of the model
- Confidence intervals on correlation coefficients

It is hoped that the scripts presented in this chapter will provide the basis for a toolset that cheminformatics practitioners will use to generate and critically evaluate predictive models.

Cheminformatics is a relatively new discipline that encompasses a number of different fields. Many people come to cheminformatics from other fields, or use cheminformatic methods as an adjunct to experimental work. Standards for how cheminformatics professionals should be trained are still emerging. Hopefully as the field matures, practitioners will begin to appreciate the importance of a firm grounding in Statistics. Without an appreciation for the statistical foundations in our methods, it is difficult to see how our field will progress. This paper should not be considered a comprehensive guide to model building or evaluation. The objective here was just to point out a few commonly encountered pitfalls. The interested reader is urged to consult any of a number of excellent Biostatistics or data analysis texts [31, 49]. Another excellent additional source of information on model evaluation and comparison is the recent work of Anthony Nicholls [50–52].

REFERENCES

1. Gleeson MP, Hersey A, Montanari D, et al. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat Rev Drug Discov* 2011;10:197–208.
2. Delaney JS. Predicting aqueous solubility from structure. *Drug Discov Today* 2005;10:289–295.
3. Wang J, Hou T. Recent advances on aqueous solubility prediction. *Comb Chem High Throughput Screen* 2011;14:328–338.
4. Hughes LD, Palmer DS, Nigsch F, et al. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P. *J Chem Info Model* 2008;48:220–232.
5. Hudelson MG, Ketkar NS, Holder LB, et al. High confidence predictions of drug-drug interactions: Predicting affinities for cytochrome P450 2C9 with multiple computational methods. *J Med Chem* 2008;51:648–654.
6. Aronov AM. Predictive in silico modeling for hERG channel blockers. *Drug Discov Today* 2005;10:149–155.
7. Song M, Clark M. Development and evaluation of an in silico model for hERG binding. *J Chem Info Model* 2006;46:392–400.
8. Smith RD, Dunbar JB, Jr, Ung PM, et al. CSAR benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *J Chem Info Model* 2011;51: 2115–2131.
9. Ceder V. *The Quick Python Book*. 2nd ed. Greenwich: Manning Publications; 2010. p 400.
10. Beazley DM. *Python Essential Reference*. 4th ed. Addison-Wesley Professional; 2009. p 717.
11. Adler J. *R in a Nutshell: A Desktop Quick Reference*. O'Reilly Media; 2010. p 636.
12. Chambers J. *Software for Data Analysis: Programming with R* (Statistics and Computing). New York: Springer; 2010. p 514.

13. Lipinski C. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 2000;44:235–249.
14. Hou TJ, Xia K, Zhang W, et al. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J Chem Info Model* 2004;44:266–275.
15. Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Info Model* 2000;40:773–777.
16. Huuskonen J, Salo M, Taskinen J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J Chem Info Model* 1998;38:450–456.
17. Tetko IV, Tanchuk VY, Kasheva TN, et al. Estimation of aqueous solubility of chemical compounds using E-state indices. *J Chem Info Model* 2001;41:1488–1493.
18. Myrdal P, Ward GH, Dannenfelser RM, et al. AQUAFAC 1: Aqueous functional group activity coefficients; application to hydrocarbons. *Chemosphere* 1992;24:1047–1061.
19. <http://www.pharmacy.arizona.edu/outreach/aquasol/index.html>. Accessed 2013 May 14.
20. <http://www.srccinc.com/what-we-do/product.aspx?id=133>. Accessed 2013 May 14.
21. Llinàs A, Glen RC, Goodman JM. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J Chem Info Model* 2008; 48:1289–1303.
22. Hopfinger AJ, Esposito EX, Llinàs A, et al. Findings of the challenge to predict aqueous solubility. *J Chem Info Model* 2009;49:1–5.
23. Guha R, Dexheimer TS, Kestranek AN, et al. Exploratory analysis of kinetic solubility measurements of a small molecule library. *Bioorg Med Chem* 2011;19:4127–4134.
24. <http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/>. Accessed 2013 May 14.
25. <http://flowingdata.com/2012/05/15/how-to-visualize-and-compare-distributions/>. Accessed 2013 May 14.
26. http://en.wikipedia.org/wiki/Box_plot. Accessed 2013 May 14.
27. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics* (Methods and Principles in Medicinal Chemistry). Weinheim: Wiley-VCH; 2009. p 1257.
28. Karelson M. *Molecular Descriptors in QSAR/QSPR*. New York: Wiley-Interscience; 2000. p 448.
29. Goldman BB, Walters WP. Machine learning in computational chemistry. *Annu Rep Comput Chem* 2006;2:127–140.
30. http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient. Accessed 2013 May 14.
31. Glantz, S. *Primer of Biostatistics*. New York: McGraw-Hill Medical; 2011. p 320.
32. http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient. Accessed 2013 May 14.
33. http://en.wikipedia.org/wiki/Root-mean-square_deviation. Accessed 2013 May 14.
34. Landrum G, Lewis R, Palmer A, et al. Making sure there's a “give” associated with the “take”: Producing and using open-source software in big pharma. *J Cheminform* 2011;3:1–1.
35. <http://www.rdkit.org/>. Accessed 2013 May 14.
36. Breiman L. Random forests. *Mach Learn* 2001;45:5–32.

37. Liaw A, Wiener M. Classification and regression by random forest. *R News* 2002;2:18–22.
38. Svetnik V, Liaw A, Tong C, et al. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Info Comput Sci* 2003;43:1947–1958.
39. Yan A, Wang Z, Cai Z. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. *Int J Mol Sci* 2008;9:1961–1976.
40. Yan A, Gasteiger J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J Chem Info Model* 2003;43:429–434.
41. Brown SP, Muchmore SW, Hajduk PJ. Healthy skepticism: Assessing realistic model performance. *Drug Discov Today* 2009;14:420–427.
42. Sheridan RP, Feuston BP, Maiorov VN, et al. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J Chem Info Comput Sci* 2004;44:1912–1928.
43. Grant J, Haigh J, Pickup B, et al. Lingos, finite state machines, and fast similarity searching. *J Chem Inf Model* 2006; 46:1912–1918.
44. Hert J, Willett P, Wilton DJ, et al. New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Info Model* 2006;46:462–470.
45. Moffat K, Gillet VJ, Whittle M, et al. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J Chem Info Model* 2008;48:719–729.
46. Muchmore SW, Debe DA, Metz JT, et al. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J Chem Info Model* 2008;48:941–948.
47. Bender A, Jenkins JL, Scheiber J, et al. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Info Model* 2009;49:108–119.
48. Nicholls A, McGaughey GB, Sheridan RP, et al. Molecular shape and medicinal chemistry: A perspective. *J Med Chem* 2010;53:3862.
49. Pearson R. *Exploring Data in Engineering, the Sciences, and Medicine*. New York: Oxford University Press; 2011. p 792.
50. Nicholls A. What do we know and when do we know it? *J Comput Aided Mol Des* 2008;22:239–255.
51. Jain AN, Nicholls A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des* 2008;22:133–139.
52. Hawkins PCD, Warren GL, Skillman AG, et al. How to do an evaluation: Pitfalls and traps. *J Comput Aided Mol Des* 2008;22:179–190.

SOURCE CODE LISTINGS

install_libraries.R

```
install.packages(c("randomForest", "Kendall", "plyr", "gplots"), dependencies = TRUE)
```

listing_1.R

```
# load the solubility data
hus = read.table("huuskonen.sol",header = T,row.names = 1)
jcim_train = read.table("jcim_train.sol",header = T,row.names = 1)
pub = read.table("pubchem_sample.sol",header = T,row.names = 1)

# label the data
hus = labelData(hus,"HUUSKONEN")
jcim_train = labelData(jcim_train,"JCIM")
pub = labelData(pub,"PUBCHEM")

# combine the 3 datasets into 1 dataframe
allData = rbind(hus,jcim_train, pub)

# define an order for the x-axis in the boxplot
allData$DATASET = factor(allData$DATASET,levels = c("HUU SKONEN","JCIM","PUBCHEM"))

# display a boxplot of LogS vs Dataset and a corresponding
# table of boxplot statistics
print(summaryTable(allData))
```

listing_2.py

```
#!/usr/bin/env python

# Listing 2
# calculate molecular descriptors using the RDKit library

import sys,string
from rdkit import Chem
from rdkit.Chem import Descriptors

if len(sys.argv) != 3:
    print >> sys.stderr,"usage %s infile.smi outfile" %
(sys.argv[0])
    sys.exit(0)

# setup a molecule supplier for the input molecules
suppl = Chem.SmilesMolSupplier(sys.argv[1],titleLine =
False)
# open the output file
ofs = open(sys.argv[2],"w")
```

```
# create a list of descriptor names for the header
nameList = ["Name"] + [x[0] for x in Descriptors.descList[3:]]
print >> ofs, string.join(nameList)
# read the input molecules
for idx,mol in enumerate(suppl):
    print >> sys.stderr, "\r", idx + 1,
    sys.stderr.flush()
    print >> ofs,mol.GetProp("_Name"),
    # write out the descriptors
    for name,desc in Descriptors.descList[3:]:
        val = desc(mol)
        print >> ofs,"% .2f" % (val),
    print >> ofs
print >> sys.stderr, "\r",idx + 1
```

listing_3.R

```
# Listing 3
# train and test a random forest model based on the
Huuskonen dataset

# load the required libraries
library(randomForest)
library(Kendall)

# calculate the root mean squared error
rmsError<-function(a,b){
  sqrt(sum((a-b)**2)/length(a))
}

# split a dataset into training and test sets
splitTrainTest<-function(dataSet,trainingFraction=0.7) {
  idxList = sample(1:nrow(dataSet))
  numTrain = floor(trainingFraction * nrow(dataSet))
  trainIdx = idxList[1:numTrain]
  testIdx = idxList[(numTrain + 1):length(idxList)]
  list("train" = trainIdx,"test" = testIdx)
}

# merge descriptors and experimental data into a single
dataframe
mergeData <-function(descriptors,logS){
  mergedData = merge(logS,descriptors,by = 0)
  mergedData = mergedData[,-c(1,3)]
  mergedData
}
```

```

# use descriptors and experimental data to train and test
# a random forest model
predictRf <- function(mergedData, trainingFraction = 0.7)
{
  ttSplit = splitTrainTest(mergedData, trainingFraction)
  rf = randomForest(LOGS ~ ., mergedData[ttSplit$train,])
  pred = predict(rf, mergedData[ttSplit$test,])
  list("train" = ttSplit$train, "test" = ttSplit$test,
  "pred" = pred, "exper" = mergedData[ttSplit$test,]$LOGS,
  "model" = rf)
}

# read the data
desc = read.table("huuskonen.rdkit", header = T, row.names
= 1)
logS = read.table("huuskonen.sol", header = T, row.names = 1)
# merge descriptors and solubility data
mergedData = mergeData(desc, logS)
# build and test the random forest model
res = predictRf(mergedData)
# plot the test set results
plot(res$exper, res$pred, xlab = "Experimental LogS", ylab
= "Predicted LogS")
# output results
cat(sprintf("      Train = %d\n", length(res$train)))
cat(sprintf("      Test = %d\n", length(res$test)))
cat(sprintf(" Pearson r = %.2f\n", cor(res$pred, res$exper)))
cat(sprintf("Kendall          tau      = %.2f\n",
Kendall(res$pred, res$exper)$tau))
cat(sprintf("          RMS      error      = %.2f\n",
rmsError(res$pred, res$exper)))

```

listing_4.R

```

# Listing 4
# train and test a random forest model based on a subset
of the Huuskonen dataset
# note that this listing is just a variation on Listing 3

# load the required libraries
library(randomForest)
library(Kendall)

# calculate the root mean squared error
rmsError<-function(a,b){
  sqrt(sum((a-b)**2)/length(a))
}

```

```

# split a dataset into training and test sets
splitTrainTest <- function(dataSet,trainingFraction =
0.7){
  idxList = sample(1:nrow(dataSet))
  numTrain = floor(trainingFraction * nrow(dataSet))
  trainIdx = idxList[1:numTrain]
  testIdx = idxList[(numTrain + 1):length(idxList)]
  list("train" = trainIdx,"test" = testIdx)
}

# merge descriptors and experimental data into a single
dataframe
mergeData <- function(descriptors,logS){
  mergedData = merge(logS,descriptors,by = 0)
  mergedData = mergedData[,-c(1,3)]
  mergedData
}

# use descriptors and experimental data to train and test
a random forest model
predictRf <- function(mergedData,trainingFraction = 0.7) {
  ttSplit = splitTrainTest(mergedData,trainingFraction)
  rf = randomForest(LOGS ~ .,mergedData[ttSplit$train,.])
  pred = predict(rf,mergedData[ttSplit$test,.])
  list("train" = ttSplit$train,"test" = ttSplit$test,
"pred" = pred,"exper" = mergedData[ttSplit$test,.]$LOGS,
"model" = rf)
}

# read the data
desc = read.table("huuskonen.rdkit",header = T,row.names =
1)
logS = read.table("huuskonen.sol",header = T,row.names =
1)
# merge descriptors and solubility data
mergedData = mergeData(desc,logS)
# create a subset with LogS between -6 and -3
realisticSubset = mergedData[mergedData$LOGS >= -6 &
mergedData$LOGS <= -3,]
res = predictRf(realisticSubset)
# plot the test set results
plot(res$exper,res$pred,xlab = "Experimental LogS",ylab =
"Predicted LogS")
# output results
cat(sprintf("      Train = %d\n",length(res$train)))

```

```

cat(sprintf("      Test = %d\n",length(res$test)))
cat(sprintf(" Pearsonr=%.2f\n",cor(res$pred,res$exper)))
cat(sprintf("Kendall           tau          = 
%.2f\n", Kendall(res$pred,res$exper)$tau))
cat(sprintf("           RMS       error      = 
%.2f\n", rmsError(res$pred,res$exper)))

```

listing_5.R

```

# Listing 5
# add simulated error to experimental data to examine the
# impact of error on correlations

# calculate the maximum correlation that can be achieved
# give a measure of experimental error
# values - the experimental data
# error - the standard deviation of the error to be added,
# mean is 0
# repeats - the number of times to repeat the simulation
maxError <- function(values,error = 0.3,repeats = 1000) {
  correlationList = c()
  for (i in 1:repeats){
    errorList = rnorm(n = length(values),mean = 0,sd =
error)
    correlationList[i] = cor(values,values + errorList)**2
  }
  mean(correlationList)
}

# load the solubility data
hus = read.table("huuskonen.sol",header = T,row.names =
1)
jcim_train = read.table("jcim_train.sol",header = T,row.
names = 1)
pub = read.table("pubchem_sample.sol",header = T,row.
names = 1)

# write the output table
cat(sprintf("%-20s   %8s   %8s   %8s   %8s", "Dataset", "Max@0.3", "Max@0.5", "Max@1.0"))
for (data in list(c(hus,"Huuskonen"),c(jcim_
train,"JCIM"),c(pub,"PubChem"))){
  name = data[2]
  maxCorrelationList = c()
  i = 1
}

```

```
for (err in c(0.3,0.5,1.)){
  maxCorrelationList[i] = maxError(data[1]$LOGS,err)
  i = i + 1
}
cat(sprintf("%-20s %8.2f %8.2f %8.2f\n",name,
maxCorrelationList[1],maxCorrelationList[2],maxCorrelationList[3]))
```

listing_6.py

```
#!/usr/bin/env python

# Listing 6
# calculate similarity between pairs of SMILES files and
# report
# the most similar training set molecule for each test set
# molecule

import sys,string
from rdkit import Chem
from rdkit import DataStructs
from rdkit.Chem.Fingerprints import FingerprintMols

# build a list of fingerprints from an input file
def buildFingerprintList(fileName):
    suppl = Chem.SmilesMolSupplier(fileName,titleLine =
    False)
    fpList = []
    for idx,mol in enumerate(suppl):
        print >> sys.stderr,"\\rGenerating fingerprints for %s "
    "% (fileName),idx + 1,
        sys.stderr.flush()
    fpList.append([mol.GetProp ("_Name"),FingerprintMols.
    FingerprintMol(mol)])
    print >> sys.stderr,"\\rGenerating fingerprints for %s "
    "% (fileName),idx + 1
    return fpList

# for each molecule in queryFpList, find and report the
# most similar molecule in refFpList
def findMostSimilar(refFpList,queryFpList):
    outList = []
    for idx,[name,fp] in enumerate(queryFpList):
        print >> sys.stderr,"\\rCalculating similarity ",idx,
        sys.stderr.flush()
```

```

simList = [[x[0],DataStructs.FingerprintSimilarity(fp,x[1])] for x in refFpList]
simList.sort(lambda a,b: cmp(b[1],a[1]))
outList.append([name,simList[0]])
print >> sys.stderr,"\\rCalculating similarity ",idx + 1
return outList

# setup training and test sets
trainingFiles = ["huuskonen_train.smi"]
testFiles = ["huuskonen_test.smi","jcim.smi","pubchem.smi"]
trainDict = {}
testDict = {}

# open the output file
ofs = open("similarity.txt","w")

# generate fingerprints
for fileName in trainingFiles:
    trainDict[fileName] = buildFingerprintList(fileName)
for fileName in testFiles:
    testDict[fileName] = buildFingerprintList(fileName)

# write the output
print >> ofs,"Query Reference Similarity Dataset"
for trainFileName,trainFpList in trainDict.iteritems():
    for testFileName,testFpList in testDict.iteritems():
        print >> sys.stderr,"Processing %s" % testFileName
        simList = findMostSimilar(trainFpList,testFpList)
        for query,[ref,sim] in simList:
            print >> ofs,query,ref,"%2f" % (sim),testFileName.
split(".") [0].upper()

print >> sys.stderr,"Results have been written to simi-
larity.txt"

```

listing_7.R

```

# Listing 7
# load data and display box plots to compare distributions

# load the required library
library(plyr)

# read the data
d = read.table("similarity.txt",header = T)

```

```
# generate the boxplot
boxplot(Similarity ~ Dataset,d,ylab = "Tanimoto Similarity
to the Huuskonen Training Set")
# generate a table of mean and median for each dataset
res = ddply(d, c("Dataset"), function(x)c(mean(x$Similarity),median(x$Similarity)))
# format the results into a table
res = data.frame(res)
names(res) = c("Dataset", "Mean", "Median")
res$Mean = round(res$Mean, digits = 2)
res$Median = round(res$Median, digits = 2)
# print the table
print(res)
```

listing_8.R

```
# Listing 8
# predict activities of 3 test sets for a training set

library(randomForest)
library(Kendall)

# calculate the root mean squared error
rmsError <-function(a,b){
  sqrt(sum((a-b)**2)/length(a))
}

# merge descriptors and experimental data into a single
# data frame
mergeData <-function(descriptors,logS) {
  mergedData = merge(logS,descriptors,by = 0)
  mergedData = mergedData[,-c(1,3)]
  mergedData
}

# build a predictive model from descriptptors and experi-
# mental data
buildModel <-function(descriptors,logS) {
  mergedData = mergeData(descriptors,logS)
  rf = randomForest(LOGS ~ .,mergedData)
  rf
}

# use a model to predict test set activity
predictTestSet <-function(model,name,suffix) {
```

```

descriptors      =   read.table(paste(name,suffix,sep      =
""),header = T,row.names = 1)
logS = read.table(paste(name,".sol",sep = ""),header
= T,row.names = 1)
mergedData = mergeData(descriptors,logS)
pred = predict(model,mergedData)
r2 = cor(mergedData$LOGS,pred)**2
err = rmsError(mergedData$LOGS,pred)
kt = Kendall(mergedData$LOGS,pred)$tau
output = list("name" = name,"r2" = r2,"rms_error" =
err,"exper" = mergedData$LOGS,"pred" = pred,"kendall" =
kt)
output
}

descriptorSuffix = ".rdkit"

# read the training set and build a random forest model
huuskonenTrainDes = read.table(paste("huuskonen_train",
descriptorSuffix,sep = ""),header = T,row.names = 1)
huuskonenTrainSol = read.table("huuskonen_train.
sol",header = T,row.names = 1)
rf = buildModel(huuskonenTrainDes,huuskonenTrainSol)

# read the test sets and predict activity
testSets = c("huuskonen_test","jcim","pubchem")
cat(sprintf("%-15      s      %8      s      %8      s      %8
s\n","DataSet","R**2","Kendall","RMS Error"))
for (fileName in testSets){
  res = predictTestSet(rf,fileName,descriptorSuffix)
  cat(sprintf("%-15      s          %8.2f          %8.2f
%8.2f\n",res$name,res$r2,res$kendall,res$rms_error))
}

```

listing_9.R

```

# Listing 9
# calculate errors for regression and plot Pearson r with
associated error bars

library(gplots)

# read the data
rc = read.table("compare_regression.txt",header = T,row.
names = 1)

```

30 WHAT ARE OUR MODELS REALLY TELLING US?

```
# generate subsets of 25, 50, and 100
subset25 = rc[0:25,]
subset50 = rc[0:50,]
subset100 = rc[0:100,]

# createList (actually vectors) to loop over
nameList = c("subset25","subset50","subset100")
dataList = list(subset25,subset50,subset100)

# print the header
header = sprintf("%-15 s %-15 s %8 s %8 s %8 s",
  "Dataset", "Method", "R", "CI LB", "CI UB")
cat(header, "\n")

# loop over the 3 subsets
res = c()
for (i in 1:length(nameList)){
  name = nameList[i]
  data = data.frame(dataList[i])
  # calculate the correlation coefficents and confidence
  intervals
  resA = cor.test(data$LOGS,data$Method_A)
  resB = cor.test(data$LOGS,data$Method_B)
  # convert from a list to a vector of numbers
  outA = (as.numeric(c(resA$estimate,resA$conf.int)))
  # print the output
  strA = sprintf("%-15 s %-15 s %8.2f %8.2f %8.2f",name,
    "Method_A",resA$estimate,resA$conf.int[1],resA$conf.
    int[2])
  res = rbind(res,list(name,"Method_A",resA$estimate,
    resA$conf.int[1],resA$conf.int[2]))
  cat(strA, "\n")
  strB = sprintf("%-15 s %-15 s %8.2f %8.2f %8.2f",name,"M
  ethod_B",resB$estimate,resB$conf.int[1],resB$conf.int[2])
  res = rbind(res,list(name,"Method_B",as.numeric(resB$e
  stimate),resB$conf.int[1],resB$conf.int[2]))
  cat(strB, "\n")
}
printDf = data.frame(res)
names(printDf)      =      c ("DATASET", "METHOD", "R", "CI_LB",
  "CI_UB")

# reformat the results for a barplot with error bars
# reformat the r values
ma = as.numeric(printDf[printDf$METHOD=="Method_A",]$R)
```

```
mb = as.numeric(printDf [printDf$METHOD=="Method_B",] $R)
plotDf = data.frame(ma,mb)
row.names(plotDf) = c("subset25","subset50","subset100")
names(plotDf) = c("Method A","Method B")

# reformat the confidence interval lower bound
la = as.numeric(printDf [printDf$METHOD=="Method_A",
] $CI_LB)
lb = as.numeric(printDf [printDf$METHOD=="Method_B",
] $CI_LB)
lbDf = data.frame(la,lb)
row.names(lbDf) = c("subset25","subset50","subset100")
names(lbDf) = c("Method A","Method B")

# reformat the confidence interval upper bound
ua = as.numeric(printDf [printDf$METHOD=="Method_A",
] $CI_UB)
ub = as.numeric(printDf [printDf$METHOD=="Method_B",
] $CI_UB)
ubDf = data.frame(ua,ub)
row.names(ubDf) = c("subset25","subset50","subset100")
names(ubDf) = c("Method A","Method B")

# plot the barplot
par(fig = c(0,0.8,0,1))
barplot2(t(plotDf),beside = T,ci.l = t(lbDf),ci.u =
t(ubDf),plot.ci = TRUE,ylim = c(0,1),ylab = "Pearson r")
# plot the legend
par(fig = c(0.4,1,0,1),new = TRUE)
smartlegend(x = "right",y = "top",legend = c("Method
1","Method 2"),ncol=1,fill=c("red","yellow"))
```


CHAPTER 2

THE CHALLENGE OF CREATIVITY IN DRUG DESIGN

AJAY N. JAIN

2.1 DRUG DESIGN HISTORY: INCREMENTALISM AND SERENDIPITY

Figure 2.1 illustrates a common strategy in drug design, that of incremental modification of an existing therapeutic to produce minor changes while creating a protectable commercial asset. This is a product of three converging phenomena. First, it is easy to conceptualize molecules as they are drawn in 2D representations, and it is natural (and often correct) to suppose that molecules that are highly similar in 2D will share similar chemical and biological properties. This can drive a bias in reasoning about the activity of molecules that is only partially correct. While it is often true that molecules with high-2D similarity share common effects, it is also the case that molecules sharing very little 2D similarity can share common effects. Second, if one seeks to engineer a new therapeutic in an area with preexisting safe and effective agents, the degree to which one can identify a compound that is *maximally similar* to the existing ones will reduce the likelihood of revealing a novel biological effect that may modulate an adverse drug reaction. While such a strategy is correspondingly unlikely to produce significant new patient benefits, it is certainly a sound approach if the primary goal is to reach a lucrative market with a safe and effective therapeutic. Third, the primary criteria in making decisions for drug approval hinge upon safety and efficacy, with issues such as the degree to which a new therapeutic brings novel benefits being considered on an ad hoc basis.

We have quantitatively shown that there is a 2D bias in drug design [1]. We considered the relative similarity of pairs of drugs that shared the same desired primary biological target (called “primary target pairs”). We also considered pairs of drugs where one drug’s effect on the primary target of another was responsible for a *side*

effect (termed “side-effect pairs”). The 2D similarity of the primary target pairs was much higher than the 2D similarity of the side-effect pairs (as in Figure 2.1). In the case of side-effect pairs (as in Figure 2.2), the striking aspect is how different the molecules looked. In fact, the 2D similarity of side-effect pairs was often as low as that observed for drug pairs that shared no known biological targets at all. The only difference between the two distributions of similarity (the primary target pair similarity vs. the side-effect pairs similarity) was that in the case of the primary target drug pairs, human designers were thinking about the target in question, but in the case of the side-effect pairs, human designers accidentally designed a drug with an interaction against a particular target. The difference derives purely from human

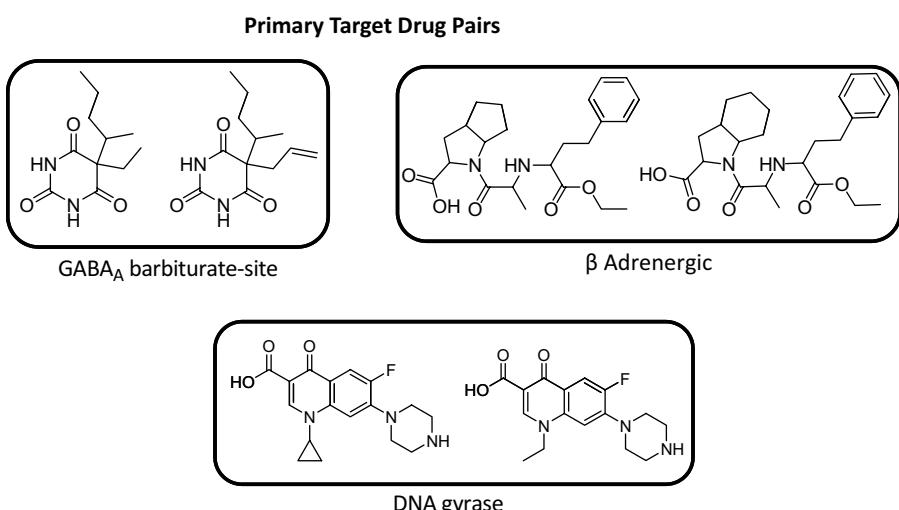


FIGURE 2.1 Examples of highly similar pairs of drugs designed to modulate the same primary targets.

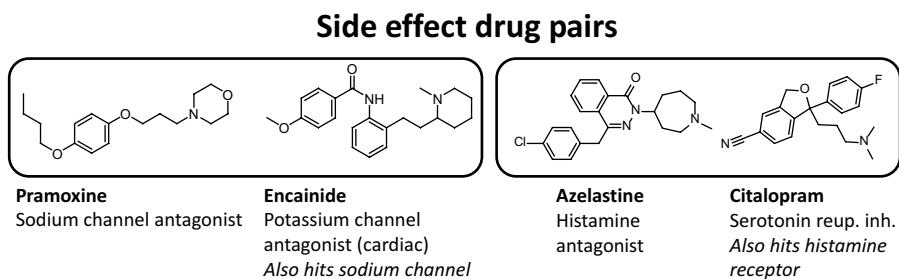


FIGURE 2.2 Typical examples of drug pairs where one modulates another’s primary target as a side effect.

design bias. As discussed earlier, this design bias is not irrational, given all of the incentives. Additional aspects of the underlying phenomena of primary and off-target activities, incentives in design, and bias in reasoning are discussed in several detailed studies [1–4].

While some degree of success can be achieved through conservative reasoning approaches, there are two important situations where nonobvious effects must be modeled and predicted accurately. The first situation arises when one would like to identify the *off targets* for a putative drug candidate. For me-too drugs that are extremely similar to pre-existing agents, this is less of a problem. But for first-in-class drugs or drugs with novel structures that may yield substantial new benefits, the off-target risks are very significant. The second situation arises in the design-phase of a first-in-class drug or one where multiple off-patent drugs exist. In the latter case, significantly new biological effects will be required in order to compete with the inexpensive off-patent medications. Such effects can be driven by structural creativity. We have shown that drug pairs with distant patent dates have much lower 2D similarity than ones with close patent dates [1], echoing this collision of economic incentive with design imperative. We have also shown that drug pairs sharing lower structural similarity, despite having identical primary targets, have much lower overlap in off-targets than drug pairs with high-2D structural similarity [3]. Clearly then, as one pursues challenging design objectives, modeling methods are presented with very serious questions.

2.2 PHYSICAL REALITY AND COMPUTATIONAL METHODS

Figure 2.3 depicts a case of structure-activity variation that encompasses a number of aspects of the physical processes governing the specific binding of a ligand to a protein-binding pocket. Four muscarinic antagonists are shown, beginning with a weakly potent ligand containing an unsubstituted furan. Substitution with either a phenyl or with a benzofuran produces great gains in potency. However, the combination of both substitutions produces a ligand with lower potency than either single substitution. A very simple physical explanation of this is cartooned in the left-hand portion of Figure 2.3. Each substitution produces a slightly different alignment, with the simple furan having ample room. Both the phenyl and benzofuran variants make favorable contacts, but at different sides of a hydrophobic cavity, with each requiring a small change in the position of the core scaffold. The combination of the phenyl with the benzofuran is simply too big to fit in the cavity without strain. In general, one should never expect perfect additivity through combination of substitutions on a common scaffold. Any degree of alignment preference difference for two substituents at different positions may produce a subadditive result, which is the typical case. However, as in this example, anti-additive behavior is also possible. Of course, though relatively less common, supraadditivity is also possible. This can occur when a substitution at one position produces a conformational change in a binding pocket that is favorable for the accommodation of a substituent at another position.

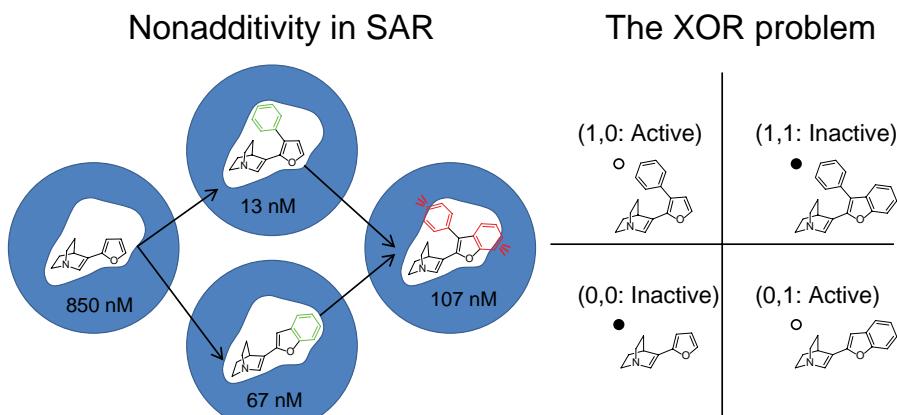


FIGURE 2.3 Four muscarinic antagonists, all on a quinuclidinene-furan scaffold, exhibit a pattern of potency variation that is both highly nonadditive and is isomorphic to the classic XOR problem. For color details, please see color plate section.

The combination of activity variations in this example can be seen to be isomorphic to the exclusive-or problem in computer science, where (1,0) and (0,1) map to TRUE and (0,0) and (1,1) map to FALSE. The significance of this problem in machine learning was pointed out in the late 1960s. By the early 1960s, linear learning models called perceptrons had been developed as a means to develop machine intelligence [5]. In 1969, Marvin Minsky and Seymour Papert [6] published a monograph in which they observed that the perceptron approach could not learn the logical function of exclusive-or. This observation was the principle factor in the decline of research on network learning models such as perceptrons. It took two decades before nonlinear modeling methods with corresponding parameter estimation regimes were shown to be able to address the XOR problem [7]. Such methods include neural networks, support-vector machines, random-forest learning, and many modern statistical machine learning algorithms (these issues are discussed in more detail in two papers [4, 8]). In retrospect, recognition of the simple fact that an entire class of models could not capture a very simple phenomenon was very good for the field of machine learning. It stimulated effort to develop new and better methods, many of which have been extremely successful.

As with the XOR problem for machine learning, we believe that it is vital for molecular modeling, as a field, to address the central dogma of physical reality in how most drug molecules exert biological effects through modulation of protein activity. Protein-ligand-binding interactions have four properties that are crucial to model in any physically sensible approach to prediction of likely targets or of binding affinity. First, the interactions are dependent on the conformation and relative alignment of both the protein and ligand. Second, that modifications on a ligand scaffold often produce changes in scaffold pose and may also produce changes in binding

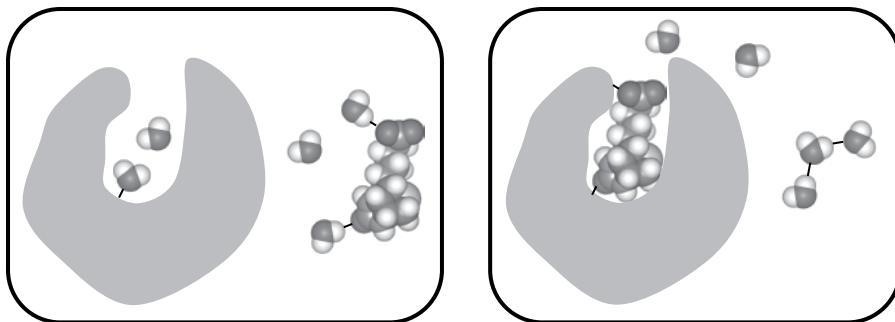


FIGURE 2.4 Modeling protein ligand binding involves the differences between solvated unbound and bound states.

pocket conformation. Third, that because of this, the problem of potency prediction is *necessarily* nonlinear. Fourth, ligands with very different underlying molecular scaffolds can bind the same protein-binding site, occupying the same volume in a competitive manner.

Approaches to molecular similarity computation and activity prediction have been developed and are in wide use that incorporate none of these aspects. Such methods can produce accurate predictions within a narrow range of structural variation, and use of such methods and design principles that echo them yield molecular data sets that can then be used post facto to validate more such methods. We believe that phenomena such as the off-target activities in Figure 2.2 and the SAR variation in Figure 2.3 are important to model and that they can be predicted effectively. In order to do so, however, it is absolutely necessary to use methods that address the central dogma of physical reality. There are three families of such methods: docking and other structure-focused physical simulation algorithms, 3D ligand-based similarity approaches, and some 3D-QSAR techniques. The following will illustrate how each method specifically addresses aspects of physical reality that are otherwise ignored by simpler techniques.

2.2.1 Protein Structure-Based Methods

The family of methods that make direct use of experimentally determined protein structures has their fundamental basis in direct physical simulation. Figure 2.4 cartoons the process of protein and ligand in a solvated and unbound state (left) moving into a bound state (right). While the cartoon may appear simple, the specific physical interactions each have multiple enthalpic and entropic effects, some of which are indirect. The association of a hydrophobic ligand surface with a complementary hydrophobic protein surface involves a direct enthalpic effect from the Van der Waals interactions. But the indirect effects include those from solvent molecules that had occupied the protein site in a semiordered state, which are released as a consequence of ligand occupancy near the protein. These solvent molecules may achieve greater entropy when free within solvent, and they may make more enthalpically favorable

hydrogen-bonding interactions outside the binding site, particularly if the binding site is highly hydrophobic in nature. Additional effects include entropic losses for the ligand on binding (possibly also for the protein). Similar considerations apply to the association of complementary polar moieties. Direct methods that make use of the partition function can be used in computationally expensive physical simulations to model these processes.

However, even molecular docking approaches that make very substantial approximations respect physical reality enough to support the discovery of non-obvious protein-ligand interactions. The field of small molecule docking was initiated by the pioneering work of Kuntz and Blaney in the 1980s [9]. While they treated both ligands and proteins as rigid bodies, the approach was still able to identify novel compounds. Practical and fully automatic methods that addressed ligand flexibility began to appear in the 1990s, with AutoDock [10, 11], GOLD [12, 13], Hammerhead [14–16], and FlexX [17, 18]. One of the key drivers of innovation and improvement in docking has come from the development and use of public benchmarks to measure pose prediction and virtual screening performance. The earliest efforts typically demonstrated very limited validation, usually with just a handful of examples of cognate ligand redocking. The publication of the 1997 GOLD validation paper [13], which reported pose prediction performance on 100 complexes, changed the scale and comprehensiveness of validation experiments. Cognate docking benchmarks were followed by those for virtual screening assessment [19]. Development of the Surflex-Dock approach (first described in 2003 [20]), the descendant of the Hammerhead system, benefited a great deal from the availability of both types of benchmarks. As multiple approaches began to exhibit strong performance on existing benchmarks, independent benchmarking of algorithms on *new* benchmarks became prevalent. Studies from Perola et al. [21] and Warren et al. [22] were particularly influential.

While ligand flexibility had been addressed early on in the docking field, practical and serious approaches for treating protein flexibility have only recently begun to appear. This has been prompted, in part, by direct evidence showing the limitations of cognate ligand redocking as a means of testing pose prediction. Studies by Sutherland et al. [23] and by Verdonk et al. [24] both showed that even relatively minor protein pocket changes could drastically affect the likelihood with which a docking algorithm could correctly predict the bound pose of a ligand. Such studies made a formal linkage between a specific physical reality and a performance consequence. Because protein pocket conformations often change on binding different ligands, and because the energy surfaces of scoring functions for docking can be quite bumpy, pose-ranking performance for noncognate ligands was poor when treating proteins rigidly. Figure 2.5 illustrates the problem and a partial solution. Sildenafil was the first-in-class PDE5 inhibitor for treating erectile dysfunction. Tadalafil was discovered and approved later, and it is not possible to fit it into the PDE5 pocket conformation that was derived by cocrystallization with sildenafil. It is possible, however, to approach the protein flexibility problem in a practical manner. Large variations in pocket configuration can be captured using multiple experimentally determined alternatives, here by using three variants of the PDE5 structure, each

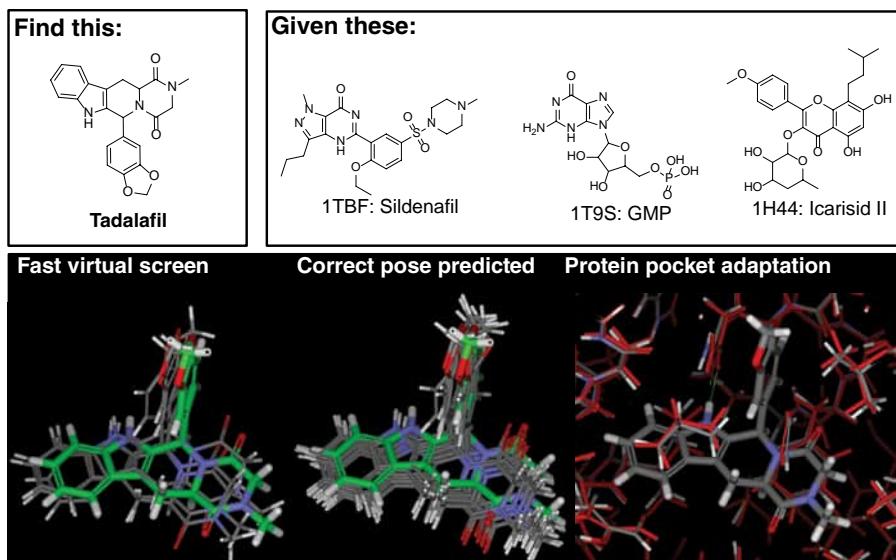


FIGURE 2.5 It is possible to make very substantial structural leaps with docking. However, it can be necessary to make use of multiple protein conformations and to flexibly adapt the pocket in order to identify a ligand and its bound pose correctly. For color details, please see color plate section.

derived with a ligand very different from tadalafil. Smaller variations, which may be important in precise scoring, can be explored by making use of a pocket adaptation and ligand rescoring protocol. By combining these approaches, it is possible to both identify tadalafil in a virtual screen *and* to correctly predict its bound pose (see [4, 25–27] for additional examples and discussion of these issues).

Within the molecular modeling field, those methods that make direct use of protein structures have made sustained progress in large part because methodological development has been spurred by the availability of benchmark data sets that clearly reveal the relationship between physical reality and algorithm performance. This has been a natural consequence of the atomistic representation of protein structures. The challenge for molecular modeling in areas that do not make direct use of protein structures is to have the discipline to maintain a respect for physical reality even when more abstract approaches may appear to yield good results.

2.2.2 Molecular Similarity

Direct comparison of small molecules can support inference: high similarity often correctly suggests shared interactions. However, 2D similarity methodology explicitly ignores conformation and alignment dependence and is incapable of inferring relatedness in cases such as those shown in Figure 2.2, where divergent scaffolds yield similar shape and electrostatic properties. Clearly, in order to address the

concerns of physical reality, one must go beyond 2D molecular comparison. Aspects of molecular shape, conformational variation, and molecular alignment must all be addressed. These requirements make physically realistic molecular similarity computation a complex endeavor, but there are now practical and widely applicable methods that fully respect the prerequisites.

One of the earliest descriptions of the use of molecular shape in relation to the biological activity of small molecules is due to Hopfinger [28]. The conceptualization of shape comparison was based on volume overlap of molecules that were modeled as collections of spheres. The notion of spherical volume overlap is the foundational concept of a family of molecular similarity approaches, best exemplified in current practice by the ROCS approach [29]. A separate line of thought characterizes molecular similarity by *surface overlap*, and one of the earliest descriptions of this notion is due to Masek et al. [30]. In that approach, molecules were characterized as having “skins” of a particular thickness, and the volume of the surface was described by the difference between a collection of spheres with standard atomic radii and one of radii made larger by the skin thickness. Similarity was measured based on the shared skin overlap between two molecules, offering some advantages over volumetric approaches, for example, when comparing molecules of very different overall sizes. The notion of molecular surface comparison is best exemplified by the Surflex-Sim approach [31], which owes its genesis to the Compass 3D-QSAR approach [32].

The appeal of using surfaces instead of volumes has two aspects. First, interactions between small molecules and proteins occur between surfaces, and there is a direct relationship between binding-free energy and encapsulated hydrophobic surface area of a ligand. So, if one accepts the basic proposition that closer approximations to physical reality are to be preferred in predictive modeling, the surface formulation is preferable. Second, as pointed out by Masek et al. [30] with their molecular skins approach, comparison of molecules with different sizes based on shared volume maximization can produce odd results (i.e., embedding a small molecule in the middle of a larger one). While conceptually attractive, the molecular skins approach was computationally burdensome. A different approach to capturing molecular surfaces was proposed during the development of the Compass 3D QSAR technique [32]. A collection of observation points (conceptually a virtual protein) was used from which to measure the minimum distance to a molecule’s surface, and this distance was compared to a learned ideal distance. This basic concept was quickly generalized to define a similarity measure that used a Gaussian reward function [33]. Similarity functions of this type correspond very closely to a surface-density function formulation of molecular shape, as follows.

$$M_i(r_i) = e^{(-(r_i - \mu_i)^2 / \gamma)} \quad (2.2.1)$$

$$E_k^P(r_k) = e^{(-(r_k - d_k)^2 / \gamma)} \quad (2.2.2)$$

$$R(r) = \left(\sum_i M_i \right) \left(\sum_k E_k^P \right) \quad (2.2.3)$$

In a volume-oriented density function such as that used by ROCS, Gaussian functions are atom-centered. In the surface-oriented formulation, the M_i of Equation 2.2.1 are Gaussians with peaks at the atomic surface (set by the atomic radii, denoted μ_i). By itself, the sum over the M_i produces *internal* molecular surfaces in addition to external ones. The E_k^P of Equation 2.2.2 defines Gaussians on local radial co-ordinates around each observer point from set P , with peaks at the *molecular* surface (set by the minimum distances from the observers to the molecule, denoted d_k). When γ is chosen carefully, the integral of the product of two molecules' surface-density functions R (defined in Eq. 2.2.3) is very closely approximated by the morphological similarity function used by Surfflex-Sim [31].

Figure 2.6 depicts the volumetric and surface-density functions for benzamidine. The molecule benzamidine was placed in a coordinate frame, such that the XY plane bisected the aromatic ring. On the left, the volume-density function of ROCS is depicted (computed to sixth order for intersections), with the red-shaded area indicating the points on the XZ plane with significant computed density. On the right, the surface-density function R is depicted, again with the significant density shown with red shading. The green curves indicate the relative value of the density functions along the X axis, penetrating two hydrogen atoms and three aromatic carbons. The area of significant volume density closely covers the collection of atomic spheres, with some smoothing at saddle points. The green curve exhibits five maxima, corresponding to each of the atomic centers upon which the Gaussians were centered. The surface-density function leaves the interior of the molecule with extremely low values, creating a peaked zone that also shows smoothing at saddle points.

These density functions form the underlying physical basis of two families of similarity computations, one which equates shape similarity with volume overlap and the other with surface congruence. In each case, molecular similarity can be

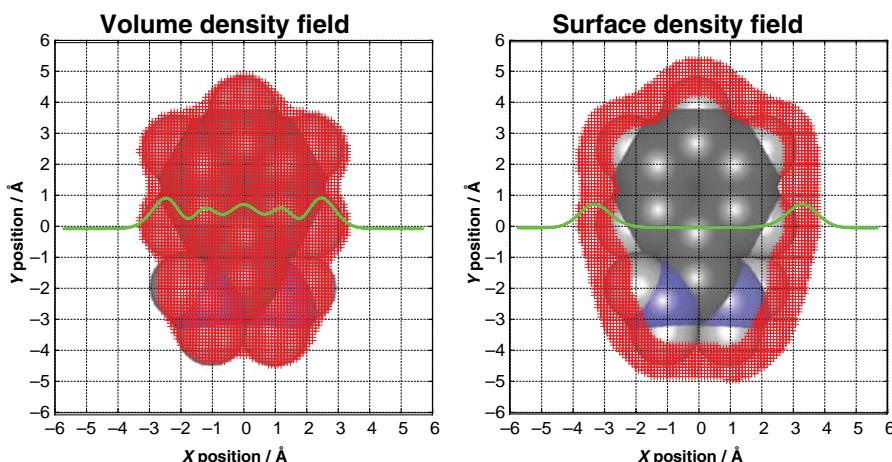


FIGURE 2.6 Volumetric and surface-based molecular density functions for benzamidine. For color details, please see color plate section.

computed based upon the integral of the product of two molecules' density functions. In the surface-density approach (Eq. 2.2.3), the density function was the product of two sums, one “lighting up” the surfaces of each atom of a molecule and the other lighting up the surfaces of spheres packed around the molecule. The product produces a function that has significantly nonzero values only at points close to the overall molecular surface, as shown in Figure 2.6. Consider two molecules *A* and *B* and one set of “observation” points *P*, giving rise to the following two density functions.

$$R^A(r) = \left(\sum_i M_i^A \right) \left(\sum_k E_k^{P,A} \right) \quad (2.2.4)$$

$$R^B(r) = \left(\sum_i M_i^B \right) \left(\sum_k E_k^{P,B} \right) \quad (2.2.5)$$

Here, the two surface-density functions are defined with respect to a *single* set of observations points *P*. The spheres that “pack” around each of the two molecules *A* and *B* share the same centers, but they have different radii, depending on the minimum distance to each molecular surface. As with the ROCS approach, one can define a similarity metric in terms of the overlap integral of the product of the two surface-density functions. This function is very closely approximated by the function computed by Surflex-Sim, simplified slightly in what follows.

$$S_k^P(d_k^A, d_k^B) = e^{\left(-(d_k^A - d_k^B)^2\right)/\sigma} \quad (2.2.6)$$

$$S_{A,B}^P = \sum_k S_k^P(d_k^A, d_k^B) \quad (2.2.7)$$

Here, the Gaussian terms are soft reward functions for concordance of the distances from the observer points *P* measured to the molecule surfaces of *A* and *B* (denoted (d_k^A, d_k^B)). When σ is roughly twice γ from Equations 2.2.1 and 2.2.2, the equivalence to the surface overlap integral holds. The intuition behind the metric is simple: when the minimum distances from each observer to each molecule are similar, the molecules must exhibit the same surface shape. The initial generalization of the Compass conceptualization of molecular surfaces to a similarity function [33] made use of two concentric spheres of observation points with radii of 6.0 and 9.0 Å, but this definition was somewhat restrictive. The morphological similarity function used by Surflex-Sim [31, 34] defines an *infinite* grid of observer points, with weights set such that a shell of observer points around each molecule subject to a comparison contribute. In practice, finite observer sets having significant weight are selected, and alignment optimization is done using the set constructed with respect to the query ligand. Similarity scores are reported using that set, another set constructed with respect to the final aligned new ligand, and a merger of the two.

Figure 2.7 shows the optimal overlay between aminomethylcyclohexane (AMC) and benzamidine according to both the volume and surface-oriented formulations just discussed using pure shape. Benzamidine is shown in green, with AMC shown in magenta

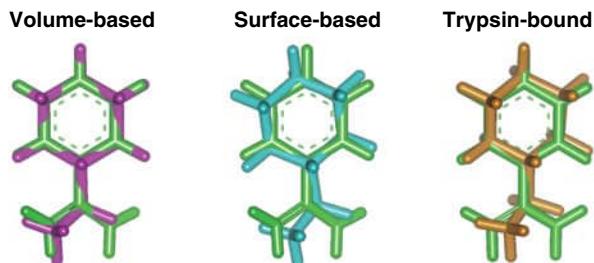


FIGURE 2.7 Volume- and surface-based alignment of aminomethylcyclohexane to benzamidine. For color details, please see color plate section.

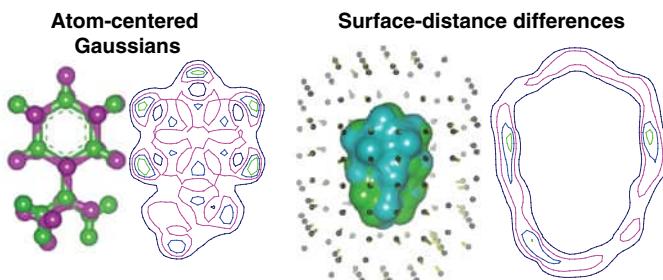


FIGURE 2.8 Relationship of molecular alignments to underlying similarity functions. For color details, please see color plate section.

(left) and cyan (middle). At right, the relationship between the ligands derived from their experimentally bound poses to trypsin is shown (PDB codes 1TNG and 3PTB). The two similarity alignments differ only by 0.3 Å rmsd, but qualitatively, there is clearly a difference. The volume-based alignment exhibits tighter congruence of atoms (and thus bonds as well), but the surface-based approach produces a slightly tilted orientation of AMC. The bottom portion of the molecules can be seen to be favoring a centered alignment based on surface considerations instead of leftward.

Figure 2.8 shows how the underlying functions that drive molecular similarity are related to the alignments of AMC to benzamidine. At left, the contour lines of the volume overlap function are shown (from the XY plane), which highlights the underlying structure of the atom-centered Gaussian functions, showing very close correspondence between the coincidence of AMC's atomic centers with those of benzamidine. At right, the respective surfaces are shown, with the observer points from the similarity computation in gray, and the differences between the distance to benzamidine's surface and that of AMC shown in yellow rods. Where possible, rods on one or the other side of the molecules tend to have similar length reflecting a balance in surface discrepancies that is averaged over all of the observer points. In the front of the display, because AMC is thicker than benzamidine, the rods

point outward from the observers, reflecting closer distances from the observers to AMC than to benzamidine. Points at which the surfaces are concordant show no yellow rods. The corresponding surface-density overlap contours are also shown, exhibiting a clear relationship to the underlying surface-density functions as seen in Figure 2.6.

This discussion has addressed only the molecular shape aspect of 3D molecular similarity, but, obviously, the degree to which the polar moieties of two molecules are congruent is important as well. Both volumetric approaches and surface-based approaches are augmented to capture electrostatic similarity. The volumetric formulation used by ROCS addresses this issue by defining atomic “colors” according to atom types, with steeper atom-centered Gaussians than those for shape and with flexibility as to weighting. The surface-based similarity approach of Surflex-Sim explicitly models hydrogen bond donors and acceptors, formal charges, and the directionality of polar interactions. Similarity is maximized according to a weighted function that combines pure shape and electrostatic considerations.

Figure 2.9 shows the optimal alignment of two competitive muscarinic antagonists using the full Surflex-Sim function, including both shape and electrostatics. The 2D structures are shown above the optimal mutual alignment, with the quinuclidine derivative shown in cyan carbons. At right, the individual molecules are shown in the same pose with atomic surfaces and rods that indicate surface areas that have

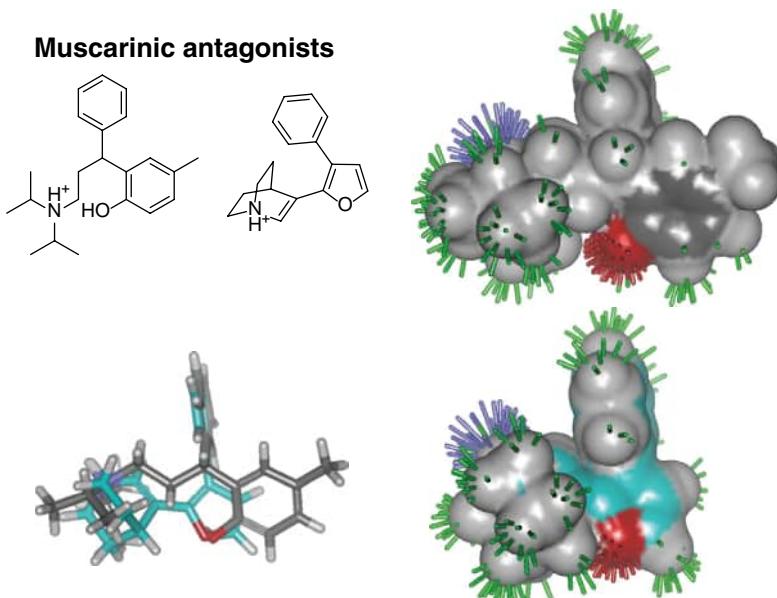


FIGURE 2.9 Optimal alignment of two muscarinic antagonists using surface shape and polarity. For color details, please see color plate section.

high similarity between the two. Green rods indicate high shape similarity, red indicate high similarity in the hydrogen-bond acceptor position and directional preference, and blue indicate concordance of the charged amine hydrogen atoms. This is an example where very high-3D similarity (0.82 on a scale of 0–1) obtains from molecules having different underlying scaffolds.

While the algorithmic complexity and computational burden of physical reality in the context of molecular similarity computation are nontrivial, the capabilities of such methods support the identification of truly surprising relationships between molecules. One particularly instructive example is that of methadone, which was synthesized as part of the World War II Nazi effort to develop a replacement for atropine. Atropine derived its utility as a nerve-gas antidote via muscarinic antagonism. Methadone was synthesized to mimic this antispasmodic effect, but it was serendipitously observed to have opiate agonist properties in a live animal assay [3]. It is possible, by both 2D and 3D molecular similarity, to identify a significant relationship between methadone and muscarinic antagonists. However, the 2D inference is supported only by similarity to muscarinic antagonists that were synthesized after methadone was made. Using physically realistic 3D molecular similarity, the relationship of methadone to morphinan-based opiate agonists can be reliably identified [3].

Because most small molecules have been synthesized as part of an explicit human design process, the effects of that process can be seen in the structures of the molecules themselves. An apparently successful molecular similarity approach may succeed only because it is able to detect aspects of intellectual ancestry from molecular structure. We have suggested two validation strategies to avoid such problems [4]. The first is temporal partitioning, where the molecules used as “knowns” have been identified *before* the molecules to be used as “unknowns.” Using such a partitioning, the methadone/muscarinic effect would be undetectable using 2D similarity. The second is “intellectual” partitioning, where the molecules used as “knowns” will have been designed intentionally to have a particular activity, and the “unknowns” will have been designed for a different activity. An intellectual partitioning would have correctly identified the question of methadone’s opiate activity as being interesting in light of the set of natural and semisynthetic morphinan opioid ligands. This question, however, can be answered only with a sophisticated 3D approach, not a 2D one.

There are two broad points concerning methods for computing molecular similarity. First, physically realistic methods for computing similarity exist and are practical for large-scale application. Second, without very careful attention paid to validation protocols and benchmarks, it is easy to demonstrate “success” for methods that, in practice, will yield little in the way of nonobvious discoveries.

2.2.3 3D QSAR: Physically Realistic Activity Prediction

As discussed earlier, methods that make use of protein structures have a natural bias toward physical realism, and molecular similarity methods exist that treat molecules as three-dimensional flexible objects. It is possible to combine the two concepts to approach the QSAR problem, as has been done with the Surflex-QMOD method

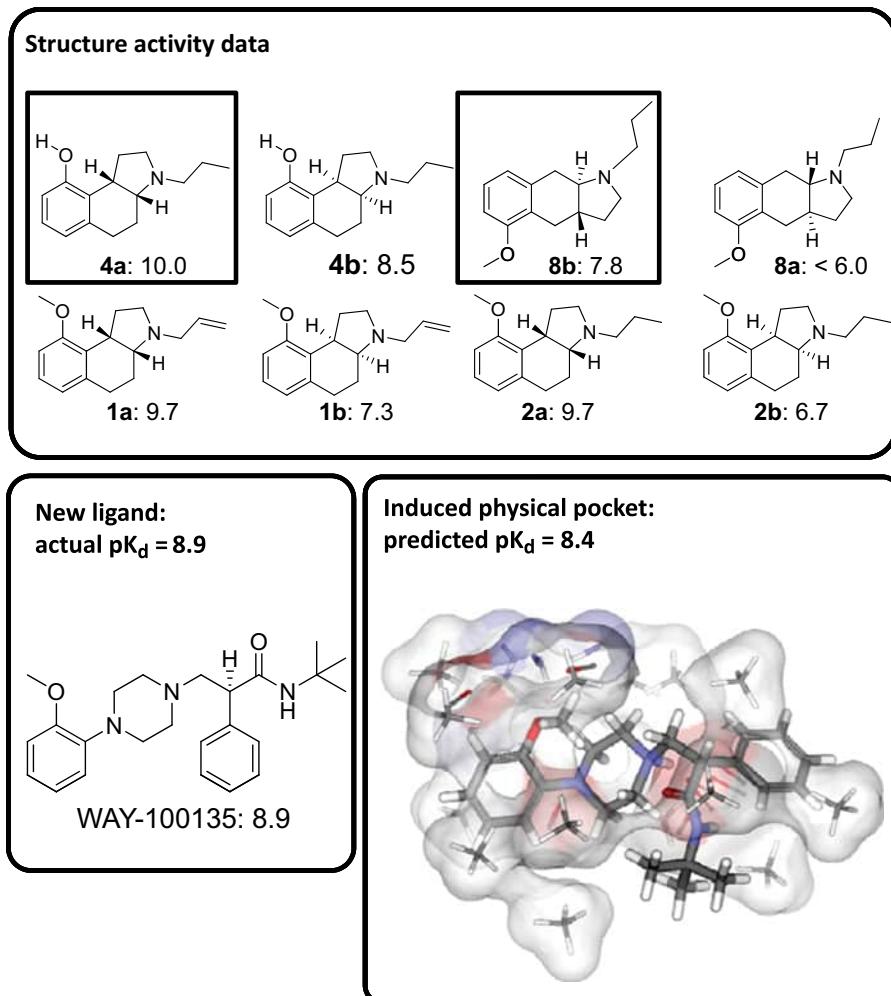


FIGURE 2.10 It is possible to construct a physically realistic model given structure-activity data. Such models can make accurate predictions even when the subject ligand is very different from any ligand on which the model was constructed. For color details, please see color plate section.

[8, 35]. In that approach, molecular similarity is used to generate hypotheses about the mutual alignment of a series of ligands with known potencies, with each ligand having many possible poses. From this initial alignment, molecular probes that are surrogates for a protein binding pocket are placed so as to make favorable interactions with at least one pose of one molecule. A complex machine-learning algorithm is employed that identifies a subset of these probes, whose fine positions are further

optimized. The score of a particular pose of a particular ligand is simply that produced by a scoring function for molecular docking (a series of intermolecular interactions along with explicit entropic fixation terms). The final “pocketmol” has the property that score of the optimal pose of each training ligand is close to the known potency. The optimal pose is that results from all-atom Cartesian optimization of the docking scoring function for all of the poses of a ligand. The learning approach is technically challenging, requiring multiple-instance machine-learning (additional details on multiple-instance learning for activity prediction and for scoring function development are available [8, 15, 32, 33, 35–39]. The key point here, however, is that the resulting models satisfy all of the aspects of the central dogma of physical reality.

Figure 2.10 shows the resulting model on a challenging example of 5-HT1a activity prediction [4], in which a careful intellectual partitioning of training and testing data was made, with 20 molecules from a single company’s effort including just two scaffolds used for training and 32 molecules with diverse scaffolds used for testing [35]. At right, the final pocketmol is shown along with the predicted bound pose of a potent aryl-piperazine 5-HT1a ligand. Despite sharing so little topological similarity, the predicted potency ($pK_d=8.4$) was very close to the experimental value ($pK_d=8.9$). Furthermore, the predicted pose, which results from “docking” into the pocketmol, matches with what has been established over many years of SAR investigation of this receptor (see the original publication for more discussion [35]).

While the QMOD method is complex, the complexity was driven by the requirements of physical reality. QSAR models of binding affinity *must* have an inter-dependence between model and molecular pose. There must also be an inter-dependence between structural variation and predicted bound pose. The models, if they are at all physically sensible, must also encapsulate nonlinear effects such as size exclusion (as in Figure 2.3). Last, ligands with very different underlying scaffolds, as in Figure 2.10, must be modeled in a manner that can recognize their capability to bind the same pocket with similar affinities.

2.3 SUMMARY

Creativity in drug design brings new scaffolds to indications with existing therapies and brings first-in-class drugs to clinical application. Such creativity, while offering the potential for substantial novel patient benefits, also creates risk. In order to support rational decision making in a drug-design endeavor that goes *beyond* purely incremental engineering, computational modeling methods must address what we have called the central dogma of physical reality that governs protein–ligand interactions: [1] molecular conformation and alignment matter; [2] structural variations, even on a common scaffold, affect ligand pose; [3] combinations of structural variations will affect binding affinity in a manner that is not strictly additive due to the interplay between structural changes and the consequent configurational adaptations; and [4] ligands with molecular scaffolds having little or no topological relationship can bind the same site. It is very easy to construct validation protocols to support the

notion that nonphysical methods work, but such methods are likely to have only a very limited window of predictive accuracy.

Respect for physical reality in methodological development, coupled with validation strategies that avoid the problems of bias in our molecular and biological data, will lead to improvements in tools for the computational support of drug design. Many of the most challenging problems in drug design absolutely require predictive modeling that goes beyond incrementalism.

REFERENCES

1. Cleves AE, Jain AN. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J Comput Aided Mol Des* 2008;22(3–4):147–159.
2. Cleves AE, Jain AN. Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem* 2006;49(10):2921–2938.
3. Yera ER, Cleves AE, Jain AN. Chemical structural novelty: On-targets and off-targets. *J Med Chem* 2011;54(19):6771–6785.
4. Jain AN, Cleves AE. Does your model weigh the same as a duck? *J Comput Aided Mol Des* 2012;26:57–67.
5. Rosenblatt F. A comparison of several perceptron models. *Self-Organizing Systems*. Washington, DC: Spartan Books; 1962, p 463–484.
6. Minsky M, Papert S. *Perceptrons*. Cambridge: MIT Press; 1969.
7. Rumelhart DE, McClelland JL. *Parallel Distributed Processing*. Cambridge: MIT Press; 1988.
8. Jain AN. QMOD: Physically meaningful QSAR. *J Comput Aided Mol Des* 2010; 24(10):865–878.
9. Kuntz ID, Blaney JM, Oatley SJ, et al. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 1982;161(2):269–288.
10. Olson AJ, Goodsell DS. Automated docking and the search for hiv protease inhibitors. *SAR QSAR Environ Res* 1998;8(3–4):273–285.
11. Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. *Proteins Struct Funct Bioinform* 1990;8(3):195–202.
12. Jones G, Willett P, Glen RC. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comput Aided Mol Des* 1995;9(6):532–549.
13. Jones G, Willett P, Glen RC, et al. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 1997;267(3):727–748.
14. Welch W, Ruppert J, Jain AN. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol* 1996;3(6):449–462.
15. Jain AN. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des* 1996; 10(5):427–440.
16. Ruppert J, Welch W, Jain AN. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci* 1997;6(3):524–533.

17. Rarey M, Kramer B, Lengauer T, et al. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 1996;261(3):470–489.
18. Rarey M, Kramer B, Lengauer T. Multiple automatic base selection: Protein–ligand docking based on incremental construction without manual intervention. *J Comput Aided Mol Des* 1997;11(4):369–384.
19. Bissantz C, Folkers G, Rogan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* 2000;43(25):4759–4767.
20. Jain AN. Surfflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* 2003;46(4):499–511.
21. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins Struct Funct Bioinform* 2004;56(2):235–249.
22. Warren GL, Andrews CW, Capelli AM, et al. A critical assessment of docking programs and scoring functions. *J Med Chem* 2006;49(20):5912–5931.
23. Sutherland JJ, Nandigam RK, Erickson JA, et al. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J Chem Inf Model* 2007;47(6):2293–2302.
24. Verdonk ML, Mortenson PN, Hall RJ, et al. Protein-ligand docking against non-native protein conformers. *J Chem Inf Model* 2008;48(11):2214–2225.
25. Spitzer R, Jain AN. Surfflex-dock: Docking benchmarks and real-world application. *J Comput Aided Mol Design* 2012;26(6):687–699.
26. Jain AN. Bias, reporting, and sharing: Computational evaluations of docking methods. *J Comput Aided Mol Des* 2008;22(3–4):201–212.
27. Jain AN. Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. *J Comput Aided Mol Des* 2009;23(6):355–374.
28. Hopfinger AJ. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J Am Chem Soc* 1980;102(24):7196–7206.
29. Rush TS, Grant JA, Mosyak L, et al. 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.
30. Masek BB, Merchant A, Matthew JB. Molecular skins: A new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins Struct Funct Bioinform* 1993;17(2):193–202.
31. Jain AN. Morphological similarity: A 3d molecular similarity method correlated with protein-ligand recognition. *J Comput Aided Mol Des* 2000;14(2):199–213.
32. Jain AN, Dietterich TG, Lathrop RH, et al. A shape-based machine learning tool for drug design. *J Comput Aided Mol Des* 1994;8(6):635–652.
33. Jain AN, Harris NL, Park JY. Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the 5-HT1a receptor. *J Med Chem* 1995;38(8):1295–1308.
34. Jain AN. Ligand-based structural hypotheses for virtual screening. *J Med Chem* 2004;47(4):947–961.
35. Langham JJ, Cleves AE, Spitzer R, Kirshner D, Jain AN. Physical binding pocket induction for affinity prediction. *J Med Chem* 2009;52(19):6107–6125.

36. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell* 1997;89(1–2):31–71.
37. Jain AN, Koile K, Chapman D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J Med Chem* 1994;37(15):2315–2327.
38. Pham TA, Jain AN. Parameter estimation for scoring protein-ligand interactions using negative training data. *J Med Chem* 2006;49(20):5856–5868.
39. Pham TA, Jain AN. Customizing scoring functions for docking. *J Comput Aided Mol Des* 2008;22(5):269–286.

CHAPTER 3

A ROUGH SET THEORY APPROACH TO THE ANALYSIS OF GENE EXPRESSION PROFILES

JOACHIM PETIT, NATHALIE MEURICE, JOSÉ LUIS MEDINA-FRANCO,
and GERALD M. MAGGIORA

3.1 INTRODUCTION

An important aspect of functional genomics research involves the identification of relationships between the expression levels of certain genes and specific biological responses. In this regard, a number of studies have been carried out to assess the underlying genetic basis of cellular responses induced by the administration of various chemicals [1–5]. The data from such studies can be represented in the form of a decision table (DT), but because the number of genes typically considered tends to be quite large, it may be difficult to identify minimal subsets of genes responsible for a given cellular response or biological function. For a number of reasons, it is desirable to identify small subsets of genes associated with biological endpoints of interest.

Rough set theory (RST), first introduced by Pawlak more than 20 years ago [6], is a set-based method that is well suited for dealing with a wide variety of discrete data that can be represented by DTs [7–11]. RST appears to be well suited to the problem at hand since, as will be seen in the sequel, it provides computationally tractable procedures for identifying minimal subsets of attributes, that is, genes, in this work [12, 13]. These subsets of attributes maintain key relationships in the data that support the generation of simple linguistic rules. This procedure is similar in many ways to dimensionality reduction techniques that are routinely employed to continuous variables to reduce the size and complexity of mathematical models [14].

Rough set theory differs from most other rule-based methods, especially those associated with artificial intelligence methods. In such methods, rules (typically called “production rules”) that encapsulate knowledge are *input* into the system and

are then used as a basis for deducing new relationships. In contrast, the primary emphasis of RST is identifying the underlying irreducible patterns in the data and transforming the information in these patterns into knowledge in the form of linguistic rules. Many other rule-based methods of various forms have been developed including decision trees [15, 16], Bayesian networks [17], and the analytic hierarchy process [18] to name a few. The book by Bender [19] also provides an excellent discussion of many types of rule-based systems.

It is important to point out that the RST-based rules obtained in this work are best considered to be *descriptive* not *predictive* rules. Descriptive rules are obtained by converting the available data into a set of IF-THEN rules. In this way, data is transformed into knowledge. Whether rules obtained in this manner are predictive is, however, quite another matter since the development of predictive rules requires independent data with which to validate the rules generated from the original dataset. This is not possible in the present work as the amount of data available is insufficient. However, measures do exist for assessing the “strength” of descriptive rules as discussed in Sections 3.2.2–3.2.5. While these are not equivalent to the type of validation required of predictive rules, they, nonetheless, provide a measure of the degree of confidence one can have that a given rule will be applicable to new data.

The present paper, which is based on the work of Sawada et al. [20], is aimed at evaluating the suitability of RST for analyzing the association of drug-induced changes in gene expression profiles with the presence or absence of the cellular pathology, phospholipidosis, in human hepatoma HepG2 cells. Phospholipidosis, a storage disorder that leads to an accumulation of phospholipids in lysosomes, is induced by a wide range of drugs [21–23] and, consequently, represents a potential hazard in drug development. Thus, it is expected that genes associated with biological processes such as lipid metabolism, transport, and proteolysis might well be involved (*vide infra*).

In this work, the rows of the DT correspond to specific drugs that induce changes in the gene-expression levels of a set of 17 genes selected by Sawada et al. [20] as potentially associated with drug-induced phospholipidosis. The columns correspond to the attributes of which there are two classes, called *condition* and *decision* attributes in RST that characterize the drugs. The condition attributes are associated with drug-induced changes in gene-expression levels and the decision attribute is associated with the cellular response to the administration of these drugs, namely, whether or not phospholipidosis occurs. The relationship of the condition attributes to the decision attributes can be captured in the form of *linguistic rules* that facilitate the description of the results of RST analyses to researchers, especially those not necessarily expert in the methodology.

In addition, as discussed, RST provides a rigorous framework for eliminating superfluous condition attributes that provide irrelevant information. This leads to a simplification of the rules since the number of condition attributes is significantly reduced, a feature of the RST method that accords well with the principle of Occam’s razor [24].

There are a number of examples of RST applications in the literature, especially in the field of medical informatics; such applications have been reviewed by Fakih and Das [25]. In addition to applications in medical diagnosis, RST has been applied

in other research fields as varied as the prediction of reaction rate constants [26], pharmacokinetics and toxicology [27], the analysis of eco-toxicological data [28, 29], and protein structural class prediction [30]. Nevertheless, the application of RST in drug discovery and bioinformatics remains a relatively open field, with comparatively few published examples. Among these, Krysiński et al. reported quantitative structure–activity relationship (SAR) studies of antielectrostatic imidazolium compounds [31], while Komorowski and coworkers have focused on a number of functional genomics applications [12, 13, 32, 33].

The basics of the RST methodology are presented in Sections 3.2.1 and 3.2.2. Because this methodology is unfamiliar, a more extensive description is provided. A simple example that is relevant to the main interest of this paper is presented in Section 3.2.3 followed by a discussion in Section 3.2.4 of the removal of superfluous information, which leads to a description of the generation of simple linguistic rules in Section 3.2.5. Analysis of the dataset describing gene-expression levels associated with drug-induced phospholipidosis of HepG2 cells is presented in Section 3.3. Since RST only deals with discretized data, a description of the discretization procedure is provided. Special emphasis is given to the linguistic rules that emerge from the analysis, with the goal of identifying a small set of simple rules that express the strongest relationships existing between specific up and down regulated genes associated with drug-induced phospholipidosis. Section 3.4 discusses the relationship of phospholipidosis to the biological processes associated with these genes. The results obtained in this study indicate that RST may be a promising tool for analyzing data produced in functional genomics experiments.

3.2 METHODOLOGY

The basics of the methodology are outlined in Sections 3.2.1 and 3.2.2. A simple example that illustrates the method is presented in Sections 3.2.3–3.2.5. Walczak and Massart [34] provide an excellent tutorial that is considerably more detailed than the account of the methodology given in this work. More extensive tutorials are given by Komorowski, Polkowski, Pawlak, and Skowron [35, 36].

3.2.1 Basic Theory

RST provides an effective means for dealing with data that can be represented in the form of a DT. The rows of the DT correspond to the set of objects, X , called the *Universe of Discourse*,

$$X = \{x_1, x_2, \dots, x_N\} \quad (3.2.1)$$

The columns correspond to the set of features or attributes, A , that characterize the objects in X . The set of attributes given in Equation 3.2.2

$$A = \{a_1, a_2, \dots, a_M\} \quad (3.2.2)$$

is typically made up of a subset condition attributes

$$C = \{c_1, c_2, \dots, c_K\} \quad (3.2.3)$$

and a subset of decision attributes

$$D = \{d_1, d_2, \dots, d_{M-K}\} \quad (3.2.4)$$

so that,

$$A = C \cup D. \quad (3.2.5)$$

Each attribute has an associated set of discrete values called its “value set”

$$V_a = \{v_1^a, v_2^a, \dots, v_p^a\}, \quad \text{where } a \in C \text{ or } D \quad (3.2.6)$$

that can be ordinal (e.g., 1, 2, ..., 1.25, 1.50, ..., etc.) or nominal (e.g., “yes” and “no,” “red,” “green,” and “blue,” ..., etc.) but not continuous; attributes with continuous values must be discretized. The family of value sets is given as $V = \{V_{a_1}, V_{a_2}, \dots, V_{a_M}\}$. Since attribute values can be either naturally ordered, such as the set of integers, or unordered, such as the set of labels red, green, blue, etc., the RST approach has considerable descriptive power. Lastly, the *ordered pair* (a, v_k^a) is called a *descriptor* in an information system, which is defined by the 3-tuple $IS = \langle X, A, V \rangle$, where each of the terms is defined earlier.

Any given subset A_k of condition or decision attributes, that is, $A_k \subseteq C$ or $A_k \subseteq D$, can induce a *partition* of X such that

$$X \xrightarrow{A_k} X(A_k) = \{X_1^{A_k}, X_2^{A_k}, \dots, X_Q^{A_k}\} \quad (3.2.7)$$

where $X_\ell^{A_k}$ is a subset of X induced by A_k . Since $X(A_k)$ is a partition of X ,

$$X = X_1^{A_k} \cup X_2^{A_k} \cup \dots \cup X_Q^{A_k} \quad (3.2.8)$$

and

$$X_\ell^{A_k} \cap X_m^{A_k} = \emptyset \quad \text{for all } \ell, m \quad (3.2.9)$$

Thus, the subsets of $X(A_k)$ correspond to *equivalence classes*, which in RST are called *indiscernibility classes* or *elementary sets*. Regardless of the nomenclature used, all of the elements within a given indiscernibility class are equivalent to or are indiscernible from each other with respect to the subset of condition or decision attributes being considered. The set of indiscernibility classes induced by a given subset of condition or decision attributes constitute the respective C - and D -spaces, of the problem. *One of the powers of RST is that it provides mathematical procedures for eliminating*

attributes that do not change the indiscernibility classes and, thus, are superfluous or redundant, greatly simplifying an RST analysis (see Sections 3.2.4 and 3.3.2).

Since the details of this topic are quite complex, they will be illustrated by the simple example provided in Section 3.2.3. As will be seen, the indiscernibility classes in C -space form a basis for representing the information contained in the indiscernibility classes in D -space. Importantly, these relationships can be described in terms of *linguistic rules*, which significantly enhance communication of the results to scientists who are not experts in RST, as illustrated by the example given in Section 3.2.5.

An important concept in RST compared to classic set theory is the notion of an *approximation set*. In RST, a set is approximated by two sets that form *lower* and *upper approximations* of the set. These approximations are given, respectively, by

$$\underline{A}_k S = \left\{ x \in X \mid X_{\ell}^{A_k} \subseteq S, \quad \ell = 1, 2, \dots, Q \right\} \quad (3.2.10)$$

and

$$\bar{A}_k S = \left\{ x \in X \mid X_{\ell}^{A_k} \cap S \neq \emptyset, \quad \ell = 1, 2, \dots, Q \right\} \quad (3.2.11)$$

where $S \subseteq X$. The region that lies between the lower and upper approximate sets, called the *boundary region*, is given by the difference set

$$B_{A_k}(S) = \bar{A}_k S - \underline{A}_k S. \quad (3.2.12)$$

Thus, Equation 3.2.12 can be rearranged as

$$\bar{A}_k S = \underline{A}_k S \cup B_{A_k}(S). \quad (3.2.13)$$

A simple graphical portrayal of lower and upper approximation sets and the boundary region between these two sets is given in Figure 3.1. As is seen in the figure, the Universe of Discourse X is partitioned into 10 indiscernibility classes X_i^A , $i = 1, 2, \dots, 10$, induced by the set of attributes A_k . The set S being approximated is indicated by the heavy dashed line. The three, light-gray shaded indiscernibility classes in the center of the diagram give the lower approximation of S , that is, $\underline{A}_k S = X_7^A \cup X_8^A \cup X_9^A$. The upper approximation of S is given by all of the indiscernibility classes bounded by the solid line and thus includes $\bar{A}_k S$, that is, $\bar{A}_k S = \underline{A}_k S \cup (X_4^A \cup X_5^A \cup X_6^A \cup X_{10}^A)$. Thus, the upper approximation includes the four indiscernibility classes shaded in dark gray in addition to the indiscernibility classes contained in lower approximation. The boundary set, which is equal to $B_{A_k}(S) = \bar{A}_k S - \underline{A}_k S = X_4^A \cup X_5^A \cup X_6^A \cup X_{10}^A$ (see Eq. 3.2.12), corresponds to the four indiscernibility classes colored dark gray in Figure 3.1. As is clear from the figure, all $x_i \in \underline{A}_k S$ also satisfy $x_i \in S$, and thus can be classified by the three indiscernibility classes of the lower approximation with complete certainty. In contrast, all $x_i \in \bar{A}_k S$ do not satisfy $x_i \in S$ and hence cannot be classified by the seven indiscernibility classes of the upper approximation with complete certainty. Thus, the indiscernibility classes located in the boundary region (shaded in dark gray in Figure 3.1) represent the uncertainty of any attempt to classify the objects of S in terms of

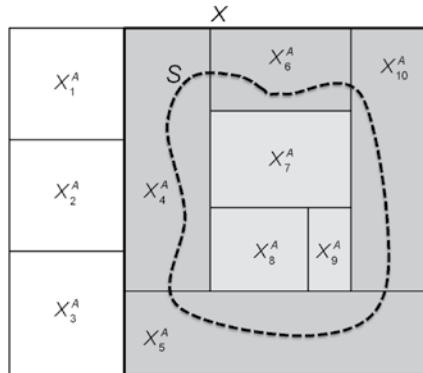


FIGURE 3.1 The Universe of Discourse X is partitioned into 10 indiscernibility classes by the set of attributes A_k . See Section 3.2.1 of the text for further discussion, three of which lie within the set S being approximated.

the indiscernibility classes of $\bar{A}_k S$. Union of the three indiscernibility classes X_1^A , X_2^A , and X_3^A constitutes the complement of the upper approximation of S , that is, $X_1^A \cup X_2^A \cup X_3^A = (\bar{A}_k S)^c$ where it is certain that $x_i \in S$ are not found.

3.2.2 Measures of Classification Accuracy and Quality

There are several measures that are related to the accuracy of the classification of a given subset S with respect to the partition induced by a subset of attributes A_k [6, 11, 34]. The most universally applied measure is the *accuracy* by which S can be approximated by the set of indiscernibility classes induced by the subset of attributes A_k ,

$$\text{Acc}(S)_{A_k} = \frac{\text{Card}(\underline{A}_k S)}{\text{Card}(\bar{A}_k S)}, \quad (3.2.14)$$

where $\text{Card}(\underline{A}_k S)$ and $\text{Card}(\bar{A}_k S)$ are the number of objects (elements) in $\underline{A}_k S$ and $\bar{A}_k S$, respectively.

More generally, consider a classification of X into Q nonintersecting subsets whose union equals X as given in Equation 3.2.8. Thus, $X(A_k)$ is a *partition* of X , which gives rise to the respective families of lower and upper approximation sets of the classification, namely,

$$\underline{A}_k X(A_k) = \left\{ \underline{A}_k X_1^{A_k}, \underline{A}_k X_2^{A_k}, \dots, \underline{A}_k X_Q^{A_k} \right\} \quad (3.2.15)$$

and

$$\bar{A}_k X(A_k) = \left\{ \bar{A}_k X_1^{A_k}, \bar{A}_k X_2^{A_k}, \dots, \bar{A}_k X_Q^{A_k} \right\}. \quad (3.2.16)$$

Now, in analogy to Equation 3.2.14, the *accuracy of classification of the partition of X induced by A_k* is given by

$$\text{Acc}[X(A_k)] = \frac{\sum_{L=1}^Q \text{Card}(\underline{A}_k X_L^{A_k})}{\sum_{L=1}^Q \text{Card}(\bar{A}_k X_L^{A_k})} \quad (3.2.17)$$

which is the ratio of the size of the family of lower-approximation sets to the size of the family of upper-approximation sets. As the family of lower and upper approximations approaches equality, that is, as $\underline{A}_k X_L^{A_k} \approx \bar{A}_k X_L^{A_k}$ for all L , the family of sets constituting the boundary region approaches the null set, that is, as $B_{A_k}(X_L^{A_k}) \rightarrow \emptyset$ for all L . When this is achieved, $\text{Acc}[X(A_k)] \rightarrow 1$. In contrast, as $\underline{A}_k X_L^{A_k} \rightarrow \emptyset$ and $\bar{A}_k X_L^{A_k} \rightarrow X$ for all L , $\text{Acc}[X(A_k)] \rightarrow 0$. Thus, $0 \leq \text{Acc}[X(A_k)] \leq 1$.

Another useful classification measure is the *quality of classification*, which is given by

$$\text{Qual}[X(A_k)] = \frac{\sum_{L=1}^Q \text{Card}(\underline{A}_k X_L^{A_k})}{\text{Card}(X)}. \quad (3.2.18)$$

This represents the ratio of all objects classified with certainty to the total number of objects in X .

3.2.3 An Illustrative Example

Consider the DT given in Table 3.1, which is intended to provide a simplified illustrative example of the chemically induced gene expression data that will be presented in much greater detail in Section 3.3 for a real dataset. In the current example, the Universe of Discourse is the set of eight hypothetical molecules under study, $Mol = \{1,2,3,4,5,6,7,8\}$, which make up the rows of the DT. The set of condition attributes are associated with three genes $G = \{\text{gene-1}, \text{gene-2}, \text{gene-3}\}$, whose expression levels may be affected by the molecules in Mol. The nominal values in the table are associated with whether a given gene is up regulated (\uparrow), down regulated (\downarrow), or normally-expressed (\blacksquare) due to the presence of a given molecule. The single decision attribute $D = \{p\}$ corresponds to whether a given pathology is observed (\oplus) or not observed (\ominus).

Examination of Table 3.2 shows that the condition attributes partition the set of compounds into five indiscernibility classes or, more specifically, C -indiscernibility classes,

$$X(G) = \{X_1^G, X_2^G, X_3^G, X_4^G, X_5^G\} \quad (3.2.19)$$

with the following members

$$\begin{aligned} X_1^G &= \{1,6\} \\ X_2^G &= \{2,5,8\} \\ X_3^G &= \{3\} \\ X_4^G &= \{4\} \\ X_5^G &= \{7\} \end{aligned} \quad (3.2.20)$$

TABLE 3.1 A Simplified DT Illustrating the Relationship of the Gene-Expression Profiles of Three Genes to the Presence or Absence of Phospholipidosis Induced by Eight Hypothetical Molecules^a

Mol	gene-1	gene-2	gene-3	p
1	↑	■	■	+
2	■	↓	↑	○
3	■	■	↑	+
4	↑	↑	↑	+
5	■	↓	↑	○
6	↑	■	■	○
7	■	↑	↑	+
8	■	↓	↑	○

^aUp regulated genes are indicated by a upwards arrow (↑), down regulated genes by a downward arrow (↓), and normally expressed genes by a filled square (■). Phospholipidosis p is indicated by a plus (+) if it occurs and an open circle (○) if it does not occur.

TABLE 3.2 The DT Derived from the Data Given in Table 3.1 with the Five Indiscernibility Classes X_1^G , X_2^G , X_3^G , X_4^G , and X_5^G Induced by the Set of Condition Attributes $G = \{gene\text{-}1, gene\text{-}2, gene\text{-}3\}$.

Class	Mol	gene-1	gene-2	gene-3	p
X_1^G	1	↑	■	■	+
	6	↑	■	■	○
X_2^G	2	■	↓	↑	○
	5	■	↓	↑	○
X_3^G	8	■	↓	↑	○
	3	■	■	↑	+
X_4^G	4	↑	↑	↑	+
X_5^G	7	■	↑	↑	+

Note that the column of the table corresponding to the decision attribute p is grayed out since its values are not considered in the determination of the indiscernibility classes induced by the conditional attributes (see text and Eq. 3.2.19 for further details). Up regulated genes are indicated by an upward arrow (↑), down regulated genes by a downward arrow (↓), and normally expressed genes by a filled square (■). Phospholipidosis p is indicated by a plus (+) if it occurs and an open circle (○) if it does not occur.

The value of each of the three gene expression levels is identical for all elements within a given indiscernibility class. For example, consider the indiscernibility class X_1^G , which is made up of molecules 1 and 6. Table 3.2 shows that *gene-1* is up regulated (expression-level=↑), while *gene-2* and *gene-3* are normally expressed (expression-level=■) for both molecules. Interestingly, the two molecules behave differently

TABLE 3.3 The DT Obtained from the Data Given in Table 3.1 with the Two Indiscernibility Classes X_1^D and X_2^D Induced by the Decision Attribute Phospholipidosis p

Class	<i>Mol</i>	<i>gene-1</i>	<i>gene-2</i>	<i>gene-3</i>	<i>p</i>
X_1^D	1	↑	■	■	+
	3	■	■	↑	+
	4	↑	↑	↑	+
	7	■	↑	↑	+
X_2^D	2	■	↓	↑	○
	5	■	↓	↑	○
	6	↑	■	■	○
	8	■	↓	↑	○

Note that the columns of the table corresponding to the condition attributes *gene-1*, *gene-2*, and *gene-3* are grayed out since their values are not considered in the determination of the indiscernibility classes (see text and Eq. 3.2.21 for further details). Up regulated genes are indicated by an upward arrow (\uparrow), down regulated genes by a downward arrow (\downarrow), and normally expressed genes by a filled square (■). Phospholipidosis *p* is indicated by a plus (+) if it occurs and an open circle (○) if it does not occur.

with respect to the decision attribute since **1** induces phospholipidosis while **6** does not, a situation that is discussed further in Section 3.2.5 on rule generation.

As shown in Table 3.3, the decision attribute partitions the set of compounds into two indiscernibility classes, called *D*-indiscernibility classes,

$$X(D) = \{X_1^D, X_2^D\} \quad (3.2.21)$$

with the following members

$$\begin{aligned} X_1^D &= \{1, 3, 4, 7\} \\ X_2^D &= \{2, 5, 6, 8\}. \end{aligned} \quad (3.2.22)$$

It is clear from Figure 3.2 that X_3^G , X_4^G , and X_5^G are subsets of X_1^D and thus their union constitutes the lower approximation of X_1^D , that is,

$$\underline{GX}_1^D = X_3^G \cup X_4^G \cup X_5^G = \{3, 4, 7\}. \quad (3.2.23)$$

Likewise, X_2^G is a subset of X_2^D and thus it constitutes the lower approximation of X_2^D , that is,

$$\underline{GX}_2^D = X_2^G = \{2, 5, 8\}. \quad (3.2.24)$$

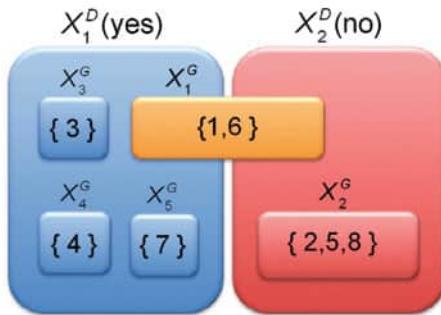


FIGURE 3.2 Schematic depiction of the indiscernibility classes of the example presented in Sections 3.2.3–3.2.5. The objects in the rectangles shaded in blue correspond to molecules that definitely induce phospholipidosis, while those rectangles shaded in blue correspond to molecules that definitely do not induce phospholipidosis. The two molecules within the orange rectangle correspond to molecules in the boundary set that may or may not induce phospholipidosis. Rules involving molecules within the indiscernibility classes depicted by the red and blue-shaded rectangles are deterministic, while those within the yellow-shaded rectangle are nondeterministic (probabilistic). For color details, please see color plate section.

The upper approximations to X_1^D and X_2^D are given, respectively, by

$$\bar{G}X_1^D = X_1^G \cup X_3^G \cup X_4^G \cup X_5^G = \{1,3,4,6,7\} \quad (3.2.25)$$

and

$$\bar{G}X_2^D = X_1^G \cup X_2^G = \{1,2,5,6,8\}. \quad (3.2.26)$$

Thus, in both cases the boundary sets are identical as shown in Equations 3.2.27 and 3.2.28.

$$\begin{aligned} B_G(X_1^D) &= \bar{G}X_1^D - \underline{G}X_1^D \\ &= \{1,3,4,6,7\} - \{3,4,7\} \\ &= \{1,6\} \\ &= X_1^G \end{aligned} \quad (3.2.27)$$

and

$$\begin{aligned} B_G(X_2^D) &= \bar{G}X_2^D - \underline{G}X_2^D \\ &= \{1,2,5,6,8\} - \{2,5,8\}, \\ &= \{1,6\} \\ &= X_1^G \end{aligned} \quad (3.2.28)$$

The boundary set is indicated by the orange rectangle in Figure 3.2.

The *accuracy of classification* given by Equation 3.2.17 becomes in this case

$$\text{Acc}[X(D)] = \frac{\text{Card}(\underline{GX}_1^D) + \text{Card}(\underline{GX}_2^D)}{\text{Card}(\underline{GX}_1^D) + \text{Card}(\underline{GX}_2^D)} = \frac{3+3}{5+5} = 0.60 \quad (3.2.29)$$

In contrast, the *quality of classification*, as given in Equation 3.2.18, is

$$\text{Qual}[X(D)] = \frac{\text{Card}(\underline{GX}_1^D) + \text{Card}(\underline{GX}_2^D)}{\text{Card}(U)} = \frac{3+3}{8} = 0.75 \quad (3.2.30)$$

This means that 75% of the drug-induced gene-expression profiles are correctly classified.

Other features of RST important in applications will be discussed in the sequel.

3.2.4 Essential and Superfluous Information

Although the example described in Section 3.2.3 is a relatively simple one, realistic examples such as the one that is the main focus of the current work are much more complex and thus may contain redundant information whose removal does not affect the overall performance of a given model to a significant extent. Specifically, redundancies exist among attributes and attribute values, which can be removed without changing the partitioning of the objects in the Universe of Discourse (*vide infra*). In order to clarify these issues as simply as possible, the following explanations are given in terms of the example presented in Section 3.2.3. More detailed discussions of the mathematical and algorithmic issues can be found in references [6, 11, 34].

3.2.4.1 Superfluous Attributes The question arises as to whether all of the condition or decision attributes are needed to obtain the correct partitioning of the objects described by a given DT into their appropriate indiscernibility classes. As will be seen in the sequel, some attributes, called *dispensable* or *superfluous attributes*, can be removed without affecting the distribution of objects among the indiscernibility classes; thus, the quality of classification is preserved. This is an important feature of RST that can lead to significant simplifications. Attributes that cannot be removed without affecting the partitioning of the objects in the indiscernibility classes are called *indispensable attributes*. Removing such attributes changes the indiscernibility classes in a way that reduces the quality of the classification.

A minimal set of indispensable condition attributes is called a *reduct*. Because reducts of condition attributes preserve the partitioning of C -space they are sometimes called C -reducts. Likewise, a subset of condition attributes that preserve the partitioning of the D -space, and thus the quality of classification, is called a *D-reduct* and is usually considered as sufficient to explain the initial variability of the D -space

as a function of the condition attributes. As will be seen in Section 3.3, D -reducts play an essential role in rule generation.

In the example presented in the previous section, removal of the third condition attribute, *gene-3*, does not affect the partitioning of the compounds in the indiscernibility classes. This can be seen by examining Table 3.2, which shows that the expression levels of the remaining two genes (*gene-1* and *gene-2*) are sufficient to maintain the original partitioning induced by all three genes.¹ Hence, *gene-3* can be considered as superfluous since its removal preserves the original C -space partitioning of the molecules into the five indiscernibility classes given in Equation 3.2.19. Thus, the set $\{\text{gene-1}, \text{gene-2}\}$ is a C -reduct. Not only is the original C -space partitioning preserved, but also the D -space partitioning, so that $\{\text{gene-1}, \text{gene-2}\}$ is also a D -reduct.

The question then arises as to whether there are other reducts. It is clear from Table 3.2 that if *gene-1* is removed the C -space partitioning is no longer preserved since $X_4^G, X_7^G \Rightarrow X_4^G \cup X_7^G$; D -space partitioning, however, is preserved. Thus, the set $\{\text{gene-2}, \text{gene-3}\}$ does not constitute a C -reduct, but does constitute a D -reduct. This case can also be seen in Figure 3.2, where it is clear that merging of X_4^G and X_7^G changes the partitioning of C -space but does not affect that of the corresponding D -space since $X_4^G \cup X_7^G \subset X_1^D$. Thus, the quality of classification is unchanged even though the C -space partitioning is no longer preserved.

Now consider the case where *gene-2* is removed. Examination of Table 3.2 shows that its removal leads to the combining of three indiscernibility classes: $X_2^G, X_3^G, X_5^G \Rightarrow X_2^G \cup X_3^G \cup X_5^G$, which is the new boundary set (cf. Figure 3.2). The lower approximation sets of X_1^D and X_2^D are changed significantly, lowering the quality of classification. Thus, the set of condition attributes $\{\text{gene-1}, \text{gene-3}\}$ is neither a C -reduct nor a D -reduct. The same holds true when any pair of condition attributes is removed.

3.2.4.2 Superfluous Attribute Values Consider the reduct $\{\text{gene-1}, \text{gene-2}\}$, which as discussed earlier is both a C -reduct and a D -reduct. In addition to removing *gene-3*, it is possible to achieve a further reduction of the DT by removing superfluous attribute values. It is seen that whenever *gene-2* is up regulated ($v_{\text{gene-2}} = \uparrow$) as is the case for *mol-4* and *mol-7*, phospholipidosis is observed (\oplus) regardless of the level of *gene-1* expression ($v_{\text{gene-1}} = \uparrow$ or $v_{\text{gene-1}} = \square$, respectively). As a consequence, the value of *gene-1* is superfluous and can be removed. This is indicated in Table 3.4 by the asterisk (*) placed to the right of the attribute values in the rows corresponding to *mol-4* and *mol-7* (shaded in light gray) associated with the condition attribute *gene-1*. Similarly, phospholipidosis is not observed (\bullet) whenever *gene-2* is down regulated (\downarrow), as is the case for *mol-2*, *mol-5*, and *mol-8*. This information is thus sufficient, and the corresponding value of *gene-1*, which is unchanged in all three cases, is superfluous, as is again indicated by asterisks (*) in the appropriate rows corresponding to the condition attribute *gene-1* (shaded in light gray). The D -reduct $\{\text{gene-2}, \text{gene-3}\}$ can be simplified in a similar manner, but is not done so because it does not provide any new insights into the methodology.

TABLE 3.4 Modified Version of Table 3.2 Explicitly Showing the Removal of the Superfluous Condition Attribute *gene-3* (Dark Gray Band) and the Superfluous Attribute Values of *gene-1*

Class	Mol	gene-1	gene-2	gene-3	p
X_1^G	1	↑	■	■	+
	6	↑	■	■	o
X_2^G	2	■*	↓	↑	o
	5	■*	↓	↑	o
	8	■*	↓	↑	o
	3	■	■	↑	+
X_4^G	4	↑*	↑	↑	+
X_5^G	7	■*	↑	↑	+

*Asterisks indicate superfluous attribute values in light-gray shaded panels. See Section 3.2.4 for additional discussion.

Although the process of identifying and removing superfluous condition attribute values is quite simple for the example presented here, which can be done “by hand,” this is not the case in general. The next works should be consulted for detailed discussion of this important issue [6, 11, 34–36]. As will be seen, in Section 3.2.5 on rule generation, the remaining attribute values in Table 3.4 form the basis for a set of linguistic rules that capture the information in the simplified data table.

3.2.5 Rule Generation

Now that all the superfluous information has been removed as shown in Table 3.4, what remains can be translated in *association rules* that represent the simplest possible rules that conserve the nonredundant information in the data. These rules are characterized by their *strength* as measured by the *support* for a given rule, which is based on the number of molecules in the dataset that support the rule. Since the example involves a very small (hypothetical) dataset, support for the rules will only be illustrative of the methodology. As will be seen in Section 3.3, a more meaningful analysis can be made on a real drug-induced phospholipidosis dataset that is focus of this work (*vide infra*). Although that dataset is also not overly large, it nonetheless represents data associated with an actual problem of interest.

The capability of RST to generate sets of rules that codify the underlying mathematical results into easy-to-understand linguistic rules enhances the transmission of information among scientists unfamiliar with the details and subtleties of RST. Although the rules obtained in the example discussed in Section 3.2.3 are quite simple, rule generation is generally a complex process, as will become clear by the analysis in Section 3.3. For the mathematically venturesome additional information can be found in the following references [37, 38].

TABLE 3.5 Rules Generated from the Example Presented in Sections 3.2.3–3.2.5

	Rules	Support
<i>Rule-1</i>	IF (<i>gene-2</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	3
<i>Rule-2</i>	IF (<i>gene-2</i> , \uparrow) THEN (<i>phospholipidosis</i> , $+$)	2
<i>Rule-3</i>	IF (<i>gene-1</i> , \uparrow) AND (<i>gene-2</i> , \blacksquare) THEN (<i>phospholipidosis</i> , $+$) OR (<i>phospholipidosis</i> , \bullet)	1,1
<i>Rule-4</i>	IF (<i>gene-1</i> , \blacksquare) AND (<i>gene-2</i> , \blacksquare) THEN (<i>phospholipidosis</i> , $+$)	1

Four rules are obtained from the example presented in Section 3.2.3 and the data given in Table 3.4, which is based on the removal of the superfluous attribute, *gene-3*, and the superfluous attribute values associated with *gene-1*. An “ordered-pair” notation is employed for compactness, where the first term corresponds to the condition or decision attribute and the second term corresponds to the value of the attribute with respect to a given indiscernibility class or object (molecule in this work) within the class. The value of the ordered pair is taken to be true and the expressions are interpreted as “IF-THEN” rules as shown in Table 3.5.²

As an example, *Rule-1* can be specified as an “IF-THEN” rule in the following way: “IF (*gene-2*, \downarrow) THEN (*phospholipidosis*, \bullet)”, which is interpreted as IF (*gene-2*, \downarrow) is true THEN (*phospholipidosis*, \bullet) is true, that is phospholipidosis does not occur.

In Table 3.5, the rules are sorted based on their level of support. As seen in the table, three and two instances, respectively, support the first two rules in the dataset. These rules indicate that phospholipidosis is not observed when *gene-2* is down regulated (*Rule-1*) but is observed when *gene-2* is up regulated (*Rule-2*). Both of these rules are *deterministic* since the result of applying them is unequivocal. *Rule-3*, on the other hand, is *nondeterministic* or *probabilistic* since the same set of condition attribute values lead to two different decision attribute values with equal likelihood. In this example, phospholipidosis can occur or not occur when *gene-1* is up regulated and *gene-2* is normally expressed. Finally, in *Rule-4* there is one instance where phospholipidosis is observed when both *gene-1* and *gene-2* are normally expressed, which is a strange result to say the least. Although *Rule-4* is deterministic, the result suggests that some other gene not in the set of condition attributes considered may be involved, but since this is only meant to be an illustrative example and is based on entirely hypothetical data, it is difficult to draw any definitive conclusions, especially since only a single example supports the rule.

Because the data presented in this example are quite limited, drawing any conclusions about the strength of the different rules is not meaningful. However, for the data analyzed in Sections 3.3.2 and 3.3.3, the strength of a given rule is an important consideration as to the generality and reliability of the rule. Rules with low levels of support do not provide consistent descriptions of new data outside of the original dataset.

3.3 DRUG-INDUCED GENE EXPRESSION AND PHOSPHOLIPIDOSIS IN HUMAN HEPATOMA HEPG2 CELLS

RST analysis is now applied to real data on the relationship of drug-induced gene expression to the occurrence of phospholipidosis in human hepatoma HepG2 cells based on the work of Sawada et al. [20]. The response of the HepG2 cells to the 30 drugs listed in Table 3.6 is characterized by the changes in expression levels of 17 genes and by whether or not the drugs induce phospholipidosis in the cells. Seventeen of the drugs, which fall into eight therapeutic classes, induce phospholipidosis, while the remaining 13 drugs, which fall into five therapeutic classes, do not. In the present work, however, all positive instances of phospholipidosis are considered to be of equal degree and phospholipidosis is taken to either occur or not occur (cf. Table 3.6). Interestingly, only the antipsychotic class of therapeutics contains molecules (Clozapine and Thioridazine) that induce phospholipidosis and a molecule (Haloperidol) that does not.

Figure 3.3 depicts the distribution of the 30 drugs in 2D chemical space. The red filled circles correspond to drugs that induce phospholipidosis, while those colored in blue represent drugs that do not induce phospholipidosis. The figure was constructed from the Tanimoto similarity matrix of the 30 drugs, each of which was

TABLE 3.6 Thirty Drugs That Induce Changes in Gene Expression Levels That May or May Not Be Associated with the Onset of Phospholipidosis [20]^a

Mol No.	Drug	Drug Class ^b	Phospholipidosis ^c	Mol. No.	Drug	Drug Class ^b	Phospholipidosis ^c
1	Amitriptyline	AD	+++	16	Loratadine	AI	+
2	Chlorcyclizine	AI	+++	17	Pentamidine	AF	+
3	Fluoxetine	AD	+++	18	Acetominophen	NSAID	-
4	Amiodarone	TH	++	19	Clarithromycin	AB	-
5	AY-9944	AC	++	20	Disopyramide	AA	-
6	Chlorpromazine	AD	++	21	Erythromycin	AB	-
7	Imipramine	AD	++	22	Flecainide	AA	-
8	Perhexiline	AG	++	23	Haloperidol	AP	-
9	Tamoxifen	AN	++	24	Levofloxacin	AB	-
10	Clozapine	AP	++	25	Oflloxacin	AB	-
11	Sertraline	AD	++	26	Procainamide	AA	-
12	Clomipramine	AD	+	27	Quinidine	AA	-
13	Thioridazine	AP	+	28	Sotalol	AA	-
14	Zimelidine	AD	+	29	Sulfamethazole	AB	-
15	Ketoconazole	AF	+	30	Sumatriptan	MH	-

^aNote that the degree of phospholipidosis indicated by the number of plus signs ("+" is not considered in this work, only whether phospholipidosis is or is not induced by a given drug molecule. The minus sign ("−") indicates that phospholipidosis did not occur.

^bAA, antiarrhythmic; AB, antibiotic; AC, anticholesteric; AD, antidepressant; AF, antifungal; AG, antianginal; AI, antiinflammatory; AN, antineoplastic; AP, antipsychotic; MH, migraine headaches; TH, thyroid.

^cThe degree of phospholipidosis induced by a given drug is indicated by the number of pluses ("+" in the fourth and eighth columns.

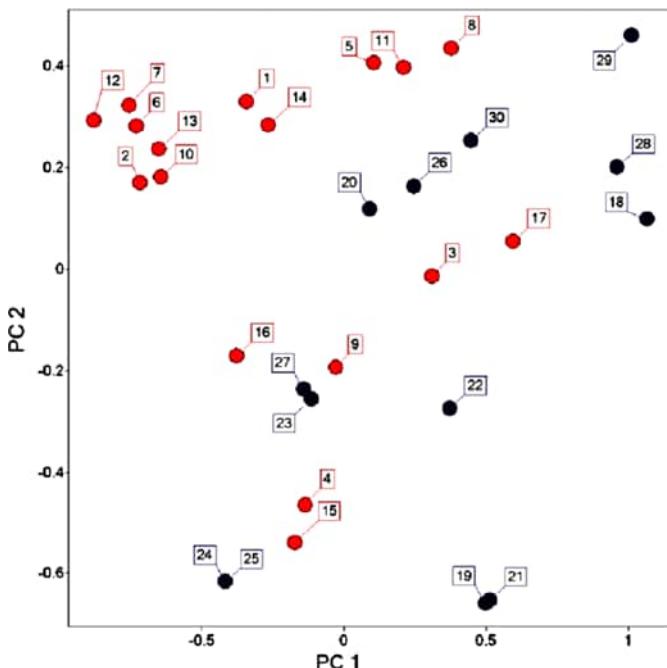


FIGURE 3.3 Two-dimensional representation of the chemical space distribution of the 30 drug molecules tested for phospholipidosis activity. The red filled circles indicate drugs that induce phospholipidosis; the blue filled circles represent drugs that do not induce phospholipidosis. The numbers indicate which specific drug (see Table 3.6) is associated with each data point. For color details, please see color plate section.

represented by its MACCS key molecular fingerprint. The similarity matrix was subjected to principal component analysis and the data points were plotted in terms of the first two principal components as described by Maggiora and Shanmugasundaram [39]. From the figure it is clear that a small number of the phospholipidosis inducing drugs are grouped together in chemical space (see the upper left corner of the figure), although the general sense of the figure shows that many of the drugs are scattered throughout chemical space without significant clustering. In this regard, it should be noted that the distribution of data points in chemical space, or any other inherently high-dimensional space for that matter, becomes less and less dense as the number of dimensions is increased from two to its true dimensionality, which in this case is 30. Thus, it is not unreasonable to expect that the minor clustering observed in Figure 3.3 will largely be dissipated as the number of dimensions increases. Because of this, it is difficult to infer SAR from such highly dispersed data, opening up the possibility for future SAR studies (see Section 3.4).

Table 3.7 lists the set of 13 candidate marker genes initially proposed by Sawada et al. [20] along with their gene symbols, gene products, and biological functions. Genes highlighted in gray were considered to be only weakly correlated with the presence of phospholipidosis and were not considered further by Sawada et al. The upward (\uparrow) and

TABLE 3.7 Candidate Marker Genes Proposed by Sawada et al. [20] and Their Gene Products.

Gene No.	Gene Symbol	Gene Regulation ^a	Gene Product	Biological Function
1	ASAHI★	↑	N-acylsphingosine amidohydrolase (acid ceramidase) 1	Lipid metabolism/phospholipid degradation
2	MGC4171★	↑	Hypothetical protein MGC4171	Lipid metabolism/phospholipid degradation
3	LSS★	↑↑	Lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)	Lipid metabolism/cholesterol biosynthesis
4	NR0B2	↑	Nuclear receptor subfamily 0, group B, member 2	Lipid metabolism/regulation of cholesterol biosynthesis
5	PHYH	■	Phytanoyl-CoA hydroxylase (Refsum disease)	Lipid metabolism/fatty acid alpha-oxidation
6	FABP1★	↑	Fatty acid binding protein 1 (Liver)	Lipid metabolism/fatty acid transport
7	INHBE	↑↑	Activin beta E	Cell cycle, proliferation, death
8	P8★	↑↑	P8 protein (candidate of metastasis 1)	Cell cycle, proliferation, death
9	HPN★	↑↑	Hepsin (transmembrane protease, serine 1)	Proteolysis and peptidolysis
10	SERPIN A3★	↑↑	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member-3	Endopeptidase inhibition
11	ASNS	■	Asparagine synthetase	Miscellaneous
12	C10orf10	↑↑	Chromosome 10 open reading frame 10	Unknown/EST
13	FLJ10055	↑↑	Hypothetical protein FLJ10055	Unknown/EST
14	FRCP1★	↑↑	Likely ortholog of mouse fibronectin type III repeat containing protein 1	Unknown/EST
15	AP1S1	■	Adaptor-related protein complex 1, sigma 1 subunit	Golgi vesicle
16	SLC2A3	↓	Solute carrier family 2 (facilitated glucose transporter), member 3	Transport
17	TAGLN★	↓	Transgelin	Smooth muscle, cell specific cytoskeletal protein

Genes highlighted in gray were considered by Sawada et al. to be only weakly correlated with the presence of phospholipidosis. Genes marked by a star (★) were determined in this work to be associated with the onset of phospholipidosis. See Table 3.8 for detailed expression data on all 17 genes.

^aThe upward arrows (↑) or downward arrows (↓) correspond to genes that are primarily, but not exclusively, up regulated or down regulated, respectively. The filled square (■) denotes a gene that is generally normally expressed.

downward (\downarrow) arrows in the third column indicate that the gene in question is predominantly up regulated or down regulated, respectively. The presence of a filled square (\blacksquare) indicates that the gene is normally expressed. The asterisk (\star) labeling the gene symbols in the second column indicates that the genes determined in this work are the most highly associated with the onset of phospholipidosis (see Sections 3.3.4 and 3.4).

3.3.1 Dataset

Table 3.8 contains the data on the effect of 30 drug molecules (cf. Table 3.6) on the gene expression levels of the 17 genes, the condition attributes (cf. Table 3.7), and whether phospholipidosis, the decision attribute, is induced by each of the drugs. The original gene-expression data is given in Table 3.6 of Sawada et al. [20] as *mRNA Scores* defined as the ratio of observed to normal or basal expression levels. The threshold mRNA Scores of 1.50 and 0.70 that separate up regulated from normally expressed genes and normally expressed from down regulated genes, respectively, are the same as those used by Sawada et al. In this work, the mRNA are discretized (as required by RST) as follows: (i) genes with mRNA Scores greater than 1.50 are classified as *up regulated* and are designated by the upwards arrow (\uparrow), (ii) genes with mRNA Scores that lie between 0.70 and 1.50 are classified as *normally expressed* and are designated by a filled square (\blacksquare), and (iii) genes with mRNA Scores less than 0.70 are classified as *down regulated* and designated by a downwards arrow (\downarrow). Cells in Table 3.8 corresponding to up regulated genes are colored dark gray, cells corresponding to normally expressed genes are uncolored, and cells corresponding to down regulated genes are colored light gray.

From the results of preliminary calculations not reported here, it was concluded that the amount of gene expression data available in this study was insufficient to support a more fine-grained discretization. Nevertheless, the current discretization is consistent with that used by Sawada et al. [20]. The data associated with the induction of phospholipidosis is found in the last column of Table 3.8 under the heading p : a plus sign ($+$) indicates the presence of phospholipidosis and an open circle (\circ) indicates its absence.

The data in Table 3.8 were analyzed with the program Rosetta [40–42]. The software is freely available for download [43]. A step-by-step RST-based analysis similar to that in Sections 3.2.3–3.2.5 is presented in Sections 3.3.2 and 3.3.3. Figure 3.4 provides a description of the rule generation process employed in this work. Details are presented in Sections 3.3.1, 3.3.2, and 3.3.4.

3.3.2 Determination of D -Reducts

Unlike the illustrative example provided in Sections 3.2.3–3.2.5, realistic DTs with larger numbers of condition attributes (17 in the present case) typically generate large sets of reducts. Moreover, obtaining sets of reducts in large information systems is a computationally intensive procedure. Fortunately, although the number of condition attributes is reasonably large, all 148 of the D -reducts can be determined. Fourteen are of length $L=1$ and $L=3$ (N.B. that there are no D -reducts of length $L=2$) and are given in Table 3.9. The remaining 134 D -reducts are of length $L=4, 5$, and 6.

TABLE 3.8 Fold Change Expression Values (“mRNA Scores”) of Candidate Phospholipidosis Marker Genes in Human Hematoma HepG2 Cells Determined By Real Time PCR and Pathology Analysis [20]^{a,b,c}

Mol. No.	Gene Number															Phospholipidosis	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	3.10	2.22	1.79	1.64	1.53	2.39	4.19	2.82	2.20	3.01	1.84	5.08	2.11	1.80	0.88	0.22	0.23
2	3.23	2.43	2.98	2.46	0.67	2.62	1.68	1.85	2.97	2.23	1.41	3.24	2.47	2.96	0.74	0.20	0.30
3	6.55	1.34	3.51	2.70	1.37	2.80	3.61	2.40	2.39	1.73	1.68	3.08	1.93	2.15	0.39	0.09	0.20
4	1.69	1.32	1.82	1.35	0.93	2.01	1.92	1.88	2.15	1.17	1.16	1.12	1.56	1.07	0.64	0.35	0.56
5	4.71	1.63	4.78	2.71	0.73	3.16	0.89	1.22	2.68	1.97	1.05	3.32	1.68	1.45	0.63	0.22	0.21
6	1.29	1.74	1.99	1.20	1.69	2.56	2.72	2.51	1.70	1.94	1.39	2.95	1.90	1.70	1.19	0.43	0.65
7	4.52	1.81	2.69	3.41	1.37	2.45	1.87	2.00	2.32	2.12	1.57	3.36	2.11	1.80	0.56	0.15	0.15
8	3.12	1.93	3.10	2.09	0.60	2.49	1.17	1.42	2.22	2.50	1.46	2.55	1.98	3.42	0.65	0.22	0.26
9	4.50	1.60	3.01	2.56	0.52	1.90	0.81	1.21	2.19	2.12	1.36	2.02	1.85	1.53	0.66	0.24	0.27
10	2.75	1.88	2.69	1.97	0.87	3.89	5.25	3.39	2.35	1.96	2.45	5.15	3.34	1.92	0.71	0.24	0.24
11	3.08	1.94	3.91	1.79	2.62	4.14	5.19	3.88	3.07	2.69	2.56	4.44	2.38	3.49	1.01	0.14	0.37
12	0.82	1.41	1.52	0.98	1.63	2.36	2.75	2.63	2.39	2.03	1.25	2.08	3.83	1.46	0.61	0.63	0.78
13	1.36	1.32	2.03	0.77	1.64	5.07	3.48	3.14	2.85	2.46	0.91	2.02	5.79	1.57	0.63	0.38	0.46
14	2.91	1.12	1.05	1.33	1.14	1.42	1.77	1.52	1.74	1.95	1.58	2.67	2.24	1.71	1.05	0.70	0.59
15	1.47	1.53	4.33	2.36	1.45	5.30	3.09	2.59	2.71	2.37	3.40	2.29	2.24	1.95	1.04	0.48	0.71
16	1.23	1.57	1.54	1.63	1.20	1.58	2.21	1.81	1.52	1.83	2.23	0.79	1.42	1.12	1.11	0.46	0.44
17	0.93	0.73	0.52	1.83	0.94	0.15	10.7	4.89	2.50	2.12	7.72	2.88	1.82	2.49	0.93	0.50	0.68
18	1.18	0.75	0.91	0.69	0.55	1.11	1.19	0.91	1.10	0.99	1.14	0.94	1.13	1.02	0.90	1.50	0.93
19	1.55	1.18	0.65	0.78	0.59	0.91	1.51	1.10	0.96	1.39	1.02	0.93	1.48	1.00	0.79	0.86	0.86
20	1.82	1.21	1.14	0.63	1.04	1.02	1.18	1.39	1.12	0.99	0.83	0.90	0.94	0.98	1.03	0.99	0.99

(continued)

TABLE 3.8 (Continued)

70

Mol. No.	Gene Number																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
21	0.90	0.86	0.87	0.90	0.77	0.73	0.86	0.84	0.87	0.85	1.10	0.71	0.78	0.92	1.03	0.73	0.85
22	2.42	1.57	1.22	1.02	1.04	1.50	2.25	1.82	1.26	1.60	2.54	1.73	2.10	1.65	1.13	0.46	0.86
23	0.77	1.24	1.52	1.49	0.92	1.50	1.01	1.50	1.43	1.31	1.25	1.38	1.32	0.93	0.95	0.44	0.52
24	1.10	0.79	0.75	0.79	0.46	0.88	0.99	0.80	0.83	0.92	0.80	0.83	0.70	0.83	0.80	1.07	0.90
25	1.17	0.77	0.78	0.81	0.75	0.75	0.79	0.74	0.72	0.82	0.67	0.79	0.86	0.68	0.79	0.94	0.74
26	0.86	1.08	0.90	0.74	0.67	1.00	0.89	0.91	1.08	0.96	0.89	0.98	1.78	0.73	0.87	0.88	0.85
27	1.43	1.20	1.30	1.03	0.77	1.38	1.73	1.20	1.21	1.22	2.01	1.06	1.27	1.26	0.96	0.37	0.63
28	1.39	0.72	0.88	0.63	0.54	0.98	1.12	0.84	0.79	0.78	0.67	0.80	0.84	0.85	0.78	1.53	0.80
29	1.26	0.78	0.82	0.82	0.52	0.85	0.95	0.87	0.89	0.98	0.95	0.70	0.84	0.87	0.92	1.11	0.99
30	1.35	0.99	1.23	0.63	0.84	1.01	1.12	1.05	1.16	0.95	0.97	1.05	1.34	1.26	1.22	0.99	1.06

^aCells shaded in dark gray represent over-expression of the gene (), uncolored cells represent normal expression of the gene () and cells colored light gray represent under-expression of the gene ().

^bSee Table 3.6 and Table 3.7 for the names of the molecules and genes, respectively, listed in this table and text for further details.

^cNote that the degree of phospholipidosis indicated by the number of plus signs ("+"") is not considered in this work, only whether phospholipidosis is or is not induced by a given drug molecule. The open circle ("○") indicates that phospholipidosis did not occur.

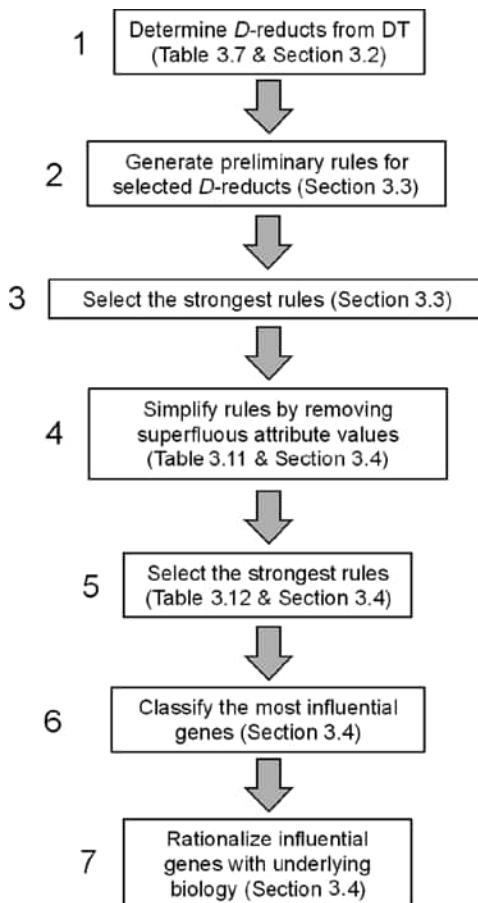


FIGURE 3.4 The process used in this work to identify gene expression levels associated with drug-induced phospholipidosis in HepG2 cells.

These D -reducts were not considered further because the goal of this work is to reduce the amount of gene expression information needed to characterize the association of gene expression with the onset of phospholipidosis. All 148 of the D -reducts, nevertheless, preserve the partitioning of D -space since the indiscernibility classes induced by each of them are subsets of the indiscernibility classes induced by the decision attribute (see the discussion of the example provided in Sections 3.2.3–3.2.5), a condition that maintains the initial quality of classification. Hence, all rules associated with these D -reducts are deterministic (*vide supra*).

The first D -reduct, D_1 , is composed of a single gene, *gene-9*. Examination of Table 3.8 clearly shows that the D -space partition is completely explained by the variability of *gene-9* alone since it is up regulated for the first 17 drug molecules, all of which induce phospholipidosis, and normally expressed for the remaining 13 drug molecules all of which do not induce phospholipidosis. In the first case, this can be

TABLE 3.9 The Set of 14 D -Reducts of Lengths $L=1,3$
Generated from the Data in Table 3.8

p	D -reducts, D_p	Length, L
1	{ gene-9 }	1
2	{ gene-6, gene-8, gene-17 }	3
3	{ gene-2, gene-6, gene-12 }	3
4	{ gene-2, gene-3, gene-8 }	3
5	{ gene-6, gene-14, gene-17 }	3
6	{ gene-1, gene-6, gene-17 }	3
7	{ gene-1, gene-10, gene-17 }	3
8	{ gene-1, gene-8, gene-17 }	3
9	{ gene-2, gene-6, gene-10 }	3
10	{ gene-6, gene-10, gene-17 }	3
11	{ gene-2, gene-6, gene-8 }	3
12	{ gene-2, gene-6, gene-14 }	3
13	{ gene-6, gene-13, gene-17 }	3
14	{ gene-6, gene-12, gene-17 }	3

written in terms of the indiscernibility class induced by the single condition attribute *gene-9* in D_1 with respect to the descriptor (*gene-9*, \uparrow), that is,

$$X_1^{D_1} = \{1,2,\dots,17\} \quad (3.3.1)$$

and by the indiscernibility class induced by the decision attribute with respect to the descriptor (*phospholipidosis*, \oplus), that is,

$$X_1^{\{p\}} = \{1,2,\dots,17\} \quad (3.3.2)$$

Clearly, the two indiscernibility classes are equal, that is, $X_1^{D_1} = X_1^{\{p\}}$, leading to the first rule given in Table 3.10, which is discussed in more detail in Section 3.4. An identical argument shows that

$$X_2^{D_1} = \{18,19,\dots,30\} \quad (3.3.3)$$

when (*gene-9*, \blacksquare), and

$$X_2^{\{p\}} = \{18,19,\dots,30\} \quad (3.3.4)$$

when (*phospholipidosis*, \bullet) again showing that $X_2^{D_1} = X_2^{\{p\}}$. Since this D -reduct only contains a single gene, no further simplification of the rule is possible (*vide infra*).

3.3.3 Generation of Preliminary Rules

As the aforementioned situation is somewhat unusual, that is, a D -reduct associated with the single condition attribute *gene-9*, attention will now be focused on the rules that can be generated from the 13 D -reducts with $L=3$ given in Table 3.9. The different

TABLE 3.10 Eleven Top Rules Describing the Relationship of Gene-Expression Levels to Drug-Induced Phospholipidosis in Hepatoma HepG2 Cells^a

Rule No.	Rule	LHS Supp	RHS Supp	RHS Acc	LHS Cov	RHS Cov
1	IF (<i>gene-9</i> , \uparrow) THEN (<i>phospholipidosis</i> , \bullet)	17	17	1.00	0.57	1.00
2	IF (<i>gene-6</i> , \uparrow) AND (<i>gene-10</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	12	12	1.00	0.40	0.71
3	IF (<i>gene-2</i> , \uparrow) AND (<i>gene-6</i> , \uparrow) AND (<i>gene-10</i> , \uparrow) THEN (<i>phospholipidosis</i> , \bullet)	11	11	1.00	0.37	0.64
4	IF (<i>gene-1</i> , \uparrow) AND (<i>gene-10</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	10	10	1.00	0.33	0.59
5	IF (<i>gene-1</i> , \uparrow) AND (<i>gene-6</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	10	10	1.00	0.33	0.59
6	IF (<i>gene-6</i> , \uparrow) AND (<i>gene-14</i> , \downarrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	10	10	1.00	0.33	0.59
7	IF (<i>gene-2</i> , \uparrow) AND (<i>gene-6</i> , \uparrow) AND (<i>gene-12</i> , \uparrow) THEN (<i>phospholipidosis</i> , \bullet)	10	10	1.00	0.33	0.59
8	IF (<i>gene-6</i> , \uparrow) AND (<i>gene-8</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	10	10	1.00	0.33	0.59
9	IF (<i>gene-1</i> , \uparrow) AND (<i>gene-8</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \bullet)	8	8	1.00	0.27	0.47
10	IF (<i>gene-2</i> , \uparrow) AND (<i>gene-3</i> , \uparrow) AND (<i>gene-8</i> , \uparrow) THEN (<i>phospholipidosis</i> , \bullet)	8	8	1.00	0.27	0.47
11	IF (<i>gene-2</i> , \uparrow) AND (<i>gene-6</i> , \uparrow) AND (<i>gene-8</i> , \uparrow) THEN (<i>phospholipidosis</i> , \bullet)	8	8	1.00	0.27	0.47

^aUp regulated genes are indicated by an upward arrow (\uparrow) and down regulated genes by a downward arrow (\downarrow) phospholipidosis (\bullet).

combinations of expression levels (up regulated, normally expressed, or down regulated) of the genes associated with each of the *D*-reducts are linked to a given value (yes or no) of the decision attribute phospholipidosis. Since our interest is primarily focused on gene expression profiles associated with drug-induced phospholipidosis, the 11 strongest rules produced by *D*-reducts with *L*=1,3 are given in Table 3.10 with respect to the descriptor (*phospholipidosis*, yes).

The rules in Table 3.10 are characterized by the following parameters, which provide a more detailed account of the rule characteristics than provided in the earlier example described in Sections 3.2.3–3.2.5. Left-Hand-Side Support (LHS Supp) is the number of instances of the antecedent (i.e., “IF part”) of a rule, while Right-Hand-Side Support (RHS Supp) is the number of instances of the consequent (i.e., “THEN part”) of a rule. Since the rules are deterministic, these values should be the same, and hence, RHS Accuracy (RHS Acc), which is the ratio of LHS to RHS Support, and is equal to unity for all of the cases considered here. LHS Coverage (LHS Cov) is the ratio of LHS Supp to the number of total number of objects (drugs), while RHS Coverage (RHS Cov) is the ratio of the RHS Supp to the number of objects (drugs) associated with the descriptor (*phospholipidosis*, \bullet). The rules are

sorted in terms of the RHS Cov values, which provide a measure of the strength of the rules; only rules with values ≈ 0.5 or greater are considered. The 11 rules in Table 3.10 are based on the expression levels of 10 genes—*gene-1*, *gene-2*, *gene-3*, *gene-6*, *gene-8*, *gene-9*, *gene-10*, *gene-12*, *gene-14*, and *gene-17*—whose expression level are predictive of drug-induced phospholipidosis.

In order to gain a better understanding of the definitions, consider the D -reduct associated with *Rule-2*, the second strongest rule in Table 3.10, which involves the subset of condition attributes $\{gene-6, gene-10, gene-17\}$ associated with D_{10} and the decision attribute phospholipidosis. Figure 3.5 provides a graphical portrayal of the indiscernibility classes induced by D_{10} , namely,

$$\begin{aligned} X_1^{D_{10}} &= \{1,2,3,5,6,7,8,9,10,11,13,16\} \\ X_2^{D_{10}} &= \{4\} \\ X_3^{D_{10}} &= \{12,15\} \\ X_4^{D_{10}} &= \{14\} \\ X_5^{D_{10}} &= \{17\} \\ X_6^{D_{10}} &= \{18,19,20,21,24,25,26,28,29,30\} \\ X_7^{D_{10}} &= \{22\} \\ X_8^{D_{10}} &= \{23,27\} \end{aligned} \quad (3.3.5)$$

and those induced by the decision attribute phospholipidosis,

$$\begin{aligned} X_1^{\{p\}} &= \{1,2,\dots,17\} \\ X_2^{\{p\}} &= \{18,19,\dots,30\}, \end{aligned} \quad (3.3.6)$$

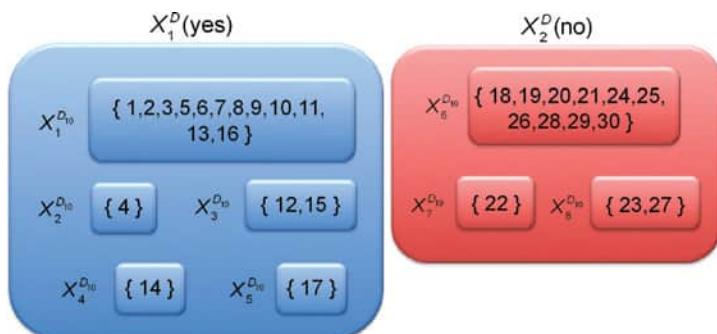


FIGURE 3.5 Schematic depiction of the indiscernibility classes induced by the D -reduct $\{gene-6, gene-10, gene-17\}$ obtained from the hepatoma HepG2 cell gene expression data set presented in Section 3.3.1. The objects in the rectangles shaded in blue correspond to molecules that definitely induce phospholipidosis, while those rectangles shaded in red correspond to molecules that definitely do not induce phospholipidosis. In this case the boundary set is null, so all of the C -indiscernibility classes are subsets of the indiscernibility classes induced by the decision attribute that indicates the presence or absence of phospholipidosis in the hepatoma HepG2 cells. Hence, all of the rules generated from this partitioning are deterministic. For color details, please see color plate section.

TABLE 3.11 Rules Derived from the Tenth D -Reduct, D_{10} , in Table 3.10 after Reduction of the Attribute Values^a

Rule No.	Rule	LHS Supp	RHS Supp	RHS Acc	LHS Cov	RHS Cov
1	IF (<i>gene-6</i> , ■) AND (<i>gene-10</i> , ■) THEN (<i>phospholipidosis</i> , ○)	12	12	1.00	0.400	0.923
2	IF (<i>gene-6</i> , ↑) THEN (<i>phospholipidosis</i> , +)	15	15	1.00	0.500	0.882
3	IF (<i>gene-6</i> , ■) AND (<i>gene-17</i> , ■) THEN (<i>phospholipidosis</i> , ○)	11	11	1.00	0.367	0.846
4	IF (<i>gene-10</i> , ↑) AND (<i>gene-17</i> , ↓) THEN (<i>phospholipidosis</i> , +)	14	14	1.00	0.467	0.824
5	IF (<i>gene-10</i> , ■) AND (<i>gene-17</i> , ■) THEN (<i>phospholipidosis</i> , ○)	10	10	1.00	0.333	0.769
6	IF (<i>gene-6</i> , ↓) THEN (<i>phospholipidosis</i> , +)	1	1	1.00	0.333	0.059

^aUp regulated genes are indicated by an upward arrow (↑), down regulated genes by a down arrow (↓), and normally expressed genes by a filled square (■). A plus sign (+) indicates that phospholipidosis has occurred; while an open circle (○) indicates that it has not occurred.

which give rise to the following values for the rule characteristics

$$\begin{aligned} \text{LHS Supp} &= \text{Card}(X_1^{D_{10}}) = 12 \\ \text{RHS Supp} &= 12 \\ \text{RHS Acc} &= \text{LHS Supp} / \text{RHS Supp} = 12 / 12 = 1.00 \\ \text{LHS Cov} &= \text{LHS Supp} / \text{Total Drugs} = 12 / 30 = 0.40 \\ \text{RHS Cov} &= \text{Card}(X_1^{D_{10}}) / \text{Card}(X_1^{\{p\}}) = 12 / 17 = 0.71 \end{aligned} \quad (3.3.7)$$

where “Card” is the number of elements in the given set.

In order to generate *Rule-2* in Table 3.11, the values of the attributes are needed (see Table 3.8) and are given here in terms of the relevant descriptors of the set of condition attributes (genes) associated with D_{10} , namely,

$$\begin{aligned} &(\text{gene-6}, \uparrow) \\ &(\text{gene-10}, \uparrow) \\ &(\text{gene-17}, \downarrow) \end{aligned} \quad (3.3.8)$$

and the decision descriptor

$$(\text{phospholipidosis}, +) \quad (3.3.9)$$

Thus, from the data in Equations 3.3.8 and 3.3.9, it is possible to construct *Rule-2* of Table 3.10.

3.3.4 Rule Simplification—Reduction of Attribute Values

The rules given in Table 3.10 can be further simplified by the removing superfluous attribute values. Since removal of these values does not influence the partitioning of D -space, the overall prediction of a given rule is unaffected (see Sections 3.2.4.1, 3.2.4.2, and 3.2.5 for additional discussion with respect to the example presented in Section 3.2.3).

Since the first rule in Table 3.10 contains only *gene-9* it cannot be simplified further. Hence, the rule simplification procedure will be illustrated using the *D*-reduct, D_{10} (see Table 3.9), associated with *Rule-2*. Removing the attributes whose values are superfluous yields the set of rules given in Table 3.11. Three of the rules, *Rule-1*, *Rule-3*, and *Rule-5*, are shaded in gray in the table. Since these rules indicate that normally expressed genes are not associated with drug-induced phospholipidosis in HepG2 cells, they are of no further interest in this work. *Rule-6*, which shows that under-expression of the gene 6 may cause phospholipidosis, is a very weak rule (RHS Supp=1, RHS Cov=0.059) and thus is neglected.

Rule-2 and *Rule-4* of Table 3.11 are definitely of interest since both are reasonably strong rules that have LHS–RHS support values of 15 and 14, respectively. *Rule-2* indicates that up regulation of *gene-6* is associated with phospholipidosis. Because there are a total of 17 cases of drug-induced phospholipidosis in the dataset, *Rule-2* is considered to be a strong rule (RHS Cov=15/17=0.882). *Rule-4* exhibits similar strength (RHS Cov=14/17=0.824), indicating that up regulation of *gene-10* and down regulation of *gene-17* are associated with drug-induced phospholipidosis. Hence, *gene-6*, *gene-10*, and *gene-17* are all strongly associated with the onset of phospholipidosis in HepG2 cells.

The remaining *D*-reducts of length $L=3$ in Table 3.9 that are associated with the rules in Table 3.10 can be analyzed in a similar fashion. This yields the set of simplified rules given in Table 3.12, sorted by their respective RHS Cov values. The table clearly shows several interesting trends with respect to drug-induced phospholipidosis. First, over-expression of gene 9 is always associated with drug-induced phospholipidosis as codified in *Rule-1*. Second, up regulation of gene 6 is also often linked to

TABLE 3.12 Most Important Rules for the Induction of Phospholipidosis^a

Rule No.	Rules	RHS Cov
1	IF (<i>gene-9</i> , \uparrow) THEN (<i>phospholipidosis</i> , \oplus)	1.00
2	IF (<i>gene-6</i> , \uparrow) THEN (<i>phospholipidosis</i> , \oplus)	0.88
3	IF (<i>gene-10</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \oplus)	0.82
4	IF (<i>gene-14</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \oplus)	0.71
5	IF (<i>gene-8</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \oplus)	0.71
6	IF (<i>gene-3</i> , \uparrow) AND (<i>gene-8</i> , \uparrow) THEN (<i>phospholipidosis</i> , \oplus)	0.71
7	IF (<i>gene-1</i> , \uparrow) AND (<i>gene-17</i> , \downarrow) THEN (<i>phospholipidosis</i> , \oplus)	0.65
8	IF (<i>gene-2</i> , \uparrow) AND (<i>gene-3</i> , \uparrow) THEN (<i>phospholipidosis</i> , \oplus)	0.65

^aUp regulated genes are indicated by an upward arrow (\uparrow) and down regulated genes by a downward arrow (\downarrow). A plus sign (\oplus) indicates that phospholipidosis has occurred; while an open circle (\bullet) indicates that it has not occurred.

the drug-induced phospholipidosis. As both of these genes, considered separately, are strongly related to drug-induced phospholipidosis, they can be considered to be of singular importance. Four other rules, namely *Rule-3*, *Rule-4*, *Rule-5*, and *Rule-7*, indicate that the under-expression of the gene 17 is associated with over-expression of one of four other genes, respectively, *gene-10*, *gene-14*, *gene-8*, and *gene-1*. *Rule-6* implicates over-expression of *gene-3* and *gene-8* in drug-induced phospholipidosis, the latter already being implicated in *Rule-5*. Lastly, *Rule-8* involves the over-expression of both *gene-3*, which is already involved in *Rule-6*, and *gene-2*, which is not involved in any other of the rules in Table 3.12.

3.4 DISCUSSION

On the basis of the rules given in Table 3.12, a subset of nine genes is associated with the onset of drug-induced phospholipidosis in human hepatoma HepG2 cells. Because RST removes redundant information, it is not unexpected that all of the 12 genes selected by Sawada et al. [20] as putative marker genes for phospholipidosis will correspond to the genes obtained in our work; rather, it is expected that they will constitute a superset, which is what is observed in Table 3.7. Specifically, eight of the nine genes identified here by stars (★) in the second column of the table correspond to genes also selected by Sawada et al. Four genes selected by Sawada et al., *gene-4* (NR0B2), *gene-12* (C10orf10), *gene-13* (FLJ10055), and *gene-16* (SLC2A3), were not obtained in the present work. One gene identified by us, *gene-8* (P8), was not found by Sawada et al. to be associated with phospholipidosis.

Examination of Table 3.8 visually confirms that there is considerable redundancy in the gene expression data. For example, many genes exhibit very similar patterns of up or down regulation. In the latter case, for example, *gene-16* (SLC2A3) and *gene-17* (TAGLN), while associated with different biological functions, nonetheless, are both similarly down regulated by most of the first 17 drugs in Table 3.8. Similar arguments can be made for the up regulated genes, but they are less obvious visually since so many have similar patterns of up regulation in response to the drugs. Because of this it is not surprising that the RST approach would identify a smaller set of putative marker genes than would the work of Sawada et al. Nevertheless, the number of genes in common identified by both approaches provides a measure of support for the marker genes determined in this work.

Further reducing the number of genes needed to characterize phospholipidosis may be beneficial, as it can simplify the analysis of potential pathways associated with drug-induced phospholipidosis, facilitating efforts to understand and characterize this biological phenomenon. In addition, it can provide a more parsimonious set of marker genes that can be used to identify the occurrence of phospholipidosis using genomics array technology to analyze relevant gene expression levels. These arguments must, however, be used with caution since a minimal set of marker genes may miss crucial mechanistic features linked to genes that are coregulated but are associated with other mechanistic pathways.

This raises an additional question, namely, whether it is possible to infer how the genes listed in Table 3.7 influence the onset of phospholipidosis? In considering this question, it is important to distinguish between those genes associated with processes that cause phospholipidosis and those genes linked to processes that are a consequence of phospholipidosis. The first two genes given in the table, *gene-9* (HPN) and *gene-6* (FABP1), are up regulated and are associated with *Rule-1* and *Rule-2*, respectively. These are the two strongest rules in Table 3.12. The work presented here suggests that the up regulation of these genes plays a significant role in drug-induced phospholipidosis. An examination of Table 3.7 shows that these genes have different biological functions; *gene-9* (HPN) is involved with proteolysis and peptidolysis while *gene-6* (FABP1) is involved in fatty acid transport. While the details of how these genes carry out their putative function(s) are unknown, they appear consistent with phospholipidosis, but this is not always the case. For example, it is difficult to understand how the up regulation of *gene-1* (ASAHI) and *gene-2* (MGC4171), which are associated with the degradation of phospholipids, leads to phospholipidosis, a condition linked to an excess of phospholipids. These two genes may well be examples of genes whose associated function(s) are influenced by phospholipidosis rather than contributing to its cause (*vide supra*). Another example involves down regulation of *gene-17* (TAGLN), which controls the amount of the cytoskeletal protein Transgelin. This observation again begs the question of whether the observed down regulation of its expression is a consequence of the onset of phospholipidosis or is one of its causes. Thus, it seems that in most cases examined in this work, identifying whether specific genes and their associated biological processes cause or are caused by the onset of phospholipidosis may be problematic. This is not surprising given the complex interplay of the various processes that is a hallmark of biological systems. Understanding the interrelationships requires considerably more data than can be obtained in gene expression experiments.

An approximate ranking of the relative importance of the genes marked with a star (★) in Table 3.7 to the onset of phospholipidosis is given in Table 3.13. The ranking is based on the observed expression levels of the genes, their frequency of occurrence in the rules, and the strengths of the corresponding rules (*vide supra*).

TABLE 3.13 Approximate Ranking of Marker Genes Determined in This Work (See Table 3.7 for Additional Details)

Rank	Gene Number	Regulation	Gene Symbol
1	9	↑	HPN
2	6	↑	FABP1
3	17	↓	TAGLN
4	10	↑	SERPINA3
5	8	↑	P8
6	14	↑	FRCP1
7	3	↑	LSS
8	1,2	↑,↑	ASAHI, MGC4171

Rules associating gene expression levels with the nonoccurrence of phospholipidosis can also be generated in an analogous fashion to what has been described for drug-induced phospholipidosis. However, the strongest rules are all related to the trivial case associated with normal levels of gene expression, and thus, discussion of these rules is not pursued further in this work.

3.5 SUMMARY AND CONCLUSIONS

A method based on the theory of rough sets has been applied to an analysis of the relationship of drug-induced changes in gene expression levels to the presence of phospholipidosis in human hepatoma HepG2 cells. A relatively brief description of the methodology is presented along with a simple example that is meant to illustrate its application to the study of drug-induced phospholipidosis and its relationship to cellular gene expression levels. The method provides a means for identifying minimal sets of attributes (genes in this work) called *D*-reducts that can generate linguistic rules that associate expression levels of specific genes with the presence or absence of the cellular pathology phospholipidosis. Our work, which is based on the experimental study of Sawada et al. [20], identifies a set of 9 genes from the original set of 17 proposed by Sawada et al. While the putative function of these genes is known, the mechanistic details of how they function in phospholipidosis are not. Nevertheless, they can serve as an effective and parsimonious set phospholipidosis marker genes. This shows that descriptive rules such as those described in this work can be useful in identifying minimal sets of marker genes associated with specific cellular phenomena such as phospholipidosis.

Although the mRNA Scores used by Sawada et al. were more highly resolved than those used here, which can be seen by comparing Table 3.6 in reference [20] with Table 3.8 in this work, the amount of available data was insufficient to support a finer resolution. To do so requires an assessment of more compounds than the 30 studied by Sawada et al. Two approaches can be used to accomplish this. One approach is based on expanding the diversity of compounds subjected to study. Another related approach is to employ a typical ligand-based virtual screening procedure [44], where compounds similar to those known to be active in inducing phospholipidosis are selected for further study. An advantage of this latter approach is that it may be possible to develop SARs, albeit local ones, which will enable the prediction of gene-expression profiles of active compounds directly without the need for gene-chip experiments. An additional benefit of such studies is that they could provide a firmer basis for assessing what genes are most important markers for phospholipidosis.

Rough set theory appears to be an ideal method for dealing with gene expression data such as that reported on here (see e.g., [12, 13]). However, the methodology has a broader range of applicability for characterizing biological systems and subsystems [45–47].

Currently there are relatively few applications of RST in drug research, although their number is growing. The simplest approach involves investigations of SARs where the condition attributes are associated with substituents at specified positions

of molecular scaffolds [48, 49]. While such an approach can handle congeneric series, it is too limited for the more general types of SAR studies that are relatively common today in, for example, combinatorial chemistry library design and in hit analysis associated with high throughput screening (HTS) campaigns. Fortunately, a number of recent studies have shown that this restriction is largely unnecessary [50–54]. Even the problem of discretizing the attribute values, which can sometimes present a problem, has been addressed by Gu et al. [55] using a modified version of the novel ChiMerge algorithm. An application of RST to the analysis of complex absorption, distribution, metabolism, excretion, and toxicity (ADMET) data has also been presented [56]. An advantage of the RST approach is that it can induce sets of linguistic rules, as described for gene expression in Sections 3.3.3 and 3.3.4. Linguistic rules can potentially improve communication among scientists, especially those not versed in the detail mathematics of many modeling methods. An additional advantage is that categorical variables, which are nominal and nonordered, can be used.

Thus, it appears that RST has a place in many aspects of modern drug discovery research. And while it has its limitations, as is the case with all mathematical models, some of its features make it a sound choice for a number of different classes of drug research problems that might otherwise be difficult to treat by more traditional methods.

NOTES

1. Blanking out the expression values of *gene-3* in Table 3.2 and observing the remaining expression values clearly shows that the overall partitioning of C-space is unchanged.
2. In formal logic the “IF part” of a proposition is called the antecedent and the “THEN part” is called the consequent.

REFERENCES

1. Heijne WHM, Lamers R-JAN, van Bladeren PJ, et al. Profiles of metabolites and gene expression in rats with chemically-induced hepatic necrosis. *Toxicol Pathol* 2005;33:425–433.
2. Leavitt J, Goldman D, Merril C, et al. Changes in gene expression accompanying chemically-induced malignant transformation of human fibroblasts. *Carcinogenesis* 1982;3:61–70.
3. Lu J, Pei H, Kaeck M, et al. Gene expression changes associated with chemically induced rat mammary carcinogenesis. *Mol Carcinog* 1997;20:204–215.
4. Lambert CB, Spire C, Renaud M-P, et al. Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. *Toxicol In Vitro* 2009;23:466–475.
5. Moggs JG, Tinwell H, Spurway T, et al. Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth. *Environ Health Perspect* 2004;112:1589–1606.
6. Pawlak Z. *Rough Sets – Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer Academic Publishers; 1991.
7. QianY, Liang J, Dang C. Converse approximation and rule extraction from decision tables in rough set theory. *Comput Mathematics* 2008;55:1754–1765.

8. Nasiri JH, Mashinchi M. Rough set and data analysis in decision tables. *J Uncertain Syst* 2009;3:232–240.
9. Xiao J-M. New rough set approach to knowledge reduction in decision table[s]. *Proc Int Conf Mach Learn Cybern* 2004;4:2208–2211.
10. Ziarko W. On learnability of decision tables. *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing; Uppsala*. New York: Springer Verlag; 2004 p 394–401. Lectures Notes in AI 3066.
11. Cios K, Pedrycz W, Swiniarski R. Rough sets. *Data Mining—Methods for Knowledge Discovery*. Dordrecht: Kluwer Academic Publishers; 1998.
12. Midelfart H, Kormorowski J, Norsett K, et al. Learning rough set classifiers from gene expressions and clinical data. *Fund Inform* 2002;53:155–183.
13. Hvidsten TR, Lægreid A, Komorowski J. Learning rule-based models of biological process from gene expression time profiles using Gene Ontology. *Bioinformatics* 2003;19: 1116–1123.
14. Lee JA, Verleysen M. *Nonlinear Dimensionality Reduction*. New York: Springer; 2007.
15. Quinlan JR. Induction of decision trees. *Mach Learn* 1986;1:81–106.
16. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
17. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press; 2000.
18. Saaty TL. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. Pittsburg: RWS Publications; 2000.
19. Bender EA. *Mathematical Methods in Artificial Intelligence*. Los Alamitos: IEEE Computer Society Press; 1996.
20. Sawada H, Takami K, Asahi S. A toxicogenomic approach to drug-induced phospholipidosis: Analysis of its induction mechanism and establishment of a novel *in vitro* screening system. *Toxicol Sci* 2005;83:282–292.
21. Reasor MJ, Hastings KL, Ulrich RG. Drug-induced phospholipidosis: issues and future directions. *Expert Opin Drug Safety* 2006;5:567–583.
22. Nonoyama T, Fukuda R. Drug induced phospholipidosis pathological aspects and its prediction. *J Toxicol Pathol* 2008;21:9–24.
23. Tengstrand-Baronas E, Lee JW, Alden C, et al. Biomarkers to monitor drug-induced phospholipidosis. *Toxicol Appl Pharmacol* 2007;218:72–78.
24. Occam's razor, Wikipedia: http://en.wikipedia.org/wiki/Occam's_razor. Accessed 2012 May 10.
25. Fakih SJ, Das TK. LEADF: a methodology for learning efficient approaches to medical diagnosis. *IEEE Trans Inf Technol Biomed* 2006;10:220–228.
26. Collette TW, Szladow AJ. Use of rough sets and spectral data for building predictive models of reaction rate constants. *Appl Spectrosc* 1994;48:1379–1386.
27. Hashemi RR, Young JF. The prediction of methylmercury elimination half-life in humans using animal data: a neural network/rough sets analysis. *J Toxicol Environ Health, Part A* 2003;66:2227–2252.
28. Chèvre N, Gagné F, Gagnon P, et al. Application of rough sets analysis to identify polluted aquatic sites based on a battery of biomarkers: a comparison with classical methods. *Chemosphere* 2003;51:13–23.

29. Chèvre N, Gagné F, Blaise C. Development of a biomarker-based index for assessing the ecotoxic potential of aquatic sites. *Biomarkers* 2003;8:287–298.
30. Cao Y, Liu S, Zhang L, et al. Prediction of protein structural class with rough sets. *BMC Bioinformatics* 2006;7:20.
31. Krysiński J, Skrzypczak A, Demski G, et al. Application of rough set theory in structure-activity relationships of anti-electrostatic Imidazolium compounds. *Quant Struct-Act Relat* 2002;20:395–401.
32. Hvidsten TR. Predicting function of genes and proteins from sequence, structure and expression data. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology, Acta Universitatis Upsaliensis, Uppsala; 2004. 63 pp.
33. Hvidsten TR, Wilczyński B, Kryshtafovych A, et al. Discovering regulatory binding-site modules using rule-based learning. *Genome Res* 2005;15:856–866.
34. Walczak B, Massart DL. Tutorial: rough sets theory. *Chemomet Intell Lab Syst* 1999;47: 1–16.
35. Komorowski J, Polkowski L, Skowron A. Rough sets: a tutorial. *Synthesis* 1999;46:2–8.
36. Polkowski L, Kormorowski J, Pawlak Z, et al. Rough set: a tutorial. In: Pal SK, Skowron A, editors. *Rough Fuzzy Hybridization. A New Trend in Decision Making*. New York: Springer-Verlag; 1999.
37. Stefanowski J. On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A, editors. *Rough Sets in Data Mining and Knowledge Discovery*. vol. 1. New York: Physica-Verlag; 1998. p 500–529.
38. Stefanowski J. The rough set based rule induction technique for classification problems. Proceedings of the 6th European Conference on Intelligent Techniques and Soft Computing EUFIT 98; 1998 Sept 7–10; Aachen. p 109–113.
39. Maggiora GM, Shanmugasundaram V. Molecular similarity measures. In: Bajorath J, editor. *Cheminformatics and Computational Chemical Biology*. New York: Humana Press/Springer Science + Business; 2011. p 39–100.
40. Øhrn A. *ROSETTA Technical Reference Manual*. Trondheim: Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU); 2000. 66 pages.
41. Øhrn A. *The ROSETTA C++ Library: Overview of Files and Classes*. Trondheim: Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU); 2000. 45 pages.
42. Komorowski J, Øhrn A, Skowron A. The ROSETTA rough set software system. In: Klösgen W, Zytkow J, editors. *Handbook of Data Mining and Knowledge Discovery*. New York: Oxford University Press; 2002.
43. Rosetta – A Rough Set Toolkit for Analysis of Data. The software can be downloaded from the website: <http://www.lcb.uu.se/tools/rosetta/index.php>. Accessed 2012 May 16.
44. Ripphausen P, Nisius B, Bajorath J. State-of-the-art in ligand-based virtual screening, *Drug Discov Today* 2011;16:372–376.
45. Hvidsten TR, Komorowski J. Rough sets in bioinformatics. *Transaction on Rough Sets VII*. Berlin: Springer-Verlag; 2007. p 225–243.
46. Cao Y, Liu S, Zhang L, et al. Prediction of protein structural class with rough sets. *BMC Bioinformatics* 2006;7:20.
47. Yellasiri R. Rough set protein classifier. *J Theor Appl Inform Technol* 2005;5:1–7.

48. Krysinski J. Application of rough sets theory to the analysis of structure-activity – relationships of anti-microbial pyridinium compounds. *Pharmazie* 1995;50:593–597.
49. Liu H, Qu L, Gao H, et al. Study of the quantitative structure relationship of C-10 substituted artemisinin (QHS)’s derivative using rough set theory. *Sci China Ser B: Chem* 2008;51:937–945.
50. Petit J, Maggiore GM. Application of rough set theory to structure-activity relationships. 229th National American Chemical Society Meeting; 2005 Mar 13–17; San Diego. Division of Chemical Information Abstr. No. 51
51. Maji P, Paul S. Rough sets for selection of molecular descriptors to predict biological activity of molecules. *IEEE Trans Syst Man Cybern* 2010;40:639–648.
52. Dong Y, Xiang B, Wang T, et al. Rough set-based SAR analysis: an inductive method. *Expert Syst Appl* 2010;37:5032–5039.
53. Petit J, Meurice N, Mousses S, et al. Applications of rough sets theory in drug discovery: Analysis of HTS data relative to the inhibition of Aurora A kinase. 236th National American Chemical Society Meeting; 2008 Aug 17–21; Philadelphia. Division of Chemical Information Abstr. No. 40.
54. Koyama M, Hasegawa K, Arakawa M, et al. Application of rough set theory to high throughput screening data for rational selection of lead compounds. *Chem-Bio Inf J* 2008;8:85–95.
55. Gu X-H, Hou D-B, Zhou Z-K, et al. Rough set based modified ChiMerge algorithm and its applications. *Proc 2005 Int Conf Mach Learn Cybern* 2005;2:1004–1008.
56. Medina-Franco JL, Maggiore GM, Goodwin JT, et al. Rule-based analysis of ADMET data using rough set theory. 232nd National American Chemical Society Meeting; 2006 Sept 10–14; San Francisco. Division of Computers in Chemistry, Abstr. No. 140.

CHAPTER 4

BIMODAL PARTIAL LEAST-SQUARES APPROACH AND ITS APPLICATION TO CHEMOGENOMICS STUDIES FOR MOLECULAR DESIGN

KIYOSHI HASEGAWA and KIMITO FUNATSU

4.1 INTRODUCTION

Quantitative structure–activity relationship (QSAR) studies express the biological activities of compounds as functions of their various chemical descriptors. Essentially, they describe how biological activity variation depends on changes of chemical structure [1, 2]. If a clearly defined relationship can be derived from the structure–activity data, the model equation allows chemists to determine with some confidence which physicochemical properties play an important role in biological activity, and thereby to attempt predictions. By quantifying physicochemical properties, it should then be possible to calculate in advance the biological activity of a novel compound.

As a multivariate statistical method, partial least square (PLS) is of particular interest in the QSAR field [3]. PLS can analyze data with strongly collinear, noisy, and numerous X variables, while simultaneously modeling several response variables Y . PLS can also provide several prediction regions and diagnostic plots as statistical measures. Using such an approach, QSAR scientists can extract the patterns embedded in the structure–activity data.

Several new 2D and 3D molecular descriptors have been invented and proposed for QSAR use [4]. However, the relatively high uncertainty associated with molecular descriptors constitutes a difficult problem. Thus, the search for more informative 2D and/or 3D molecular descriptors has been of major concern in QSAR research. Using variable selection from a very large pool of descriptors, PLS can detect the informative molecular descriptors [5].

The omics fields have become an increasingly large component of biological research. Omics studies encompass multiple disciplines across heterogeneous scientific fields and include, for example, chemogenomics, proteomics, and metabolomics [6]. Because omics experiments typically generate vast amounts of data, the data matrices X and Y , whose correlation is to be analyzed, are typically large and undertake highly complex interactions through another matrix Z . Omics data have necessitated the evolution of standard PLS to cope with the multiple demands of complex data structures [7].

In this chapter, we focus on chemogenomics and the application of bimodal PLS to aminergic G protein-coupled receptor (GPCR) inhibitory activity data. A special bimodal PLS approach, termed L-shaped PLS (LPLS) [8], was used to connect ligand and protein descriptors to biological activities. The LPLS approach explores relationships between both matrix columns and rows by building bimodal or bifocal models. Besides constructing a regression model for the descriptor matrix X and the response matrix Y , LPLS builds a further regression model connecting the weights or loadings of X to another matrix Z .

To date, chemogenomics modeling has adopted the kernel approach combined with a nonlinear method [9, 10]. However, chemical interpretability requires a more comprehensive approach. Chemical interpretation is valuable for generating hypotheses or knowledge, which is the final goal of molecular design. The LPLS method is suitable for chemical interpretation, and the corresponding regression coefficient matrix can guide the design of novel inhibitors against an orphan GPCR target.

4.2 MATERIAL AND METHODS

4.2.1 Aminergic GPCR Inhibitory Activity Data

We collected a dataset of human aminergic GPCR inhibitors from the GVK database (GVK DB) [11]. The inhibitory activity against human aminergic GPCR was expressed as the logarithm of the reciprocal IC_{50} value (pIC_{50}), where IC_{50} represents the micromolar concentration at which 50% inhibition is achieved.

The inhibitory activity data stored in the GVK DB are incomplete and some data are missing for each molecule and GPCR target. To generate the missing data, we performed an appropriate PLS analysis against each GPCR target. In these cases, the extended-connectivity fingerprints of depth 6 (ECFP_6) were used as chemical descriptors [12]. PLS models with Q^2 values greater than 0.500 were selected and used to predict the missing data for each GPCR target. The components of each PLS model were determined by 10-fold cross-validation [3]. For consistency, observed inhibitory activity values were replaced by their predicted values. Table 4.1 shows the selected aminergic GPCR targets used in subsequent LPLS analysis. The total dataset comprised a matrix of 6185 compounds against 16 GPCR targets.

The ECFP_6 and PLS analysis with cross-validation was performed using Pipeline Pilot of Accelrys [13].

TABLE 4.1 Selected Aminergic GPCR Targets Used in Subsequent LPLS Analysis

No.	Targets	No. of Compounds	PLS Components	R^2	Q^2
1	5HT1B	68	3	0.834	0.642
2	5HT1D	88	11	0.923	0.640
3	5HT1F	175	4	0.699	0.571
4	5HT2A	1270	17	0.648	0.515
5	5HT2B	252	8	0.713	0.513
6	5HT2C	1954	18	0.660	0.594
7	5HT4	108	4	0.809	0.702
8	M3	1203	10	0.746	0.659
9	A1B	696	8	0.664	0.507
10	A1D	657	8	0.688	0.573
11	A2A	334	4	0.659	0.535
12	A2C	276	3	0.756	0.674
13	B1	99	9	0.889	0.659
14	B3	208	5	0.690	0.569
15	D1	509	9	0.688	0.520
16	H1	289	5	0.686	0.510

R^2 and Q^2 represent squared and leave-one-out cross-validated correlation coefficient values, respectively.

4.2.2 Ligand and Protein Descriptors for LPLS Analysis

The ECFP_6 ligand descriptors used to construct the full aminergic GPCR inhibitory activity data were also used in the LPLS analysis. After filtering by 0.05 variance cutoff, 149 fingerprints were eligible from a generated total of 36,252.

To construct the protein descriptors, z -scales were used. z -scales, originally developed as a descriptor of amino acids, contain three variables designated z_1 , z_2 , and z_3 [14]. The z parameters were determined by principal component analysis of 29 physicochemical parameters characterizing 20 natural amino acids. The first, second, and third principal components, corresponding to z_1 , z_2 , and z_3 , respectively, could be tentatively interpreted as hydrophobicity, steric, and electronic properties.

Using this approach, the amino acid sequences of the aminergic GPCR proteins were translated into vectors of numbers. Because the sequence portions of the cavity of aminergic GPCRs can be aligned unambiguously, comparisons between resulting vectors should directly identify protein variations between aminergic GPCRs. The z -scales of each aligned sequence residue were combined into a uniform protein descriptor matrix. Table 4.2 shows the cavity-based alignment proposed by Rognan et al. [15]. Among 30 amino acid residues forming the cavity, three residues (AA_11, AA_22, and AA_23) are strictly conserved and were not considered further. For each of the remaining 27 amino acids, three z -scales were assigned to yield a total of 81 z -scales as the protein descriptors.

TABLE 4.2 Cavity-Based Alignment of 16 Aminergic GPCRs

GPCR_Labels	AA_Labels	5HT1B	5HT1D	5HT1F	5HT2A	5HT2B	5HT2C	5HT4	M3	A1B	A1D	A2A	A2C	B1	B3	D1	H1
1.35	AA_01	L	L	V	S	A	P	L	I	V	L	M	A	T	L		
1.39	AA_02	L	L	L	T	L	L	L	T	L	A	L	L	L	L		
1.42	AA_03	I	I	V	T	T	F	A	V	T	T	T	T	T	T	I	
1.46	AA_04	T	T	V	V	V	S	V	M	S	G	A	K				
2.57	AA_05	V	M	M	S	S	M	F	S	E	V	A	E	L			
2.58	AA_06	M	S	I	I	W	W	W	Y	L	D	D	D	D	D		
2.61	AA_07	S	T	W	W	L	A	D	D	V	V	V	S	S			
2.65	AA_08	T	W	L	D	D	D	V	C	C	I	I	I	I	Y		
3.28	AA_09	W	W	D	I	V	S	C	I	V	V	V	V	V	I		
3.29	AA_10	L	L	D	D	V	T	S	I	I	V	V	V	V	V		
3.32	AA_11	D	I	C	S	S	S	I	I	I	P	P	P	P	P		
3.33	AA_12	I	I	C	I	V	V	I	I	V	P	P	P	P	P	I	
3.36	AA_13	C	C	C	I	I	I	V	I	I	V	Y	Y	Y	Y	F	
3.40	AA_14	I	I	I	I	P	P	I	I	V	A	V	A	V	A	K	
4.56	AA_15	I	I	I	I	V	V	F	F	V	A	A	S	S	S	T	
4.60	AA_16	P	Y	S	S	F	F	M	G	S	T	S	C	S	S	A	
5.38	AA_17	Y	T	T	T	V	V	G	S	A	A	S	S	S	S	N	
5.39	AA_18	T	S	S	S	G	S	A	A	F	F	F	F	F	F	F	
5.42	AA_19	S	T	T	T	V	F	F	N	W	W	W	W	W	W	W	
5.43	AA_20	T	A	A	A	S	S	S	N	F	F	F	F	F	F	F	
5.46	AA_21	A	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
6.44	AA_22	F	W	W	W	F	F	F	F	W	W	W	W	W	W	W	
6.48	AA_23	W	F	F	F	F	F	F	F	F	F	F	F	F	F	F	
6.51	AA_24	F	S	E	N	N	N	N	N	V	L	L	Y	Y	N	N	
6.52	AA_25	F	F	F	N	N	N	N	N	W	W	W	F	F	F	H	
6.55	AA_26	S	E	N	N	N	N	N	N	Y	Y	Y	N	N	N	F	
7.35	AA_27	F	F	S	L	L	L	L	L	Y	F	F	F	F	F	H	
7.39	AA_28	T	T	A	V	V	V	Y	Y	Y	F	F	F	F	N	V	
7.43	AA_29	Y	Y	Y	S	S	S	S	S	N	N	N	Y	Y	Y	Y	
7.45	AA_30	N	N	N	N	N	N	C	N	N	N	N	N	N	N	N	

The alphabets from column 3 to 17 are one-letter descriptions of amino acids.

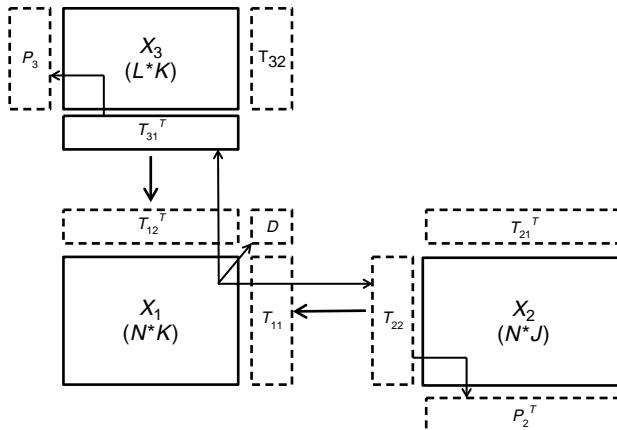


FIGURE 4.1 Architecture of endo-LPLS.

4.2.3 L-Shaped PLS

The LPLS approach was introduced by Martens et al. for exploring consistent patterns of covariation between three data matrices arranged in an L-shaped system, where X_2 and X_3 provide additional descriptors to the rows and columns of X_1 , respectively [16]. Two types of LPLS approach are used, depending on whether the matrix X_1 is the regressand (endo-LPLS) or regressor (exo-LPLS) [8]. In our study, the endo-LPLS was used to explain the inhibitory activity data on the basis of the ligand and protein descriptors. The “endo” prefix reflects the inward-pointed regression of a single response matrix X_1 from two outer regressors X_2 and X_3 . The architecture of endo-LPLS is shown in Figure 4.1. The matrices X_1 , X_2 , and X_3 contain the inhibitory activity data, the ligand descriptors, and the protein descriptors, respectively. In Figure 4.1, the lengths of vectors N , K , J , and L are 6185, 16, 149, and 81, respectively.

The original endo-LPLS algorithm was based on singular value decomposition (SVD) of the matrix product of X_1 , X_2 , and X_3 . As an alternative to SVD, the nonlinear iterative partial least squares (NIPALS) algorithm [17] can be used to extract single vectors from three data matrices. Here, LPLS is introduced as the NIPALS algorithm with sequential extraction of latent structures (T s). Analogous to ordinary PLS, a relatively small set a of latent structures is assumed sufficient for capturing the majority of the variability in the response variables. The small set a is termed the number of components of LPLS and is typically determined by cross-validation [3].

The endo-LPLS algorithm proceeds as follows [8]:

For latent vectors extraction $a = 1, \dots, A$

- Find t -vectors t_{22}^a and t_{31}^a by the NIPALS algorithm iterating through $X_1^{a-1,a-1}$, X_2^{a-1} , and X_3^{a-1} :

$$t_{22}^a = X_1^{aa} t_{11} \left(t_{11}^T t_{11} \right)^{-1}, \quad t_{31}^a = X_1^{aa} t_{12} \left(t_{12}^T t_{12} \right)^{-1} \quad (4.2.1)$$

Let $T_{22} = (t_{22}^1, \dots, t_{22}^a)$ and $T_{31} = (t_{31}^1, \dots, t_{31}^a)$.

2. Compute X_2 - and X_3 -loadings (P_2 and P_3) by projection onto orthogonal column matrices T_{22} and T_{31} :

$$P_2 = \left(X_2^0 \right)^T T_{22} \left(T_{22}^T T_{22} \right)^{-1}, P_3 = X_3^0 T_{31} \left(T_{31}^T T_{31} \right)^{-1} \quad (4.2.2)$$

The kernel loadings matrix for X_1 , D ($a \times a$), is defined as

$$D = \left(T_{22}^T T_{22} \right)^{-1} T_{22}^T X_1^{00} T_{31} \left(T_{31}^T T_{31} \right)^{-1} \quad (4.2.3)$$

3. Decompose the data matrices by the contribution of the scores identified to form residual matrices:

$$X_1^{aa} = X_1^{00} - T_{22} D T_{31}^T \quad (4.2.4)$$

$$X_2^a = X_2^0 - T_{22} P_2^T, X_3^a = X_3^0 - P_3 T_{31}^T \quad (4.2.5)$$

The double-centered response matrix X_1^{00} is expressed in terms of the latent components and the kernel loadings matrix D :

$$X_1^{00} = T_{22} D T_{31}^T + E_1^A \quad (4.2.6)$$

where the E -matrix contains the residual variation in the observed variables that is not accounted for by the orthogonal latent variables in T_{22} and T_{31} .

Alternatively, the model for X_1^{00} may be expressed in terms of the original variables:

$$X_1^{00} = X_2^0 C X_3^0 + E_1^A \quad (4.2.7)$$

where C is a ($J \times L$) regression coefficient matrix estimated by

$$C = V_1 D V_3^T \quad (4.2.8)$$

where $V_1 = T_{21} \left(P_1^T T_{21} \right)^{-1}$, $V_3 = T_{32} \left(P_{31}^T T_{32} \right)^{-1}$.

The LPLS analysis was implemented by a program developed by Dr. Solve Sæbø (Norwegian University of Life Science). The program was implemented as the R package on a Linux machine.

4.2.4 Atom Colorings Derived from Regression Coefficient Matrix

The regression coefficient matrix (C) obtained from the LPLS model provides useful information on how fingerprints of chemical compounds are related to the z -scales of the amino acids of aminergic GPCRs. The original numerical format is of limited utility; thus the regression coefficient matrix was transformed into atom colorings used in Bayesian analysis of CYP3A4 substrate/nonsubstrate classification [18].

Analogous to the Bayesian classification [18], the atom score was derived from the regression coefficient of each substructure. The regression coefficient of each ECFP_6 substructure was divided by the number of heavy atoms present in the substructure, and the resulting score value was assigned to each atom. To determine the average atom scores of the test molecules, substructures of the test molecules were identified, and the respective scores of atoms in the substructures were summed up and divided by the number of occurrences of the atoms. Of the three z -scales constructed in this study, only the z_1 -scale, which represents the largest component, was considered for atom coloring [14]. As the hydrophobicity of amino acid increased, the z_1 -scale became increasingly negative. Thus, in the atom-coloring visualization, negative values of the regression coefficient matrix alone were highlighted (blue coloring).

4.3 RESULTS AND DISCUSSION

4.3.1 LPLS Analysis

Preceding the LPLS analysis, the columns of X_2 and the rows of X_3 are typically centered (X_2^0 and X_3^0). The corner matrix X_1 is subject to double centering across both rows and columns (X_1^{00}). The 10-fold cross-validation method was used to determine the number of significant LPLS components [17]. Twelve significant LPLS components ($A=12$) so obtained explained 63.8% of the variance in X_1 . The LPLS model explained 48.3% and 99.6% of variances in X_2 and X_3 , respectively. The corresponding root mean square error of the prediction value was 0.778.

From the established LPLS model, the regression coefficient matrix (C) was obtained. The heat map of z_1 -scales derived from the regression coefficient matrix is shown in Figure 4.2. The X-axis and Y-axis represent the 149 ECFP_6 and the 27 z_1 -scales, respectively. Red and blue indicate positive and negative regression coefficients, respectively. Because strongly hydrophobic amino acids are characterized by large negative values on the z_1 -scale, we focus on the blue areas of the heat map.

4.3.2 Atom Colorings and Support by Molecular Modeling

The performance of the LPLS model was assessed by two validation studies. Firstly, we investigated a specific 5HT_{2C} GPCR inhibitor (GVK_ID: 3090943). The atom colorings of amino acid residues 26, 27, and 28 of 3090943 after docking into the 5HT_{2C} GPCR are shown in Figure 4.3a. Atoms with regression coefficients below a chosen threshold of -0.0004 are highlighted in blue font. In the case of AA_26, the blue-highlighted portion is dichlorobenzene. In contrast, the colored areas are shifted to the fused six- and five-membered ring system in AA_28. An intermediate state is observed in AA_27. Figure 4.3b shows the putative docking mode of 3090943 in the 5HT_{2C} GPCR. In Figure 4.3b, color transitions roughly match the docking pose of 3090943 in 5HT_{2C} GPCR. The 3D structure of 5HT_{2C} GPCR was modeled by homology based on the X-ray crystal structure of the human beta2 adrenoreceptor

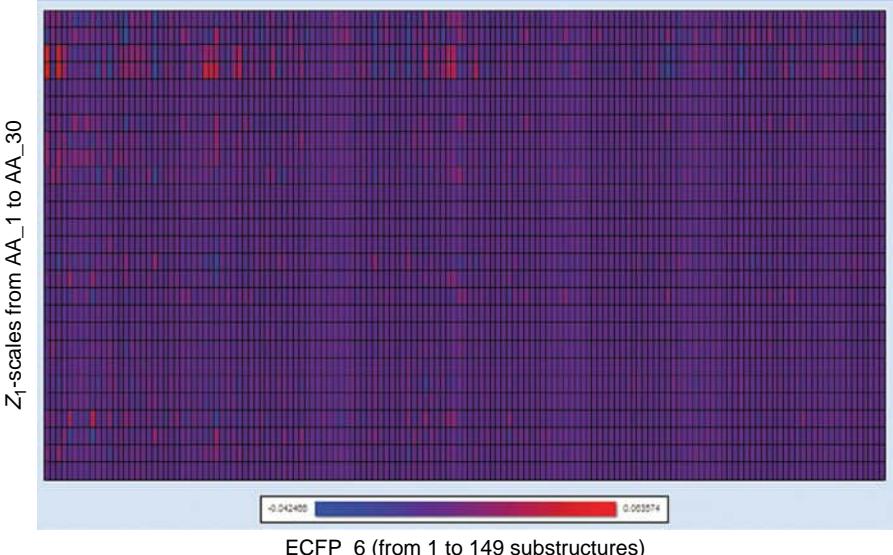


FIGURE 4.2 Heat map of z_i -scales derived from the regression coefficient matrix. For color details, please see color plate section.

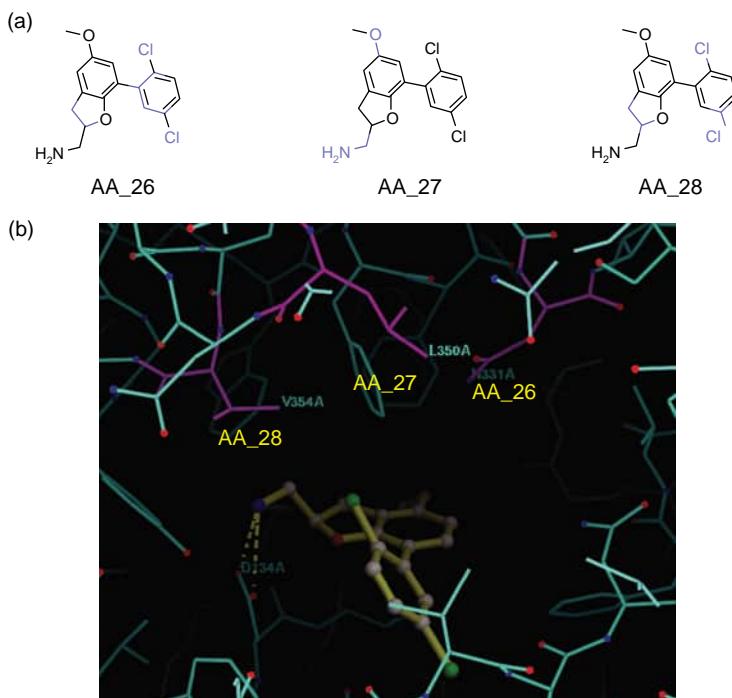


FIGURE 4.3 (a) (Left to right) Atom colorings of AA_26 (Asn), AA_27 (Leu), and AA_28 (Val) of GPCR inhibitor 3090943 docked into 5HT_{2c} GPCR. Strong hydrophobic interactions between each amino acid residue and the GPCR are highlighted in blue. (b) Putative docking mode of 3090943 in the homology model of 5HT_{2c} GPCR. Reprinted from Ref. [19]. For color details, please see color plate section.

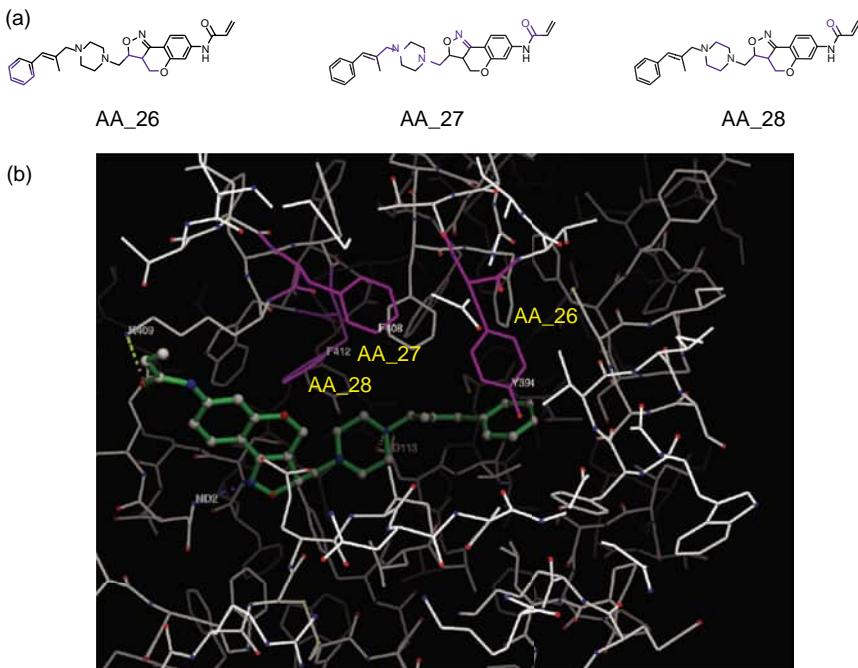


FIGURE 4.4 (a) (Left to right) Atom colorings of AA_26 (Tyr), AA_27 (Phe), and AA_28 (Phe) of GPCR inhibitor 3844318 docked into A_{2A} GPCR. Strong hydrophobic interactions between each amino acid residue and the GPCR are highlighted in blue. (b) Putative docking mode of 3844318 in the X-ray crystal structure of A_{2A} GPCR. For color details, please see color plate section.

(PDB_code: 2R4S). This model was extracted from the literature [19], and docking was performed using the program “Glide” with default settings [20].

Secondly, we evaluated a specific A_{2A} GPCR inhibitor (GVK_ID: 3844318). Similar to the $5H_{2C}$ GPCR inhibitor, the atoms highlighted in blue are consistent with the putative docking mode of 3844318 in the A_{2A} GPCR. The X-ray structure of the A_{2A} GPCR (PDB_code: 3EML) was used as a template for docking of 3844318 using “Glide” with default settings [20] (Figure 4.4).

4.4 CONCLUSION

In this study, we applied LPLS analysis to the inhibitory activity data of 16 aminergic GPCRs based on their ligand and protein descriptors. The ECFP_6 fingerprints were used as ligand descriptors to represent specific fragments of a molecule. The protein descriptors were z -scales of amino acids forming the cavity of aminergic GPCRs. From the resulting LPLS model, a regression coefficient matrix, describing ligand–protein interactions, was determined. The regression coefficient matrix was transformed into atom colorings, from which the fragments that interact with specific

amino acid residues in GPCRs could be readily identified. The utility of this approach was confirmed by two validation studies on ligand docking modes identified in 3D structures of aminergic GPCR proteins. This study, being the first application of LPLS to QSAR or chemogenomics, provides a significant scientific advance.

Specifically, we applied LPLS modeling to ligand–protein interactions using atom colorings derived from the regression coefficient matrix. The original PLS design is well suited for graphical inspection of major patterns of covariation in data tables (loading or score plots). This is also the case for the LPLS method. It is important to acquire both a detailed interpretation of the patterns within each of the three data tables and an integrated overview of how the patterns within the tables are related. From the integrated overview, we aim to derive loading and score plots, which would further elucidate ligand–protein interactions. This study is underway, and the results are expected in the near future.

Another potential benefit of our approach is that data gaps may be patched using the NIPALS algorithm of LPLS. In chemogenomics, missing data are the norm than the exception. In this study, prior to LPLS analysis, we filled missing data based on the prediction of each PLS model. Via the NIPALS algorithm, the secondary score vector t_2 can be computed from the incomplete data matrix X and the initial score vector t_1 during the training phase. That is, t_2 is estimated by projection onto nonmissing data in t_1 regressed against the nonmissing descriptors in X . This approach represents a new direction for chemogenomics studies and should be further pursued.

4.5 ACKNOWLEDGMENTS

We would like to thank GVK Biosciences, India, for allowing the use of GVK DB under an academic license. We would also like to thank Dr. Solve Saebo at the Norwegian University of Life Science for providing the R package implementation of LPLS.

REFERENCES

1. Gedeck P, Lewis RA. Exploiting QSAR models in lead optimization. *Curr Opin Drug Discov Dev* 2008;11:569–575.
2. Yap CW, Li H, Ji ZL, et al. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini Rev Med Chem* 2007;7:1097–1107.
3. Hasegawa K, Funatsu K. Advanced PLS techniques in chemometrics and their applications to molecular design. In: Lodhi H, Yamanishi Y, editors. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. Hershey: IGI Publishing; 2011. p 145–168.
4. Gasteiger J, Engel T. *Chemoinformatics*. Weinheim: Wiley-VCH; 2003.
5. Hasegawa K, Funatsu K. Partial least squares modeling and genetic algorithm optimization in quantitative structure-activity relationships. *SAR QSAR Environ Res* 2000;11:189–209.

6. Ho RL, Lieu CA. Systems biology: An evolving approach in drug discovery and development. *Drugs RD* 2008;9:203–216.
7. Wold S, Trygg J, Berglund A, et al. Some recent developments in PLS modeling. *Chemom Intell Lab Syst* 2001;58:131–150.
8. Saebo S, Martens M, Martens H. Three-block data modeling by endo- and exo-LPLS regression. In: Vinzi VE, Chin WM, Henseler J, et al., editors. *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Heidelberg: Springer; 2010. p 359–379.
9. Mahe P, Vert J-P. Virtual screening with support vector machines and structure kernels. *Comb Chem HTS* 2009;12:409–423.
10. Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;24:232–240.
11. <http://www.gvkbio.com/informatics.html>. Accessed 2013 May 14.
12. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;50:742–754.
13. <http://accelrys.co.jp/>. Accessed 2012 May 14.
14. Hellberg S, Sjostrom M, Skagerberg B, et al. Peptide quantitative structure-activity relationships: A multivariate approach. *J Med Chem* 1987;30:1126–1135.
15. Surgand J-S, Rodrigo J, Kellenberger E, et al. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* 2006;62:509–538.
16. Martens H, Anderssen E, Flatberg A, et al. Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR. *Comput Stat Data Anal* 2005;48:103–123.
17. Wold S, Albano C, Dun WJ, III, et al. Pattern recognition: Finding and using patterns in multivariate data. In: Martens H, Russwurm H, Jr., editors *Food Research and Data Analysis*. London: Applied Science Publishers; 1983. p 147–188.
18. Hasegawa K, Funatsu K. Bayesian classification of cytochrome P450 3A4 substrates/non-substrates and color mapping for chemical interpretation. *J Comput Aided Chem* 2010;11:19–24.
19. McRobb FM, Capuano B, Crosby IT, et al. Homology modeling and docking evaluation of aminergic G protein-coupled receptors. *J Chem Inf Model* 2010;50:626–637.
20. <http://www.schrodinger.com/productsguide/>. Accessed 2013 May 14.

CHAPTER 5

STABILITY IN MOLECULAR FINGERPRINT COMPARISON

ANTHONY NICHOLLS and BRIAN KELLEY

5.1 INTRODUCTION

Molecular similarity is unarguably the most useful concept in molecular modeling. If we could apply physical theory, for example, Newton's laws of motion, to accurately calculate properties, we need to advance drug discovery, for example, affinity, solubility, stability, toxicity, and so on. However, the complexity and uncertainty of applying "hard" science to molecular design is well known [1]. As such, the use of similarity within our field is widespread and profound. Broadly speaking, there are two types of use: structure–activity relationships (SARs) and quantitative structure–activity relationships (QSARs) [2]. SAR relies on a *fixed* method of molecular similarity that attempts to lead from one or more molecules with a particular property to other molecules that might also have this property and is essentially qualitative, although the degree with which similar behavior may be expected can be quantified [3, 4]. QSAR methods create new, *flexible*, similarity measures, for example, by linear regression over a set of properties to create a weighted sum of such [2], or by variable selection such that locality in parameter space reflects locality in activity space (e.g., kNN QSAR [5]). QSAR attempts to quantitatively assign or predict a particular property, typically by training over a set of molecules of known activity, whereas SAR seeks to find molecules likely to be of similar activity. While QSAR can be more informative, it also relies on having substantial data, whereas SAR can make do with a single compound. Also, QSAR can be subject to over-parameterization, that is, overtraining, that can be hard to spot. Both methods suffer from limitations of domain, that is, if a test molecule is unlike any in the training set, neither fixed nor flexible, similarity measures will be very effective. This division is not hard and fast; in fact, a case can be made [6] that the type of similarity measure used ought to be chosen based on the nature of the property under investigation, that is, fixed can also be somewhat flexible!

In this study, we shall be concerned with fixed similarity measures, that is, can we take a definition of molecular similarity and use it to find molecules of analogous behavior. The most common of such fixed methods are “fingerprints,” that is, deconstructions of molecules into either arrays of bits, signifying the presence or absence of a feature, or values, for instance molecular properties. As such, the field is now vast [7–9], we shall concentrate on three types of bitwise fingerprints, specifically path-based, feature-based, and lexicographic. Path-based fingerprints come in many varieties, but the three types considered here are a linear path-based method, akin to those from Daylight Chemical Information Systems, a circular path-based method, akin to the Extended Connectivity Fingerprints from Scitegic [10], and a tree-based method of our own construction. Our example of a feature-based fingerprint is an implementation of the 166 feature MACCS key fingerprint [11]. Finally, we examine the LINGO fingerprint method [12] that is lexicographic in nature.

Although fingerprints are abstractions of the underlying nature of a molecule, a nature that is essentially 3D, such “2D” methods are very powerful, particularly for four reasons. First, they require only a single representation of a molecule, whereas most molecules adopt multiple 3D conformations. Secondly, they are very fast to compare. Optimization of the process of comparison [13–15] has produced methods that can match millions of fingerprints per second with everyday computers. Thirdly, the representations can be very compact; in fact, the LINGO method only requires a simplified molecular input line entry system (SMILES) [16] representation. Finally, there are simple and intuitive measures that arise from comparing such strings. For instance, the ubiquitous Tanimoto, which, for a bitwise representation, is the number of bits in common divided by the total number of bits, goes from 0 for completely dissimilar molecules to 1 for essentially identical molecules. The list of other possible measures, for example, Cosine, Dice, Tversky, and so on, is long, but in this work we shall stick mostly with the Tanimoto both because of its widespread usage and because it elegantly folds in aspects of both the difference and similarity of two molecules.

One of the assumed downsides to 2D methods is that they are less likely to provide “jumps” in *chemical* similarity—a facility useful in avoiding patents or developing alternate pharmacology. However, even though 2D methods are designed to encapsulate *chemical* similarity, such similarity is often coarse enough to still retrieve molecules that satisfy intellectual property purposes. This is, in part, because patent law has a very vague notion of chemical similarity—some molecules may appear very similar to a chemist or his or her 2D metric, yet be considered “nontrivial” by the patent office. But it is also because, just as computer programs can now produce nonobvious winning moves in chess, 2D methods can capture aspects of similarity that are nonobvious to a chemist or patent examiner. As such, 2D methods can sometimes make molecules that seem different “to the eye” to be rated as similar. 2D methods are all imperfect casts at capturing the underlying physics of molecular character and each, typically, captures different aspects of this information. As such, a part of the utility of such similarity measures is actually their *imprecision*.

However, there is also an inevitable downside to the imperfections of 2D methods—one of which is *fragility*. A small change in a molecule can sometimes lead to a large change in its fingerprint representation. An illustration is provided in Figure 5.1.

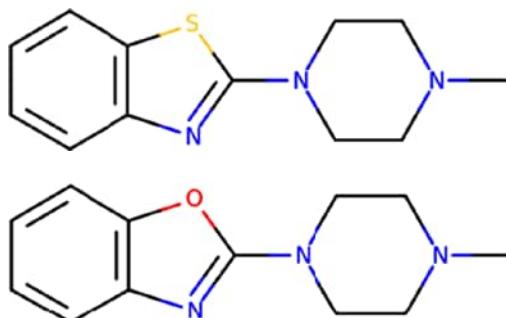


FIGURE 5.1 The structure on the left (CC1=CC=C2=C1C(=N1CCCC1)N(CCCC1)C2=CS) is sterically and electrostatically similar to the structure on the right (CC1=CC=C2=C1C(=N1CCCC1)N(CCCC1)C2=O), yet the Daylight Tanimoto between the two structures is only 0.44, well below the threshold typically expected for molecules likely to have similar activity. As explored in the chapter, atoms that are contained in many paths lead to sensitivity to small changes, more so for smaller molecules. For color details, please see color plate section.

A small, relatively conservative change taking the benzoxazole moiety to a benzothiazole causes the Daylight similarity (path-based) to crash from 1.0 (exact similarity) to 0.44. It has been shown that a Daylight similarity of above 0.85 gives a chance of similar affinity of about 30% [3], and so a value of 0.44 is substantially below what might be expected for a conservative change often classified as “bioisosteric.” Fragile events occur when two molecules are reported as being dissimilar when to the eye they are not. This, then, gives a false negative aspect to any kind of virtual screening—molecules that ought to have been included as being likely active are instead regarded as too different to be so. In addition, they represent a confusing aspect of any SAR analysis, that is, molecules with similar activities but apparently distant similarity. Also, 2D fragility interferes with effective clustering of structures, for instance, in constructing representative sets, and singleton evaluation in high-throughput screening (HTS)/screening analysis. There is also a converse problem: *over-robustness*. We expect two molecules that look very different will be scored as different; however, it is possible that a measure may cast two very different molecules as similar, that is, do not reflect large-scale differences. This is probably an under-appreciated aspect of similarity by fingerprints and leads to false positives in any kind of screen, that is, suggesting quite different molecules are likely to be active.

In this chapter, we shall examine these two concepts. In doing so, we are constrained by the lack of a rigorous definition of similarity. Initially we had hoped to use diverse metrics such that outliers from a predicted “average” similarity would serve our purpose. However, as has been noted [17], methods are sufficiently poorly correlated as to make this approach unworkable. In fact, the lack of correlation led us to examine, as have others [18–20], fusing of information from different measures in virtual screening, and the implications of fingerprint stability on this effort. To overcome the issue of a “gold standard,” we turned to an approach that generates isosteres, that is, molecules with the same graph but different composition. For this we used the program WABE [21].

An elegant aspect of WABE is that one can precisely count the number of changes in going from one molecule to another and hence have something closer to an edit distance congruent with visual inspection. With this measure in hand, we are better able to see both qualitative and quantitative aspects of fingerprint stability and to suggest approaches to avoid or take advantage of such in practical modeling.

5.2 METHODS

5.2.1 2D Methods

Here, we will concern ourselves with five well-established 2D methods: MACCS keys, a path-based fingerprint, a version of the circular fingerprints originally developed by Scitegic, a variant we call “tree-based,” and finally the LINGO measure that evaluates the lexicographic similarity between two SMILES strings. All methods are as implemented in the “GraphSim” toolkit from OpenEye Scientific Software, Inc.

MACCS keys are a feature-based fingerprint, originally from the company MDL Information Systems, where a bit is turned on if a given chemical grouping is present. These feature sets were originally proprietary; however, a reduced set of 166 features, which include such things as the presence of large or small rings, different types of bond, nature of rings, and so on, were published and these are straightforward to detect using standard chemical processing. There is a degree of information “folding” in a MACCS keys fingerprint, that is, if a molecule has multiple instances of a feature, it is still recorded just once. Thus, two fingerprints may be identical and yet the underlying molecules may be quite different. In theory, we might expect MACCS keys to be insensitive to small changes in a molecule since such changes can only add or subtract in a relatively linear manner.

Path-based fingerprints were first popularized by Daylight Chemical Information Systems. Essentially each linear path through the chemical graph of a certain length sets a bit in the fingerprint. As the number of such paths grows exponentially with the length of the path, such bit patterns are then typically (but not always) “folded,” that is, condensed into a bit pattern of a given, fixed length, for instance 1024 bits. As with MACCS keys, two different molecules can, in theory, produce the same fingerprint. However, clever hashing algorithms, that is, mappings from paths to bits turned on, can make this relatively unlikely. Path-based fingerprints can be susceptible to fragility because if many of the paths traverse a single atom or groups of atoms, then alterations to this group may cause a disproportionate number of changes to the hashed fingerprint.

Circular fingerprints take the idea that rather than taking linear paths across the molecule, circular paths radiate out from each atom. Each circle goes a certain number of bond connections, and then each “fragment” is assigned an integer. These integers can then be kept as a list or, as with linear path fingerprints, hashed down to a bit in a string of given length. Circular fingerprints come closer to capturing local substructure information. Fragility in circular fingerprints could arise if an atom, or small group of atoms, impinges on many different “circles.” This might be the case if these differences are centrally located in a relatively compact molecular structure.

Tree-based fingerprints are a variant of circular- and path-based fingerprints where, rather than insisting that branched paths are equidistant, as with circular fingerprints, tree fingerprints look at all subgraphs, branched or not, of a given number of atoms. These then capture local substructures but are not restricted to having a central core atom. Tree fragments are then be hashed or stored as lists. As with path and circular fingerprints, we chose to examine the behavior of the more typical, fixed-length fingerprints. Tree fingerprints may be sensitive to small changes if many trees cover the same small group of atoms—possibly the case if some atoms are “highly connected.”

The *LINGO* method is a lexicographical approach where a SMILES string is decomposed into contiguous sets of characters, for example, the first four characters in the SMILES, then the second to the fifth, the third to the sixth, and so on until the last character is the terminating character. The comparison is made, then, between each molecule’s set of character subsets, incrementing a count for each match. A Tanimoto is formed by dividing this count by the counts from applying this method to each molecule with itself, minus the comparison count. The LINGO method seems absurdly simple compared to the extensive use of chemical graphs in the other methods, and yet it often performs as well. It has an obvious potential fragility in that SMILES strings for similar molecules may look quite different, for instance if the atom chosen as the first, or “root,” in the SMILES string differs. To some extent canonicalization can improve this situation, but not always, for example the first atom in a canonicalization may be different in two slightly different molecules.

5.2.2 Generation of Molecular Isosteres: WABE

WABE is a program that combines single heavy-atom replacements within a molecule to produce new molecules with the same underlying connection table, that is, where each heavy atom forms an equivalent node in a graph and each bond an edge between nodes. The hybridization state at each node remains unchanged. An example of WABE output as applied to aspirin is shown in Figure 5.2.

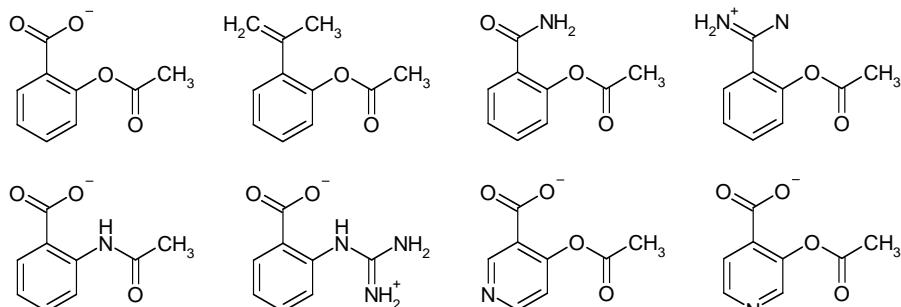


FIGURE 5.2 Examples of variants of aspirin (top left) produced by WABE, wherein the atomic composition is varied but the (uncolored) graph structure is maintained.

Molecules produced by WABE have essentially the same volume, up to small changes from element differences, and the same shape, differing only in their distribution of electrostatic potential, which can be significant from a pharmacological viewpoint. Here, the advantage is that WABE can generate large numbers of structures and that the differences between WABE isosteres can be directly enumerated in terms of the number of edits.

5.2.3 Tanimoto and Significance

Most similarity measures are presented in the form of a “Tanimoto.” A Tanimoto can be defined between any pair of ordered set of numbers as the inner product of those sets divided by the sum of the self-products minus that inner product. When all numbers are all greater than or equal to 0, the Tanimoto ranges from 0 to 1.0, with 1.0 representing identity, making it a natural concept for similarity. Its advantages are that it folds in both the features in common and any difference in size, that is, a molecule with few features all of which are shared with a molecule with many features is not evaluated as similar. This can also be seen as a disadvantage, that is, where such “partial” similarity is important alternate measures as Tversky are popular. It also has the character of being less discriminating of dissimilarity—molecules have many more ways to have a Tanimoto near 0 as near 1.0. Finally, different methods may have quite different expected values of Tanimoto, that is, a Tanimoto of 0.4 may mean two molecules are completely unrelated by MACCS keys or similar by circular fingerprints.

As such, it can be useful to transform a Tanimoto to a significance, or “ z -score,” that is, to subtract from a similarity measure the expected mean of that measure over pairs of randomly chosen molecules, and to divide the result by the standard deviation of such numbers. It should be noted that as the transformation from Tanimoto to z -score is linear all correlations between measures subject to such scaling are unchanged. As shown by others [22], distributions of Tanimoto values are not particularly Gaussian, in particular, in their long-tail behavior; yet such a scaling is still useful, especially near mean values where the distribution can be Gaussian-like. A further criticism would be that the nature of the distribution depends on the nature of the query—for instance, as has been noted [22–24] both the mean and the character of the distribution depends on the size of the molecule. In the work here on isosteres, this will be of less importance.

Table 5.1 shows the mean and standard deviations for the 2D measures employed here by looking at the distribution of similarity scores for 100,000 randomly chosen pairs of molecules from the ZINC database [25].

TABLE 5.1 Expectation Values and Standard Deviations of the Five Fingerprint Methods Used in This Study, Extracted from 100,000 Randomly Chosen Pairs from the ZINC Database

Method	MACCS166	LINGO	Path	Tree	Circular
Mean	0.381	0.118	0.095	0.096	0.076
Stdev	0.121	0.069	0.044	0.049	0.027

5.3 RESULTS

With five fingerprint methods at our disposal, our initial concept was to look for outliers within the comparisons for a given pair of molecules, that is, if all the four methods suggested that a couple of molecules were similar at roughly the same level of significance yet a fifth suggested a radically different similarity, this might be an example of fragility (if the latter significance was much lower) or over-robustness (if higher). In this way, we use the fourfold consensus as a replacement for “seems similar by eye.” However, as has been commented on before [26], there is generally very little agreement between different fingerprints, that is, we usually failed to find a fourfold consensus value from which a fifth might deviate. Taken over the same set of 100,000 pairs randomly chosen from ZINC, Table 5.2 shows the square of Pearson correlation coefficient [27] between each method. Only the path and tree methods are strongly correlated, with circular method somewhat similar to path and tree. Fingerprints based on MACCS keys and LINGO are both dissimilar to each other and to the other three methods.

In general, a poor correlation between methods that each work well, that is, retrieve actives, might indicate that they are good candidates for data fusion, that is, if they carry orthogonal information. To test this hypothesis, we looked at a dataset provided by Martin and Muchmore in their paper on lead-hopping [28]. In this work, the authors looked at the ability of different approaches to find ligands that were similarly active but substantially different in structure. The concept of “different” here was a combination of the expertise of authors in choosing the systems and chemotypes, but also by using Daylight fingerprints, that is, two similarly active molecules were considered a “hop” if less than 0.7 Tanimoto. In total, the authors defined a total of 166 lead-hop pairs. Decoy hops were formed from pairs of molecules that were not in the same activity class. Although this approach to generating decoys has some issues (e.g., 3D comparison of some decoy pairs suggested some decoys might have been active if so tested), it also avoids some of the mistakes commonly made in virtual screening evaluations, such as lack of basic drug-like properties. The conclusions from the Muchmore and Martin paper were that

1. Daylight fingerprints still did quite well, despite being used as the filter for “not similar,” that is, there is considerable signal in path-based methods even in the penumbra of their signal.
2. 2D methods did better than most 3D approaches.

TABLE 5.2 R^2 Correlation Coefficients Between the Fingerprint Methods Used Here, as Calculated for the Same Set of Molecular Pairs as in Table 5.1

	LINGO	Path	Tree	Circular
MACCS166	0.11	0.21	0.21	0.2
LINGO		0.34	0.38	0.32
Path			0.87	0.49
Tree				0.52

TABLE 5.3 Enrichments (at 1%) and AUC (Area Under the ROC Curve) for the Muchmore–Martin Set of Lead-Hops

AUC/1%/Enrichment	MACCS	Path	LINGO	Circular	Tree
MACCS	0.80/12	0.8	0.81	0.82	0.82
Path	15	0.75/15.6	0.77	0.79	0.77
LINGO	18.7	16.9	0.77/21.8	0.8	0.79
Circular	18.7	18.1	21.25	0.81/20.6	0.81
Tree	15.6	16.3	18.1	20.0	0.78/16.3

The diagonal elements contain the AUC/enrichment values for each method; the off-diagonal elements contain the AUC (upper triangle) and enrichment values (lower triangle) from a merge (use the best *z*-score) pairs of measures.

We used the Muchmore and Martin dataset to test whether combinations of our diverse set of fingerprints would improve standard metrics, such as AUC, the area under the curve, for the receiver–operator characteristic curve, or the enrichment, that is, increase in the number of actives over that expected by random chance when 1% or 2% of the decoys had been discovered. Table 5.3 shows the results of using a “merge” fusion, that is, the Tanimoto that is most significant between the two measures for a given pair of compounds are used. Here the AUC is essentially unchanged, that is, higher than either method alone in six cases, lower in four, but the enrichment is poorer in 9 out of 10 cases. In fact, in the tenth case, the enrichment is identical when blending the tree- and path-based methods that are already highly correlated. This mirrors previous observations of ours with regard to the fusion of 3D types of data in virtual screening, for example, ligand-based and structure-based methods. Here, also, AUC is improved because the probability of one such method being essentially random is quite high, but the probability of two, different, methods failing is rare. However, the actual performance of combined methods does not improve exactly where one would most like it to, that is, in early performance. One explanation of this is simply that as well as protecting against “fragility,” that is, true positives being ranked unexpectedly lower, fusion introduces more chances for “over-robustness,” that is, lack of discrimination of pairs which ought to be considered different. This hypothesis was tested in our next experiment, namely, on the set of isosteres generated by the program WABE.

As our test case molecule for WABE, we took methotrexate, the well-known anti-cancer drug. This molecule was the focus of the original WABE publication, wherein it was shown that many of the observed SAR trends could be rationalized by the evaluation of a continuum theory estimation of binding affinity. The principal advantage of using this set here is that we can exactly enumerate the number of simple changes made between two molecules. Over a set of 32000 such isosteres, there were between 1 and 15 changes from the root structure of methotrexate. Some of these changes amount to tautomeric shuffling of protons, some to changes of heavy atom character (see Fig. 5.5). Of immediate interest was to calculate the Pearson correlation coefficient, or *r*-squared, between the similarities of the WABE set to methotrexate by different methods to see if these correlations were similar to those shown in Table 5.2 over random pairs from ZINC. The results are shown in Table 5.4.

TABLE 5.4 R^2 Correlation Coefficients for the Five Fingerprint Methods Over the WABE Set of 32,000 Variants of Methotrexate, with the Number of WABE Changes Treated as a Sixth Similarity Measure (But Where a Smaller Number Means More Similar)

	LINGO	Path	Tree	Circular	# WABE Changes
MACCS166	0.27	0.38	0.40	0.19	0.38
LINGO		0.24	0.24	0.22	0.44
Path			0.97	0.68	0.67
Tree				0.72	0.68
Circular					0.48

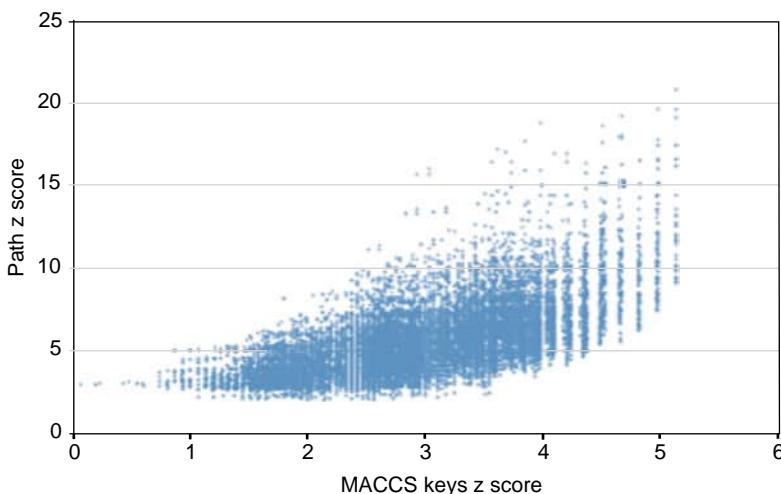


FIGURE 5.3 A plot comparing two similarity measures, MACCS keys and path-based, of methotrexate to 32,000 variants. Note that both approaches recognize that some variants are very different to methotrexate (lower z-scores), but only MACCS keys get close to 0, that is, indistinguishable from a random structure. Path-based approaches still claim significant similarity for structures MACCS keys claim have no similarity to methotrexate.

It is clear that the r -squared values in Table 5.4 follow a similar pattern to those shown in Table 5.2, namely, that the tree and path fingerprints are highly similar with some significant similarity to circular fingerprints, but that LINGO and MACCS keys give rise to relatively orthogonal representations. There are differences, however, that are worth noting, in particular how much higher the similarities are between circular, path, and tree than for the random set, MACCS Keys generally higher, while LINGO tends to be lower. It is instructive to look at one of the scatter plots between MACCS keys or LINGO and a path, tree, or circular fingerprint. Figure 5.3 shows the general trend, that is, with a MACCS key fingerprint or a LINGO similarity, it is possible to have a WABE structure become essentially “random” compared to methotrexate (i.e., the z-score goes to 0), while

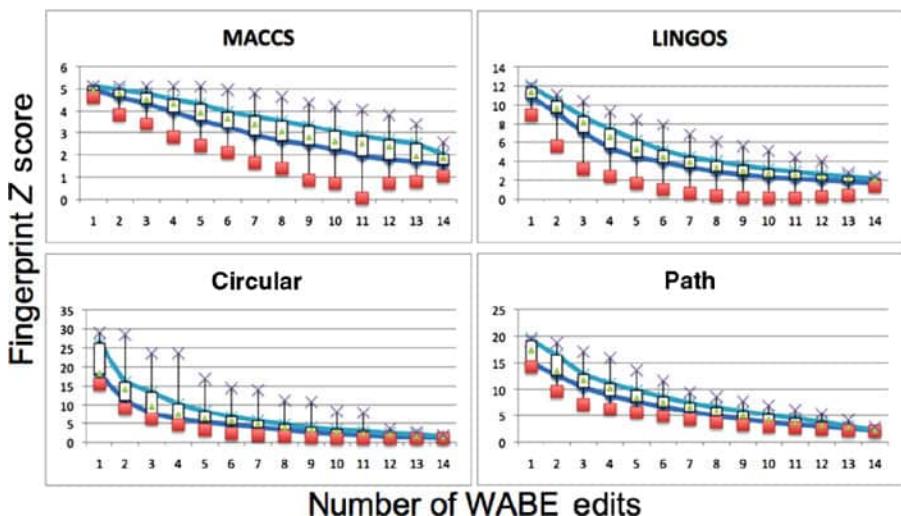


FIGURE 5.4 Graphs of the median, 95% range about the median, and maximum and minimum z -score similarity of the five methods as a function of the number of WABE edits. For color details, please see color plate section.

for circular, path, and tree the lowest values are 0.95, 2.4, and 2.35, respectively. What this means is that circular, path, and tree are encoding the graph of the molecule as well as its features and no matter how those features are changed the isostere does not register as a random compound compared to methotrexate. We suspect this is how most chemists might see matters, that is, their eyes are drawn to the frame as much as the functionalities of a molecule. Further support for this is seen in the final column of Table 5.4, wherein the correlation between each measure and the unambiguous number of WABE changes is reported. Here, a clear difference is seen between path and tree and the other measures, that is, both path and tree give quite a good correlation to the number of changes. Combined together, the reproduction of the number of edits on a molecule plus the carrying of information of the general scaffold does much to explain the popularity of path-based fingerprints as a “common-sense” fingerprint.

Having considered average behaviors, we next return to the subject of outliers. To do so, we looked to see what structures most deviated from the mean similarity for a measure, given a fixed number of WABE edits. An example of such is given in Figure 5.4.

In this figure, the median similarities are shown as a triangle within a box that indicates 95% of the structures about the median. The maximum and minimum similarities are highlighted with a cross and a square, respectively. The graph for the tree fingerprint is not shown as it adds nothing new to the graph for path-based fingerprints. These graphs are a rich source for understanding of graph similarities. First, it is very clear that not only do the path, tree, and circular fingerprints carry a residual similarity due to the graph conservation between isosteres, but the similarity of molecules with small

TABLE 5.5 Average, Maximum, and Minimum Similarity Values (z-Scores) Over a Subset of the WABE Methotrexate Set When There Are Exactly Three WABE Changes

z-Scores	MACCS166	LINGO	Path	Tree	Circular
Average	4.5	7.7	12.0	11.3	10.8
Minimum	3.4	3.2	7.2	7.2	6.2
Maximum	5.1	10.4	17.1	15.1	23.6

numbers of changes have much higher *z*-scores, that is, are deemed much more distinct from random molecules than either LINGO or, especially, MACCS keys would suggest. The difference in scale of the similarities is instructive when considering the minimum similarities (indicating fragility) or maximum similarities (indicating over-robustness), for example, while it appears MACCS keys or LINGO have a wider range of similarities, actually each are comparatively stable compared to circular or path-based. We tabulate some of these observations for the subset of isosteres with exactly three edits in Table 5.5.

As Table 5.5 shows, the actual variation for MACCS keys is very small—less than two standard deviations between the most similar and least similar. We anticipated this because a MACCS keys fingerprint only relies on feature counts and so is less likely to be fragile to small changes than a path-based or lexicographic approach. The other methods have increasingly wide ranges, with LINGO less than path, tree, and circular and circular most varied of all. An example of the variability in circular fingerprints is shown in Figure 5.5, which shows methotrexate (a), followed by its most similar structure with three changes (b), and followed by the structure most dissimilar with three changes (c). The most similar structure only changes terminal atoms, hence only a limited number of elements in the circular fingerprint are changed. The most changed structure involves the movement of a proton from the pteridine ring nitrogen to the terminal acid, but more importantly, changing the central linker nitrogen to a carbon. As many paths and circular elements include this linker atom, large changes are seen in these similarity measures.

There is a case to be made that the sensitivity of mean similarity derived from circular or path fingerprints to small changes is, in some circumstance, an advantage, that is, such fingerprints discriminate rapidly between very similar structures, something that neither LINGO nor MACCS keys can do. This may well explain their power in retrospective virtual screening evaluations, that is, since the structures typically used for such evaluations are from congeneric series, the sensitivity of path or circular may be in its favor. Conversely, such methods, in particular, circular fingerprints, seem very good at maintaining similarity even when a number of peripheral changes have been made. Again, since such virtual screening sets often contain medicinal chemistry variants that leave the core untouched, this may explain the success of such in retrospective studies.

One approach to visualize the importance of different sites on a molecule to different similarity methods is simply to replace each heavy atom in turn with a dummy atom and then calculate the similarity to the parent molecule. Figure 5.6 illustrates this for our five methods. Here, each atom is colored more densely if the

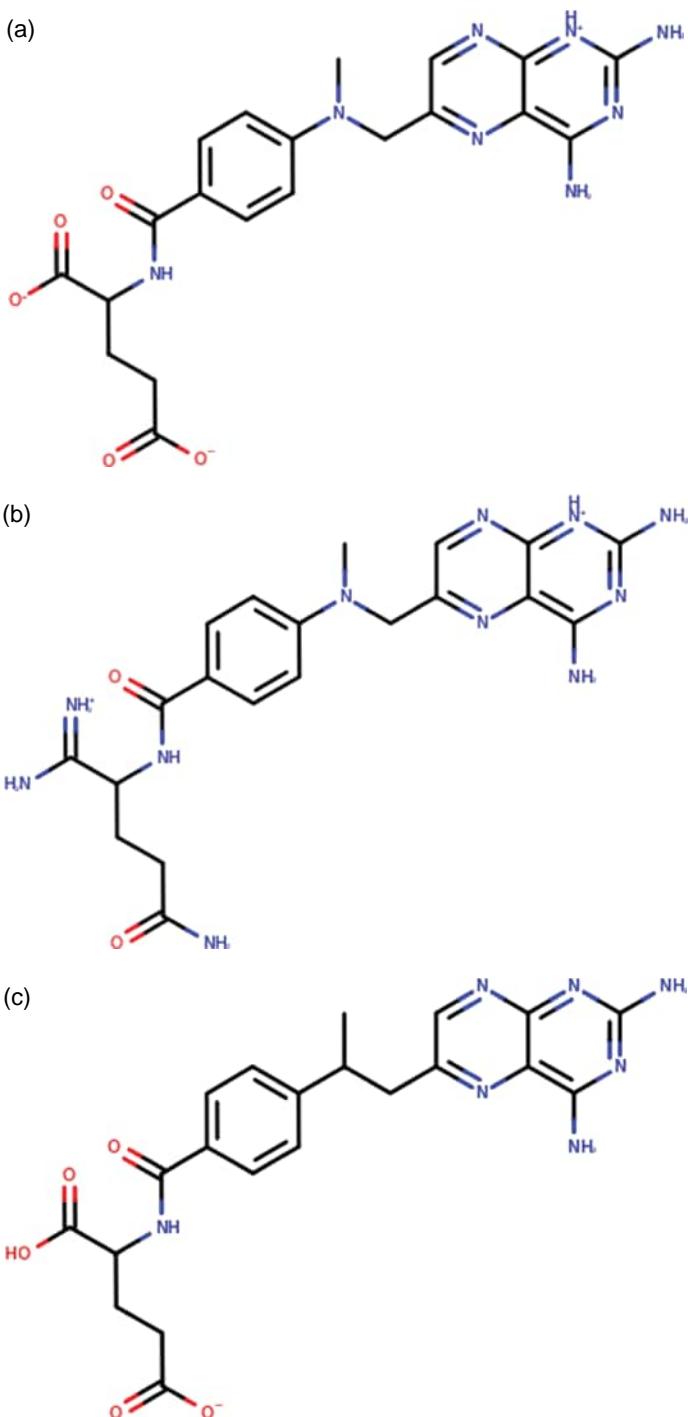


FIGURE 5.5 Structures of methotrexate (a) and the WABE variants most similar (b) and most dissimilar (c) by circular fingerprints when there are three WABE changes. Structure (c) appears very different because circular fingerprints are particularly sensitive to changes in linkers ($C \rightarrow N$) and rings ($N \rightarrow N^+$ in the pteridine ring). For color details, please see color plate section.

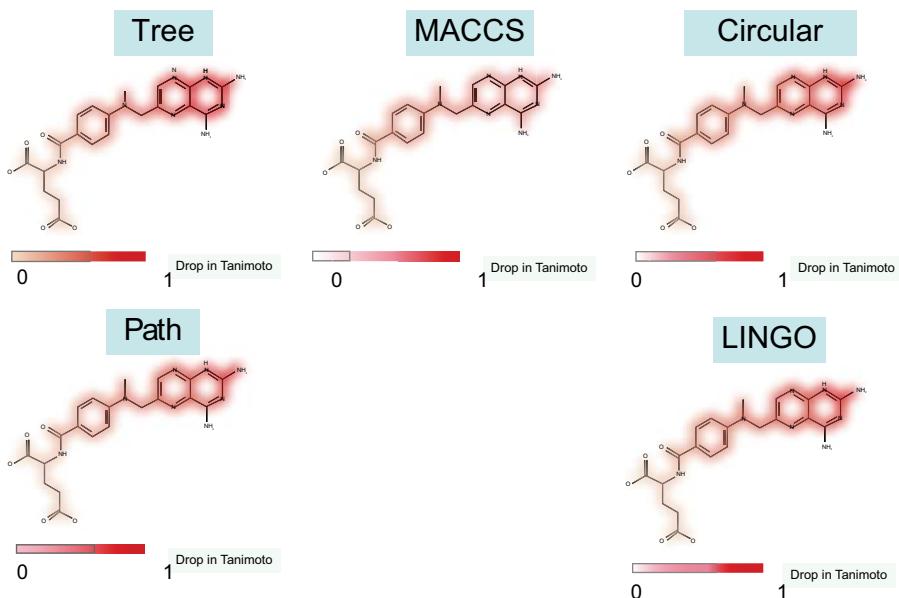


FIGURE 5.6 A graphical illustration of the similarity of methotrexate to a structure where a single atom has been changed (to a “dummy” atom). Atoms are colored more heavily the less the similarity to methotrexate. This figure illustrates many of the conclusions found in this chapter as to the sensitivity (or lack) of different fingerprint methods. For color details, please see color plate section.

similarity measure is more sensitive. As expected, the MACCS key method is relatively unaffected by which atom is changed, whereas the path-based method, and also the LINGO method, shows more sensitivity in the rings and linkers. The circular fingerprint is particularly sensitive to changes in the linker atoms between the pteridine ring and the rest of the molecule, as seen in the example shown in Figure 5.5.

5.4 CONCLUSIONS AND DIRECTIONS

Our work here suggests some recommendations for the use of fingerprints in practical applications of molecular similarity.

1. Fingerprint methods are pretty uncorrelated. This is a disadvantage in some respect because the field of molecular modeling could use a more canonical measure of similarity and has been criticized as such [26]. However, it can also be seen as an advantage. Fusing methods based on significance scores do not appear to improve performance on traditional retrospective tests, at least

not in terms of early enrichment, but it is likely to give quite different structures at the top of the list. In practice, we have observed that successful modeling is often not related to the accuracy of a particular method but to the ability of modelers or medicinal chemists to improvise. As such, we would recommend at the very least looking at similarity lists from a path-based method (path, tree, or circular) and a feature or lexicographic method (MACCS keys or LINGO), ideally from three methods, either both a path/tree and a circular method, and a feature-based method, or a path method and two feature-based methods.

2. A practitioner should be aware of the different expected values of a Tanimoto. Path-based fingerprints, because of their popularization by Daylight, have come to define a range of values likely to mean physical similarity, yet this range is quite different for other measures.
3. The position of changes can matter within a molecule. In general, both path and circular fingerprints are sensitive to changes in linker regions and rings, whereas feature-based methods are largely refractory to such changes.
4. Lacking a clear edit distance between two molecules, the path-based method appears, at least in this study, to give the strongest, average correspondence to the number of changes a chemist might notice. This, in addition to the latent weight given to the scaffold in path-based and circular methods, likely explains their popularity with chemists.
5. Circular fingerprints appear most sensitive to small numbers of changes, while at the same time maintaining similarity despite noncore, peripheral changes.

Our analysis does suggest directions for the development of perhaps more stable fingerprints. Firstly, one of the reasons fingerprints became so popular is that they enabled searching of large corporate databases in “real time,” that is, in few seconds. This was not a trivial advance; algorithmic performance that allows a user to maintain a conscious connection and commitment to the task at hand has always altered workflows and practices—for instance, the rendering speeds for tessellated surfaces meant that researchers moved away from ball and stick representations to molecular interfaces, as exemplified in the GRASP visualization program [29]. However, computation speeds have continued to advance, either by clock speed or by parallel execution. Getting back 2D similarities to a database in 0.1 seconds, rather than 1 second, is not a practical improvement, although these speeds may be, of course, useful in clustering large datasets. Perhaps what the field actually needs are *slower* methods, slower but richer in information. An example of this might be to edit distance metrics between two molecules, that is, just what changes are required to transform molecule A to molecule B. In our WABE example, this was easy to compute because the graph remained the same. However, general approaches to graph similarity via graph edit distances are possible, for instance, using the “Hungarian” algorithm for graph assignment [30], and feasible given today’s computing resources. Another example might be a similarity score based on maximum common substructure, yet another

might be to compare string representations, as with LINGO, but where the starting atom in the representation is permuted, that is, find the minimum distance between many representations of structure A and many of structure B. This gives up one of the original advantages of fingerprints, that is, a single representation; however, it removes the disadvantage of such a representation failing because of a change in ordering of the representation of another molecule.

A final conclusion from our work is that perhaps the differential stability of similarity based on single atom changes can be addressed and even factored into a better fingerprint. There have been suggestions in the past as to the value of different metrics to offset fingerprint limitations, such as the difference in complexity of fingerprints with size [26], or the variation in frequency of bit occupancy [31]; here, an obvious direction would be to weigh the component of a fingerprint such that those components that encompass a sensitive atom, or sets of atoms, are underweighted. In this way, a fingerprint similarity would change equally for any single atomic modification. This could be done “on-the-fly” for a particular structure, simply by comparing the similarity between the query molecule and one-atom changes to that molecular, as illustrated in Figure 5.6. An alternate possibility would be to overweigh certain changes—for instance, if the user actually wants to find molecules with a different core structure, it might be advisable to weigh the periphery, not the core. Finally, there is nothing to stop a user generating his/her own weighting scheme, based perhaps on known SAR or preferences for diversity in cores, linkers, or functional groups.

Clearly there is much that can still be done with simple molecular fingerprints and much that current technology increasingly enables. It is entirely possible that the most useful days of molecular fingerprints are still ahead!

REFERENCES

1. Faver JC, Benson ML, He X, et al. *J Chem Theory Comput* 2011;7:790–797.
2. Hansch C. *Acc Chem Res* 1969;2:232–239.
3. Martin YC, Kofron JL, Traphagen LM. *J Med Chem* 2002;45:4350–4358.
4. Muchmore SW, Debe DA, Metz JT, et al. *J Chem Inf Model* 2008;48:941–948.
5. Golbraikh A, Bonchev D, Tropsha AJ. *Chem Inf Comput Sci* 2001;41:147.
6. Nikolova N, Jaworska J. *QSAR & Comb Sci* 2004;22:1006–1026.
7. Willet P. *Methods Mol Biol* 2011;672:133–158.
8. Stumpfe D, Bajorath J. *Comput Mol Sci* 2011;1:260–282.
9. Johnson MA, Maggiora GM. *J Mol Struct* 1992;269:376–377.
10. Rogers D, Hahn M. *J Chem Inf Mod* 2010;50:742–754.
11. Durant JL, Leland BA, Henry DH, et al. *J Chem Inf Comput Sci* 2002;42:1273–1280.
12. Vidal D, Thormann M, Pons M. *J Chem Inf Model* 2005;45:386–393.
13. Grant JA, Haigh JA, Pickup BT, et al. *J Chem Inf Model* 2006;46:1912–1918.
14. Haque IS, Pande VS, Walters WP. *J Chem Inf Model* 2010;50:560–564.
15. Kristensen TG, Nielsen J, Pedersen CNS. *J Chem Inf Model* 2011;51:597–600.

16. Weininger D. J Am Chem Soc 1988;28:31–36.
17. Bender A, Jenkins JL, Scheiber J, et al. J Chem Inf Model 2009;49:108–119.
18. Fechner U, Schneider G. ChemBioChem 2004;5:538–540.
19. Whittle M, Gillet VJ, Willett P, et al. J Chem Inf Model 2006;46:2206–2219.
20. Hert J, Willett P, Wilton DJ, et al. J Chem Inf Model 2006;46:462–470.
21. Sayle R, Nicholls A. JCAMD 2006;20:191–208.
22. Baldi P, Nasr R. J Chem Inf Model 2010;50:1205–1222.
23. Flower D. J Chem Inf Comput Sci 1998;38:379–386.
24. Wang Y, Bajorath J. J Chem Inf Comput Sci 2008;48:75–84.
25. Irwin JJ, Sterling T, Mysinger MM, et al. J Chem Inf Model 2012;52:1757–1768.
26. Eckert H, Bajorath J. Drug Discov Today 2007;12:225–233.
27. Pearson K. Philos Mag Ser 6, 1901;2:559–572.
28. Martin YC, Muchmore S. QSAR Comb Sci 2009;28:797–801.
29. Nicholls A, Sharp KA, Honig B. Proteins Struct Funct Bioinform 1991;11:281–296.
30. Kuhn HW. Nav Res Log Quart 1955;2:83–97.
31. Arif SM, Holliday JD, Willett P. J Chem Inf Model 2010;50:1340–1349.

CHAPTER 6

CRITICAL ASSESSMENT OF VIRTUAL SCREENING FOR HIT IDENTIFICATION

DAGMAR STUMPFE and JÜRGEN BAJORATH

6.1 INTRODUCTION

Virtual (compound) screening (VS) has been defined in different ways. For example, in 1998, VS was described as “automatically evaluating very large libraries of compounds using computer programs” [1]. The identification of novel hits indeed is the primary goal of VS, analogously to experimental high-throughput screening (HTS). Over the years, many different VS methods, with rather different degrees of sophistication and computational complexity, have been introduced with the aim to further improve hit identification potential. However, many of these VS methods are probably never practically applied in the search for new active compounds, because benchmark calculations using known active compounds are extremely popular in chemoinformatics as a purely computational exercise.

Basically, VS can be divided into ligand-based virtual screening (LBVS) [2] and structure-based virtual screening (SBVS) [3]. Molecular docking is by far the most popular SBVS approach (in fact, it is currently the single most popular VS methodology). SBVS relies on the use of 3D structures of biological targets as templates for ligand docking or structure-based pharmacophore screening. On the other hand, LBVS makes use of known active compounds as reference molecules and attempts to extrapolate from them to identify new hits. Regardless of methodological details, LBVS is generally based on the principle of molecular similarity (as a hypothetical measure of activity similarity), whereas SBVS is based on the principle of molecular complementarity (between a ligand site and a binding site). LBVS approaches can generally be divided into similarity search and compound classification methods [4]. In similarity searching, database compounds are compared to a set of reference molecules and the compounds are

TABLE 6.1 Overview of Exemplary LBVS Methods

Dimensionality	Methodology
2D	Substructure searching Similarity searching
2D and/or 3D	Clustering/partitioning Mapping Distance functions Machine learning
3D	Volume/surface matching Pharmacophore searching

ranked according to their calculated similarity values to the references. Thus, similarity is established on the basis of pairwise compound comparison. By contrast, in compound classification, regardless of whether unsupervised or supervised learning methods are applied, molecules are typically distinguished on the basis of class labels (i.e., “active” or “inactive”). For compound classification, machine learning methods become increasingly popular. In addition to class label predictions on the basis of models derived from training sets, ranking of database compounds according to the likelihood of activity is also possible. LBVS can employ both 2D and 3D molecular representations. 2D representations are calculated from molecular graphs (e.g., topological or fragment fingerprints) and 3D representations from (hypothetical or known) bioactive conformations (e.g., shape queries or pharmacophore models). Exemplary LBVS approaches are given in Table 6.1.

In addition to new VS methods and benchmark investigations, hundreds of practical VS applications resulting in the identification of new hits have been reported over the past decade. These studies have addressed a variety of biological targets, mostly proteins. However, these reports of “success stories” are rather heterogeneous in nature and frequently lack proper computational and/or experimental assessment. In fact, some of these studies call into question whether or not VS should be regarded as a scientifically sound and serious exercise, considering the nature of “novel” hit compounds that are reported and the often obscure computational procedures that are applied.

In this chapter, different VS studies and their results are discussed and it is attempted to critically evaluate the opportunities and limitations of VS, with a particular focus on its practical relevance and value. A general perspective upfront: As proponents and developers of computational methods and also as VS practitioners, our view of the VS field should not come across as intentionally negative. Rather, we understand the critique as a plea for this field. However, we are convinced that it will be essential to raise the scientific standards of VS and the awareness of its limitations in order to provide a foundation for further growth of the field and for higher impact on drug discovery research.

6.2 FACTORS AFFECTING THE OUTCOME AND EVALUATION OF VIRTUAL SCREENING CAMPAIGNS

6.2.1 General Scientific Factors

Many publications are available that report practical VS applications using a broad spectrum of computational methodologies leading to the identification of active compounds. The good news is that an increasing number of such practical applications is actually reported, in addition to typical benchmark calculations; the bad news is that these “success stories” do not provide a realistic view of the state of the art in the VS field, at least for two reasons. First and foremost, only successful applications are reported, but not the many failures that outnumber the successes. Currently, there is no serious scientific forum available for the communication of unsuccessful computational studies, although one would like to rigorously investigate why methods fail, rather than succeed, and focus on their scientific limitations; a situation that plagues many different scientific fields beyond computational chemistry and chemoinformatics. Second, many VS applications are carried out in pharmaceutical environments and are not published. As a consequence, a disproportionately large number of academic applications are available in the scientific literature. These applications are often fairly remote from drug discovery research. Compared to practical applications, many more theoretical VS benchmark studies are available that are also frequently communicated by groups from industry. We will not consider theoretical benchmark investigations for the purpose of our discussion, because it is in general difficult to draw firm conclusions from these studies. For example, the strong compound class-dependence of VS calculations and the strong influence of different molecular representations on the results have repeatedly been pointed out [5, 6], yet little progress has been made over the past decade to tackle these problems. Awareness of these problematic issues is often limited to experts in the field. In addition, benchmark calculations are generally artificial in nature and consistently overestimate the performance of VS methods. Clearly, benchmarking is a requirement (if not “necessary evil”) for method development and assessment, but the information content of these calculations is generally limited. To be a bit provocative, the quality of benchmark studies we continue to observe in the literature ranges from informative over boring and bewildering to scientifically meaningless (with many studies falling into the latter categories). Moreover, there are also general limitations of VS that go beyond benchmarking. For example, methodological complexity does not scale with VS success rates and different approaches and calculation protocols often produce only very little or no overlap in compound rankings [7], reflecting again the strong case dependence of VS. Furthermore, LBVS methods generally rely on the principles of molecular similarity [8], but there are no generally applicable quantitative relationships between calculated molecular and observed biological similarity of compounds. As a consequence, compound selection from LBVS rankings is heavily influenced by knowledge and subjective criteria, similar to the situation with docking.

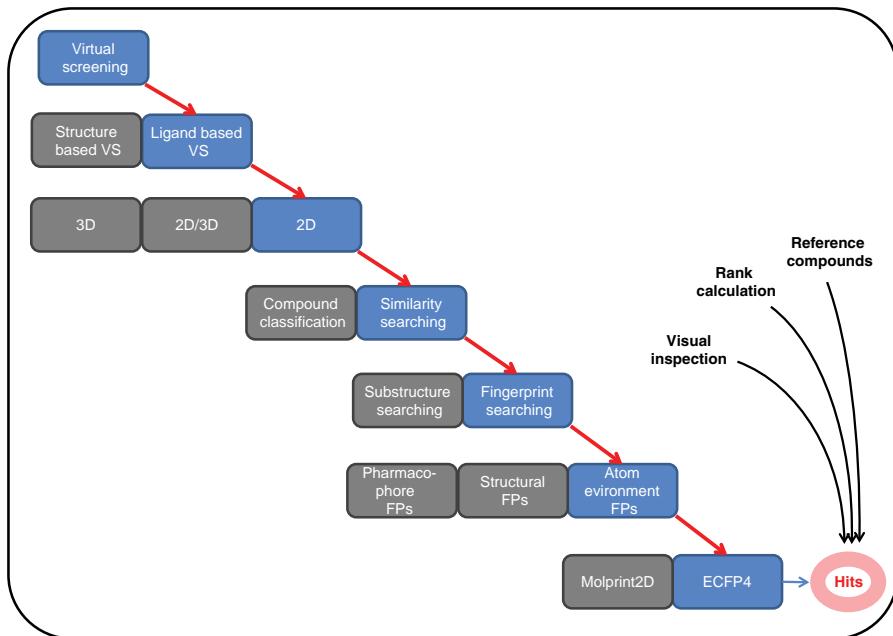


FIGURE 6.1 Exemplary workflow of a practical VS project. A possible VS application is illustrated. Different methodological choices can be made. Additional factors such as visual inspection, rank calculation, and the choice of reference compounds also strongly influence final compound selections. For color details, please see color plate section.

Figure 6.1 outlines a prototypic VS application. The blue boxes represent chosen approaches and the gray boxes possible alternatives. Changes in methodology will most likely affect the results of a VS effort, resulting in typically very little overlap of compound rankings or classifications produced by different methods, as discussed earlier. In addition, differences in the choice of reference compounds, applied calculation protocols, or ranking schemes (e.g., data fusion) [9] will affect the results. Simply put, there are no generally preferred molecular representations, algorithms, or candidate compound selection methods, and it is usually not possible to predict which descriptors or approaches are best suited for a given class of active compounds or a target. We reemphasize that knowledge and intuition play an important role in VS and the selection of candidate compounds [10], given the generally limited sensitivity and specificity of the calculations and the lack of defined relationships between molecular and activity similarity, which complicates predictions.

6.2.2 Characteristics of Practical Applications

Being aware of general factors that limit success rates of VS and a realistic assessment, it is well worth considering the scientific quality of prospective VS applications in greater detail. We have recently collected 250 SBVS and 115 LBVS

applications from the scientific literature and analyzed them in detail [10, 11]. It was determined that at least 10% of published LBVS and SBVS studies had substantial limitations. Often computational and/or experimental procedures were described in insufficient detail and the studies were not reproducible. Furthermore, candidate compounds were frequently evaluated in inappropriate assays that did not rigorously test the VS target and/or lacked required controls. In addition, many studies had minor shortcomings with respect to one or another evaluation criterion. Of course, practical VS applications strongly influence the perception of VS in the scientific community and the current lack of rigorously enforced standards for publication of practical VS studies presents a significant problem. As a consequence, first attempts have been made by popular journals in the field including the *Journal of Medicinal Chemistry* and the *Journal of Chemical Information and Modeling* to clearly formulate acceptance criteria for VS, raise the bar, and synchronize standards across journals [12]. These steps only represent the beginning of establishing scientific quality criteria and publication standards that ultimately need to be consistently applied in this field.

6.3 HOW TO EVALUATE VIRTUAL SCREENING PERFORMANCE?

What are the important criteria for VS success? The number of new hits? The potency of hits? The number of new scaffolds among hits? In practical applications, the number of identified hits or the corresponding hit rate might be used to judge VS performance. The hit rate is calculated by dividing the number of identified hits through the number of tested compounds. The number of candidate compounds selected for experimental evaluation often varies significantly. Especially for small compound numbers, hit rates are not very informative. For example, is there a significant difference if one new active compound is identified by testing 10 or 20 candidates? The hit rate doubles in the former case, but at the end, a single new active compound has been identified by testing a very small number of database compounds. This situation is similar to a search for “needles in haystacks.” What if 200 or 300 compounds had been tested? If more hits would have been identified, enrichment characteristics of different VS calculations could perhaps be compared in a more meaningful manner. Another critical point is the potency of newly identified hits. Our recent survey of 379 published LBVS and SBVS projects revealed that 97 of these applications, that is, more than a quarter, identified hits with maximal potency in the range of more than 10–100 μM (only the most potent hits from any publication were considered) [11]. Indeed, it is rather common that successful VS campaigns report weak hits with potency in the low to mid micromolar range. Of course, one would consider the identification of nanomolar hits (which are not often reported) a more significant success, at least from a medicinal chemistry perspective. However, it is often not considered that compound potency is not utilized as a search parameter in LBVS, neither is it utilized as a calculation parameter for docking (the occasionally attempted correlation of docking scores with compound affinities is scientifically not meaningful). Hence, the identification of more or less

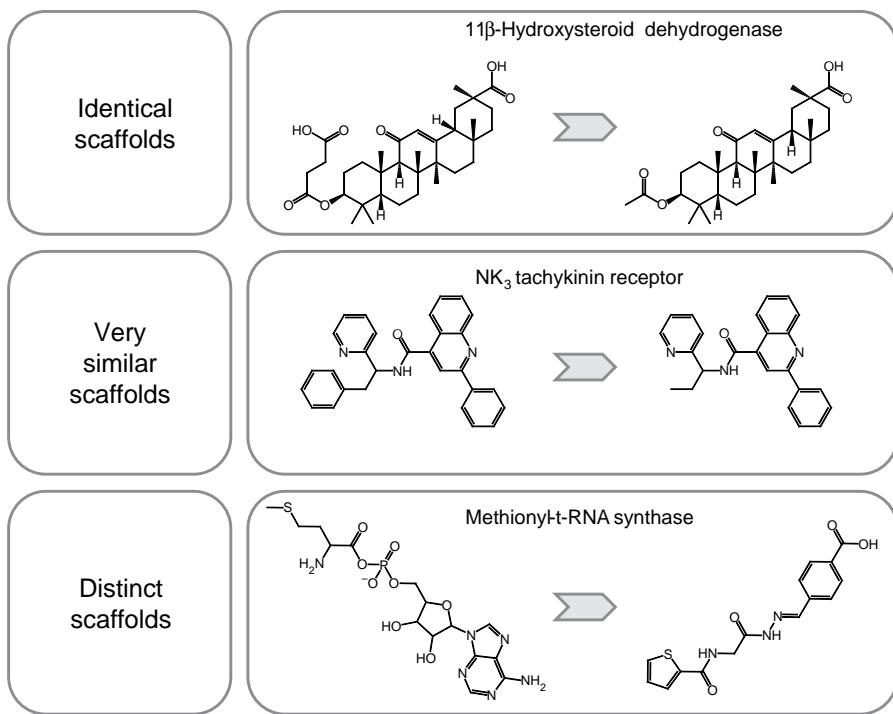


FIGURE 6.2 Exemplary reference compound and hits. From the top to the bottom, structures of pairs of reference compounds (left) and hits (right) become increasingly dissimilar. Examples taken from Schuster et al. and the REPROVIS database [17, 18].

potent compounds through VS can not be rigorously used as a performance measure; it is the luck of the draw and/or intuition in compound selection. In addition to hit numbers/rates or potency, structural novelty of active compounds is often a focal point of the evaluation. In fact, the identification of novel active scaffolds, that is, the “scaffold hopping potential” [13–15] of VS, is usually considered the most important measure of success. The more previously unobserved active scaffolds a VS campaign identifies, the more valuable the approach is thought to be. Intuitively, this makes sense. The caveat is that scaffolds might be defined in rather different ways and, in addition, be more or less closely related to each other [14]. In the literature, the term scaffold is often used in a fairly loose manner. Furthermore, formally defined scaffolds might cover a wide range of structural similarity. For example, the exchange of a single heteroatom in core structures of two analogs defines a new scaffold according to Bemis and Murcko [16], following the currently probably most widely applied scaffold definition. A potential scaffold hop formed by two such molecules, which are visibly similar, is much easier to facilitate than a hop formed by two structurally distantly related or distinct compounds. Figure 6.2 compares reference compounds with hits originating from three different VS projects [17, 18]. Even for a nonexpert, differences are obvious. The first two examples represent reference compounds and

hits with the same or very similar scaffolds. For the detection of such structural relationships, sophisticated VS methods or search protocols are not required; a simple analog search would be sufficient. Yet, pharmacophore or shape similarity searching were applied to identify these compounds [17, 18]. The last example in Figure 6.2 is different. In this case, the reference compound and hit contain distinct scaffolds. Hence, this VS exercise has indeed generated structural novelty, which illustrates opportunities of VS calculations.

Returning to our original question, how should one then evaluate VS success in practical applications? Clearly, the more structurally novel hits are identified, the more successful the application is. Because VS calculations typically (but not always) identify only limited numbers of weakly to moderately potent hits, structurally novelty is indeed a key factor. As a general guideline, reemphasizing the point made earlier, if hits are visibly similar to already known active compounds, they can most likely be identified using simple similarity search techniques (if not “by eye”). The identification of such compounds does not justify the application of highly complex LBVS protocols (or SBVS methods). High potency of newly identified active compounds certainly is a bonus, but cannot be considered a primary measure of success, as discussed earlier. In fact, hits that are structurally distinct from reference compounds are usually not very potent, because the potency of compound series is typically improved during chemical optimization, regardless of the source of the hits. However, this does not exclude the possibility that highly potent compounds might occasionally be identified through VS. As an interesting side note, although docking currently is by far the single most popular VS methodology, the average potency of docking hits is generally lower than of hits identified by LBVS methods, which are less frequently used [19].

6.4 VIRTUAL VERSUS HIGH-THROUGHPUT SCREENING

We next address key questions that are of fundamental relevance for the VS field. How should VS be positioned relative to HTS? And, even more importantly, given the large HTS capacities that are often available nowadays, do we really need VS? In the following, we will discuss different aspects that characterize relationships between HTS and VS.

6.4.1 Do We Need Virtual Screening?

At least in academia, HTS capacity is generally limited. In many instances, VS is the only available approach to search for new active compounds. Furthermore, in cases where assays are difficult, time consuming, and/or expensive and hence not straightforward to format for HTS, VS presents a fast and comparably cheap alternative. It is evident that HTS as an individual discipline has not solved the bottlenecks of drug discovery, despite massive investments. Many HTS campaigns are plagued with false-positives and hit sets that are often difficult to deconvolute. There are a number of examples where VS has identified first-in-class active compounds including cases

where HTS failed [20]. Even if these were exceptional success stories, there are principal justifications for VS that go beyond situations where HTS might have shortcomings. Computationally, it is feasible to process the chemical information content of millions of molecules and prioritize those that are likely to exhibit specific biological activities. Neglecting this information would not be very careful, regardless of whether HTS is used or not. Furthermore, VS methods can be efficiently applied for hit or lead expansion by screening the chemical neighborhood of interesting compounds, taking structural or pharmacophore constraints into account, without the need to experimentally test thousands of compounds. Moreover, for extrapolating from a given active compound and identifying increasingly diverse structures that might still have similar activity (i.e., the scaffold hopping exercise), VS methods are well suited, if applied in an appropriate manner. It might be noted that this exercise is quite popular in pharmaceutical research trying to circumvent intellectual property positions of competitors, a savvy form of “me-too-ism.” Also in this case, computational methods can be applied to search for novel structures in a highly focused manner, without the immediate need to test many new compounds. Clearly, if the VS field works against its scientific heterogeneity and occasional science fiction approaches but rather focuses on sound science, VS will have its place as a discipline. And there is more to come because VS does not yet live up to its full potential.

6.4.2 Underutilized Strengths

Many VS studies attempt to find “needles in haystacks,” as discussed earlier. Very small compound sets are selected in order to identify novel hits. Sometimes, this strategy is essential, for example, when only small numbers of active compounds can be tested in complex biological assays (or if an academic lab with limited resources can only afford acquiring or synthesizing a few candidate compounds). However, the needles in haystacks scenario does not play into the strength of VS approaches, given their limited accuracy (which applies to both molecular similarity and structure-based methods). Rather, a strength of VS approaches is their enrichment capacity. Hence, while it is generally difficult to select 10, 20, or perhaps 100 candidates from a very large database and identify novel hits (although it is an intellectually stimulating exercise), VS methods must be capable of significantly enriching active compounds within larger database subsets. For example, if one starts with a database of a million compounds, VS approaches must display a statistically significant enrichment of active molecules (if available) within selection sets of 1, 5, or 10% of the database (otherwise, the methods would essentially be useless). Many LBVS and SBVS approaches yield enrichments of active compounds at this level. Following VS, these database fractions would amount to 10,000–100,000 candidate compounds, that is, far too many for conventional bioassays. However, in combination with HTS, likely candidates that might possess a specific biological activity contained in a large screening database can be evaluated by testing only a fraction of the database compounds. If several subsequent iterations of VS and HTS are carried out, starting with, for example, 1% of the database and taking information from newly identified hits into account at each stage, the screening process might be efficiently streamlined.

This iterative process represents the “sequential screening” paradigm that has already been discussed more than a decade ago [21]. Yet, sequential screening has thus far not received a high level of attention in drug discovery. This is due to several reasons including, for example, infrastructure requirements associated with cherry picking of compounds from screening plates (which would be necessary for sequential screening). However, this situation is beginning to change in the pharmaceutical industry. In addition, there are also philosophical (and reward structure) differences between HTS and chemoinformatics investigators who are trained to promote high-throughput and reductionist approaches, respectively. Clearly, for VS, the formation of viable interfaces with biological screening is a primary growth area. Exploiting the enrichment characteristics of VS calculations is scientifically more meaningful than pushing the technology over its accuracy limits. For pharmaceutical research, further development opportunities exist by focusing on a complementary use of computational and biological screening techniques; another reason why VS is here to stay.

6.5 STRUCTURAL NOVELTY REVISITED: EXEMPLARY CASES

An important prerequisite of VS success, regardless of the methods applied, is the nature of the target. Simply put, “good” small molecular targets are highly preferred, as for any hit identification approach. The majority of reported successful prospective VS studies involve prominent drug targets such as protein kinases or G-protein coupled receptors [10, 11, 19]. For those targets, many active compounds (hundreds to thousands) are already known, which provides a rich source of chemical knowledge (and an abundance of reference compounds). In such cases, the relevance of VS campaigns must be carefully considered. If already a great variety of active compounds is available, how difficult might it be to identify yet another weakly active hit? Probably not very difficult. This is a question that becomes critical for the credibility of VS studies and their scientific merits, especially if “methodological overkill” is involved. For any target, but especially heavily investigated ones, structural novelty of newly identified hits is a crucial criterion for VS relevance and must be clearly demonstrated. Other criteria such as functional relevance of newly identified active compounds or the presence of evolvable and sustainable structure–activity relationships (SARs) are beyond the control of VS calculations.

Figure 6.3 shows a spectrum of known actives for two different targets that were available when a VS campaign was initiated as well as new hits that were then identified [22]. These compounds and associated SAR information are presented in network-like similarity graphs (NSGs) [23]. In these networks, molecules are depicted as nodes that are connected by an edge if they are pairwise similar above a predefined threshold. The potency of a compound is reflected by a node’s color using a continuous color spectrum ranging from green (lowest potency) over yellow to red (highest potency in a dataset). Additionally, nodes are scaled in size based on calculated compound (local) SAR discontinuity scores [23]. A compound makes a strong contribution to local SAR discontinuity if its potency significantly differs

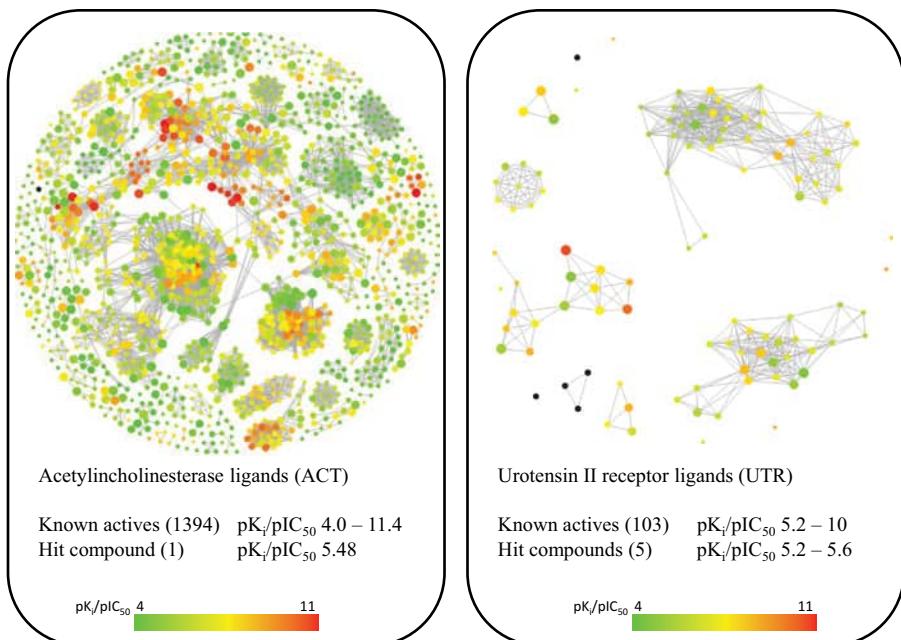


FIGURE 6.3 Network-like similarity graphs for known actives of two exemplary targets. Hits from two VS projects targeting either acetylcholinesterase (ACT) or the urotensin II receptor (UTR) are shown together with all active compounds known prior to VS. In network-like similarity graphs (NSGs), nodes represent compounds and edges pairwise similarity relationships (here fingerprint Tanimoto similarity above a given threshold value). Nodes are color-coded according to potency values using a continuous color spectrum from green (low potency) over yellow to red (high potency) and scaled in size according to compound SAR discontinuity scores. Black nodes represent hits. For color details, please see color plate section.

from its structural neighbors. For both targets, known active compounds were taken from the ChEMBL database [24] (freely available at the time of the VS projects). The two NSGs shown in Figure 6.3 reveal major differences. For acetylcholinesterase (ACT), most of the 1418 displayed ligands have only low potency and these compounds cover a broad structural spectrum, as reflected by the presence of small to moderately sized node clusters and many singletons (nonconnected nodes). From an SAR point of view, several interesting clusters are formed by green, yellow, and red nodes (representing discontinuous local SARs). By contrast, for the urotensin II receptor (UTR), only relatively few known actives with limited structural diversity are available. Here, only a few node clusters are formed. One of these clusters represents a strongly discontinuous local SAR.

In both NSGs, hit compounds identified by VS are shown in black. The single ACT hit contains a previously unobserved scaffold and is a singleton in the graph, that is, structurally dissimilar to any known actives. Hence, in this case, the VS effort only identifies an individual hit, but produces structural novelty for an already well-explored

small molecular target. Furthermore, the five hits identified for UTR are also not structurally related to any of the known actives. Thus, this VS effort produces more novel hits for a target for which only a limited number of known actives are available.

The likelihood of identifying hits should be larger for ACT than for UTR, as indicated by the much larger number of known ACT actives (1417) compared to UTR (108). For both targets, VS was carried out using 3D pharmacophore searching that identified the lowly to moderately potent hits. The much larger compound knowledge base available for ACT favors pharmacophore query development. Although the newly identified ACT hit is structurally novel, the NSG representation shows that many ACT ligands are structurally distinct (singletons). Thus, ACT is permissive to structurally diverse ligands and, from this point of view, an “easy” target.

The situation is different for UTR. In this case, only a limited number of active compounds are known (less than 10% compared to ACT) and most of these compounds form structural relationships. However, the VS hits are unrelated to known actives and thus present a starting point for the development of new structural classes of UTR inhibitors. One might conclude that UTR is a more difficult (or at least less well explored) small molecular target than ACT and that the identification of the new UTR hits is a more significant accomplishment. Although scientific views might differ in such cases, these examples illustrate that the assessment of structural novelty of hits, which we consider a key criterion for VS performance evaluation, is more complex than often thought and strongly context-dependent with respect to target and compound information.

6.6 EXPECTATIONS AND SELECTED APPLICATIONS

Having discussed thus far strategic aspects of VS and principal limitations, we next review selected application examples from our laboratory to highlight opportunities. On the basis of our experience with practical VS (mostly LBVS) over more than a decade, we have learned that the majority of projects we have undertaken (but certainly not all) yielded at least one or a few active compounds that were structurally novel (to varying degrees). Examples of practical VS applications that were carried out in our laboratory over the past few years are summarized in Table 6.2. Because most of these projects were carried out in collaboration with pharmaceutical companies, target names are not provided (they are also not essential for our discussion). An important take-home message from these studies has been, at least for us, that it is generally difficult to expect more from VS than the identification of a few weakly potent compounds. However, these compounds might still be useful for hit-to-lead expansion and optimization, if structurally novel. Hits originating from two of the projects listed in Table 6.2 went through lead optimization and were transformed into pre-clinical candidates. The results in Table 6.2 provide a fairly realistic view of what we typically expect from our VS campaigns. However, occasionally there are also unusually successful VS projects, which by themselves might justify the approach. In the following, two examples are discussed.

TABLE 6.2 Practical VS Applications

Target	VS Hit Rate, %	Active/Assayed cpds
1	0.7	1/135
2		0/143
3	4.0	1/25
4	7.6	6/79
5	11.7	17/145
6	4.4	5/113
7	2.0	1/50
8	2.1	1/48
9		0/22
10	7.1	1/14
11	6.6	5/76
12	4.4	8/183
13	18.8	3/16
14	17.9	26/145
15	25.5	13/51

6.6.1 Inhibitors of Multifunctional Proteins: Cytohesins

Cytohesins are small guanine-nucleotide exchange factors that control various cellular regulatory networks implicated in, for example, vesicle trafficking, integrin activation, or insulin signaling [25–28]. At the time of our VS campaign, only a single cytohesin inhibitor and a few subsequently generated derivatives were known. This molecule, termed SecinH3, a triazole derivative, was identified in a biological screen and had an IC_{50} value of 11.4 μM [29, 30]. Attempts to further improve this compound were not successful and therefore a VS campaign was launched to search for new active chemotypes (structural classes). The applied LBVS protocol consisted of parallel as well as sequential applications of different methods. Initially, two different strategies were applied in parallel: support vector machine (SVM) modeling [31] and similarity searching using three different 2D fingerprints (that were also used as descriptors for SVM calculations). Despite the only very limited active compound information that was available, we trained SVM models for compound classification and ranking. Although the same fingerprint descriptors were used for similarity searching and SVM predictions, there was no overlap between highly ranked ZINC database [32] compounds. From 145 candidate compounds that were assembled from individual search lists, 26 were found to be active in different assays with higher potency than SecinH3 and all of these hits contained new scaffolds [30]; a truly unexpected result. The most potent compound, termed Secin16 (with an IC_{50} value of 3.1 μM), was identified by SVM and subsequently used as reference compound for a second round of similarity searching [33]. Figure 6.4a reports the workflow of VS campaign in more detail. The second similarity search based on Secin16 led to the identification of a more potent compound named SecinB7, with an IC_{50} value of 440 nM. Hence, this LBVS effort successfully extrapolated from a small set of structurally related compounds and

identified an unexpectedly large number of structurally diverse and more potent hits, as illustrated in Figure 6.4b. In addition, Figure 6.4c depicts the stepwise expansion and screening of chemical space around the known active compounds. On the left, cytohesin inhibitors are shown that were available at the time of VS round 1 (top row), round 2 (middle row), and after the VS campaign (bottom row). On the right, selected candidate compounds that were inactive or less potent than the most potent reference compound are shown together with qualifying hits. The first candidate selection focused on a broad range of structurally diverse compounds in order to search for new active chemotypes. More than 20 structurally distinct compounds were found to be more active than SecinH3. During the second round, structural neighborhood analysis of a preferred inhibitor then identified more potent compounds.

In addition to potency improvements, the VS campaign revealed hits with different inhibitory profiles with respect to the three biological activities of cytohesins: guanine-nucleotide exchange [29], reduction of gene expression in the *Drosophila* insulin-like peptide signaling pathway (analogously to the insulin signaling pathway in vertebrates) [26], and cell adhesion [28]. For example, Secin69 was predominantly active in the inhibition of nucleotide exchange, whereas Secin107 was an effective inhibitor of cell adhesion. Furthermore, Secin16 was a “consensus hit” that interfered with all three cytohesin functions under investigation [30]. The identification of new cytohesin inhibitors with differentiated inhibitory profiles was an unexpected, but very interesting result, because it provided the basis for the use of small molecular probes to differentiate between cytohesin functions. The ability of LBVS calculations to identify these compounds was not expected. Such compound characteristics were essentially unpredictable and there was no “computational magic” involved; neither were the applied LBVS protocols particularly complex. Global molecular similarity methods were sufficient to identify new inhibitors with differentiated functions at an unexpectedly high rate.

6.6.2 First-in-Class Inhibitor for Ecto-5'-Nucleotidase

Ecto-5'-nucleotidase (ecto-5'-NT) is a member of the metallophosphoesterase superfamily and catalyses the hydrolysis of adenosine monophosphate to adenosine and phosphate [34]. Increasing levels of extracellular adenosine activate adenosine receptors that play important roles in several therapeutically relevant biological processes [35]. Due to observed antiangiogenic effects associated with ecto-5'-NT interference, the enzyme has been implicated in cancer become an emerging oncology target [36–39]. In addition to the natural molecules adenosine di- and triphosphates (ADP, ATP) and number of ADP and ATP analogs, only anthraquinone derivates have thus far been known to inhibit ecto-5'-NT. However, anthraquinones are charged under physiological conditions and do not penetrate cell membranes. Hence, they are not suitable for therapeutic intervention. An SBVS project was initiated using 70,000 ZINC database compounds that were preselected by 2D fingerprint similarity searching using anthraquinones as templates. These compounds were flexibly docked into a comparative model of human ecto-5'-NT (enzymes from different species have

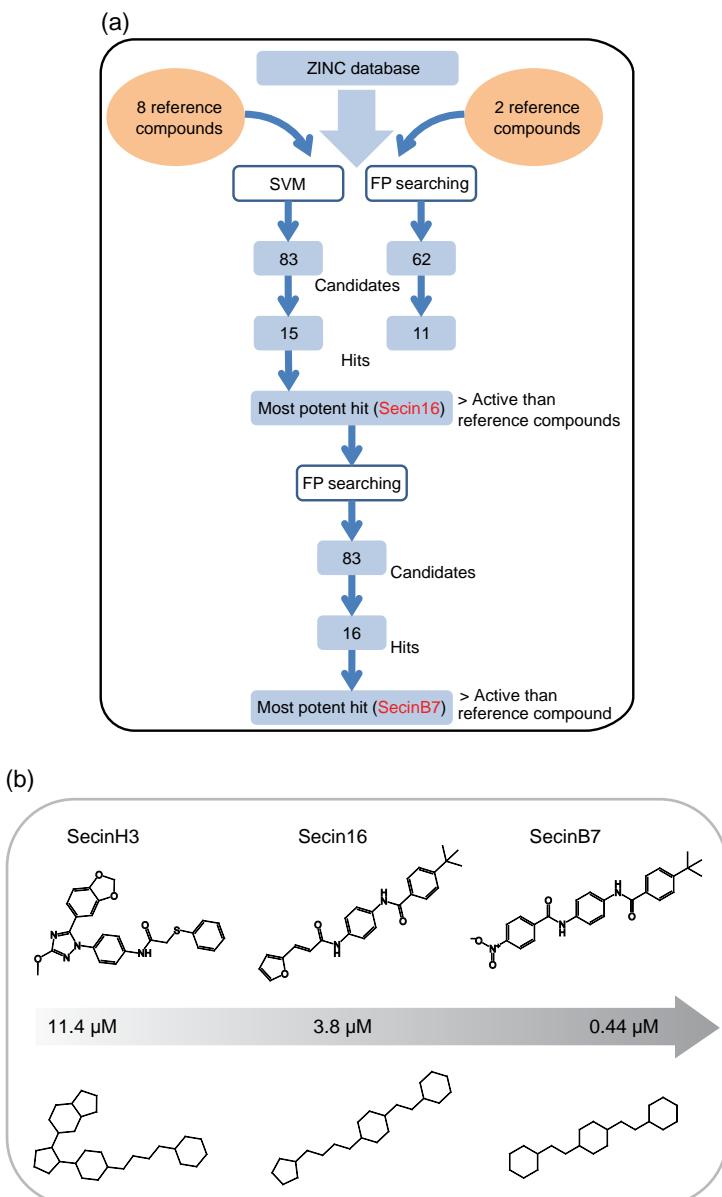


FIGURE 6.4 Virtual screening for cytohesin inhibitors. (a) VS workflow. Detailed description of the workflow leading to selected candidate compounds and experimentally confirmed hits. (b) Structures of reference compound and preferred hits. The reference compound SecinH3 and the best hits from each of two VS rounds, Secin16 and SecinB7, respectively, are shown together with their carbon skeletons. The potency of each compound is also reported. For color details, please see color plate section.

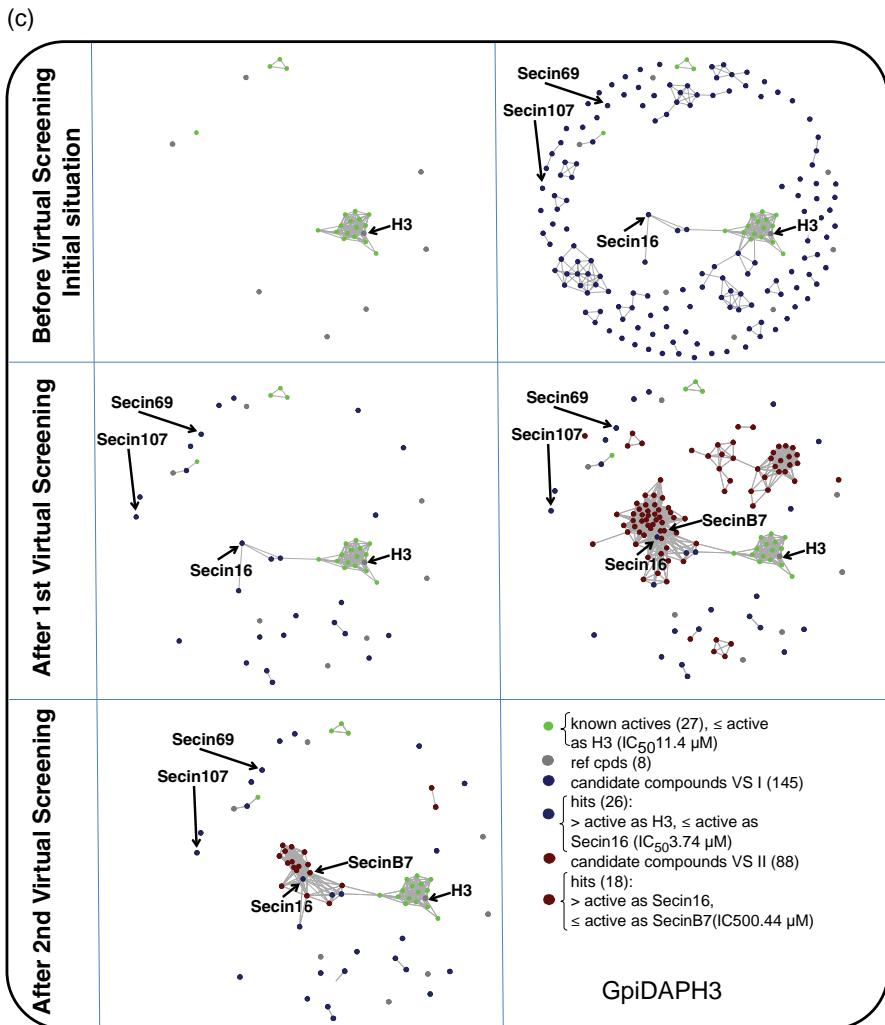


FIGURE 6.4 (Continued) (c) NSGs capturing different stages of the cytohesin inhibitor VS projects. NSG representations are shown to capture active compounds and their similarity relationships prior VS and following the first and second round of VS. Similarity calculations were carried out using a 2D pharmacophore-type fingerprint termed GpiDAPH3 (Molecular Operating Environment, Chemical Computing Group, Montreal, Canada). For color details, please see color plate section.

conserved active site regions) [40]. The 3500 top-ranked compound were visually inspected and 372 compounds were selected after the first round that displayed a reasonable degree of complementarity with the modeled active site of human ecto-5'-NT. Using these compounds, a second round of visual analysis was carried out evaluating the hypothetical binding modes in greater detail. On the basis of this evaluation, 128 candidates were selected, 51 of which were commercially available. These candidates

were experimentally tested in enzyme inhibition assays. Thirteen of the 51 tested compounds were found to inhibit ecto-5'-NT and six of these inhibitors yielded IC_{50} values below 10 μM . The most potent compound had an IC_{50} value of 1.90 μM and a K_i value of 1.58 μM (comparably potent to the best anthraquinone derivates) [40]. The most potent compounds identified on the basis of SBVS and extensive visual inspection were sulfonamides containing a nucleoside-mimicking substituted benzene moiety. These compounds were more drug-like than anthraquinones and, importantly, uncharged under physiological conditions and thus membrane-permeable and suitable for further pharmaceutical development.

The cytohesin and ecto-5'-NT examples represent unusually successful VS applications. They emphasize that VS calculations are capable of identifying rather interesting active compounds, often unexpectedly so. These examples also illustrate that highly complex methods and calculation protocols are not necessarily required and that chemical knowledge and intuition often play an instrumental role in candidate compound selection. Complementing approximate calculations with knowledge-based approaches is completely appropriate and much better science than promoting “computational overkill” or overinterpreting computational results.

6.7 CONCLUSIONS: WHAT IS POSSIBLE? WHAT IS NOT?

We have emphasized that VS approaches are not a form of “computational magic.” Rather, it is straightforward to understand their opportunities and scientific approximations and limitations. If applied carefully, VS can make an impact on pharmaceutical research, although its full potential is still not utilized, especially at the interface with HTS. It is not the strength of approximate VS calculation that have only limited sensitivity and specificity to select small numbers of candidate compounds with the aim to identify attractive new hits. Nevertheless, such projects also succeed on a fairly regular basis, depending on the targets and compound classes under study. A typical successful VS project results in the identification of a few active compounds with micromolar potency. Hence, structural novelty becomes a crucial criterion for the potential impact of such findings. As also discussed herein, the evaluation of VS results is often context-dependent and structural novelty might be viewed differently depending on already available compound information.

For the future of VS, it will be crucial to further improve its scientific rigor and establish general acceptance criteria and scientific standards for publication of VS studies. Initial steps in this direction have been taken. Furthermore, there are unsolved scientific problems of high significance and practical relevance that must be addressed including, among others, the strong dependence of VS calculations on compound classes and molecular representations, the absence of well-defined relationships between calculated molecular similarity and observed activity similarity, or the limitations of scoring functions.

If VS is positioned as a rigorous scientific discipline, it will be here to stay. It is an interesting field. Of course, VS should be more than an intellectually stimulating exercise, and it certainly has the potential to impact pharmaceutical research.

REFERENCES

1. Walters WP, Stahl MT, Murcko MA. Drug Discov Today 1998;3:160–178.
2. Bajorath J. Curr Opin Drug Discov Devel 2002;2:24–28.
3. Klebe G. Drug Discov Today 2006;11:580–594.
4. Stahura F, Bajorath J. Comb Chem High Throughput Screen 2004;7:259–269.
5. Sheridan RP, Kearsley SK. Drug Discov Today 2002;17:903–911.
6. Eckert H, Bajorath J. Drug Discov Today 2007;12:225–233.
7. Stumpfe D, Bajorath J. Wiley Interdiscip Rev: Comput Mol Sci 2011;1:260–282.
8. Johnson MA, Maggiora GM. *Concepts and Applications of Molecular Similarity*. New York: John Wiley Sons, Inc.; 1999.
9. Salim N, Holliday J, Willett P. J Chem Inf Comput Sci 2003;43:435–442.
10. Ripphausen P, Stumpfe D, Bajorath J. Future Med Chem 2012;4:603–613.
11. Ripphausen P, Nisius B, Bajorath J. Drug Discov Today 2011;16:372–376.
12. Bajorath J. J Med Chem 2012;55:3593–3594.
13. Schneider G, Neidhart W, Giller T, et al. Angew Chem Int Ed Engl 1999;38:2894–2896.
14. Hu Y, Stumpfe D, Bajorath J. J Chem Inf Model 2011;51:1742–1753.
15. Brown N, Jacoby E. Mini Rev Med Chem 2006;6:1217–1229.
16. Bemis GW, Murcko MA. J Med Chem 1996;39:2887–2893.
17. Schuster D, Maurer EM, Laggner C, et al. J Med Chem 2006;49:3454–3466.
18. Ripphausen P, Wassermann A, Bajorath J. J Chem Inf Model 2011;51:2467–2473.
19. Ripphausen P, Nisius B, Peltason L, et al. J Med Chem 2010;53:8461–8467.
20. Stumpfe D, Ripphausen P, Bajorath J. Future Med Chem 2012;4:593–602.
21. Bajorath J. Nat Rev Drug Discov 2002;1:882–894.
22. Ripphausen P, Nisius B, Wawer M, et al. J Chem Inf Model 2011;51:837–842.
23. Wawer M, Peltason L, Weskamp N, et al. J Med Chem 2008;51:6075–6084.
24. ChEMBL, European Bioinformatics Institute (EBI). 2010. Available at <http://www.ebi.ac.uk/chembl/>. Accessed 2011 Jan 10.
25. Klarlund JK, Guilherme A, Holik JJ, et al. Science 1997;275:1927–1930.
26. Fuss B, Becker T, Zinke I, et al. Nature 2006;444:945–948.
27. Ogasawara M, Kim SC, Adamik R, et al. J Biol Chem 2000;275:3221–3230.
28. Kolanus W, Nagel W, Schiller B, et al. Seed Cell 1996;86:233–242.
29. Hafner M, Schmitz A, Grüne I, et al. Nature 2006;444:941–944.
30. Stumpfe D, Bill A, Novak N, et al. ACS Chem Biol 2010;5:839–849.
31. Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, editors. *Kernel Methods: Support Vector Learning*. Cambridge: MIT Press; 1999.
32. Irwin JJ, Sterling T, Mysinger MM, et al. J Chem Inf Model 2012;52:1757–1768.
33. Bill A, Blockus H, Stumpfe D, et al. J Am Chem Soc 2011;133:8372–8379.
34. Knöfel T, Sträter N. J Mol Biol 2001;309:239–254.
35. Colgan SP, Eltzschig HK, Eckle T, et al. Purinergic Signal 2006;2:351–360.

36. Stagg J, Divisekera U, McLaughlin N, et al. Proc Natl Acad Sci USA 2010;107: 1547–1552.
37. Clayton A, Al-Taei S, Webber J, et al. J Immunol 2011;187:676–683.
38. Stagg J, Divisekera U, Duret H, et al. Cancer Res 2011;71:2892–2900.
39. Stagg J, Beavis PA, Divisekera U, et al. Cancer Res 2012;72:2190.
40. Ripphausen P, Freundlieb M, Brunschweiger A, et al. J Med Chem 2012;55:6576–6581.

CHAPTER 7

CHEMOMETRIC APPLICATIONS OF NAÏVE BAYESIAN MODELS IN DRUG DISCOVERY: BEYOND COMPOUND RANKING

EUGEN LOUNKINE, PETER S. KUTCHUKIAN, and MEIR GLICK

7.1 INTRODUCTION

Although naïve Bayesian models (NBMs) date back to the eighteenth century, they gained increasing popularity in drug discovery primarily in the past decade. The key driver behind this change was the need to leverage “big data” in drug discovery and more specifically *in silico* lead finding [1–3]. The size of state-of-the-art screening collections grew from tens of thousands in the 1990s into millions. Such collections were routinely screened to identify new tools to probe the biology or starting points for lead optimization [2, 3]. This resulted in hundreds of millions of data points with an unknown number of false positives and negatives. The search for a classification approach that scales linearly with the number of data points and is also tolerant to stochastic noise led to the usage of NBM in high-throughput screening (HTS) data analysis [1]. The implementation of NBM in commercial chemoinformatics data pipelining packages with sparse 2D chemical descriptors turned out to be an easy IT solution that enabled scientist to build and deploy classification models for large HTS datasets in a matter of minutes [4–6]. NBMs were routinely applied to answer specific questions such as Which chemical features are associated with the bioactivity of a compound? Which compounds are worth retesting or purchasing? What kind of chemical matter is missing from the collection? Which chemical features are associated with assay artifacts? NBM was further used in other applications that were not HTS centric: prediction of chemical attractiveness of compounds based on chemists’ voting, prioritization of compounds

from high-throughput docking, or prioritization of small molecules for screening based on fragment screens [7]—just to name a few.

The next step in the evolution of NBM came from the transition from a single-target to a multitarget view, for example, in phenotypic assays in drug discovery. The chemogenomics paradigm opened new avenues in the application of NBM on a sparse data matrix of millions of compounds over hundreds of targets. Here the data outlived the projects and additional millions of structure–activity relationship (SAR) points were available in commercial and noncommercial biologically annotated databases. Apart from the tolerance to noise and its speed, the ability of NBM to deal with missing data proved to be extremely useful. Multicategory Bayesian models were routinely built to model the entire bioactivity data of compound sets [4–7]. This enabled the prediction of the activities of compounds against a spectrum of targets that they had never been assayed against. Scientists were better able to answer more complex questions such as How selective are the hits predicted to be? Are they likely to hit on-target or off-target? What is the compounds' mode of action (MOA)? How are targets related to each other based on the similarity of the ligands that bind them?

Today NBM is used both before and after HTS. In the former case, NBM is used to design a compound set to answer a specific biological question. The compounds, for example, may have certain probabilities of modulating various nodes in a biological pathway. If the compound subset was able to address the question, then a full HTS may no longer be required, saving time and money. Post HTS, multicategory Bayesian models are used to assess the robustness of the assay, determine whether the assay is capturing compounds with expected MOA, and to develop a hypothesis for follow-up experiments.

We first give a brief overview of the mathematical foundations for NBM. Then we will discuss routine applications of NBM in drug discovery; we conclude with novel applications of NBMs that go beyond classification of compounds based on their chemical structure.

7.1.1 Naïve Bayesian Models

In chemoinformatics, an NBM typically assesses the likelihood of a compound to belong to a defined class, for example, the likelihood of being active at a specific biological target. This likelihood is calculated on the basis of multiple, often binary, descriptors. The classifier is called naïve because it assumes independence of the individual variables. For example, when the presence or absence of different substructures is used as descriptors, the classifier assumes that the presence of one substructure is not influenced by the presence of any other feature. Obviously this assumption does not take into account substructure relationships between the different features, and ignoring (or “shadowing”) features [8] that are substructures of larger features can improve classifier performance. Nevertheless, NBMs have proved to be stable and resistant to noise in the data [1]. Although the naïve Bayesian classifier has been applied to both binary and continuous features in combination [9], a commonly used implementation [6, 10] uses binary features only,

and bins continuous variables accordingly. Here we give a brief overview of how the classifier is derived with a particular focus on feature weights and how they are estimated.

Assuming independence, the conditional probability that a compound is active given a set of features can be written as the product:

$$P(\text{Active} \mid \text{Features}) = P(\text{Active}) \times \prod_i \left(P(\text{Feature}_i \mid \text{Active}) / P(\text{Feature}_i) \right).$$

Thus, the probability of the compound being active in the first place, $P(\text{Active})$, is multiplied by the product of conditional probabilities that the compound is active given a particular Feature_i, divided by the probability that Feature_i is present irrespective of activity. The quotient $P(\text{Feature}_i \mid \text{Active})/P(\text{Feature}_i)$ can be smaller, equal to, or greater than one. If it is one, then the probability of encountering a particular feature among active compounds is the same as encountering the feature in any compound. If it is greater than one, then the probability of finding the feature in an active compound is higher than finding the feature in any compound, and vice versa. This ratio can be estimated from a training set with a total number of molecules T , A active compounds, and their counterparts containing a feature F : AF—active compounds with the feature, and TF—any compounds with the feature. Then the probabilities can be estimated as follows:

$$P(\text{Feature} \mid \text{Active}) = \text{AF}/\text{A},$$

$$P(\text{Feature}) = \text{TF}/\text{T},$$

and hence

$$P(\text{Feature} \mid \text{Active}) / P(\text{Feature}) = \text{AF}/(\text{TF} \times (\text{A}/\text{T})).$$

This corresponds to the proportion of actually observed number of active compounds with the feature over the expected number of such observations, $(\text{TF} \times (\text{A}/\text{T}))$, assuming the independence of activity and feature. If there is more than one class present, the classifier can be efficiently trained on multiple classes in parallel, changing only A and AF for each class. In order to correct for under-sampled features, pseudo-counts are added, yielding the Laplace-corrected estimator [4, 5]:

$$P(\text{Feature} \mid \text{Active}) / P(\text{Feature}) = (\text{AF} + 1) / (\text{TF} \times (\text{A} / \text{T}) + 1)$$

This ensures that for rare features the ratio will tend toward 1. As an extreme example, consider a feature that is present in only one molecule in a training set of 1,000,000 molecules with 100 actives, and this molecule happens to be active; the uncorrected estimate is

$$1 / (1 * (100 / 1,000,000)) = 1 / 0.0001 = 10,000$$

which is clearly an overestimation of the importance of a feature that only occurs once in the entire training set. In contrast, the corrected estimate is $2 / 1.0001 \approx 2$.

For computational efficiency, the logarithms of the ratios are often added, rather than multiplying the ratios themselves. For memory efficiency, features with log-ratios equal to or very close to 0 can be omitted from the model; especially when sparse fingerprints such as ECFP_4 are used, most features will be found with approximately equal probability in both active and inactive compounds [5]. The sum of the log-ratios of features present in test compounds serves as a score to rank them according to their likelihood to be active:

$$\text{Bayesian Score} = \sum \ln(P(\text{Feature}_i | \text{Active}) / P(\text{Feature}_i))$$

Note that this sum does not take into account the probability of a compound being active in the first place (the prior), rather, it reflects if the molecule contains more class-characteristic features or more features that are not present in the class. An ensuing limitation is bias toward high-molecular weight compounds. For example, polyhydroxylated natural products tend to receive too high scores if features encoding for hydroxyl groups in different chemical contexts are enriched in an activity class.

7.2 VIRTUAL SCREENING USING BAYESIAN MODELS

The aim of virtual screening using NBM often is to find compounds active at a defined target. NBM in combination with extended connectivity fingerprints (ECFP_4 and ECFP_6) have been shown to be particularly suited for prioritizing active compounds. Rather than comparing molecules to reference compounds, as is done in molecular similarity approaches [11], the NBM assesses the features of the compound independently: The more activity-class-enriched features a molecule possesses, the more likely it is assumed to be active. Bayesian models have proven useful in identifying compounds that are active at diverse targets [5, 12].

While the compound score does not reflect the actual probability of a compound being active, it is often interpreted as the relative likelihood of being active and can be used for ranking compounds. In particular, this score does not *a priori* correlate with potency. If compound A has a score of 100 and a compound B has a score of 50, one cannot say, based on these scores, that compound A may be more potent than compound B [6]. One way to include potency information into NBM is to introduce multiple classes for potency, for example, high-, medium-, and low-potent compounds. Sometimes, overlapping classes are used, such as low, low + medium, medium + high, and high. The final compound classification is a combination of the different scores. However, models trained on low-activity fragments have been shown to enrich for highly potent HTS hits due to the naïve treatment of each feature as independent [7].

In addition, alternate virtual screening methods can incorporate activity-class-characteristic features. For example, by counting the number of activity-characteristic features in test compounds, active molecules were identified that possessed new combinations of activity-characteristic features that were not observed in reference compounds, resulting in chemotypes that were distinct from any of the reference compounds [13]. Similarly, virtual fragment linking [7] used features enriched

in fragment-based screening hits in combination to prioritize larger compounds for HTS. One limitation of this approach is the underrepresentation of active singletons. Such low-scoring active compounds can be used to identify SAR holes in the collection.

7.2.1 Reverse Virtual Screening: Target Fishing

In conventional virtual screening, multiple compounds are scored for activity at a defined target; however, there are an increasing number of use cases for scoring multiple targets against one compound. The shift of focus from a single target toward an ensemble of often distinct targets, which includes (desired) polypharmacology and (unwanted) off-target activities, has put the emphasis on annotation of compounds for as many targets as possible. Multi-category NBMs have been demonstrated to predict bioactivity at multiple targets for individual compounds [4]. Scoring a compound against many different targets has also been used to generate Bayes affinity fingerprints [14], which substitute biological activity descriptors with Bayesian scores and can be used to compare compounds in a chemogenomics space defined by target models.

Predicted targets are usually followed up by *in vitro* experiments that test the prediction for selected compounds, such as representatives of a chemical series. These data can be used to evaluate and improve the current model. A well-designed informatics infrastructure with data normalization must be present in order to automate such a process (c.f. Section 7.3).

7.2.1.1 Target Identification in Phenotypic Screening Campaigns

One practical application of reverse virtual screening that is carried out at Novartis on a regular basis is the identification of potential targets for hits coming from phenotypic screens. In addition to known targets of the compound, potential targets are predicted. Thus, NBMs complement the known compound-target activity matrix. As a follow-up, these targets may be mapped to biological pathways and/or processes. Technically, targets are represented by collections of Entrez gene IDs, and they are mapped to Gene Ontology processes or similar classification from commercial sources, such as Thompson Reuters GeneGo Metabase [15]. The resulting processes can then be interpreted in the context of the phenotype in question, and targets can be prioritized that are likely to modulate phenotype-relevant pathways.

Different databases and known compound classes can be combined in order to generate an MOA hypothesis. Here, we exemplify this using a historical Novartis antiviral screen where NBM were utilized to elucidate MOA. One million compounds were screened in a whole-cell live-virus infection assay to measure each compound's ability to prevent influenza virus-induced cell death. A total of 167 confirmed hits were identified with $IC_{50} < 40 \mu M$ that did not interfere with the luminescence assay readout (assessed by a counterscreen). First, a multiclass model was trained on chemogenomics data assembled from external and internal databases, which predicted 2435 mammalian targets for the 167 compounds, with up to 128 predicted targets/compound (Figure. 7.1). Although the Novartis screening collection

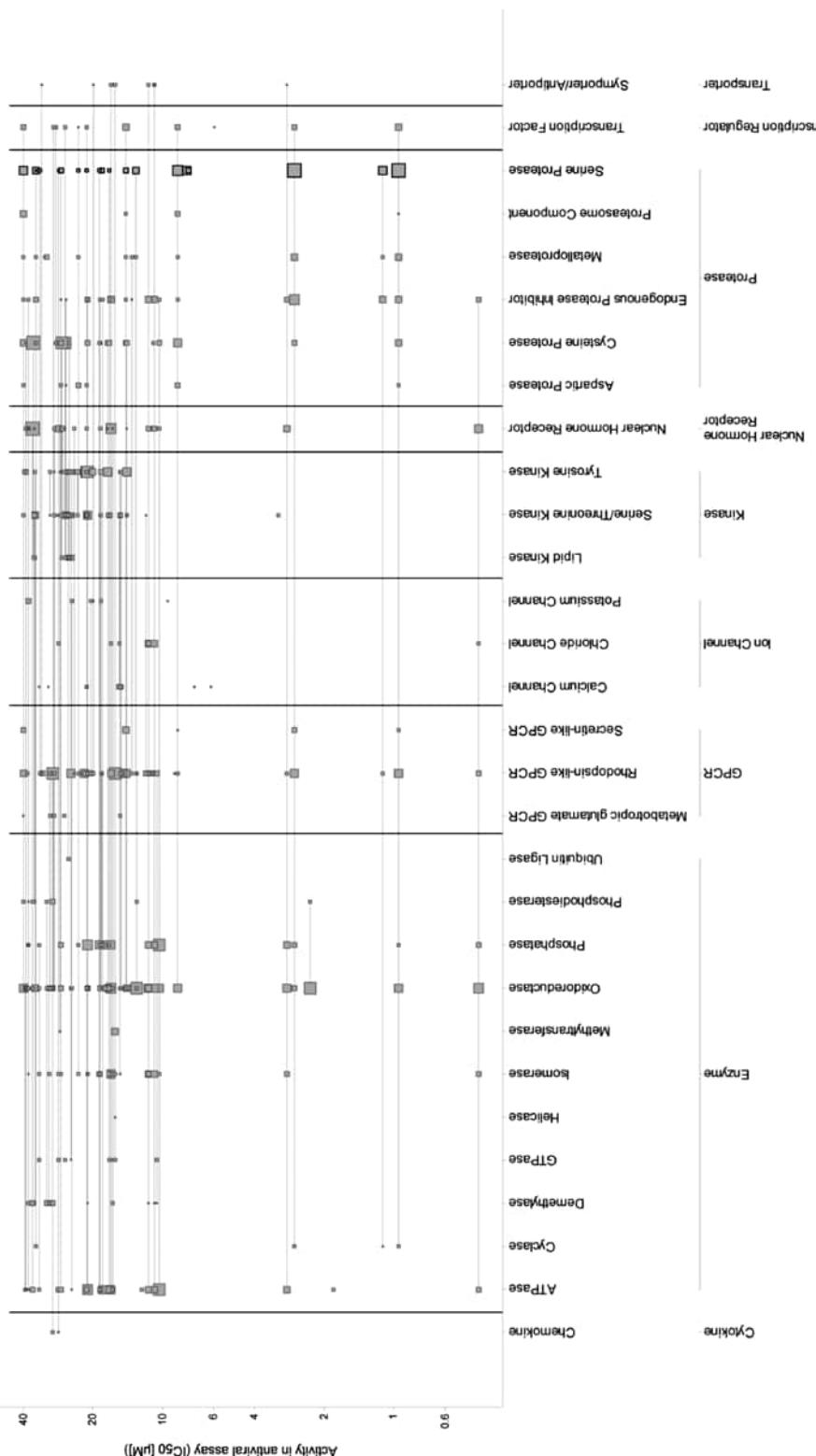


FIGURE 7.1 Target prediction for phenotypic assay hits. Predicted target classes (x-axis) for compounds active in the influenza assay (y-axis: primary IC₅₀ in micrometer) are shown. Each horizontal line corresponds to one compound. Marker size reflects maximal Bayes scores for each compound-target class pair. Highly scoring serine protease predictions are enriched for highly potent compounds and are highlighted. Vertical separators indicate target families.

is enriched for kinase chemotypes, the most prevalent predicted targets were (serine) proteases, especially among highly active compounds, suggesting protease inhibitory activity, known to interfere with virus entry and pathogenicity [16]. In an orthogonal follow-up analysis, a multiclass NBM was built using MDL Drug Data Report on more general MOA classes that were not always linked to a particular protein target. This approach also identified several compounds as potential protease inhibitors; in addition, three compounds were identified as potential antivirals. Indeed, these compounds were structurally similar to known antiviral drugs. Identification of such straightforward MOAs serves two purposes: first, it validates both the assay and the model. Second, these compounds can be selected and optimized for other properties like absorption, distribution, metabolism, and excretion (ADME) to improve their efficacy and safety compared to drugs on the market. Alternatively, if the goal of the project is to identify novel MOAs, as in this case, such predictions can serve as a filter to exclude compounds with known MOA and focus on compounds whose activity in the phenotypic assay is more surprising.

7.2.1.2 Off-Target Prediction and Alert About Potential Liabilities Individual targets can be linked to clinical adverse drug reactions (ADRs) that can range from comparably mild ADRs such as nausea and headaches (histamine receptor H₂ [17]) to serious adverse reactions, such as arrhythmias (hERG [18]). Of particular interest are targets for which *in vitro* activity can predict serious ADRs that manifest late in the clinic, and thus are likely to be missed in clinical trials. For example, serotonin 5-HT_{2B} receptor agonists can cause valvular cardiopathy through activation of pathways leading to hypertrophy and stiffness of the valves [19]. This irreversible effect accumulates over years and is likely to be missed in short-term clinical studies. In contrast, early *in vitro* assessment of agonistic activity at 5-HT_{2B} can help mitigate this liability by increasing compound selectivity or choosing a different lead compound without this liability. Links between individual targets and ADRs have led to the creation of *in vitro* safety panels containing targets of interest [20, 21], and lead candidates are routinely screened against multiple targets to identify potential liabilities early on [21].

A critique that is often raised generally against computational models that predict off-target activity is that since compounds are going to be screened routinely anyway, there is no value in predicting off-targets. However, *in vitro* panel screens are expensive and typically only promising lead candidates are subjected to such testing. Conversely, thousands of (virtual) compounds can be scored for a panel of targets. This has a direct application in the early prioritization of chemotypes extracted from HTS screens—off-target models can identify chemotypes that may be less potent at the primary targets, but also much cleaner in terms of potential off-target activities. These compounds can then serve either as lead candidates or diverse backup compounds. If off-target effects are not taken into consideration early on, then the backup series may be subject to similar or additional, unforeseen liabilities as the primary lead compound.

7.2.2 Comparison to Other Molecular Representations and Machine Learning Techniques

NBMs have most widely been used with sparse extended connectivity fingerprints [5, 7, 12, 14, 22]. Other descriptors also have been explored, such as 2D pharmacophore triplets [23] and atom types [24]. Molecular fingerprints were also combined with continuous descriptors [25]. Although widely applied naïve Bayesian implementations bin continuous descriptors, the Bayesian framework naturally allows for incorporation of continuous descriptors characterized by a probability density, rather than a probability mass function, as in the binned case.

Compared to other machine learning techniques, the naïve Bayes classifier is very fast: because of assumption of independence, it scales linearly with the number of descriptors (features) and compounds. From our experience [26], and as reported by other groups [27], computationally more expensive machine learning techniques, such as random forests (RFs) tend to perform equally well or better, but need a much longer time to be trained (minutes for Bayesian models vs. hours to days for non-parallelized RF training). Given its stability, speed, and interpretability [5, 26], NBM in combination with extended connectivity fingerprints remain a primary choice for machine learning in our group.

7.2.2.1 Combining Bayesian Models with Molecular Similarity

A common strategy for SAR elucidation of hits is the clustering of compounds based on chemical criteria with subsequent assessment of their bioactivity [28]. However, standard molecular representations weigh all molecular features equally, which can introduce bias toward functional groups, such as halogens, that are often used to diversify compounds or solve ADME issues, but are not related to the biological activity. By contrast, parts of the pharmacophore may be missed by automatic clustering approaches, and compounds are rarely automatically grouped based on activity-characteristic molecular cores. This bias is also known in virtual screening, where complex reference compounds yield artificially high similarity values due to high feature density [29, 30]. In order to address this problem, recently we combined molecular similarity with the traditionally complementary NBM [31].

In standard molecular similarity approaches, the similarity of two molecules is assessed by calculating, for example, the Tanimoto coefficient (T_c). The T_c assesses how many features two molecules have in common and relates that number to the total number of unique features present in either molecule. When each feature is given the same weight, complexity effects as described earlier occur and compounds may be clustered based on features that are not characteristic of activity.

We extracted feature weights from NBM trained on active versus inactive compounds and weighted bit positions accordingly. We then compared the molecules using these weighted fingerprints and a general version of the T_c that allows for real-valued descriptors. This procedure corresponded to comparing only parts of molecules that were characteristic of active (vs. inactive) compounds. Clusters calculated using weighted similarity were defined by activity-characteristic molecular

cores, whereas compounds sharing common diversifying features that were also present in inactive compounds were separated by our clustering scheme [31]. This approach is particularly useful for SAR elucidation of large and diverse activity sets (e.g., HTS hits), where compounds are not organized in well-defined chemical series. Rather, activity-weighted clustering can help identify chemotypes based on activity-characteristic molecular cores.

7.3 DATA TYPES AND DATA QUALITY REQUIREMENTS

In order to define compound (activity) classes as input for any computational models, and NBM in particular, compounds, as well as their annotations, need to be standardized. This is particularly important when more than one database is utilized, for example, external databases like ChEMBL, DrugBank, and GVKBio together with internal, proprietary data. In our experience, the following standards and normalization procedures proved useful.

7.3.1 Compound Structure

The IUPAC International Chemical Identifier (InChI) is a layered linear representation of molecular structure [32]. Different layers encode the chemical sum formula, connectivity, stereochemistry, and tautomers. Since it is computed from the chemical structure, molecules from disparate databases can be represented in a coherent way. A 27 character string, the InChIKey, which is computed using a hashing algorithm from the original InChI allows for efficient database storage and query. The layered structure of the InChI is carried along in the InChIKey; for example, the first 14 characters are computed from the connectivity layer. This is useful for finding compounds with same connectivity, but distinct stereochemistry. For chemogenomics data, this is often necessary because correct assignment (or assignment at all) of stereochemistry is a weak spot of many databases [33].

7.3.2 Biological Activity

To uniquely identify targets across databases, in our experience, it has proved useful to represent each target by the Entrez gene ID of the gene coding for the target. Protein as well as RNA targets can be thus represented and used across different databases. This level of granularity does not account for distinct binding sites of the same gene product, but this information is rarely present for a large enough number of compounds to build a reliable model anyway. A technical advantage of the Entrez gene ID over other identifiers is that it does not contain text and thus can be stored as an integer in relational databases, improving lookup speed. Target families as well as biological pathways can be conveniently represented as lists of gene IDs.

Quantitative annotation, such as K_i , IC_{50} , EC_{50} , and related data, can be stored in micro/nanomolar or log unit (pIC_{50}). Percent inhibition/activation can also be used,

but since the concentration at which compounds have been measured can vary, this is often only meaningful for consistent (in house) screening data. For many biological questions, it is important whether a compound is an agonist or antagonist at a certain target and for some targets enough functional data are available to incorporate this information into the model. Functional annotations can come in many varieties, but we found that for our purposes often three categories are sufficient: [1] “Positive interaction” including agonists, partial agonists, enhancers, and so on; [2] “negative interaction” including antagonists, inhibitors, blockers, inverse agonists, and so on; and [3] “binding” including all data with no further functional classification.

7.3.3 Binning of Potency and Guidelines for Multiclass Bayesian Models

Because discrete classes need to be defined in order to train a (multiclass) NBM, the training set of active compounds is often defined by an activity cutoff, such as $1\text{ }\mu\text{M}$. Many databases, both public and commercial, sometimes contain qualitative annotations, in addition or instead of potency information. NBMs have been shown to be very tolerant to noise [1] and from our own experience providing as many actives as possible in a training set benefits the model more than limiting it to highly active compounds only. To illustrate these trends, for the purpose of this chapter, we have simulated virtual screening using a multiclass NBM for 89 targets covering five distinct target families from ChEMBL_11 [34]. We monitored how well a model trained on 50% of the compounds for any target using a cutoff of 0.1, 1, $10\text{ }\mu\text{M}$, or no potency cutoff at all would perform in retrieving actives—also defined using these different cutoffs—among the other 50% (Figure. 7.2). To evaluate performance, we recorded the area under the operator receiving characteristic curve (ROCAUC) for all targets and 25 independent simulations. Corroborating our general observations, a conservative cutoff of $0.1\text{ }\mu\text{M}$ performed worse than more generous cutoffs in retrieving any category of compounds, because only a limited number of highly active compounds were available for model training. Conversely, using all activity annotations, irrespective of the potency, performed overall well. Moreover, all models generally were best in enriching for highly potent compounds compared to less stringent cutoffs. This suggests that features characteristic of highly active compounds are also enriched in larger sets of compounds including highly and moderately potent ones. Combining many such enriched features prioritizes highly potent compounds. This observation is in line with the notion that combining different features preferentially found in active compounds identifies chemically distinct active molecules that contain combinations of activity-class characteristic features not found in reference compounds [13]. A general guideline for assembling a training set thus may be:

Use as many active compounds as possible to train an NBM, rather than choosing a conservative potency cutoff.

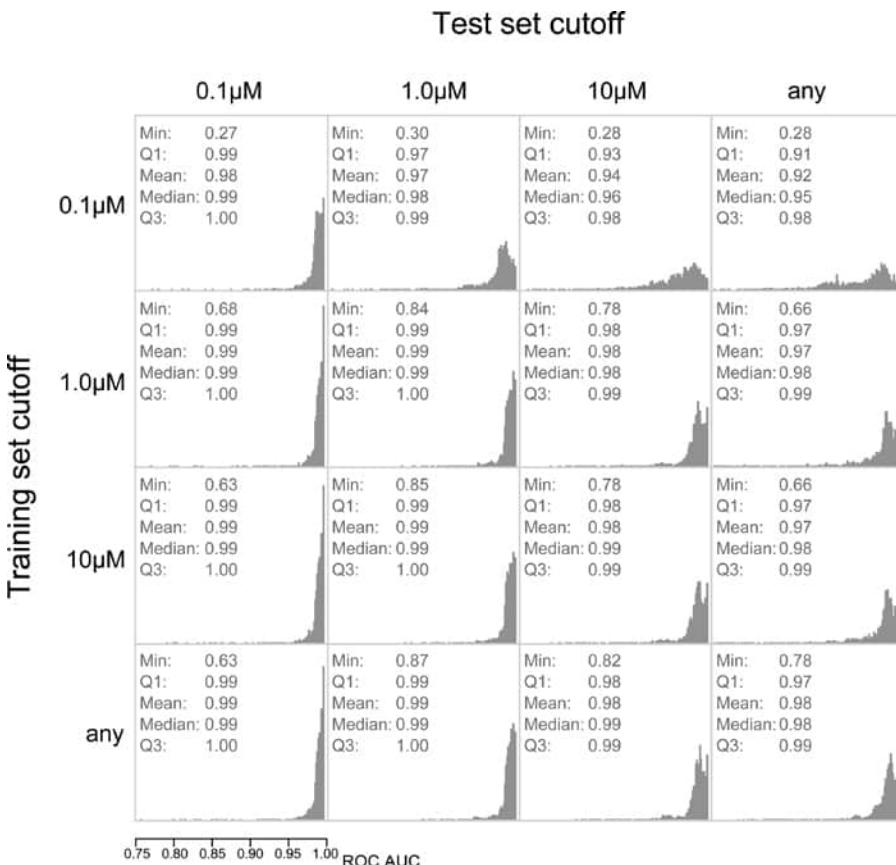


FIGURE 7.2 Guidelines for multiclass NBM for target predictions. Using different activity cutoffs to define training and test sets, we monitored the NBM performance of retrieving active compounds from ChEMBL. ROCAUC score distributions are shown for each training-test cutoff pairing. Q1, Q3: first and third quartiles. Training with less conservative cutoffs improves performance and in particular retrieves potent compounds across many activity classes.

7.4 TARGET AND PHENOTYPE COMPARISON IN CHEMICAL AND BIOLOGICAL ACTIVITY SPACE

The observation that features enriched in sets of active compounds irrespective of potency enables retrieval of highly potent compounds also supports the usage of feature weights extracted from NBM to identify compound-class characteristic molecular features. So far, we have discussed only activity classes, but any way to group molecules can be used to identify features that preferentially occur in desired groups of compounds. Other examples of compound class definition include ADME characteristics [24], toxicity [21], clinical ADRs [35], or chemical

characteristics relating to compound synthesis. Recently, a Bayesian idea generator has been described [36] that predicts the chemical library a compound is likely to come from in order to aid in the synthesis of analogs and thus facilitate SAR exploration. NBMs represent these distinct classes of compounds in form of weighted feature vectors. If the same descriptors have been used to define the chemical or biological activity space, for example, if biospectra have been used for training the model, then each model projects a compound class into a uniform chemical and/or biological space, where often disparate domains can thus be compared.

7.4.1 Comparison of Compound Classes Using Bayesian Weights

Technically, individual models have been compared using the Pearson correlation coefficient on feature weights [35, 37]. Using this metric, two models are considered similar if the same features receive high or low weights, respectively. Thus, very distinct concepts can be compared, as long as they are defined by sets of compounds. Because the models abstract from the concrete structures, the different underlying compound sets do not have to overlap, and individual compound pairs do not have to be overall similar. It is sufficient if the same chemical features are characteristic of the two different compound classes.

Multiple approaches to comparing targets based on their ligands have been described that differ in their level of abstraction. The most direct way is to connect targets that share ligands [38]. However, this requires that a large number of compounds have been screened across the entire target panel. Although this may be feasible for “prominent” compounds such as drugs, usually the data matrix is relatively sparse. The next level of abstraction is to look at ligand similarity and connect two targets if they share one or more ligands that are structural analogs. The similarity ensemble approach (SEA) [39] introduced a further abstraction step as it identifies ligand sets that are overall more similar to each other than would have been expected by chance, even if this similarity is not reflected by any single ligand pair. Whereas SEA still relies on pairwise comparison of individual ligands, NBM comparison allows for the highest level of abstraction: enrichment of chemical features is compared for different targets. By design, these features are treated independently of each other and thus can occur in different combinations in ligands of one or the other target. This allowed, for example, redefining relationships between proteases that were more informative for selectivity assessment than their phylogenetic distances [37].

Going beyond targets, phenotypes have also been related to each other using Bayesian models. For example, clinical adverse events of drugs have been projected into chemical space using weights derived from NBM trained on adverse event classes [35]. This allowed correlation of ADRs with each other as well as with (off-) targets. Again, individual drugs did not have to be tested across all targets in order to make these connections, because chemical features, rather than individual drugs, were used to derive relationships.

7.5 MINING FOR ENRICHED FEATURES AND INTERPRETING THEM

With increasing amount of bioactivity data available for compounds, compound activities can also serve as features. For example, *in vitro* binding can serve as a fingerprint, where each target represents a feature and a compound has the feature present if its activity at the target falls below a threshold, for example, 10 μM. Because NBM can handle a large number of descriptors, recently amino acid residues of kinases have been combined with retrosynthetically derived ligand fragments to mine for enriched amino acid–fragment pairs [40]. Rather than predicting potential targets from compound structure, existing knowledge about the compounds can thus be utilized. Because a screened compound rarely is highly selective for one single target only, NBM-derived feature (i.e., target) weights can be used to identify targets characteristic of compounds that are active in a particular phenotypic assay. This approach is taken on a regular basis in our group to elucidate MOA of phenotypic assay hits with rich bioactivity annotations. For example, imagine some kinase inhibitors with overlapping yet distinct selectivity profiles inhibit the growth of a cancer cell line, whereas other kinase inhibitors are inactive. Then feature (target) weights derived from a model trained on active versus inactive compounds will point to the common protein target that separates active from inactive compounds. NBMs therefore allow for less stringent selectivity criteria for tool compounds at the cost of extensive knowledge about the tool compound’s activity profile and a larger number of compounds screened.

Descriptor selection as described earlier can easily be generalized to other biological activity, for example, descriptors derived from high content screening, secretomics, or, in the case of drugs and drug candidates, clinical data. The derived models may not always be useful for the typical application, that is, compound scoring, because they may be over-fitted if only small numbers of compounds are present for training. However, they still can be used in a descriptive fashion to rank and focus the attention on properties that discriminate between wanted and unwanted compounds.

Although a model can yield many information-rich features that well separate a class of compounds from another, sometimes it is hard to find an overarching property subsuming the enriched features. Even mapping them onto molecules may be of limited use if one seeks to find rule of thumb-like [41] properties that describe a class in an intuitive manner. Next, we describe an approach that utilizes NBM in combination with well-defined “toy classifiers” that each focus on a set of descriptors representing an intuitive property that reflect medicinal chemists’ preferences and can be used as a medchem guideline for selecting follow-up candidate molecules [42].

7.5.1 Understanding Chemist’s Chemical Preferences

In addition to the many well-defined properties of a putative lead compound that might make it desirable, such as affinity for a target, solubility, or size, there is a relatively intangible aspect of a lead compound which can prove just as important: whether a medicinal chemist will be interested in carrying the compound forward

and perform exploratory chemistry around it. It would be desirable to know *a priori* whether discovery chemists would be willing to work on a scaffold, as this knowledge might impact the design of chemical libraries for both virtual and experimental screening. This question was recently addressed at Novartis by asking chemists to assess fragments ($MW < 300$) and select compounds that they would be willing to carry forward in a drug discovery campaign [42]. Semi-naïve Bayesian models (SNBMs) were then built to understand their decisions. SNBMs have been developed in order to retain the interpretability of their predecessor, NBM, while allowing for attributes to be considered jointly [43]. In this case, the categories of compounds that the models were tasked with predicting were defined by the chemists as selected (desirable) or unselected (undesirable).

One goal of this project was to illuminate the most salient, well-defined properties of compounds that impacted individual chemist's selections. This is in contrast to the more conventional goal of training an accurate model using many different features and then investigating the enriched ones. We sought to answer questions such as whether a chemist selected compounds based on specific parameters such as hydrogen bonding groups, size, or ring topology. As such, the descriptors that we used to build models were mapped to one or more general parameters (e.g., number of atoms and molecular weight both mapped to the parameter "size"). A feature subset selection was used to ensure that models only used the parameters that were significant in selections. Furthermore, many of the descriptors were considered jointly, allowing complex patterns that result from dependencies between attributes to be perceived by the models.

In our approach, 192 models were built for each chemist, based on their compound selections. Each of the models was built using one or more medicinal chemistry relevant descriptors. From these models, the most predictive one-parameter model (i.e., having the greatest ROCAUC for a test set) was initially selected (Figure. 7.3a). If a two-parameter model was significantly more predictive (ROCAUC increased >0.009), then it was selected and replaced the previous model. Similarly, models with more parameters were selected only if the ROCAUC significantly increased when compared to the current model. This ultimately resulted in a model for each chemist that only consisted of parameters that were statistically significant in their selections. We found that an increase in ROCAUC score of >0.009 when adding a parameter to an existing model was a useful cutoff based on identifying the correct parameters for the simulated classifiers (described in the following).

In order to validate this approach, we first tested it on simulated classifiers meant to simulate chemist selections by selecting desirable compounds based on only a few predefined parameters (e.g., size) and value preferences (e.g., all compounds with >10 atoms are desirable). A given amount of noise was also included in each classifier, to mimic human error. SNBM were then built to try to reproduce each simulated classifier's selections, and to extract what parameters were actually used by the classifier. Importantly, this strategy was able to extract all the parameters that were actually used by the simulated classifiers in an automated fashion, and did not identify any parameters that were not used as important (Figure. 7.3b).

Applying this method to the actual chemists' selections afforded a number of striking findings. We learned that out of dozens of possible parameters, chemists

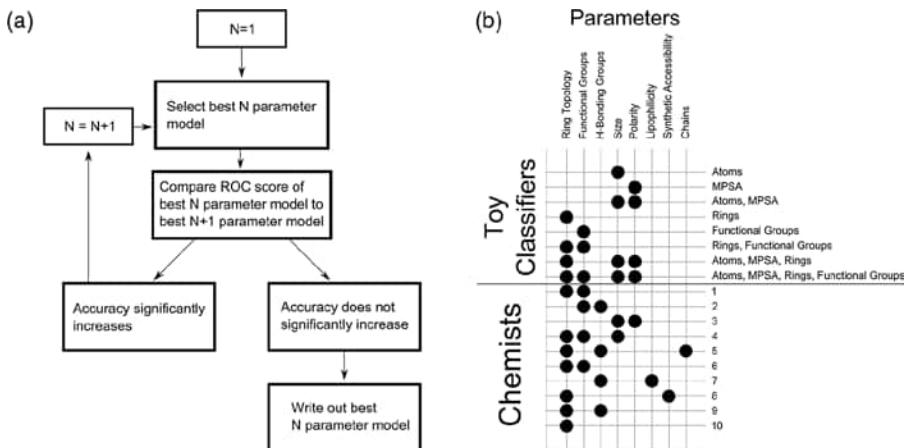


FIGURE 7.3 (a) Selection of minimal Bayesian model. N is set to 1, and the best N parameter model is selected. It is then compared to the best $N+1$ parameter model. If the ROC score of the best $N+1$ parameter model is significantly more accurate than the current best N parameter model (difference >0.009), then N is incremented, and the process is repeated. If not (difference <0.009), then the current best N parameter model is selected. (b) The parameters extracted from the most accurate minimal Bayesian models for Toy Classifiers (top) and Chemists (bottom). The properties used by each Toy Classifier (such as the number of atoms) are reported. For each Toy Classifier, the parameters identified by the minimal Bayesian models are in good agreement with the properties that the classifier used. The minimal Bayesian models built on each chemist [1–10] reveal that most chemists use 1–2 parameters when selecting compounds.

typically used 1–2 parameters (Figure. 7.3b), reducing a massively complex problem to a few dimensions. Furthermore, we identified an interesting pattern of agreement and disagreement between chemists: they tended to use the same parameters to select compounds (such as ring topology), but differed in their value preferences for these parameters, which ultimately led to an overall lack of agreement in the selected compounds. Finally, we observed that chemists were largely unaware of the parameters that were most predictive of their selections, and on average reported far more parameters (~7) than our models identified (~2) as important for distinguishing desirable from undesirable compounds.

7.6 SHORTCOMINGS OF NBM

While NBM has a number of strengths that have secured its place in the current toolbox of drug discovery, there is a major shortcoming of NBM that a discovery scientist should be aware of in order to apply NBM in the most beneficial manner: problems that can be addressed using NBM are constrained by the amount and type of relevant data that can be used for model training. This has a number of

consequences, of which the following are of utmost relevance to practical drug discovery in the pharmaceutical industry.

First, if the most likely biological target of a chemical with a *novel* chemotype is sought, ligand-based NBM is likely to perform poorly. This is especially detrimental when a new chemotype is proposed (e.g., by a chemist, from a *de novo* design effort, or from virtual library enumeration), and the task is to determine the biological relevance of the chemotype. Similarly, if chemical matter is sought for a new target with no known ligands, NBM is not directly applicable. This can be especially undesirable as pharmaceutical companies shift their discovery efforts to new target classes. For example, if a company that has historically focused on kinases shifts its focus to RNA targets, models trained on historical in-house data are no longer applicable. There are creative ways of dealing with new targets, for example, if there is screening data on a similar target an NBM could be built with this data and used to suggest compounds that might perturb the target of interest. Another way to address new targets is to run a small experimental screen with a diverse set of compounds against the unknown target, and train an NBM that can be used to identify compounds for a follow-up screen.

In other cases, the target class or chemotype might not be entirely missing, but specific targets or chemotypes are not uniformly sampled in the data. For example, if one chemical series has been screened against one target, and another series has been screened against another target, the respective NBMs may differ not only because of the innate differences in the targets, but also because of the differences between the two chemical series. Similarly, if a chemotype is underrepresented in the training set, the resulting NBM will not be able to accurately predict how this chemotype will interact with biological targets. Likewise, if activity information is preferentially available for some targets and not others, compounds will be more often predicted to bind these targets. As an extreme example, let us assume a company uses the same library of compounds on all in-house screens, but focuses the bulk of screening efforts on 100 kinases, and only screens 5 select GPCRs. A model trained on these data will more often predict activity at a kinase rather than a GPCR, simply because GPCRs are not covered by the training set.

From the earlier discussion, it is clear that breadth, diversity, and uniformity of data are desirable. As a result, resources must be carefully considered when applying NBM. Indeed, for practical purposes, the accuracy and applicability of an NBM is related to the data that a user has direct access to. For example, if a multi-class Bayesian model is built at a large pharmaceutical company using rich proprietary in-house data, large commercially available databases, and freely available databases, then its accuracy and breadth of applicability will probably be much greater than a model built using only freely available data. Even so, there will always be some bias in the underlying data for most practical applications of NBM.

7.7 SUMMARY

In this chapter, we have touched upon the many faces and applications of NBM. Although NBM is but one machine learning technique among many, and is sometimes surpassed in accuracy by more elaborate approaches, its robustness, speed, and

versatility make it a good choice for real-world virtual screening, target fishing, and MOA elucidation. Together with other methods [38, 39], NBMs have enabled the reorganization of drug targets based on their ligands [37] and the linking of phenotypes such as clinical adverse events to molecular structure [35]. At the heart of NBMs lie enriched features, which can be of interest themselves, or can be combined with molecular similarity and knowledge-based approaches to elucidate SAR and other structure–property relationships.

ACKNOWLEDGMENTS

PSK is a presidential postdoctoral fellow supported by the Education Office of Novartis Institutes for Biomedical Research.

The authors thank Jeremy L. Jenkins and Florian Nigsch for helpful discussions on data quality requirements.

REFERENCES

1. Glick M, Klon AE, Acklin P, et al. *J Biomol Screen* 2004;9:32–36.
2. Glick M, Jacoby E. *Curr Opin Chem Biol* 2011;15:540–546.
3. Mayr LM, Bojanic D. *Curr Opin Pharmacol* 2009;9:580–588.
4. Nidhi MG, Davies JW, Jenkins JL. *J Chem Inf Model* 2006;46:1124–1133.
5. Rogers D, Brown RD, Hahn M. *J Biomol Screen* 2005;10:682–686.
6. Xia X, Maliski EG, Gallant P, et al. *J Med Chem* 2004;47:4463–4470.
7. Crisman TJ, Bender A, Milik M, et al. *J Med Chem* 2008;51:2481–2491.
8. Biniaishvili T, Schreiber E, Kliger Y. *J Chem Inf Model* 2012;52:678–685.
9. Vogt M, Godden JW, Bajorath J. *J Chem Inf Model* 2007;47:39–46.
10. Kappler MA. *Curr Opin Drug Discov Devel* 2008;11:389–392.
11. Bender A, Glen RC. *Org Biomol Chem* 2004;2:3204–3218.
12. Bender A. *Methods Mol Biol* 2011;672:175–196.
13. Hu Y, Lounkine E, Bajorath J. *Chem Biol Drug Des* 2009;74:92–98.
14. Bender A, Jenkins JL, Glick M, et al. *J Chem Inf Model* 2006;46:2445–2456.
15. Tiikkainen P, Franke L. *J Chem Inf Model* 2012;52:319–326.
16. Kido H, Okumura Y, Takahashi E, et al. *Biochim Biophys Acta* 2012;1824:186–194.
17. Bhargava KP, Dixit KS, Palit G. *Br J Pharmacol* 1976;57:211–213.
18. Curran ME, Splawski I, Timothy KW, et al. *Cell* 1995;80:795–803.
19. Rothman RB, Baumann MH, Savage JE, et al. *Circulation* 2000;102:2836–2841.
20. Hamon J, Whitebread S, Techer-Etienne V, et al. *Future Med Chem* 2009;1:645–665.
21. Nigsch F, Lounkine E, McCarren P, et al. *Expert Opin Drug Metab Toxicol* 2011;7:1497–1511.
22. Rogers D, Hahn M. *J Chem Inf Model* 2010;50:742–754.
23. Watson P. *J Chem Inf Model* 2008;48:166–178.

24. Sun H. J Med Chem 2005;48:4031–4039.
25. Vogt M, Bajorath J. Chem Biol Drug Des 2008;71:8–14.
26. Glick M, Jenkins JL, Nettles JH, et al. J Chem Inf Model 2006;46:193–200.
27. Chen B, Sheridan RP, Hornak V, et al. J Chem Inf Mod 2012;52:792–803.
28. Wawer M, Lounkine E, Wassermann AM, et al. Drug Discov Today 2010;15:630–639.
29. Wang Y, Bajorath J. J Chem Inf Model 2008;48:75–84.
30. Scior T, Bender A, Tresadern G, et al. J Chem Inf Model 2012;52:867–881.
31. Lounkine E, Nigsch F, Jenkins JL, et al. J Chem Inf Model 2011;51:3158–3168.
32. Heller S, McNaught A, Stein S, et al. J Cheminform. 2013;5:7.
33. Williams AJ, Ekins S. Drug Discov Today 2011;16:747–750.
34. Gaulton A, Bellis LJ, Bento AP, et al. Nucleic Acids Res 2011;40(Database issue): D1100–1107.
35. Scheiber J, Jenkins JL, Sukuru SCK, et al. J Med Chem 2009;52:3103–3107.
36. van Hoorn WP, Bell AS. J Chem Inf Model 2009;49:2211–2220.
37. Sukuru SCK, Nigsch F, Quancard J, et al. Protein Sci. 2010;19:2096–2109.
38. Yildirim MA, Goh K-I, Cusick ME, et al. Nat Biotechnol 2007;25:1119–1126.
39. Keiser MJ, Roth BL, Armbruster BN, et al. Nat Biotechnol 2007;25:197–206.
40. Niijima S, Shiraishi A, Okuno Y. J Chem Inf Model 2012;52:901–912.
41. Gleeson P, Bravi G, Modi S, et al. Bioorg Med Chem 2009;17:5906–5919.
42. Kutchukian P, Vasilyeva NY, Xu J, et al. PLoS One. 2012;7:e48476.
43. Konenko I. Semi-naïve Bayesian classifier. *EWSL-91: Proceedings of the European Working Session on Learning on Machine Learning*. Heidelberg: Springer; 1991. p 206–219.

CHAPTER 8

CHEMOINFORMATICS IN LEAD OPTIMIZATION

DARREN V. S. GREEN and MATTHEW SEGALL

8.1 HISTORICAL INTRODUCTION

Lead optimization is a critical component of the drug discovery process. Typically, a “hit” or a “lead” molecule is found by some type of screening process (the diversity-based methods of high-throughput screening and fragment screening or some kind of knowledge-driven computational selection). The lead will have some affinity for the target protein or pathway, but will often not satisfy the plethora of other attributes required for the molecule to be a medicine: solubility and dissolution rate, drug metabolism and pharmacokinetics (DMPK), and safety or stability. *Ab initio* prediction of such parameters, particularly in humans, is a long-term aspiration and a monumental challenge. For the foreseeable future, leads will need to undergo some form of optimization, commonly defined as

The synthetic modification of a biologically active compound, to fulfill stereo electronic, physicochemical, pharmacokinetic and toxicological clinical usefulness. [1]

Typically, this will involve the synthesis of some hundreds of analogs, in tens of iterative steps [2]. In this chapter, we will explore how iterative lead optimization might be guided and made optimally efficient by thinking of the process in terms of a mathematical problem. A significant impediment to doing so has been the historical association between lead optimization and the actual discipline of medicinal chemistry defined as:

A chemistry-based discipline, also involving aspects of biological, medical and pharmaceutical sciences. It is concerned with the invention, discovery, design, identification and preparation of biologically active compounds, the study of their metabolism, the

interpretation of their mode of action at the molecular level and the construction of structure-activity relationships. [1]

Thus, the discipline of medicinal chemistry is actually a collection of different scientific disciplines, lead optimization being just one of them. Worse, the theoretical aspects of optimization are routinely conflated with, and subjugated to, the synthetic aspects of medicinal chemistry because recruitment of medicinal chemists is almost exclusively from the ranks of synthetic organic chemistry schools. Therefore, unlike fields such as engineering, it is uncommon to see discussions on lead optimization as a separate science. Corwin Hansch, one of the founders of a scientific discipline that is very relevant to this discussion, quantitative structure–activity relationships (QSARs), wrote an exasperated final paragraph to his definitive textbook:

Shortly after QSAR was introduced in 1962 it quickly became clear to many people interested in the subject that considerable attention must be given to the design and synthesis of chemicals for biological testing. It is amazing to peruse the journals over 30 years later to see how poorly most data sets are constructed. Without proper design it is simply impossible to attain real understanding of structure-activity relationships. So many of those people controlling the synthesis have little appreciation for experimental design. Hunches and intuition still control much of the synthetic side of drug research. [3]

Fortunately, there are signs of a resurgent interest in the field [2, 4, 5], and these approaches are discussed in Section 8.3. First, it is necessary to review the early work in this area and to understand why such methods did not initially have greater impact on the practice of lead optimization.

The development [6] and application [3] of QSAR methods quickly led researchers to explore whether such techniques might be used to plan iterative synthetic modifications in order to optimize the probability that the analogs might yield new SAR information [7, 8]. This thinking yielded standard tools for a generation of medicinal chemists in the Craig plot [9] and Topliss Tree [10]. The Craig plot (Figure 8.1) is a simple visualization of physicochemical property space (typically the Hansch π and σ parameters) that can inform the user as to which substituents are similar and which would elicit new SAR by probing very different chemical space.

The Topliss Tree is an enduring tool for the iterative optimization of a chemical series where it is not possible to synthesize all close analogs (i.e., for all medicinal chemistry projects); Figure 8.2 illustrates the practical application of the method. The continued popularity of the Topliss Tree (although, in the authors' view, the tool is still underused by medicinal chemists) is perhaps explained by the simplicity and visual application of the approach. However, the method is without a strong mathematical framework, which limits its applicability to more complex problems, such as multiparameter optimization (MPO).

The work of Martin and Panas [11] sought to introduce such a mathematical basis and laid down some criteria for efficient series design:

1. The analogs should be synthetically feasible.
2. The series should contain enough variation in the properties, which may influence potency.

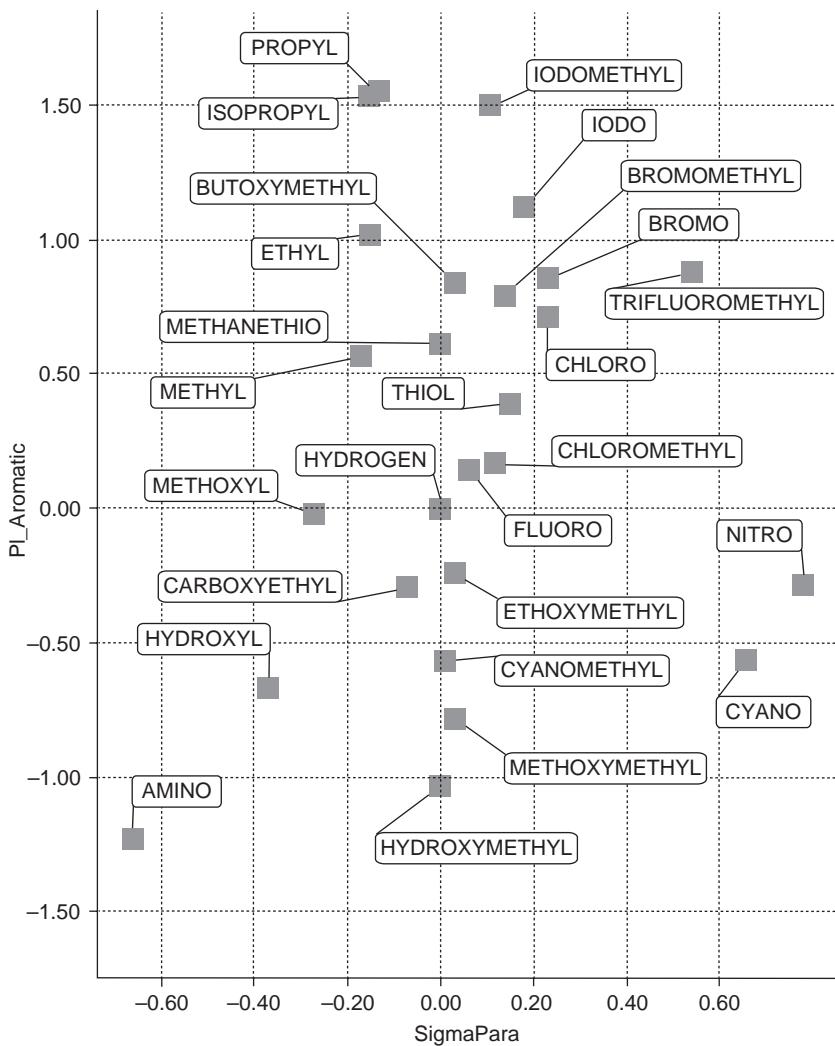


FIGURE 8.1 A Craig plot showing the Hansch σ and π values for aromatic ring substituents.

3. These properties should be varied independently of each other.
4. The series should be the minimum acceptable size, that is, each analogue should contribute unique information.

Austell combined these concepts with the statistical method of factorial design to produce a systematic method for lead optimization [12]. This method follows classical response surface modeling [13] used in other industries and allows an optimal experimental design of substituents on a molecule, which (assuming the Hansch parameters of σ and π are related to the biological response) will provide nonredundant SAR information that can be used to design the next iteration of analogs (Figure 8.3).

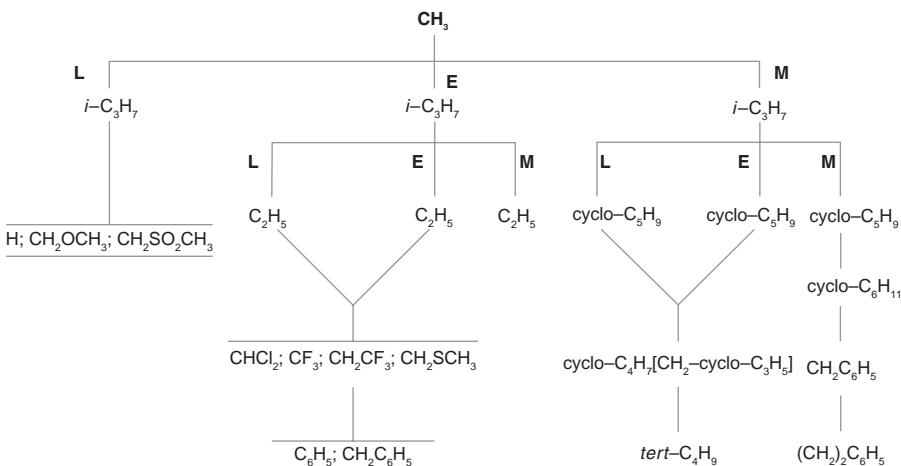


FIGURE 8.2 A Topliss Tree for aliphatic side chain substitutions. The Topliss schemes were constructed by consideration of hydrophobic and electronic factors and are designed such that the optimum substituent may be found as efficiently as possible. It is assumed that the methyl substituted compound has been made, tested, and compared to the unsubstituted compound. There are three possibilities: the analogs will have less (L), equal (E), or more (M) activity and this determines which branch of the tree should be followed next.

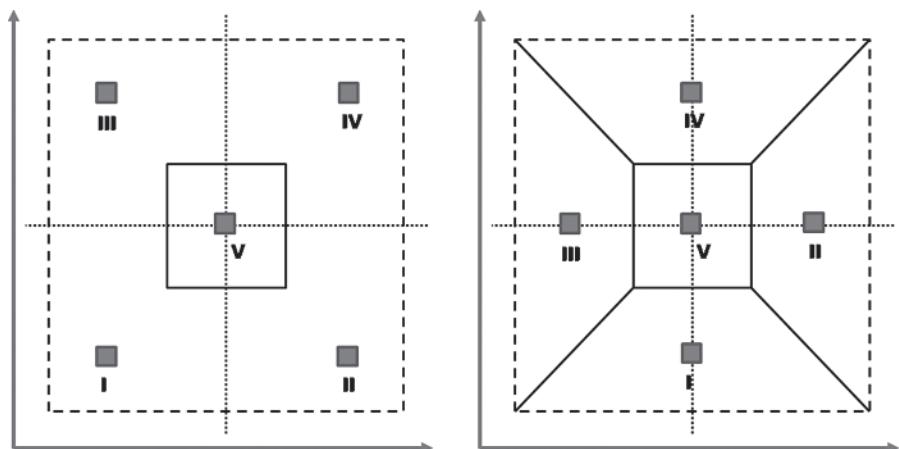


FIGURE 8.3 An illustration of the rotated factorial design scheme. The left-hand side is a traditional factorial design. One substituent would be sampled from each area marked I–V. For common parameters, for example, molecular size and lipophilicity, it is not possible to populate parts of the graph (e.g., a very small compound with very high or low lipophilicity). However, rotation of the design by 45° (right-hand side) makes the statistical design amenable to chemical applications.

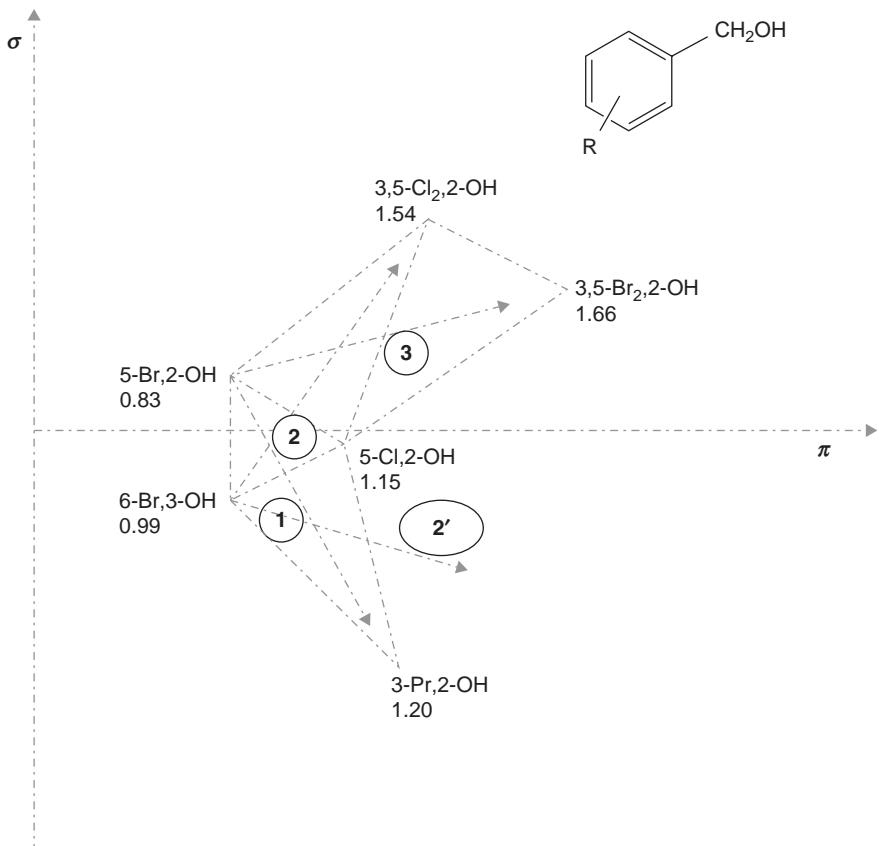


FIGURE 8.4 An illustration of the simplex optimization method for bacterial benzyl alcohols using σ and π parameters. The 5-bromo,2-hydroxyl,6-bromo,3-hydroxyl and 5-chloro,2-hydroxyl analogs have been synthesized and tested. Projection [1] from the least active of the three molecules (5-Br,2-OH) leads to the σ - π space for 3-propyl,2-hydroxy, which is found to be the most active compound yet. Projection (2') from the base of the new triangle leads to an area of space that cannot be occupied using available substituents, and therefore an alternative projection [2] is used, leading to the 3,5-dichloro,2-hydroxy analog. This analog is found to have increased activity and therefore defines a new direction [3] leading to the most active compound 3,5-dibromo,2-hydroxy benzyl alcohol.

Darvas [14] had previously taken an important conceptual leap and applied a mathematical technique for optimization—the simplex optimization method—to the lead optimization of natriuretic sulphonamides. By describing molecules by their Hansch parameters (σ and π), a common molecule space could be created to describe the series and this space could be “walked” by the optimization algorithm (Figure 8.4) using the following steps:

1. A “parent compound” is chosen whose activity is to be improved, and the sites of change (substitution) are selected in the molecule.
2. The parent compound is located in the appropriate (σ - π) coordinate system or other coordinate systems used in the Hansch-type procedures.
3. Points corresponding to derivatives, which presumably possess the desired effect, are determined in the coordinate system. In compiling the list of substances, the special aspects of the synthetic work, e.g., biopharmacology, can be taken into account.
4. Two compounds are selected from the derivatives near the parent compound, prepared, and tested. These are chosen following the Topliss approach [10]; that is, by looking for points where one of the parameters is systematically increased or decreased with respect to the original substance, the other remaining constant as far as possible.
5. On the basis of the activities of these three compounds, a decision is made about the next derivative to be prepared. The basis of the decision is always the simplex, that is, the triangle formed by the three substances. The directional derivative is determined practically by a plane geometrical procedure; the point of the less effective derivative is connected with the midpoint of the opposite side of the triangle, and the new, supposedly more effective compound is searched for in this direction.
6. The mechanical repetition of this procedure leads to the preparation of substances having outstanding activities as compared to their neighbors. This maximum is surrounded stepwise by simplexes, since the point with the highest effect in the triangle is always involved. If we wish to continue the optimization procedure, it is most preferable to include new substances in the optimization map around this maximum and to resume the optimization on a smaller scale.

Therefore, by the early 1980s, researchers had identified mathematical techniques that could make lead optimization more efficient—indeed, data from Abbott [15] suggested that by employing these techniques alongside the intuition of experienced medicinal chemists, the average number of analogs required to investigate a single physical property was reduced from 11 to 4. Why, then, did these techniques not progress to make lead optimization a mathematically driven, predictable process?

8.2 LEAD OPTIMIZATION IS A LARGE, COMPLEX, MULTIOBJECTIVE PROCESS

Many of the early QSAR-driven methods made assumptions about the behavior of the datasets to be studied and/or generated. A significant assumption is one of additivity—the effect of a substituent is constant when combined with all other types of substituent or when added to all other chemotypes. This was quickly found to be untrue in many cases. An example is two series of H₂-antagonists, which both gave rise to marketed drugs (ranitidine and cimetidine) [16], shown in Figure 8.5.

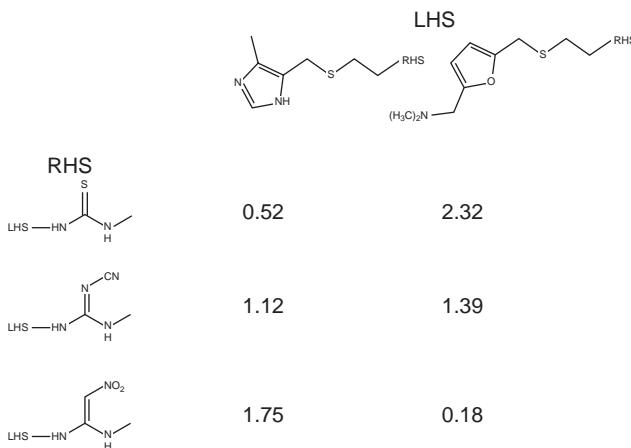


FIGURE 8.5 Nonlinear SAR from the cimetidine and ranitidine series of H2-antagonists. Activity is expressed as milligram per kilogram in a perfused rat stomach preparation.

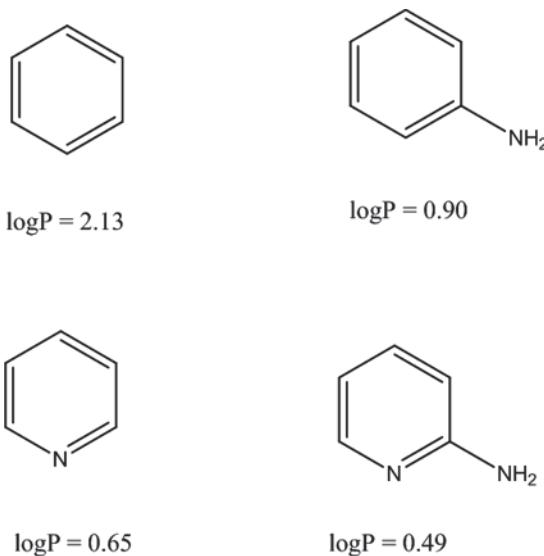


FIGURE 8.6 An illustration of how even simple molecular systems lead to nonadditive SAR. Here an aniline moiety is added to benzene and pyridine. The measured $\log P$ decreases by over an order of magnitude (1.23) for benzene, but only marginally (0.16) for pyridine.

The phenomenon is not restricted to interaction with proteins and is manifest in the properties of even the most simple chemical structures. For example, the addition of an aniline substituent to benzene and pyridine produces very different reductions in the $\log P$ of the molecules (Figure 8.6).

Dealing with such cases requires more sophisticated statistical techniques and larger datasets that were not available before the advent of higher throughput assays.

Nonetheless, these effects can be modeled by modern methods [17]. However, the adoption of higher throughput assays and the widespread embrace of molecular biology led to a different problem, in that the number of parameters to be modeled increased. Previously, when data would often come from a lower throughput, phenotypic or tissue-based assay, parameters such as solubility, membrane permeability, and such like were incorporated into the biological measurement—if the molecule could not be dissolved, or could not reach the site of action, no response would be produced. In a world of isolated, recombinant, protein assays performed in solvents such as DMSO, it would be necessary to produce QSARs for all necessary responses. Often, the simple Hansch-type molecular descriptors would not be able to produce predictive models for such a variety of molecular interactions and thus the type of simple coordinate space used in the simplex optimization procedure was difficult to produce. In order to produce more predictive QSARs, modelers began to use a wide variety of molecular descriptors—topological, binary fingerprint, shape, electrostatic fields, and so on [18]. On top of all of these challenges is the combinatorial nature of lead optimization, a particular problem as the size of molecules made in medicinal chemistry has increased. For example, using only 100 possible substituents and only three of the seven positions on a quinoline ring leads to a possible 35 million analogs. Therefore, the lead optimization problem became an explosion of endpoints to be modeled using a myriad of molecular descriptors and modeling techniques, across an almost infinite number of molecules. Even with higher throughput assays and synthesis, only a fraction of the potential molecular space may be explored.

However, it is now understood that the problem is even harder than was envisaged by the QSAR pioneers. The rich datasets that are now produced in lead optimization projects allow us to understand the true nature of the drug-receptor response surface. Understanding the response surface is critical to a correct choice of optimization strategy, for example, the simplex optimization method relies on a smooth, continuous response so that it may gradually map and climb to the maximum of the response. Lead optimization data, unfortunately, do not behave in this way. Medicinal chemists will speak of the “magic methyl” whereby a small change in the molecular structure can make a very large change in the biological response. For example, we can return to ranitidine [18] (Figure 8.7).

Although the magic methyl phenomenon is not rife in lead optimization data (if it were, the entire premise that a lead may be optimized by synthesis of close analogs would be discredited), recent studies on large datasets now reveal a consistent but low-probability effect of potency increases on the addition of a methyl group [19], the presence of activity cliffs in many biological response surfaces [20], and the extent to which assumptions of additivity are valid [21]: only 50% of the (single parameter) datasets studied exhibited additivity.

The existence of activity cliffs (a sharp change in biological activity in response to a small change in chemical structure of the ligand) in SAR is well known to practicing medicinal chemists. However, it is only recently that the phenomena have been quantified [22], in some case rationalized [23], and tools developed to aid visualization and navigation of this difficult response surface. Modern chemoinformatics methods are now able to visually articulate the effect of the phenomena on SAR data (Figure 8.8), even if it is likely to remain unpredictable.

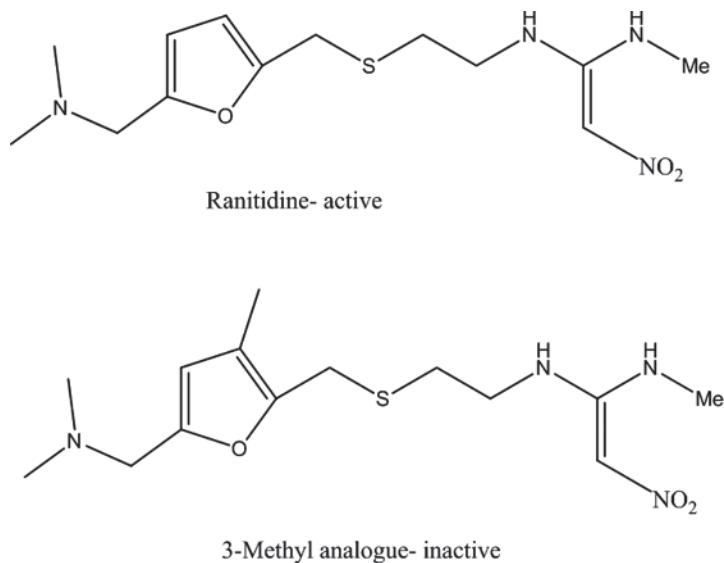


FIGURE 8.7 A simple example of the “magic methyl” effect in ranitidine. Addition of the 3-methyl substituent to the furan ring caused dramatic loss of biological activity due to a conformational effect.

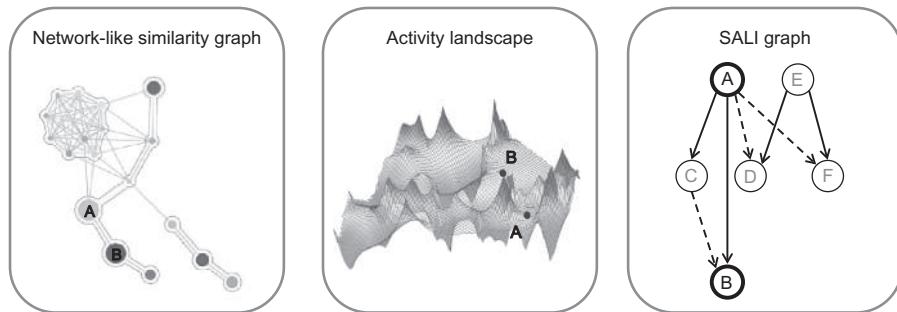


FIGURE 8.8 Three different visualizations of “activity cliffs” in a series of close analogs A–E. Compound A is mildly potent while B is highly potent, leading to discontinuity in local SAR. This is expressed by colors (red = high potency, green = low potency) and for the network-like similarity graph, size is used to show where activity cliffs occur. (Reprinted with permission from D. Stumpfe and J. Bajorath, *J. Med. Chem.* **2012**, *55*, 2932–2942, Figure 1). For color details, please see color plate section.

Therefore, the lead optimization problem can now be described as multiple endpoints to be modelled using a myriad of molecular descriptors and modeling techniques, on an almost infinite number of molecules, across a response surface that is discontinuous and nonlinear. Or, as operational research scientists would describe it [24], a *complex multiobjective optimization problem in a very large search space*.

8.3 CHEMOINFORMATICS METHODS FOR MULTIOBJECTIVE OPTIMIZATION

Fortunately, complex multiparameter (or objective) optimization problems are not unique to drug discovery and there are a wealth of techniques that have been developed in other disciplines, including engineering, quality control, aerospace, and economics. Many of these techniques have been adapted for application in the drug discovery arena, where additional challenges apply. As discussed earlier, the optimization space in lead optimization is discontinuous; while it is possible to smoothly modify variables such as temperature, pressure, mass, and so on. The smallest possible change to a molecule is the addition, removal, or substitution of a single atom and the resulting change in biological properties may be large and unpredictable. Furthermore, the complexity of biological systems means that the data on which the optimization is based have significant uncertainty due to both experimental variability and predictive error. In engineering disciplines, it is possible to measure and predict responses to parts per million. In contrast, data from biological experiments often have variabilities of a *factor* of 2, or more, while uncertainties of the order of one log unit (a *factor* of 10) are common for predictions.

In the following sections, we will discuss a range of different MPO approaches that have been applied to lead optimization, ranging from simple rule-based approaches to sophisticated search or directed optimization algorithms [4]. For each, we will discuss the relative pros and cons and give illustrative examples of their application.

8.3.1 Rules of Thumb

Numerous rules have been proposed that define criteria for simple compound characteristics to guide the design of compounds and improve their chance of success, or at least reduce the chance of encountering common problems downstream. The most widely recognized of these is, undoubtedly, Lipinski's Rule of Five [25] that identifies criteria for four properties, the octanol–water partition coefficient ($\log P$), the molecular weight (MW), and number of hydrogen bond donors and acceptors (HBD and HBA) that are satisfied by the majority of orally available drugs, specifically:

- $\log P < 5$
- $MW < 500$
- $HBD < 5$
- $HBA < 10$

Other rules have been developed for characteristics such as the number of rotatable bonds (RotB), polar surface area (PSA) [26], and fraction of sp³ carbons [27]. These have not only been related to oral bioavailability but also endpoints such as *in vivo* toxicity [28] and “developability” [27]. The rules are often considered to define the properties of “drug-like” molecules because they are often derived by examination of the properties that successful drugs have in common.

The clear advantages of these simple “rules of thumb” are that they are easy to understand and apply. The characteristics on which they are based are straightforward to calculate, violations of the rules are easy to spot, and the steps necessary to optimize a compound in order to meet the rules are clear.

However, the simplicity of these rules belies the dangers associated with overzealous application. The simple characteristics on which the rules are based have a relatively poor correlation with the biological properties with which they are associated. Intuitively, having similar characteristics to successful drugs will reduce the chance of failure by avoiding the risk associated with exploration of unprecedented areas of chemical space. However, meeting the rules is far from a guarantee of success; the vast majority of compounds that meet the rules are not drugs. Furthermore, hard cutoffs make an artificial distinction between similar compounds; does a compound with an MW of 499 have a significantly higher chance than one with an MW of 501?

It is also notable that some of the most common rules of thumb have been derived from the analysis only of the properties of successful compounds, without reference to unsuccessful compounds synthesized in the course of drug discovery projects. Therefore, they do not necessarily highlight the key characteristics that distinguish a successful compound from the wide diversity of chemistry that may be explored in lead optimization. For example, if the distribution of a characteristic is the same for both the drugs and nondrug compounds, then it does not provide any information about what makes them different; in this scenario, being “similar” to known drugs does not convey an advantage [29].

Therefore, these rules of thumb should be treated as rough guidelines that indicate when chemistry is straying beyond the bounds of characteristics for which there is a strong precedence of success, not as hard-and-fast design constraints.

8.3.2 Filters

Another very commonly applied approach for the simultaneous optimization of multiple parameters is to apply a series of filters to remove compounds that do not meet all of the defined property criteria. This approach has the advantage of being very simple to understand and apply, and the properties requiring optimization are clearly identified. However, again, the apparent simplicity hides a number of potential pitfalls.

As discussed earlier in the context of rules of thumb, applying hard cutoffs to reject compounds may draw an overly harsh distinction between similar compounds. This is particularly so in the case of experimental or predicted property data, where the uncertainty in the property values further blurs the distinction between compounds. For example, it does not make sense to reject a compound outright with a predicted solubility of 5 μM , on the basis of a minimum cutoff of 10 μM , when the uncertainty in such a prediction may be a *factor* of 10 (1 log unit). The potential for incorrectly rejecting compounds due to experimental or predicted error is compounded when combining multiple filters in sequence. For example, the chance of an ideal compound passing five property filters that are each 80% accurate is only 33%, that is, an ideal compound is twice as likely to be incorrectly rejected than accepted by this process. The opportunity cost of inappropriately rejecting good compounds on the basis of

overly hard property criteria or due to experimental or predicted error is hard to measure; we rarely have the opportunity to discover the ultimate outcome for a rejected compound. However, while the focus has traditionally been on reducing late stage failures, the opportunity cost of missed drugs may be a similar order of magnitude.

When applying filters it is also difficult to associate different degrees of importance to failures of different criteria. For example, it may be critical to meet the criterion for one property, for example, potency against the therapeutic target, while it may be appropriate to compromise on other properties in order to achieve good outcomes for critical factors. It is rare to find a “perfect” compound and, therefore, it is important to explore acceptable trade-offs.

Closely related to sequential filtering is the “traffic light” visualization of multi-parameter data, as illustrated in Figure 8.9, in which data are color-coded by whether they pass (green), fail (red), or are “close” to (yellow) the criterion for that property. This is a convenient way in which to get an overview of a set of compound data, for example, an entirely red column would indicate a consistent failure for a chemical series. However, for high-dimensional data, this approach becomes less useful; while the appropriate course of action is clear for a compound that is “green” for all properties, this is a rare result in practice and how would one choose between a compound that was “red” for two properties and another that was “red” for one and “yellow” for two? The answer depends on the importance of each of the properties and the confidence in the data on which the outcomes are based.

8.3.3 Desirability Functions

A more flexible method for combining multiple factors to compare the quality of different options was originally developed by Harrington in 1965 [30]. In this approach, a “desirability function” is defined for each parameter that maps the possible values of the parameter onto a scale of zero to one, depending on the desirability of the outcome. An ideal value for a parameter corresponds to a desirability of 1, and a totally unacceptable value has a desirability of 0. However, the desirability may take any value between these extremes, avoiding hard cutoffs or filters. Some examples of desirability functions are shown in Figure 8.10.

The desirability functions for individual properties map the property values, which may be defined in different units and ranges, onto a common scale. This allows the individual desirability values to be combined into a single, overall, measure of the quality of the compound, described as a “desirability index.” There are two common approaches for the calculation of the desirability index: additive and multiplicative. In an additive approach, the desirability values for the individual properties are added together to calculate the overall desirability index and this is often divided by the total number of properties to give an average between zero and one, that is,

$$D = \frac{1}{N} \sum_{i=1}^N d_i(x_i),$$

where D is the desirability index, N is the total number of properties, and $d_i(x_i)$ is the desirability of the value x_i of property i .

Name	Structure	pIC ₅₀	Selectivity (log)	Solubility (μM)	HLM (%loss @ 40 min)	RLM (%loss @ 40 min)
1 XXX572		9.88	1.05	138	38.5	81.8
2 XXX518		5.76	0.67	148	4.33	38
3 XXX582		6.01	1.07	137	65.1	79.9
4 XXX295		6.25	0.99	146	53	77
5 XXX311		8	0.87	183	55.8	71.8
6 XXX509		6.18	1.13	197	95.6	64.6
7 XXX292		6.28	1.22	192	84	64
8 XXX313		6.8	1.18	881	71.4	53.5
9 XXX274		5.81	0.89	124	91.9	49.2
10 XXX025		5.89	0.71	138	54.2	77.8

FIGURE 8.9 Example of a simple “traffic light” display. Here, good property values are colored green, bad values are colored red, and intermediate values are yellow. But which compound is best? For color details, please see color plate section.

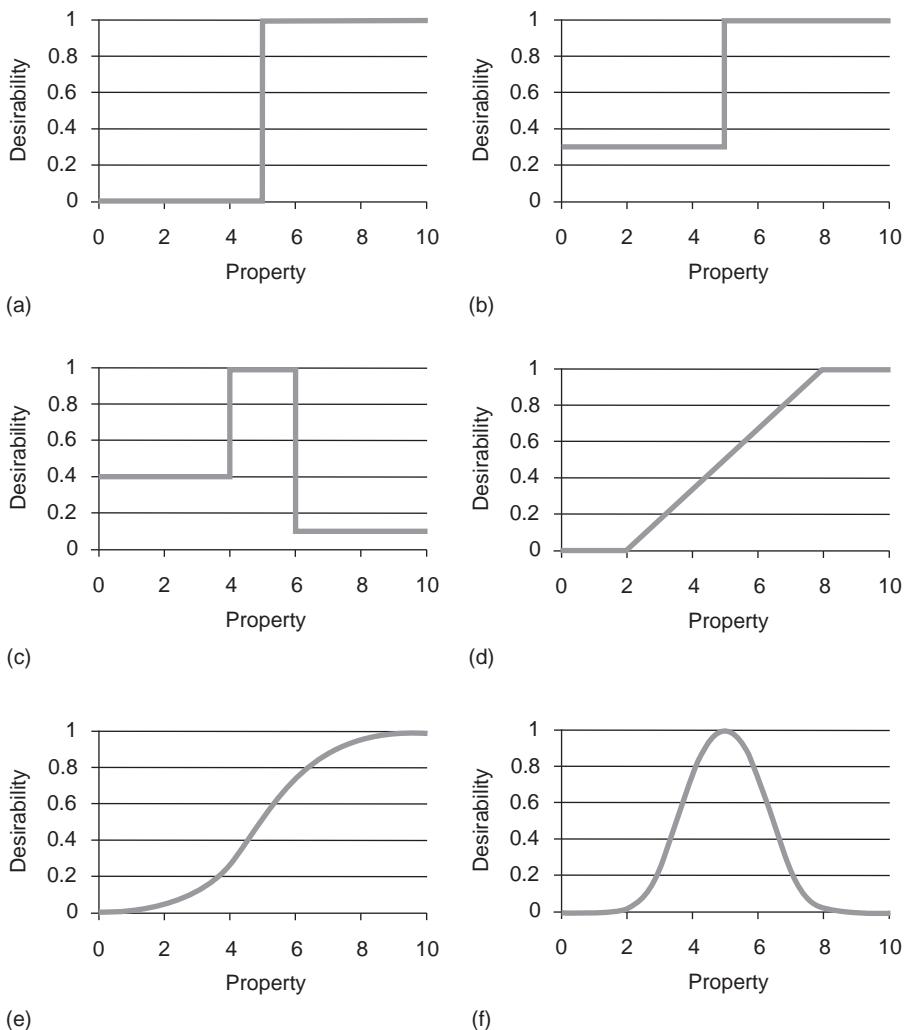


FIGURE 8.10 Examples of desirability functions. (a) Illustrates a desirability function representing a simple filter with a threshold of 5, compounds with a property value of 5 would be accepted and those with a value below 5 would be rejected; (b) represents a similar threshold, but in this case compounds with values below 5 would be less desirable, but not rejected outright; (c) corresponds to an ideal property range between 4 and 6, compounds with values above 6 are less desirable than those with values below 4. The example shown in (d) illustrates a linear increase in desirability from a value of 2 to an ideal value of 8 or above. The examples in (e) and (f) illustrate nonlinear desirability functions; (e) is a sigmoid with an inflection point at a value of 5 while (f) indicates an optimum value of 5 with a Gaussian desirability function with a standard deviation of 1 unit.

A multiplicative scheme combines the individual desirability values as a product and often the geometric mean is calculated to normalize the resulting desirability index with respect to the number of properties, that is,

$$D = \sqrt[N]{\prod_{i=1}^N d_i(x_i)},$$

which may, equivalently, be written as

$$D = e^{\frac{1}{N} \ln \left(\sum_{i=1}^N [d_i(x_i)] \right)}$$

In some schemes, a weighted average of the desirability values is used to reflect the relative importance of each property.

A multiplicative scheme provides greater flexibility to reflect the relative importance of the individual properties, in particular if there are critical criteria that must be achieved. In a multiplicative scheme, a “failure” to meet a critical criterion will result in a very low, or even zero, desirability index, which will result in the compound being “killed.” For example, if a compound has no potency against the therapeutic target, it will not usually be of interest even if the ADME properties are ideal. However, when interpreting the results it is important to note that, in a multiplicative scheme, the overall desirability index will decrease exponentially with the number of “failures,” therefore a compound that fails multiple low-importance criteria may receive the same desirability index as a compound that is perfect except for a moderate outcome for a single critical criterion. Qualitatively, it may be easier to “fix” a compound with one poor property than to solve multiple minor issues simultaneously. An additive scheme can help with this interpretation, because the desirability index will decrease linearly with the number of “failures.”

One example of the application of desirability functions is the calculation of the quantitative estimate of drug-likeness (QED) developed by Bickerton et al. [31]. In this method, desirability functions were constructed to reflect the probability distributions of eight simple characteristics of marketed, oral drugs: log P, MW, HBA, HBD, the number of RotB, the number of aromatic rings (Arom), PSA, and the number of structural alerts for potential toxic or reactive groups (Alerts). The individual desirability values were combined using a multiplicative scheme to calculate the QED. The authors explored various weighting schemes to identify the characteristics with the greatest information content. Furthermore, they demonstrated that the QED has a good correlation with subjective opinions of medicinal chemists regarding the attractiveness of compounds for optimization. The QED provides a significant improvement over traditional rule-based definitions of drug-likeness, as it eliminates the use of artificially hard cutoffs for the individual properties and allows the drug-likeness of compounds to be compared on a continuous scale. However, similar caveats apply to the interpretation of QED as to other measures of drug-likeness: The simple characteristics employed in the calculation of QED have a limited correlation with the relevant biological properties and a greater similarity to known drugs does not necessarily indicate a higher probability of becoming a drug. It is somewhat incongruous, then, that a much earlier study of drug-likeness by Gillet et al. [32] is an example of the derivation and application of complex data-derived

functions, explicitly trained to distinguish between drugs and nondrugs. Simple molecular properties were used: hydrogen-bond donors and acceptors, the numbers of RotBs and aromatic rings, the molecular weights, and the ${}^2\kappa_{\alpha}$ shape descriptors. It was found that the resulting models were very effective across a number of different datasets. Encouragingly, the models even performed well when used with empirical data, for example, assessments of attractiveness of compound structures by medicinal chemists. Recently, a general extension of desirability functions has been published which explicitly considers the relative likelihood of a compound becoming a drug given a set of characteristics [29].

8.3.4 Probabilistic Scoring

None of the MPO methods discussed earlier account for the uncertainty of the underlying data on the basis of which compounds are prioritized. The use of desirability functions can mitigate this, to some extent, by avoiding hard cutoffs associated with filters or rules. However, it does not provide information regarding the confidence in the determination of the overall score or desirability index. The probabilistic scoring approach [33] builds on desirability functions to explicitly include the uncertainty of the data in the calculation of an overall score for each compound. This clearly separates the intrinsic desirability of a property value from the accuracy with which it can be determined, which may vary from compound to compound. In doing this, it avoids inappropriate rejection of compounds based on uncertain data and the corresponding opportunity cost. Furthermore, an explicit uncertainty in the overall score can be calculated so that it is clear when compounds may be distinguished or when higher quality data are required to make a confident decision.

In the probabilistic scoring approach, a “scoring profile” is defined to reflect the profile of properties required for an ideal compound in the context of a project (an example is shown in Figure 8.11). This profile may include simple calculated characteristics, predicted properties, or experimental endpoints. Underlying each of the property criteria is a desirability function that defines the importance of the property to the overall objective of the project and the acceptable compromises. These desirability functions are defined in terms of the impact of the property on the chance of success of the compound; a low desirability indicates a low chance of success, or equivalently a high risk, due to the value of the property. Thus, the overall score will reflect the best estimate of the overall chance of success of a compound. As a probability, the overall score will be between zero and one and is multiplicative with respect to the contributions of the individual properties.

The probabilistic scores may be visualized and interpreted in many ways. For example, the plot shown in Figure 8.12 shows scores for the compounds in a dataset along with error bars indicating the confidence in each score (one standard deviation). This allows the best compounds and those that are not confidently distinguishable to be easily identified. In addition, the contributions of each property to the overall score of a compound may be visualized, either as a histogram for each compound or as a heat map as shown in Figure 8.13. The latter is a generalization of the traffic light scheme, discussed earlier, containing much more information; a property will only

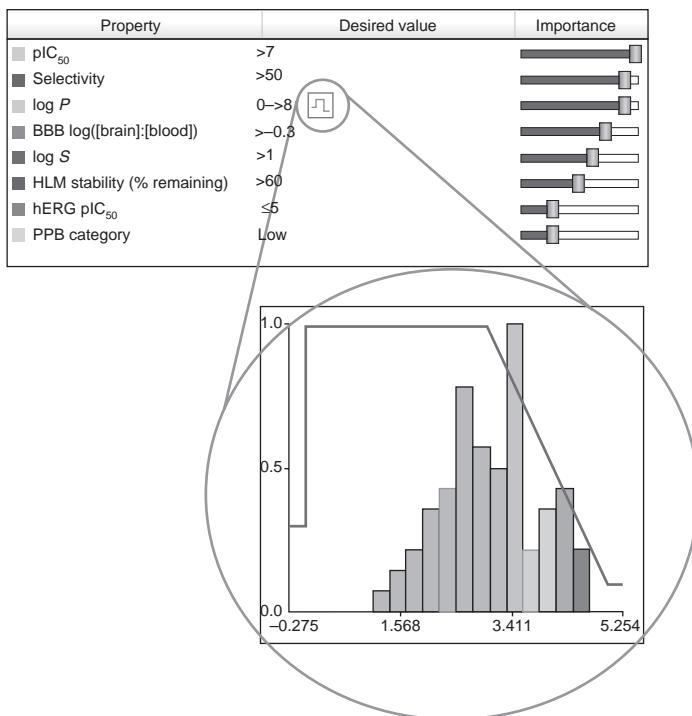


FIGURE 8.11 Example of a scoring profile. This defines the properties of interest, the ideal desired values, and relative importance of each property. Underlying each criterion is a desirability function, an example of which is shown for $\log P$. For color details, please see color plate section.

be red if the value fails to meet a very important criteria with high confidence, while a green property indicates a high confidence of achieving the ideal outcome for that property. Both of these visualizations help to guide the optimization of compounds by clearly indicating the properties for which optimization will have the greatest impact on improving the chemistry's chance of success.

An example application of probabilistic scoring is given in Section 8.4.

8.3.5 Finding the “Best” Balance of Properties

All of the methods described earlier help to guide the optimization of compounds on the assumption that the required property profile is known *a priori*. It is often challenging to determine the “best” profile for a given therapeutic objective. In general, this relies on the knowledge and previous experience of the scientists on the project team. Where the appropriate profile is not known *a priori*, informatics approaches can also help to explore potential multiparameter trade-offs. While these algorithms do not directly guide the optimization of new compounds, they help to define the property criteria and their relative importance that can, in turn, guide the design of compounds for a given therapeutic objective.

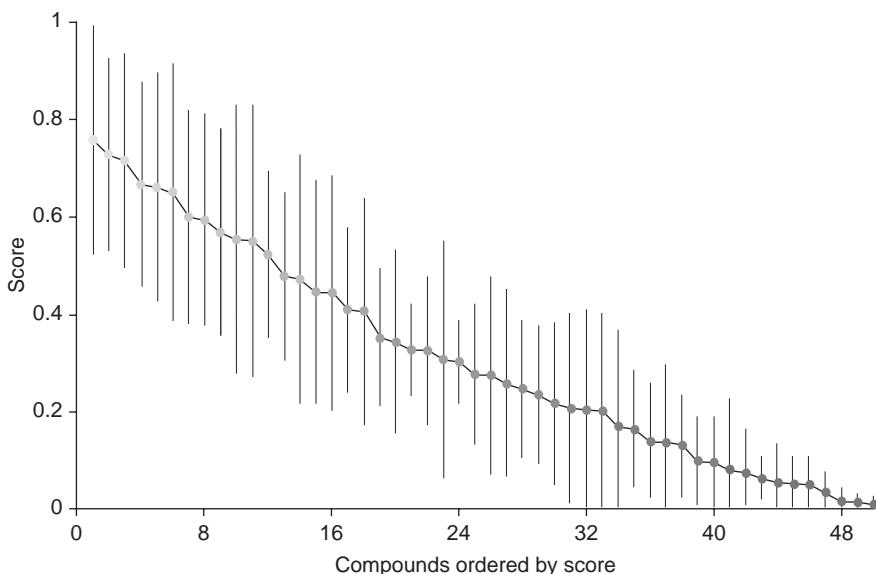


FIGURE 8.12 An example output from probabilistic scoring for 50 compounds. The compounds are ordered from left to right along the *x*-axis in order of their score and the overall score for each compound is plotted on the *y*-axis. The overall uncertainty in each score (1 standard deviation), due to the uncertainty in the underlying data, is shown by error bars around the corresponding point. In this case, the error bars of approximately the top 20 compounds overlap indicating qualitatively that these compounds cannot be confidently distinguished by the available data. For color details, please see color plate section.

Pareto optimization [34], originally conceived by the Italian economist Vilfredo Pareto, is a method for selection of multiple solutions (corresponding to compounds in our application) that each represent different, optimal combinations of parameters. A Pareto optimal solution is one that is not bettered in all parameters by any other solution. The family of Pareto optimal solutions defines a surface in the MPO space (known as the “Pareto front”), as illustrated in Figure 8.14. Selecting compounds from the Pareto front for further investigation explores the trade-offs between the different parameters. Studies with downstream experiments can reveal the “best” balance to achieve the objective of the project, which can then be used to design or select additional compounds when the ideal values of all parameters cannot be attained simultaneously. One example, illustrated in Figure 8.14, is the optimization of *in vitro* potency and membrane permeation to achieve activity in a cell-based assay for an intracellular target. Ideally we would like to identify a highly potent compound with high permeability (e.g., MDCK P_{app}); however, often optimization of potency is found to have a negative impact on membrane permeability and vice versa. In these cases, it is often not clear if optimization of potency is more, or less, important than optimization of permeability. Selecting the Pareto optimal compounds and testing these in a cell-based assay can reveal the best balance of these factors for future optimization.

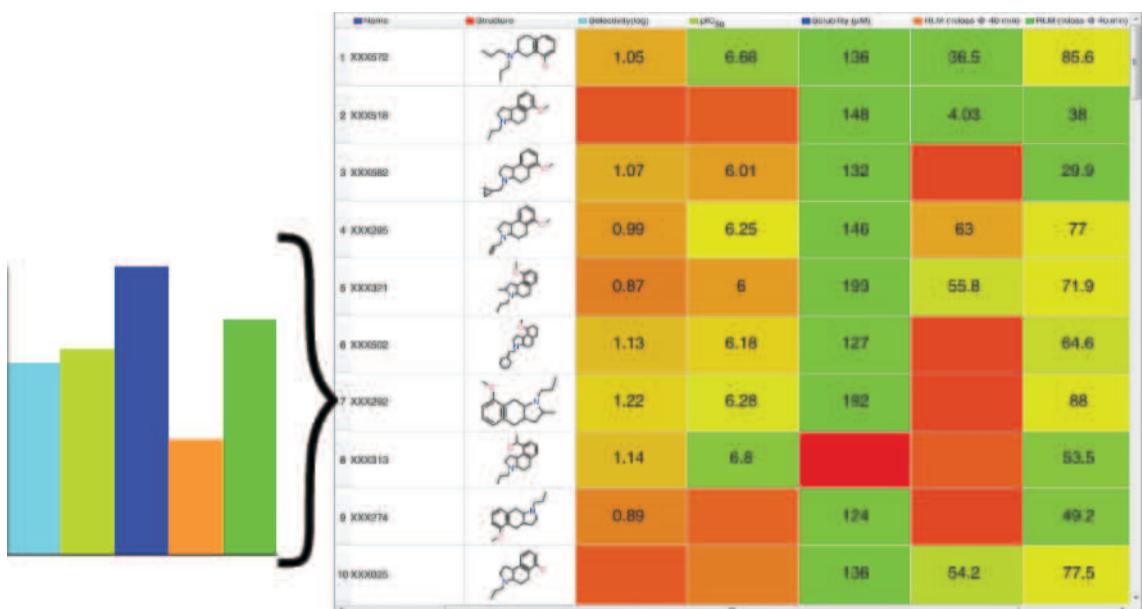


FIGURE 8.13 Examples of approaches to visualize the results from probabilistic scoring. The colors on the heat map on the right reflect not only the values of the data relative to the success criteria but also the importance of the property and the confidence of the outcome. A red cell indicated a poor result for an important property with high confidence. A green cell represents a good outcome for a property with high confidence. For comparison, this is the same data as shown in Figure 8.9. The contributions of each property to the score for a single compound can also be represented as a histogram, as illustrated to the left. Here the height of the bar reflects the impact of the property and, similarly, a low bar indicates that a significant issue has been identified with confidence; in this case, the orange bar indicates an issue which requires attention, corresponding to human liver microsomal stability (HLM). For color details, please see color plate section.

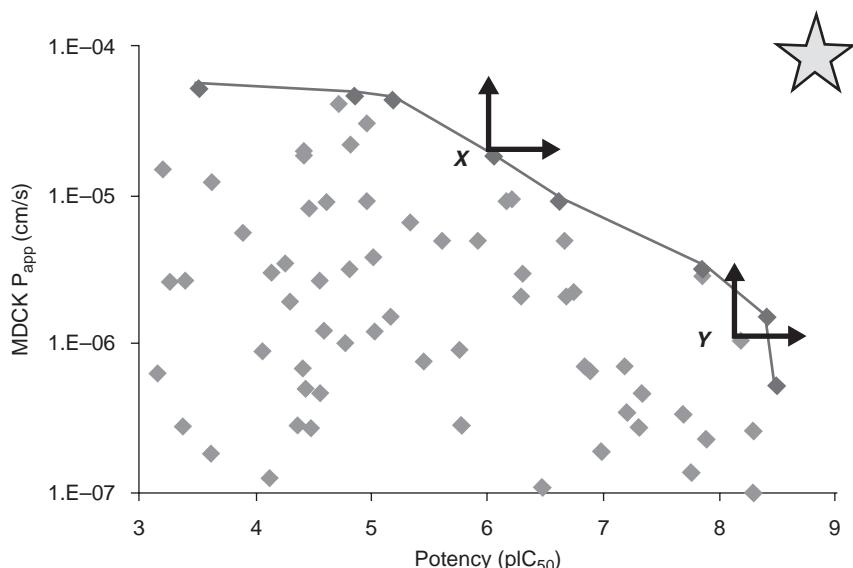


FIGURE 8.14 Illustration of Pareto optimal compounds for two-dimensional optimization of potency (pIC_{50}) and permeability ($\text{MDCK P}_{\text{app}}$). Each point represents the potency and permeability of a compound. The “ideal” compound would have both high potency and permeability as represented by the gold star. The red points, for example the point labeled X, are Pareto optimal, that is, there are no compounds better in both properties. The point labeled Y is not Pareto optimal, the point discussed earlier and to the right is better in both properties. The Pareto optimal compounds define the Pareto front, shown by the red line. For color details, please see color plate section.

A limitation of Pareto optimization is that the number of compounds on the Pareto front increases exponentially with the number of parameters, making it impossible to evaluate all of the optimal solutions. In practice, this limits the routine use of Pareto optimization to approximately four or less simultaneous parameters. One approach to overcome this challenge is to combine multiple, related properties into a single optimization parameter, for example, multiple, ADME-related properties can be combined using a desirability index and the balance of ADME score with potency explored with Pareto optimization to find the best balance to achieve *in vivo* efficacy.

An alternative approach to determining a good property profile with which to identify high quality compounds is by the analysis of historical data. For example, the QED and/or rules of thumb discussed earlier were derived from careful statistical analysis of the common characteristics of successful compound, in some cases contrasting these with compounds that failed for the same objective. However, manual analysis of complex, high-dimensional data to identify the key criteria for success is intractable and informatics approaches can help to guide this analysis. Methods described as “bump hunting” [35] have been developed in the field of machine learning that identifies regions in multidimensional parameter spaces with a higher probability of achieving a desired outcome. The application of these to drug discovery

data is at an early stage [36] and poses significant challenges due to the relative sparseness and variability of the available historical data.

Finding an appropriate property profile to identify high-quality compounds is similar to building a statistical model to predict the outcome for a compound from a set of measured or calculated properties. However, the goals differ in two important ways: Unlike a “black box” model that outputs a prediction for a compound with little explanation, the property criteria should be easily interpretable, for example, an optimal property range. This helps to guide the optimization of new compounds to improve their chance of success. Furthermore, easily interpretable and adjustable criteria allow experts, with an intimate understanding of the underlying biological and disease mechanisms, to influence the process and achieve a blend of scientists’ experience with computers’ ability to analyze complex data. Secondly, the resulting profile should indicate the relative importance of each property criterion. This helps to identify the critical data that distinguish good compounds from poor early in the process and prioritize the most important experiments, while identifying unnecessary measurements that add little value to compound optimization decisions. This, in turn, will further improve the efficiency of lead optimization.

8.4 CASE STUDIES

8.4.1 Retrospective Analyses

There are considerable difficulties in the implementation of the mathematical optimization procedures described earlier in a medicinal chemistry program. As mentioned in Section 8.1, the capricious nature of synthetic chemistry can confound the best experimental design. There are additional practicalities that also need to be confronted, for example, the necessary use of screening cascades, where more expensive or lower throughput assays are only performed on a subset of molecules. Any project with an *in vivo* model will also be constrained by ethical considerations to only test molecules with a good chance of progressing to be a drug candidate. The different capacities and turnaround times of assays therefore add complexity to the design of the optimization iterations. In order to keep the project progressing at a reasonable pace, some molecules may need to be synthesized “at risk,” that is, before useful information is available [37]. A retrospective study of several lead optimization programs at GlaxoSmithKline [38], alongside a refreshingly honest account from two senior chemists [39], provides some insight into the behavior of experienced research teams in a pharmaceutical company. The Pareto front of the candidate profile was monitored during the lead optimization process. For one project, the eventual candidate was identified quite early. However, the team continued to work for a long period afterwards. During this time, very few molecules [5] extended the Pareto front (Figure 8.15). On investigation, a dramatic drop in the number of compounds with a complete profile was found to have occurred (Figure 8.16). This happened because the team, having found molecules with a very high potency, decided to impose a higher potency cutoff before further profiling the compound. This decision locked

Count

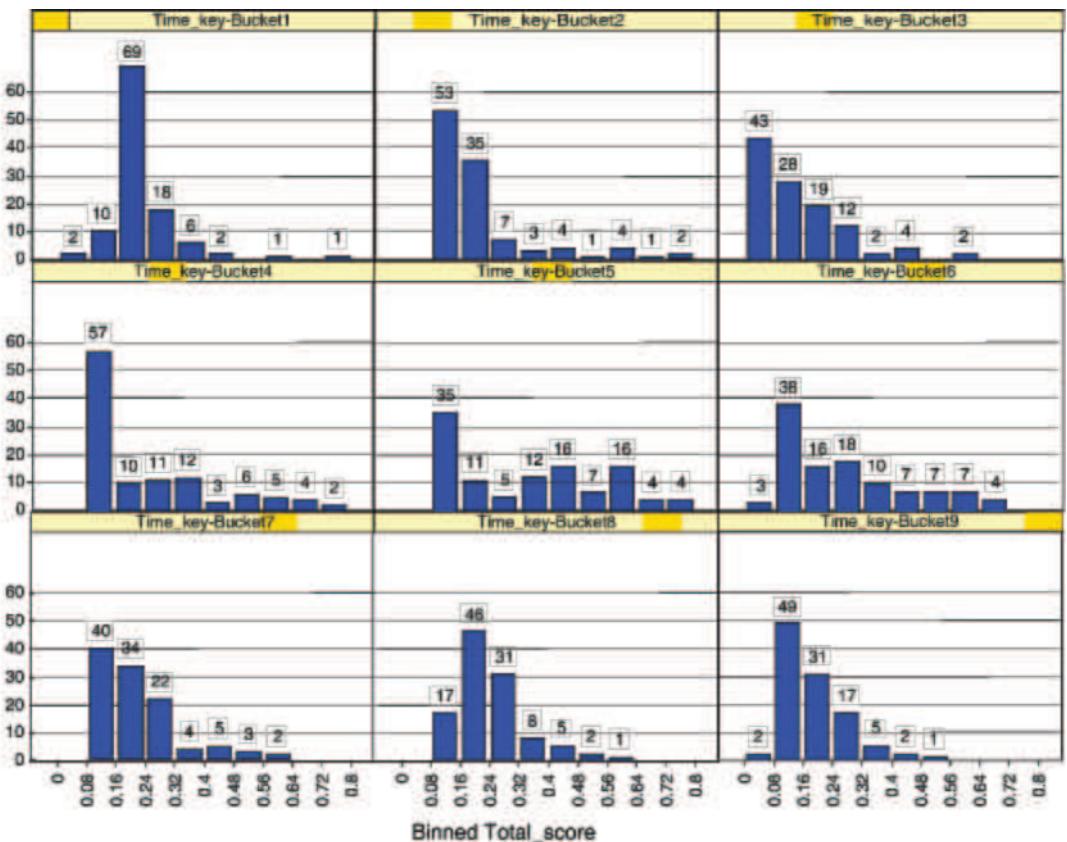


FIGURE 8.15 An example of the change in Pareto front during a lead optimization project. At each time step, the nondominated solutions are in Bin 0. It can be seen that after Time Bucket3 the project ceased to produce new molecules which were true improvements over those previously discovered. In this case, the injudicious application of a potency threshold caused the optimization to stall. For color details, please see color plate section.

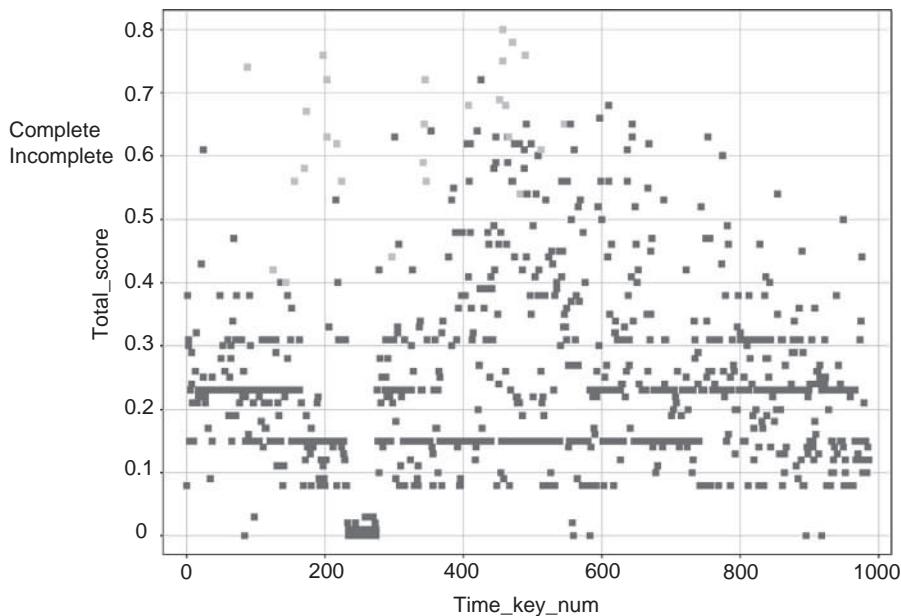


FIGURE 8.16 This figure illustrates the effect of setting too high a potency threshold for progression to other assays (e.g., ADMET evaluation). About halfway through the project, no further molecules generated a full molecular profile as they did not meet a high potency threshold. These molecules could have been superior in every other criteria but potency, and the project would never discover this key information. For color details, please see color plate section.

the series into a local minimum from which it could not escape as every change that would have improved the other properties of the molecule would have reduced the potency below the cutoff. In order to optimize in a multiobjective manner, a temporary loss of potency would need to be accepted.

In some cases, it was deemed impossible to optimize the molecule to reach all desired properties. In particular, the concept of competing objectives is well established in multicriteria optimization case studies. Competing objectives are those where an improvement in one objective results in the deterioration in the other. Parallel coordinate plots, a simple tool that can be used to identify competing objectives, were found to be a quick and useful aid. Figure 8.17 illustrates the case of two competing objectives, solubility and *in vitro* enzyme potency: the crossed lines clearly identify the objectives are in competition.

Unfortunately, research in the area of optimization strategies is hampered by a lack of datasets on which to test new methods. Ideally the data would comprise not only what was made in the project but what could have been made—a “full rank” dataset of every R-group at each position with every R-group at the other positions. Recently, one of us published a study on such a dataset (with the full rank data available as supplementary material) [40]. A series of MMP-12 inhibitors, found by a high-throughput screening, was elaborated at two positions by 50 substituents each (Figure 8.18).

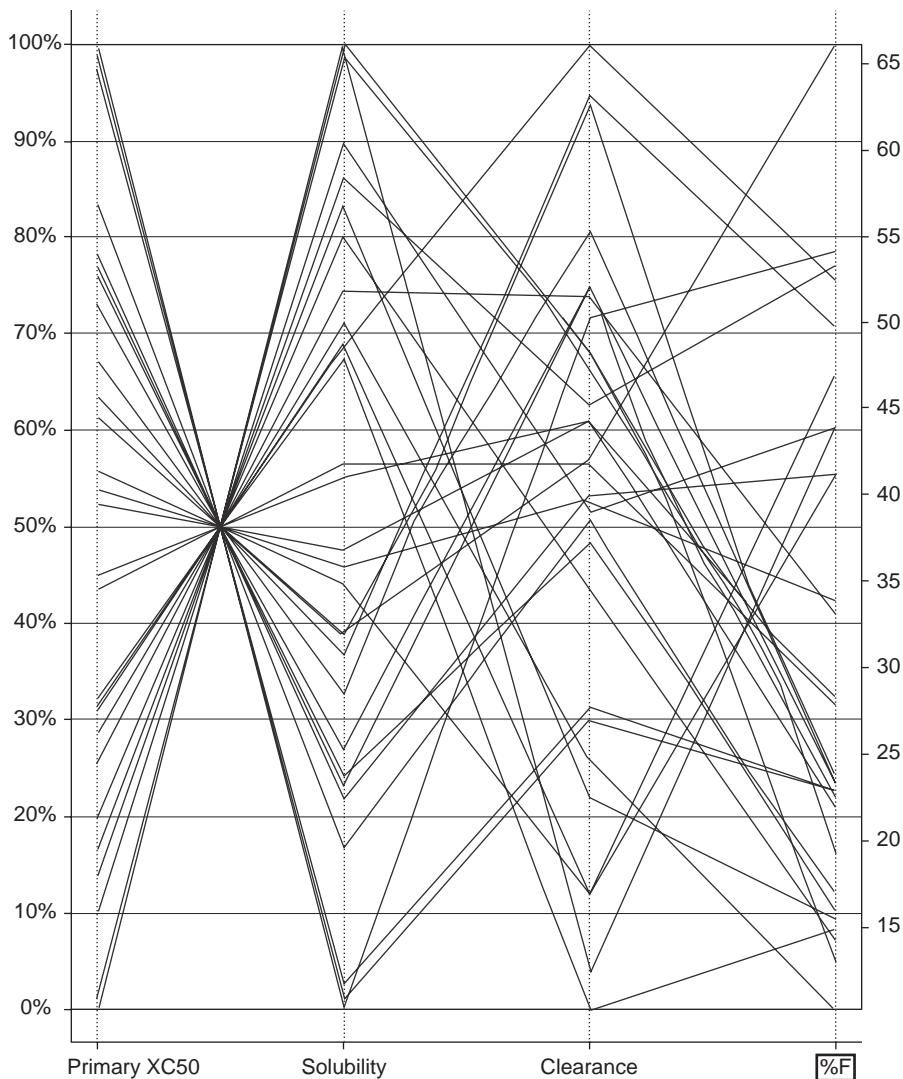


FIGURE 8.17 Parallel coordinate plots allow detection of competing objectives in optimization processes. This is an extreme example, where solubility and *in vitro* potency (“Primary XC50”) are competing. When solubility increases, potency decreases and vice versa. This results in the characteristic crossed-line plot shown here.

The team attempted to synthesize the full 2500 set of compounds—620 could not be made with a reasonable time investment, a sobering statistic. All compounds were tested against MMP-12 to derive the full dataset. Independently, a series of iterative experiments were made, guided by a purely computational optimization approach, in this case a genetic algorithm (GA). At each iteration, 14 molecules were made and tested. The results were added to the known data, and the GA was

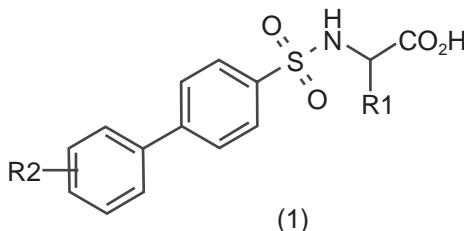


FIGURE 8.18 The series of MMP-12 inhibitors used by Pickett et al. [40] to demonstrate automated, iterative, lead optimization techniques. Reprinted with permission from Pickett et al. [40], © 2011 American Chemical Society.

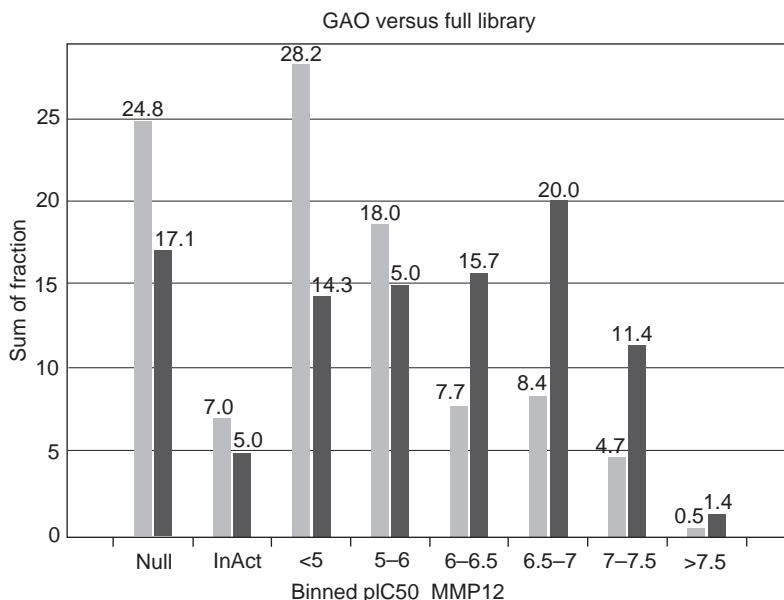


FIGURE 8.19 Proportion of compounds sampled by the GAO algorithm (blue) compared to the full dataset (red), binned by the primary assay. It can be seen that the algorithm samples compounds from all parts of the activity spectrum, but is very efficient at sampling the most potent compounds. Reprinted with permission from Pickett et al. [40], © 2011 American Chemical Society. For color details, please see color plate section.

used to select new molecules using only the existing results, and the simplest of all descriptions of the molecules—which R groups had been used (each R group had a unique number, not related to its chemical structure). Using 10 cycles of iterative design/synthesis (i.e., just 140 compounds), the algorithm was able to find the most active compounds in the set (Figure 8.19). Such an approach may be easily extended to multiple objectives [41] and some chemical intelligence introduced so that, for example, the algorithm understands which R groups are similar in structure or Hansch parameters.

8.4.2 MPO-Guided Hit to Candidate

Looking for a good balance of properties from the earliest stages of a project can help to ensure rapid progress through lead optimization, avoiding multiple, lengthy iterations, or the need to “hop” to new lead series to overcome critical liabilities. This example demonstrates how a project with a goal of an orally dosed compound for a cardiovascular target applied MPO approaches from the earliest stages to target chemistries with a good balance of properties, resulting in rapid progress through lead optimization.

A plot of the “chemical space” explored by the initial screening library of the project is shown in Figure 8.20a. This plot shows the structural diversity of the compound library of approximately 500 compounds screened for activity against the therapeutic target; the measured activity is indicated by the color. From this, it can be seen that good activity was identified for a wide diversity of chemistry.

These experimental data were combined with *in silico* predictions for a range of ADME properties, using the probabilistic scoring method described earlier, and assessed against the profile shown in Figure 8.21a. The resulting scores are plotted in the scoring plot and chemical space shown in Figure 8.20b, which shows that only a small number of compounds in restricted areas of the chemical space are likely to exhibit both good activity and appropriate ADME properties.

To confirm this finding, ~40 compounds were selected for primary *in vitro* ADME assays to measure their solubility and human liver microsomal stability. These are shown in the chemical space in Figure 8.20c and, while these focus on those compounds predicted to have the best balance of properties, a wider range of diversity was explored to confirm the predicted hypothesis. The compounds studied experimentally were, in turn, scored based only on the *in vitro* data using the profile shown in Figure 8.21b; the results are indicated by the colors of the points and the scoring plot shown Figure 8.20c. This *in vitro* analysis reinforced the high-quality chemistry identified by the *in silico* analysis.

Further chemistry expansion was undertaken in lead optimization, along with detailed *in vitro* and *in vivo* studies. However, in common with the project discussed in the previous case study, the eventual development candidate was one of the compounds identified early in the project, indicated by the blue point highlighted in Figure 8.20b and c.

8.5 CONCLUSION

In this chapter, we have discussed a variety of different chemoinformatics approaches to aid the efficient optimization of high-quality compounds in lead optimization and provided some examples of their application. It is clear that these approaches, appropriately applied, have the potential to significantly improve the lead optimization process. No algorithm can wholly replace the skills of an experienced scientist, but the objective rigor that computational approaches bring can augment this expertise to ensure that a wide variety of opportunities are explored and synthetic and

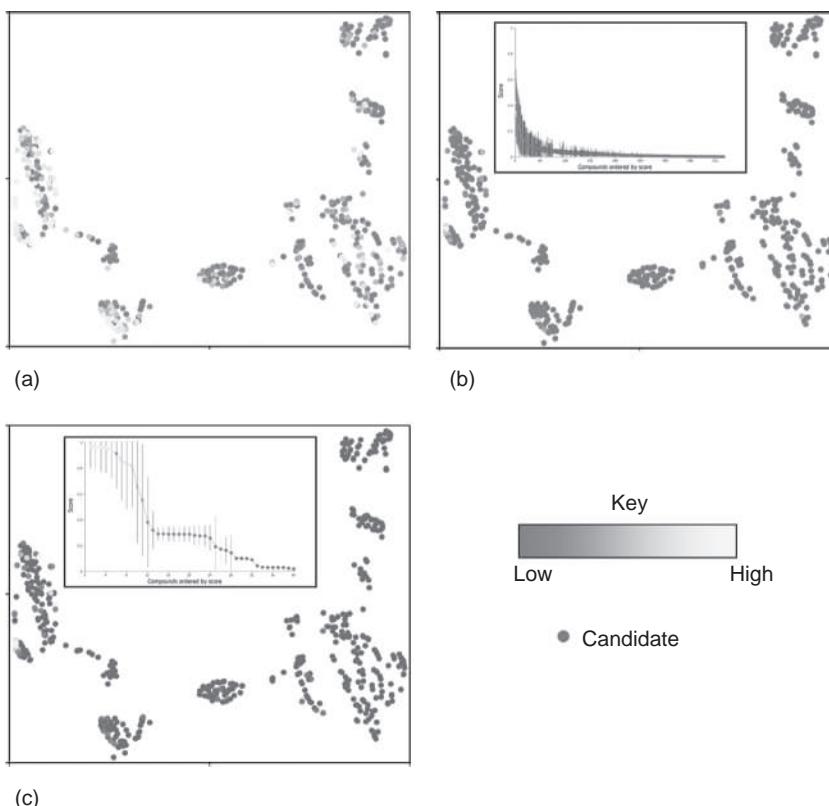


FIGURE 8.20 Chemical space plots illustrating the diversity of the library compounds screened for activity against the project's therapeutic target. Each point represents a compound and the distance between points represents the structural similarity (defined by Tanimoto similarity of 2D fingerprints). The points in (a) are colored by activity (% inhibition) of each compound against the target, showing that active compounds were identified for a wide diversity of chemistry. The colors in (b) show the score of each compound against the profile shown in Figure 8.21a chosen to identify compounds with a good balance of experimental potency and predicted ADME properties. A plot of the scores is shown inset, indicating that only a small number of compounds, representing a small number of similar chemistries, are likely to achieve this desired profile. Finally, (c) shows the compounds selected for initial *in vitro* ADME studies, focusing on the chemistries most likely to have a good balance of properties. The compound scores, based on the *in vitro* data, against the profile shown in Figure 8.21b, are indicated by the colors and plotted in the inset graph. The compound ultimately selected as the development candidate is highlighted in blue in plots (b) and (c). For color details, please see color plate section.

experimental efforts distributed appropriately to focus on the chemistries with the highest chance of success.

In the future, informatics methods can go beyond guiding the tactical decisions of which compounds to synthesize and test next. Decision analysis techniques can

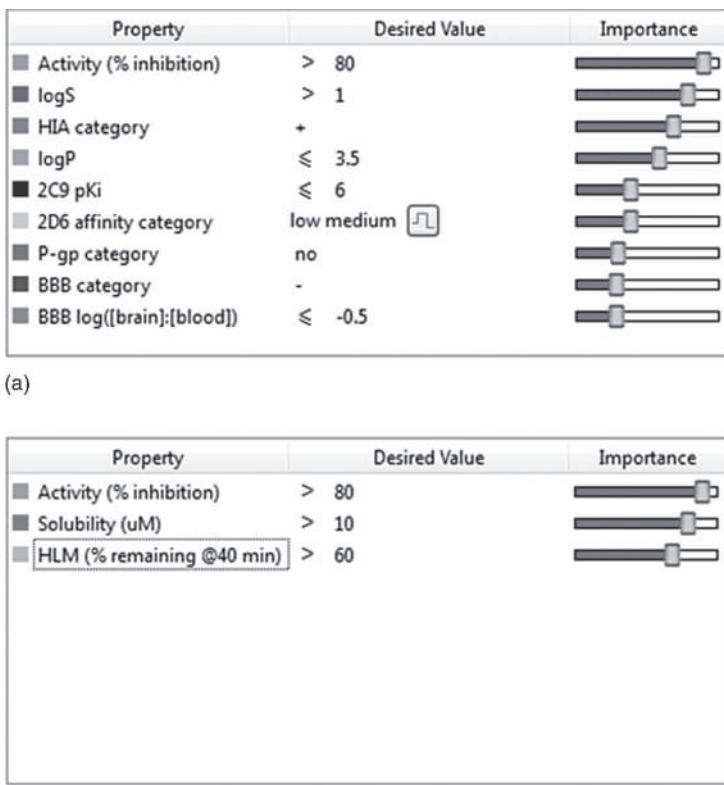


FIGURE 8.21 Scoring profiles used for prioritization of compounds intended for oral dosing against a cardiovascular target. The profile shown in (a) combines the experimentally measured target activity (as a percentage inhibition) with *in silico* predictions of solubility ($\log \mu\text{M}$), human intestinal absorption (HIA), $\log P$, inhibition of cytochrome P450s CYP2D6 and CYP2C9, active transport by P-gp, and blood–brain–barrier penetration (BBB). The profile shown in (b) combines the experimental activity with the primary *in vitro* ADME assay results for solubility in micrometer and human liver microsomal (HLM) stability measured as percentage remaining after a 40 min incubation. For color details, please see color plate section.

help to analyze the many strategic directions that could be taken to progress a project and identify the most important experiments to perform or compounds to synthesize. For example, in compound optimization, there remains a tension between synthesizing the compounds that are expected to be “best” and synthesizing those that will provide valuable information to understand and predict the SAR across multiple properties. Often, we believe too much focus is placed on the short-term goal of finding the best compounds for a given property, which can lead to inefficient search strategies and many more design-test iterations than necessary. We envisage a future in which tools will be available that will analyze potential future courses for

optimization and provide guidance on the appropriate balance of resources between performing experiments on existing compounds and synthesis of new compounds to both identify high-quality compounds and inform future optimization across multiple parameters.

The final key to the successful adoption of chemoinformatics methods is their accessibility to all of the decision-makers in a project team. The majority of these scientists are not computational experts, so the tools and their results must be presented in an intuitive and user-friendly way to encourage routine use. Many computational tools for drug discovery are designed by computational scientists, for computational scientists, presenting a barrier to use by the majority of chemists and biologists; this must change to achieve widespread adoption. Encouragingly, more attention is now being given to the effective presentation of data and analyses [42, 43].

REFERENCES

1. Proudfoot J, Nosjean O, Blachard J, et al. *Pure Appl Chem* 2011;83:1129–1158.
2. Lusher S, McGuire R, Azevedo R, et al. *Drug Discov Today* 2011;16:555–568.
3. Hansch C, Leo A. *Exploring QSAR*. Washington, DC: American Chemical Society; 1995. p 542.
4. Segall M. *Curr Pharm Des* 2012;18:1292–1310.
5. Ekins S, Honeycutt JD, Metz JT. *Drug Discov Today* 2010;15:451–460.
6. Hansch C, Fujita T. *J Am Chem Soc* 1964;86:1616–1626.
7. Hansch C, Unger SH, Forsythe AB. *J Med Chem* 1973;16:1217–1222.
8. Martin YC. *Quantitative Drug Design: A Critical Introduction*. New York: Marcel Dekker; 1978.
9. Craig PN. *J Med Chem* 1971;14:680–684.
10. Topliss JG. *J Med Chem* 1972;15:1006–1011.
11. Martin YC, Panas HN. *J Med Chem* 1979;22:784–791.
12. Austell V. *Quant Struct Act Relation* 1983;2:59–65.
13. Box GEP, Wilson KB. *J Royal Stat Soc* 1951;13:1–45.
14. Darvas F. *J Med Chem* 1974;17:799–804.
15. Martin YC. *J Med Chem* 1981;24:229–237.
16. Maliski EG, Latour K, Bradshaw J. *Drug Des Discov* 1992;9:1–9.
17. van de Waterbeemd H, editor. *Advanced-Computer Assisted Techniques in Drug Discovery*. New York: VCH Publishers; 1994.
18. Maliski EG, Bradshaw J. QSAR and the role of computers in drug design. In: *Medicinal Chemistry: The Role of Organic Chemistry in Drug Research*. London: Academic Press; 1993. p 94–102.
19. Livingstone D. Characterising chemical structures using physicochemical descriptors. In: *Drug Design Strategies: Quantitative Approaches*. Cambridge: Royal Society of Chemistry; 2012. p 220–241.
20. Stumpfe D, Bajorath J. *J Med Chem* 2012;55:2932–2942.

21. Patel Y, Gillet VJ, Howe T, et al. J Med Chem 2008;51:7552–7562.
22. Peltason L, Bajorath J. J Med Chem 2007;50:5571–5578.
23. Sisay MT, Peltason L, Bajorath J. J Chem Inf Model 2009;49:2179–2189.
24. Steur RE. *Multiple Criteria Optimization: Theory, Computation and Application*. New York: John Wiley & Sons; 1986.
25. Lipinski CA, Lombardo F, Dominy BW, et al. Adv Drug Deliv Rev 1997;23:3–25.
26. Veber D, Johnson S, Cheng H, et al. J Med Chem 2002;45:2615–2623.
27. Lovering F, Bikker J, Humbert C. J Med Chem 2009;52:6752–6756.
28. Hughes J, Blagg J, Price JD, et al. Bioorg Med Chem Lett 2008;18:4872–4875.
29. Yusof, I, Segall M. Drug Discov. Today 2013;18(13/14):659–666.
30. Harrington E. Ind. Qual. Control 1965;21:494–498.
31. Bickerton G, Paolini G, Besnard J, et al. Nat Chem 2012;4:90–98.
32. Gillet VJ, Willett P, Bradshaw J, et al. J Chem Inf Comput Sci 1998;38:165–179.
33. Segall M, Champness E, Obrezanova O, et al. Chem Biodivers 2009;6:2144–2151.
34. Jaffe W. J Econ Lit 1972;10:1190–1201.
35. Friedman J, Fisher N. Stat Comput 1999;9:123–143.
36. Hashimoto T, Segall M. Finding drug discovery “rules of thumb” with bump hunting. American Chemical Society Fall National Meeting; 2010; Boston.
37. Petrillo EW. Drug Discov World 2007;8:9–16.
38. Loo J. Tracing small molecules in lead optimisation [M.Sc. dissertation]. Sheffield: Sheffield University; 2006.
39. MacDonald SJF, Smith PW. Drug Discov Today 2001;6:947–953.
40. Pickett SD, Green DVS, Hunt DL, et al. ACS Med Chem Lett 2011;2:28–33.
41. Gillet VJ, Khatib W, Willett P, et al. J Chem Inf Comput Sci 2002;42:375–385.
42. Ritchie T, Erlt P, Lewis R. Drug Discov Today 2011;14:65–72.
43. Champness E. Innov Pharm Technol 2010;34:32–38.

CHAPTER 9

USING CHEMOINFORMATICS TOOLS TO ANALYZE CHEMICAL ARRAYS IN LEAD OPTIMIZATION

GEORGE PAPADATOS, VALERIE J. GILLET,
CHRISTOPHER N. LUSCOMBE, IAIN M. McLAY,
STEPHEN D. PICKETT, and PETER WILLETT

9.1 INTRODUCTION

Array chemistry has become a routine way of exploring the chemical space around lead compounds during the lead optimization stage of drug discovery. A lead compound is one that has shown activity against the target of interest, for example, in high-throughput and secondary activity screening, and which has sufficient potential for further development in terms of chemical modification possibilities, novelty and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties [1, 2]. The aim of the lead optimization process is to explore the chemistry space around the lead in order to identify compounds with improved properties and potential for further development. Traditionally, lead optimization followed a sequential path in which potency and selectivity were optimized first followed by pharmacokinetic and toxicity issues [3]. However, this sequential approach suffers from serious limitations since there is no guarantee that levels of potency and selectivity will be maintained as ADMET properties are optimized. Indeed, factors that improve potency often have a detrimental effect on ADMET properties, for example, increased lipophilicity often leads to stronger binding but is detrimental to solubility [4]. Therefore, a multiobjective optimization approach is now favored over the sequential paradigm [5–7] with the expectation that this will lead to compounds with a better balance of properties and a reduction in attrition during clinical trials.

Array chemistry is the process whereby a series of related compounds is synthesized simultaneously by varying structural components around a scaffold using a

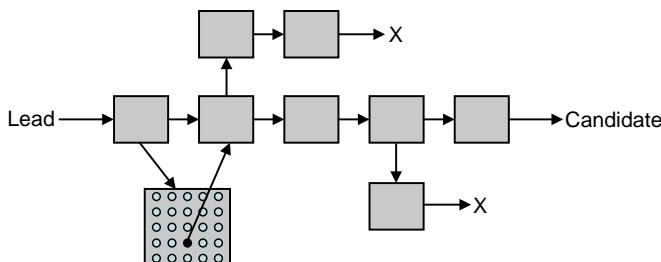


FIGURE 9.1 A schematic of the organization of arrays during the lifetime of a lead optimization project. The relationships between different arrays (gray boxes) and project milestones, such as the lead, seed (black circle), and drug candidate structures, as well as dead ends (marked by X's), are shown.

common reaction scheme and a variety of different reagents. Array synthesis is usually applied in an iterative manner: The first array is designed around the lead compound, the compounds are then synthesized, purified, and tested; the most promising compound from among the array members is then selected to be used as the starting point for the design of the next array. This iterative procedure is repeated, ideally until a clinical drug candidate is discovered, or until the lead optimization project is terminated. Typically, large and structurally diverse arrays are synthesized at the start of the process, when medicinal chemists are exploring large areas of the chemical space. Conversely, toward the end of a project the structure–activity relationships (SARs) are more evident, enabling the design of smaller, more focused arrays, which aim to fine-tune the property profiles of the compounds [8].

A schematic of the use of arrays in the lead optimization process is shown in Figure 9.1. The starting point is a lead compound from which a seed compound would be defined as the basis for the construction of an array (the lead compound itself may not be suitable for the design). The compounds in the array will then be tested and one or more new compounds selected as seeds for further modification. Some of the paths explored are likely to lead to dead ends (marked by the X's) where it is not possible to improve upon the compound on which the array was designed.

Array synthesis has been a key technology in lead optimization for several years and pharmaceutical companies have built large archives of compounds that were synthesized using arrays, together with their biological screening data. These archives represent a significant investment in corporate resources and are potentially valuable for the extraction of knowledge that could be used to assist prospective array design. Mining historical data to derive rules for prospective design has become an increasingly popular activity made possible by the growth in size of databases of compounds and their properties. For example, matched molecular pairs (MMPs) analysis, which was first suggested by Leach and colleagues [9], has been used to predict the effect of changing one substituent for another on many different properties of compounds [10] (our own contribution to this area is to consider the effect of context on an MMP [11]). These analyses are based on whole collections of compounds regardless of the purpose for which they were synthesized and tested.

In this chapter, we describe efforts to analyze the performance of arrays in retrospective lead optimization projects within GlaxoSmithKline (GSK) with the aim of deriving rules to guide prospective array design. The process of lead optimization can be described as identifying the change in structure, ΔS , required to bring about some desirable change, ΔP , in the property (or properties) of interest. Framing the problem in this way illustrates the difficulty of the task facing the medicinal chemist carrying out a lead-explosion study, since quantitative structure–activity relationship (QSAR) studies typically make predictions about a change in property given some specific change in structure, that is, the aim is to predict ΔP given ΔS rather than the converse “inverse QSAR” [12]. However, the existence of an archive of previous array experiments could provide a basis for tackling the “inverse array” problem by deriving predictive rules to direct future lead-explosion projects. Specifically, a lead-explosion study is characterized by the lead molecule, the intermediate molecules created by a sequence of one or more array syntheses, and the optimized candidate, that is, the final output from the study. An analysis of previously synthesized arrays could provide information on issues such as: the structural changes required to produce a desired change in physicochemical properties/potency; the nature of the relationship between the size and configuration of an array and corresponding changes in physicochemical properties/potency (e.g., what size of library is required to increase potency by $x \log$ units); and the extent to which generally applicable, rather than project-specific, guidelines can be derived (e.g., respiratory-diseases drugs can be administered orally or inhaled, these modes requiring the optimization of different physicochemical properties).

We begin by using chemoinformatics tools to analyze the performance of arrays in two historical lead optimization projects within GSK. The analysis focuses on the chemical and property space covered by arrays, the progress of arrays relative to their seed compounds and to the eventual candidate compound, and the variation in property profiles as a project proceeds over time. While these examples serve to illustrate that valuable insights can be gained from such an analysis, the derivation of generally applicable rules would require the analysis of a much greater number of lead optimization projects than these two. However, inspection of the GSK archive revealed that although the individual compounds are stored along with project details, information relating to the array in which they were synthesized is typically not stored, and, furthermore, the decisions taken during the lead selection and optimization process are generally not well documented (this is not unique to GSK and has been observed previously by others [13]). We then describe efforts to automatically mine array-centric data from the GSK archive. We conclude that the data currently captured during lead optimization is insufficient to enable such a larger scale analysis to be carried out and that this presents a significant barrier to the extent to which the wealth of data can be exploited.

9.2 LEAD OPTIMIZATION PROJECTS

The analysis was based on two lead optimization projects. Project A represents a lead optimization campaign against an enzyme target that spanned approximately 4 years, between 2004 and 2007, and involved the synthesis and testing of 2154 unique

TABLE 9.1 Characteristics of Projects A and B

Project A	Project B (Subset)
2154 compounds	387 compounds
109 chemical arrays	21 chemical arrays
1733 compounds synthesized in an array	258 compounds synthesized in an array
1–45 compounds per array	1–60 compounds per array
23 compounds on average in each array	20 compounds on average in each array
94 arrays containing more than 3 compounds	19 arrays containing more than 3 compounds
6 distinct chemical classes	1 chemical class

compounds. Of these, 1733 belonged to one of the 109 chemical arrays. The project was initiated in 2004 on two lead structures, and ultimately the proposed drug candidate compound was discovered in September 2006, and belonged to array 78. Project B was a lead optimization campaign against a protein target and was live at the time of the study, having spanned approximately 10 months, between 2008 and 2009. A subset of 387 compounds was selected, all belonging to the same chemical class. Of these compounds, 285 were synthesized in an array. The proposed candidate compound for this chemotype (pre-candidate) was discovered in November 2008 and did not belong to an array. Table 9.1 summarizes the characteristics of Projects A and B.

9.3 COVERAGE OF CHEMISTRY AND PROPERTY SPACE (ΔS – ΔA PLOTS)

Figure 9.2 shows plots of the structural dissimilarities, ΔS , and biological activity differences, ΔA , between the seeds and the array members for three arrays in Project A. The shape of an array plot depends on (1) the choice of the descriptor/fingerprint and (2) the choice of the seed compound used for the dissimilarity and activity difference comparisons. In this case, pairwise structural dissimilarities were calculated using Tanimoto dissimilarity and MolPrint 2D fingerprints [14] using Pipeline Pilot [15] and the corresponding activity differences are in the primary assay pIC_{50} values.

Plots such as this can be very informative. For example:

- By focusing on the activity difference ΔA , an array can be assessed on the improvement in activity achieved compared to the seed: in arrays 83 and 68 almost all the array members were worse in activity than their respective seed.
- By focusing on ΔS , the extent of exploration of chemical space around the seed can be evaluated: array 83 exhibits a more thorough exploration of chemical space, with smoothly distributed dissimilarity values; whereas in array 68, all the compounds are equidistant from the seed in terms of structural dissimilarity.

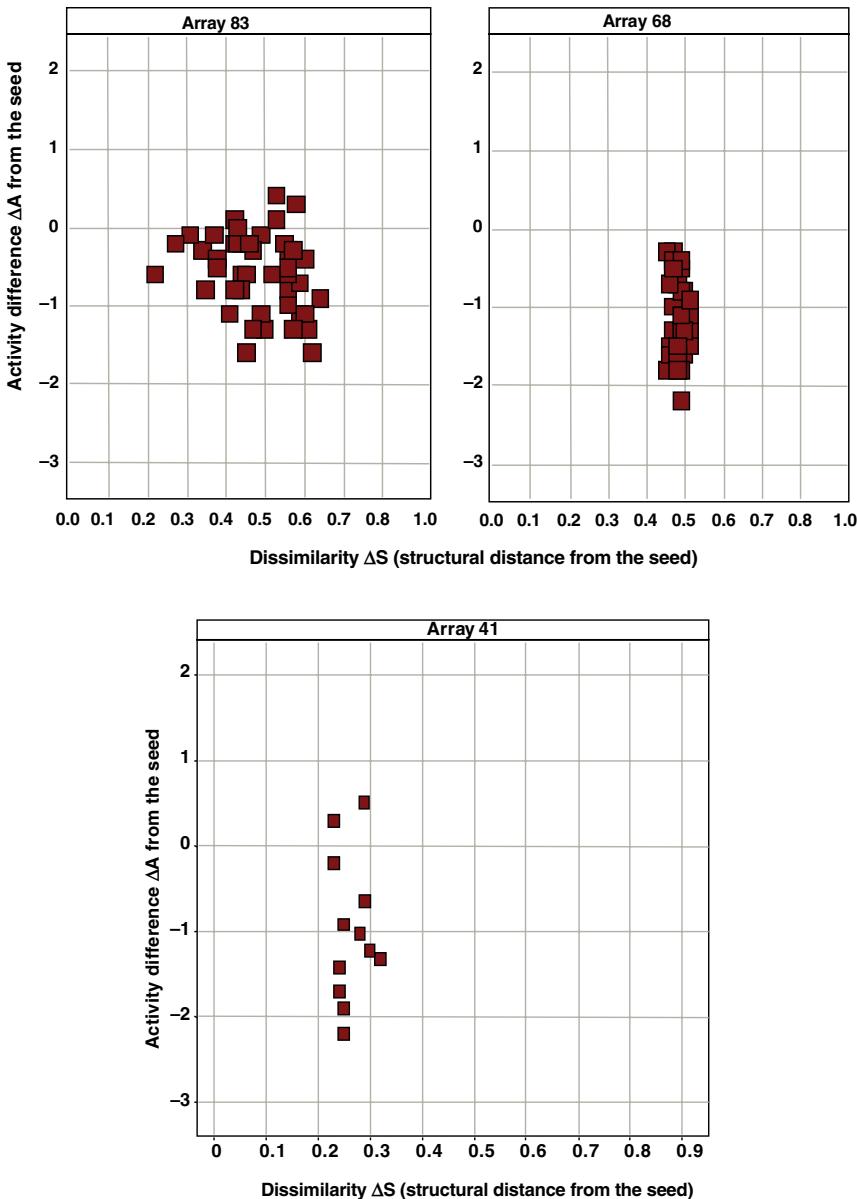


FIGURE 9.2 Seed – array (ΔS – ΔA) scatter plots for three arrays in Project A. The x -axis is the dissimilarity ΔS between each member of the array and the array’s seed. The y -axis is the corresponding activity difference ΔA between each member of the array and the array’s seed.

- The existence of data points toward the top or bottom left of a plot indicates compounds that are structurally similar to the array’s seed and have a large activity difference from it, that is, they represent “activity cliffs” in the local

SAR landscape [16, 17]. For example, in array 41, the 12 array members are structurally similar (>0.65 Tanimoto similarity) to their seed. However, the majority of them have significantly lower activity values, being 10 to 100 times less potent than the seed of the array.

Figure 9.3 presents a summary of the average ΔS - ΔA relationships across multiple arrays and enables trends between attributes such as array size and array performance

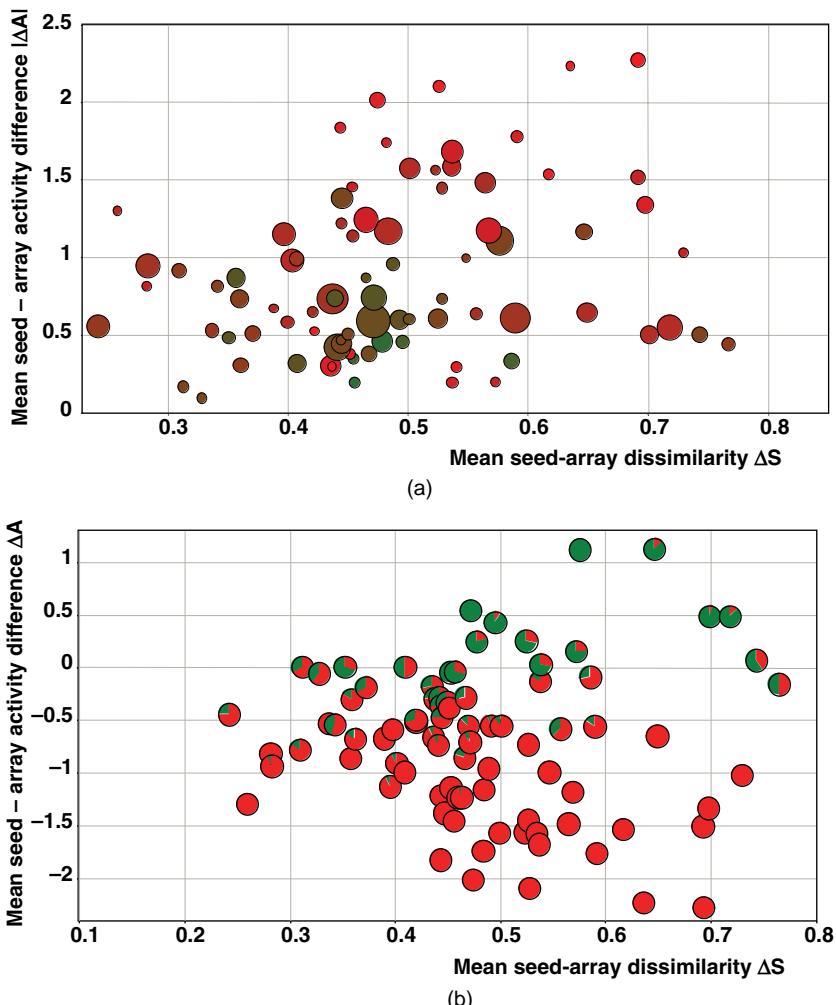


FIGURE 9.3 Each data point represents an array. In (a) and (c), the circles are color-coded by potency pIC_{50} value, ranging from red ($pIC_{50}=5$) to green ($pIC_{50}=8.6$) and the size of the circles corresponds to the size of the array (3–45 members). In (a) and (b), the x - and y -axes denote the average dissimilarity ΔS and absolute average property distance ΔA between the seed and each array member, respectively. In (b), the pies illustrate in green the proportion of the array molecules that have better potency than the seed.

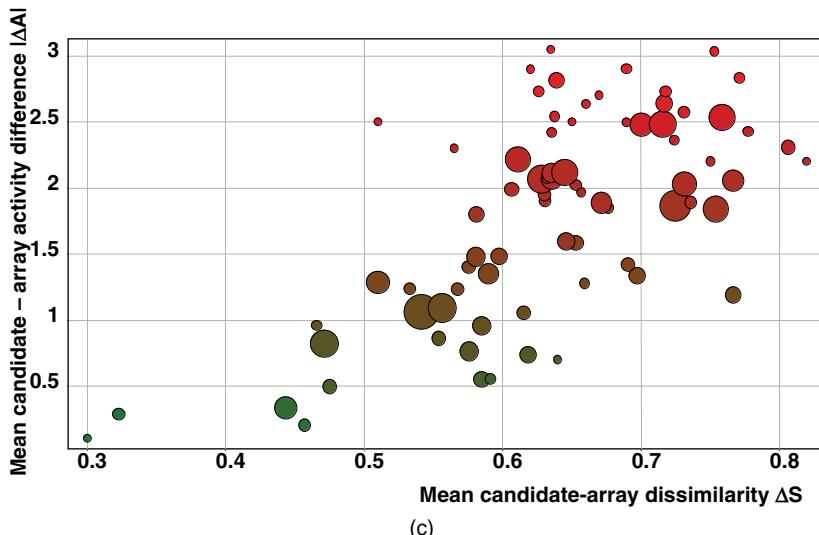


FIGURE 9.3 (Continued) In (c), the x - and y -axes denote the average dissimilarity ΔS and absolute average property distance ΔA between Project A's proposed drug candidate ($pIC_{50} = 8.2$) and each array member, respectively. For color details, please see color plate section.

to be identified for a whole project. For example, in Figure 9.3a and b, each data point represents an array plotted as average ΔS and average ΔA defined as the average structural dissimilarity and activity difference between the seed of the array and each of the array members, respectively. The size of each point corresponds to the size of the array. In Figure 9.3a, the circles are color coded accordingly to the average potency values and in Figure 9.3b they are colored according to the percentage of molecules in an array that are more potent than the seed. The roughly triangular shape of both of these plots confirms the similarity principle at the array level. In other words, the more structurally similar the array members are to their seed, the closer they are to the seed in activity values. There are four or five noteworthy exceptions here (the pies/arrays on the bottom left corner of Figure 9.3a) where although the array members are very similar on average to their seed, the corresponding activity differences are above 0.5 log units. This is an indication that the particular seeds and arrays form “activity cliffs” where the similarity principle is no longer applicable, and thus not useful to array design. Figure 9.3a also shows that there is no evidence to suggest that larger arrays lead to more potent compounds.

Figure 9.3b shows that even though the similarity principle holds, the more successful arrays (i.e., pies with a larger green part) tend to appear when the similarity “jump” from the seed is relatively large. This suggests that chemists might need to take greater risks during array design by synthesizing compounds dissimilar to their seed. Furthermore, there are notably more red colored pies than green ones. This reflects the sad but true fact that in real-life lead optimization most of the arrays fail to produce compounds with increased potency compared to their seed. It should be noted here that increased potency against the primary biological target is not

always the main aim of array design, as there are other objectives considered in lead optimization such as selectivity, or ADMET properties.

Finally, Figure 9.3c shows the average structural difference between the compounds in an array and the actual drug candidate (with $\text{pIC}_{50}=8.2$) for the project. This plot leads to two striking observations. First, there is a linear trend across the average candidate-array ΔS and ΔA values: the closer an array is structurally to the candidate, the more similar its average activity value is to the candidate and therefore the more potent (colored in green) are the compounds within it. Furthermore, there are no arrays in the bottom right corner of the plot, which indicates that all the arrays consisting of compounds that are structurally dissimilar to the candidate failed in terms of potency. The second striking observation is that there are only two small arrays in the structural neighborhood of the candidate, depicted in the bottom left corner with average dissimilarity value lower than 0.4. This suggests that chemical space around the candidate had not been thoroughly explored in the arrays that were analyzed. (In subsequent discussions with the medicinal chemists, it was revealed that the exploration around the candidate was still ongoing.)

9.4 TEMPORAL ANALYSIS OF LEAD OPTIMIZATION

Compounds synthesized during lead optimization can be considered as time-ordered with properties that evolve over the course of the project. Thus it is possible to conceive of a project trajectory as a path through descriptor space that varies with time. The idea that compound properties change as a project advances has been used to develop the concept of pharmaceutical lead-likeness where it has been shown that chemical structures become, on average, larger (increase in molecular weight) and greasier (increase in $\log P$) as lead optimization progresses from lead to drug candidate [1]. Here we examine how multiple properties varied over time using property profiles. As discussed earlier, a multiobjective approach to lead optimization is now considered preferable to the traditional sequential approach with a balance in properties being sought (assuming the likely conflicting nature of the objectives) [8, 18].

We have used two different multiobjective methods to score the compounds produced during the course of lead optimization, namely, weighted desirability scoring and Pareto ranking. The main questions that were explored are (a) did the use of a linear cascade lead to compounds being overlooked by the medicinal chemists even though these compounds had a better multiobjective score? and, more generally, (b) does lead optimization gradually lead to better overall compounds as the project progresses in time or is there no such trend? The second question was examined at both the array and chemotype levels.

In the weighted desirability approach [19], each objective is first transformed into a dimensionless desirability value, d , in the range 0–1, where a value of 0 indicates the objective value is unacceptable and a value of 1 indicates that the objective has exactly the target value. The value of d increases monotonically with the desirability of the corresponding objective value [20]. The individual desirability functions were

TABLE 9.2 Acceptance Criteria According to the Cascade of Project A

Criterion	Desired Values	Weight
Enzyme assay A	$\text{pIC}_{50} > 8.0 \pm 0.5$	0.75
Enzyme assay B	$\text{pIC}_{50} < 6.0 \pm 0.5$	0.70
Cell assay A	$\text{pIC}_{50} > 7.0$	0.65
Cell assay B	$\text{pIC}_{50} > 6.8$	0.60
Clearance Cl	$<40 \pm 10 \text{ ml}/(\text{min}\cdot\text{kg})$	0.20
Oral bioavailability F	$>30\% \pm 10\%$	0.50

The relative weights used for the weighted desirability method are also listed.

then combined into a single score by using a weighted-sum approach where the importance of the individual objectives can be varied.

In Pareto ranking, no prioritization or weighting of the individual objectives is required. Each objective is considered separately and ranking is based on the concept of dominance. A nondominated individual (compound) is one for which there is no better compound when all of the objectives are taken into account. In the approach used here, each compound is assigned a rank according to the number of compounds by which it is dominated. Pareto ranking has been used extensively in the field of chemoinformatics; examples include library design according to multiobjective criteria using a genetic algorithm [21], docking-based virtual screening [22], fragment-based *de novo* ligand design [23], as well as optimizing (Q)SAR models [24, 25].

The compounds in Project A were scored according to six acceptance criteria that were defined at the onset of the project by the medicinal chemists involved and which are shown in Table 9.2. The weighted desirability scores were calculated using an in-house tool called Adamantis Pro. A linear desirability function was defined for each criterion according to the desired ranges listed in Table 9.2 and the individual scores were combined using the weights shown in the third column that were also defined by the medicinal chemists. Missing values were handled by using a value that is lower than the worst value seen over the whole dataset.

The property profiles of the compounds synthesized during Project A are illustrated in Figure 9.4 using parallel coordinates plots, where each column on the *x*-axis represents an optimization objective and the *y*-axis shows the desirability functions, with the combined function in the right most column. Each line in the plot presents a compound and the line is colored according to the combined final score (green for high values; red for low values). The plot demonstrates that most of the compounds were tested in the first two enzyme assays; however, most were not tested against the remaining four criteria, such as the cell assays and the two ADMET properties. This is typical in a linear cascade approach since compounds which perform poorly in one assay will not proceed to the next. Moreover, the difficulties associated with optimizing the different properties are readily apparent as only 51 of the 2154 compounds synthesized during Project A had a weighted score above 0.6. Finally, a trade-off between the two leftmost properties (enzyme assays A and B) can be observed (the crossed lines show that good performance in one property is coincident

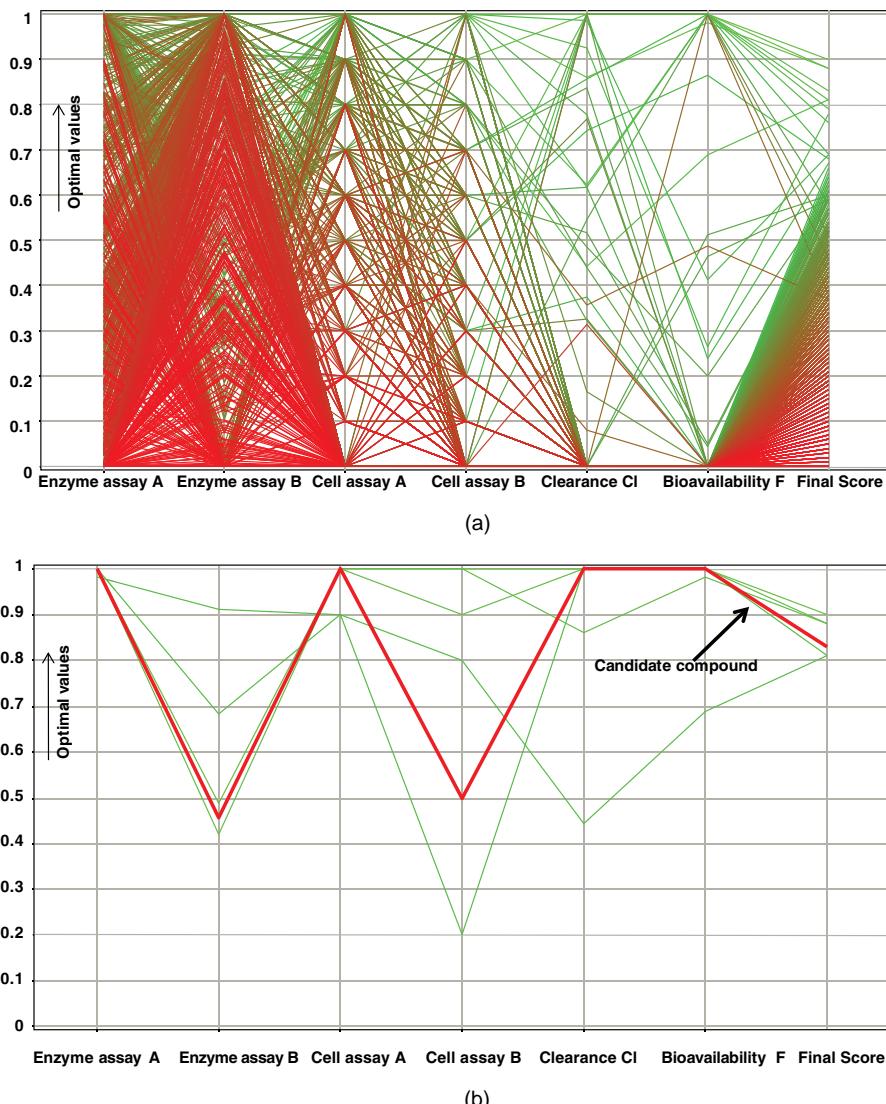


FIGURE 9.4 (a) Parallel coordinates plot of Project A compounds' six most important experimental properties (normalized from 0 to 1) and weighted desirability score. Higher scores are depicted in green. (b) Parallel coordinates plot of the six compounds having a weighted desirability score above 0.8. The actual candidate of Project A is highlighted in red and has a score of 0.83. For color details, please see color plate section.

with poor performance in the other), indicating that the medicinal chemists were faced with selectivity issues.

Figure 9.4b shows the six molecules with a weighted score above 0.8. The candidate compound proposed by the medicinal chemists at the end of the project came

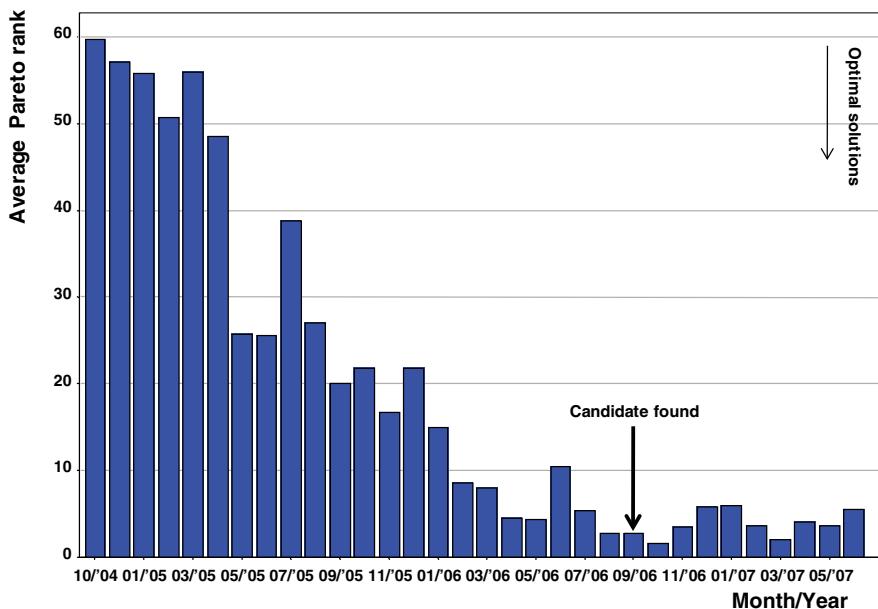


FIGURE 9.5 Histogram of the average Pareto rank binned by the chronologically ordered date of experiment. The Pareto rank reaches lower (=better) values as the project proceeded in time (2004–2007), demonstrating that optimization was effective overall; however, the candidate compound was a Pareto suboptimal solution.

fourth with a score of 0.83 (profile highlighted in red line), while three other molecules had a better overall score. Notably, these three molecules were synthesized and tested six months after the discovery of the candidate. Based on this chart, it is clear that although the candidate has optimal values in enzyme and cell assay A, clearance and bioavailability, it has comparatively lower values in properties such as enzyme and cell assay B.

The Pareto ranks were also calculated for all compounds synthesized in the project using Adamantis Pro. Compounds with missing property values were treated in the same way as for the desirability scoring approach. In order to examine the effectiveness of the optimization during Project A, the Pareto ranks were binned by the month and the year that the compounds were synthesized, and an average Pareto rank calculated for each bin. Figure 9.5 plots the average Pareto rank against time and shows that the profile of the compounds improves as the project progresses. However, in both cases (desirability function and Pareto ranks), the way that missing data are handled gives an inherent bias to compounds synthesized later in the project since they are more likely to have complete property profiles. The candidate compound was a Pareto suboptimal solution, having a Pareto rank of 1. Remarkably, there were 32 compounds on the Pareto front, having a rank of 0. This highlights the limitations of the linear cascade approach, which failed to identify the compounds with the optimal property trade-off.

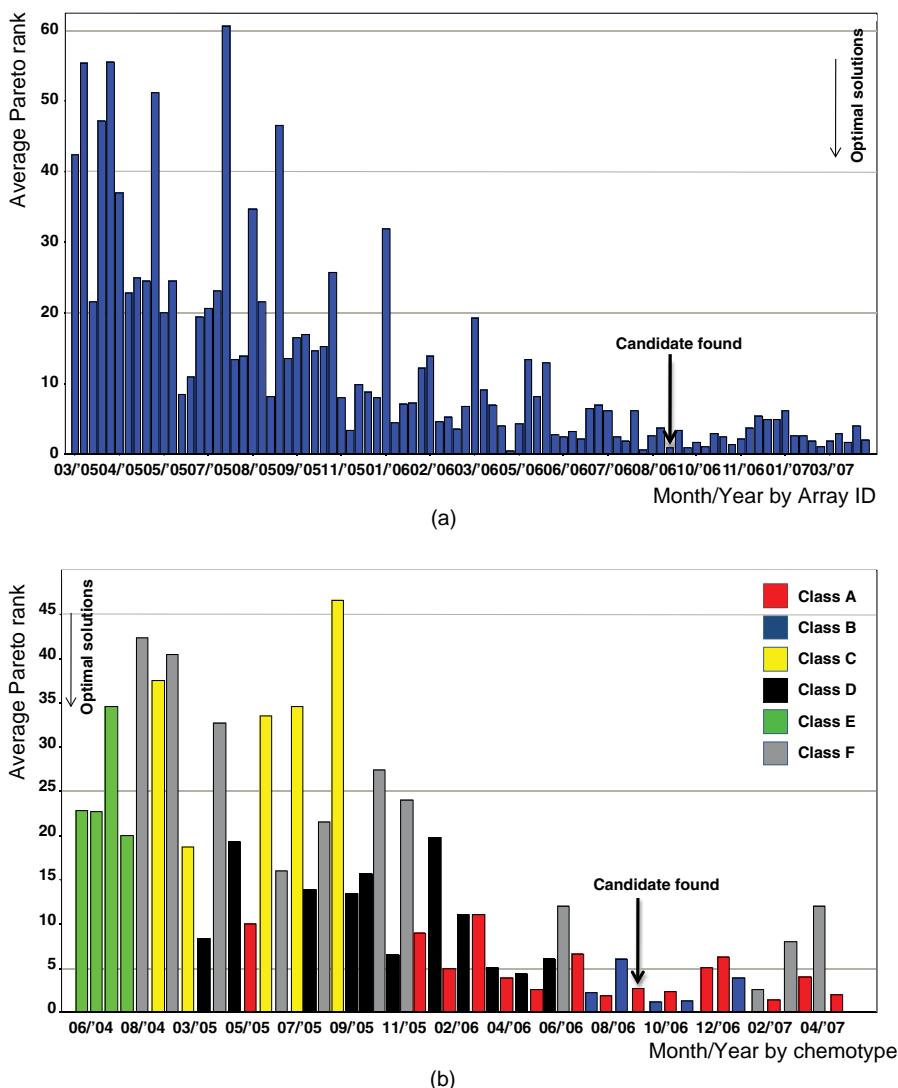


FIGURE 9.6 (a) Average Pareto rank binned by chronologically ordered chemical arrays. (b) Average Pareto rank binned by date and color-coded by chemotype. For color details, please see color plate section.

Two additional plots were generated to examine the Pareto rank performance of individual chemical arrays and chemotypes over time. Figure 9.6a illustrates the average Pareto rank for the 94 arrays of Project A that contain more than three compounds using their chronological ordering. Each array was assigned the date that the chronologically penultimate compound in the array was first tested against the project's primary assay. The plot largely follows the trend identified in Figure 9.5 and shows an improvement in

the balance of the six acceptance criteria over the course of the optimization. The drug candidate was synthesized in an array of eight compounds, exhibiting a very low average Pareto rank of 0.875. Aforementioned, the candidate molecule itself was a sub-optimal solution having a Pareto rank 1 rather than the optimal value of 0. Further visual inspection revealed that during the course of the project there were arrays which had an equivalent or even better average Pareto rank, both before and after the discovery of the drug candidate. Most likely, the medicinal chemists applied further acceptance criteria to justify their candidate choice, which were not considered during this analysis (e.g., $c \log P$, solubility, hERG, synthetic tractability, etc.). Nevertheless, the application of a linear cascade is once again demonstrated to be incapable of prospectively identifying compounds with a well-balanced property profile.

Shifting the focus to chemotypes rather than arrays, Figure 9.6b summarizes the overall effectiveness of the six different chemotypes explored over the timeline of Project A. This plot illustrates the chemotypes that gave consistently better Pareto ranks. For example, the exploration around the chemical classes C and E (colored in yellow and green, respectively) did not deliver and were therefore abandoned early in the project. Chemotype D gave comparatively better compounds but it was also abandoned after the discovery of chemotype A. Conversely, chemotypes A and B (red and blue) were consistently successful in terms of average Pareto ranks; it is no surprise that the drug candidate came from chemotype A.

9.5 MODELING LEAD OPTIMIZATION AS A SELF-AVOIDING RANDOM WALK

Delaney [26] recently proposed that the temporal trajectory of a lead optimization project in chemical space can be modeled by a self-avoiding random walk (SAW). A random walk (RW) is a mathematical formalization of a trajectory which is created by successively moving a point in discrete jumps in a random direction [27]. An SAW is a popular variation of the RW, whereby it is forbidden to visit the same place twice during the walk, that is, the SAW trajectory cannot intersect with itself [28]. Indeed, there are several direct analogies between the chemical space trajectory of a lead optimization project and an SAW, for example,

- Lead optimization starts from a well-defined point in chemical space (the lead structure).
- Each compound synthesized is *structurally similar* to the previous one, since lead optimization projects do not randomly hop around the entirety of chemical space. Therefore, each step between two molecules in chemical space is small.
- Once synthesized and tested, no compound is considered twice during the advance of a lead optimization project. Thus the trajectory in chemical space is self-avoiding.
- The target point in chemical space (the drug candidate) is *a priori* unknown. A lead optimization project terminates only once the successful compound has been synthesized and tested against the desired property profile criteria.

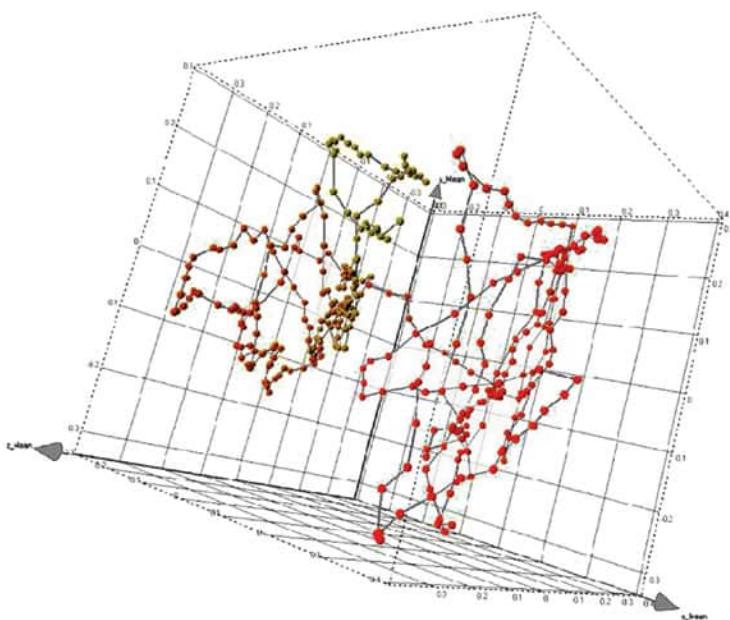
We applied Delaney's [26] methodology to visualize the progression of the 382 compounds in Project B using the following steps:

1. The compounds were represented using MolPrint 2D circular 1024-bit fingerprints, generated in Pipeline Pilot.
2. The original 1024 dimensions were reduced to three using the Sammon mapping [29] algorithm implemented in MATLAB [30], using Soergel distance (1-Tanimoto) as the distance measure, following the approach of Clark et al. [31]. Sammon mapping retains the inter-object similarities within a dataset, so that similar objects remain close in the reduced dimension space.
3. The compounds were ordered chronologically by their sequential registration numbers (e.g., GSK-12333). Then a moving average was applied to each of the three reduced dimensions, as well as to the corresponding potency values. A moving average takes a fixed-length window on a sequence of numbers, averages the values in that window, and moves the window on by one step and repeats. In this study, the moving window size was set to 10.
4. The resulting averages were plotted in a 3D scatter plot. Chronologically successive points were connected by a line in order to reveal the SAW trajectory. The final plot was generated in Spotfire [32], color-coded by the corresponding average potency against the project's primary assay (Figure 9.7).

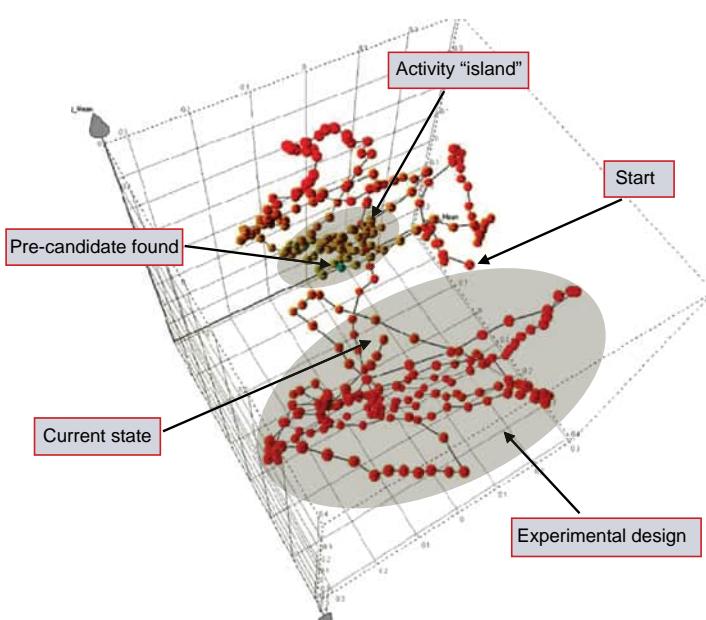
The shape of the plot in Figure 9.7a is very similar to Delaney's SAW plots, both his simulated plots and plots generated using Syngenta lead optimization data. The plot in Figure 9.7b is annotated with project-specific milestones, such as the start, current state, and pre-candidate discovery points in time. Soon after the start of the exploration of the particular chemical series, the project reached an activity "island," that is, a region in chemical space featuring molecules with consistently high potency values (shaded area). A molecule from that region was chosen as a pre-candidate compound (highlighted in light green). The project continued and around the middle of the timeline the exploration took a notably different direction in chemical space (shaded area) when the medicinal chemists explored a much larger region in space but, as indicated by the red color in that area, without success. After discussions with the medicinal and computational chemists involved in the project, it was revealed that this region of chemical space corresponded to a change in the chemical array strategy. Specifically, the newer arrays were designed using the principles of experimental design [33], a chemometrics method that allows for a statistically controlled variation of R-groups attached to the chemical core.

9.6 INSIGHTS FROM THE DATA ANALYSIS

The analyses described previously were found to be extremely informative by both computational and medicinal chemists at GSK and were the source of many constructive discussions. The visualization and multiobjective scoring of the entire set of compounds was thought to provide a useful retrospective documentation and



(a)



(b)

FIGURE 9.7 3D SAW-like trajectories revealing the chronological progress of a lead optimization project in chemical space. Every point is the average of the x , y , and z coordinates of 10 chronologically ordered molecules. Chronologically successive points are connected by a line to reveal the trajectory in 3D chemical space. In (a) points are color-coded by their corresponding average potency value. Red indicates weak activity, whereas green indicates strong activity against the project's primary assay. In (b) the plot is annotated by project-specific milestones such as the start, finish, pre-candidate discovery (shown in light green color), and change in array chemistry approach (gray shaded). An identified activity "island" is also highlighted in gray. For color details, please see color plate section.

evaluation of a project. The addition of the chemical array and chemotype dimensions allowed for a more finely grained comparative examination of the performance and effectiveness of the project across a multitude of criteria and objectives. The visualization of an optimization profile during a project was also thought to provide valuable information to help guide resources management and decision making.

The SAW-like plots allowed for the identification of milestones and other “hotspots” along the timeline of the project, and provided an annotated map of the lead optimization journey in chemical space. Additionally, the medicinal chemists could identify regions that were thoroughly explored and regions that exhibit poor values of an experimental property. Furthermore, different array design strategies could be visually compared, in terms of coverage of chemical space or identification of activity “islands.” The combination of three crucial lead optimization parameters, namely, chemical space, time, and biological activity, in a single plot provides a very informative, high-level summary, which serves as a documentation of the progress of the project. Such a plot can be very helpful not only to the medicinal chemists involved in the particular lead optimization project but also to scientists or line managers who were not directly associated with it.

9.7 EXTRACTING INFORMATION ON ARRAYS FROM THE ARCHIVE

The data on Projects A and B used in the earlier analyses was compiled manually with some data coming directly from the GSK database and some being gathered directly from the chemists that worked on the projects. At the outset of the study, the hope was that the data analysis methods could be applied more widely to lead optimization projects within the GSK archive to allow generalized rules to be devised such as those described in Section 9.1. However, the data gathering process was too time consuming to be extended to more projects and, furthermore, would be impossible in most cases due to the chemists either no longer being available or being unable to remember all of the decisions made during a project and the exact sequence of events that took place. The following section describes our attempts to develop a series of algorithms to allow an array-centric reconstruction of projects. The extent to which this could be achieved was evaluated using Project A.

9.7.1 Annotating by Arrays

In GSK, all biological data are stored in a main database, containing structures, experimental assay results, calculated properties, and other relevant information, such as dates of experiments as well as a specific project name and biological target (e.g., oral PDE4 inhibitors), under which the experiments were carried out. Every project is assigned a *project code*, which is not necessarily unique for every biological target, that is, there can be many project codes for a single biological target. When an array of compounds is synthesized, the members of the array are registered in an electronic lab notebook (eLNB), and given reference numbers of the format Nxxxx-xxx-xx, where x is a numeric digit (e.g., N1035-175-29). The first two parts of the

reference number (e.g., N1035-175) refer to the particular “page” of the eLNB where the experiment is registered for the first time.

In theory at least, since the members of an array are synthesized combinatorially as a batch, they should all be registered on the same eLNB page and have the same first two parts of their eLNB reference code in common. This is, therefore, a straightforward way to identify arrays and classify molecules into their respective arrays. In practice, however, it became apparent that the eLNB definition of an array was not always consistent. In fact, the chemists would sometimes break the array into two parts for practical reasons and register them in two separate eLNB pages. As this was completely undocumented, there was no definitive way to go back and trace when this had happened in order to try to reallocate molecules into their original arrays.

9.7.2 Automatic Chemotype Detector

A very important concept during lead optimization is that of the chemotype, also known as the chemical scaffold, core, framework, or template [34]. Chemotypes are the structural cores upon which different types of substituents can be attached. According to common lead optimization practice, chemists focus their work on a specific congeneric chemical series (or structural class), that is, a family of structures which share the same chemotype or scaffold. The members of a chemical array are very likely to share the same core structure or chemotype. An algorithm was therefore developed to automatically and efficiently detect sets of compounds that share the same chemotype without any prior knowledge of a particular project. There is no universal definition of a chemotype; however, some general observations are as follows:

- They almost always consist of rings with well-defined substituents.
- The rings are usually aromatic and/or unsaturated.
- The rings often contain heteroatoms.

These three simple observations were the main drivers for the implementation of the automatic chemotype detector. The algorithm was developed in Pipeline Pilot [15]. It takes a set of molecules as input and then outputs a list of ring systems sorted in order of structural complexity (SC). These ring systems are then mapped back to the original set of molecules.

The algorithm is as follows:

1. The chemical structures for a given biological target and project code are retrieved from the main database and preprocessed (SMILES canonicalized, duplicates, and invalid entries removed).
2. Each input structure is fragmented into ring assemblies, that is, contiguous ring systems (Figure 9.8). The ring assemblies from all the compounds are collected and their unique occurrence (based on the canonical SMILES of their structure) is added to a list. Ring assemblies which occur infrequently (e.g., they are present in less than 10 of the input molecules) are removed from the

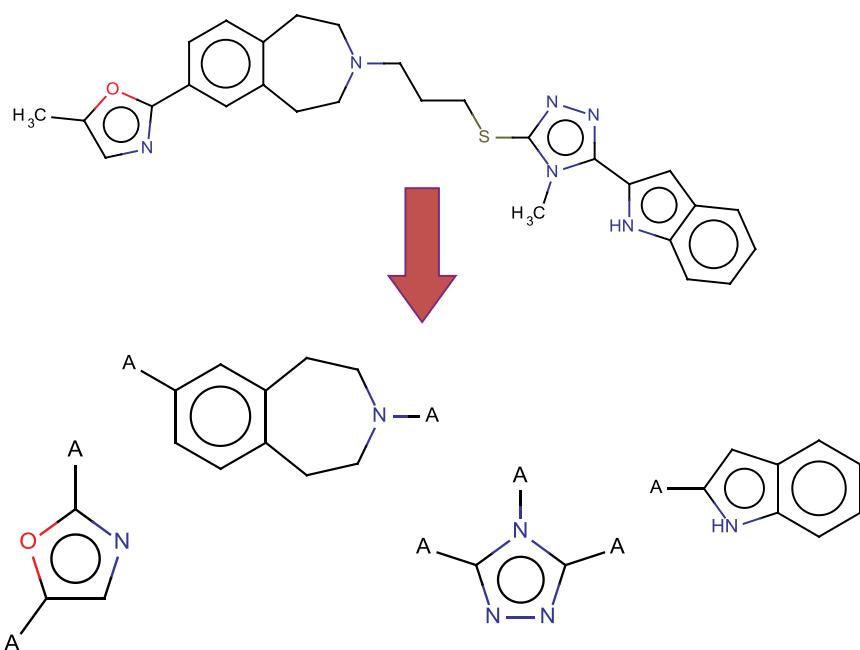


FIGURE 9.8 The input molecule is split in ring assemblies (i.e., contiguous ring systems). Side chains are marked by “A.” For color details, please see color plate section.

list. For each remaining ring assembly, a complete list of the original compounds which contain it is also stored.

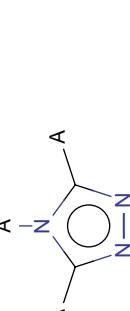
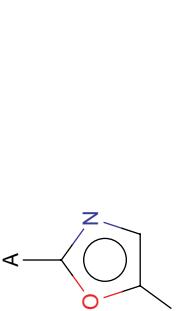
3. The ring assemblies are scored according to their SC score. This score is effectively a sum of five simple descriptors, namely, number of side chains, number of rings, number of linker side chains, number of heavy atoms, and number of heteroatoms (N, O, and S) (Table 9.3). The ring assemblies are sorted by their SC score and those with the highest scores are output candidate chemotypes.

The chemotype detection algorithm was evaluated on a publicly available dataset of 2644 compounds consisting of literature compounds tested against hERG channel blocking and approved drugs [35]. A total of 1995 distinct ring assemblies were identified, of which 61 had SC score above 12 and frequency above 20 and were suggested as chemotypes; some examples are shown in Table 9.4. When the automatic scaffold detector was applied to Project A, the six known chemotypes were successfully detected.

9.7.3 Detecting Seed Compounds

According to the lead optimization workflow illustrated in Figure 9.1, the most important reference structures during a lead optimization project are the lead(s), seed(s), and candidate(s) compounds. A seed compound is a promising compound with regard to one or more desired properties, and is used as a prototype for the design and synthesis

TABLE 9.3 Calculation of the Structural Complexity Score for Each of the Four Ring Assemblies of Figure 9.8

Ring Assembly	Number of Chains	Number of Rings	Number of Linkers	Number of Heavy Atoms	Number of Heteroatoms	SC Score
	2	2	2	11	1	18
	1	2	1	9	1	14
	3	1	2	5	3	14
	2	1	1	5	2	11

The score is the sum of five descriptor counts, namely number of side chains, rings, linker chains, heavy atoms, and heteroatoms.

TABLE 9.4 Chemotypes Identified in the hERG Publicly Available Dataset [35] along with Their Frequency of Occurrence and the Final Complexity Score

Chemotype	Frequency	SC Score
	20	26
	106	15
	20	22
	36	19
	41	20

Familiar scaffolds such as triazoles, quinolines, and penams were identified.

of a new array of analogous compounds with the aim of exploring the chemical space around it. The seed compounds serve as milestones for the local SARs and can shift the focus of a project to different chemical families or subfamilies. Therefore, in a retrospective analysis of arrays, the notion of the seed is crucial in understanding the purpose of the array and whether or not it was successful in achieving its aims, for example, producing a change in structure that corresponded to the intended or desired change in property. However, accurately identifying the seed for an array is a very difficult, and often impossible, task. This is because the information about seeds is not reported in any of the available databases in GSK. Even in the highly sophisticated

eLNB system, which is a corporate-wide detailed account of every scientific experiment carried out in GSK, the seed structure of a new array design is usually not explicitly recorded. Even if it is, gathering this information requires browsing manually through a large archive of experiments, which is a very time-consuming exercise. It seems that the only way to retrieve a complete list of the seeds of a project would be either to have personal interviews with the medicinal chemists responsible for the project, or to delve into the huge number of reports and presentations produced during a project. Both of these are impossible in practice. In the first case, either the chemists will have already forgotten about the project or they may not even work for the same company. In the second case, it is a Sisyphean task for one person to accurately reproduce the history of a project from reports and the presentations produced at the time.

An attempt was therefore made to identify automatically the seed compounds in real project data. Three main criteria were defined for “seed-likeness” based on the traditional array design workflow:

- At least modest potency against the primary biochemical assay
- Structural relevance to the array members
- Chronological precedence compared to the array members

The rationale for these is as follows: The seed should be a promising compound and thus exhibit good potency against the drug target, otherwise it would not be considered at all. It should be of the same chemotype/series as the array members since the chemists explored the space around it with analogous structures. Finally, since it was used as the starting idea, it should have been first synthesized earlier than the array members.

The protocol was designed to be fast and general, and therefore largely chemistry-agnostic, for example, it does not consider any R-group fragmentation in particular, nor does it try to apply a reaction scheme such as the RECAP methodology [36]. The seed detector algorithm works as follows:

1. Two pools of compounds form the input. The first pool contains molecules which were synthesized in chemical arrays and tested against a primary biochemical assay during a lead optimization project. The second pool is a superset of the first and contains all the compounds (both array members and singletons) that were ever tested against the same biochemical assay.
2. For each chemical array in the first pool of structures, the maximal common subgraph is found, that is, the largest set of atoms and bonds in common, among the array members [37]. This subgraph is then abstracted to a Murcko scaffold [38]. This allows the identification of the general structural framework of the array members. The scaffold is then submitted as a substructure search query against the second, larger pool of compounds. The compounds that have the same substructure as the query are retrieved. These first steps ensure the second criterion for seed-likeness (structural relevance to the array members) is met.
3. The retrieved compounds are filtered by potency and date. Only relatively active compounds (i.e., having a $\text{pIC}_{50} > 5.5$) are selected and from them only those

that are older than the oldest member of the array are retrieved. This filtering deals with the remaining seed-likeness criteria (potency and chronology).

- In case of more than one candidate seed, all of them are reported back for the user to decide.

An example of the workflow of the algorithm is illustrated in Figure 9.9.

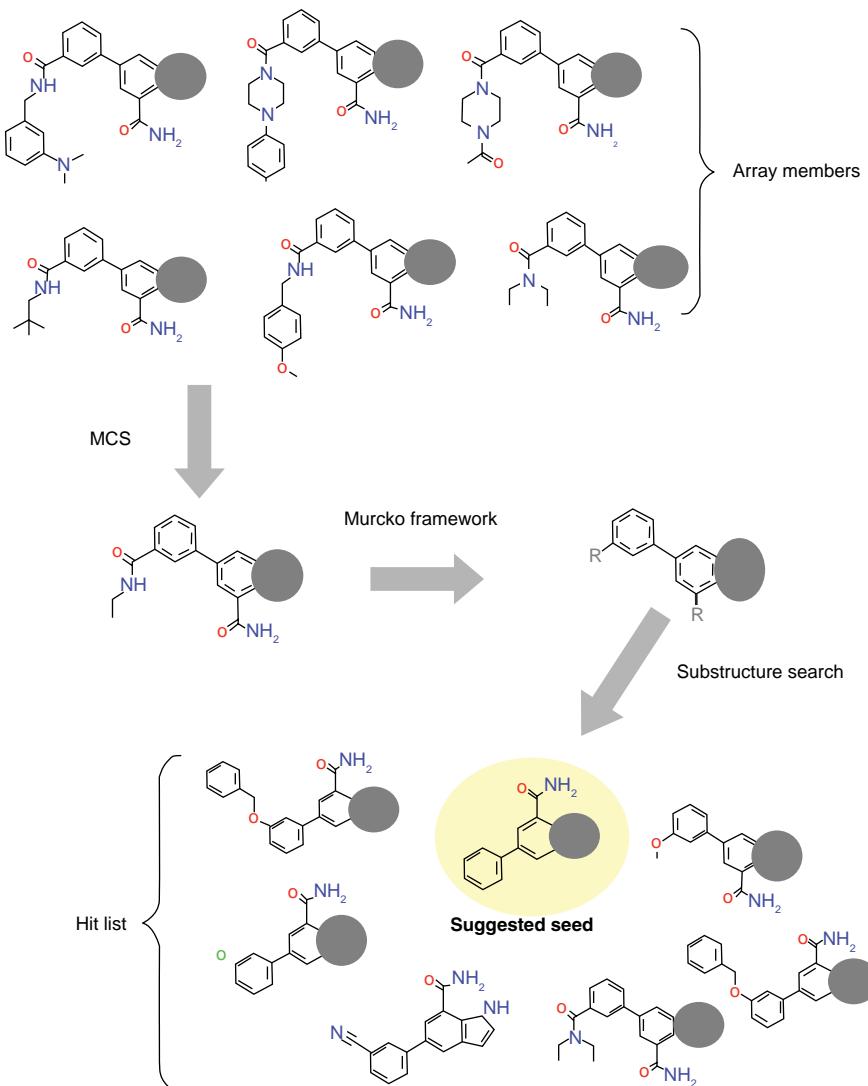


FIGURE 9.9 The seed detector workflow applied to real lead optimization data. The algorithm seeks to identify the seed of each array by applying the seed-likeness criteria to all the relevant molecules. For legal reasons, the molecules depicted in this example have a part of their structure covered. For color details, please see color plate section.

The seed detector workflow was implemented in Pipeline Pilot [15] and applied to Project A. The first pool of structures contained 1733 molecules belonging to 109 chemical arrays and 94 of them had more than three members. Finding and collating the second pool of structures was a challenging task. In an effort to be as inclusive as possible, several small datasets were collated. Each came from the main database of biological data and had been assigned a different project code. This was because there were several project codes associated with the single protein target and in reality the chemists can take ideas and use seeds that belong to a different but related project. The final second pool contained 7207 unique structures.

In total, the algorithm suggested three candidate seeds for each of the 94 chemical arrays. The lead medicinal chemist on the project was asked to assemble a list of the actual seeds. This list contained a total of 25 unique structures for the 94 arrays. In 83 out of the 94 arrays, the actual seed was among the three structures suggested by the seed detector algorithm. Thus although it was possible to produce a candidate list of seed compounds that was likely to contain the seed for most of the arrays, it was not possible to uniquely identify the seed.

9.8 CONCLUSIONS

Although the data analysis of the retrospective lead optimization projects was thought to be very informative by both computational and medicinal chemists, it proved too difficult to automate the accurate reconstruction of lead optimization projects. This severely limited the extent to which the historical data could be exploited through data mining and restricted the extent to which generalized rules could be devised to assist in the prospective design of arrays. Furthermore, the medicinal chemists also pointed out that an array-centric approach would not capture important information about singleton (i.e., nonarray) compounds synthesized during a lead optimization project which contribute equally and could give new and significant information about the local SARs. Thus the model of lead optimization shown in Figure 9.1 proved to be somewhat idealized and that in Figure 9.10 can be considered to be more representative of real-life array design.

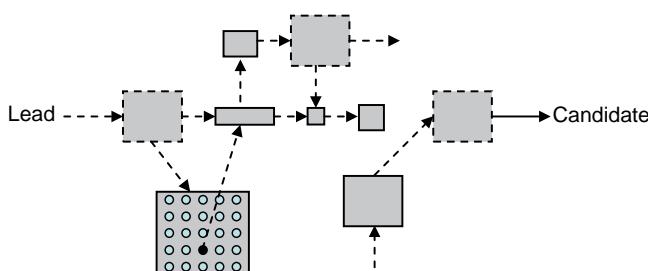


FIGURE 9.10 Real-life array design for a given chemical series: Neither the seeds nor the arrays can be authoritatively defined most of the time. The precise relationships and context of the arrays (dashed lines) cannot be extracted from the existing data. The actual array succession is also uncertain.

Another factor that limited the analysis is that many of the decisions made during a project are not documented, including the particular goal of an array. Thus, it was not possible to define an array as successful or otherwise, for example, an array which results in a decrease in potency relative to the seed may not indicate failure if the goal was to increase selectivity or to reduce hERG liability. Delaney [26] also reported severe difficulties with the extraction of crucial project-related information, such as the desired property profile for the lead and candidate structures, the prerequisite nature of the SARs around the lead, and even the time allowed for the projects to run-on [26]. MacDonald and Smith [39] have noted the common practice in large pharmaceutical and agrochemical companies whereby important parameters and milestones in projects are kept in spreadsheets and shared only by the members of a department. Given the fact that there are tens of scientists involved, assigned to different departments and specialties (physical chemistry, medicinal chemistry, combinatorial chemistry, process chemistry, pharmacology, pharmacokinetics, toxicology, etc.), any attempt to collectively document lead optimization projects retrospectively is impossible. Oprea [13] has also noted that decision making and procedures for lead selection and optimization are typically not well documented and has highlighted the value of documenting information such as the prerequisites and milestones for lead optimization. In a similar vein, Lipinski [40] has suggested a “debriefing” procedure for retiring scientists, so that their experiences and tacit knowledge of lead optimization may be captured.

Systems aimed at capturing knowledge during drug discovery are beginning to appear: for example, the CODD [41] and PFAKT [42] systems focus on capturing compound-specific knowledge; ActWiki [43] and ROCK [44] are designed to capture more general medicinal chemistry knowledge; and the DEGAS [45] system captures information with the aim of enabling more effective communication between multiple partners. As the number of projects that make use of these systems increases, there will be increased opportunities to mine the data and apply lessons learned to future efforts. We also advocate a more systematic approach to proactively documenting the goals in lead optimization and in annotating the data appropriately in order to maximize the benefits for future data mining efforts. The information that would be required to extend the analysis carried out here includes: the array a compound belongs to; the seed compound from which an array is constructed; the order in which arrays were designed and the progression from one array to another; important milestones such as candidates and dead-ends; and the design goals for each array.

ACKNOWLEDGMENTS

We gratefully acknowledge Gianpaolo Bravi, Anthony Cooper, Simon Macdonald, and John Pritchard for very many useful discussions that contributed to this work. The work was funded by the Engineering and Physical Sciences Research Council and GlaxoSmithKline.

REFERENCES

1. Oprea TI, Davis AM, Teague SJ, et al. J Chem Inf Comput Sci 2001;41:1308–1315.
2. Valler MJ, Green D. Drug Discov Today 2000;5:286–293.
3. Egan WJ, Walters WP, Murcko MA. Curr Opin Drug Discov Devel 2002;5:540–549.
4. Waring MJ. Expert Opin Drug Discov 2010;5:235–248.
5. Oprea TI. J Comput Aided Mol Des 2002;16:325–334.
6. Bleicher KH, Bohm HJ, Muller K, et al. Nat Rev Drug Discov 2003;2:369–378.
7. Baringhaus KH, Matter H. Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters. In: Oprea TI, editor. *Chemoinformatics in Drug Discovery*. Weinheim: Wiley-VCH; 2004. p 333–379.
8. Gillet VJ. Curr Opin Chem Biol 2008;12:372–378.
9. Leach AG, Jones HD, Cosgrove DA, et al. J Med Chem 2006;49:6672–6682.
10. Griffen E, Leach AG, Robb GR, et al. J Med Chem 2011;54:7739–7750.
11. Papadatos G, Alkarouri M, Gillet VJ, et al. J Chem Inf Model 2010;50:1872–1886.
12. Lewis RA. J Med Chem 2005;48:1638–1648.
13. Oprea TI. Chemoinformatics in lead discovery. In: Oprea TI, editor. *Chemoinformatics in Drug Discovery*. Weinheim: Wiley-VCH; 2005.
14. Bender A, Mussa HY, Glen RC, et al. J Chem Inf Comput Sci 2004;44:1708–1718.
15. Accelrys Inc. 2010. Pipeline Pilot. Available at <http://accelrys.com/products/pipeline-pilot/>. Accessed 2010 Mar 15.
16. Peltason L, Bajorath J. J Med Chem 2007;50:5571–5578.
17. Guha R, Van Drie JH. J Chem Inf Model 2008;48:646–658.
18. Segall MD, Beresford AP, Gola JM, et al. Expert Opin Drug Metab Toxicol 2006;2:325–337.
19. Derringer G. Qual Prog 1994;27:51–60.
20. Li J, Ma C, Ma Y, et al. Appl Microbiol Biotechnol 2007;74:563–571.
21. Gillet VJ, Khatib W, Willett P, et al. J Chem Inf Comput Sci 2002;42:375–385.
22. Li H, Zhang H, Zheng M, et al. BMC Bioinformatics 2009;10:58.
23. Dey F, Caflisch A. J Chem Inf Model 2008;48:679–690.
24. Brown N, McKay B, Gasteiger J. J Comput Aided Mol Des 2006;20:333–341.
25. Birchall K, Gillet VJ, Harper G, et al. J Chem Inf Model 2008;48:1558–1570.
26. Delaney J. Drug Discov Today 2009;14:198–207.
27. Edwards AM, Phillips RA, Watkins NW, et al. Nature 2007;449:1044–1048.
28. Hayes B. Am Sci 1998;86:314–319.
29. Sammon JW. IEEE Trans Comput 1969;18:401–409.
30. Mathworks. 2008. Matlab for technical computing [Online]. Available at <http://www.mathworks.com/>. Accessed 2008 July 21.
31. Clark RD, Patterson DE, Soltanshahi F, et al. J Mol Graph Model 2000;18:404–411.
32. TIBCO. 2010. Spotfire. version v8.0. [Online]. Available at <http://spotfire.tibco.com/>. Accessed 2012 May 14. TIBCO Software Inc., Palo Alto, CA.
33. Leardi R. Anal Chim Acta 2009;652:161–172.
34. Katritzky AR, Kiely JS, Hébert N, et al. J Comb Chem 1999;2:2–5.

35. Doddareddy M, Klaasse E, Shagufta, et al. ChemMedChem 2010;5:716–729.
36. Lewell XQ, Judd DB, Watson SP, et al. J Chem Inf Comput Sci 1998;38:511–522.
37. Raymond JW, Willett P. J Comput Aided Mol Des 2002;16:521–533.
38. Bemis GW, Murcko MA. J Med Chem 1996;39:2887–2893.
39. Macdonald SJF, Smith PW. Drug Discov Today 2001;6:947–953.
40. Lipinski CA. Technical and people disconnects hindering knowledge exchange between chemistry and biology. 227th ACS National Meeting, American Chemical Society, 2004: Anaheim, CA.
41. Robb G. Hypothesis-driven drug design using wiki-based collaborative tools. In UK-QSAR and ChemoInformatics Group Meeting; 2009 May 14; 2011. Pfizer: Sandwich, UK.
42. Brodney MD, Brosius AD, Gregory T, et al. J Chem Inf Model 2009;49:2639–2649.
43. Sander T, Freyss J, von Korff M, et al. J Chem Inf Model 2009;49:232–246.
44. Mayweg A, Hofer U, Schnider P, et al. Drug Discov Today 2011;16:691–696.
45. Lee M-L, Aliagas I, Dotson J, et al. J Chem Inf Model 2012;52:278–284.

CHAPTER 10

EXPLORATION OF STRUCTURE–ACTIVITY RELATIONSHIPS (SARs) AND TRANSFER OF KEY ELEMENTS IN LEAD OPTIMIZATION

HANS MATTER, STEFAN GÜSSREGEN, FRIEDEMANN SCHMIDT,
GERHARD HESSLER, THORSTEN NAUMANN, and
KARL-HEINZ BARINGHAUS

10.1 INTRODUCTION

The efficient conversion of lead structures with promising biological properties into viable drug candidates fulfilling a multitude of criteria is a challenging step in medicinal chemistry with high impact toward the identification of a clinical candidate [1, 2]. Here, drug design is at the central stage in this research workflow, as this discipline generalizes key learnings from the current body of data and information, which will be prospectively applied to design novel molecules.

For an efficient optimization of lead structures toward preclinical candidates, the development of reasonable structure–activity relationships (SARs) for compound series is of utmost interest. The widespread application of modern technologies like parallel synthesis and combinatorial chemistry [3, 4] in the pharmaceutical industry has led to an exponential increase of molecules and data points from corresponding biological assay. Both the amount and the quality of data have changed enormously over the past decades in scope as well as in degree of precision. The extraction of relevant information and knowledge to support informed decisions and—at the same time—to respond to the cost pressure in the industry is therefore clearly one of the challenges at this stage in drug discovery projects.

The focus of this chapter is to review current approaches for SAR analysis in lead finding and optimization. We will first present methods to analyze SARs and to

extract relevant knowledge from chemical series. This discussion includes the similarity principle, activity distributions, and the activity landscape concept. Here, visualization is a key concept to extract knowledge and correlate chemical changes to biological responses. In Section 10.2, we outline methods for rescaffolding based on different molecular representations to identify novel series beyond the currently explored ones. Once such a novel series has been established, the transfer of SAR elements between those series is of high interest for an efficient optimization. At the same time, it is also important to avoid unwanted activities for secondary targets. Consequently, we outline and show the applications for a concept to address and modulate those undesirable biological antitarget activities to arrive at compounds with fewer side effects [5].

Computational approaches are nowadays indispensable prerequisites for SAR analysis, as they help to formulate and capture the rules for optimizing chemical series for specific biological targets. This is in particular true, if the datasets under investigation become larger and more diverse. Intrinsically, some approaches split a molecule into a scaffold and functional group decoration. Such a conceptual separation of several regions in a molecule is often motivated by chemical considerations, that is, route of synthesis and strategic areas in a molecule, where optimization by rapid side-chain replacement is easily possible. Other approaches express the relationship between molecules using molecular descriptors, which provide a more holistic comparison using statistical techniques. Both concepts find the application in different SAR analysis approaches and will therefore be presented here.

One of the fundamental assumptions of SAR analysis and transfer is that similar and active compounds in a chemical series interact via a consistent binding mode with their target binding site. Furthermore, the protein-binding site is assumed to adopt similar conformations upon binding similar ligands. Any SAR analysis could therefore suffer from compounds, which do not fulfill these prerequisites upon binding to their target protein, thus producing outliers. As we will show in the following chapter, it is important to detect and analyze those compounds and also molecules with small changes leading to unfavorable interactions with the protein-binding site in order to identify structure–activity cliffs [6] in chemical series.

In general, there are many motivations for SAR transfer in lead optimization, which are primarily aiming to overcome particular liabilities in an actual chemical series. In some instances, a lot of SAR information has been accumulated for a single chemical series (chemotype), while profiling might reveal series-related side effects or ADMET issues, which could be attributed predominantly to the chemical scaffold. This situation often prompts a research team to change the underlying chemical scaffold of a particular series (rescaffolding). In order to rapidly arrive at potent molecules in the novel series, key SAR elements from the first series are then transferred to the novel scaffold. This SAR transfer could also only be applied under the assumption that related series by 3D-similarity or binding site interaction similarity will also bind to the protein-binding site in a similar manner.

10.2 METHODS FOR SAR ANALYSIS

10.2.1 Similarity Principle

The majority of SAR analyses in today's medicinal chemistry projects are based on the *similarity principle*, as formulated by Johnson and Maggiora in 1990 [7]. It states that chemically similar molecules are likely to exhibit similar physicochemical and biological properties. Of course, this fundamental assumption was already explored in medicinal chemistry decades earlier and even before the advent of computational tools, resulting in the discovery of many clinical development candidates.

The similarity principle provides a conceptual framework for a diverse range of chemoinformatics data analysis approaches [8] like similarity searching [9] for series enrichment, virtual screening [10], clustering [11], and quantitative structure–activity relationship (QSAR) [12]. In particular k -nearest neighbor QSAR statistical techniques [13, 14] for model building directly apply the similarity principle, as the activity of novel compounds is estimated as weighted average of a predefined number (k) of chemically similar members of the training set.

However, the identification of similar compounds also prompts for a critical assessment of descriptors and metrics to quantify chemical similarity [15]. A detailed review of descriptors and a comparison of two-dimensional versus three-dimensional descriptors has been given elsewhere [16, 17, 18]. Often information encoded by different descriptors shows a significant degree of overlap and thus might be employed interchangeably to a certain degree [18–20]. Nevertheless, it is important to appreciate that different descriptors could result in alternative interpretations of an SAR within a series. According to D.J. Livingstone and the authors' own opinion, “it is very likely that the best properties to use are heavily dependent on the nature of the intended application” [18]. But these considerations also reveal the necessity to explore different types of descriptors in order to arrive at a solid SAR interpretation with relevant rules for future design.

Traditional SAR investigations often tend to interpret the effect of structural changes on activity in a qualitative sense. Synthesis efforts to explore an SAR might also often involve structural changes to completely destroy biological activity [21]. QSAR approaches, on the other hand, attempt to quantify gradual changes and relate structural changes to interpretable descriptors. In order to merge traditional SAR and QSAR, both viewpoints have to be integrated and physicochemical changes related to structural and biological changes should be considered in a complementary fashion [21].

10.2.2 Molecular Scaffolds

In general, the term “scaffold” is applied in a multitude of ways to describe a molecular core structure, for example, as a common element of a chemical series or for building blocks in synthetic chemistry. For closely related compounds, such a scaffold could be defined by their maximum common substructure (MCS) [22].

Several graph-based algorithms for the determination of the MCS have been described; recent improvements can also efficiently handle large sets of molecules [23]. However, the MCS identification does not necessarily result in chemically meaningful substructures with intact ring systems and functional groups. The absence of a general definition of scaffolds is a problem, when comparing success stories from virtual screening and scaffold hopping [24]. This inconsistency can also be extended to the problem of designating substructures, fragments and functional groups, which are by no means consistently defined [25]. The common RECAP approach [26] offers a pragmatic possibility to split chemical structures into building blocks with one attachment point and scaffolds having multiple attachment points guided by pre-defined synthetic disconnections. Those synthetically accessible moieties could then serve as building blocks for library design [27], but also as privileged substructures, if enriched in active molecules for protein targets or target families.

Scaffold abstraction approaches like topological frameworks or reduced graphs were first proposed by Bemis and Murcko [28] and further explored by Xu and Johnson [29]. Groupings into series by some fuzzier similarity are obtained by using different abstraction levels for atom types, bond types, and ring topologies.

These descriptions have recently been integrated into structure classification systems for molecules, as summarized in an informative review by Schuffenhauer and Varin [30]. The most prominent organization schemes for scaffolds based on hierarchies and relationships between scaffolds include the molecular equivalence number structural classification system (*Meqnum*) from Xu et al. [29, 31], the hierarchical scaffold clustering system *HierS* from Wilkens et al. [32], and the *scaffold trees* from Schuffenhauer et al. [33], first applied toward a hierarchical organization of natural products [34]. The primary motivation for those systems is to facilitate the interactive analysis of large databases like high-throughput screening (HTS) sets [30]. Such a hierarchy provides a generally applicable and useful reference framework for analysis and scaffold design. The resulting substructures at each node of the hierarchy are chemically meaningful and provide intact motifs.

One interesting solution to generate entire scaffold hierarchies and 2D structure-based classification schemes is implemented in the program *Distill* [35]. Here, a dendrogram is generated from a hierarchical MCS analysis, where each node represents a substructure of the entire dataset and the total set of compounds containing that substructure. These nodes can be color-coded according to properties, such as average activity, to relate characteristic substructures plus their child and parent nodes to that property. Such an annotated dendrogram could serve to illustrate how incremental structural change might influence activity, as well as how combinations of changes may provide a pathway to activity. However, the way, how these hierarchical substructure trees are generated, does again not necessarily produce chemically relevant, intact moieties.

10.2.3 Privileged Substructures

The prominent concept of privileged substructures in medicinal chemistry summarizes a view on molecular scaffolds or significant parts of chemical structures with a high propensity to consistently interact with related protein targets [36, 37]. Here, the

existence of readily available chemical building blocks offers an interesting perspective for focused library design, namely in the lead identification phase.

The existence of topologically related areas in target families like the hinge-binding region in protein kinases allows to understand the preference of chemical motifs with complementary hydrogen-bonding pattern and flatness to interact with this well-characterized protein motif [38]. This situation could be different in other families like serine proteases. From a topological view of the protein site, it depends, whether the privileged substructure or binding motif in a substituent is situated in a subpocket with only a single attachment vector or in a central region with multiple vectors pointing to different protein subsites.

In some cases, the link between chemical substructures and protein-target families has been overemphasized, as it was assumed that some motifs interact exclusively with a family. However, the rapid accumulation of broader arrays of protein–ligand interaction data from numerous profiling assays is revealing that privileged substructures for particular protein families are in fact also represented in active molecules against totally unrelated targets [39]. Those relationships were often not discovered due to data incompleteness. Due to the typical sparseness of selectivity data, there is sometimes a misleading statistical evidence for target selective scaffolds [25]. It is thus reasonable to consider the relative enrichment of privileged substructures in different target families rather than their exclusive occurrence [25, 38].

This entire concept was extended by mining [40, 41] public databases such as BindingDB [42], PubChem [43], or ChEMBL [44]. In such a study, all human target protein pairs were extracted, which share at least five active compounds [40]. This selectivity-centric analysis of protein–ligand data unveils more than 200 scaffolds that appear to be selective for certain target families. A subset of these scaffolds refers to highly selective molecules compared to neighboring proteins [40].

10.2.4 Investigating the Outliers: Activity Cliffs

Experience in medicinal chemistry has taught that there are significant exceptions to the similarity principle in many series. In fact, often small changes in a lead structure completely destroy biological activity or—in a more favorable manner—result in a boost of potency for a target. This well-known phenomenon has recently been coined “activity cliff” [6] or “activity hotspot” [5]; we will therefore use both terms in our discussion.

Y.C. Martin et al. reported an instructive HTS data analysis for 19,533 compounds tested in one or more of 115 biological assays at Abbott to investigate the correlation between chemical and biological similarities [45]. This analysis revealed a significantly higher frequency of activity cliffs than expected from only analyzing lead series for a particular protein target. For this mid-sized set of experimental data, there is only a 30% chance that a compound with a similarity larger than a Tanimoto coefficient [46] of 0.85 to a previously known active is active as well on the same target. From this study, one can conclude that similar compounds do not necessarily interact with their target in the same manner for multiple reasons.

These observations provide compelling evidence that the underlying structure–activity landscapes are unevenly distributed. Structure–activity landscapes can be

rationalized as biological response surfaces in chemical space [47]. A visual representation for these SAR landscapes can be obtained by adding a biological activity dimension to a 2D-projection of the chemical space, which sometimes appears to be highly variable. The resulting “biological response surface” produces maps related to topological maps [47]. Any manual or automated SAR analysis is attempting to identify key derivatives, which only slightly differ chemically, but show a significantly different biological behavior. Medicinal chemistry programs in early SAR exploration are therefore systematically probing key functional groups in order to possibly identify a set of essential features, which are typically summarized as “pharmacophore.”

Many instructive exceptions of the similarity principle were summarized by Kubinyi [48, 49]. Some of these differences could be related to unexpected 3D-binding modes of ligands in the protein cavity after only minor chemical modifications. These examples show that it is very difficult to describe chemical similarity in an objective and global manner without considering the protein environment. In fact, lead optimization often takes advantage on these surprises and further explores such “activity cliffs” for biological activity. Regions around early hit structures with a “flat SAR” are typically less often explored.

Global 2D-descriptors are often not suitable to capture these changes. Here more advanced descriptors, sometimes relying on 3D-treatments, are more adequate to model a particular SAR heterogeneity. However, the most valuable sources of information to rationalize activity cliffs are protein–ligand X-ray structures, providing a detailed atomistic understanding of essential protein–ligand interactions [50].

10.2.5 Quantification of Activity Cliffs

The introduction of a structure–activity landscape index (SALI), as proposed by Guha and Van Drie [51], helps to identify and quantify activity cliffs. This SALI is defined as follows:

$$\text{SALI}_{i,j} = \frac{|A_i - A_j|}{1 - \text{sim}(i, j)}$$

Here, A_i and A_j denote the biological activity of compounds i and j , given as pKi, pIC₅₀, or pEC₅₀ values. The chemical similarity between any pair of molecules is then expressed by $\text{sim}(i, j)$. The SALI-value is high for pairs of chemically very similar molecules showing large differences in biological activity. Therefore this index provides a quantitative measure of information captured in informative graphical SAR visualizations like PLS [52–54] $t-u$ plots and neighborhood plots [55].

Compound pairs detected as informative activity cliffs often illustrate key chemical features for activity. These pairs, however, may also often be detected as apparent statistical outliers in quantitative SAR analysis methods [56], since the assumption of SAR continuity is fundamental for QSAR model building and affinity prediction.

For this purpose, we employ internally two model diagnosis plots, as exemplified using a CoMFA [57] model of a series of 107 3-oxybenzamides as factor Xa inhibitors

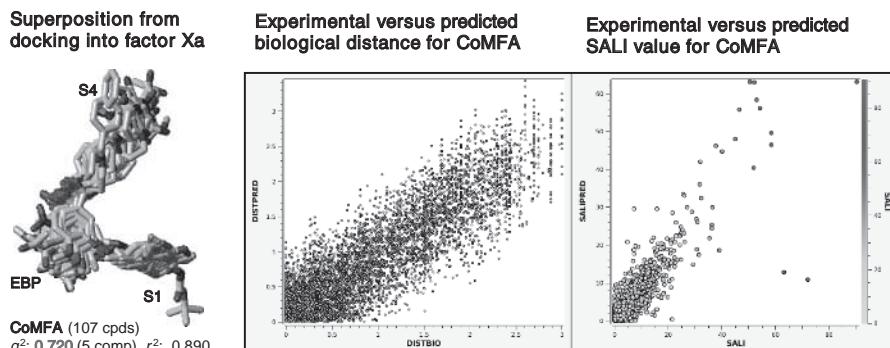


FIGURE 10.1 Model diagnosis plots for 107 3-oxybenzamides as factor Xa inhibitors [58, 59]. Left: Alignment from docking into factor Xa binding site. Middle: Experimental versus predicted biological differences (ΔpK_i) for molecular pairs above a similarity threshold for a CoMFA model (q^2 0.720, r^2 of 0.890, 5 components). Right: Experimental versus predicted SALI values for CoMFA model predictions; informative pairs with high SALI values indicated in red. For color details, please see color plate section.

[58, 59]. In the first scatter plot (Figure 10.1, middle), the experimental biological activity difference (factor Xa pK_i) is plotted versus the predicted difference from applying this PLS-based model. We typically only consider related molecular pairs above a certain similarity threshold to focus on informative analogs for model and SAR interpretation. Except from few outlying pairs, this example model is able to capture all SAR trends in this dataset. This behavior is of course not always observed, when analyzing other models. Additionally, we usually compare the SALI index from experimental data with the SALI index for activity differences from model predictions (Figure 10.1, right). Compound pairs with a high SALI value are colored in red. Except for two pairs in the lower right corner of this plot, this particular QSAR model predicts all relevant SAR trends correctly and thus appears to be useful for affinity predictions. This analysis is for us an informative alternative to recently introduced SALI curves [60]. A closer investigation of underlying physicochemical and structural relationship of such a pair of molecules and potential sources of experimental uncertainty often supports an interpretation of an activity cliff.

10.2.6 Matched Molecular Pairs

SAR is often expressed in an additive manner to indicate that adding a single functional group will add a certain degree of activity within a series. Therefore pairs of chemical structures have been investigated systematically, which differ by only one single small chemical substitution. The specific term “matched molecular pair” (MMP) for those close analogs was introduced by Kenny et al. [61]. An MMP can thus be defined as a pair of molecules with minor structural changes at distinct positions. The subsequent MMP analysis can be focused on single-point or multiple-point transformations. The detection of MMPs was then exploited by Leach et al. in

particular for chemical substitutions favorably influencing distinct physicochemical and ADMET properties [62].

Multiple efficient algorithms [63, 64] for computing MMPs have been reported in the literature. MMPs were applied for investigating different molecular properties such as primary activity, liver microsomal stability, or other ADMET-related properties [62,65–70].

In their informative review, Griffen et al. have compared some key methods for the identification of MMPs [71]. Small and focused MMP datasets with a common scaffold are often analyzed using different variations and abstraction levels of traditional R-group tables. Some formalization of this concept can be done by defining a scaffold or core and plotting substituents at a given position against the property of interest [72].

Supervised approaches on larger, more heterogeneous datasets rely on the identification of predefined transformations for analyzing biological trends. Typically a transformation pair is defined using flexible chemistry line notations like SMILES and SMARTS [73].

Unsupervised approaches provide the most general way to find multiple relationships in a larger dataset. The fastest and most efficient way is to systematically decompose molecules into fragments, which are then indexed to support rapid sorting and database retrieval approaches. Fragmentation schemes often rely on Murcko and Bemis related scaffold definitions [28] plus bond fragmentation schemes. Those often include all exocyclic single bonds and are sometimes extended with double- and triple-bond fragmentations. In this approach, each molecule is just disconnected once and the following process of matched pair identification is based on pairing molecules having a common core structure. The fast approach described by Hussain and Rea [64] for example allows analyzing very large datasets within a reasonable computational time.

The principle concern of MMPs for chemical structures is how the trade-off between specificity and general trend is achieved [66]. A very specifically defined environment for a chemical transformation results in a much narrower dataset for statistical analysis. The extracted trend might then be clearer for this particular problem, but based on a smaller dataset. Hence, any further generalization might be difficult. If on the other hand, the molecular context is completely unconstrained, then the effect of a particular substitution is likely to reflect only simple trends, mainly related to lipophilicity as one of the cardinal properties in medicinal chemistry [62, 66, 67]. The simpler the property for MMP analysis is, the more likely is a direct link to molecular structure, as shown by Leach et al. in their study on solubility and plasma protein binding [62]. Typical control parameters for stable MMP trends are thus related to a minimum number of matched pairs for analysis and a certain degree of structural diversity upon all members in a dataset, excluding of course those linked by a molecular transformation [71].

Interesting applications of this concept include a study from Boström et al. [74] exploring property changes with clear impact on medicinal chemistry [75]. They compared pairs of molecules differing by the topology of their central five-membered heterocyclic moiety, namely 1,2,4- and 1,3,4-oxadiazoles. This minor structural

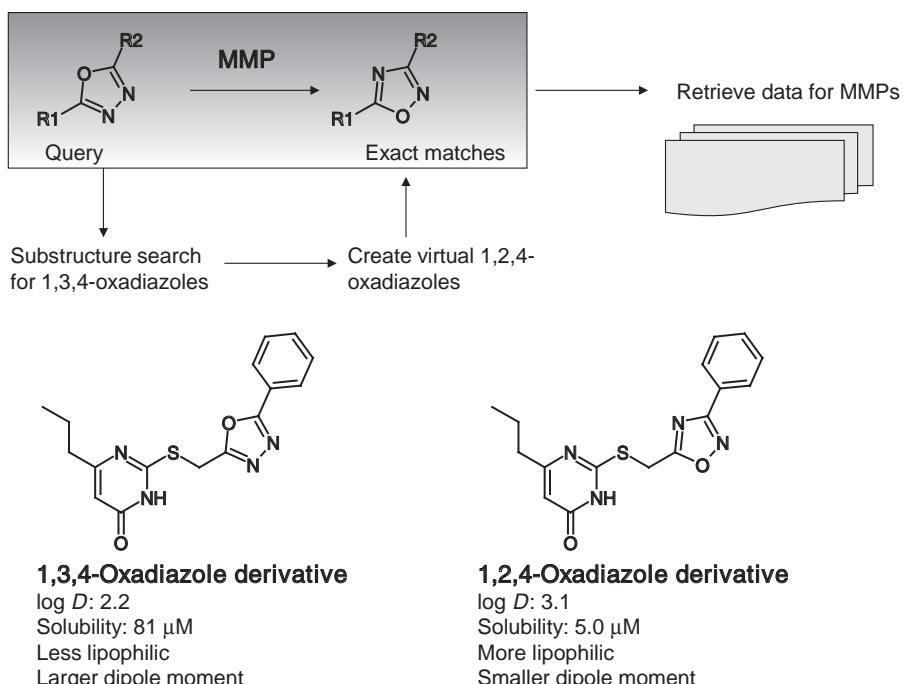


FIGURE 10.2 Matched molecular pair of oxadiazoles differing only by oxygen to nitrogen exchange: 1,3,4-oxadiazole (left) and 1,2,4-oxadiazole (right) with experimental data on lipophilicity and solubility [74].

modification refers topologically only to a positional interchange of an oxygen and nitrogen atom without changing the shape of the molecule. The analysis of ~140 structurally diverse MMPs from a part of the AstraZeneca data repository reveals the striking trend that 1,3,4-oxadiazoles were systematically more polar compared to their regioisomers ($\Delta \log D \sim 1$) [74], as shown for one typical pair in Figure 10.2. Several other ADMET trends (metabolic stability, hERG inhibition, and aqueous solubility) were also found to depend on the oxadiazole topology, always favoring the 1,3,4-isomer. The study shows that differences in the profiles both for regioisomers as matched pairs can be rationalized by their intrinsically different charge distribution, which, for example, is reflected in dipole moment differences [74].

We have implemented an in-house version for detection of MMPs with some post-processing enhancements compared to the original algorithm from Hussain and Rea [64]; relevant MMPs are then filtered by substituent size. To keep only relevant pairs, the number of non-hydrogen atoms in the fragment is required to be smaller than in the scaffold of the molecule. In addition, the number of non-hydrogen atoms of the variable fragment is limited. To get only relevant MMPs, the activity difference is also stored in the database and used to reject pairs of similar activity for some applications.

The successful MMP applications will certainly stimulate studies involving multiple and more diverse datasets, while data-mining investigations in larger public

databases were already reported meanwhile. An analysis by Hu and Bajorath revealed MMP transformations, which have the potential to alter the biological activity profile of a molecule based on 754 target proteins considered [76]. This concept was also used to identify chemical transformations, which are likely to introduce activity cliffs across different compound classes and targets [77]. Another study then reported bioisosteric changes that are specific for a protein target family [78].

Hence, the use of MMPs for SAR analysis is a powerful tool in drug design, as it organizes and presents experimental data in a manner, which directly stimulated ideas about next synthetic target molecules.

10.2.7 Exploration of Activity Cliffs for SAR Analysis

The systematic exploration of activity cliffs from various representations of activity landscapes is summarized in an interesting article by Stumpfe and Bajorath [79]. These representations could include either distinct chemical transformations (MMPs) or similarity-based relationships between molecules. In a classical medicinal chemistry approach, activity cliffs are extracted from R-group tables, as shown in Figure 10.3 for 3-oxybenzamides as factor Xa inhibitors [58, 59]. This has to be repeated for each scaffold of interest with a predefined view on informative attachment points at the molecular core. Hence this approach is feasible for smaller datasets with only a few informative attachment points for SAR investigation, typically a narrow series in lead optimization.

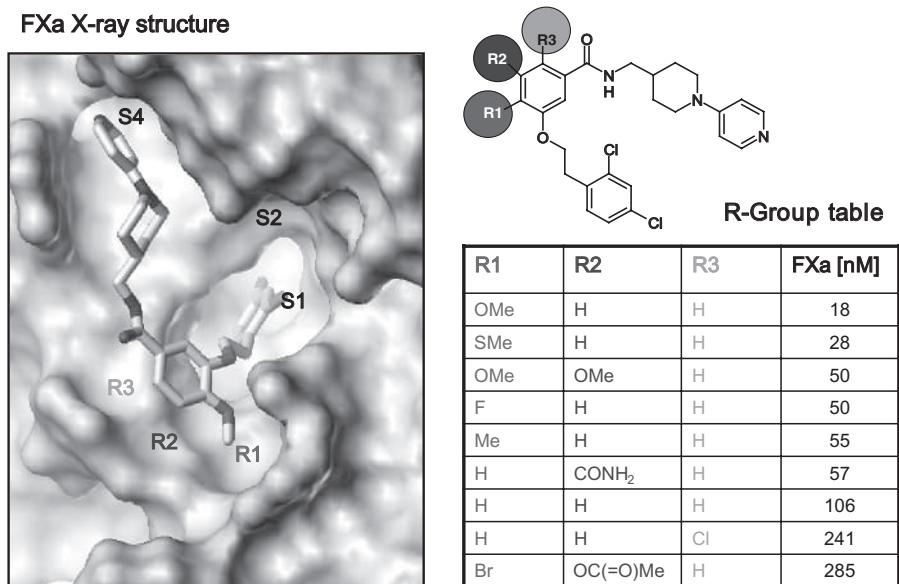


FIGURE 10.3 R-group plot for 3-oxybenzamides as factor Xa inhibitors [58, 59] (right) with binding pose from X-ray crystallography (PDB 2BMG, resolution 2.7 Å). For color details, please see color plate section.

For larger and more diverse datasets, graphical representations of activity cliffs in various activity landscape plots have been proposed. Those approaches require the quantitative description of activity cliffs in a consistent manner, as outlined above with the introduction of the SALI index [51]. Different representations to facilitate the detection of cliffs by visual inspection were then introduced, namely the network-like similarity graph (NSG) [80], 3D-activity landscape models [81], and SALI graphs [51]. For this purpose, we use an internal program named SaliExplorer (Giegerich C, Müller M, Klabunde T. SaliExplorer 2010, personal communication) for sorting and organizing molecular pairs. A heatmap representation can then be interactively browsed to find and save key derivatives as activity cliffs, as shown in Figure 10.4.

For further characterization of activity landscapes, numerical SAR analysis functions were also proposed [47, 82]. Those functions globally characterize continuous and discontinuous SAR regions.

The SAR index (SARI) provides a composite score of individual SAR continuity and discontinuity functions, which have been introduced by Peltason et al. [83, 47]. However, in order to identify activity cliffs, local scoring functions focusing directly on SAR discontinuity in a local environment around a chemical structure of interest are more informative than global activity landscape overviews. Regions of SAR discontinuity provide a lot of information on the SAR of a particular chemical series for deriving a pharmacophore hypothesis. For this case, the compound discontinuity score was introduced [80]. This variation of the SARI formalism is aiming to quantify individual molecular contributions to local SAR discontinuity. Typically, the

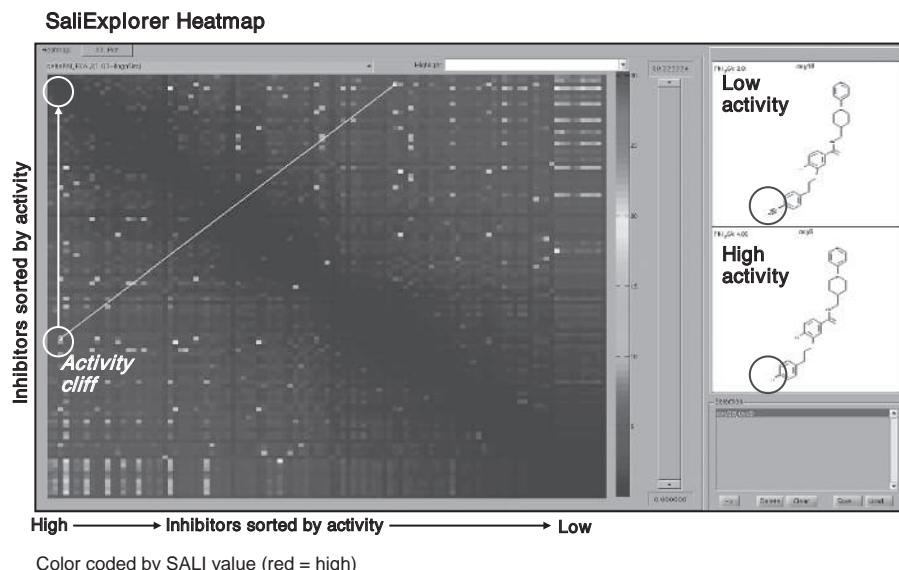


FIGURE 10.4 SaliExplorer heatmap color coded by experimental SALI value (red = high, blue = low) for 107 3-oxybenzamides as factor Xa inhibitors [58, 59]. Inhibitors are sorted by activity on both x- and y-axes; clicking on a point displays the molecular pair. For color details, please see color plate section.

summation of contributions is performed for molecular pairs with a chemical similarity greater than a threshold [79, 80]. Molecules exhibiting a high compound discontinuity score close to the maximal value of 1 indicate activity cliffs. This score thus reflects the potential participation of a molecule in local activity cliffs.

In a recent work, activity cliffs were also investigated using MMPs instead of using 2D-similarity to define relevant compound pairs [5, 84]. Open points relate not only to the molecular descriptors but also to the quality and consistency of biological data for analysis. Those should preferably have high quality and should originate from a single assay, thus reflecting preferably a single mechanism of action.

Recently the activity cliff concept has been extended further. Consensus activity cliffs were introduced by Medina-Franco et al. to highlight consistent activity cliffs independent from the employed chemical descriptors [85]. R-Cliffs were proposed as activity cliff from a graphical representation of R-group tables [86]. Selectivity cliffs [87], mechanism cliffs [88], and multitarget activity cliffs [89, 90] were also described based on the same formalism using different biological data to detect cliffs influencing multiple targets or modes of action for a single target.

An investigation of the distribution of activity cliffs in public datasets like BindingDB [42] and ChEMBL [44] reveals 12% of all bioactive compounds to be involved in cliffs, while only 4% of those were multitarget cliffs [89]. Activity cliffs with different selectivity pattern for multiple targets tend to be rare, which suggests that it might be difficult to modify molecules to display differential selectivity versus multiple targets [79]. The well-known difficulties to arrive at highly selective molecules and the challenge to alter selectivity profiles in protein target families like kinases and GPCRs is providing additional evidence to this analysis.

10.2.8 Visualization to Support SAR Analysis

The organization of data in R-group tables, SAR tables [91], or SAR Maps [92] is a first step to detect essential chemical features and develop initial hypotheses on key structural motifs in a series. The use of predefined or systematically generated substructures or fragments is also frequently combined with activity annotations. Here, the program LeadScope [93] uses a well-designed collection of drug-like fragments for SAR mining and visualization. Those fragments are hierarchically linked and annotated with the activities of compounds containing this substructure. In contrast, Distill [35] applies a systematic fragmentation to produce a complete fragment hierarchy for a dataset, however, containing chemically less meaningful fragments for analysis (see earlier). The SAR Map [86, 92] approach uses heatmaps to visualize distributions of biological properties around a common core structure in a series. The structure–activity report in MOE [91, 94] combines a scaffold hierarchy [30, 33] with SAR table features related to SAR Maps, thus providing a detailed view of the distribution of multiple biological properties among a dataset.

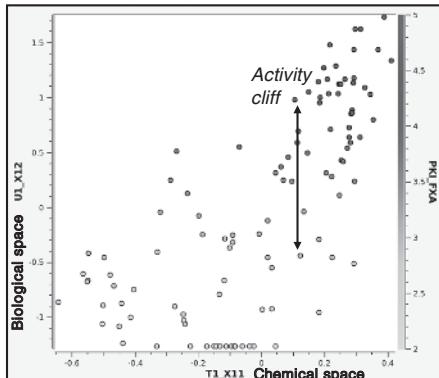
Dimensionality reduction is an approach often employed to condense complex information into only a few representative dimensions (often two or three). Typically the reduced representations are then combined with standard graphical visualization techniques like scatter plots, preferably with a direct link to chemical structure visualization and underlying data in a directly associated table for convenient

interactive analysis and browsing. Different mathematical approaches for data reduction are available; those differ in the degree, how much variance of the original dataset is captured by the reduced representation and whether the distance relationships of objects in the high-dimensional space are preserved in the reduced dimensionality transformations and derived plots.

Principal component analysis (PCA) condenses the larger multidimensional dataset into a few explanatory linear-combinations of the original data; so-called principal properties or PC scores [95–97]. A small number of the PC scores then can explain a significant portion of the original dataset by a linear, statistically interpretable model. PCA is an unsupervised approach, which does not take the biological activity into account. In contrast to that, the supervised PLS approach [52–54] uses biological activity as dependent variable to build a linear model from a few orthogonal linear combinations of descriptors. The first two PLS scores can be used like PC scores to visualize chemical space relationships. In particular the PLS $t-u$ plot is a useful plot, which could also be used as diagnosis tool, as shown in Figure 10.5 on the left. For the first PLS component, the chemical similarity of this projection is usually plotted on the x -axis versus the biological similarity on the y -axis. Points in this scatter plot with similar coordinates for the chemical space and different coordinates for the biological space (i.e., large vertical difference indicated by the arrow) indicate activity cliffs and thus are useful for diagnosis. Typically vertical traces in this plot for the first component indicate series of similar molecules in the chemometrical

PLT t-u plot

Points represent single molecules
Color coded by activity (pK_i)



Neighborhood plot

Points represent pairs of molecules
Color coded by SALI value

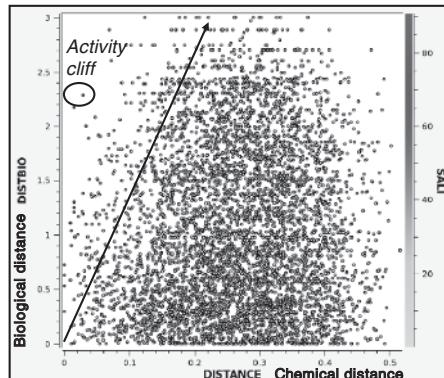


FIGURE 10.5 Left: PLS $t-u$ plot of the first PLS component for the 3-oxybenzamide-based CoMFA model; chemical similarity is plotted on the x -axis versus biological similarity on the y -axis. The vertical arrow indicates an activity cliff with similar coordinates for the chemical space and different coordinates for the biological space. Right: Neighborhood plot for all $n \times (n - 1)/2$ pairs of 107 3-oxybenzamides as factor Xa inhibitors [58, 59] color coded by SALI values (red = high); chemical similarity on the x -axis with most similar compounds on the left is correlated to biological differences on the y -axis. Compounds with highest SALI values are located in the upper left triangle with respect to the arrow defining the neighborhood radius. For color details, please see color plate section.

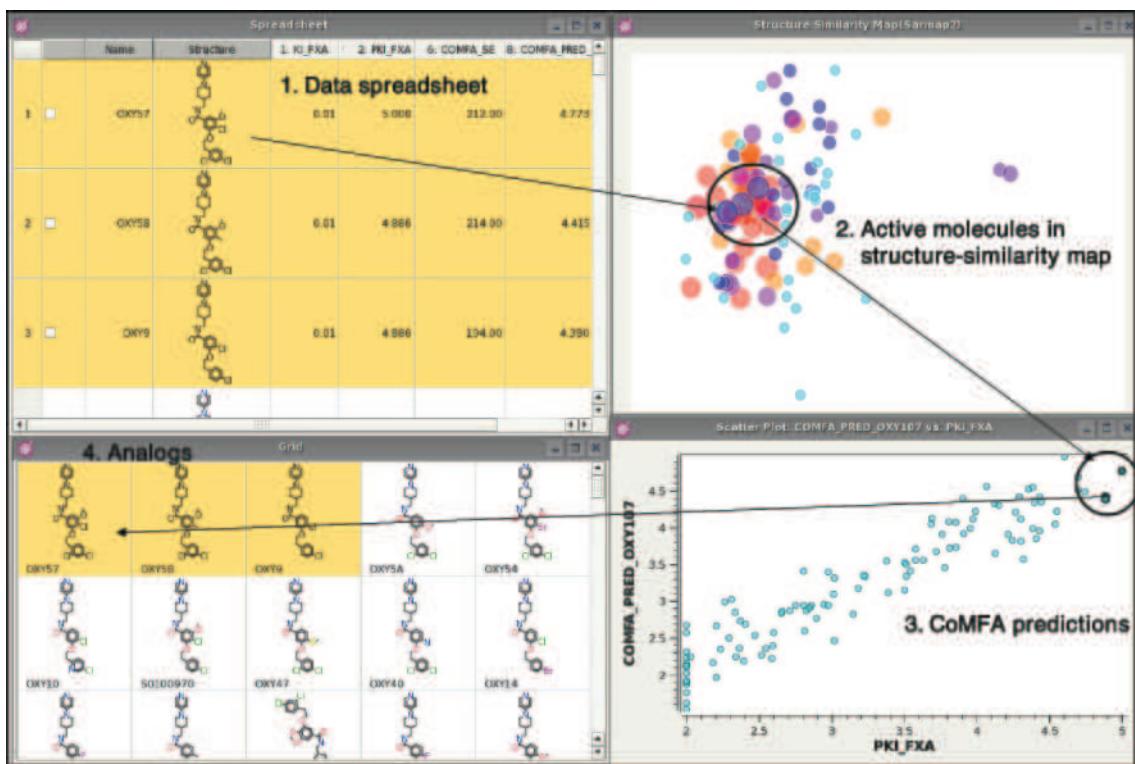


FIGURE 10.6 Interactive exploration of structure-similarity map for factor Xa inhibitors [58, 59]. From the data table (upper left) a structure-similarity map colored and sized by activity (large and red = high pKi) is derived. Interesting regions can be interactively explored and highlighted in a scatter plot with experimental versus predicted activities for a CoMFA model of 107 3-oxybenzamides (lower right), which also allows to display analogs in a 2D-grid plot (lower left). For color details, please see color plate section.

space. Additional PLS components then provide corrections to activity cliffs detected in the first PLS component, if such a correction is possible without overfitting, as it will be detected by other statistical methods.

Multidimensional scaling [98], nonlinear mapping [99], and other nonlinear dimensionality reduction approaches [100] intend to maintain pairwise distance or similarity relationships from the original high-dimensional space. This sometimes can better preserve close local relationships in a dataset, while detailed relationships between more dissimilar compounds are not maintained.

Some available software tools therefore generate such maps, like “structure similarity maps” in the Molecular Data Explorer [101]. In this 2D-plot, the similarity map is an optimal projection of the N -squared similarities between all molecules. Points around the edge of the map represent compounds without nearest neighbors in the dataset above a certain cutoff; those are excluded from computation and displayed on the border of the structure similarity map. To generate the map, a PCA is first conducted using 2D-fingerprints. The first two principal components then serve as initial coordinates for the map. A subsequent nonlinear mapping algorithm minimizes the overall fractional error and preserves the actual distances in many dimensions when plotting in fewer dimensions.

One example is given in Figure 10.6 for the previous series of factor Xa inhibitors [58, 59]; here the original data table (upper left) is used to derive a structure-similarity map colored and sized by activity (large and red = high pKi). Activity cliffs are characterized by large red next to small cyan points. Interesting regions can be interactively explored and propagated to a scatter plot with prediction from the example CoMFA model (lower right, see the earlier example), which also allows displaying analogs in a 2D-grid plot (lower left).

Kohonen networks or self-organizing maps (SOMs) are obtained by a complex, highly nonlinear mathematical approach for dimensionality reduction [102]. After training those networks produce a 2D-map with regions containing similar molecules. As a result of the complex mathematical formalism, models produced by nonlinear projection approaches are often more accurate maintaining the local environment of a molecule, while straightforward interpretation is less obvious.

An advanced modeling approach to SAR landscape visualization toward interpretability was developed by Reutlinger et al. [103]. Their *LiSARD* approach generates interactive graphics for compound design with a focus on the conservation of the local environment around a compound. Stochastic neighborhood embedding (SNE) [103] was found to perform best with respect to preserving the local neighborhood from the high-dimensional space. 2D-projections of the chemical space plus a third dimension with the biological space were used. Furthermore, a continuous color surface is derived based on all data points; local activities are computed from experimental data by Gaussian kernel regression with adaptive bandwidths. This approach resulted in a highly interactive, continuous SAR landscape representation [103].

The integration of structural similarity and activity within a single visualization approach provides a consistent framework for the elucidation of SAR trends, continuity regions, and activity cliffs in a dataset. Again dimensionality reduction approaches are essential for mapping the chemical space onto a 2D-plot. The first

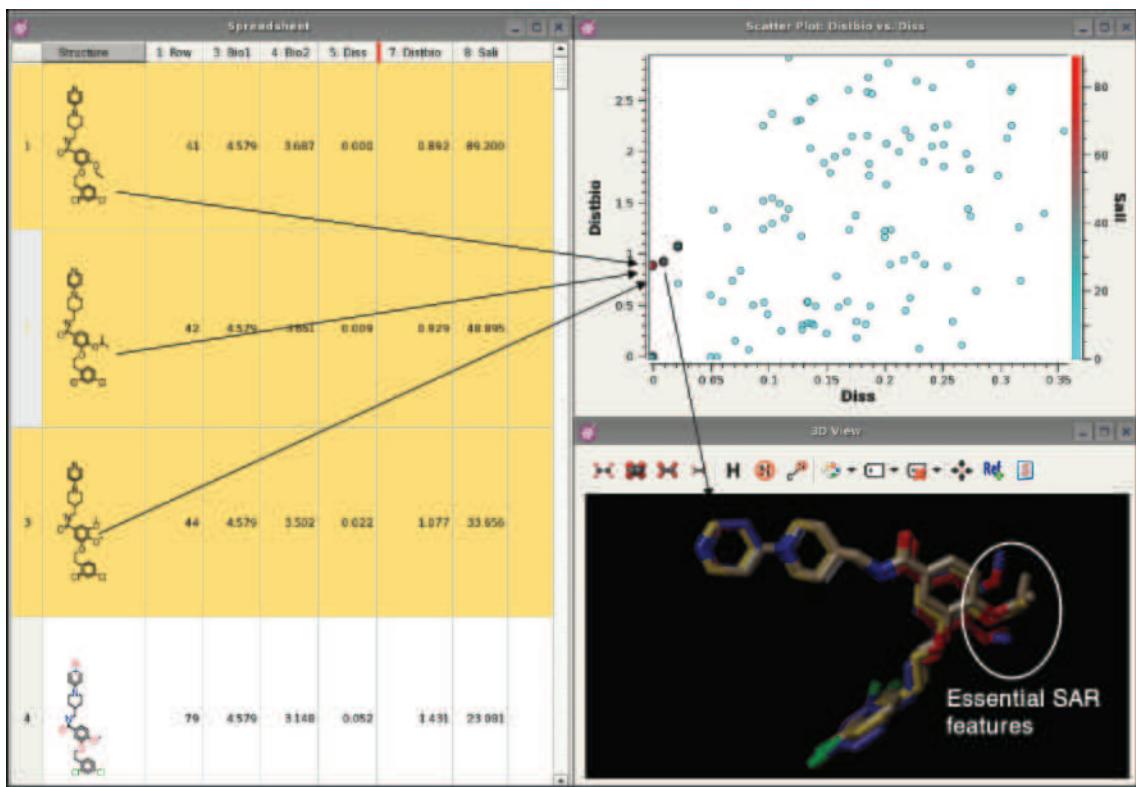


FIGURE 10.7 Local neighborhood plot (upper right) for 3-oxybenzamides as factor Xa inhibitors [58, 59] and the compound in table row 1 (left) as query. Pairs related to the query are displayed with chemical similarity given on the x-axis versus biological difference on the y-axis (upper right), color coded by SALI values (high = red). Interactive exploration by a 2D-table view (left) and a 3D-window (lower right) allows focusing on essential SAR features for activity. For color details, please see color plate section.

method depends directly on the SALI index for the detection of activity cliffs (see above). SALI networks can be constructed such that each node represents a molecule and connections are drawn, once the SALI value for this particular molecular pair is larger than a threshold value to focus on key derivatives useful to understand a particular SAR [51]. Therefore, the connected pairs typically represent essential, but rare aspects of the SAR. Such a SALI map does not reveal directly any SAR trends, but supports the interactive SAR elucidation process. Each informative pair then forms a hypothesis, which could be supported by the existence of related activity cliffs in the same dataset.

A particularly useful way toward a graphical representation of chemical similarity versus biological similarity was introduced using the so-called “structure–activity similarity” (SAS) maps [104]. These maps provide a graphical and numerical tool for the analysis of SAR; on their *x*-axis, the chemical similarity of compound pairs is plotted versus the biological similarity on the *y*-axis. Based on the pairwise comparison of chemical structure and biological activity in a compound series, regions with different SAR characteristics can be identified.

This SAS map is conceptually similar to neighborhood plots introduced earlier for the validation of different descriptors [55]. In their pioneering work to unveil the relationship between chemical and biological similarities, Patterson et al. investigated multiple datasets based on congeneric series with activities for multiple targets [55]. In a neighborhood plot, as displayed in Figure 10.5 on the right, the chemical similarity is plotted on the *x*-axis with most similar compounds shown on the left versus biological difference on the *y*-axis. This graphical representation was first employed to estimate a local radius of similarity for different molecular descriptors (neighborhood radius) [15, 55]. Compound pairs in the upper-left triangle of this plot represent activity cliffs.

Figure 10.7 shows a local neighborhood plot for 107 3-oxybenzamides as factor Xa inhibitors [58, 59]. All compound pairs related to a query molecule located at the coordinate origin [0;0] are displayed with chemical similarity on the *x*-axis versus biological activity difference on the *y*-axis (upper right), color-coded by experimental SALI values (high = red). Interactive exploration by combining a 2D-table view (left) and a 3D-window allows focusing on essential SAR features for activity (lower left, white circled area). In this case, our analysis is indicating SAR elements interacting with a specific pocket in factor Xa and thereby modulating activity.

This environment of query compounds can also be explored using the “spiral view” method [105], which organizes derivatives based on their similarity to the reference molecule, while connections indicate differences in activity. Chemical neighborhood graphs [106] also visualize the similarity and activity distribution around a reference molecule and share many features with those. In addition, LASSO graphs have been proposed for data mining based on a canonical organization scheme employing a molecule—scaffold—skeleton hierarchy in analogy to hierarchical scaffold trees discussed above [107].

NSGs [80] are intended to combine those pairwise molecular relationships for exploring local chemical environments with a global view on the entire dataset

[8, 108]. Nodes in this, NSGs are color coded by activity and scaled in size by compound discontinuity score to reveal activity cliffs.

Based on their implementation of NSGs, the Bajorath group has also proposed different data structures for mining the information present in these rich graphical representations. SAR pathways visualize the effect of small structural changes on activity as sequences of pairwise similar compounds [109]. Different pathways that have particular compounds in common at their start or end point are summarized as SAR trees. Some of these interesting graphical tools are integrated in the publically available program Saranea for interactive exploration of NSGs [110]. These SAR network–based approaches are typically employing 2D-fingerprint-based similarity calculations, while extensions are based on MMPs to define pairs of related compounds [111].

10.2.9 Solutions in Pharmaceutical Industry

Tools for integrated data access, analysis and decision support are essential to rapidly retrieve and analyze large internal datasets. The pharmaceutical industry has significantly invested in testing large compound libraries; retrieving those data and making optimal use of it is therefore directly maximizing these investments. Commercial tools like D360 [112] or internal solutions in different companies like Integrated Project Views (IPV) [113], VlaaiVis [114], and others greatly help in retrieving and managing those data.

However, data quality, harmonization and standardization within a company are also major concerns to obtain reliable datasets for analysis and interpretation. Specifically for cell-based assays and phenotypic assays, data variability can be high and multiple underlying mechanisms might add a significant level of noise and uncertainty to the data. This problem is even larger, when relying on data analysis from public sources only, as here it is rarely the case that different assays or read-outs for a particular target are standardized. The impact of experimental uncertainties on *in silico* model building has been analyzed by Kramer et al. [115]. Although beyond the scope of this review, data management, access, and integration are the most critical parts to settle before any SAR interpretation and comparison of datasets can be started.

SAR analysis is often individually performed for single compound series. In those cases, R-group analysis (Figure 10.3) and MMPs as extension of this concept are very useful, preferably coupled in an interactive manner with informative plots. Project-specific local QSAR models often support data analysis within such a series and might also detect outliers (Figure 10.6).

For larger and more heterogeneous datasets, those requirements for SAR analysis typically change, as medicinal and computational chemists cannot longer manually compare and analyze relevant compound pairs. Often also a quick overview on a past or transferred project has to be provided. This kind of retrospective SAR analysis coupled with clear criteria of compound attractiveness [2, 116, 117] is also recommended to detect further starting points in larger datasets, once the project teams needs to change the current subseries for any reason.

The presence of activity cliffs in the dataset then prompt for further compound synthesis in this area with further variations around a lead structure to sample this region more densely [47]. Exploration in the hit-to-lead phase and early lead optimization are seeking to find those regions, where significant activity improvements are likely to occur [118]. Once a potent molecule has been found, the detection of attachment points allowing to fine-tune other molecular properties without significantly influencing activity for the primary target are then clearly more desirable.

10.3 SAR TRANSFER IN RESCAFFOLDING

10.3.1 Concepts of Rescaffolding

During lead optimization, particular liabilities of a compound series are sometimes evolving in parallel to the SAR for the primary target. If these liabilities appear to be linked to the chemical scaffold, this scaffold needs to be changed (rescaffolding), while attempts are made to retain beneficial decoration from the original series (SAR transfer). The following section will describe the first step on this approach, namely approaches to replace the scaffold. Depending on the context, this process has also been cited in the literature as “scaffold hopping” [119, 120] and “lead hopping” [121, 122]. The subsequent chapter then will integrate SAR analysis and rescaffolding toward SAR transfer.

All rescaffolding approaches are built on two fundamental concepts: The first one is the similarity principle, the other one is the concept of “bioisosterism” [123–125], which was formally introduced by Friedmann [126], based on work by Langmuir [127] and Erlenmeyer [128]. Here, those compounds are regarded as bioisosteres that show some structural differences but still share similar biological activity. Later this concept was employed for the replacement of functional groups with the aim to improve ADMET properties of compounds [129, 130].

In fact it is the combination of chemical similarity and bioisosterism—the appropriate balance between similarity and dissimilarity—that enables rescaffolding to work. All rescaffolding approaches are based on different ways for encoding chemical structures and describing their similarities [131, 132]. The challenge for chemoinformatics-based 2D-rescaffolding approaches is to provide a suitable way of encoding chemical structures for the right balance between similarity for molecular recognition and SAR description on the one hand and dissimilarity to identify novel chemical matter on the other. Molecular similarity considerations cannot provide a well-defined analysis threshold, but three arbitrarily defined ranges could be defined instead: First, there is a high similarity zone, where close analogs in the same chemical series are located. On the opposite, there is the dissimilarity zone, where compounds are truly unrelated to the query. However, most important for scaffold hopping is the zone of medium similarity. Here, compounds share some features necessary for molecular recognition, yet they are dissimilar enough to constitute novel, nonobvious chemical matter. Chemoinformatics approaches for rescaffolding need to be most efficient in this medium similarity zone, where typical 2D-fingerprints or substructure keys fail.

10.3.2 2D-Based Approaches Beyond 2D-Fingerprints

2D-methods are classified as such because the comparison of molecules is done based on their 2D-molecular graph. 3D-aspects of the molecular geometry are ignored, which makes them computationally very fast and suitable for virtual screening of very large compound databases. But 2D-methods have some more advantages over 3D-methods. First, the conformations of the ligand are not considered. This means that the conformational ligand ensemble is treated implicitly rather than explicitly as for 3D-methods. However, in order to be effective for rescaffolding, a reweighting of the molecular description is necessary, such that the specific topology of the individual scaffold is somewhat hidden and the essential features are exposed.

One way of blurring the topology of a molecule is the chemically advanced template search (CATS) introduced by Schneider et al. [133]. This method belongs to the class of topological atom–pair descriptors. The CATS descriptor is generated by combining a distance matrix based on pairs of atoms and their mutual distance in terms of number of bonds and a pharmacophore matrix that is generated from assigning simple pharmacophoric types to atoms into a correlation vector. This correlation vector represents the frequency of the different pairwise combinations of pharmacophore types at different distances. A first prospective validation study [133] was performed starting from the T-type Ca^{2+} channel blocker mibepradil, where 9 out of 12 compounds tested showed significant activity ($\text{IC}_{50} < 10 \mu\text{M}$). All of them showed a different molecular scaffold, while retaining the essential pharmacophoric functionality. In the meantime, some further applications and extensions of the method have been published [134]. Also, we have assessed the suitability of the CATS descriptor for scaffold hopping in the context of fragment-based *de novo* design [135].

With the concept of *feature trees*, a different approach was developed by Rarey et al. [136]. A feature tree is a graph-based description of molecules, where the structure of the molecule is converted into a tree of simplified topology. Rings and small molecular fragments are converted into a single graph node that is characterized by certain attributes such as hydrophobicity or hydrogen bond donor/acceptor properties. The pairwise comparison of different molecules is done on the basis of the split-search and match-search algorithms. The “multiple feature tree model” (MTree) method [137], which is an extension of the feature tree method to include multiple active molecules simultaneously in one search, was used in a retrospective study to identify novel scaffolds for ACE and α_1 _A receptor datasets.

The “topomer searching” method, developed by Cramer et al. [138], was one of the first methods dedicated to scaffold hopping. The method is characterized by the splitting of the 3D-structures of molecules into fragments along rotatable single bonds followed by the rule-based conversion of each fragment into a “topomeric” conformation. Based on the topology of the molecular graph, the dihedral angles of each rotatable bond in a fragment are adjusted following predefined rules such that preferably a most extended conformation is obtained. This topomeric conformation does not represent an individual energetic minimum on a force-field-based hypersurface, but rather a representative sample of the entire conformational ensemble of the substructure, as it is generated in a consistent manner for all fragments. As such, the

topomer approach is not a classical 3D-method, as it relies mainly on the molecular topology and therefore includes an implicit treatment of the conformational ensemble. The fragments are described by CoMFA steric fields [57, 138] and differences of the fragments can be calculated as Euclidian distance of their fields using the split bond vector as alignment rule. This works well, if the split bonds are located at a central position within the molecule. In describing the fragments by steric fields the topomer approach was initially based on molecular shape alone a later extended to account for simple pharmacophoric properties.

The usefulness of topomer searching and the topomer distance measure for rescaffolding has been demonstrated in several prospective studies reported in the literature [122, 139]. Furthermore, initial results on Topomer CoMFA [140] seem very promising for SAR-transfer.

10.3.3 3D-Ligand-Based Approaches

3D methods can provide molecular representations that are independent of the underlying molecular topology. However, in order to compare molecules in 3D-space, two fundamental problems need to be addressed: A suitable alignment rule has to be defined to identify the relevant, that is, bioactive conformation. This requires full accounting for the conformational degrees of freedom of the molecules. Therefore, all methods discussed in this section perform a conformational analysis in the beginning.

Three-point pharmacophore fingerprints (*pharmacophoric triplets*) [141, 142] were implemented to provide a fast way for calculating 3D-similarities between molecules to assess the diversity of compound libraries. Here, the alignment problem is solved by encoding the 3D-information of the molecules in a large fingerprint, where each bit represents the presence or absence of a specific combination of three pharmacophoric features in specific mutual distances. Six groups of pharmacophoric features are derived from a given molecular conformation: hydrogen-bond donor and acceptor, acidic and basic centers, as well as hydrophobic centers and aromatic rings. Accounting for the conformational flexibility is done by generating a representative selection of conformations. However, it turned out that the number of conformations should not exceed 25–50 as an increased number of conformations add noise to the descriptor [143].

3D-ligand-based methods that create *pharmacophore models* capture the SAR by identifying common pharmacophoric features within a set of active molecules. These models are composed of the input molecules in a joint 3D-alignment that is based on those common features and not on 2D-topology. Hence, this approach enables the possibility to find compounds that share the same features, but are based on a different bioisosteric scaffold. This approach is being widely used in both academy and industry and has been extensively reviewed [144, 145].

A further class of ligand-based 3D-methods focuses on the comparison of the shape of molecules. This is done either on the basis of a solvent accessible surface or the shape is approximated by atom-centered soft Gaussian functions. Maximizing the overlay of these Gaussian functions maximizes the overlap between a query molecule and a single conformation of the target molecule. This is used in the Rapid

Overlay of Chemical Structures (ROCS) program, which has become very popular in recent years [146, 147]. It also includes pharmacophore features by assigning a “color force field” to the atoms based on the work of Mills and Dean [148]. A number of successful applications have been published [149]. For instance Boström et al. reported a new series of CB1 receptor antagonists based on replacing the methylpyrazole scaffold in Rimonabant [150]. *EON* [151] is an extension of ROCS in that way that it determines the similarity of molecules not just on the basis of shape, but takes the electrostatic potential of a molecule into account as well.

Recently, a successful rescaffolding application of ROCS has led in our company to the identification of novel inhibitors of DGAT-1, starting from a series of 2-aminothiadiazole series represented by compound **1a** (Figure 10.8) [152]. Unfortunately, the benzoylaminocacid scaffold in this structure suffers from low cellular permeability. Different linker proposals were prioritized using in-house *in silico* ADMET models and ROCS-based alignment using compound **1a** as template.

Two aligned examples are shown in Figure 10.8. Both **1b** and **1c** are permeable, but only compound **1c** is highly active [152]. The difference between both can be explained by comparing alignments with **1a**. Between **1a** and **1c** the thiadiazole scaffolds overlap significantly and form a negative patch alongside the thiadiazole nitrogen atoms indicating an essential feature for recognition. This is not the case for

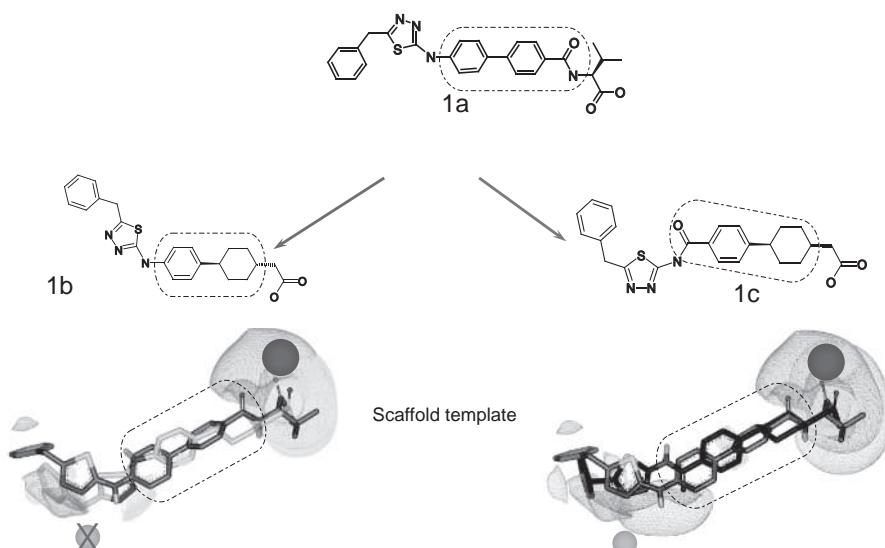


FIGURE 10.8 Rescaffolding of a thiadiazole series of DGAT1 inhibitors using shape-based alignment. Scaffold template elements are illustrated by a dash-dotted line. Replacement of the biphenylbenzoylamide scaffold **1a** by benzoylcyclohexyl **1c** leads to a bioactive compound with full pharmacophoric match, including both the negative (blue) and the acceptor features (green). The replacement by phenylcyclohexyl **1b** leads to a non-bioactive compound with only partial pharmacophoric match of the thiadiazole nitrogen acceptor (green). For color details, please see color plate section.

1b. Here, a poor overlap of the thiadiazole scaffolds reveals significant differences in the distance between the aniline hydrogen and the carboxylate carbon atom, as well as the negative patches of the scaffold extending in different areas in space. All alignments were performed using our in-house alignment program MARS [5, 153] based on ROCS. The MARS procedure starts from a given conformer and systematically scores each conformer of new compounds by aligning them to the optimally matched conformer. This results into a matrix of similarity scores between each conformer to the others. The scores matrix is analyzed to optimize alignment sets to result in a superposition with one conformer per compound. The best ranked alignment—as determined by the maximum sum of scores—is being reported.

10.3.4 3D-Protein-Based Approaches

The highest degree of abstraction from the topology of the starting scaffold is found in 3D-protein-based approaches. In this case, requirements for binding are deduced directly from the protein-binding site instead from a set of active ligands as in the case of ligand-based pharmacophore approaches. Protein crystal structures or validated homology models are required; those are often available in an industrial setup.

The Breed approach that was developed by Pierce et al. for systematic structure-based design [154] capitalizes from the availability of multiple X-ray crystal structures. This method starts from different X-ray complexes of known ligands that are aligned based on the binding site. Then a systematic analysis is performed to identify single bonds from different ligands that overlap in space. New structures are generated by using these single bonds to split the ligands to exhaustively swap scaffolds and side chains between them. After systematically splitting aligned molecules into fragments, those are systematically recombined to allow for an efficient SAR-transfer and rescaffolding.

10.4 ADDRESSING ANTITARGET ACTIVITY

10.4.1 SAR Transfer in Lead Optimization

Transferring SAR between chemotypes is a challenging task in medicinal chemistry, which could be prompted by the necessity to replace the central core due to undesirable ADMET properties or patentability issues. In a previous chapter, we have outlined different approaches to analyze SAR in a chemical series. Furthermore, we have introduced rescaffolding approaches to replace the central scaffold by a topologically equivalent solution. Ideally, a project team then seeks to transfer key SAR elements from the old to the new series in order to rapidly improve activity and other properties.

Often the computational approach for rescaffolding is also applied to identify attractive substitution points for SAR transfer in a prospective design setting. If a 3D-pharmacophore or shape analogy model led to identifying any new chemical series, the 3D-alignment suggests topological regions for adding key substituents beneficial for activity. 3D-alignments based on protein–ligand X-ray structures and

sometimes also supported by docking poses allow to visually identifying those essential regions for activity, additionally validating them using protein-binding site requirements and applying them for SAR transfer also in a prospective manner. This concept has been integrated in automated design workflows and SAR transfer programs like Breed (see earlier) [154].

Often the knowledge of existing transfer series for a particular biological target in larger corporate and public databases also provides information on previously possibly unexplored communalities with the intention to further extend those analogies and transfer more information modulating activity from one to the other series. A prerequisite for assessing the SAR transfer potential is to identify parallel series with pairwise analogs differing only in the chemical scaffold. Recently, a computational approach based on a molecular-network representation was applied to mine exhaustively for SAR transfer series in larger datasets [155]. First, parallel series differing only by the core were identified in a network-like graphical structure, where relationships between compounds and series were expressed using MMPs. The SAR transfer potential of these parallel series was then explored using a scoring system based on activity distribution in corresponding pairs from both series. This approach allows detecting existing SAR transfer steps in historic data, which might enable a project to fully explore the corresponding chemical series and transfer more SAR elements to the new scaffold or alternatively to design novel scaffolds resembling key features of existing scaffolds and add critical SAR elements for decoration.

10.4.2 Identification and Application of Antitarget Activity Hotspots

In a recent study, we have demonstrated the use of antitarget activity cliffs extracted from larger public or internal databases for the optimization of undesirable antitarget activity in chemical series [5]. Antitargets are receptors, ion channels, or enzymes modulating unwanted pharmacological effects of a drug. The optimization of promising lead series often requires addressing those undesirable side activities. To this end, we have developed an approach to extract antitarget activity hotspots with reported activities (e.g., IC₅₀ values) from SAR databases and to transfer this SAR knowledge as informative pairs of molecules carrying characteristic structural modifications onto novel chemical series in a prospective manner resulting in design proposals for medicinal chemistry.

These antitarget activity hotspots are captured as pairs of activity cliffs, which are chemically similar with different biological activity for this antitarget. Collecting activity cliffs of highly relevant antitargets represents valuable information for the optimization of side effects in pharmaceutical research, namely for human ether-à-go-go-related gene (hERG) and CYP3A4 inhibition. We have identified activity cliffs for antitargets from multiple datasets and apply this knowledge to a novel optimization problem.

In order to guide compound optimization, a link between the antitarget activity cliff and the new scaffold is established by 3D-shape comparison [156]. The entire workflow (Figure 10.9) serves primarily to generate ideas in compound optimization, if an antitarget liability has been experimentally observed.

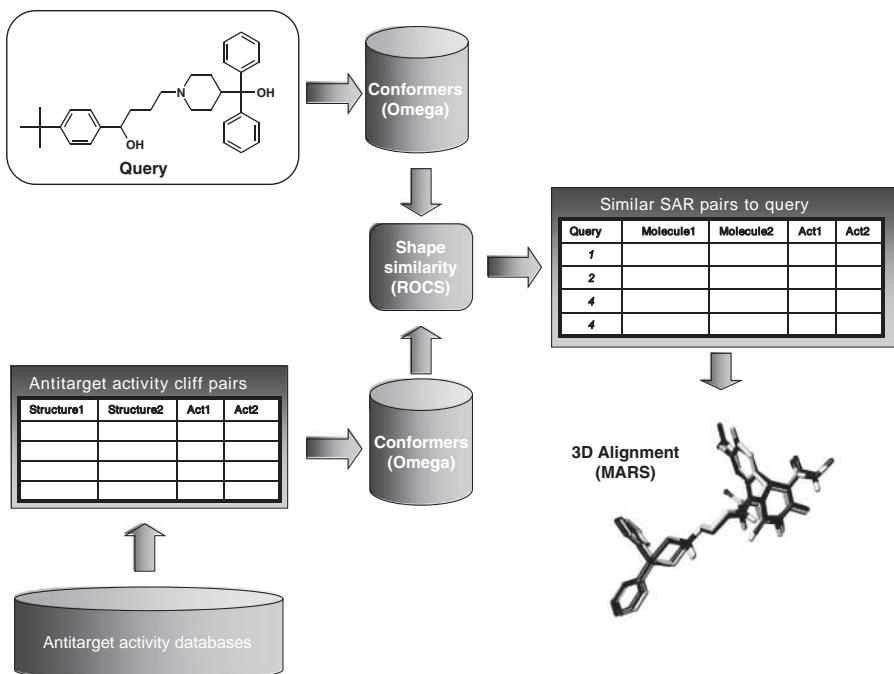


FIGURE 10.9 Workflow to identify antitarget activity hotspots and transfer those between chemical series for optimizing antitarget related issues. A query molecule is compared to a database of informative antitarget activity cliffs using 3D-shape similarity. Resulting hits are aligned in 3D, which allows transferring the essential chemical modification to lower antitarget activity from one to another series.

As preparation step, all structures with biological data for a particular antitarget including assay information are retrieved from public and internal databases like our corporate database or the AurSCOPE database [157]. The latter contained 12,556 entries related to the modulation of the hERG channel in 2010 as an example. Considering only *in vitro* data based on human cell lines and limiting the assay categories to binding, electrophysiology, and ion-flux assays, a total of 4393 SAR data points were extracted.

Subsequently, relevant activity cliffs are identified using two complementary approaches described above, namely high SALI values from pairwise molecular similarity comparisons [51], or MMPs to unveil single structural changes for relevant antitarget activity cliffs. The MMPs are filtered by the relative number of atoms in comparison to their MCS to focus on small, informative transformations. Using the MMP algorithm described above and additional constraints, we obtained 2755 informative SAR pairs for hERG from the Aureus database with the MMPs algorithm. For all molecules in the final list of activity cliffs, we generate conformers using Omega [158].

The conformations for each query molecule are then matched to all SAR pairs by ROCS [159–161] to identify those pairs, which match by shape and chemical functionality to our query. This following pairwise alignment procedure is the essential part to reduce undesirable antitarget activity; our implementation integrates Omega and ROCS in a program named PARI (pairwise alignment for reduction of antitarget inhibition) [5]. In our experience, this ROCS approach is particularly useful for scaffold-hopping.

As hits, we retrieve the most similar antitarget activity cliffs to our query; the query and the activity cliff pair are then subjected to a high-quality alignment using our internal program MARS [5, 153] based on a combinatorial implementation of ROCS 3D-alignments. After visual inspection of the alignment, a proposal for transferring substituents between series on the basis of 3D-similarity and a 2D-activity cliff is obtained, thus providing a clear rational to incorporate a functional group known to influence antitarget activity at a reasonable position. Due to the high similarity between those pairs, it is reasonable to assume a consistent 3D-alignment, which highlights the position of an antitarget activity hotspot to be transferred to a new series.

10.4.3 Application Examples to Address hERG and CYP3A4 Inhibition

This approach was applied to relevant antitargets, namely hERG and CYP3A4 using query molecules with known antitarget side activity. The identified molecular pairs are not necessarily similar to our query in 2D. The corresponding 3D-alignment with the key region for the SAR transfer step is indicated in Figure 10.10. This superposition suggests variations for the query to lower antitarget activity. For the following cases, we have examined, whether literature reported antitarget activity for the activity cliff, the query and the potential design result with reduced antitarget activity were available and obtained for each 2D-related pair from the same biochemical assay.

One important antitarget is the hERG potassium channel [162], which is associated with withdrawal of some marketed drugs due to prolongation of the QT interval. Several examples for activity hotspots to control binding to hERG have been compiled by Jamieson et al. [163].

In the first example (Figure 10.10a), the query *clozapine 2a* with hERG binding activity of 320 nM [164] is found similar to the N1-phenylindole **2b** (IC_{50} 11 nM) [165]. The corresponding activity cliff for **2b** is the nonbasic phenylindole derivative **2c**, in which the replacement of the basic nitrogen by a methyl group, lowered hERG binding affinity significantly (IC_{50} 26,000 nM). The high-quality 3D-alignment is shown with the query **2a**, suggesting that removing or modulating the basic piperazine nitrogen in **2a** in accordance with literature suggestions [163]. For this design compound **2d** can serve as result, showing that the removal of the basic nitrogen by introducing an N-oxide leads to a significantly reduced hERG IC_{50} value of 133,000 nM.

Inhibition of CYP3A4 is another undesirable mechanism resulting in potential undesirable drug–drug interaction [166]. We retrieved a total of 3150 informative

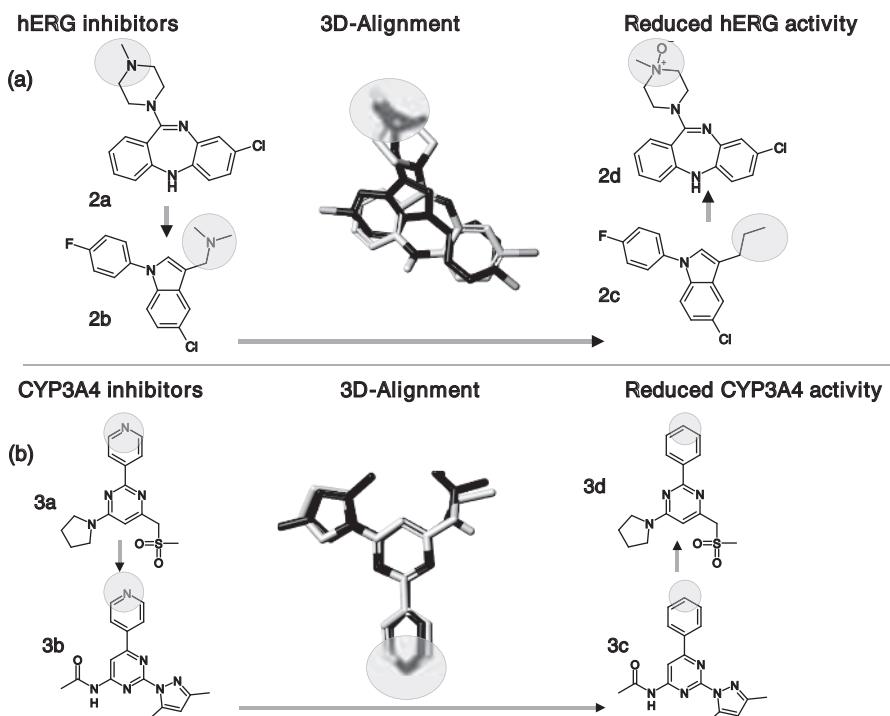


FIGURE 10.10 Application examples to reduce hERG binding activity (a) and CYP3A4 inhibition (b) by transfer of antitarget activity hotspots for diverse query molecules using the transfer of activity hotspot approach. These chemical transformations led to a significant reduction of hERG or CYP3A4 affinity.

SAR pairs from the Aureus database based on MMPs. In the second example (Figure 10.10b), we investigated the pyridyl-pyrimidine **3a** showing a CYP3A4 IC₅₀ value of 1300 nM [167] for the human recombinant enzyme. The 3D-similarity search identified the related pyridine-derivative **3b** (IC₅₀ 8300 nM) [168] with superimposed pyridine rings. The informative SAR pair capturing the antitarget activity cliff is represented by **3c**, where the pyridine is replaced by a phenyl ring (IC₅₀ 43,000 nM). Thus, replacing the pyridine in CYP3A4 inhibitors lowers CYP3A4 inhibition, possibly due to disrupting the potential interaction between the aromatic nitrogen atom and the heme iron atom [167, 169]. In particular, nitrogen-containing heterocyclic aromatic molecules are known to inhibit cytochromes by coordination to the heme iron in the substrate CYP-binding site, which is known as type-II inhibition [170]. Related interactions are present in high-resolution CYP3A4 X-ray structures in the PDB database (3NXU [171], 2VOM [172]). Transferring this design concept back onto the query **3a** suggests the phenyl-pyrimidine derivative **3d**, which is known to be inactive with a reported CYP3A4 IC₅₀ value > 40,000 nM in the same assay as the query **3a** [167].

10.4.4 Integration in Lead Optimization Projects

The optimization of undesirable antitarget activity is a challenging task in lead optimization. Several examples produced using our workflow demonstrate the ability of 3D-similarity searching to identify isosteric scaffolds, for which antitarget activity hotspots are known and to transfer this information back to lower undesirable activities. Any application for prospective design requires a valid structural link between known antitarget activity cliffs and new chemotypes, which is an ideal task for 3D shape-based methods. The entire workflow serves as idea generator in early optimization, if a potential liability has been detected.

For this approach, the quality of the alignment deserves careful visual inspection, before using it toward developing a design concept. One fundamental limitation is the intrinsic assumption that both chemotypes share a common binding mode for a particular antitarget, which cannot be validated in most cases.

The applicability of this workflow to novel chemotypes also depends on size and quality of the underlying database holding activity cliffs. When building these databases for our internal standardized assays, data across a larger collection of chemotypes become directly comparable for use in current projects, while this is more problematic for data from literature and public databases.

As final recommendation, it has to be carefully evaluated whether the SAR for the target and antitarget are closely related for a particular chemotype. In this case, any attempt to introduce a useful functional group extracted from an antitarget activity hotspot will also lower activity for the desired target. Hence, these cases clearly cannot be subjected to any rational optimization strategy. Often only a rescaffolding approach might result in novel chemical matter devoid of undesirable antitarget activity.

10.5 CONCLUSION

The challenge associated with the exploration of SAR still is one of the major bottlenecks during hit-to-lead and lead optimization phases in drug discovery, mainly due to the intrinsic multidimensionality of this problem. Efficient tools for navigating the activity landscape for a single target are essential. Moreover, the ultimate quality criterion for SAR analyses in lead optimization clearly is their impact on decision making in medicinal chemistry for designing next synthetic target molecules.

SAR exploration for a target includes first the detection of informative molecules in relevant areas of the chemical space. For many applications, the detection of activity hotspots or cliffs is important to understand trends by investigating the local environment around a lead. Furthermore, it is useful to highlight regions with flat SAR, indicating that further substitutions are not likely to result in more active molecules. These tasks could be facilitated by SAR visualization approaches, which are primarily supporting the display, correlation and interpretation of small differences in chemical space and considering associated biological effects. In particular, those approaches linking chemical and biological similarity are useful to explore trends.

SAR interpretation does not require any statistical model for quantification. In contrast, those approaches are data driven and allow an intuitive analysis of relevant features for activity, for example, the powerful MMP concept. Interactive use of data, subset selections and visualization techniques is very important to support computational and medicinal chemists in order to explore SAR for a particular target.

In practice, often the combination of qualitative SAR analysis with local or global models capturing SAR trends on a more quantitative basis are effective in the design phase for novel derivatives. If these models are based on local chemical environments coupled with interpretable chemical descriptors and preferentially a linear statistical approach for data modeling, those could potentially have a high impact on decisions taken during lead optimization and progressing a project effectively toward the next milestone.

Often, secondary constraints require changing the chemical series within a project by a process known as “rescaffolding.” The efficient extraction of knowledge and transfer of critical SAR information onto the new chemical series therefore is important to allow for a rapid progression of this chemotype. On the other hand, it might also be possible to identify antitarget activity cliffs from larger corporate or public data collections and transfer those to the series of interest.

In addition, these SAR analysis approaches have also led to methods for extracting antitarget activity hotspots from larger databases and transferring this knowledge onto new chemical series by 3D-similarity. While for most project teams, the SAR for the desired target often becomes clear after a few rounds of chemical synthesis and testing, this might be completely different for many common antitargets in drug discovery projects. Here, the impact of this and related approaches in the context of multidimensional optimization is obvious. Therefore, we have implemented a workflow to support design in early optimization for lowering antitarget activities based on the analysis of local similarities from activity cliff databases. In particular, this transfer of an essential activity cliff to a 3D-related chemotype might be promising, where no global antitarget QSAR model can reproduce the SAR trend. This information from internal and external sources is mined systematically and applied to solve problems for other chemotypes, which might help to avoid typical pitfalls in future projects.

Collectively these approaches for SAR analysis, interpretation and transfer are important tools in drug discovery projects toward an efficient project progression. These tools particularly support the conversion of huge amounts of available data from multiple internal and external sources into knowledge for optimizing primary target activities as well as undesirable side effects.

ACKNOWLEDGMENTS

The authors gratefully thank our colleagues C. Giegerich, T. Klabunde, M. Müller, and L.H. Wang for discussions and software implementation and P. Mougenot, M. Nazaré, C. Philippo, H. Schreuder, V. Wehner, and D.W. Will for many SAR discussions.

REFERENCES

1. Wess G, Urmann M, Sickenberger B. Medicinal chemistry: Challenges and opportunities. *Angew Chem Int Ed Engl* 2001;40:3341–3350.
2. Baringhaus K-H, Matter H. Efficient strategies for lead optimization by simultaneously addressing affinity, selectivity and pharmacokinetic parameters. In: Oprea TI, editor. *Chemoinformatics in Drug Discovery*. Weinheim: Wiley-VCH; 2004. p. 333–379.
3. (a) Gordon EM, Kerwin JF, editors. *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*. New York: Wiley; 1998. (b) Jung G, editor. *Combinatorial Chemistry*. Weinheim: Wiley-VCH; 1999.
4. (a) Gallop MA, Barrett RW, Dower WJ, et al. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J Med Chem* 1994;37:1233–1251. (b) Gordon EM, Barrett RW, Dower WJ, et al. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J Med Chem* 1994;37:1385–1399. (c) Geysen HM, Schoenen F, Wagner D, et al. Combinatorial compound libraries for drug discovery: An ongoing challenge. *Nat Rev Drug Discov* 2003;2:222–230.
5. Hessler G, Matter H, Schmidt F, et al. Identification and application of antitarget activity hotspots to guide compound optimization. *Mol Inf* 2011;30:996–1008.
6. Maggiora GM. On outliers and activity cliffs – Why QSAR often disappoints. *J Chem Inf Model*. 2006;46:1535.
7. Maggiora GM, Johnson MA. *Concepts and Applications of Molecular Similarity*. New York: Wiley; 1990. p 99–117.
8. Wawer M, Lounkine E, Wassermann AM, et al. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov Today* 2010;15:630–639.
9. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci* 1998;38:983–996.
10. Oprea TI, Matter H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* 2004;8:349–358.
11. Willett P, editor. *Similarity and Clustering in Chemical Information Systems*. Letchworth: Research Studies Press; 1997.
12. Hansch C, Leo A, editors. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. Washington, DC: American Chemical Society; 1995.
13. Sharaf MA, Illman DL, Kowalski BR. *Chemometrics*. New York: Wiley; 1986.
14. Zheng W, Tropsha A. A novel variable selection QSAR approach based on the k-nearest neighbor principle. *J Chem Inf Comput Sci* 2000;40:185–194.
15. Matter H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem* 1997;40:1219–1229.
16. Downs, GM. Molecular Descriptors. In: Bultinck P, De Winter H, Langenaeker W, Tollenaere JP, editors. *Computational Medicinal Chemistry for Drug Discovery*. New York: Marcel Dekker, 2004, p. 515–537.

17. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics*. 2 volumes. Weinheim: Wiley-VCH; 2009.
18. Livingstone DJ. Characterising chemical structure using physicochemical descriptors. In: Livingstone DJ, Davis AM, editors. *Drug Design Strategies: Quantitative Approaches*. Cambridge: RSC Publishing; 2012. p 220–241.
19. Benigni R, Passerini L, Pino A, et al. The information content of the eigenvalues from modified adjacency matrices: Large scale and small scale correlations. *Quant Struct Act Relat* 1999;18:449–455.
20. Benigni R, Gallo G, Giorgi F, et al. On the equivalence between different descriptions of molecules: Value for computational approaches. *J Chem Inf Comput Sci* 1999;39:575–578.
21. Martin YC. Development of QSAR. In: Livingstone DJ, Davis AM, editors. *Drug Design Strategies: Quantitative Approaches*. Cambridge: RSC Publishing; 2012. p 60–87.
22. Armitage JE, Lynch MF. Automatic detection of structural similarities among chemical compounds. *J Chem Soc C* 1967;7:521–528.
23. Hariharan R, Janakiraman A, Nilakantan R, et al. MultiMCS: A fast algorithm for the maximum common substructure problem on multiple molecules. *J Chem Inf Model* 2011;51:788–806.
24. Geppert H, Vogt M, Bajorath J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J Chem Inf Model* 2010;50:205–216.
25. Hu Y, Stumpfe D, Bajorath J. Lessons learned from molecular scaffold analysis. *J Chem Inf Model* 2011;51:1742–1753.
26. Lewell XQ, Judd DB, Watson SP, et al. RECAP – retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* 1998;38:511–522.
27. Matter H. Computational approaches towards the quantification of molecular diversity and design of compound libraries. In: Hillisch A, Hilgenfeld R, editors. *Modern Methods of Drug Discovery*. Basel: Birkhäuser; 2003. p 125–156.
28. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996;39:2887–2893.
29. Xu YJ, Johnson M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J Chem Inf Comput Sci* 2001;41:181–185.
30. Schuffenhauer A, Varin T. Rule-based classification of chemical structures by scaffold. *Mol Inform* 2011;30:646–664.
31. Xu Y-J, Johnson M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J Chem Inf Comput Sci* 2002;42:912–926.
32. Wilkens SJ, Janes J, Su AI. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J Med Chem* 2005;48:182–193.
33. Schuffenhauer A, Ertl P, Roggo S, et al. The scaffold tree – Visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* 2007;47:47–58.
34. Koch MA, Schuffenhauer A, Scheck M, et al. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 2005;102:17272–17277.

35. Distill is a module in SybylX2.0, available from Tripos/Certara, St. Louis, MO, USA.
36. Evans BE, Rittle KE, Bock, et al. Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J Med Chem* 1988;31: 2235–2246.
37. Müller G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov Today* 2003;8:681–691.
38. Aronov AM, McClain B, Moody CS, et al. Kinase-likeness and kinase-privileged fragments: Toward virtual polypharmacology. *J Med Chem* 2008;51:1214–1222.
39. Schnur DM, Hermsmeier MA, Tebben AJ. Are target-family substructures truly privileged ? *J Med Chem* 2006;49:2000–2009.
40. Hu Y, Wassermann AM, Lounkine E, et al. Systematic analysis of public domain compound potency data identifies selective molecular scaffolds across druggable target families. *J Med Chem* 2010;53:752–758.
41. Hu Y, Bajorath J. Exploring target-selectivity pattern of molecular scaffolds. *ACS Med Chem Lett* 2010;1:54–58.
42. Liu T, Lin Y, Wen X, et al. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007;35:D198–D201.
43. PubChem. National Center for Biotechnology Information: Bethesda, US, 2012. <http://pubchem.ncbi.nlm.nih.gov/>. Accessed 2013 May 14.
44. ChEMBL. European Bioinformatics Institute (EBI): Cambridge, UK, 2012. <http://www.ebi.ac.uk/chembl/>. Accessed 2013 May 14.
45. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? *J Med Chem* 2002;45:4350–4358.
46. Willett P, Winterman V. Comparison of some measures for the determination of intermolecular structural similarity. *Quant Struct Activ Relat* 1986;5:18–25.
47. Bajorath J, Peltason L, Wawer M, et al. Navigating structure-activity landscapes. *Drug Discov Today* 2009;14:698–705.
48. Kubinyi H. Similarity and dissimilarity – A medicinal chemists view. *Perspect Drug Discov Des* 1998;11:225–252.
49. Kubinyi H. Chemical similarity and biological activities. *J Braz Chem Soc* 2002;13: 717–726.
50. Bissantz C, Kuhn B, Stahl M. A medicinal chemist's guide to molecular interactions. *J Med Chem* 2010;53:5061–5084.
51. Guha R, Van Drie JH. Structure-activity landscape index: Identifying and quantifying activity cliffs. *J Chem Inf Model* 2008;48:646–658.
52. Wold S, Albano C, Dunn WJ, et al. In: Kowalski B, editor. *Chemometrics: Mathematics and Statistics in Chemistry*. Dordrecht: Reidel; 1984. p 17–95.
53. Dunn WJ, Wold S, Edlund U, et al. Multivariate structure-activity relationship between data from a battery of biological tests and an ensemble of structure descriptors: The PLS method. *Quant Struct Act Relat* 1984;3:31–137.
54. Geladi P. Notes on the history and nature of partial least squares (PLS) modelling. *J Chemom* 1988;2:231–246.
55. Patterson DE, Cramer RD, Ferguson AM, et al. Neighborhood behavior: A useful concept for validation of molecular diversity descriptors. *J Med Chem* 1996;39:3049–3059.

56. Johnson S. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J Chem Inf Model* 2008;48:25–26.
57. Cramer RD, III, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 1988; 110:5959–5967.
58. Nazaré M, Matter H, Klingler O, et al. Novel factor Xa inhibitors based on a benzoic acid scaffold and incorporating a neutral P1 ligand. *Bioorg Med Chem Lett* 2004; 14:2801–2805.
59. Matter H, Will DW, Nazaré M, et al. Structural requirements for factor Xa inhibition by 3-oxybenzamides with neutral P1 substituents: Combining X-ray crystallography, 3D-QSAR, and tailored scoring functions. *J Med Chem* 2005;48:3290–3312.
60. Guha R, Van Drie JH. Assessing how well a modelling protocol captures a structure-activity landscape. *J Chem Inf Model* 2008;48:1716–1728.
61. Kenny PW, Sadowski J. Structure modification in chemical databases. *Methods Princ Med Chem* 2005;23:271–285.
62. Leach AG, Jones HD, Cosgrove DA, et al. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem* 2006;49:6672–6682.
63. Raymond JW, Watson IA, Mahoui A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J Chem Inf Model* 2009;9:1952–1962.
64. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model* 2010;50:339–348.
65. Haubertin DY, Bruneau P. A database of historically-observed chemical replacements. *J Chem Inf Model* 2007;47:1294–1302.
66. Papadatos G, Alkarouri M, Gillet VJ, et al. Lead optimization using matched molecular pairs: Inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J Chem Inf Model* 2010;50:1872–1886.
67. Hajduk PJ, Sauer DR. Statistical analysis of the effect of common chemical substituents on ligand potency. *J Med Chem* 2008;51:553–564.
68. Lewis ML, Cucurull-Sanchez L. Structural pairwise comparisons of HLM stability of phenyl derivatives: Introduction of the Pfizer metabolism index (PMI) and metabolism-lipophilicity efficiency (MLE). *J Comput Aided Mol Des* 2009;23:97–103.
69. Gleeson P, Bravi G, Modi S, et al. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg Med Chem*. 2009;17:5906–5919.
70. Birch AM, Kenny PW, Simpson I, et al. Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. *Bioorg Med Chem Lett* 2009;19:850–853.
71. Griffen E, Leach AG, Robb GR, et al. Matched molecular pairs as medicinal chemistry tool. *J Med Chem* 2011;54:7739–7750.
72. Andrews DM, Gibson KM, Graham MA, et al. Design and campaign synthesis of pyridine-based histone-deacetylase inhibitors. *Bioorg Med Chem Lett* 2008;18:2525–2529.
73. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.

74. Boström J, Hogner A, Llinàs A, et al. Oxadiazoles in medicinal chemistry. *J Med Chem* 2012;55:1817–1830.
75. Müller K. The power of MMPA and a teaching lesson in medicinal chemistry. *J Med Chem* 2012;55:1815–1816.
76. Hu Y, Bajorath J. Chemical transformations that yield compounds with distinct activity profiles. *ACS Med Chem Lett*. 2011;2:523–527.
77. Wassermann AM, Bajorath J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J Chem Inf Model* 2010;50:1248–1256.
78. Wassermann AM, Bajorath J. Identification of target family directed bioisosteric replacements. *MedChemComm* 2011;2:601–606.
79. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J Med Chem* 2012;55:2932–2942.
80. Wawer M, Peltason L, Weskamp N, et al. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J Med Chem* 2008;51:6075–6084.
81. Peltason L, Iyer P, Bajorath J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J Chem Inf Model* 2010;50:1021–1033.
82. Peltason L, Bajorath J. Systematic computational analysis of structure-activity relationships: Concepts, challenges and recent advances. *Future Med Chem* 2009;1:451–466.
83. Peltason L, Bajorath J. SAR index: Quantifying the nature of structure-activity relationships. *J Med Chem* 2007;50:5571–5578.
84. Hu X, Hu Y, Vogt M, et al. MMP-cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *J Chem Inf Model* 2012;52:1138–1145.
85. Medina-Franco JL, Martinez-Mayorga K, Bender A, et al. Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *J Chem Inf Model* 2009;49:477–491.
86. Agrafiotis DK, Wiener JJM, Skalkin A, et al. Single R-group polymorphisms (SRPs) and R-cliffs: An intuitive framework for analyzing and visualizing activity cliffs in a single analog series. *J Chem Inf Model* 2011;51:1122–1132.
87. Peltason L, Hu Y, Bajorath J. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* 2009;4:1864–1873.
88. Iyer P, Stumpfe D, Bajorath J. Molecular mechanism-based network-like similarity graphs reveal relationships between different types of receptor ligands and structural changes that determine agonistic, inverse-agonistic, and antagonistic effects. *J Chem Inf Model* 2011;51:1281–1286.
89. Wassermann AM, Dimova D, Bajorath J. Comprehensive analysis of single- and multi-target activity cliffs formed by currently available bioactive compounds. *Chem Biol Drug Des* 2011;78:224–228.
90. Dimova D, Wawer M, Wassermann AM, et al. Design of multi-target activity landscapes that capture hierarchical activity cliff distributions. *J Chem Inf Model* 2011;51:256–288.
91. Clark AM, Labute P. Detection and assignment of common scaffolds in project databases of lead molecules. *J Med Chem* 2009;52:469–483.

92. Kolpak J, Connolly PJ, Lobanov VS, et al. Enhanced SAR maps: Expanding the data rendering capabilities of a popular medicinal chemistry tool. *J Chem Inf Model* 2009;49:2221–2230.
93. Richon A. LeadScope: Data visualization for large volumes of chemical and biological screening data. *J Mol Graph Model* 2000;18, 76–79.
94. MOE (version 2009.10) Available from Chemical Computing Group (CCG), Montreal, Canada.
95. Dillon WR, Goldstein M. *Multivariate Analysis: Methods and Applications*. New York: Wiley; 1984.
96. Cramer III, RD. BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J Am Chem Soc* 1980;102:1837–1849.
97. Wold S, Albano C, Dunn WJ, III, et al. Multivariate data analysis in chemistry. In: Kowalski BR, editor. *Chemometrics: Mathematics and Statistics in Chemistry*, NATO, ISI Series C 138. Dordrecht: D. Reidel Publ. Co.; 1984. p 17–96.
98. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika* 1952; 17:401–419.
99. Domine D, Devillers J, Chastrette M, et al. Non-linear mapping for structure-activity and structure-property modelling. *J Chemom.* 1993;7:227–242.
100. Reutlinger M, Schneider G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J Mol Graphics Model* 2012;34:108–117.
101. Molecular Data Explorer (MDE) is integrated in Sybylx2.0, available from Tripos Inc., St. Louis, MO, USA.
102. (a) Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982;43:59–69. (b) Gasteiger J, Zupan J. *Neural Networks in Chemistry and Drug Design*. Weinheim: Wiley-VCH; 1999. (c) Yan A. Application of self-organizing maps in compounds pattern recognition and combinatorial library design. *Comb Chem High Throughput Screen* 2006;9:473–480. (d) Schneider P, Tanrikulu Y, Schneider G. Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Curr Med Chem* 2009;16:258–266.
103. Reutlinger M, Guba W, Martin RE, et al. Neighborhood-preserving visualization of adaptive structure-activity landscapes: Application to drug discovery. *Angew Chem Int Ed Engl* 2011;50:11633–11636.
104. Shanmugasundaram V, Maggiora GM. Characterizing property and activity landscapes using an information-theoretic approach. 222nd ACS National Meeting, Chicago, IL, 2001, CINF-032.
105. Smellie A. General purpose interactive physico-chemical property exploration. *J Chem Inf Model*. 2007;47:1182–1187.
106. Wawer M, Sun S, Bajorath J. Computational characterization of SAR microenvironments in high-throughput screening data. *Int J High Throughput Screen* 2010;1:15–27.
107. Gupta-Ostermann D, Hu Y, Bajorath J. Introducing the LASSO graph for compound data set representation and structure-activity relationship analysis. *J Med Chem* 2012;55: 5546–5553.
108. Stumpfe D, Bajorath J. Methods for SAR visualization. *RSC Adv* 2012;2:369–378.
109. Wawer M, Bajorath J. Systematic extraction of structure-activity relationship information from biological screening data. *ChemMedChem* 2009;4:1431–1438.

110. Lounkine E, Wawer M, Wassermann AM, et al. SARANE: A freely available program to mine structure-activity and structure-selectivity information in compound data sets. *J Chem Inf Model* 2010;50:68–78.
111. Wawer M, Bajorath J. Local structural changes, global data views: Graphical substructure-activity relationship trailing. *J Med Chem*. 2011;54:2944–2951.
112. D360, available from Tripos/Certara, St. Louis, MO, USA.
113. Baede EJ, den Bekker E, Boiten J-W, et al. Integrated project views: Decision support platform for drug discovery project teams. *J Chem Inf Model* 2012;52:1438–1449.
114. Howe TJ, Mahieu G, Marichal P, et al. Data reduction and representation in drug discovery. *Drug Discov Today* 2007;12:45–53.
115. Kramer C, Kalliokoski T, Gedeck P, et al. The experimental uncertainty of heterogeneous public Ki data. *J Med Chem* 2012;55:5165–5173.
116. Segall M, Beresford A, Gola J, et al. Focus on success: Using in silico optimization to achieve an optimal balance of properties. *Expert Opin Drug Metab Toxicol* 2006;2:325–337.
117. Segall M, Champness E, Obrezanova O, et al. Beyond profiling: Using ADMET models to guide decisions. *Chem Biodivers* 2009;6:2144–2151.
118. Duffy BC, Zhu L, Decornez H, et al. Early phase drug discovery: Cheminformatics and computational techniques in identifying lead series. *Bioorg Med Chem* 2012;20:5324–5342.
119. Böhm H-J, Flohr A, Stahl M. Scaffold hopping. *Drug Discov Today Technol* 2004;1:217–224.
120. Schneider G, Schneider P, Renner S. Scaffold-hopping: How far can you jump? *QSAR Comb Sci* 2006;25:1162–1171.
121. Martin YC, Muchmore S. Beyond QSAR: Lead hopping to different structures. *QSAR Comb Sci* 2009;28:797–801.
122. Cramer RD, Jilek RJ, Güssregen S, et al. “Lead hopping”. Validation of topomer similarity as a superior predictor of similar biological activities. *J Med Chem* 2004;47:6777–6791.
123. Burger A. Isosterism and bioisosterism in drug design. *Prog Drug Res* 1991;37:288–362.
124. Patani GA, LaVoie EJ. Bioisosterism: A rational approach in drug design. *Chem Rev* 1996;96:3147–3176.
125. Meanwell NA. Synopsis of some recent tactical application of bioisosteres in drug design. *J Med Chem* 2011;54:2529–2591.
126. Friedman HL. Influence of isosteric replacements upon biological activity. *Natl Acad Sci* 1951;206:295–358.
127. Langmuir I. Isomorphism, isosterism and covalence. *J Am Chem Soc* 1919;41:1543–1559.
128. (a) Erlenmeyer H, Berger E. Studies on the significance of structure of antigens for the production and the specificity of antibodies. *Biochem Zeitschrift* 1932;252:22–36. (b) Erlenmeyer H, Berger E, Leo M. Beziehungen zwischen der Struktur der Antigene und der Spezifität der Antikörper. *Helv Chim Acta* 1933;16:733–738.
129. Carini DJ, Duncia JV, Aldrich PE, et al. Nonpeptide angiotensin II receptor antagonists: The discovery of a series of *N*-(biphenylmethyl)imidazoles as potent, orally active antihypertensives. *J Med Chem* 1991;34:2525–2547.

130. Allen FH, Groom CR, Liebeschuetz JW, et al. The hydrogen bond environments of 1H-tetrazole and tetrazolate rings: The structural basis for tetrazole–carboxylic acid bioisosterism. *J Chem Inf Model* 2012;52:857–866.
131. Brown N, Jacoby E. On scaffolds and hopping in medicinal chemistry. *Mini Rev Med Chem* 2006;6:1217–1229.
132. Mauser H, Guba W. Recent developments in de novo design and scaffold hopping. *Curr Opin Drug Discov Dev* 2008;11:365–374.
133. Schneider G, Neidhart W, Giller T, et al. “Grundgerüstwechsel” (scaffold-hopping) durch topologische Pharmakophorsuche: ein Beitrag zum virtuellen screening. *Angew Chem* 1999;111:3068–3070.
134. Renner S, Fechner U, Schneider G. Alignment-free pharmacophore patterns – A correlation-vector approach. In: Langer T, Hoffmann RD, editors. *Pharmacophores and Pharmacophore Searches (Methods and Principles in Medicinal Chemistry)*. Weinheim: Wiley-VCH; 2006. p 49–80.
135. Krueger BA, Dietrich A, Baringhaus K-H, et al. Scaffold-hopping potential of fragment-based de novo design: The Chances and Limits of Variation. *Comb. Chem. & High Throughput Screen.* 2009;12, 383–396.
136. Rarey M, Dixon SJ. Feature trees: A new molecular similarity measure based on tree matching. *J Comput Aided Mol Des* 1998;12:471–490.
137. Hessler G, Zimmermann M, Matter H, et al. Multiple-ligand-based virtual screening: Methods and applications of the MTree approach. *J Med Chem* 2005;48:6575–6584.
138. Cramer RD, Clark RD, Patterson DE, et al. Bioisosterism as a molecular diversity descriptor: Steric fields of single “topomeric” conformers. *J Med Chem* 1996;39: 3060–3069.
139. Cramer RD, Poss MA, Hermsmeier MA, et al. Prospective identification of biologically active structures by topomer shape similarity searching. *J Med Chem* 1999;42: 3919–3933.
140. Cramer RD. Topomer CoMFA: A design methodology for rapid lead optimization. *J Med Chem* 2003;46:374–388.
141. Ashton MJ, Jaye MC, Mason JS. New perspectives in lead generation II: Evaluating molecular diversity. *Drug Discov Today* 1996;1:71–78.
142. Pickett SD, Mason JS, McLay, IM. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J Chem Inf Model* 1996;36:1214–1223.
143. Matter H, Pötter T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J Chem Inf Comput Sci* 1999;39:1211–1225.
144. Langer T, Hoffmann RD, editors. *Pharmacophores and pharmacophore searches*. In: *Methods and Principles in Medicinal Chemistry*. Weinheim: Wiley-VCH; 2006.
145. Güner OF, editor. *Pharmacophore Perception, Development, and Use in Drug Design*. IUL Biotechnology Series. La Jolla: International University Line; 2000.
146. Grant JA, Pickup BT. A Gaussian description of molecular shape. *J Phys Chem* 1995;99:3503–3510.
147. Nicholls A, Grant JA. Molecular shape and electrostatics in the encoding of relevant chemical information. *J Comput Aided Mol Des* 2005;19:661–686.
148. Mills JEJ, Dean PM. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J Comput Aided Mol Des* 1996;10:607–622.

149. Rush TS, Grant JA, Mosyak L, et al. A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J Med Chem* 2005;48:1489–1495.
150. Boström J, Berggren K, Elebring T, et al. Scaffold hopping, synthesis and structure–activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: A novel series of CB1 receptor antagonists. *Bioorg Med Chem* 2007;15:4077–4084.
151. Nicholls A, MacCuish NE, MacCuish JD. Variable selection and model validation of 2D and 3D molecular descriptors. *J Comput Aided Mol Des* 2004;18:451–474.
152. Mougenot P, Namane C, Fett E, et al. Thiadiazoles as new inhibitors of diacylglycerol acyltransferase type 1. *Bioorg Med Chem Lett* 2012;22:2497–2502.
153. Klabunde T, Giegerich C, Evers A. MARS: Computing three-dimensional alignments for multiple ligands using pairwise similarities. *J Chem Inf Model* 2012;52:2022–2030.
154. Pierce AC, Rao G, Bemis GW. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J Med Chem* 2004;47:2768–2775.
155. Gupta-Ostermann D, Wawer M, Wassermann AM, et al. Graph mining for SAR transfer series. *J Chem Inf Model* 2012;52:935–942.
156. Nicholls A, McGaughey GB, Sheridan RP, et al. Molecular shape and medicinal chemistry: A perspective. *J Med Chem* 2010;53:3862–3886.
157. AurSCOPE database; Aureus Sciences: Paris, France. <http://www.aureus-pharma.com>. Accessed 2013 May 14.
158. Omega; OpenEye Scientific Software: Santa Fe, NM, USA. <http://www.eyesopen.com>. Accessed 2013 May 14.
159. Rush III, TS, Grant JA, Mosyak L, et al. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 2005, 48:1489–1495.
160. Hawkins PCD, Skillman AG, Nicholls A. comparison of shape-matching and docking as virtual screening tools. *J Med Chem* 2007, 50:74–82.
161. ROCS version 3.1.1.; OpenEye Scientific Software: Santa Fe, NM, USA. <http://www.eyesopen.com>. Accessed 2013 May 14.
162. Sanguinetti MC, Tristani-Firouzi M. hERG potassium channels and cardiac arrhythmia. *Nature* 2006;440:463–469.
163. Jamieson C, Moir EM, Rankovic Z, et al. Medicinal chemistry of hERG optimizations: Highlights and hang-ups. *J Med Chem* 2006;49:5029–5046.
164. Ekins S, Crumb WJ, Sarazan RD, et al. Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J Pharmacol Exp Ther* 2002;301:427–434.
165. Pearlstein RA, Vaz RJ, Kang J, et al. Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg Med Chem Lett* 2003;13:1829–1835.
166. Fleishaker JC, Herman BD, Carel BJ, et al. Interaction between ketoconazole and almotriptan in healthy volunteers. *J Clin Pharmacol* 2003;43:423–427.
167. Ahlström MM, Zamora I. Characterization of type II ligands in CYP2C9 and CYP3A4. *J Med Chem* 2008;51:1755–1763.

168. Zhang X, Tellew JE, Luo Z, et al. Lead optimization of 4-acetylaminoo-2-(3,5-dimethylpyrazol-1-yl)-6-pyridylpyrimidines as A2A adenosine receptor antagonists for the treatment of Parkinson's disease. *J Med Chem* 2008;51:7099–7110.
169. Verras A, Kuntz ID, Ortiz de Montellano PR. Computer-assisted design of selective imidazole inhibitors for cytochrome P450 enzymes. *J Med Chem* 2004;47:3572–3579.
170. Ballard SA, Lodola A, Tarbit MH. A comparative study of 1-substituted imidazole and 1,2,4-triazole antifungal compounds as inhibitors of testosterone hydroxylations catalysed by mouse hepatic microsomal cytochromes P-450. *Biochem Pharmacol* 1988; 37:4643–4651.
171. Sevrioukova IF, Poulos TL. Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir. *Proc Natl Acad Sci USA* 2010;107:18422–18427.
172. Ekroos M, Sjøgren T. Structural basis for ligand promiscuity in cytochrome P 450 3A4. *Proc Natl Acad Sci USA* 2006;103:13682–13687.

CHAPTER 11

DEVELOPMENT AND APPLICATIONS OF GLOBAL ADMET MODELS: IN SILICO PREDICTION OF HUMAN MICROSOMAL LABILITY

KARL-HEINZ BARINGHAUS, GERHARD HESSLER, HANS MATTER,
and FRIEDEMANN SCHMIDT

11.1 INTRODUCTION

Absorption, distribution, metabolism, excretion, and toxicology (ADMET) properties determine the pharmacokinetic and safety behavior of compounds and thus to a large part, how effective a compound acts as a drug, how the compound has to be dosed, and what safety window has to be considered. Due to the decisive role of ADMET properties in lead optimization and drug development, these parameters are investigated very early by *in vitro* systems. For example, Caco-2 cell lines are used to estimate permeability of compounds, while liver microsomes are used to study metabolic stability of novel compounds *in vitro* [1, 2].

ADMET experiments and *in silico* predictions allow for the timely identification of potential liabilities in order to discard compounds with unfavorable ADMET properties early according to the “fail-early-fail-cheaply” paradigm. In the hit and lead finding phase, *in vitro* and/or *in silico* ADMET studies play a major role in the identification of critical parameters that need to be considered to reach the next milestone in discovery. Compounds are usually preferred as hits and in particular as leads, which exhibit the least liabilities to be optimized. A first analysis should indicate clearly that liabilities appear to be optimizable, for example, strong target-related activity of compounds can be kept while structure-ADMET-properties are improving.

In the lead optimization phase, *in vitro* ADMET assays play in particular an important role to monitor, how potential liabilities evolve and to drive their optimization to an overall acceptable compound profile.

In his pioneering work, Lipinski analyzed a set of compounds (marketed or phase III molecules) and derived “the rule-of-5” for molecules exhibiting drug-like behavior [3]. This analysis has stimulated numerous additional studies on related datasets, coming up with very similar property ranges [4–8]. In lead optimization, very often size and lipophilicity of compounds are significantly increased during the optimization process [9], thus it appears to be advisable to begin lead optimization with smaller, more polar compounds. This lead-like property range was termed the rule-of-3 [10].

In any case, rules provide only a meaningful and chemically interpretable framework to establish drug-like chemical space. They are derived from statistical analysis and thus summarize trends. They do not necessarily apply to a particular compound-indication profile. However, such rules give guidance for individual compounds to be optimized.

Compound optimization requires more sophisticated (quantitative) approaches. Large efforts have been undertaken to develop predictive *in silico* models, which support the progress of compound series in hit evaluation and in lead optimization.

Local ADMET models are derived from molecules belonging to one specific chemotype and are only applicable for the prediction of compounds belonging to the same scaffold. They are frequently used in lead optimization and depend on the availability of a balanced set of molecules with data reflecting a single mechanism of action on the target or ADMET property of interest. The number of compounds used to build local models is therefore typically much smaller than for global models.

The development of global predictive *in silico* models, which are not limited to a single series of molecules, requires large ADMET datasets covering the chemical space appropriately. Global *in silico* models are generated using standard computer-aided drug design techniques. Typically, 2D-quantitative structure–activity relationship (QSAR) or machine learning approaches are used for training global models.

This chapter briefly describes structure- and ligand-based ADMET modeling and then focuses in detail on building a global model for metabolic stability of compounds including a recent application in a research project.

11.1.1 Structure-Based ADMET Models

Structure-based approaches appear to be beneficial, when a 3D-structure of the protein responsible for the ADMET liability is available. For numerous proteins, which are involved in toxicological endpoints or in metabolic degradation of compounds, 3D-structures are available directly or reasonable homology models can be built based on the X-ray structure of close homologs [11].

Frequently, 3D-ligand-based approaches are combined with protein structure or homology models, in order to derive more detailed insights into the protein–ligand interaction. In some cases, an automated docking procedure into the protein can be used to align the compounds for the 3D-QSAR approach. Several predictive models of the human ether-à-go-go-related gene (hERG) channel, for example, were built by combining homology models with 3D-QSAR [12–14].

The pregnane X receptor (PXR) is a nuclear receptor, regulating the expression of several metabolizing enzymes [15–17]. In particular, expression of CYP3A4 can be

significantly increased by the activation of PXR resulting in increased degradation of drugs metabolized by CYP3A4. X-ray structures of PXR without [18] and with ligands [19–24] are available and have been used in docking studies to develop binding hypothesis of PXR activators [25]. Due to the large lipophilic, flexible binding site, two hypotheses were followed up, with one abolishing PXR activation for newly synthesized compounds.

Numerous X-ray structures are also available for human cytochrome P450s (CYPs), involved in metabolic degradation of drugs, for example, CYP3A4 [26], CYP2D6 [27], and CYP2C9 [28]. Docking studies are used to predict metabolites or to propose putative binding modes for CYP-inhibitors in order to rationalize key interactions with the enzymes and decrease their inhibition [29–32].

Structural information about CYP enzymes is also used within the program MetaSite [33], which is able to predict the site of metabolism (SOM) for compounds. The structure of the CYP is used to calculate an interaction fingerprint based on GRID-derived pharmacophoric features of the binding site [34]. This fingerprint is compared with the corresponding interaction fingerprint of the compound to identify which sites of the molecules might be exposed to the heme and potentially could be oxidized. The final ranking of potential SOM is done by taking into account the chemical reactivity of the corresponding functional group with respect to radical abstraction. Overall, MetaSite ranks potential SOM within a molecule. About 80% of the MetaSite predictions are in accordance with experimental data.

11.1.2 Ligand-Based ADMET Models

For many properties in the ADMET field, ligand-based approaches are applied. Among them, 3D-pharmacophores are quite powerful as they identify molecular features which are responsible for undesired properties of a series. For example, pharmacophores have been developed for the Kv11.1 potassium ion channel (hERG). Most hERG pharmacophores contain a spatial arrangement of one positively ionizable group surrounded by some lipophilic features [35–39].

A large database of pharmacophores was developed by Inte:Ligand [40]. These models were automatically derived from available protein structures with the software LigandScout [41] or were built manually for selected targets, for which no structural information is available. This pharmacophore database is suitable for *in silico* profiling of compounds yielding early on potential liabilities of molecules, which have to be proven experimentally.

In recent years, more and more approaches emerged for *in silico* screening of compounds against a panel of antitargets. These approaches employ different techniques, for example, the similarity principle [42–45], Bayesian modeling [46–48], or different machine learning techniques [49, 50]. These approaches can be used to predict affinity fingerprints of compounds allowing the identification of antitargets and putative side affinities [51].

Very often, statistical approaches have been used to establish a correlation of chemical structures to experimental data. This requires an adequate description of structures by molecular descriptors and robust statistical techniques to identify a correlation to

ADMET data. Typically, a large set of different descriptors is entered into these models, ranging from 1D-descriptors such as molecular weight, $\log P$, etc. to 2D-descriptors, which are computed from a topological representation of the molecules [52], such as chemical fingerprints [53], chemically advanced template descriptors (CATS) [54, 55], or pharmacophoric fingerprints [56, 57]. These descriptors are combined with different algorithms for regression or for classification models. Prominent machine learning algorithms, which have been successfully used for ADMET models are random forests [58, 59], neuronal nets [60], support vector machines [61], and decision tree approaches [62]. Machine learning approaches, for example, have been used recently for the generation of global permeability models [63].

11.2 CASE STUDY ON METABOLIC LABILITY

Compounds with a high rate of metabolism suffer from short half-life which is often detrimental for a suitable (e.g. once daily) dosing. Therefore, optimization of the metabolic behavior of compounds is an important task in drug discovery.

Drug metabolism typically involves different steps. In phase I, compounds are chemically modified in a way to increase hydrophilicity of the compounds. In phase II metabolism, endogenous and polar groups are conjugated to the drug compounds to further increase hydrophilicity and aqueous solubility [64, 65]. Phase I metabolism is mainly done by oxidizing enzymes such as CYPs, amongst which CYP3A4, CYP2D6, and CYP2C9 are the most important ones [66]. CYP3A4 is considered to be responsible for the metabolism of about 50% of all drugs [67]. Phase II metabolism is driven by conjugating enzymes, such as glutathione S-transferase and sulfatases, or by glucoronidases.

Due to its important role for the pharmacokinetics of drugs and drug candidates, metabolic lability is studied early in the drug discovery process; for example, during the identification of suitable chemical hit series. In lead optimization, monitoring and optimization of metabolic lability becomes important to reach an acceptable metabolic stability for the desired dosing scheme. In every phase of early discovery, usually more compounds are available than can be tested for the ADMET parameters. In this situation, predictive *in silico* models can help to prioritize compounds for experimental testing or to guide compound optimization.

Cytochrome enzymes are major players in phase I transformations. They recognize and oxidize a broad panel of different substrates. Different CYP enzymes also have overlapping substrate specificity and show complex kinetic behavior. For the most important CYP enzymes for drug metabolism crystal structures are available, such as for CYP3A4 [26], CYP2D6 [27], or CYP2C9 [28]. The structural analysis of these structures show that they typically have a large lipophilic binding site and exhibit significant structural flexibility, as seen for example from the X-ray structures of CYP3A4 bound to ritonavir or metyrapone. These observations partly explain the observed ligand promiscuity of CYP enzymes.

Oxidative metabolic clearance has many facets, which renders modeling approaches to cytochrome-mediated metabolism rather challenging. Accordingly,

only few approaches have been described so far, which attempt to predict the metabolic clearance for larger sets of compounds.

One method uses the similarity principle to predict the metabolic lability in a k nearest neighbors (kNN) approach [68] based on MolconnZ [69] descriptors and atom pair descriptors [70]. At Eli Lilly, regression and classification models for metabolic stability of more than 12,000 compounds were built using multiple linear regression, discriminant analysis, and recursive partitioning [71]. Researches from Pfizer have trained random forest and Bayesian models from about 14,500 compounds measured in human liver microsomes (HLM) [72]. Recently, Gaussian Process models were developed using DRAGON descriptors [73] on a larger compound set from Bayer Schering [74]. Typically, more than 80% correct predictions could be obtained for the training set, while performance for the test sets slightly dropped.

All described methods aim to develop global models applicable to a large chemical space. It should be kept in mind however that the applicability domain of models is important. Series-specific (local) models are precise for local SAR trends but narrow in their applicability, while global models built from large chemical space offer the advantage of cross-series transfer of knowledge albeit with less accuracy.

The following part of this chapter describes in detail the development of a global model for metabolic lability based on a large dataset from a human liver microsomal assay. The software package Cubist [75] was applied as it offers a variety of different decision tree-based machine learning techniques.

11.2.1 Model Building

Driven by the mechanistic complexity of metabolism, a broad evaluation of the most successful combination of different machine learning algorithms and molecular descriptors was made. The focus is on machine learning methods, which are capable of dealing with different mode of actions within a dataset. Decision trees partition the dataset into different subsets at certain property thresholds of the descriptors, if overall separation into active or inactive compound improves. The C5 implementation developed by Quinlan [76] as provided by Rulequest in the software package Cubist was selected as it offers a variety of algorithms in combination with decision trees. Regression trees combine a decision tree with multiple linear regression models, which are derived for the data points in the terminal nodes of each leaf [77]. The predictive power of decision trees can be strengthened by an analysis of nearest neighbors in the training dataset (kNN approach). The prediction for a new data point is corrected by the deviations between prediction and actual values for the kNN in the training set [78]. Cubist does also allow for the generation of committee models. Committee models within Cubist are built in a way, that subsequent models correct errors made by the previous models. Thus, the first model is generated by the standard approach, but a second model corrects for the mispredictions of the first one. An optional third model then corrects for the errors of the two previous models, and so on.

Metabolic degradation of compounds by cytochromes comprises two important aspects. Initially, the ligand has to be recognized by the enzyme and needs to bind at its active site. Following that, a ligand site is exposed to the heme-oxygen complex

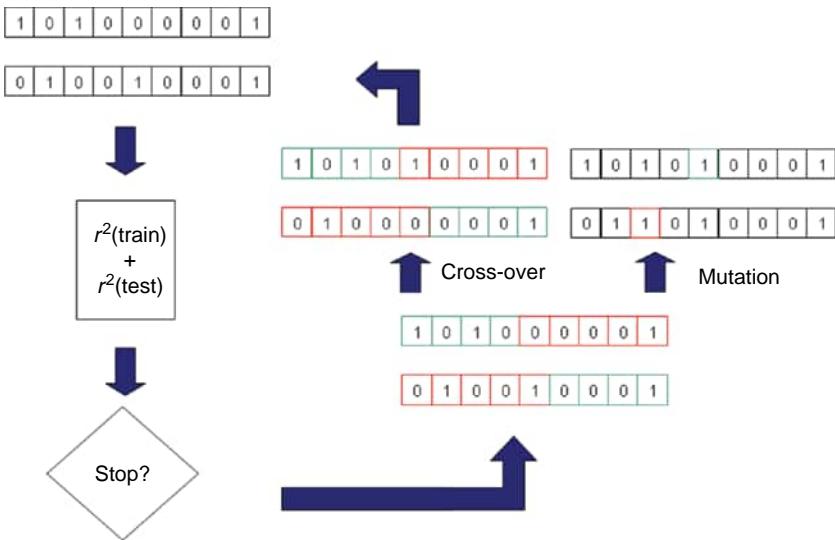


FIGURE 11.1 Workflow of the genetic algorithm used for feature selection to reduce the number of molecular descriptors for model building. The bit string encodes whether a descriptor is used in model building. The stop criterion is derived from the goodness of the fit between predicted and experimental values for the model. For color details, please see color plate section.

and undergoes an oxidative reaction. Therefore, chemical reactivity of the ligand and the CYP enzyme is the second important driver for metabolism of compounds.

To account both for molecular recognition and for reactivity, different descriptors encompassing physicochemical properties, surface properties, shape-related properties, or pharmacophoric patterns were employed. These descriptors were derived from different software packages and algorithms like DRAGON [73], MOE [79], and CATS [54, 55]. In addition, PARASURF [80] provides descriptors from quantum-chemical calculations, which reflect the electronic structure of the ligands and also capture some of the reactivity trends of ligands. These descriptors were supplemented by MACCS keys [81] based on frequent chemical fragments and by physicochemical properties calculated with QikProp [82].

Typically, in large descriptor sets, numerous descriptors show only slight variability over the dataset or are strongly correlated among each other. Therefore, the number of descriptors should usually be reduced to avoid random noise. To this end, a genetic algorithm (GA) for feature selection, driven by the quality of the generated model, was applied (cf. Figure 11.1). The individual chromosomes encode the descriptors used for model calculation. Initially, an ensemble of 100 chromosomes is randomly generated. While the best solutions are preserved from each generation (survival of the fittest), a child generation is produced by crossover and by mutations. At every step of the iteration, the selected descriptors are submitted for model building. The corresponding fitness score for the selected descriptors results from the correlation coefficients r^2 between experimental and predicted values for the training

set ($r^2(\text{train})$) and for the test set ($r^2(\text{test})$). Up to 1000 generations are generated by crossbreeding and pruning, until the final model is selected. All models were built as quantitative models, capable of predicting the percentage of compound metabolized. When models for metabolic lability are used as global models for the ranking of lead compounds, a classification is often sufficient. Therefore, thresholds are derived to convert the percentage of metabolized compounds into three classes, namely metabolically stable, moderately labile, and labile.

11.2.2 Dataset

Metabolic lability is routinely tested in an *in vitro* assay, using human liver microsomes (hlm). The amount of parent drug remaining is determined after an incubation time of 15 min. In aiming at a global model, a large Sanofi dataset obtained from different in-house laboratories was collected. All data was obtained with a harmonized in-house assay protocol to ensure consistency of the dataset. Comparability of data from different labs is routinely checked by a set of several tool compounds. Accordingly, data consistency was observed for more than 80% of repeated measurements showing deviation smaller than 10% in absolute values.

Data points with high variability or low contribution of CYP enzymes to the observed metabolic lability were excluded. Compounds were also omitted, for which repetitive measurements showed standard deviations larger than 10% in absolute values. The experimental studies are done with and without the addition of NADPH. Significant deviations between these values indicate only a minor influence of CYPs on the metabolic lability. Thus, corresponding compounds were also discarded from the dataset.

Numerous chemical filters were applied to focus the training space to relevant drug space. Therefore, reactive fragments, such as aliphatic sulfates, as well as rarely populated chemical fragments were excluded.

The final dataset consisted of ~6400 compounds. The chemical diversity of the dataset was analyzed using a modified version of Murcko frameworks [83], which have been originally defined as contiguous ring systems including all chain linkers between rings. Initially, all side chains are clipped from the original molecule, but exocyclic double bonds are maintained to keep the nature and aromaticity of ring scaffolds. In a subsequent step, all carbon chain linkers $n \geq 3$ between rings are being clipped, and the most complex ring system in terms of number of atoms and number of heteroatoms is reported. Whenever less than two rings are present, the original molecule is reported. The framework analysis of the microsomal lability dataset yielded almost 3300 different fragments illustrating chemical diversity in this set.

For model building and validation, the dataset was randomly split into a training set of 5151 compounds and an internal test set of 1288 molecules.

11.2.3 Results

A stepwise strategy to find the most suited combination between the different machine learning approaches and the different descriptor sets was employed. In the first step, different descriptor sets available from various software packages were

TABLE 11.1 Comparison of Different Descriptor Sets for Modeling of Metabolic Liability from Human Liver Microsomes

	$r^2(\text{train})$	$r^2(\text{test})$	Number of Descriptors
DRAGON	0.71	0.41	590
MOE	0.60	0.29	92
CATS	0.57	0.22	108
PARASURF	0.56	0.26	38
MACCS	0.54	0.26	95
MOE/PARASURF	0.68	0.43	131
CATS/PARASURF	0.68	0.43	128
MACCS/PARASURF	0.68	0.41	130

Models were generated with GA-based feature selection in combination with regression tree models from Cubist.

evaluated using GA-based feature selection coupled to a regression tree-based approach. In the second step, different machine learning algorithms were compared using the set of descriptors from step one. Model quality was assessed using the Pearson correlation coefficient between calculated and experimental data for training and test set. For the training data, the correlation coefficient reaches values between 0.6 and 0.7, while for the test data the correlation coefficients drop to values below 0.4.

Significant differences between individual descriptor sets were observed (see Table 11.1). The best correlation between predicted and experimental data was obtained from a large set of descriptors calculated by DRAGON version 1.4. In general, the GA reduced the number of descriptors significantly.

In particular, combinations of PARASURF descriptors with molecular surface-based descriptors were of interest. Descriptors like the local ionization potential or the local electron affinity, provided by PARASURF, should capture aspects of chemical reactivity, while surface or pharmacophore-related descriptors could have the potential to describe the recognition of the molecule by cytochromes. Hence, various descriptor sets were combined with quantum-chemical descriptors from PARASURF yielding improved models. For example, the sole set of 2D-MOE descriptors shows an r^2 of 0.6 for the training set, while r^2 rises to 0.68 for the MOE/PARASURF combinations.

The number of variables, which have entered the final models, could be significantly reduced by feature selection with a GA. Overall, less than 150 descriptors appear to be sufficient to obtain reasonable models.

DRAGON descriptors and the combination MOE/PARASURF show the best correlations between predicted and experimental data. In the second step, the comparison of different machine learning algorithms was performed in combination with descriptors calculated by MOE and PARASURF. It turned out that the predictivity improves by the correction approaches available within Cubist. If the model predictions are corrected based on a kNN approach with five neighbors giving best results, correlation coefficients rise to 0.97 for $r^2(\text{train})$ and 0.47 for $r^2(\text{test})$ compared

TABLE 11.2 Confusion Matrix for the Cubist Committee Model Based on DRAGON Descriptors

	Training Set Prediction			Test Set Prediction		
	Stable (%)	Moderate (%)	Labile (%)	Stable (%)	Moderate (%)	Labile (%)
Stable	83	36	3	73	32	8
Moderate	12	30	6	15	31	12
Labile	5	34	92	12	36	80

Classification is done by application of thresholds at 20% and 40%. (Predicted data have been normalized to % of experimental class.)

to 0.69 for r^2 (train) and 0.41 for r^2 (test) for the regression tree approach. Similar values are found with a committee model consisting of five models. While r^2 (test) reaches 0.43, r^2 (train) is 0.78. For DRAGON descriptors, almost identical values were achieved (r^2 (train)=0.79, r^2 (test)=0.44).

Committee models based on DRAGON descriptors or the descriptor combination from MOE and PARASURF appear to result in comparable models with good statistical quality. Since the calculation of PARASURF descriptors employs a semiempirical calculation step, it is significantly more time consuming than the calculation of DRAGON descriptors. Hence, the subsequent classification committee models were built only with DRAGON descriptors.

Classification of compounds into three classes is often suitable for many applications, for example, the prioritization of compounds for experimental testing in a hit evaluation phase, the removal of potential metabolically unstable from planned chemical libraries, and so on. The classification performance of the final model is illustrated by a confusion matrix (Table 11.2). Overall, 66% are classified correctly into the given three classes, while the reliability for the metabolically labile compounds is slightly better (80%) than for the stable compounds (73%). Only 10% of the compounds are mispredicted over two classes.

Typically, QSAR models work reasonably well within their domain of application, but the predictivity breaks down significantly outside this domain. In pharmaceutical industry, the chemical estate is permanently increased by synthesizing or purchasing novel compounds, which are not necessarily well predicted by static models. Especially, if compounds are excluded from further activities based on *in silico* models, it is important to assess the validity of predictions.

Many different approaches have been described for estimation of the applicability domain of a model: chemical similarity to the training set compounds [84], distance to the used descriptor space, significant deviations in the range of the descriptors from the training set, and so on [85–89]. In this particular example, the similarity of a compound to the training set is used to identify molecules, which are most likely outside the domain of applicability. In order to establish the best similarity thresholds, the average prediction error for novel compounds depending on their similarity to the closest compound in the training set was analyzed using newly collected experimental

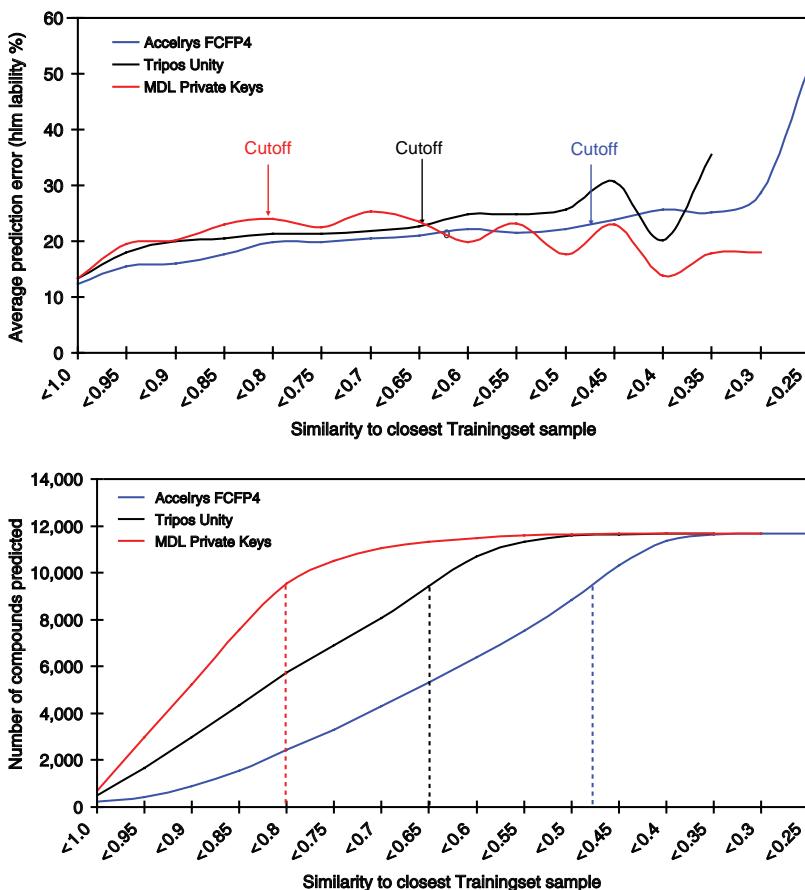


FIGURE 11.2 Analysis of the prediction error depending on the similarity of predicted compound to the closest compound in the training set. The top panel shows the prediction error, the bottom panel plots the number of compounds, which can be predicted at the respective similarity threshold. For color details, please see color plate section.

data and compared to the number of compounds predicted (cf. Figure 11.2). In general, the prediction error rises, if no similar compound is in the training set according to Unity fingerprints [90] and FCFP4 fingerprints [91]. If the similarity cutoff is chosen too restrictive, only few compounds can be predicted. We selected a similarity cutoff of 0.65 for Unity fingerprints as a reasonable compromise between prediction error and number of compounds being predictable. Compounds below the similarity threshold do not have close neighbors in the training set and are flagged as out-of-the-model applicability domain.

For model update, newly collected experimental data were pooled together with the original data, resulting in about 15,000 training and 3,000 test compounds. The final Cubist committee model comprises 602 DRAGON descriptors. Due to the

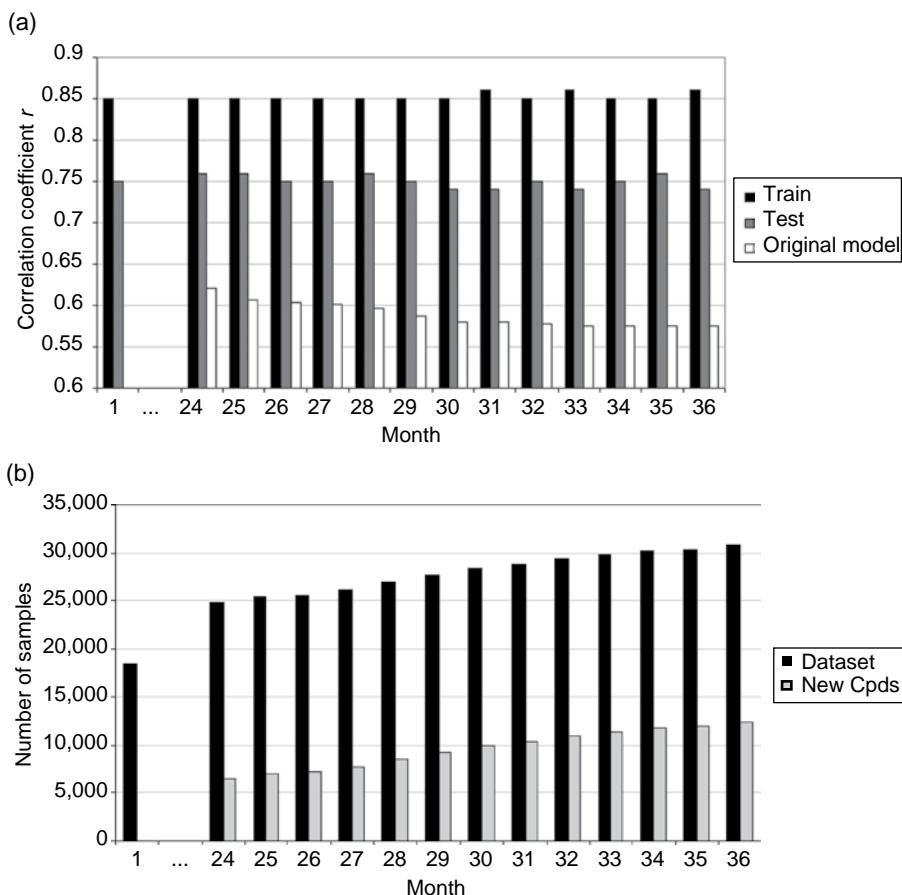


FIGURE 11.3 Analysis of model performance over time. Every month the analysis is repeated with all newly obtained data from the in-house microsomal lability assay. (a) The model performance is given by the correlation coefficient between experimental and predicted values for each new dataset. The light bar shows the performance of the original model, the dark bars show the performance of a model that was updated on a monthly base. Model update is done by adding all previously determined data to the training dataset. (b) The initial model was derived from a combined training and test set of 18,000 samples. After 36 months, the number of available samples has increased by 12,500 samples.

increased coverage of the chemical space, the training set performance remains constant ($r^2=0.73$), but the test set performance improves ($r^2(\text{test})=0.56$). The rate of correct classifications increases from 66% to 67%.

As a final validation of the applicability of the model, its performance was monitored over a certain time of application. Every month, all newly tested compounds were collected and predicted by the model. All compounds have been excluded from prediction and analysis, for which the applicability of the model is not guaranteed. The correlation coefficients for all external compounds were determined and analyzed

(Figure 11.3a). After a period of 2 years, the correlation coefficient of more than 6000 new samples has already dropped to 0.63, and in parallel to the arrival of new samples (cf. Figure 11.3b) it continues to decline steadily for the following 12 months until a correlation coefficient of 0.56 is obtained.

In parallel, regular updates of the model were generated by adding novel compounds to the training set. Since the training procedure is quite time consuming, the GA-based model building was not repeated. Instead, novel compounds were added to the training set and only the descriptor weights were recalculated. By this method, the correlation coefficient of the test set remains stable at 0.75 allowing more compounds to be predicted, since the training space is permanently increased. This procedure assures quick model update cycles by feeding in novel synthesized compounds resulting in a continuous increase of the validity domain of the model.

This case study described model building for the prediction of metabolic lability of novel compounds. The analysis of different descriptors and machine learning algorithms shows that the chosen descriptor set, as well as the machine learning algorithm influence the predictivity of the model. Committee models, as implemented within Cubist, include an inherent error correction mechanism, which improves predictivity.

The final model returns the metabolic lability of novel compounds. With defined thresholds, the quantitative prediction is converted into a classification. This classification is used for flagging of compounds for experimental testing (e.g., in hit or lead evaluation). On the other hand, quantitative predictions are of value to guide compound optimization, if the applicability of the model has been demonstrated. This approach is illustrated in the following example of a multidimensional optimization program.

11.2.4 Application of a Global Model for Metabolic Lability in the Optimization of DGAT1 Inhibitors

Diacylglycerol-acyl-transferase 1 (DGAT1) is an enzyme, which is involved in the metabolism of fatty acids. DGAT1 is membrane-bound to the endoplasmatic reticulum and catalyzes the final step in the triglyceride synthesis, thus it is considered relevant for metabolic disorders and for the treatment of obesity [92, 93]. During the course of lead identification and optimization, *in silico* models have been successfully applied to identify new (DGAT1 inhibitors. The class of 2-aminothiazols was discovered as promising lead structures for inhibition of DGAT1 [94], represented by compound 1 which displayed an IC_{50} of 0.25 μM in the enzymatic assay.

Compound 1 shows encouraging functional activity, but at the same time it suffers from high oxidative metabolism accompanied by rather high metabolic lability in HLMs (lability=53%). Compound 1 is highly lipophilic, with a $\log D$ of 5.87 at pH=7.4 ($\text{PSA}=115.38\text{\AA}^2$). A conceptually simple strategy to reduce the rate of metabolism within a congeneric series of lipophilic compounds employs the reduction of overall lipophilicity of the compound. Thus, preferably polar functional groups, such as amines, amides, hydroxyls, sulfones, sulfonamides, and carboxylic acids, were introduced at various sites of the scaffold. These variations, which had to be amenable to synthesis, were screened upfront *in silico* for their compatibility to the DGAT1 pharmacophore of the series and for the rate of oxidative microsomal metabolism predicted with the Cubist committee model.

Following the concept of “model applicability domains,” a validation of the applicability domain of a global model is strictly advisable for a reasonably sized compound set ≥ 10 . Within this series, an excellent correlation of the predicted metabolic lability to the experimental data was observed with a correlation coefficient of $r^2=0.80$ (Figure 11.4). Thus, the global model for human microsomal lability is applicable for the optimization of this chemical series. Metabolically stable compounds were finally obtained by introduction of a benzoylvaline moiety yielding compound 2 with an IC_{50} of 30 nM in the enzymatic test (metabolic lability of 1%, $\log D=3.0$ at $pH=7.4$, $PSA=132.45\text{ }\text{\AA}^2$).

The introduction of polar, particularly charged, functional groups often leads to a switch of the primary SOM; it may even induce a different way of clearance, such as renal or biliary clearance, or disbalance other ADME optimization parameters, such as passive permeability. Within this series of DGAT1 inhibitors, the carboxylic acid turned out to be essential for metabolic stability, but also led to extremely low passive permeability. This permeability issue was observed with all negatively charged derivatives of compound 1, particularly for benzoyl-aminoacids, mainly driven by a high polarity, resulting in a multidimensional optimization problem dealing with activity, metabolic stability and permeability.

Thus, a benzoylvaline surrogate had to be identified in order to retain good inhibitory activity of DGAT1 and improve permeability. Chemical rescaffolding attempts [95] ultimately led to the phenylcyclohexyl acetic acid series (compounds 3 and 4, Table 11.3).

Compound 3, a potent inhibitor of DGAT1 with an IC_{50} of 40 nM, has a high metabolic lability of 41%, and the oxidative metabolism observed in microsomes was found to be mainly driven by CYP3A4 (contribution: 69%). Compound 4, which is a 19 nM inhibitor of DGAT1, exhibits a metabolic lability of 37% with a dominant CYP3A4 contribution of 63%. In order to guide further chemical synthesis, information relating to metabolic “hot spots” within unstable compounds was required, which cannot be directly extracted from 2D-quantitative structure–property relationship (QSPR) models alone. *In silico* tools such as MetaSite support the identification of potential SOM in a molecule once the main CYP enzyme responsible for degradation is known. However, the current version of MetaSite cannot predict rates of metabolism.

MetaSite has repeatedly proven useful in the optimization of metabolic properties of lead-like compounds [95, 96] (see earlier) and it was used to study potentially labile sites of compound 4 in detail.

The CYP450 3A4 model based on the ketoconazole crystal structure was used within MetaSite3 to predict SOM for compound 4 (Figure 11.5a). Built-in corrections for reactivity of the cytochrome and the ligand were applied. Both for compounds 3 and 4, the predicted primary SOM is the benzylic alpha position to the thiadiazole ring (Figure 11.5b). The SOM predicted by MetaSite next in hierarchy, the hydroxylation of the benzylic position of the cyclohexyl moiety, was not observed experimentally, nor was the aromatic oxidation of the benzothiazole sulfur. Instead, *in vitro* metabolism studies are in agreement with the primary SOM and confirmed a labile cyclopentyl-ethyl moiety, but a stable thiadiazol scaffold and a stable

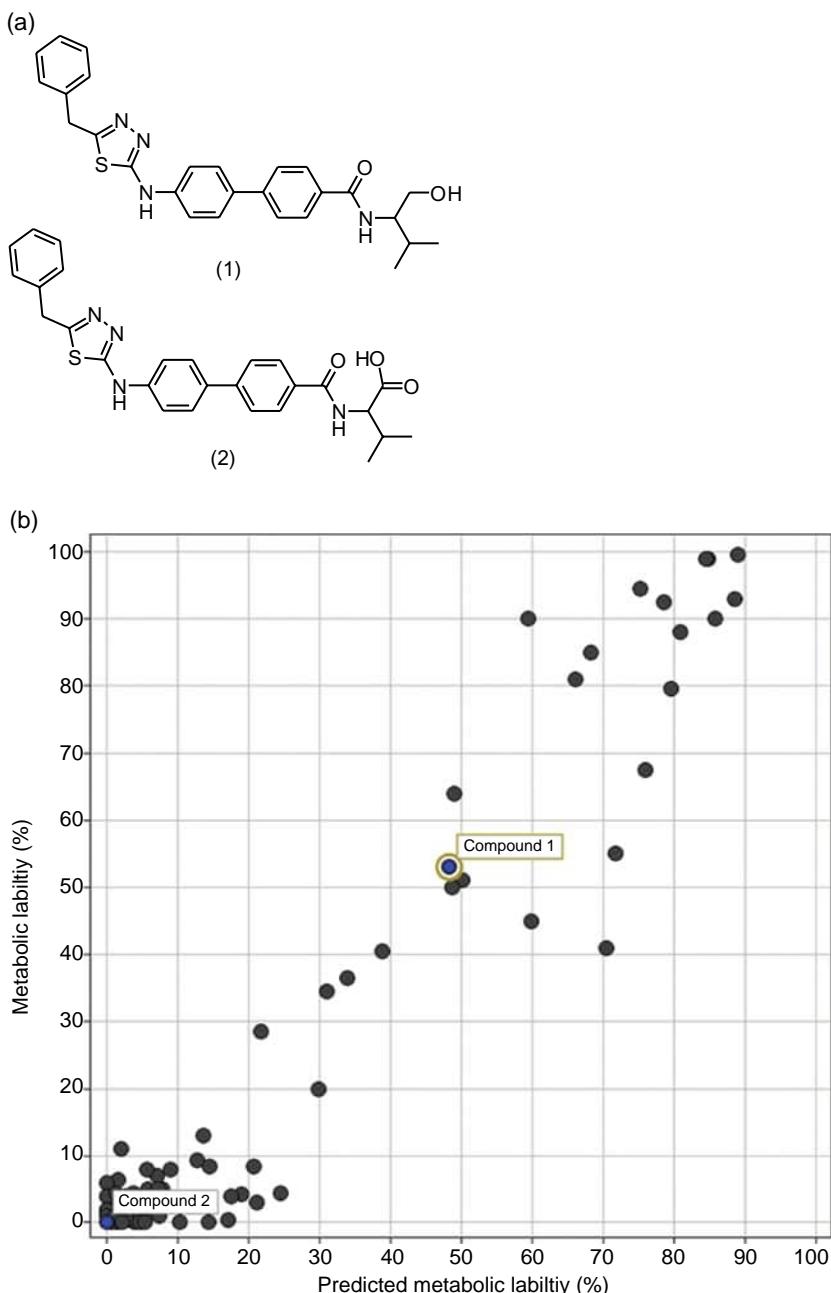
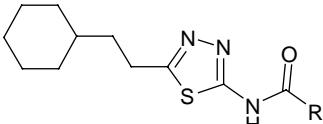
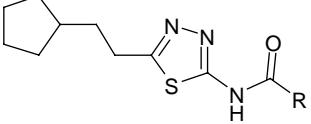
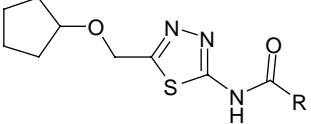
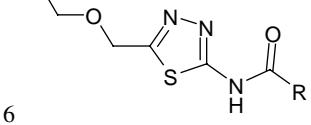
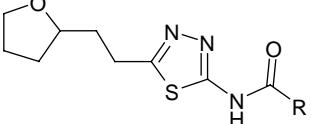


FIGURE 11.4 Predicted and experimental metabolic lability of a series of DGAT1 inhibitors. The predictions were made with a 2D-regression tree committee model based on DRAGON 1.4 descriptors.

TABLE 11.3 Experimental and Predicted Metabolic Liabilities in Human Liver Microsomes of a Series of DGAT1 Inhibitors with a Cyclohexylacetic Acid (R) as Common Pharmacophore

Compound	Metabolic Liability (Predicted) (%)	Metabolic Liability (Experimental) (%)
	38.9	41
3		
	34.0	37
4		
	7.4	4
5		
	14.3	0
6		
	14.5	9
7		

The metabolic liability in human liver microsomes was predicted with a 2D-QSAR model.

phenyl-cyclohexyl linker unit. Thus, the structure contribution graph predicted by MetaSite was used to identify atomic contribution essential for substrate enzyme interactions. It suggests a dominant contribution of the carboxylic function, responsible for orienting the aliphatic linker towards the heme and making it accessible for a hydroxylation reaction (Figure 11.5c). A classical medicinal chemistry approach towards improving metabolic stability is to chemically block SOM by chemically inert groups [97]. In general, metabolism can be reduced by introduction of stable

functional groups, such as inert halogens, at the SOM, or by addition of polar isosteres in close proximity to the SOM that modulate the substrate recognition.

For a quantitative prediction of microsomal lability, our global QSAR model was employed again with the purpose to restrict the chemical synthesis of new compounds

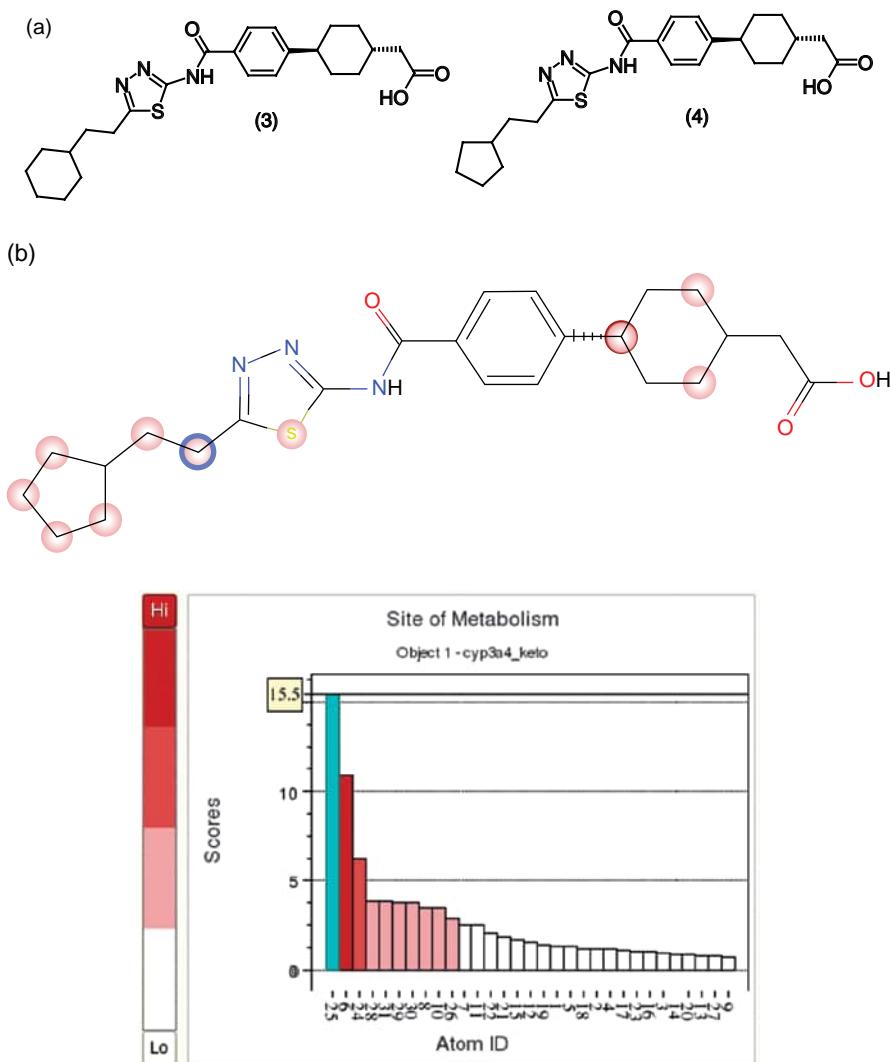


FIGURE 11.5 (a) A series of DGAT1 inhibitors with a benzylamidothiadiazole scaffold sharing a cyclohexylacetic acid as common pharmacophore. (b) Analysis of compound 4 with MetaSite3 [99]. The proposed main metabolite is the aliphatic hydroxylation product at the SOM. The scores plot ranks the thiadiazole alpha position on top of all substrate sites (blue bar). Most probable SOM are color coded in a structure diagram (blue circle). For color details, please see color plate section.

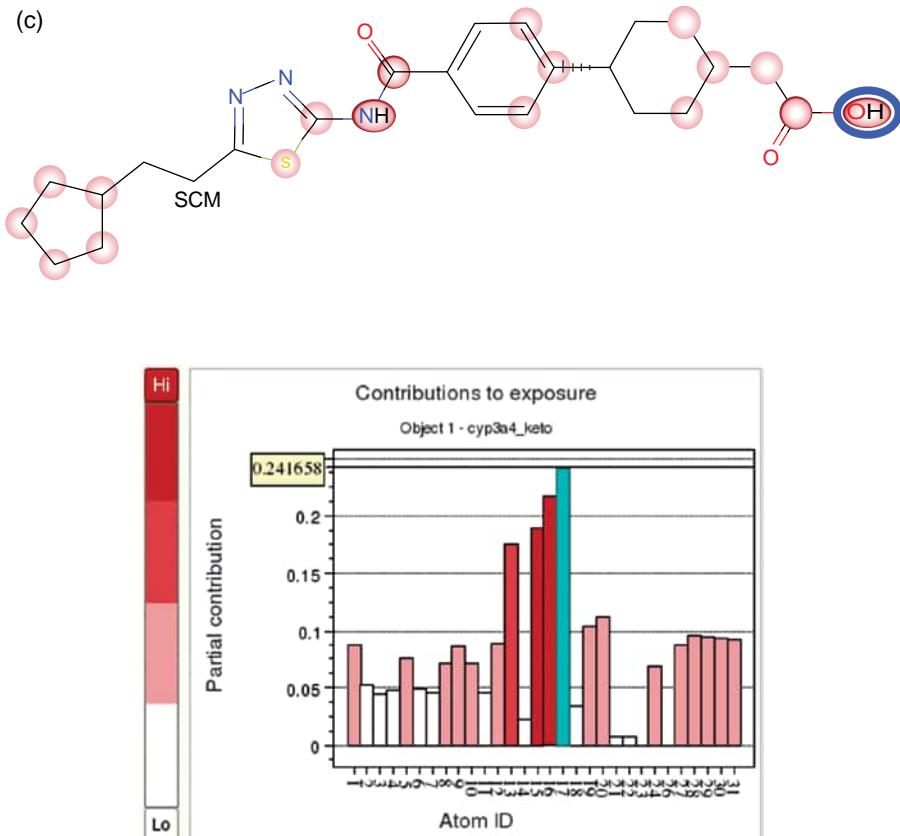


FIGURE 11.5 (Continued) (c) Depending on the SOM, atomic contributions to the substrate recognition are analyzed in a second graph. A dominant contribution to the orientation of compound 4 is made by the top-ranked carboxylic acid function (blue circle) and the second-ranked amide function. For color details, please see color plate section.

to the most promising synthesis proposals. Prediction with the QSAR model leads to the suggestion that replacing the labile methylene position by a chemically inert difluoromethylene group may block the primary SOM but would not improve the total rate of metabolism of this compound.

Introduction of polarity proximal to the cyclopentyl-ethyl function is accommodated by a reduction of the predicted total metabolism observed. Thus, aliphatic ethers were introduced, which are also tolerated by the DGAT1 pharmacophore.

Table 11.3 demonstrates a very reasonable correlation between the measured intrinsic lability in HLMs and the *in silico* predictions for this series. Within the experimental error, the metabolic stability is almost quantitative, if the ether function is introduced proximal to the SOM (compounds 5 and 6). Introduction

of a tetrohydrofuran function still has a substantial effect and improves the metabolic stability by a factor of 4. This experimental trend is correctly predicted by the 2D-QSAR model.

Finally, a series of cyclopentyl ethers was identified as potent, metabolically stable derivatives leading to the discovery of a next generation lead series.

11.3 CONCLUSION

A global model for the prediction of human liver microsomal lability was described. Model building was focused to machine learning algorithms, which are principally capable of dealing with multiple modes of actions. The software Cubist provides such a set of machine learning tools employing decision tree-based algorithms. A global model for metabolic lability was built by systematically testing large descriptor sets in combination with different machine learning algorithms. The finally used committee model based on DRAGON descriptors allows efficient prediction of novel compounds. The model returns quantitative predictions for metabolic lability, which can be translated into a classification scheme by defined thresholds. The applicability of the model was underpinned by the optimization of a metabolically labile DGAT-1 lead structure. For the early evaluation of novel chemical matter in hit exploration attempts the categorizing model gives a warning flag, which then triggers experimental testing. In both application scenarios, it is of key importance to proof the applicability of the model to the series of interest. This was accomplished with a Unity fingerprint based similarity criteria. In compound optimization early validation of the model by correlating predicted versus observed metabolic lability is recommended.

Monitoring of model performance over time clearly shows that the predictivity deteriorates with more and more novel chemotypes becoming available. Therefore, monthly update procedures were implemented resulting in updated descriptor weights. With this update procedure, the applicability domain permanently increases due to the broader chemical space of the training set.

Overall, the illustrated example of modeling metabolic lability shows that global *in silico* ADMET models are successful tools to optimize ADMET parameter. Future predictions of ADMET models might be added into complex tools such as physiological-based pharmacokinetic models yielding an *in silico* engine for ADMET support [98].

REFERENCES

1. Nomeir AA. ADME strategies in lead optimization. In: Cayen MN, editor. *Early Drug Development*. Hoboken: John Wiley & Sons, Inc.; 2010. p 27–88.
2. Korfomacher W. Strategies and techniques for higher throughput ADME/PK assays. In: Wang PG, editor. *High-Throughput Analysis in the Pharmaceutical Industry*. Boca Raton: CRC Press; 2009. p 205–231.

3. Lipinski CA, Lombardo F, Dominy BW, et al. *Adv Drug Deliv Rev* 1997;23:3–25.
4. Vieth M, Siegel MG, Higgs RE, et al. *J Med Chem* 2004;47:224–232.
5. Wenlock MC, Austin RP, Barton P, et al. *J Med Chem* 2003;46:1250–1256.
6. Leeson PD, Springthorpe B. *Nat Rev Drug Discov* 2007;6:881–890.
7. Morphy R. *J Med Chem* 2006;49:2969–2978.
8. Oprea TI. *J Comput Aided Mol Des* 2002;16:325–334.
9. Oprea TI, Davis AM, Teague SJ, et al. *J Chem Inf Comput Sci* 2001;41:1308–1315.
10. Congreve M, Carr R, Murray C, et al. *Drug Discov Today* 2003;8:876–877.
11. Moroy G, Martiny VY, Vayer P, et al. *Drug Discov Today* 2012;17:44–55.
12. Pearlstein RA, Vaz RJ, Kang J, et al. *Bioorg Med Chem Lett* 2003;13:1829–1835.
13. Durdagi S, Duff HJ, Noskov SY. *J Chem Inf Model* 2011;51: 463–474.
14. Taboureau O, Jorgensen FS. *Comb Chem High Throughput Screen* 2011;14:375–387.
15. Kliewer SA, Moore JT, Wade L, et al. *Cell* 1998;92:73–82.
16. Xie W, Barwick JL, Simon CM, et al. *Genes Dev* 2000;14:3014–3023.
17. Bertilsson G, Heidrich J, Svensson K, et al. *Proc Natl Acad Sci USA* 1998;95:12208–12213.
18. Watkins RE, Wisely GB, Moore LB, et al. *Science* 2001;292:2329–2333.
19. Watkins RE, Maglich JM, Moore LB, et al. *Biochemistry* 2003;42:1430–1438.
20. Watkins RE, Noble SM, Redinbo MR. *Curr Opin Drug Discov Dev* 2002;5:150–158.
21. Chrencik JE, Xue Y, Orans JO, et al. *Mol Endocrinol* 2005;19:1125–1134.
22. Xue Y, Redinbo MR. *Bioorg Med Chem* 2007;1:2156–2166.
23. Watkins RE, Davis-Searles PR, Lambert MH, et al. *J Mol Biol* 2003;331:815–828.
24. Teotico DG, Bischof J, Jason J, et al. *Mol Pharmacol* 2008;74:1512–1520.
25. Gao Y-D, Olson SH, Balkovec JM, et al. *Xenobiotica* 2007;37:124–138.
26. <http://www.rcsb.org/>, Codes 1TQN, 1WOE, 1WOF, 1WOG, 2JOD, 2VOM, 3NXU. Accessed 2013 May 14.
27. <http://www.rcsb.org/>, Code 2F9Q. Accessed 2013 May 14.
28. <http://www.rcsb.org/>, Codes 1OG2, 1OG5, 1R9O. Accessed 2013 May 14.
29. Tie Y, McPhail B, Hong H, et al. *Molecules* 2012;17:3407–3460.
30. Kirchmair J, Williamson MJ, Tyzack JD, et al. *J Chem Inf Model* 2012;52:617–648.
31. Stjernschantz E, Vermeulen NPE, Oostenbrink C. *Expert Opin Drug Metab Toxicol* 2008;4:513–517.
32. Tarcsay A, Kiss R, Gyoergy KM. *J Comput Aided Mol Des* 2010;2:399–408.
33. Cruciani G, Carosati E, De Boeck B, et al. *J Med Chem* 2005;48:6970–6979.
34. Goodford PJ. *J Med Chem* 1985;28:849–857.
35. Taboureau O, Joergensen FS. *Comb Chem High Throughput Screen* 2011;14(5):375–387.
36. Durdagi S, Subbotina J, Lees-Miller J, et al. *Curr Med Chem* 2010;17:3514–3532.
37. Ekins S, Crumb WJ, Sarazan RD, et al. *J Pharmacol Exp Ther* 2002;301:427–434.
38. Yamakawa Y, Furutani K, Inanobe A, et al. *Biochem Biophys Res Commun* 2012;418:161–166.
39. Recanatini M, Cavalli A. QSAR and pharmacophores for drugs involved in hERG blockade. In: Vaz RJ, Klabunde T, editors. *Methods and Principles in Medicinal Chemistry*. Weinheim: Wiley-VCH; 2008. p 109–126.

40. Steindl TM, Schuster D, Wolber G, et al. J Comput Aided Mol Des 2006;20:703–715.
41. Wolber G, Langer T. J Chem Inf Model 2005;45:160–169.
42. Keiser MJ, Shoichet BK. Nat Biotechnol 2007;25:197–206.
43. Keiser MJ, Shoichet BK. Nature 2009;462:175–181.
44. Mestres J. J Chem Inf Model 2006;46:2725–2736.
45. Schuffenhauer A. J Chem Inf Comput Sci 2003;43:391–405.
46. Nidhi A, Glick M. J Chem Inf Model 2006;46:1124–1133.
47. Nettles JH, Bender A. J Med Chem 2006;49:6802–6810.
48. Nigsch F, Bender A. J Chem Inf Model 2008;48:2313–2325.
49. Poroikov VV. J Chem Inf Comput Sci 2000;40:1349–1355.
50. Yabuuchi H, Niijima S, Takematsu H, et al. Mol Syst Biol 2011;7:472.
51. Garcia-Serna R, Mestres J. Expert Opin Drug Metab Toxicol 2010;6:1253–1263.
52. Gozalbes R, Doucet JP, Derouin F. Drug Targets Infect Disord 2002;2:93–102.
53. Duan J, Dixon JF, Lowrie W. J Mol Graphics Model 2010;29:157–170.
54. Fechner U, Paetz J, Schneider G. J Comput Aided Mol Des 2003;17:687–698.
55. Schuffenhauer A, Floersheim P, Acklin P, et al. J Chem Inf Comput Sci 2003;43:391–405.
56. McGregor MJ, Muskal SM. J Chem Inf Comput Sci 1999;39:569–574.
57. Mason JS, Morize I, Menard PR, et al. J Med Chem 1999;42:3251–3264.
58. Breiman L. Mach Learn 2001;45:5–32.
59. Svetnik V, Liaw A, Tong C, et al. J Chem Inf Comput Sci 2003;43:1947–1958.
60. Zupan J, Gasteiger J. *Neural Networks in Chemistry and Drug Design*. Weinheim: Wiley-VCH; 1999.
61. Christianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. New York: Cambridge University Press; 2000.
62. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
63. Geerts T, Heyden YV. Comb Chem High Throughput Screen 2011;14:339–361.
64. Testa B, Krämer SD. *The Biochemistry of Drug Metabolism: Principles, Redox Reactions, Hydrolysis*. Weinheim: Wiley-VCH; 2008.
65. Testa B, Krämer SD. *The Biochemistry of Drug Metabolism: Conjugations, Consequences of Metabolism, Influencing Factors*. Weinheim: Wiley-VCH; 2010.
66. Evans WE, Relling MV. Science 1999;286:487–491.
67. Clark SE, Jones BC. Human cytochromes P450 and their role in metabolism-based drug-drug interaction. In: Rodrigues AD, editor. *Drug-Drug Interactions*. New York: Informa Healthcare; 2003. p 55–88.
68. Shen M, Xiao Y, Golbraikh A, et al. J Med Chem 2003;46:3013–3020.
69. MolConnZ, version 3.5; Hall Associates Consulting: Quincy, MA, 1998.
70. Carhart RE, Smith DH, Venkataraghavan R. J Chem Inf Comput Sci 1985;25:64–73.
71. Gombar VK, Alberts JJ, Cassidy KC, et al. J Comput Aided Mol Des 2006;2:177–188.
72. Lee PH, Cucurull-Sanchez LJ, Lu YJ, et al. J Comput Aided Mol Des 2007;21:665–673.
73. Todeschini R, Consonni V, Mauri A, et al. 2006. Dragon for Windows and Linux 2006. Available at <http://www.talete.mi.it/>. Accessed 2013 May 14.
74. Schwaighofer A, Schroeter T, Mika S, et al. J Chem Inf Model 2008;48:785–796.

75. RuleQuest Research Pty Ltd, 30 Athena Avenue, St Ives NSW 2075, Australia.
76. Quinlan JR. J Artif Intell Res 1996;4:77–90.
77. Breiman L, Friedman JH, Olshen RA, *Stone Classification and Regression Trees*. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
78. Quinlan JR. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*. San Mateo: Morgan Kaufmann Publishers; 1993. p 236–243.
79. Molecular Operating Environment (MOE), 2011.10; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite 910, Montreal, QC, Canada, H3A 2R7, 2011.
80. Parasurf10, Cepos InSilico Ltd, Bedford MK42 8BQ, UK.
81. Ahrens EKF. Customization for chemical database applications. In: Warr WA, editor. *Chemical Structures*. Berlin: Springer; 1988. p 97–111.
82. QikProp, version 3.3, Schrödinger, LLC, New York, NY, 2010.
83. Bemis W, Murcko MA. J Med Chem 1996;39:2887–2893.
84. Sheridan RP, Feuston BP, Maiorov VN, et al. J Chem Inf Comput Sci 2004;44: 1912–1928.
85. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. ATLA, Alternatives to Laboratory Animals 2005, 33, 445–459.
86. Tetko IV, Bruneau P, Mewes H-W, et al. Drug Discov Today 2006; 11:700–705.
87. Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? J Chem Inf Model 2012;52(6):1413–1437.
88. Hewitt M, Ellison CM. Issues in Toxicology 2010;7(In Silico Toxicology):301–333.
89. Cronin MTD. Issues in Toxicology 2010, 7(In Silico Toxicology):275–300.
90. Tripos International, 1699 South Hanley Road, St. Louis, MO 63144-2319 USA
91. PipelinePilot, Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA
92. Subauste A, Burant CF. Curr Drug Targets Immune Endocr Metabol Disord 2003;3: 263–270.
93. Chen HC, Farese Jr, RV. Arterioscler Thromb Vasc Biol 2005;25:482–486.
94. Mougenot P, Namane C, Fett E, et al. Bioorg Med Chem Lett 2012;22:2497–2502.
95. Ahlström MM, Ridderström M, Zamora I, et al. J Med Chem 2007;50:4444–4452.
96. Caron G, Ermondi G, Testa B. Pharm Res 2007;24:480–501.
97. Böhm HJ, Banner D, Bendels S, et al. Chem Biol Chem 2004;5:637–643.
98. Paixao P, Gouveia LF, Morais JAG. Int J Pharm 2012;429:84–98.
99. Molecular Discovery Ltd., Via Stoppani, 38, 06135 – Ponte San Giovanni – Perugia, Italy http://www.moldiscovery.com/soft_metasite.php. Accessed 2013 May 14.

CHAPTER 12

CHEMOINFORMATICS AND BEYOND: MOVING FROM SIMPLE MODELS TO COMPLEX RELATIONSHIPS IN PHARMACEUTICAL COMPUTATIONAL TOXICOLOGY

CATRIN HASSELGREN, DANIEL MUTHAS, ERNST AHLBERG,
SAMUEL ANDERSSON, LARS CARLSSON, TOBIAS NOESKE,
JONNA STÅLRING, and SCOTT BOYER

12.1 INTRODUCTION

For a compound to pass approval as a drug product, it is not enough to show efficacy, the compound also has to have an acceptable safety profile. The key goal of the pre-clinical safety package done within drug discovery and development programs is to be able to understand the relationship between a compound and its potential adverse effect in humans, and thereby being able to avoid them. The preclinical program is heavily driven by the regulatory requirements for obtaining Investigational New Drug status before clinical trials can begin and to assure the safety of the volunteers. It involves a multitude of tests, some being mandatory, with complexity ranging from preclinical *in vivo* studies, tissue- and cell-based, as well as biochemical *in vitro* assays. There are also requirements involving computational modeling as any potential genotoxic impurity (PGI) that is associated with manufacture of the drug are required to be assessed using such tools. With safety remaining a common reason for compound attrition [1], exploiting the data generated throughout the discovery and development process to its fullest is of utmost importance to ensure that the most promising candidates are progressed at each stage. In this context, computational techniques and informatics are becoming increasingly important.

However, attempts to relate a compound's toxicity to its chemical properties is far from novel. As early as 1868, Crum-Brown and Fraser investigated the potential of a set of alkaloid salts to cause narcosis, and how this varied with the chemical composition [2]. Work by Fujita and Hansch [3], relating a set of physicochemical properties of a compound to the biological activity using multiple linear regression and the introduction of Free-Wilson analysis, have paved way for much of the computational chemistry employed today. Computational tools are nowadays heavily integrated into the whole process from bioinformatics-based target assessment and high-throughput safety filters used in hit identification [4], all the way to screening for PGIs postmarketing [5]. These methods are highly attractive because of their speed and relative low cost compared to traditional *in vitro* and *in vivo* experiments. Since both the number of compounds tested and the complexity of the experimental output vary dramatically at different stages of drug discovery and development programs, the relevant safety issues that can be addressed vary accordingly. To support this broad scope of modeling, a wide variety of tools and methods are employed, and awareness of the benefits and shortcomings is imperative for efficient and appropriate use. This awareness is not limited to technical details alone, but also to the understanding of the available toxicological data and the endpoint in question. Conceptual understanding is imperative because of the multitude of ways in which a compound can exert a high-level toxic effect (e.g., liver necrosis), be it either via a specific reactive chemical feature of the compound-like reactive metabolite formation in diclofenac [6], a primary or secondary pharmacological effect, or through the interaction with a biological pathway [7]. Another important factor to consider is the availability of relevant safety data and the quality of that data. For certain well studied targets, for example, the hERG (human Ether-á-go-go-Related-Gene) channel, there are tens of thousands of compounds tested in a rather standardized fashion for inhibition, whereas other targets have much less chemical coverage and incomparable experimental setups highlighting the challenge of aggregating even this type of low-level data. Furthermore, looking at higher level data, other important questions arise concerning the difference in severity of the toxic finding, at what dose/concentration the effect was seen (i.e., Is it of clinical relevance?) and in what species. Finally, there may be only a handful of compounds showing a specific finding in pre-clinical studies. It is clear from this discussion that the area of safety modeling is highly complex, and care has to be taken both in collecting relevant data and in selecting the appropriate modeling approach.

Due to the aforementioned discrepancy in data availability (especially relevant to translation of toxic effect) and the fact that many clinical endpoints are multi-mechanistic, it is important to stress that each computational step should be well defined and model small steps, for example, a traditional quantitative structure-activity relationship (QSAR) approach based on chemical structure is probably relevant to distinguish hERG binders from nonbinders, but not relevant to model a small set of diverse compounds associated with a complex endpoint such as drug induced liver injury (DILI). A second important factor to consider when constructing *in silico* safety models is the intended use of the model, and the potential cost associated with false positives versus false negatives from the model. For instance, there is zero

tolerance for genotoxic findings in PGIs, requiring the false negative rate to be kept low, whereas screening for potential off-targets or screening out genotoxic risks for library design purposes might be more accepting. The rest of this chapter will illustrate how AstraZeneca utilizes all the available data to provide safety modeling at all stages of the drug discovery and development and how we try to ensure that these models are used appropriately in our projects.

12.2 DATA-DRIVEN MODELING

12.2.1 Linking Chemical Structure to *In Vitro* Results

The scientific strategies for avoiding safety issues during drug development dictate how computational tools can best be used. Early experimental work is performed with *in vitro* assays used as indicators of potential toxicity that might arise in preclinical *in vivo* studies. These assays are in some cases standardized within the industry (e.g., the Ames assay) based on regulatory requirements, but they may also employ quite varied experimental protocols if they are not part of a mandatory submission package. As the standardized *in vitro* assays are applied in a wide range of projects, the generated datasets are commonly chemically diverse, providing a solid foundation for model building. Conversely, most assays are not standardized, and only a limited set of compounds are tested, producing smaller data which may need to be supplemented with commercial and public data. This poses challenges to any informatics approach concerned with combining and analyzing the data with questions such as, how should IC_{50} values be treated versus K_i values? How should binding data and functional data be compared? Can activities reported from different cell lines be combined? Therefore, to be able to select the best method to model the data, great care and diligence is required in data collection and aggregation, as well as tailoring the computational method to the nature and quality of data, depending on the intended application.

12.2.1.1 Safety QSAR Modeling of *In Vitro* Endpoints As part of the internal strategy to minimize costs and improve compound quality, *in vitro* datasets are used to design computational filters to assess safety properties even before a molecule is synthesized. The filters are usually based on the identification of potentially liable fragments related to the endpoint in question (structural alerts, which will be discussed in more detail later in the chapter) or QSAR models. Such filters are mainly used for library design and experimental prioritization. QSAR models incorporating large, diverse datasets are often referred to as “global” models [8] differentiating them from models built on congeneric series. It is inherently challenging to build such models due to both the vastness of chemical space and the multi-mechanistic nature of most safety assay responses. To assure the greatest possible model applicability, as much data as possible is incorporated into a global model. As new experimental data become available, it is beneficial to continuously monitor the performance of the model [9] and to expand the coverage of chemical

space by automated model updating protocols where all new data are retrieved and used in an automated model rebuilding and validation process on a regular basis. Internal validation strategies such as cross validation and hold-out validation complement external and temporal test sets in the assessment of the generalization accuracy.

12.2.1.2 QSAR and Machine Learning in Practice The early QSAR models were mainly based on derived physicochemical molecular properties, used in conjunction with multiple linear regression. However, the often nonlinear nature of the relationship between a molecular property and a biological response is generally recognized [10, 11], and elaborate machine learning (ML) methods, such as artificial neural networks, support vector machines, instance-based methods, decision trees, and ensemble techniques such as Random Forest, are increasingly applied and are the methods most commonly used in AstraZeneca.

Independently of how robustly a QSAR model is built and validated, it is important that the user is aware of the applicability limitations of the model in terms of confidence in a prediction and the intended usage. Models designed as coarse filters for library purchasing and library design would, for example, usually not be very useful for predicting small changes in activity within a structurally similar series. In terms of confidence in a prediction, this relates to how well the query compound is represented in the model and assessing this in the best possible way is a very active and debated area of research related to QSAR modeling [12–14]. Multiple methods have been proposed, relying on fundamentally different information such as descriptor value distributions [15], instance density [16], near neighbors [17, 18], and model perturbations [19]. Within the QSAR community, the near neighbor-based methods, also referred to as “distance-to-model” or “leverage,” have been the most popular [20] but as of now, there is no consensus on how this is done in the most reliable way. It is, however, extremely important for the user to understand how reliable a prediction is for each particular compound both when a model is used alone but especially when a model prediction is combined with other types of information, which will be discussed in Section 12.2.3.

12.2.1.3 Promoting the Mechanistic Understanding Despite their recognized importance in accounting for nonlinear dependencies between descriptors and the response, nonlinear machine learning algorithms are often criticized for providing no interpretation or understanding of the underlying reasons for a particular prediction. This is however not entirely true. In addition to statistically ranking descriptor importance in the training set of the model, specific mechanistic insight for individual predictions can also be obtained by calculating numerical partial derivatives of the predicted response with respect to each descriptor [21]. The partial derivatives describe the characteristics of the model in the local neighborhood at the point of the prediction. For models based on structural fragments, the greatest numerical partial derivative signifies the structural fragment with the highest contribution to the model prediction [21]. This is illustrated in Figure 12.1.

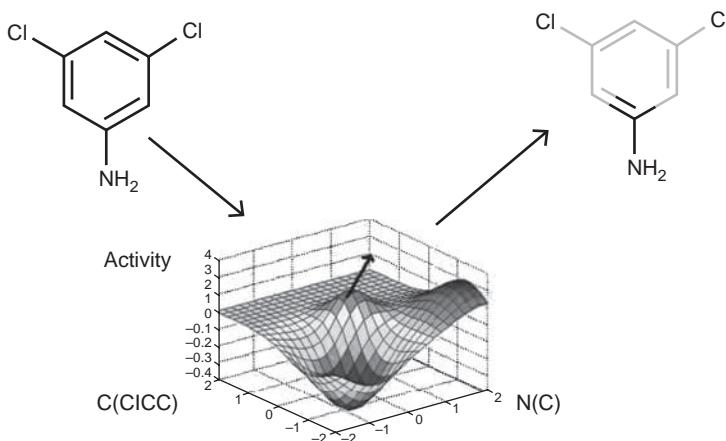


FIGURE 12.1 Schematic representation of how the descriptor influence on each prediction can be extracted from the models response surface. The most significant descriptor fragment can then easily be highlighted in the structure for visual inspection. For color details, please see color plate section.

In practical applications, the user is provided with the molecular structure image of the query compound where the most influential fragment is highlighted. This method of identifying significant substructures has been successfully applied to decision support systems within AstraZeneca in a few different contexts. In cases where the model predicts the compound to have an unfavorable property, the method is able to present suggested structural modifications to the compound that would improve the situation. An extension of this method, useful in *de novo* design, where a number of compounds are automatically modified to produce a series of new compounds with the fragments contributing the most to unfavorable properties replaced with fragments that would contribute to better compounds, has been described by Helgee et al. [22]. The method is there applied to compounds predicted to be active in the Ames test being altered to produce compounds that are instead predicted to be inactive. As illustrated, this method is very powerful as it not only provides the user with an answer, but also provides an insight into why the compound is assumed to have the predicted activity and a suggestion of how the situation can be rectified, if this is desired.

12.2.1.4 Reducing the Risk for Cardiac Liabilities Using a QSAR Filter Applying a QSAR model can have significant impact if it is implemented in alignment with a scientific strategy, such as that related to cardiac safety. This is governed by the regulatory authorities and has one part relating to monitoring of the potential interaction with the voltage gated hERG potassium channel. This stems back to the withdrawal of cisapride [23] and terfenadine based on their inhibition of hERG resulting in cardiac arrhythmia (torsades de pointes) due to QT interval prolongation. The pharmaceutical industry has as a consequence modified the preclinical safety assessment strategies for cardiac liabilities by the development of

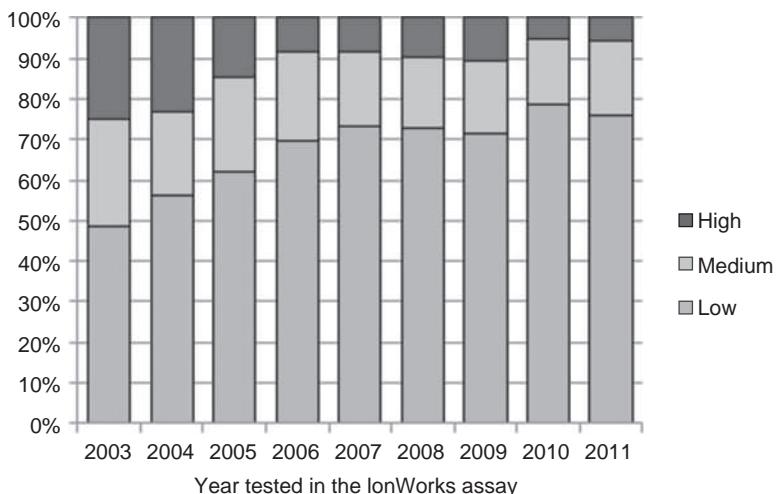


FIGURE 12.2 Histogram of compounds tested for hERG activity in the IonWorks assay from 2003 to 2011, categorized as having “low” >30 μM (green), “medium” 3–30 μM (amber), or “high” <3 μM (red) activity. For color details, please see color plate section.

high throughput *in vitro* functional cardiac ion channel assays such as the patch clamp assay developed by IonWorks Inc. [24]. The assay has been routinely applied in early discovery at AstraZeneca since 2003, resulting in a hERG dataset currently encompassing around 50,000 compounds. As part of the same cardiac safety strategy, a computational filter in the form of a QSAR model built using the internally tested compounds has been in use to prioritize compounds for experimental testing since 2003. This model is also used for library design, hit selection, and chemical design.

Implementation of a mandatory computational filter, together with the accumulated intuitive understanding of the hERG SAR among chemists, has reduced the fraction of liable compounds tested in the IonWorks assay, as displayed in Figure 12.2. In particular, the fraction of compounds experimentally categorized as having “high” activity has decreased from ~25% to 5–10% between 2003 and 2011 showing that significantly fewer compounds are being synthesized that could induce potentially lethal cardiac arrhythmias in patients.

12.2.2 Chemical Structure and Using All Available Data

There are a multitude of ways to combine data synergistically to maximize information content and to use otherwise low resolution data not suitable for quantitative modeling. In this section, methods other than QSAR used to relate chemical structure with activity are described and examples are given of applications using such methods. The section continues with discussing the combination of data to enable an overall assessment of risk based on all the underlying pieces of information.

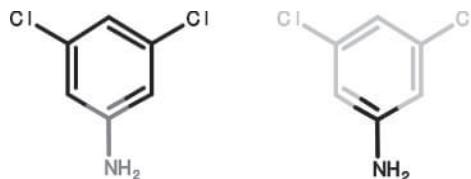


FIGURE 12.3 A general aromatic amine alert highlighted in red on the left. The amine is in general correlated with positive activity in the Ames test. However, when combined with the chlorine substituents in the meta positions, which are correlated with inactivity in the Ames test, the overall assessment is that the compound would be inactive. For color details, please see color plate section.

12.2.2.1 Structural Alerts One of the first toxicological endpoints in modern times to be systematically investigated in a structure–activity context was mutagenicity in studies by Ashby and Tennant [25, 26] with the aim of avoiding carcinogenicity. They successfully identified substructures that frequently occurred in compounds that were mutagenic and the term “structural alert” was introduced. Since then, this has been a very common approach for the applications of chemically based toxicity, that is, toxicity that is related to reactivity of the compound. There are several commercial systems (e.g., Leadslope [27], MultiCASE [28, 29], and Derek [30]) utilizing this concept. The benefits of using structural alerts are that it is usually intuitive for the user to understand the alert and the alerts themselves are often based on a mechanistic understanding of the chemical reactivity of the compound. A drawback is that a lot of focus is usually placed on active compounds with the assumption that the lack of an alert is a prediction of the compound being inactive when it could just as well be lack of information, and absence of evidence does not imply evidence of absence. Structural alerts are very useful in cases where there is not enough high quality data to build a robust quantitative predictive model. There are several different ways of deriving structural alerts, the most common being based on subjective mechanistic understanding of which parts of a structure are responsible for the observed toxicity. These can be tailored according to how much of the SAR can be supported with experimental evidence and can be as general or specific as desired. It is, however, also beneficial to systematically mine datasets to extract substructures correlated either positively or negatively with the observed activity [31]. This has the benefit of not being biased based on the user’s prior experience by providing a way to split molecules into substructures (not necessarily based on functional groups) in order to see how each part is correlated with activity. To illustrate, a molecule contains an aromatic amine, which is a well-known structural alert for mutagenicity (Figure 12.3) [32], with further inspection of the substructures in the molecule, we find that the phenyl ring that the amine is bound to also has chlorine substituents in the meta positions. The chlorine substituted phenyl in itself constitutes a substructure that is negatively correlated with mutagenicity and the overall decision on whether the aromatic amine is a valid alert will have to take into account the environment it is situated in. In this case, the overall assessment is that the compound is inactive in the Ames test which is also the result that is found experimentally. The manually derived structural alert

implies that the compound should be predicted active but applying a statistical algorithm to compare the prevalence of activity for the two substructures allows the correct conclusion that the compound is actually inactive, to be drawn. This type of application resembles using structural descriptors in a traditional QSAR model but in a very intuitive way.

12.2.2.2 Read-Across Another suitable method for utilizing low-resolution data is “read-across,” where we draw conclusions for one compound based on experimental results from a different compound based on their chemical similarity. Of course, this has to be done with care and requires informed intellectual input from the user but is very beneficial and can be applied both across endpoints and across species. Looking at experimental data from structurally similar compounds will usually allow the user to start thinking about SAR and ideas for the modifications to virtual compounds in order to enhance properties.

12.2.2.3 Predictive Secondary Pharmacology In addition to building predictive models of *in vitro* data, it is also common to mine large sets of compounds for structurally similar compounds that have reported bioactivity against a high number of targets such as kinases, proteases, cytochrome P450s, G-protein coupled receptors (GPCRs), and so on. The data are rarely of high enough quality to build robust predictive models and the aim of usage is also different. Statements suggesting that about 75% of all adverse drug reactions (ADRs) are dose dependent and predictable from preclinical safety pharmacology studies [33] underpin the importance of investigating secondary pharmacology effects that could potentially result in side effects. This work should be done early in discovery and encompass appropriate assessment and mitigation of any identified risk, as described in an internal application example further down. Within AstraZeneca, an approach termed predictive secondary pharmacology (PSP) has been developed for the purpose of mining data derived through systematical identification and generalization of all possible compound related off-target effects (*in vitro* bioactivity data) based on public and internal data sources. Combining such vast amounts of data collected using different protocols, stored in different formats and with varying experimental quality involves challenges in data handling. Cutoffs for when a compound is considered active may vary between data sources and requires standardization which may or may not be feasible. This might result in restructuring data into classification categories instead of continuous values or other compromising treatments.

PSP is based on the assumption that similar compounds have a similar bioactivity profile [34–36]. This method is conceptually distinct to the QSAR approaches detailed before since the current implementation does not aim to predict an endpoint specific activity (see the earlier hERG example) but is rather designed to analyze all available target data for all available compounds. Hence, the challenges are different for a PSP search compared to those for QSAR modeling. Similarly to QSAR models, the relevance of the retrieved near neighbors depends on the choice of descriptors to encode chemical information, as well as the quality of the (often) aggregated activity results and how they are standardized. Not less challenging is the fact that secondary

pharmacology analyses often deal with sparse compound datasets, that is, not every compound is tested against all targets, and large potential compound target interaction sets. When successful linking of target profiles to ADRs can be done and conclusions for compound specific side effects are drawn, the relevance of the results also depends on the ADR ontology. A secondary pharmacology analysis is usually an integral part of a risk assessment and is the one part that more than any other, requires cross-functional (toxicologists, chemists, and safety pharmacologists) expertise to fully interpret the (sometimes considerable) model output.

There have been several publications demonstrating that ADRs can be predicted from a chemical structure via preclinical safety pharmacology targets. For instance, Krejsa et al. [37] showed that it was possible to predict ADRs for benperidol employing only *in vitro* similarity. In a similar fashion, Fliri and coworkers used hierarchical clustering of *in silico* profiles to link compounds, and saw that compounds having similar bioprofiles also tended to have similar side-effect profiles [38]. These efforts required the experimental profile to be known, and therefore several attempts to predict such profiles *in silico* have been performed. For instance, Bender et al. used ECFI fingerprints and a Naïve Bayes classifier to link chemical fragments to target activities, which were in turn used to generate predictive models for adverse drug reactions [39]. Thereafter, an interesting extension to this approach was introduced by the same group combining knowledge-based annotation of ADRs to chemical structures and knowledge-based annotation of pathways [40]. Another interesting approach, known as similarity ensemble approach (SEA), was presented by Keiser et al. [41]. By calculating the chemical similarity between known target ligands and adaptation of the BLAST algorithms they relate targets and compounds. Mestres and coworkers have in a series of publications also shown that target profiles can be predicted by using distance weighted similarity metrics, and they have also shown that drugs connected by similar side effects share affinities by multiple targets [42, 43].

Applying PSP to a query compound comprises a search against “ChemistryConnect” [44], an internally aggregated data warehouse, in order to identify chemically similar compounds (target molecules) and their bioactivity profiles on multiple targets. The PSP approach facilitates three different types of searches: (1) an exact match search (retrieving data only for the specified query molecule), (2) a similarity search (based on a 2D fingerprint and the Tanimoto similarity coefficient) to retrieve similar compounds, and (3) a substructure search to retrieve compounds sharing a defined chemical substructure. In addition to the bioactivity profile, any known ADRs for the near neighbors are retrieved.

The following example of a PSP analysis may show the reader the benefit of such investigations. A potent compound from a class 1 GPCR project was subjected to a routine PSP search (Figure 12.4). A similar external compound was identified, which had high reported activity on the Histamine 3 (H_3) receptor [45]. As activity on H_3 is related to insomnia as a potential side effect, the internal compound was tested in an H_3 binding assay where it was active, followed by a functional assay where it turned out to be a potent H_3 receptor antagonist (<100 nM). As a result, H_3 was introduced as a standard target to screen for all new compounds synthesized in this project.

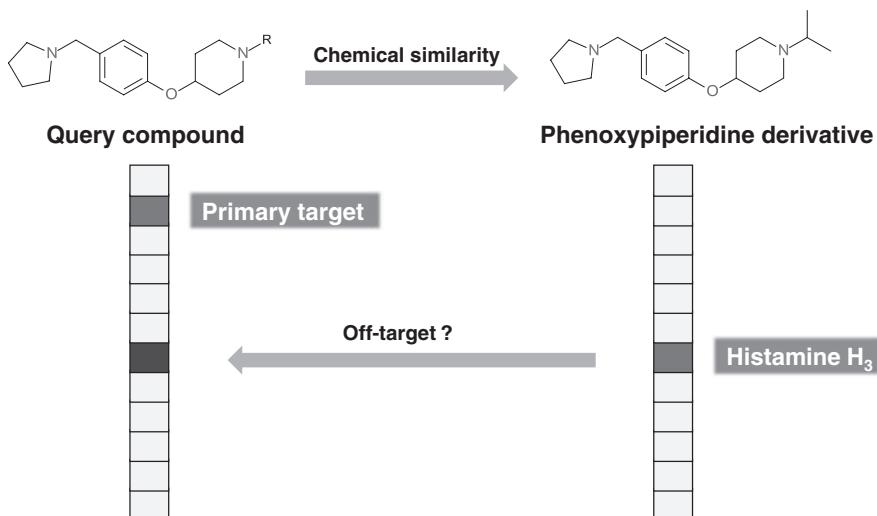


FIGURE 12.4 Concept of predictive secondary pharmacology exemplified with a query compound in the upper left corner. The phenoxyphiperidine to the right was one compound identified by the similarity search. The vertical bars denote the compound specific target profiles. For color details, please see color plate section.

In addition to safety risk assessments, a secondary pharmacology analysis can be carried out for problem-solving purposes, where a compound that failed at some stage of discovery due to observed side effects can retrospectively be analyzed to see if the secondary profile can explain the observed ADRs. If a sufficient number of similar compounds with test data are available, it might be possible to link the adverse findings to a substructure in the molecule. PSP analyses are therefore also suited for drug repurposing as it is obviously as easy to look for desired or potentially useful off-target interactions as it is to look for undesired ones.

12.2.2.4 Utilize Everything: Introducing Warning Systems In our view, the optimal way of using all the available data for a specific type of toxicity is by combining all the different ways of analyzing and outputting data into what we refer to as “warning systems.” These are designed to be as comprehensive as possible and the AstraZeneca Genetox Warning System (GWS) [46] is a good example. The GWS accesses an internally built database containing internal and external genetic toxicology data from the Ames mutagenicity, Mouse Lymphoma (MLA), Micronucleus (MN), Chromosome Aberration, Comet Assay, and Carcinogenicity tests. The database contains ~12,000 compounds of which ~3000 are proprietary to AstraZeneca. When a user submits a compound, a series of events occur including checking if the compound already has any experimental results. The system will run the query compound through an Ames QSAR model and display the substructure most significant for the prediction (this feature was described in Section 12.2.1.3), it will identify chemically similar compounds experimentally tested in

any genetic toxicology assay (read-across) and identify any structural alerts present in the query compound. If the query compound contains any structural alerts, the underlying compounds for the alert in question will be displayed together with their activities to enable the user to understand if the alert is relevant in the chemical context of the query compound. The GWS system was the first computational tool in AstraZeneca to replace an *in vitro* assay as part of the scientific strategy. Historically, a SOS/umu test was used prior to the Ames assay but it was found that the computational model was more predictive towards the Ames assay than the SOS/umu test was. Needless to say, it is also much faster and cheaper.

This type of system can be, and is in fact used in all phases of drug discovery and development. In the early lead identification (LI) phase, it can be used to risk assess and compare potential lead compounds and later, lead series. In lead optimization (LO) and prior to candidate selection it is crucial to look at all the data available and think ahead of the coming experimental testing strategy. The benefit of using the GWS in this context was very well demonstrated in a recent internal LO project. The most promising compound had been selected and had already been run through an Ames test. As the compound showed no Ames activity, indications were that there would be no genotoxicity liabilities. However, when examining the available data from other assays for structurally similar compounds, it was clear that although there was no Ames activity, there were other compounds that had showed activity in the MLA test, despite being Ames negative. As a consequence, the project decided to frontload the *in vitro* MLA test and the *in vivo* MN and Comet assays for their potential candidate to avoid unnecessary experimental work on the compound. In a similar way, one can use the system to determine if potential metabolites (e.g., aromatic amines) have been experimentally tested already or what their predicted activity would be.

As mentioned earlier in this section, depending on how much and what quality data one has access to, the unique format of a warning system will vary. Reactive metabolites are another area where structural alerts have been used extensively. Reactive metabolites can arise in several different classes of compounds but have in common that they are formed through bioactivation of the parent compound, resulting in an intermediate species that is reactive with the propensity to bind to macromolecules or DNA *in vivo* or in humans. They are extremely problematic for the pharmaceutical industry as they are often idiosyncratic in nature and as such are identified in the late stage clinical trial unless this risk is mitigated through deliberate strategies in discovery and early development. Such strategies pose the problem of drawing a link between an *in vitro* and an *in vivo* trapping experiments and actual clinical toxicity. As far as computational approaches, structural alerts are often used to highlight the structural moieties that are known to have the propensity to be bioactivated such as quinone methides. Similarly to the GWS, we have built a Reactive Metabolite Warning System (RMWS), where we include the different assays used to trap reactive species (glutathione, methoxylamine, cyanide, etc.). For this endpoint, we are not able to build any robust predictive model and use only structural alerts together with site of metabolism predictions based on the MetaPrint2D software [47, 48]. The user is able to assess not only the highlighted structural alert but also how this overlaps with the most probable sites of metabolism with the assumption

that there is a lower risk of reactive metabolite formation if there is no overlap. The included alerts have all been derived from either public or internal compounds exerting toxicity where reactive metabolite formation is known to occur. This is a quite simple form of warning system and the intended usage is consequently of a simpler nature than GWS. The RMWS is a hazard identification tool that can be followed up by an earlier *in vitro* assay.

A third warning system to exemplify how data sources can be combined is the Testicular Toxicity Warning System (TTWS). Testicular toxicity is the sixth most common target organ toxicity causing project delay within AstraZeneca [49]. There are probably a multitude of mechanistic reasons for testicular toxicity. Testicular toxicity is traditionally not detected until the pivotal 1 month rat and dog *in vivo* studies are performed, and as such, there is a strong need to highlight any potential risks using informatics as early as possible. It has been shown recently that an *in vitro* RAR α antagonist assay is able to identify a certain class of testicular toxicants and this has now been incorporated into the regular screening strategy and these data are also modeled and included in the TTWS as a predictive model. In addition, the system includes manually curated *in vivo* data, initially retrieved using text mining, and structural alerts. The user is presented with any experimentally tested (*in vivo* or *in vitro*) structural near neighbors and also any structural alerts that are present in the query compound. Similarly to the previously described warning systems, structural alerts have been derived based both on mechanistic understanding and on automatically identified substructures that are statistically correlated with activity.

All the warning systems have in common that they are as comprehensive as the available data allow. They exemplify how you can combine *in vivo* data, *in vitro* data, mechanistic understanding, or very limited understanding in a way that is useful and can aid decision making or risk assessment and mitigation. In some cases, they may not be able to present the user with a prediction of activity for virtual compounds but they may prevent the projects from repeating past mistakes, which in itself is a significant saving in terms of time and money and fulfills the need to make responsible use of human and animal safety data.

12.2.3 Combination of Evidence

Generally, methods and data are combined for an endpoint such that a multitude of outputs are obtained, as exemplified with the different warning systems. This of course ensures that all available data is utilized, but at the same time raises the question of how to interpret and analyze the different outputs. Interpreting each outcome separately may result in conflicting evidence and objectively making an overall assessment from a multitude of outputs is not an easy task for the user. To address this issue it is desirable to obtain one single answer based on and describing, the combined evidence using all available outputs. When combining the different types of information, the confidence related to each prediction will have a bearing on the overall assessment and should be weighted accordingly.

Combination of evidence is the mathematical formulation that allows a combination of different sources of evidence to a single “predictor” under certain

assumptions. Enabling combination of evidence requires a weighting of all sources of information, to obtain a comparable measure. There are several ways that this can be done, all with different implications and the task is often referred to as weight of evidence (WoE). Traditionally, obtaining this overall estimate has been performed using weighting and consensus approaches, such as average or majority voting [50], or where possible, Bayesian probability theory [51]. The voting approach may be sufficient when combining data sources of similar types and predictive ability, which is not the common situation. Dempster–Shafer theory (DST) [52], which can be interpreted as a generalization of Bayesian probability theory [53], was applied for combining evidence. In contrast to Bayesian modeling requiring well-defined probability distributions, DST is concerned with “masses” associated with an outcome or a subset of outcomes. This approach provides a lower bound (belief, i.e., total mass of all outcomes that imply the predicted outcome) and an upper bound (plausibility, i.e., the sum of all masses that does not contradict the predicted outcome) of the probability of each outcome. By explicitly including the “unknown state,” interpreted as any possible outcome, DST provides an intuitive way to work with varying levels of confidence in evidence in a rigorous manner.

WoE using DST is used in the GWS and has also previously been successfully applied when combining evidence from processed electronic healthcare records and text mining of literature data for the EU-ADR project [54]. In the GWS, all the outputs are combined with their individual weights. In practice, the components to be combined are: the predictions from a QSAR model weighted by the accuracy of the model with respect to the predicted class (active/inactive), the structural alerts where each alert is weighted by the fraction of the number of actives over the total number of hits in the experimental database that match the alert, and the presence of structural near neighbors, weighted according to their experimental activity. The mathematical combination of the components provides an overall assessment that considers all underlying uncertainties and gives the user a clear understanding on how reliable the assessment actually is.

DST provides a means to effectively combine “evidence” into a single predictor in a transparent way and can be applied to all the different types of evidence that are encountered in the applications described in this chapter. The different stages in the process are easily visualized giving the user an opportunity to trace the evidence and uncertainty back to the individual sources. This brings two benefits: one is to enable a more complete risk assessment based on all the data available, the other is to provide the user with as much information as possible concerning the reliability of that assessment, also including information on the possible source of error for that assessment.

12.2.4 Focusing on Biological Data

12.2.4.1 Making Use of Biological Data

Up until now the chapter has focused on relating a chemical structure to certain toxicological endpoints. However, there is also a growing amount of biological data available for safety modeling both within industry and the public domain. For instance, large investments have been

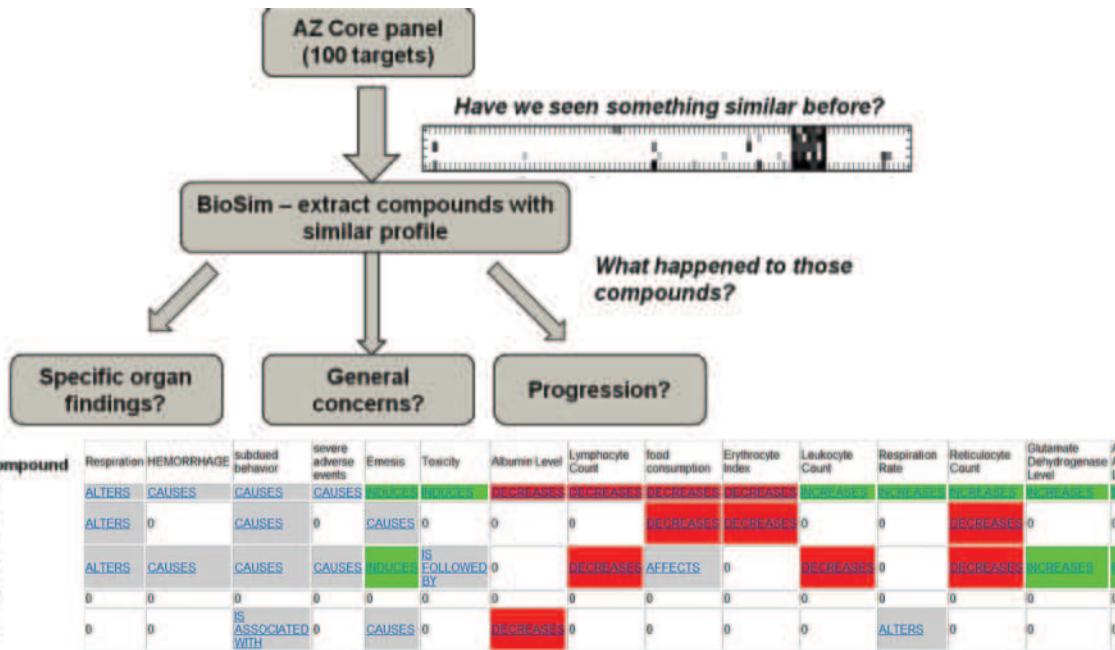


FIGURE 12.5 Schematic overview of the BioSim application. After a compound (A) has been profiled in a broad pharmacology, its profile is compared to all other tested compounds to retrieve the most similar profiles (originating from compound B–E). Among other information, shared *in vivo* findings are highlighted (with links to original documents). For color details, please see color plate section.

made to generate broad chemogenomics databases (e.g., Bioprint [37], ChEMBL [55], and Pubchem BioAssay [56]), and work is ongoing to structure preclinical animal studies to make them amendable for modeling studies [57]. These efforts will allow for new ways of relating a compound to a toxic effect.

12.2.4.1.1 Making Use of *In Vitro* Profiles As stated earlier, the pharmacological effect of a compound is to a large extent dependent on the compound's interactions with different target macromolecules. It has become clear that most compounds interact with several different targets and the term polypharmacology is frequently used. There are two major aspects of modeling associated with polypharmacology; firstly, it is of interest to try and predict all potential targets for a specific compound and secondly how do these interactions relate to the pharmacological or toxicological effect seen. The first aspect was addressed while discussing the PSP strategy within AstraZeneca, and this section will focus on the latter. To increase understanding of the polypharmacology of compounds in discovery and development programs, both pharmaceutical companies, contract research organizations, and academia (Bioprint [58], DrugMatrix [59], and ToxCast [60]) perform broad *in vitro* screening on a regular basis. It has been shown that the generated *in vitro* profiles contain additional information compared to regular chemical fingerprints commonly used in QSAR modeling, and they have been used both on their own and in conjunction with structural information to build both models of both target endpoints and high-level endpoints [61]. Increasing attention has been given to the observed patterns of interactions, and pathway perturbation as opposed to single target activity [43, 62]. At AstraZeneca, secondary pharmacology screening today comprises a set of *in vitro* assays, covering over 90 targets representing all major target families. The fact that a large collection of compounds have been tested in a standardized fashion makes this an information rich set for relating secondary pharmacology to *in vivo* responses. To tap into this resource a similarity tool named BioSim has been developed (manuscript in preparation). BioSim (Figure 12.5) relates compounds to each other based on the similarity of their profiles in the broad pharmacology panel and presents the identified compounds together with known *in vitro* safety, DMPK, and *in vivo* data. This allows for a good overview of the known potential secondary pharmacology issues associated with a specific profile. By focusing on the pharmacological similarity, it becomes easier to hypothesize about the pharmacological origin of certain observations. This provides a more comprehensive view of the observed secondary pharmacology and how it might translate into *in vivo* effects. There are, of course, also major challenges in making this link. Comparison of *in vivo* findings is complicated by nonstandardized *in vivo* protocols and reporting of findings. Especially the identification of confirmed true negatives poses a challenge due to the fact that not all compounds have been tested in the same *in vivo* setting and that a severe toxicity finding might "shadow" other findings that would be present at higher exposures. Although primarily derived from a safety perspective, this approach is also useful in making chemically unintuitive links between projects sharing target profiles, relevant for both drug repurposing and idea generation.

12.2.4.1.2 Exploiting Preclinical Data Preclinical *in vivo* data are the most expensive data generated before a compound reaches clinical development but also the most information rich with respect to translating to human toxicology. Although there is a large concordance between preclinical and clinical findings for defined endpoints with clear preclinical counterparts [63], there are many clinical endpoints that today lack established preclinical signals. One of the reasons that this vast data source is rather underexploited is that this data has until recently not been well structured for safety modeling. To change this situation several companies have launched initiatives to make this data more easily available to be able to leverage present knowledge into future projects. There are also joint efforts such as the Innovative Medicines Initiative (IMI) eTOX project [57] that aims to generate a large database of preclinical legacy reports to provide data supporting chemoinformatic and bioinformatic modeling of preclinical safety endpoints. Such efforts make it possible to identify patterns of findings that can be used for translational research as well as for generating mechanistic models for the preclinical findings.

One recent example of how this type of data can be used is the identification of gastrointestinal preclinical side effect profiles that could indicate a clinical risk of nausea [64]. Despite being a very common finding in both clinical studies (seen in ~30% of Phase I studies) and post-marketing, there is no single commonly accepted preclinical marker. By collecting all the gastrointestinal preclinical observations seen for a set of 86 marketed drugs and reshaping them into a binary vector representation, it was possible to cluster drugs based on the similarity of their preclinical profiles. Interestingly, the generated clusters largely corresponded to the presence or absence of clinical nausea, although this information was not part of the clustering. The cluster generation was also validated using a set of 20 AstraZeneca development compounds having reached Phase I. Based on the clustering a prediction accuracy of 90% was achieved in this balanced and blinded set, indicating the general applicability of the nausea profiles. To inspect the generality of this approach it was also applied to cluster compounds associated with dizziness from nondizziness compounds. Preliminary findings show that the profiles do contain information that can help separate the high risk compounds from those with reduced risk.

To speculate about future directions for this line of research, the generated clusters might not only be able to discriminate between clean and risk compounds. Since these endpoints usually are multi-mechanistic, the clusters might also provide insight into specific mechanisms underlying the observed preclinical phenotype, common for the members of that cluster. From that information, one can hypothesize and potentially find targets or chemical features that can be evaluated *in vitro* or *in silico* to help design away from the final clinical risk. Considering the now well-known mechanistic endpoints, it almost always started with the identification of a set of compounds that displayed a certain *in vivo* phenotype. The more carefully the *in vivo* “toxicology” phenotype was described, the easier it was to decide which chemical structures could be included in the “active” group. This type of high-level analysis will help feed back information to levels that have more available compounds, allowing for more quantitative modeling and filter generation for early risk detection. And

most importantly, doing it this way will substantially increase our knowledge about drug safety since it will provide us with testable hypothesis.

12.3 DELIVERING IMPACT: BRINGING IT TO THE CUSTOMER

Everything discussed so far becomes useless if the information is not easily accessible for the projects and efficiently used to enhance the quality of propagated compounds. Hence, the user interface needs to be well structured to address project concerns and the systems must be able to deliver the information and predictions to the end user in a timely fashion. It is therefore of utmost importance that the systems are robust and kept updated.

12.3.1 Technical Solution

An important aspect is that of data and model stability. The raw data, for example, assay data, need to be updated regularly as new results become available. This process is tedious and time consuming and once a valid protocol for data acquisition and validation has been established, it is suitable for automation. The automatic procedure requires extensive data integrity checks (e.g., do all data points have valid structures and experimental results, how to handle conflicting results from different experiments) and a formalized automated normalization process. This involves answering questions such as how to handle isomers, do experimental *in vivo* results have precedence over *in vitro* results or should all results be shown, should more confidence be put in results from Good Laboratory Practice (GLP)-studies, and so on. The models are then automatically rebuilt using the updated data and auto-updated as was mentioned in the QSAR section. A model may be promoted to use in the production system if it passes some defined validation tests, specific for each model, which was also mentioned earlier.

At AstraZeneca, this system relies almost exclusively upon internally developed and Open Source codes. For example, many QSAR models are built using the AZOrange package [65], which is an internal development of the Orange [66] ML platform. The Open Source foundation reduces license costs and gives full access to the source code, facilitating further development. AZOrange, specifically tailored to meet the requirements of QSAR modeling, is in turn made available to the Open community through a public repository.

The development should follow common practices in software development, such as keeping the source code in a revision control system [67], automated deployment, using automated tests, using a bug tracking system (can also handle feature requests and assist in release planning), and having accurate documentation integrated with the code. The agile programming practices (such as extreme programming [68] and SCRUM [69, 70]) are specifically developed to work in an environment of constant changes in requirements and are hence very suitable in a research setting. It is very costly both in time and money to introduce a change in a software that is about to be shipped or in production use because it would require

extensive testing to make sure the code change will not break other parts of the software. That problem is addressed by the agile methods by the use of automated testing or even Test-Driven Development (TDD) [71], where the tests for a piece of functionality are written before the code that implements it. When these tests are in place, it is much easier to make changes without introducing bugs in other parts of the code because the tests will expose such bugs.

As mentioned earlier, automated tests are an integral part of development of the systems. They come in the range from unit tests of a single function to whole-system tests. These tests are run before any piece of code is checked into the revision control system. A continuous integration [72] server can also be set up, that continuously checks out code from the revision control system and runs all the tests to detect errors at the system level. If a bug is found, regression tests that expose the bug are written before the bug is corrected. The regression test is then included in the test suite that are run at regular intervals, to detect if the bug reappears.

Production systems need to be monitored continuously so that system administrators are notified if a system becomes unresponsive. It is often advantageous if the systems are set up in a modular fashion as independent services that can be combined and visualized together or independently. Integrating different services is much easier if open, implementation agnostic standards, such as SOAP [73] or REST [74], are used. Since third party software packages are frequently used, and sometimes required by regulatory authorities, it is often useful to wrap these in services to make it easier to integrate these. This is easier if the third party software provides an Application Programming Interface (API) or at least a command line interface.

12.3.2 Facilitate Usage

To meet the varying needs of different categories of users it is important to provide the same information but with tailored views. The project toxicologist may be interested in different parts of the information compared to scientists in pharmaceutical development. Users may also be accustomed to specific applications or ways of viewing the information, so it may be good to integrate the predictions in the applications they already use. This implies that systems should be developed in a generic fashion to enable easy access via several interfaces. This ensures that it does not matter which interface is used to access the information, these are always extracted from the same underlying system and will always provide the same answer. It is vital that users do not get different answers to the same question depending on which interface they access. The results can be subsets of the entire information available, but they should never be conflicting. This can otherwise happen when databases or predictive models are updated at different time points, for example. At AstraZeneca, most of the prediction services have been integrated into several of the tools users use in the various phases, bringing the predictions to their fingertips no matter how they like to work [75]. The modular service-oriented architecture brings benefits in this context as well.

All the models and tools described in the previous sections have been implemented in an integrated informatics approach to systematically evaluate all available

safety-relevant data and predictions through a single platform, referred to as Plato. Plato has been designed to be used by the project toxicologist and chemist together to facilitate discussion, risk assessment, hypothesis generation, and problem-solving activities. The strength of the tool lies in that it is a one-stop-shop for risk assessment at all stages. In the early phases of LI to LO, it is suitable to run a series representative from each series to compare the risk associated with each series. In LO, usage is more specific and focus should be directed at individual compounds and more in-depth searches for off-target effects. At later stages, problem solving is the primary task and can take on many different forms depending on the character of the observed toxicity, how many compounds it is observed in, and if there is public knowledge available. To ensure a complete project penetration all key users are regularly trained on this application in a face-to-face set up. Surveys help the developers to collect feedback and identify information gaps in the tool or lack of training amongst users.

The usage of Plato as a risk assessment tool is also aligned with the internal safety strategies and is required at certain check points or milestones during the drug development process. Certain central parts from each endpoint evaluated in Plato have been combined as an integrated risk assessment score and the calculation of this score and the associated flags have been integrated into the company standard design tool for discovery purposes. The calculation is automatically triggered when compounds are selected and marked to be synthesized. The reason for doing this in an automated way is twofold, it simplifies the workload for the project but, most importantly, it ensures that the risk assessment is actually done before compounds are synthesized in order to avoid unnecessary costs and animal usage associated with experimental testing.

Developing computational tools is sometimes easier than getting users to consistently incorporate them into their everyday workflows. Every user group has a slightly different way of working and different views on what is important. To succeed, tools have to be reliable and cater to what the user needs. They should also be aligned with the scientific strategies such that they have natural check points where the results are required so that usage is seen as beneficial to the project.

12.4 SUMMARY AND OUTLOOK

Safety data provides several challenges to modelers; the available data is largely scattered and nonstandardized, many endpoints are not well understood mechanistically and there are a vast number of unique toxic endpoints that need to be covered. Since safety remains an important reason for drug attrition in clinical development, it is imperative that we generate as good an understanding as possible to be able to mitigate as many safety liabilities as early as possible in the development of a new drug. To achieve this, a wide range of chemoinformatics tools are employed, ranging from traditional QSAR modeling to analysis of biological data in order to identify the potential mechanism of toxicity. For a better way of dealing with data of different quality, warning systems have been developed that encompass all available knowledge

concerning a specific liability and in addition, WoE approaches have been applied to obtain an overall view on potentially disagreeing data. In order to be of use in the drug discovery and development setting, it is necessary to bring tools into the project teams in a robust and easily accessible manner. This calls for good system architecture assuring stability, highly flexible solutions that can be accessed from as many and diverse user interfaces as needed, and finally tools that enable the user to make better design decisions or generate mechanistic understanding and provide the user with a better means to assess the risk associated with their compound.

The future of safety modeling will focus much more on system and high-content data, such as phenotypical and genotypical screens, and making it accessible for modeling. A few key aspects of this is to make sure that data can flow easily between clinical and preclinical development, development of good ontologies to enable the exploration and curation of the vast amount of clinical and preclinical data that is today stored in paper folders, unstructured text documents, and highly skilled coworkers tacit knowledge of the field. This, together with data and understanding of metabolism, pharmacokinetics, and of the biological system will constitute a solid ground for the generation of mechanistic hypothesis that can be tested in prospective studies. Chemoinformatics is today playing an important role in predicting safety liabilities, and it will continue to do so more in the future, in helping identify, quantify, and understand the links between chemistry and biology.

ACKNOWLEDGMENTS

Part of this work has been performed under the Innovative Medicines Initiative Joint Undertaking under Grant agreement no. 115002 (eTOX) and the EU-ADR project no. 215847.

REFERENCES

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004;3:711–715.
2. Crum-Brown A, Fraser TR. On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebata, codeia, morphia, and nicotia, Royal Society of Edinburgh (1868).
3. Fujita T, Hansch C. Analysis of the structure-activity relationship of the sulfonamide drugs using substituent constants. *J Med Chem* 1967;10:991–1000.
4. Nigsch F, Lounkine E, McCarron P, et al. Computational methods for early predictive safety assessment from biological and chemical data. *Expert Opin Drug Metab Toxicol* 2011;7:1497–1511.
5. Dobo KL, Greene N, Fred C, et al. *In silico* methods combined with expert knowledge rule out mutagenic potential of pharmaceutical impurities: An industry survey. *Regul Toxicol Pharmacol* 2012;62:449–455.

6. Boelsterli UA. Diclofenac-induced liver injury: a paradigm of idiosyncratic drug toxicity. *Toxicol Appl Pharmacol* 2003;192:307–322.
7. Toyoda Y, Endo S, Tsuneyama K, et al. Mechanism of exacerbative effect of progesterone on drug-induced liver injury. *Toxicol Sci* 2012;126:16–27.
8. Helgee EA, Carlsson L, Boyer S, et al. Evaluation of quantitative structure-activity relationship modeling strategies: Local and global models. *J Chem Inf Model* 2010;50:677–689.
9. Gavaghan CL, Arnby CH, Blomberg N, et al. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput Aided Mol Des* 2007;21:189–206.
10. Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. *Int J Mol Sci.* 10 (2009) 1978–1998.
11. Michielan L, Moro S. Pharmaceutical perspectives of nonlinear QSAR strategies. *J Chem Inf Model* 2010;50:961–978.
12. Tetko IV, Bruneau P, Mewes H, et al. Can we estimate the accuracy of ADME-Tox predictions?. *Drug Discov Today* 2006;11:700–707.
13. Netzeva TI, Worth A, Aldenberg T, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern Lab Anim* 2005;33:155–173.
14. Sahigara F, Mansouri K, Ballabio D, et al. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 2012;17:4791–4810.
15. TOPKAT OPS, 2000: <http://accelrys.com/products/discovery-studio/admet.html>. Accessed 2013 May 24.
16. Gray A, Moore A. Very fast multivariate kernel density estimation using via computational geometry. In *Proceedings of Joint Statistics Meeting 2003*. Alexandria: The American Statistical Association; 2003.
17. Weaver S, Gleeson v. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 2008;26:1315–1326.
18. Sushko I, Novotarskyi S, Korner R, et al. Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J Chem Inf Model* 2010;50:2094–2111.
19. Bosnić Z, Kononenko I. Estimation of individual prediction reliability using the local sensitivity analysis. *Appl Intell* 2008;29:187–203.
20. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* 2003;22:69–77.
21. Carlsson L, Helgee EA, Boyer S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J Chem Inf Model* 2009;49:2551–2558.
22. Helgee EA, Carlsson L, Boyer S. A method for automated molecular optimization applied to Ames mutagenicity data. *J Chem Inf Model* 2009;49:2559–2563.
23. Walker BD, Singleton CB, Bursill JA, et al. Inhibition of the human ether-a-go-go-related gene (HERG) potassium channel by cisapride: Affinity for open and inactivated states. *Br J Pharmacol* 1999;128:444–450.
24. Sorota S, Zhang X, Margulis M, et al. Characterization of a hERG screen using the IonWorks HT: Comparison to a hERG rubidium efflux screen. *Assay Drug Dev Technol* 2005;3:47–57.

25. Ashby J, Tennant RW, Zeiger E, et al. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutat Res Genet Toxicol Test* 1989;223:73–103.
26. Ashby J, Tennant RW. Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat Res Genet Toxicol Test* 1988;204:17–115.
27. <http://leadscope.com/>. Accessed 2013 May 14.
28. I Multicase, MC4PC, (2006): <http://www.multicase.com/products/prod01.htm>. Accessed 2013 May 24.
29. <http://multicase.com/>. Accessed 2013 May 14.
30. <https://lhasalimited.org/>. Accessed 2013 May 14.
31. Ahlberg Helgee E. *Improving Drug Discovery Decision Making using Machine Learning and Graph Theory in QSAR Modeling*. Göteborg: Chalmers Reproservice; 2010.
32. Neumann H. Aromatic amines: mechanisms of carcinogenesis and implications for risk assessment. *Front Biosci Landmark Ed* 2010;15:1119–1130.
33. Redfern WS, Wakefield ID, Prior H, et al. Safety pharmacology—A progressive approach. *Fundam Clin Pharmacol* 2002;16:161–173.
34. Johnson M, Lajiness M, Maggiora G. Molecular similarity: A basis for designing drug screening programs. *Prog Clin Biol Res* 1989;291:167–171.
35. Bostrom J, Hogner A, Schmitt S. Do structurally similar ligands bind in a similar fashion? *J Med Chem* 2006;49:6716–6725.
36. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity?. *J Med Chem* 2002;45:4350–4358.
37. Krejsa CM, Horvath D, Rogalski SL, et al. Predicting ADME properties and side effects: The BioPrint approach. *Curr Opin Drug Discov Devel* 2003;6:470–480.
38. Fliri AF, Loging WT, Thadeio PF, et al. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc Natl Acad Sci USA* 2005;102:261–266.
39. Bender A, Scheiber J, Glick M, et al. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2007;2:861–873.
40. Scheiber J, Bender A, Azzaoui K, et al. Knowledge-based and computational approaches to in vitro safety pharmacology. *Methods Princ Med Chem* 2009;43:297–322.
41. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;25:197–206.
42. Mestres J, Martin-Couce L, Gregori-Puigjane E, et al. Ligand-based approach to *in silico* pharmacology: Nuclear receptor profiling. *J Chem Inf Model* 2006;46:2725–2736.
43. Garcia-Serna R, Mestres J. Anticipating drug side effects by comparative pharmacology. *Expert Opin Drug Metab Toxicol* 2010;6:1253–1263.
44. Muresan S, Petrov P, Southan C, et al. Making every SAR point count: The development of chemistry connect for the large-scale integration of structure and bioactivity data. *Drug Discov Today* 2011;16:1019–1030.
45. Dvorak CA, Apodaca R, Barbier AJ, et al. 4-phenylpiperidines: Potent, conformationally restricted, nonimidazole histamine H3 antagonists. *J Med Chem* 2005;48:2229–2238.
46. Glowienke S, Hasselgren C. Use of structure activity relationship (SAR) evaluation as a critical tool in the evaluation of the genotoxic potential of impurities. In: Teasdale T,

- editor. *Genotoxic Impurities: Strategies for Identification and Control*. Hoboken: John Wiley & Sons; 2010. p 97–120.
47. Boyer S, Arnby CH, Carlsson L, et al. Reaction site mapping of xenobiotic biotransformations. *J Chem Inf Model* 2007;47:583–590.
 48. Carlsson L, Spjuth O, Adams S, et al. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinf* 2010;11:362.
 49. Redfern WS. SOT 2010, Poster 1081. *Toxicologist* 2010;114.
 50. van Erp M, Nici N, Vuurpijl L, et al. An overview and comparison of voting methods for pattern recognition. In *Proceedings of the 8th Annual Frontiers in Handwriting Recognition Workshop*; Washington, DC; 2002. p 195–200.
 51. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 2nd ed. Boca Raton: CRC Press; 2003.
 52. Liu L, Yager RR. Classic works of the Dempster-Shafer theory of belief functions: An introduction. *Stud Fuzziness Soft Comput* 2008;219:1–34.
 53. Dempster AP. A generalization of Bayesian inference. *J R Stat Soc B* 1968;30:205–247.
 54. Triffiro G, Fourier-Reglat A, Sturkenboom MC, et al. The EU-ADR project: Preliminary results and perspective. *Stud Health Technol Inform* 2009;148:43–49.
 55. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;40:D1100–D1107.
 56. Wang Y, Xiao J, Suzek TO, et al. PubChem’s bioAssay database. *Nucleic Acids Res* 2012;40:D400–D412.
 57. Briggs K, Cases M, Heard DJ, et al. Inroads to predict in vivo toxicology—An introduction to the eTOX project. *Int J Mol Sci* 2012;13:3820–3846.
 58. Krejsa CM, Horvath D, Rogalski SL, et al. Predicting ADME properties and side effects: The BioPrint approach. *Curr Opin Drug Discov Devel* 2003;6:470–480.
 59. Ganter B, Tugendreich S, Pearson CI, et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J Biotechnol* 2005;119:219–244.
 60. Dix DJ, Houck KA, Martin MT, et al. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 2007;95:5–12.
 61. Mason JS, Migeon J, Dupuis P, et al. Use of broad biological profiling as a relevant descriptor to describe and differentiate compounds: Structure *in vitro*–*in vivo* (safety) relationships. *Methods Princ Med Chem* 2008;38:23–52.
 62. Millan MJ. Dual- and triple-acting agents for treating core and co-morbid symptoms of major depression: novel concepts, new drugs. *Neurotherapeutics* 2009;6:53–77.
 63. Olson H, Betton G, Robinson D, et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol* 2000;32:56–67.
 64. Parkinson J, Muthas D, Clark M, et al. Application of data mining and visualization techniques for the prediction of drug-induced nausea in man. *Toxicol Sci* 2012;126:275–284.
 65. Stalring JC, Carlsson LA, Almeida P, et al. AZOrange—High performance open source machine learning for QSAR modeling in a graphical programming environment. *J Cheminform* 2011;3:28.
 66. Cerk T, Demsar J, Xu Q, et al. Microarray data mining with visual programming. *Bioinformatics* 2005;21(3):396–398.
 67. <http://svnbook.red-bean.com/>, Accessed 2013 May 14.

68. Beck K. *Extreme Programming Explained: Embrace Change*, US ed., Reading: Addison-Wesley; 1999.
69. <http://www.scrumalliance.com.>, Accessed 2013 May 14.
70. Sutherland JV, Schwaber K. *Business Object Design and Implementation: OOPSLA '95 Workshop Proceedings*. London: Springer; 1995. p 118.
71. Beck K. *Test-Driven Development by Example*. Reading: Addison Wesley; 2002.
72. Fowler M. 2006. <http://www.martinfowler.com/articles/continuousIntegration.html>. Accessed 2013 May 14.
73. <http://www.w3.org/TR/soap12-part1/>, Accessed 2013 May 14.
74. Richardson L, Ruby S. *RESTful Web Services: Web Services for the Real World*. Sebastopol: O'Reilly Media; 2007.
75. Cumming JG, Winter J, Poirrette A. Better compounds faster: The development and exploitation of a desktop predictive chemistry toolkit. *Drug Discov Today* 2012; 17(17–18):923–927.

CHAPTER 13

APPLICATIONS OF CHEMINFORMATICS IN PHARMACEUTICAL RESEARCH: EXPERIENCES AT BOEHRINGER INGELHEIM IN GERMANY

BERND BECK, MICHAEL BIELER, PETER HAEBEL,
ANDREAS TECKENTRUP, ALEXANDER WEBER, and NILS WESKAMP

13.1 INTRODUCTION

Although many people are working in the field of Cheminformatics (CI) and a lot of exciting scientific papers are published year by year, there is no clear, commonly accepted definition. It starts with the discussion whether this discipline is called “cheminformatics” or “chemoinformatics”—to avoid any conflicts we will use the term CI from now—and ends with every pharmaceutical company interpreting CI differently depending on where it fits best in their drug discovery process.

At Boehringer Ingelheim (BI) in Germany, CI has been established as part of the combinatorial chemistry group in 1999. Combinatorial chemistry again belonged to the Department of Lead Discovery, which was also home of other enabling disciplines, namely high-throughput screening (HTS) and compound logistics. Logically, from the very beginning the focus of CI has been on designing combinatorial libraries and intelligently analyzing huge amounts of HTS results, knowing that a prioritization approach that is better than just sorting by activity must exist.

Today, CI as a discipline can look back over a 10+ year’s history at BI in Germany. Since then, the main focus of CI was always (and continues to be) to provide an optimal support of research projects. Three aspects turned out to be essential during the years: First, CI at BI has always been embedded in an open-minded network of researchers. Second, from the beginnings CI has been well equipped and has been working based on very well-maintained state-of-the-art hardware and software. And last but not least, through several collaborations with leading CI experts in academia,

CI has established a continuous access to the latest scientific developments. Together with academic partners, methods have been adapted to better address the “real” CI questions of the drug discovery process.

This chapter is organized as follows: First, we will start by giving a short overview over a number of important CI tools and systems that form the basis of our daily work. Subsequently, we show a number of different application examples for CI methods. Our intention is to show the broad range of topics and questions to which CI as a discipline can contribute in an industrial environment.

13.2 INFRASTRUCTURE AND SYSTEMS

13.2.1 General Overview

Main focus of most pharmaceutical research organizations is to generate or acquire novel compounds and to characterize these in terms of various properties. The generated compound profiles serve as a basis for the conception of a next round of—hopefully—improved molecules. This cycle continues until a clinical candidate has been identified or a project has to be terminated. Most organizations have established a research informatics infrastructure to support these processes. A number of commercial software products are available today in this application area. Depending on the available resources, typical infrastructures rely on one or a combination of multiple commercial platforms that have been customized and extended to fit the specific needs of each organization [1–4].

At BI, all internal research compounds are registered in a central compound database (CDB). The registration procedure includes a quality check, certain structure normalization steps, and the automatic assignment of unique compound identifiers. Additionally, all experimentally determined results are reported into the CDB to assemble a comprehensive profile for all research compounds in an easily searchable form. While the CDB serves as a central data warehouse and makes it possible to collect and distribute structures and results from different research projects at different sites, it is not intended to provide sophisticated data analysis capabilities or to support the logistics and collaboration within individual research projects. This is the purpose of a number of specialized systems and applications built on top of the CDB. These systems cover a range of technologies and also exhibit a certain degree of overlap in their functionality. Instead of a one-size-fits-all approach, our aim is to ensure optimal support for the diverse needs of different research projects. Our research infrastructure thus has a flexible, modular architecture and consists of a number of independent and weakly coupled information systems (cf. Figure 13.1 for an overview).

In the remainder of this section, we describe some of these systems, their characteristics, and some of the motivations that led to their development. From a CI perspective, the CDB system has—by policy—two important limitations: First, it is only supposed to store experimentally determined results (as opposed to molecular descriptors or calculated properties). Second, only compounds that have been synthesized or used internally are registered in the CDB. This does not include compounds that could be purchased externally (e.g., vendor catalogs) or virtual

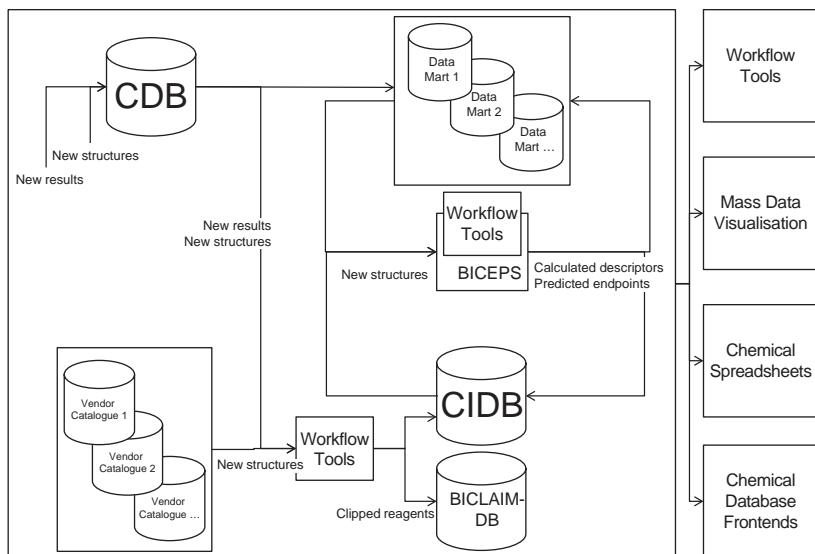


FIGURE 13.1 Schematic view on the most important data flows and information systems in our CI landscape. Various sources for new molecular structures and experimental results exist internally and externally. Different update processes import new structures into our infrastructure, process new results, and ensure that descriptor calculations and model predictions are generated and stored for later usage. A range of different front-ends allows CI experts and end users to access the infrastructure according to their respective needs and tasks.

compounds that could easily be synthesized using available reagents and known chemistry protocols. This led to the development of two different database systems specifically serving these purposes: The Cheminformatics database (CIDB, cf. Section 13.2.2) and the BI Comprehensive Library of Accessible Innovative Molecules (BICLAIM) database (cf. Section 13.2.6.). For individual research projects, only a small fraction of the structures and the results in the CDB system are usually of interest. Therefore, individually tailored, project-specific data marts are set up for most of the research projects (cf. Section 13.2.5). The data marts not only contain a subset of the data from the CDB, they are also complemented with, for example, calculated properties. These types of calculations are typically carried out in workflow systems (cf. Section 13.2.3) and a specialized infrastructure has been established that allows for connecting project data marts and other front-ends to workflow systems (cf. Section 13.2.4).

13.2.2 Cheminformatics Database (CIDB)

For many CI applications, it is useful to access all relevant compound collections in a simple, uniform, and efficient way. We therefore built a CIDB containing normalized structures from the CDB and also from other compound collections. A chemistry cartridge [5] allows for efficient substructure and similarity searches in the different

compound sets. Automatic update processes ensure that the compound collections in the CIDB are synchronized with their respective sources. The import procedure for external collections contains a filtering step that excludes structures containing unwanted substructures or potentially controlled substances (cf. Section 13.3.5).

For all compounds in the CIDB, a number of pre-calculated properties or predicted endpoints are stored. These pre-calculated properties allow, for example, for an efficient assembly of property-filtered subsets of the different compound collections. Additionally, important parameters have to be calculated only once for each compound and can then be used multiple times. This saves computational resources and ensures that all users rely on standardized structures and descriptor values that have been calculated in a consistent way. The structures from the CIDB are therefore also used for updates of computational chemistry tools that maintain compound sets internally (e.g., pharmacophore search software). The calculation of the properties is facilitated by a number of workflows that are automatically triggered whenever structures are added or updated in a compound collection. For some structures, not all properties can be calculated successfully—this case is captured by an error tracking mechanism. Furthermore, a version tracking procedure notes which version of a property calculator was used to generate certain property value and permits recalculations of properties if necessary.

For many specialized CI tasks, data from various laboratory systems or other information sources are needed. Here, the CIDB serves as an interface to the research application landscape. This way, the CI infrastructure is decoupled from the corporate Information systems (ISs) landscape and an efficient access to all relevant data is ensured. In some cases, the CIDB actually replicates data from partner systems to protect these from the high query loads generated during some CI activities. During the analysis of large datasets, the CIDB can be used to store temporary data.

A schematic view of the setup and usage of the CIDB is shown in Figure 13.1.

Initially, it was planned to implement a separate interactive front-end for the CIDB to allow for a native access to the database contents. However, our experience shows that such a front-end is rarely used in practice and that it is sufficient to provide access via workflow tools (e.g., Pipeline Pilot [6] and KNIME [7]). Extracts of the data are made available to end users through the project data marts described in Section 13.2.5.

13.2.3 Workflow Systems

A computational chemist is frequently concerned with the handling and processing of data. Before it is possible to analyze certain dataset, it is often necessary to retrieve data from various heterogeneous sources, to convert between different file formats, to merge different datasets, and to handle outliers and missing values. Workflow tools such as Pipeline Pilot [6] or KNIME [7] can significantly simplify and automate these cumbersome tasks. They provide generic components for important operations such as database querying or file I/O and basic data transformations. Most existing workflow tools rely on the concept of visual programming and allow the user to create workflows in a simple and intuitive way. They are thus accessible to a wide range of

end users with different skill sets and levels of programming experience. The generated workflows serve as a documentation of the performed processing steps and can be exchanged among users to facilitate knowledge sharing.

Most workflow systems can be easily integrated into existing application landscapes using web services and other generic interfaces. Furthermore, expert users typically have an option to implement custom nodes through a programming interface and to extend the workflow tool according to their organization's needs.

At BI, workflow systems are currently the main tools for CI applications. They support us in our daily work (e.g., to perform ad-hoc analyses to answer questions arising from research projects) and also help us to automate repetitive standard tasks or to share adaptable solution templates for frequent work packages with our colleagues.

13.2.4 BI Chemical Property Structure Planning System (BICEPS)

Important examples for automated CI-services at BI that are frequently developed and deployed using workflow tools are calculators for simple molecular descriptors, different physicochemical properties, and also quantitative structure–activity relationship (QSAR) models for predicting a number of absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints [8–10]. The low success rates of pharmaceutical compounds experienced during clinical development have led to a plethora of publications examining the influence of molecular properties on clinical success [11–17]. Consequently, a number of scores and “rules” have been proposed that should aid in selecting the most promising compounds from a list of starting points or to prioritize synthesis ideas. The high popularity of many of these predictors among scientists and managers calls for making them conveniently accessible in a consistent way among various tools and front-ends. Figure 13.2 shows an example for a front-end that is currently being used at BI: The BI Chemical Property Structure Planning System (BICEPS) helps medicinal chemists to characterize their synthesis ideas. Virtual compounds are registered in a planning database, which conveniently stores all relevant compounds for a particular research project. Upon registration of a new compound, descriptor calculations, similarity searches, and QSAR model predictions are triggered and their results are made available in the front-end within a couple of minutes. This allows the chemist to focus synthesis efforts on the most promising compounds from a set of ideas.

Nowadays, many different versions and “flavors” of even the most basic molecular descriptors are available. Even for a simple and intuitive descriptor as the “number of rotatable bonds” of a molecule, different implementations will yield significantly differing values for the same structure. Since such descriptors form the basis of many *in silico* predictions and support strategic decisions (e.g., the prioritization of one compound over another), it is desirable to standardize their calculation within a pharmaceutical research organization. This can be achieved easily if a single implementation of the respective descriptor is used and made available as a central service for all relevant clients and front-ends. This service has to be maintained and kept stable over time since descriptor values calculated for older compounds are frequently used as reference points and compared to more recent results.

Structure		Sample Identifier					Library ID	Chemist	Class	Subclass	Main PhysChem Prop Dose PPB Pemetrexed Hydrogen Rule
		AT-20090114									
MolWeight	Molformula		Purpose		Successor of						
145.16	C ₈ H ₇ N ₂										
Num. Comment 1	Num. Comment 2	Num. Comment 3	Num. Comment 4	Num. Comment 5							
Text Comment 1	Text Comment 2	Text Comment 3	Text Comment 4	Text Comment 5							
18982											
CLogP	ClogP message	Weight	# Acceptors	# Donors	TPSA	# Rotatable Bonds	# Lipinski violations				
0.76	All fragments measured:	145	2	1	52	0	0				
CNS MPG Score	Veber Message	HIA Message	Andrews bonding energy	# pos. N-atoms	Abbott Bioavailability	Net charge @ pH 5					
5.83	BA ok	good	-5.50	0	0.55	0.0					
pKa											
Structure		Identical Sample Codes		Nearest Neighbors		Caco-2 Permeability ab [1E-6 cm/s]		bc [1E-6 cm/s]		ETraffic	
				5		Ver. from to		Ver. from to		Ver. from to	
						V12 40.0 80.0		V07		V12 1.5 3.0	
										V07 45.0 60.0	
Metabolic stability											
HLM (%Qh)		RLM (%Qh)		NLIM (%Qh)		Solubility thermodynamic		Int. [1E-6 cm/s]			
Ver. from to		Ver. from to		Ver. from to		Ver. from to		Ver. from to			
V08		V08 40.0 60.0		V04 60.0 75.0							
V02A 75.0 100.0											
HERG (Inv. IC50 [uM])											
n-o/w		CL - rat (%Qh)		VSS - rat (kg)		pIC50 (nM)		pK (Vol. 10 g/L)		pH (Vol. 10 g/L)	
Ver. from to		Ver. from to		Ver. from to		Ver. from to		Ver. from to		Ver. from to	
V07 10.0 100.0		V07		V09 2.0 5.0							
Phospholipidosis											
Human (%Binding)		Rat (%Binding)		HTSOL (uM)		Risk		pH		Ver. from to	
Ver. from to		Ver. from to		Ver. from to		low		2.2 V01 200.0 10000		4.9 V01 200.0 10000	
V01 0.0 95.0		V01 0.0 95.0						6.8 V01 200.0 10000			

FIGURE 13.2 An example for a BICEPS planning database that is used to collect and characterize synthesis ideas (virtual compounds) for a research project. A number of molecular properties are predicted and used to prioritize synthesis ideas. BICEPS planning databases are a special type of project data mart.

We currently use a generic mechanism to deploy such standardized descriptor calculations, QSAR-models, and other services that have been implemented in different workflow tools to a number of different front-ends. This ensures that the end user sees identical descriptor values regardless of the front-end. The mechanism works completely asynchronously and uses a database as intermediate communication infrastructure. The database layer ensures that a request is stored persistently until it is processed—it thus automatically provides a buffering mechanism for large requests or computationally expensive calculations and guarantees that no requests are lost even if one of the calculation engines should be unavailable.

13.2.5 Project Data Marts

All newly synthesized structures and all experimental results are stored in the CDB. The data model of this data warehouse is necessarily very complex and flexible, which makes it often difficult and slow to directly retrieve data in a useful format. Therefore, a specialized project data mart is set up that contains only the relevant structures and results for a particular project. This includes sample-related data such as availability or analytics results, but also project-related assay data as well as experimental physicochemical and ADMET results. All results are pivoted to a suitable format, normalized to a certain unit, and presented in a conveniently readable form. This includes the calculation of mean values (and other statistical parameters) in the case of multiple measurements and conditional formatting according to critical thresholds. An automated daily update procedure retrieves new structures and results from the CDB and adds them to the data mart. In addition to experimental results, the BICEPS mechanism is used to provide access to calculated properties (cf. Section 13.2.4). Since project data marts serve as central data exchange and communication platform within research projects, they are a natural reporting front-end for many other CI tasks such as the results from HTS data analyses (cf. Section 13.3.1) or the BioProfile of a compound (cf. Section 13.3.2). The communication within the project is facilitated by a number of manual comment fields that can be used by the team to add comments or suggestions and to track the status of requests for assay testing.

As these project data marts are stored in an Oracle [18] database instance, there are several possibilities to access the data. Analysis of the data can be done with workflow tools like Pipeline Pilot [6] or KNIME [7]. Visualization in mass data viewers like Spotfire [19] is also possible. Currently, medicinal chemists often use ISIS/Base [20], using a predefined set of standard forms, which can be adjusted if necessary. An example for such a form is shown in Figure 13.2.

13.2.6 Database of Virtual Combinatorial Libraries (BICLAIM-DB)

Within the drug discovery process, different compound collections are used to identify new starting points for a specific target. Based on defined filter criteria, compound collections with up to 10^4 entities were compiled for fragment-based screening approaches [21, 22]. Corporate compound collections amenable to HTS include traditionally about 10^6 entities. All these collections can be tested with biochemical

assays and other detection techniques [23–25]. External compound collections with up to 10^7 entities usually complement the internal corporate dataset with respect to chemical motifs and structural diversity. Ligand- or structure-based virtual screening approaches are traditionally used to identify and select a subset of interesting new molecules from external compound collections [26, 27]. With respect to the total number of compounds, there is an obvious gap to the postulated maximum number of chemical drug-like entities, the so-called “chemical space” with up to 10^{60} molecules [28, 29]. One way to fill this gap of chemical entities is the incorporation of virtual combinatorial libraries. A virtual combinatorial library is defined by one or multiple sets of building blocks and a number of rules defining how these building blocks can be combined to form new molecules.

BI's internal pool of virtual combinatorial libraries is called BICLAIM [30]. It distinguishes two types of building blocks: library cores and reagents. In our terminology, a core has multiple functional groups (or attachment sites) that can be substituted with reagents possessing only a single functional center. Figure 13.3 shows an example: The shown key intermediate contains an aryl-iodide motif and a primary aniline functional group. To transform this intermediate into a library core, the identified functional groups are clipped to virtual linking atoms, assigned as R1 and R2. These linking atoms indicate variation points for chemical synthesis. In the given example, boronic acids could be introduced in R1, whereas carboxylic acids, sulfonyl chlorides, and isocyanates could be used for variations in R2 [31]. The clipped library core and the lists of suitable reagent types are stored in a library information table.

The first version of BICLAIM consisted of about 200 manually defined libraries. The structures of the library cores together with the main library information were summarized in an Excel sheet. Since then, new interesting libraries were added year by year. Synthesized combinatorial libraries were added to the virtual space as well as new

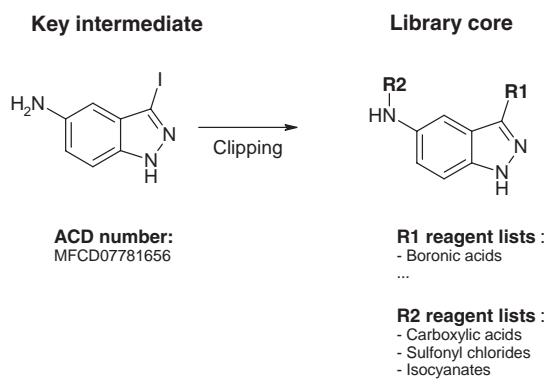


FIGURE 13.3 Structure of virtual combinatorial libraries in BICLAIM space. Certain functional groups are detected and clipped to linking atoms to yield a virtual library core. The different attachment points are annotated with the different types of reagents that can be attached.

library proposals from literature. In addition to these manually defined library protocols, automated routines were established to detect interesting key intermediates of new potential BICLAIM libraries. The implemented assignment routine at BI includes about 30 different functional groups for library core detection. Additionally, about 30 reagent lists were defined that could be combined with the detected core functional groups.

This new procedure expands the BICLAIM space to several thousands of new virtual library cores resulting in up to 10^{13} enumerated virtual molecules. With respect to this expansion, technical improvements are necessary to cope with the increased library space. Computational systems and tools that are able to store and screen such large virtual combinatorial libraries without fully enumerating all possible products are of great interest. As a consequence, the old Excel version of the BICLAIM space was substituted with an Oracle [18] database, called BICLAIM-DB. Structural information of the library key intermediates as well as the structure of the clipped library core is imported into this database. In addition, important library properties were connected to each library entity: availability of key intermediates, source of key intermediate (internal, external), linkage to chemical synthesis protocols, and calculated properties, such as molecular weight of the library core. All structure tables can be used for substructure and similarity searches, supporting the navigation through the virtual space of BI's combinatorial libraries.

Different ways to access the data content of the BICLAIM-DB are available (cf. Section 13.3.4). KNIME [7] and Pipeline Pilot [6] workflows were implemented to extract and screen the BICLAIM-DB content. A graphical user interface was established for manual navigation through the data content. All BICLAIM-DB front-ends offer the possibility to select and export subsets of the BICLAIM-DB content in different file formats (Excel, SDF, and CSV).

13.3 METHODS AND APPLICATIONS

In the previous section, an overview of our current CI infrastructure, its most important tools and systems was given. In this section, we now describe which applications formed the motivation for the development of these tools and show a selection of typical work packages. We start by giving a high-level description of BI Mining and Exploration of Screening Hit (BIMESH), our approach for mining results from HTS campaigns (cf. Section 13.3.1). An important foundation for BIMESH and other applications is BioProfile, a data collection containing all available experimental results for a given compound in an easily accessible, standardized, and annotated form (cf. Section 13.3.2). Another important aspect of our work concerns the exploration of hits to extend known lead classes and to improve the understanding of SAR within classes (cf. Section 13.3.3). Virtual combinatorial compound libraries pose a particular challenge to most CI methods. In Section 13.3.4, we describe BI Split Substructure Search (BI S3 or BISCUBE), one of the approaches used to provide access by means of substructure searches to these enormous compound collections. In Section 13.3.5, we show an example of how CI can aid in ensuring legal compliance in a highly regulated industry.

13.3.1 BIMESH: The HTS Data Analysis

A central concept in pharmaceutical research projects is that of a “lead”—a compound (or, preferably a class of related compounds) that possess certain desired activity together with acceptable values in most other relevant parameters. Such a lead can serve as a promising starting point for a full optimization project aiming at further improvement of the overall profile of the compound class. As previously mentioned (cf. Section 13.2.6), leads are typically generated by some form of systematic screening: A collection of compounds is subjected to a cascade of tests with increasing reliability and complexity [32–34]. After each step, hits may be selected and prioritized for the following steps. A variety of different screening strategies has been proposed in the literature, whereas differences typically exist in the size and composition of the screening collection and in the applied testing principle. Typical testing principles include biochemical, biophysical, and phenotypical assay formats [21, 25, 35, 36]. Of course, also *in silico* screening methods relying on ligand- or structure-based techniques can be applied [37, 38]. The tested compound collection can consist of up to millions of compounds in the case of classical HTS, but it can also consist of only a few carefully selected compounds in the case of focused or fragment-based screening. If *in silico* screening is used, it is often also possible to consider virtual compound collections, which requires a purchasing or synthesis step subsequently in the screening campaign. The details of the applied screening strategies differ significantly from organization to organization. It may even be advisable to use an adjusted lead-identification strategy for each novel research project. A general observation is, however, that the complexity of typical lead-identification programs increases over time. New, complementary approaches and techniques are used in addition to older methods. Learnings from failed lead candidates lead to additional filtering and characterizing steps (e.g., technology counter-screens). Improved knowledge about target families is used to test for selectivity and anti-targets early in the process.

All these factors require an elaborate and complex analysis of the many and diverse test results generated during a screening campaign. For example, it is necessary to keep track of which compound was tested where, under which conditions, what result was generated, and how that influenced the further progression of the compound. At critical decision steps, it is necessary to integrate all relevant information (regardless of its type or source) into a single view to support decision making. In addition to experimental results, an *in silico* profile based on calculated properties and predicted properties is included. In the following sections, we will describe BIMESH—a collection of KNIME [7] workflows we developed for this purpose. It contains workflows covering all major work packages addressed during a screening campaign. BIMESH serves as a template, not as a tool: While the workflows in BIMESH are set up to cover the “standard case,” most real screening campaigns require some modifications of at least some workflows to fit their specific needs.

The workflows in BIMESH are organized in four subsequent stages: data retrieval, preprocessing, clustering, and reporting. Most workflows read in data files that were created in earlier stages of the analysis or generate new files that will be used in later stages. Additional folders exist for recently added, not yet fully integrated workflows and for optional work packages.

Aim of the “data retrieval” stage is to extract all relevant raw data from their respective sources. The user is usually able to execute these workflows as they are and has nothing to care about the technical details of the various data sources. Test results for biological activity can be retrieved in standardized format from the corporate compound database (CDB, cf. Section 13.2.1). Other data, such as the BioProfile (cf. Section 13.3.2), the library information for compounds originating from combinatorial screening libraries or hit quality check results have to be retrieved directly from different laboratory information systems or from the CIDB(cf. Section 13.2.2). Once all the relevant data have been retrieved, it is stored conveniently as a set of files. This easily accessible collection serves as foundation for all subsequent analyses.

After all relevant data have been retrieved; it is brought into a suitable format during the “preprocessing” stage. Although this is mainly a technical step involving pivoting, grouping and filtering steps, it is usually the first step that requires a manual, project-dependent adaptation (e.g., to account for the various readouts of different assay types and technologies). A very central part of the whole BIMESH process is the “clustering” stage. It aims at grouping the various hits found during the screening campaign into structural classes of related compounds. The structural class concept lies at the very heart of CI and medicinal chemistry in general, at the same time it is inherently vague and fuzzy. We therefore integrated various similarity measures, clustering algorithms, and fragmentation schemes [39–43] into BIMESH that are applied on a regular basis to generate a number of different cluster assignments. In our experience, no single algorithm is able to generate satisfactory clustering results in each and every project. Inspection, selection, and refinement of the different alternative cluster assignments are thus again manual, project-dependent steps that have to be performed during a BIMESH analysis. The final stage of BIMESH is “reporting.” It aims at aggregating all relevant data using the previously prepared clustering to generate a profile for each structural class. This also includes the known BioProfile of the contained compounds in aggregated form (cf. Section 13.3.2). The per-class overview (including all available measured and calculated data) is used to generate a couple of reports (in HTML or PDF formats) using the reporting capabilities of Pipeline Pilot. Additionally, the cluster assignments and some statistical numbers are reported to the project data marts (cf. Section 13.2.5). Again, the reporting stage usually requires a manual, project-dependent adaptation to include and highlight those parameters that are decision-relevant for the respective project.

The outcome of a BIMESH analysis is thus a convenient information package that is distributed to the project team and used to prioritize and select the most promising structural classes for further profiling and analysis. The static reports provide an overview of the various structural classes while the project data mart can be used to retrieve detailed information about the different molecules in a given cluster. This interplay between the different outputs allows project teams to get a quick and substantiated overview of the results of a screening campaign.

After the most promising lead classes have been identified by the team, one of the next steps is to extend the structural class through analog searches (hit exploration) and to improve the understanding of the SAR within this class (cf. Section 3.3).

13.3.2 BioProfile

In the last 10–15 years, many new technologies and approaches have been implemented in pharmaceutical research; these include HTS or combinatorial chemistry, which result in a rapidly growing amount of biological assay and structural data in the corporate databases. Figure 13.4 shows the increase of the assay data stored in the CDB (cf. Section 13.2.1) within the last decade.

Efficient use of this growing data mountain is a key success factor. We want to provide as much knowledge as possible, as early as possible, and therefore enable research teams to make the best possible decision whenever this decision can be supported by stored data. In this section, an approach termed BioProfile is described that is used within BI to generate knowledge from the in-house assay results.

During prioritization of HTS hit classes, the following questions are frequently addressed:

- Are there cross-reactivity or selectivity issues for a given compound/compound class?
- Is a given compound a real hit or a frequent hitter or an artifact of the assay technology?
- Can we identify selectivity targets not known in advance for the HTS hit set or an interesting hit class?
- Are there known toxic effects?

The total workflow for the BioProfile analysis consists of three parts; automated and regular data retrieval from the corporate database for all new data, preprocessing of the data, and storage of the preconditioned data in the CIDB (cf. Section 13.2.2).

For the analysis of single dose measurements, we currently concentrate on values from HTS campaigns and on dose–response data from our corporate database. During first trials with the data retrieved a few years ago, we soon discovered that it would be very helpful if we were able to obtain some additional assay specifications



FIGURE 13.4 Development of the number of stored single point (percent of control) and dose–response measurements in the CDB over the last decade. Reprinted from *Bioorganic & Medicinal Chemistry*, Vol. 20, Bernd Beck, BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds, 5428–5435, 2012, with permission from Elsevier.

TABLE 13.1 Classification of the Target and Technology Types

Assay Technology	Target Type
Absorption	Enzyme
AlphaScreen [57]	GPCR
Delfia [57]	Ion channel
FlashPlate [57] blue	Kinase
FlashPlate [57] red	Nuclear receptor
FLINT (fluorescence intensity)	Nucleic acid/protein binding
FP (fluorescence polarization)	Phosphatase
FLIPR [58]	Polymerase/nuclease/helicase
HTRF [59]	Protease
HCS (high-content screening)	Protein binding
LANCE [57]	Transporter
Luminescence	Other
SPA (scintillation proximity assay) blue	
SPA (scintillation proximity assay) red	
FRET (fluorescence resonance energy transfer)	
Other	

Reprinted from *Bioorganic & Medicinal Chemistry*, Vol. 20, Bernd Beck, BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds, 5428–5435, 2012, with permission from Elsevier.

from our colleagues from the HTS units. This includes the information whether a result originates from a screen for agonistic or antagonistic effects, the hit threshold, the mean value, and the standard deviation of the screens and also information about the target type and the assay technology as shown in Table 13.1.

These kinds of annotations are currently available for more than 220 primary screens. A weekly update retrieves new single dose data for the stored assays from the corporate database. For the dose–response data, we retrieve all IC₅₀, EC₅₀, Ki, Kd, pEC₅₀, and pIC₅₀ values stored in the central database. Again, an automated weekly update process retrieves new data. For most of the assays with reported dose–response values, we also add information about target type and the assay technology as for the primary screening data. For older methods this data was annotated manually.

Preprocessing of the retrieved data is an essential part of the whole process. Single dose data and dose–response data are handled separately.

For the primary screen data, we retain from all the retrieved percent-of-control values from the database, the minimum values for antagonistic screens and the maximum ones for agonistic screens if multiple measurements per compound and method exist. In this way, we generated up to now approximately 20 million agonist and 150 million antagonist data points.

For the dose–response data from the corporate database, we also need some preprocessing. First of all, values are converted to micrometer. For multiple measurements (without an operator) per compound per method, we calculate the median value. This results in more than 4.6 million data points from more than 4000 different assays. All preconditioned data are finally stored in the CIDB (cf. Section 13.2.2).

We rely mainly on workflows built in KNIME workflows to analyze this data. The per-compound data are also available through our project data marts as described and shown in Section 13.2.5. In the following, we describe two application examples. The first one is a compound-based analysis using primary assay data to identify frequent hitters. The second example is an analysis based on the dose–response data of a complete hit set.

For frequent hitter analysis, we defined a frequent hitter score that depends on the number of screens in which a compound participated and on the number of screens where this compound was a hit. We aimed at identifying a simple, empirical score that allows us to rank compounds with respect to their promiscuity, also in cases where compounds were tested in a different number of assays. A biological assay system is modeled as a biased coin that yields “hit” or “non-hit” with certain probabilities and the various assays to which a compound is subjected as a sequence of independent coin flips. Thus, we use a binomial distribution function to estimate the relative probability of identifying a compound as a hit n times in k independent assays by chance. The probabilities for the events “hit” and “non-hit” were estimated empirically from a set of assays.

During the discussion of the first results, it became clear that we need an additional frequent hitter score that takes the number of different technologies and target types for which a given compound was found as a primary hit into account. For example, a common kinase inhibitor that was tested in many kinase assays will have a high-frequent hitter score even if it was only found as hit in kinase projects. This compound is not a real frequent hitter. We therefore now also use a modified score “NewScore” that includes this information.

The analysis is started by simply joining the pre-calculated scores with the number of screens in which a compound has participated and the count how often it was a primary hit. An example output with some public domain structures is shown in Table 13.2. For each compound, it is reported how often the compound was in a primary screen campaign, how often it was found as a hit, and in how many different assay technologies and different target types it was found as a hit. The hit set–based analysis is done in order to gain a quick overview on potential cross reactivity issues. The results can be delivered as a table or a graph. An example is shown in Figure 13.5.

In the plot, there are regions in which only a few values or only values with “>” operator (not plotted) can be found (e.g., around assay ID 5850). We are not able to draw any conclusions in these areas. On the other hand, there are several assays in which we have a large overlap with the current hit set (e.g., assay IDs between 4550 and 4750). In these cases, we only found yellow and green symbols for an assay in which the compounds of the actual hit set were not active, so that we have no issue here. If we find many red or dark red compounds for an assay (e.g., assay ID 4890), we need to check the assay more closely as a possible selectivity counter-screen. Analyzing the plot in more detail, for example for assay 4890, one can determine whether compounds are already available that are selective against the target and therefore obtain some first hints about how to address the cross-reactivities.

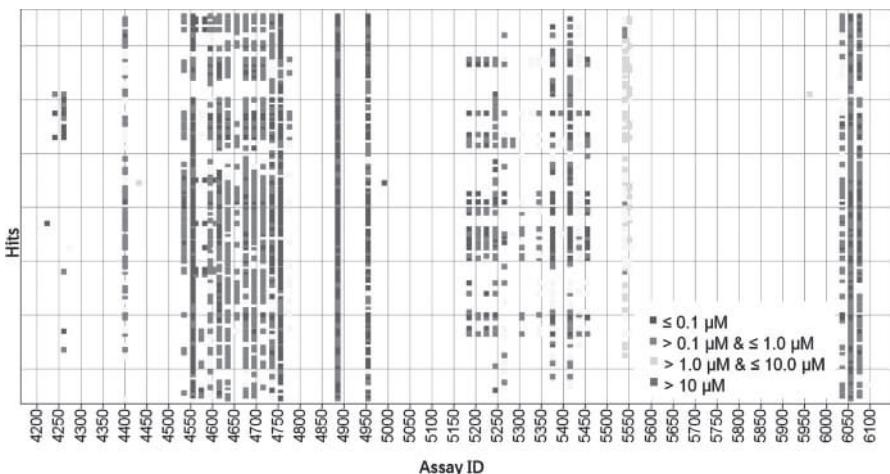


FIGURE 13.5 Analysis of the DR data for a compound class or complete hit set. It can be easily analyzed which other assays overlap with the given hits. The color coding is from green (inactive) to dark red (very active) compounds. Reprinted from *Bioorganic & Medicinal Chemistry*, Vol. 20, Bernd Beck, BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds, 5428–5435, 2012, with permission from Elsevier. For color details, please see color plate section.

Our BioProfile approach makes it easily possible for research project teams to access information about cross-reactivities within a given hit set. This information can be used to prioritize compounds or compound classes. It is also used to check for potential selectivity targets or counter-screens and to identify screening artifacts. Identification of frequent hitters in the screening collection is another example of use.

More details for the BioProfile analysis are found in the recent publication [44].

13.3.3 SAR Analysis

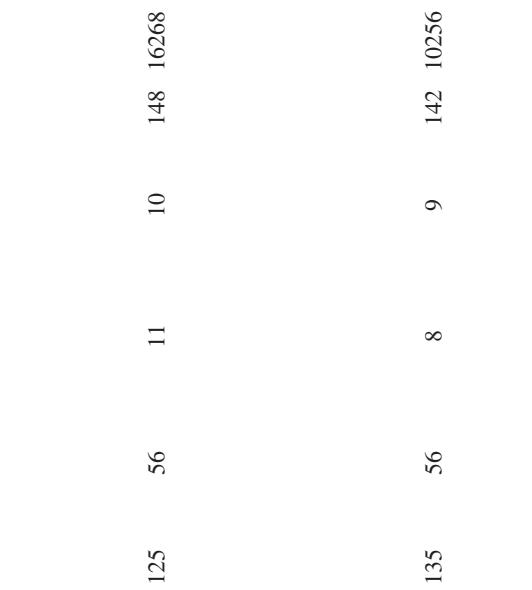
13.3.3.1 Introduction Activity data for thousands of compounds are generated during lead identification and lead optimization phases of research projects. These data contain a wealth of information about structure–activity relationships (SARs), which are used to guide medicinal chemists. One of the major tasks of a computational chemist is to analyze and visualize this SAR information, and to design improved compounds based on the extracted SAR knowledge.

The optimization cycle starts with the determination of biological activities for a set of synthesized compounds. SAR information is extracted from the activity data and guides the design of a new set of compounds. Compound synthesis is followed by another round of biological testing and the cycle begins anew. The primary optimization parameter is biological activity on the target protein in a biochemical or cellular assay. Early on, additional optimization parameters, like physicochemical properties (solubility, lipophilicity, and stability) and ADMET properties, for

TABLE 13.2 Example of a Frequent Hitter Analysis Result

ID	Structure	Primary						New					
		Screens Partic.	#Screens Primary Hits	Found as Primary	Hit in # Different Technologies	Different	Target Types	Score	Score	Score	Score	Primary Hit in # Different Target Types	Score
MFCDD00071920			159	78	9		10	214	19295				
MFCDD00011750			204	79	10		10	193	19333				

MFCD00004021



MFCD00602221



Reprinted from *Bioorganic & Medicinal Chemistry*, Vol. 20, Bernd Beck, BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds, 5428–5435, 2012, with permission from Elsevier.

example, inhibition of the human ether-à-go-go related gene (hERG) channel, are considered. This leads to a multiparameter optimization problem with complex structure–property relationships (SPRs). Multiparameter compound optimization is a highly interactive process that involves specialists from a range of disciplines, including medicinal chemists, biologists, pharmacologists, structural biologists, and computational chemists. There is a great demand for tools that analyze multi-parameter SARs and that present the generated SAR hypotheses in a way that is intuitive to all scientists involved. A wide range of approaches to analyze SAR is available ranging from simple R-group tables to high-level machine learning algorithms [45]. A small subset of software tools recently added to our toolbox will be discussed in this section, including in-house workflows built in KNIME and Pipeline Pilot.

13.3.3.2 Activity Landscapes HTS campaigns deliver single point activity data for several hundreds of thousands of screening compounds. Additionally, dose-response data for the hit set—usually consisting of hundreds to a few thousands of compounds—is generated. Despite the fairly large degree of uncertainty resulting from single point measurements, these data provide an initial blueprint of the activity landscape surrounding the dose–responsive screening hits. Analysis of the activity landscape can provide first answers to two central questions—what structural modifications are allowed or even favorable and what modifications are unfavorable or completely disallowed?

An analysis of structurally related compounds with varying activities provides SAR information on allowed and favorable structural modifications and is the approach of choice to extract SAR. An alternative view on favorable SAR can be derived from a ligand efficiency (LE)–based analysis in which activities normalized with respect to molecular complexity are utilized. High LEs indicate a high mean contribution of all atoms to the binding event. Thus, the LE-based SAR view often provides a sharp picture of favorable structural motifs, as it focuses on smaller compounds with fewer, but on average very efficient, interactions.

Unfavorable structural modifications can be deducted from chemically related compounds with large differences in activity that form steep cliffs in the activity landscape [46]. Activity cliffs are rich in SAR information and help to focus the optimization activities by reducing the options for structural modifications. The identification of activity cliffs critically depends on the availability of information on inactive compounds. This information mainly exists in the form of single-point data. In the past, this valuable information has often been neglected, because single point measurements are significantly less reliable than dose–response data. However, retesting of key analogs can help to confirm the extracted SAR relationships.

13.3.3.3 Matched Molecular Pairs A very intuitive way to extract and present SAR information is based on the matched molecular pairs (MMPs) approach. MMPs describe a pair of structurally related molecules that share a well-defined constant region (or regions) varying only in one particular structural element. Examples of MMPs are two compounds with a side chain variation, or two compounds

with identical substitution patterns, but different cores. Several approaches to identify MMPs have been described in the literature, in particular from industrial research groups [47–49]. The fragmentation-based method published by Hussein and Rea [50] scales well with large datasets and is widely used in different variants at BI.

Multiple MMPs with an identical constant region can be grouped into an MMP series of related molecules. Members of an MMP series can be sorted by different molecular properties (e.g., potency) resulting in an intuitive presentation of SAR information closely resembling the commonly used SAR tables. In contrast to R-group decompositions, no prior knowledge (e.g., in form of predefined substructures) is required to perform the MMP fragmentation. The MMP analysis often identifies unexpected SARs, like core variations and linker exchanges, because there is no user induced bias in the analysis. The visual SAR interpretation is simplified by a sharp separation between variable and constant structural elements.

We use MMP approaches to analyze primary SAR, selectivities, and anti-target SARs. Results of an MMP analysis are presented in the form of HTML or PDF reports, which show the constant region of the molecular series on the left-hand side followed by the variable region of each individual compound (cf. Figure 13.6) sorted by the property of interest. Activity can be substituted by any other property that is part of the multiparameter optimization problem, for example, solubility or ADMET endpoints. Data interpretation is facilitated by color coding.

From our perspective, the MMP approach is an interesting new development in the field of CI. The simple and intuitive interpretation of MMP data adds to the popularity of the approach. In particular, pharmaceutical companies have large datasets of analog compounds from combinatorial libraries waiting to be explored by MMP approaches.

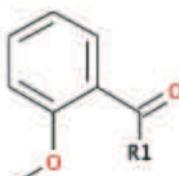
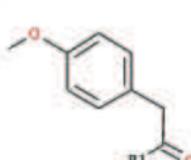
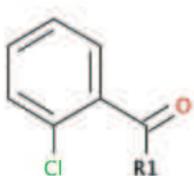
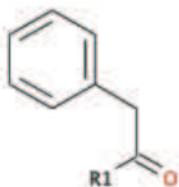
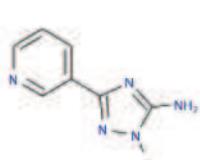
13.3.4 Searches in BICLAIM-Space

BICLAIM, the chemical space of synthesizable combinatorial libraries described in Section 13.2.6, comprises about 10^{13} virtual molecules. The enormous number of compounds precludes time consuming, systematic searches and reduces the repertoire of computational techniques to those tools that are explicitly optimized to handle combinatorial libraries. Complete enumeration of the entire compound set is impossible due to the time needed to perform complete enumeration and due to the storage size of the enumerated chemical space. However, even if the storage problem could be solved, there are no tools to perform a search in such a huge compound space within a satisfying time frame. Therefore, applicable tools are typically optimized to work within un-enumerated fragment spaces.

Feature Trees [51] is one example for those software tools that make use of the fragment space by comparing molecules based on their features and their connectivity. This tool enables us to perform similarity searches within the BICLAIM space. Only a limited number of tools allow for performing substructure searches [52–54]. The search for chemical substructures in molecules is a widely used CI technique in Drug Design. Applications range from filtering of molecules containing unwanted substructures like toxicophores to the analysis of substance pools. Sometimes, a

Thrombin SAR Update - 29.10.2012

Class 3A - Series 7

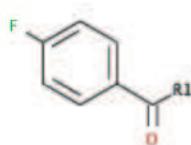
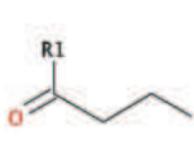
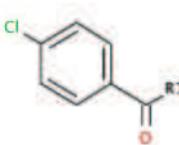
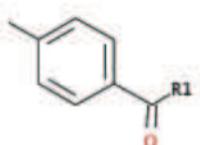


CPD4262570 : 0.078 μ M		
IC50	Caff1	Caff2
0.078	0.044	0.107
MW	TPSA	clogP
219	86	23
CYP3A4	rMcSub	rOH
8.1	62%	57%

CPD24793640 : 0.083 μ M		
IC50	Caff1	Caff2
0.083	0.044	0.108
MW	TPSA	clogP
209	86	23
CYP3A4	rMcSub	rOH
8.1	70%	59%

CPD4261529 : 0.269 μ M		
IC50	Caff1	Caff2
0.269	11.65	19.61
MW	TPSA	clogP
209	95	23
CYP3A4	rMcSub	rOH
8.7	41%	81%

CPD7965471 : 0.486 μ M		
IC50	Caff1	Caff2
0.486	8.85	11.83
MW	TPSA	clogP
209	95	23
CYP3A4	rMcSub	rOH
8.6	40%	69%



CPD14731064 : 0.543 μ M		
IC50	Caff1	Caff2
0.543	25.229	42.138
MW	TPSA	clogP
279	86	23
CYP3A4	rMcSub	rOH
8.4	56%	22%

CPD24823247 : 0.781 μ M		
IC50	Caff1	Caff2
0.781	12.502	25.318
MW	TPSA	clogP
269	86	23
CYP3A4	rMcSub	rOH
8.7	69%	40%

CPD14736080 : 0.986 μ M		
IC50	Caff1	Caff2
0.986	24.142	40.304
MW	TPSA	clogP
261	86	23
CYP3A4	rMcSub	rOH
8.6	58%	49%

CPD24796909 : 2.009 μ M		
IC50	Caff1	Caff2
2.009	20.219	38.788
MW	TPSA	clogP
263	86	23
CYP3A4	rMcSub	rOH
7.7	37%	56%

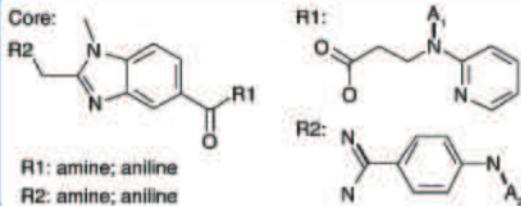
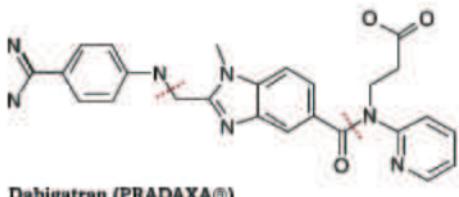
FIGURE 13.6 SAR report based on a matched molecular pairs series. The MMP algorithm efficiently identifies closely related molecules with a conserved constant region (upper left). Molecules within one MMP series can be sorted with respect to the property of interest, for example, IC50. The example shows an MMP series from a public thrombin dataset (PubChem [56], AID 1215), containing IC50 values [μ M]. For color details, please see color plate section.

substructure search is also part of a virtual screening workflow with the aim to identify new active molecules for a target receptor. By definition, a substructure search identifies parts of a molecule (substructure) which are equivalent to a query structure. When searching in large chemical spaces, queries that cover only common substructure motifs (e.g., phenyl or benzyl) result in millions or even billions of substructure hits. Thus, the visualization of enumerated hits is not feasible anymore. A systematic representation of these hits for the purpose of visualization is one of the reasons why we decided to implement our own algorithm for substructure searches. In the following, we will describe our implementation for substructure searches that is optimally adapted to the way the BICLAIM space was designed.

We will describe an artificial example for substructure searches in BICLAIM space to elucidate the implemented algorithm. For this example, let us assume the direct thrombin inhibitor dabigatran (cf. Figure 13.7a) [55] could be synthesized by combinatorial chemistry. Figure 13.7b shows the fragmentation of dabigatran into an artificial core element with two open residues (R1 and R2) and the corresponding two substituents with A1 and A2 indicating the attachment positions for residue 1 and residue 2, respectively. Let us further assume, in position R1 and R2 of the combinatorial library any primary or secondary aniline or any primary or secondary amine could be placed. Considering the current BICLAIM space, the combinatorial library would comprise hundreds of millions of enumerated molecules (#Cores x #R1-Substituents x #R2-Substituents).

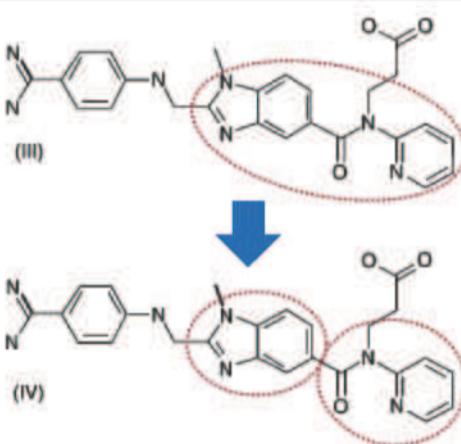
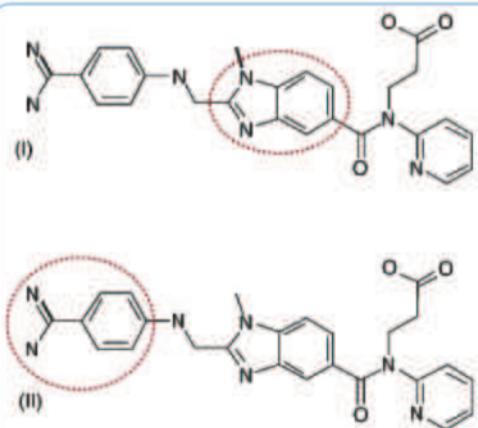
The principle of our algorithm can be shown in Figure 13.7c: A substructure is either part of the core (case I), part of a residue (case II), or combines core and residue (case III). In case I, it is sufficient to search only in the list of cores, and in case II, it is sufficient to search only in the list of residues. Case III is more complicated, but we can split the substructure query into two parts and can reduce the problematic case III to the easy cases I and II. Thus, we can reduce a substructure search that would only map enumerated molecules to a case where we can search in the fragment space of cores and residues, using query fragments. This procedure of splitting substructure queries into fragments and perform the substructure search with each query fragment in the respective subset of the BICLAIM space is the central concept of our algorithm and motivated its name: We call it “BI Split Substructure Search” or BI S3 (spoken as BISCUBE).

The entire algorithm is implemented as a workflow in Pipeline Pilot [6]. Obviously, the crucial step in the implementation is the intelligent splitting of the substructure query. Here, we can make use of some simplifications, given by the construction of the BICLAIM space. As mentioned earlier in Section 13.2.6, the BICLAIM space consists of library core and corresponding residues. A core can be attached to up to three residues via a single bond. We explicitly exclude ring formation during the enumeration of cores and substituents. Therefore, the first operation in the workflow is splitting the substructure query and analyzing the resulting split schemes. This is done in three steps: In the first step, all bonds in the substructure query that could map core-residue-bonds in the enumerated molecules are identified. This list contains all single, noncyclic bonds. A query can contain bonds that are specified by query features like “ring bond.” As those bonds shall map rings, they obviously will



(a)

(b)



(c)

FIGURE 13.7 Example for the application of BISCUBE. A hypothetical synthesis scheme for Dabigatran (a) consists of a core with two residues and corresponding residue lists (b). To do substructure searches in the respective combinatorial library, one can search in the list of cores (c-I) or in the R2 residues (c-II). Query (c-III) combine parts of the core and parts of the R1 residues and therefore can only be found in enumerated libraries. BISCUBE splits this query into two sub-queries (c-IV) that can be used to search in cores and R1 residues, successively.

not map core-residue-bonds. Therefore, they are excluded from the list. In the second step, we generate split schemes by multiplying the substructure query and deleting up to three of the identified bonds per query. In the end, we have N query structures where N is the number of all possible combinations of up to three bonds from the list. Each of the queries consists of up to four fragments. The last step is the validation of the split schemes. A reasonable split scheme contains either two fragments, three fragments, or four fragments where one of the four fragments has three attachment atoms. The two-fragment scenario represents the case where one of the two fragments is the part of the core and the second is part of the residue. In the three-fragment scenario, the fragment with two attachment atoms must map the core and the remaining two fragments map the residues. The scenario with four fragments can be divided into one case where two fragments have two attachment atoms and one case where one fragment has three and the three remaining fragments have one attachment atom. The first case is not a valid split scheme as it is not covered by the design of the BICLAIM space.

The second operation is the analysis of the generated split schemes and the allocation of the fragments in each split scheme to the list of cores or residues of the BICLAIM space. If a split scheme contains two fragments, both fragments could map a BICLAIM core and the corresponding fragment should map the respective list of residues. Hence both fragments have to be checked whether they map a core and in case of a positive map, the respective second fragment of the split scheme has to be checked if it maps onto the respective residues. In the case of a split scheme containing three or four fragments, only the fragments with more than one attachment atom have to be checked against the list of cores. This way, we generate lists of fragments that have to be checked against all cores and lists of corresponding fragments that, in the case of a positive core match, have to be checked against the BICLAIM residues, associated with the respective core. Residue fragments can be assigned easily by the index of the split scheme and the attachment position to the core fragment of the split scheme.

In the case of dabigatran, the splitting operation generates 155 different and unique split schemes that are in agreement with the design of the BICLAIM space, generating 168 query fragments for search operations in the cores. Only one of the split schemes (indicated by the dotted lines in Figure 13.7a) contains a core that matches the library core, defined in Figure 13.7b. In summary, when using dabigatran as a substructure query to search in the combinatorial library defined in Figure 13.7b, 729 million comparisons have to be performed when searching in the enumerated space. Using BISCUBE, we search with 168 cores in the list of library cores (in the example only one) and finally with the associated R1 query fragment of the only core query fragment hit in the list of R1 residues and with the corresponding R2 query fragment in the list of R2 residues. This way, we have to perform only a few thousand search operations (#Cores + #R1-Substituents + #R2-Substituents), which is roughly a factor 10,000 less than the search in the enumerated space.

As the current version of BICLAIM covers about several thousand cores, the number of substructure hits can be pretty huge. This implies that the hits cannot be enumerated in all cases. Therefore the output of the results of the substructure

search can be selected by the user. To get a first impression, it is recommended to not enumerate the final hits, but to calculate the number and presenting the different library suggestions together with the number of hits from the respective library and one example for an enumerated molecule. This typically gives the user a first impression of the diversity of libraries, covered by the substructure query and of the hit distribution over all matching libraries. The user may now decide to refine the query by making it more specific or to enumerate the hits for a selected subset of libraries. Even the partial enumeration of a certain fraction of the entire hit set would be possible.

A typical substructure search with BISCUBE lasts about 30 min up to a few hours for less specific queries. Thus, it provides easy access to a huge space of synthesizable compounds optimized for combinatorial chemistry for the generation of ideas and for the design of project-related combinatorial libraries.

13.3.5 Automatic Annotation of Controlled Substances

The proper handling of controlled substances compliant to all legal restrictions is a task confronting all manufacturers, distributors, and pharmaceutical companies. Federal law limits the handling of these substances and by regulation requires that they be properly safeguarded at all times. Especially for companies operating in more than one country and with routine shipments between the different sites, this challenging task requires familiarity with the local laws of different countries. The industry is responsible for establishing and maintaining effective controls and procedures to prevent misappropriation. This does include purchase and shipment of compounds as well as their synthesis.

Obviously, chemists know very well about the chemical structures and regulatory implications of controlled substances by education. Nevertheless there are two reasons for an automated annotation workflow: The list of controlled substances changes casually, thus it might be difficult to stay up to date for all members of a large organization. Furthermore, the local law of other countries is not necessarily part of the education. To aid being consonant with the applicable law, we used Pipeline Pilot to establish a process that automatically monitors BIs compounds and checks for clashes with controlled substances. Some points have to be taken into account when setting up such a workflow:

- The identification component of the workflow has to be up-to-date all the time.
- All possible sources for compounds have to be identified.
- Depending on the source, it has to be ensured that the identification process is done either on-demand or continuously.

Beside all controls, procedures and practices, the first and most important step remains to reliably identify controlled substances. Luckily this step can be reduced to the task of a simple substructure search. For each controlled substance or for a group of controlled substances, specified in the respective schedule of the legal text, a substructure can be defined that can be used to flag potential controlled

substances. However, the devil is in the details: Rather than explicitly mentioning each single controlled substance, the legal text includes for most controlled substances "... their isomers, esters, ethers, salts, and salts of isomers, esters and ethers, whenever the existence of such isomers, esters, ethers, and salts is possible within the specific chemical designation." This expansion of the explicitly denoted examples of controlled substances is quite a challenge. So, the scope is to define a set of substructures that reliably matches all controlled substances while not tagging structurally similar derivatives. As the definition of the substructure pattern is the most precarious part, it has to be done by an authorized person with the necessary qualification. The monitoring of controlled substances is not restricted to compound synthesis. The purchases of vendor compounds as well as the ordering of reagents for the synthesis of compounds should be checked, too. While the latter two sample sets are updated manually every few weeks and therefore can be checked on demand and flagged in the corresponding databases, the synthesis of compounds should be observed continuously. Typically, chemists create a new experiment in their electronic laboratory notebook (ELN) before they start any synthesis. We thus ensure that every new entry from the ELN is sent to the established process which compares it against all defined substructure queries and sends the result back to the ELN.

The identification of controlled substances is a good example where the usage of workflow tools is extremely helpful to automate well defined and recurrent processes. Once all data flows are defined and the workflow is set up and productive, the annotation is done completely automated. But in the end it still is the human individual who finally has to decide how to proceed with the findings.

13.4 DISCUSSION

In this chapter, we described several different CI systems that are currently in use at BI, together with some application examples for CI methods. Many of the presented methods were intentionally selected to be relatively simple and straightforward, particularly in comparison to the complex and elaborate methods usually published in contemporary literature. In an industrial environment, the focus is usually not on the development of completely novel methods, but on the evaluation, validation, and adaptation of methods according to the specific needs of an organization. In many cases, a significant amount of work is spent on evaluating promising new methods using internal datasets. To establish a new method within a large research organization, it is necessary to gain a significant amount of experience and to slowly build up trust into the new method. A new method will often only be accepted for daily project work if a couple of positive internal examples exist and also a good understanding of the methodological limitations is given. For very abstract methods that are remote from the usual thinking of experienced medicinal chemists, it is often difficult to get acceptance while intuitively understandable methods can be adopted very quickly. In our experience, this is—for example—a big advantage of SAR methods relying on MMPs (cf. Section 13.3.3). On the other hand, one should keep

in mind that it can be difficult to contribute novel or unexpected insights to a research project if one focuses too strongly on methods that resemble just the intuitive thinking of medicinal chemists.

Some of the mentioned methods (e.g., from the area of SAR analysis) were developed in collaboration with academic groups. Through collaboration with external partners, it is possible to stimulate new, tailor-made methodological developments in the field. Both academia and industry benefits: Industry gets timely access to academic innovation while academia receives feedback from experienced practitioners.

Many recent publications from the field of CI are based on a competitive mindset and aim at ultimately replacing experiments with computations. In our experience, this rarely ever happens. Instead, computational approaches are typically seen as additional and complementary sources of information that can in some cases be used to prioritize experiments or to move experiments to a later phase. Thus, we expect of lot of impact from methods that combine and integrate computational and experimental approaches in intelligent ways. To achieve this, CI researchers will have to embrace experimental errors and uncertainties, but also biases and simplifications inherent to *in silico* methods. Along these lines, computational methods could, for example, be used to explain why certain experiments fail or generate inconsistent results while tailored experiments could be performed to derive correction factors for certain computations.

Another highly relevant area of research concerns the analysis of multiparameter SAR. In the course of a lead optimization project, the number of considered endpoints and parameters grows considerably. This makes it likely to observe tradeoffs between different parameters that cannot be optimized separately due to parallel SAR. Furthermore, one is often forced to work with incomplete or even sparse datasets—most compounds are characterized with respect to some endpoints, but a full profile is available only for a small fraction of the compound set. Methods that support the analysis of such complex optimization scenarios in a convenient, efficient, and intuitive way are highly welcome.

An important step is to integrate newly established methods into the existing CI infrastructure. Due to strict timelines and high workloads in daily project work, preference is often given to methods that work reliably, predictably and that can be applied and analyzed in a convenient way. Therefore, it is often necessary to integrate CI-methods seamlessly with tools and systems supporting the practical parts of the research process. In our view, this explains the great success of workflow tools within pharmaceutical industry to a large extent. They enable the combination of independent functions into larger protocols: For example, the output of an analog search module can be fed into a component checking the physical availability of compounds.

If a new CI method is established internally, an important question is whether it should remain an “expert technology” that has to be applied and interpreted by the computational chemist as a service for colleagues from other disciplines or if it is suitable for deployment to a wider user base. This discussion is facilitated by the fact that modern workflow tools allow for a simple exchange of workflows and protocols

between experienced and less experienced users and for a rapid prototyping and deployment of new tools. In this chapter, we showed examples for both cases: Services such as the calculation of molecular descriptors, QSAR predictions, or the classification of controlled substances are made available as self-service tools to all interested end users. This makes it essential to include a robust error handling and interpretation aids into the services. For example, in deployed QSAR models, applicability domain checks have to be performed to avoid the generation and usage of irrelevant predictions. Additionally, it is important to supplement deployed services with a comprehensive documentation that should also include a contact person. Other services, such as the analysis of HTS campaigns (BIMESH) are considered to be too complex and to require too much manual, project-specific intervention to be deployed as an automated tool. The role of a computational chemist within a research project should be to ensure that all relevant data are found, made available, and used for decision making. This requires not only a good overview of the available systems and databases (internal and external), but also a sound understanding of methods from the fields of data mining, statistics, and visualization. Additionally, one should not underestimate the persistence needed. For example, it usually takes multiple rounds of analog searches and biological activity testing until a new, interesting compound is identified or a sufficient understanding of the SAR within a structural class is obtained.

In summary, we hope that the examples in this chapter could show the broad range of topics a CI researcher is confronted with in an industrial environment. There is a variety of ways how CI can contribute significantly to pharmaceutical research projects. CI remains an interesting area with a number of open challenges to be solved, limited understanding of SAR/SPR in all of its complexities being only one of them.

REFERENCES

1. Agrafiotis DK. Advanced biological and chemical discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J Chem Info Model* 2007;47: 1999–2014.
2. Sander T, Freyss J, von Korff M, et al. OSIRIS, an entirely in-house developed drug discovery informatics system. *J Chem Inf Model* 2009;49(2):232–246.
3. Stahl M, Guba W, Kansy M. Integrating molecular design resources within modern drug discovery research: The Roche experience. *Drug Discov Today* 2006;11:326–333.
4. Gobbi A, Funeriu S, Ioannou J, et al. Process-driven information management system at a Biotech Company: Concept and implementation. *J Chem Inf Comput Sci* 2004;44(3): 964–975.
5. Accelrys Inc. SYMYX Direct 7.0. San Diego: s.n.; 2011.
6. Accelrys Software Inc. Pipeline Pilot, Version 8.0. San Diego: s.n.; 2011.
7. Berthold MR, Cebron N, Dill F, et al. KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization* (GfKL 2007), Freiburg. Heidelberg: Springer; 2007.

8. Kramer C, Beck B, Clark T. Insolubility classification with accurate prediction probabilities using a MetaClassifier. *J Chem Inf Model* 2010;50(3):404–414.
9. Kriegl JM, Arnhold T, Beck B, et al. A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J Comput Aided Mol Des* 2005;19(3):189–201.
10. Kramer C, Beck B, Kriegl JM, et al. A composite model for HERG blockade. *ChemMedChem* 2008;3(2):254–265.
11. Lipinski CA, Lombardo F, Dominy BW, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 2001;46:3–26.
12. Lovering F, Bikker J, Humbel C. Escape from Flatland: Increasing saturation as an approach to improving clinical success. *J Med Chem* 2009;52(21):6752–6756.
13. Lipinski CA. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discov Today Technol*. 2004;1(4):337–341.
14. Oprea TI, Davis AM, Teague SJ, et al. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* 2001;41:5.
15. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* 2007;6:881–890.
16. Proudfoot JR. The evolution of synthetic oral drug probe. *Bioorg Med Chem Lett* 2005;15(4):1087–1090.
17. Veber DF, et al. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* 2002;45(12):2615–2623.
18. Oracle, Inc. [Online] www.oracle.com.
19. Spotfire AB. Spotfire DecisionSite 9.1.1. 2008.
20. MDL. ISIS Base 2.5/ SR 3. 2003.
21. Hajduk PJ. *J Med Chem* 2006;49:6972–6976.
22. Hajduk PJ. *Nat Rev Drug Discov* 2007;6:211–219.
23. Hajduk PJ. *J Am Chem Soc* 1997;119(25):5818–5827.
24. Swann S. *ACS Med Chem Lett* 2010;1(6):295–299.
25. Navratilova I. *ACS Med Chem Lett* 2011;2(7):549–554.
26. Klebe G. *Drug Discov Today* 2006;11:580–594.
27. Koeppen H. *Methods Princ Med Chem* 2011;48:61–85.
28. Fink T. *J Chem Inf Model* 2007;47:342–353.
29. Reymond JL. *MedChemComm* 2010;1(1):30–38.
30. Lessel U. *J Chem Inf Model* 2009;49(2):270–279.
31. Herdemann M. Optimisation of ITK inhibitors through successive iterative design cycles. *Bioorg Med Chem Lett* 2011;21:1852.
32. Macarron R, Banks BN, Bojanic D, et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* 2011;10:188–195.
33. Fox S, Farr-Jones S, Sopchak L, et al. High-throughput screening: Update on practices and success. *J Biomol Screen* 2006;11(7):864–869.
34. Bleicher KH, et al. A guide to drug discovery: Hit and lead generation: Beyond high-throughput screening. *Nat Rev Drug Discov* 2003;2:369–378.
35. Carr R, Jhoti H. Structure-based screening of low-affinity compounds. *Drug Discov Today* 2002;7(9):522–527.

36. Milligan G. High-content assays for ligand regulation of G-protein-coupled receptors. *Drug Discov Today* 2003;8(13):579–585.
37. Bajorath J. Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* 2002;1:882–894.
38. Koeppen H. Virtual screening – What does it give us? *Curr Opin Drug Discov Devel* 2009;12(3):397–407.
39. Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD-96); 1996 Aug 2–4; Portland. p 226–231.
40. Stahl M, Mauser H, Tsui M, et al. A robust clustering method for chemical structures. *J Med Chem.* 2005;48(13):4358–4366.
41. Stahl M, Mauser H. Database clustering with a combination of fingerprint and maximum common substructure methods. *J Chem Inf Model* 2005;45(3):542–548.
42. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 1996;39(15):2887–2893.
43. Wilkens SJ, Janes J, Su AI. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J Med Chem* 2005;48(9):3182–3193.
44. Beck B. BioProfile—Extract knowledge from corporate databases to assess. *Bioorg Med Chem* 2012;20(18):5428–5435. <http://dx.doi.org/10.1016/j.bmc.2012.04.023>. Accessed 2013 May 14.
45. Wawer M, Lounkine E, Wassermann AM, et al. Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov Today* 2010;15(15):630–639.
46. Peltason L, Iver P, Bajorath J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *JCIM.* 2010;50(6):1021–1033.
47. Sheridan RP, Hunt P, Culberson JC. Molecular transformations as a way of finding and exploiting consistent local QSAR. *JCICS* 2006;46:180–192.
48. Leach AG, Jones HD, Cosgrove DA, et al. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem.* 2006;49:6672–6682.
49. Warner DJ, Griffen EJ, St-Gallay SA. WizePairZ: A novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *JCIM.* 2010;50(8):1350–1357.
50. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *JCIM* 2010;50(3):339–348.
51. Rarey M, Dixon JS. Feature trees: A new molecular similarity measure based on tree matching. *J Comput Aided Mol Des* 1998;12(5):471–490.
52. Ehrlich HC, Rarey M. Searching substructures in fragment spaces. *J Chemoinf* 2011;3(Suppl 1).
53. ChemAxon Ltd. Markush Search and Enumeration. <http://www.chemaxon.com/products/markush-ip/>. Accessed 2013 May 14.
54. Digital Chemistry. Markush structures and combinatorial library analysis. http://www.digitalchemistry.co.uk/prod_markush.html. Accessed 2013 May 14.

55. Huel NH, Nar H, Priepeke H, et al. Structure-based design of novel potent nonpeptide thrombin inhibitors. *J Med Chem* 2002;45(9):1757–1766.
56. Wang Y, Xiao J, Suzek TO, et al. PubChem's bioassay database. *Nucleic Acids Res* 2012; 40(Database issue):D400–D412.
57. PerkinElmer. Registered Trademark of PerkinElmer. Waltham: s.n.
58. Molecular Devices, LLC. Registered Trademark of Molecular Devices, LLC. Sunnyvale: s.n.
59. CisBio Bioassays, IBA group. Registered Trademark of CisBio Bioassays. Louvain-la-Neuve: s.n.

CHAPTER 14

LESSONS LEARNED FROM 30 YEARS OF DEVELOPING SUCCESSFUL INTEGRATED CHEMINFORMATIC SYSTEMS

MICHAEL S. LAJINESS and THOMAS R. HAGADONE

14.1 INTRODUCTION

Following World War II, pharmaceutical companies started to invest in the synthesis of relatively large numbers of new compounds in the search for new therapeutic agents. Associated with this increase in numbers was the need to maintain records on these molecules and their biological properties. Records were initially in paper-based systems that were labor intensive to maintain and almost impossible to search. However, by the mid-1970s, computer hardware and software technology had advanced to the point where it was possible to envision a graphics-based system for storing and searching a company's collection of chemical structures and discovery biology data. There were no commercially available cheminformatics database systems (CIDBSs) at the time; so a few intrepid companies started internal projects to construct in-house systems. Even today, there are very few commercial CIDBSs available and startlingly few examples of proprietary solutions that are well regarded, have stood the test of time, and are considered to be successful.

Three examples of successful systems are Cousin, ChemLink, and Mobius, all of which were developed, maintained, and supported by the authors over the past 30+ years. There have been no other systems that we are aware of that have been in nearly continuous successful operation for as long a time period. Thus, it seems appropriate in this forum to examine some of the history behind these successful developments and relate some of the lessons learned along the way.

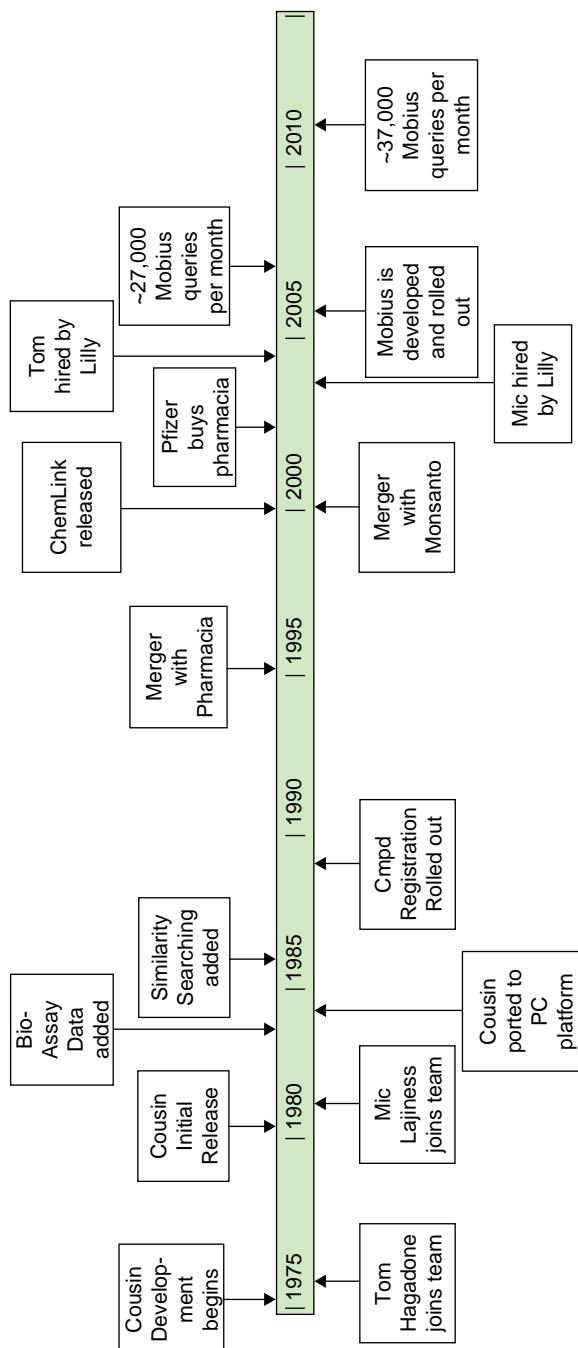


FIGURE 14.1 Timeline (1975–2012). Milestones in the development of Cousin, ChemLink, and Mobius.

In this journey, we will first cover Upjohn/Pharmacia's systems: Cousin (1981–2001) and ChemLink (2001–2003). We will then discuss several less successful attempts at development including Pfizer's RGATE (2003–2004) and Lilly's Beacon Software (2001–2005). Finally, we will describe the successful development and deployment of Lilly's Mobius system (2005 till present). We will then summarize and highlight some of the most important lessons learned over the many years concerning the development and deployment of CIDBSs. To facilitate a better understanding of this history of development, a timeline is included in Figure 14.1.

14.2 HISTORY

14.2.1 Cousin: 1981–2001

One of the earliest examples of a proprietary CIDBS was begun at The Upjohn Company in 1974. The initial goal was a basic one; create a system that would allow the company's 50,000 structures to be drawn in and stored in a database in their native graphical form and searched via graphically entered full structure and sub-structure search queries. Search response was expected to be interactive (i.e., less than 30 sec per substructure search).

The system that was developed, Cousin, consisted of state-of-the-art (at the time) hardware, Digital PDP-11 graphics workstations with tablets for structure drawing (see Figure 14.2 for a picture of an original Cousin Workstation and example display), an IBM 370 mainframe for structure storage, and a separate dedicated PDP-11 with a relatively fast disk for structure searching. The structure drawing software was based on the work of the LHASA project [1]. It was sophisticated for its time and provided intuitive chemical structure drawing and graphical capabilities for defining complex R-group queries [2]. Structures, registry numbers, and associated data were stored on the mainframe in an experimental relational IBM database system called System-R that included a new relational query language called Sequel (later changed to Structured Query Language (SQL)) [3]. Structure and substructure searching software was developed using a set of structure searching algorithms optimized for the PDP-11 hardware [4].

The initial version of Cousin and its associated database were built over a period of 6 years, long by today's standards, and released to the user community in 1981. Even though there were only two of the relatively expensive graphics workstations for the entire chemistry community, the system proved to be popular. Cousin was continuously expanded to include additional discovery research data and functionality over a 20-year period from 1981 through 2001. Some of the key developments during this time were

- 1983—Integration of bioassay data into Cousin via the associated relational database system.
- 1984—Replacement of the PDP-11 workstations with widely available IBM PCs resulting in a large increase in usage.

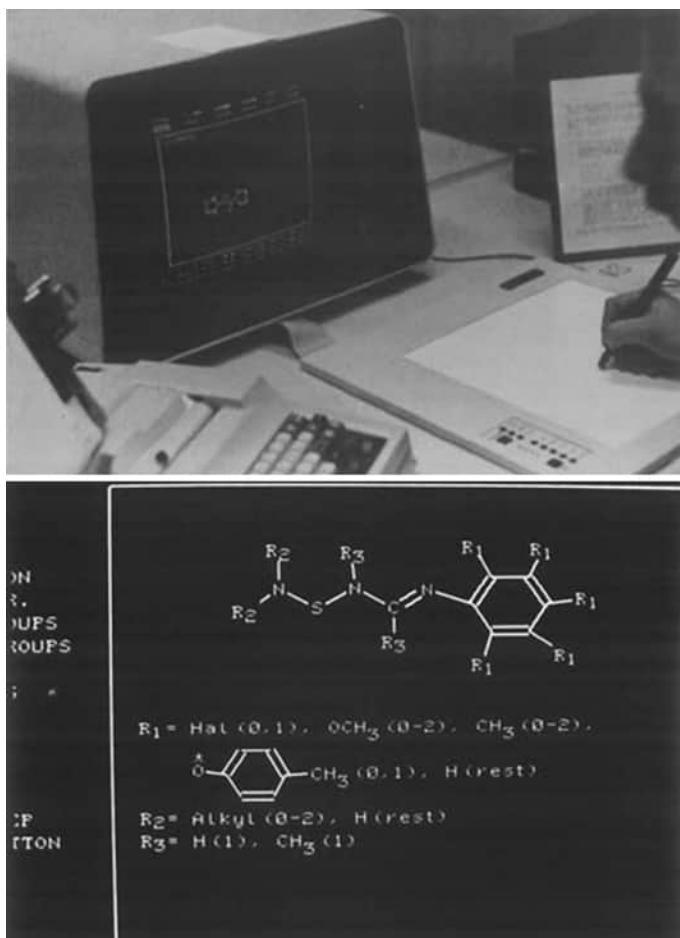


FIGURE 14.2 Original Cousin Workstation. The workstation consisted of a Digital PDP-11 graphics workstation with a Talos tablet for structure drawing.

- 1984—Formalization of the integration of the chemical structure data type into the relational database system as a data type on par with the built-in types [5]. This approach is a standard practice today and an active area of development in the open source community but was novel at the time.
- 1984—Support for publication-quality drawing of synthetic flow schemes and other chemistry documentation.
- 1984–1990—Conversion of the application code from PL/I, FORTRAN, and assembler to C++ and Visual Basic. Conversion of the relational database system from System-R to SQL/DS to Oracle.
- 1986—Integration of full-structure similarity searching [6].

- 1988—Online interactive registration of new compounds by end users.
- 1991—Integration of substructure similarity searching which locates compounds containing structures similar to a substructure query using an optimized MCS algorithm [7].
- 1995—Publication of a free Microsoft Access add-in based on Cousin that allowed structures to be stored, searched, and retrieved using the Microsoft Access database engine and user interface [8].

Twenty years is a relatively long time for a custom in-house-developed system of this type to survive. This is particularly true since viable commercial chemical database software came into existence in the late 1970s in the form of the MACCS system, which was widely adopted by industry, but the object of a great deal of user dissatisfaction. Cousin was regularly reviewed by both inside and outside reviewers against other available systems and remained the recommended choice. We believe that some of the reasons for this longevity were the following:

- The system *evolved rapidly* to include new data types and functionality and was *fast and reliable*. This was greatly facilitated by being a proprietary system.
- The user interface was *easy-to-use* for novices but included features for the advanced user. It evolved incrementally over the 20-year period with no major disruptions to its look and feel.
- The early incorporation of chemical structures into the relational database and SQL provided a *strong technical foundation* to build on.
- Having biological data alongside chemistry data was a critical step forward in the understanding of structure–activity relationships.
- The ability to perform ad hoc queries was a powerful way to facilitate decision making by allowing a user to quickly define and execute both simple and complex queries.
- The development and support staff, which ranged from two to five people over the years, maintained a *high level of discovery research, computer science, and chemistry–software expertise*. The team was stable in membership and was very responsive to user requests over its nearly 20 years of existence.
- Cousin development and support was always *organizationally on the science side of research* rather than on the information technology side. This allowed the Cousin team to be more science-oriented and protected it from several potentially disruptive IT reorganizations.
- Becoming operational in the 1980s, it became the de facto standard for viewing pharmaceutical discovery data within Upjohn.
- The company was stable and merger-free until 1995.

14.2.2 ChemLink: 2001–2003

In 1995, Upjohn merged with Pharmacia to form Pharmacia and Upjohn and then, in 2000, Pharmacia and Upjohn merged with Monsanto to form a new company, Pharmacia. In each of these mergers, the discovery chemistry and biology databases

were merged to create an integrated global company database. Reviews of each of the companies' database structures and interfaces were performed by the scientific community to select a design for the new merged companies.

At this point, the aging Cousin drawing program and substructure search engine really offered no significant advantages over the commercially available ISIS components. However, the way Cousin integrated multiple data sources, the associated ad hoc user query interface, and its extensibility features were seen as continuing to add value. A decision was made to develop a new interface, called ChemLink, which was based on the Cousin design but used ISIS components for structure drawing and searching. From this point forward, data and tool integration and the associated user interface became the main focus for continued development. Following the second merger, this decision was revisited and confirmed. ChemLink was completed and was well-received by the scientists from all three predecessor companies. It was heavily used for 3 years until the next acquisition occurred.

Proprietary software allows one to choose what to develop. In developing ChemLink, we were able to integrate a number of cheminformatic tools that helped scientists to better exploit their data. Examples of tools that were developed and integrated into ChemLink included Dissimilarity (Diversity) Searching [9] and Structure–Activity Landscape Analysis—the predecessor to SALI [10].

An interesting aspect of the change from Cousin to ChemLink was that the GUI now better-supported mouse navigation in place of keyboard control. This more modern style of input was met by some initial resistance but was soon embraced by all. Overall, Cousin/ChemLink enjoyed a high level of usage. Figure 14.3 gives a graph of the usage statistics for Cousin/ChemLink.

Another interesting thing that occurred during this time of mergers was how people tended to embrace the system that they knew, at least for a time. Scientists at Pharmacia and then at Monsanto were used to their local CIDBSs, but once familiar with Cousin/ChemLink they were quick to embrace it. Interestingly, but not surprisingly, those in charge of developing the CIDBSs were not in favor of the alternate software and fought hard to isolate it. Getting the teams from the merged companies to work together on a common system was very difficult to achieve.

Some of the reasons that ChemLink continued to be the key CIDBS at Pharmacia include the following:

- By continuously, but judiciously, adding user requested and other enhancements, we were able to keep the application and user interface relevant and “fresh.”
- Increased use of standard external components (like ISIS/Draw) allowed the developer(s) to focus on strategic enhancements.
- Integration of additional data sources provided access to an ever-expanding collection of data that made ChemLink the de facto source for discovery data.
- The ability to adapt to changing business environments was facilitated by well-designed and compartmentalized code.
- A stable and well-coordinated support and development staff is essential for navigating disruptive business conditions.

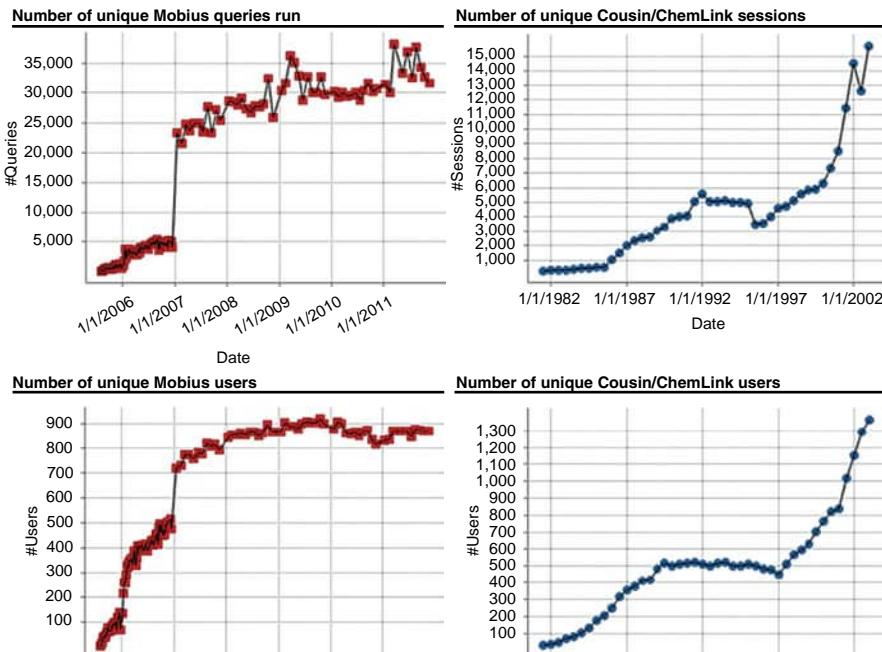


FIGURE 14.3 Usage Statistics for Mobius and Cousin/ChemLink. Usage statistics from Cousin/ChemLink and Mobius. Number of sessions are reported for Cousin/ChemLink versus number of queries for Mobius. On average in our experience there are approximately 2.3 queries run per session. For color details, please see color plate section.

14.2.3 RGate: 2003+

In 2003, Pfizer acquired Pharmacia. Once again, corporate databases were merged and the chemical database interfaces were reviewed by research teams from the various company sites. A new design, based mainly on ChemLink, called RGate, was created and a large IT team was charged with developing the new application. RGate was developed as a Java-based web application even though the application which it most sought to emulate, ChemLink, was a client–server design.

The authors left Pfizer before RGate was complete; however, it is our understanding that the initial implementation of RGate was deemed unacceptable and that it was eventually rewritten as a client–server application.

Some of the lessons learned from the RGate experience include

- A large team of developers is not necessarily better than a small well-focused team.
- One needs to choose the right technologies (e.g., client–server) at the right time to ensure successful development.
- Totally redesigning and rewriting a large, successful application entails significant cost and risk of failure.

14.2.4 Beacon Projects (~2000–2004)

Around 2000 Eli Lilly made the decision to replace their aging ISIS-forms-based system, Icaris, with something more modern. Two attempts were made to develop custom replacement applications, neither of which proved successful. Both attempts employed a formal IT approach and utilized extensive lists of requirements from diverse user groups including many high-level scientists and managers. The first system, Beacon, was based on Spotfire Decision Site and was innovative in that it integrated graphical visualizations along with traditional tabular data retrieval. However, two issues ultimately led to the demise of the project: the inability to effectively integrate a range of different data sources, and, the inability to adequately deal with chemical structure information.

These issues alienated the chemistry community which then prompted the development of a second Beacon Project, Beacon forms. Since the existing Icaris system was built using the form-oriented ISIS/Base and the first Beacon project did not have forms and it “failed” for chemistry, the lesson was assumed to be that chemists must like and need a forms-based system. Thus another project was formed to create a java application that implemented an optimized forms-based environment. Unfortunately, the first production version was extremely slow, awkward to use, fragile, and was not able to access the full range of important data sources. Coincidentally, a stealth project was underway at the time to develop a proof-of-concept CIDBS for Lilly based on ChemLink that eventually became Mobius.

Some lessons to be learned from the Beacon experiences include

- A formal IT process and long list of requirements does not compensate for a lack of scientific expertise in the developers.
- It is more important to have tools in hand that can be used together to forge a solution rather than a set of disparate functions that coordinate poorly.
- One needs to ensure that you have the right expertise to evaluate the best companies to partner with and the right mix of technologies to ensure successful development.

14.2.5 Mobius: 2005 Till Present

The authors moved to Lilly in circa 2004 and almost immediately started a side (stealth) project to develop a small proof-of-concept application along the lines of Cousin/ChemLink. When the second Lilly effort was deemed unsuccessful, this proof of concept was shown to management as a possible path forward and was given the go-ahead for further development. This effort, called Mobius, was initially completed and put into production at Lilly in late 2005.

Mobius presented a somewhat different situation than the Cousin/ChemLink scenario. With Cousin/ChemLink, there were no other CIDBs in existence—thus, users had NO prior experience or expectation, but with Mobius, several different systems had been in use prior to its introduction. In addition, Mobius employed a decidedly

different ad hoc form of user interface and experience. This necessitated a somewhat careful approach to its rollout.

When Mobius was first put into production, it accessed about five different data sources. Our thought at that time was to get Mobius out as fast as possible to begin to get value and build enthusiasm. Subsequently, we planned to add additional data sources as time and resources permitted. Since prior CIDBSs did not access a wide variety of data, multiple independent systems had been developed across the corporation to access these islands of data. Once the Mobius “beachhead” was established, it became increasingly easy to get permission to integrate additional sources of data and as the number of data sources grew the added value of this integration became apparent.

The initial introduction of Mobius to the user community utilized the apostle approach: select highly motivated and influential individuals; give them extensive training; and then let them go forth and spread the gospel. Early on in the rollout, user experiences were documented and rapid code changes occurred. Daily updates were not uncommon. Effective interactions with prospective users are, of course, extremely important but to be useful, they have to be perceived to be potentially useful by the prospective user. Thus, rather than require training, the Mobius philosophy was to allow access to any valid user but to provide access to training and training materials in a variety of formats. These include periodic large group training lectures that cover introductory, intermediate, and advanced issues; Mobius News—an online publication that provides information on enhancements and changes; short training videos; on-line help information; and training sessions tailored for specific groups such as Toxicology, Analytical Chemistry, and so on. Using this approach, we have reached a high level of penetration in the Discovery user community and have achieved a high-level of customer satisfaction.

Interactions pertaining to support and maintenance are by their nature more sporadic but no less important. Systems changes occur frequently and related software bugs sometimes require quite a bit of effort to resolve. Problems or issues that arise that involve the Mobius system are dealt with as fast as possible. Often however, the underlying problem lies outside Mobius and involves faults in related databases and other non-Mobius infrastructure. Regardless of the origin of the issue, the most important thing to keep in mind is fast problem resolution. We have found that when users have an issue and they contact you that it is a great opportunity. It is a chance to show that you care about them and that resolving their problem is important to you. When resolving difficulties it is also a great opportunity to understand how they are using the system and to offer suggestions, comments, and perhaps a bit of training to help them do their jobs better. It does take valuable time, but the benefits of having experienced expert help trumps outsourcing the problem resolution function. One benefit of this approach is that often during user interactions ideas for new enhancements come up. It is a great time to explore these ideas and record them. Users are often impressed when you take the time to capture their thoughts and explore their ideas. It helps them feel “ownership” in the software. Turning a problem into a happy user experience can pay big dividends as attested to by the Mobius usage statistics in Figure 14.3.

Some of the lessons learned from the Mobius experience include the following:

- Developing software like Mobius “in-house” may have to be done at least partially in stealth-mode to avoid early political battles.
- Our experience once again illustrates that CIDBS development can often best be done within a science-based organization rather than in an IT organization.
- Development needs to be done by people that know the science, the business process, are technically gifted, and are very much customer focused.
- An ad hoc query interface can offer significant advantages over a forms-based interface.
- A user interface that is truly easy to navigate coupled with access to a broad range of data translates to greater usage and a happier user experience.
- It is surprising how many people seem to believe that a system like Mobius should be developed and then “transitioned to support” rather than being allowed to continuously evolve to meet the ever-changing needs of research.

14.3 KEYS TO THE SUCCESS OF MOBIUS: A TECHNICAL PERSPECTIVE

In many ways, Mobius is the culmination of many years of software development, experience, and observation. Thus, it is instructive to use Mobius as an example to discuss some critical aspects of a CIDBS from a technical point of view including

- Data sources
- Metaview
- Query engine
- Ad hoc query interface
- Software components

14.3.1 Data Sources

The most important attribute of a CIDBS in a pharmaceutical research setting is the range and completeness of the data sources that it provides and the way those sources are organized and presented to the user. When a user thinks about a CIDBS, they are really thinking about the data that it encompasses. The user interface and its associated software features are really secondary. The data sources appropriate for inclusion in a CIDBS that were present at Lilly were typical of those available at a large pharmaceutical company that had been doing research for many years. They include the following:

- A large collection of proprietary company registry numbers, chemical structures, and associated submission information (e.g., who, when, where, third parties involved, restrictions, library identifiers)

- Inventory information on which compounds were available for testing (e.g., quantities, formats, locations, restrictions)
- Measured physicochemical properties (e.g., solubility, pK_a , $\log P$, mass spectroscopy, NMR) and co-crystal x-ray structures
- In vitro screening and secondary assay results
- In vivo preclinical animal ADME and toxicology data
- In silico predictions for physicochemical properties, lead/drug models, ADME/toxicology models, and activity models
- Project information describing hypotheses, where compounds are at in a project and decisions made with respect to templates and particular compounds
- Miscellaneous special purpose databases (e.g., ACD, GeneGo, ChEMBL, custom databases of literature data built under contract, external chemistry synthesis tracking database, a high throughput reagent database)

Each of the aforementioned data sources was created and evolved independently over time based on the needs and demands associated with each individual system. In many cases, each system had its own interface in addition to its own data and queries across systems and data sources were difficult or impossible. Most of the data were available in Oracle databases although a significant amount existed only in Excel workbooks and other non-database files. The Oracle data were spread across over 100 database schemas stored on 24 individual database instances. Because of time and resource constraints and our design philosophy, a decision was made to employ a federated approach in which data would be accessed from its original location whenever possible but integrated into a unified view via the Mobius software.

The data in Excel workbooks and other files presented a particular challenge. Because it was stored in files, often on individual's personal machines, these data did not provide acceptable reliability, availability, and serviceability/scalability/security for a shared CIDBS. There were three possible solutions: (1) do not integrate the file-based data, (2) integrate the data in a semiformal way managed by the user, or (3) integrate the data as a full scale IT project. The first solution was unacceptable since it did not meet our goal of maximum data integration. It would just perpetuate the Excel-as-a-database-system problem that we were trying to eliminate. The third solution was not considered feasible for reasons of resource and time constraints.

We chose the second solution and implemented it in the form of "annotation tables." From the user's perspective, an annotation table is simply a table of rows and columns that reflects the data stored in an associated Excel worksheet or a file. An annotation table normally contains a registry number column that is used to link the table to the rest of the database and a set of additional columns in the source file that define the data of interest. To create an annotation table, the user simply supplies the name of the file, a project that it should be linked to, and a data security setting to Mobius. The file is then imported into a set of Oracle tables that hold annotation table data where it is then available for use on par with the other available data sources. Mobius can check for changes to the source file and update the table in Oracle whenever the source file changes if desired. Additionally, annotation table data can be

entered and edited directly within Mobius query results views. This feature has proven to be very useful and we typically see approximately 1500 active annotation tables in use each year.

An additional data source that Mobius allows users to create is the calculated field. A calculated field is simply a computed expression of arbitrary complexity involving one or more existing database fields; for example, the ratio of two IC₅₀ values for two different assays. The user can optionally apply a classification method to the output value of a calculated field to categorize the result. For example, the base calculated field could simply consist of a chemical structure column and the classification could assign a chemical template name to the structure based on a set of user-defined substructure query classification rules. As with annotation tables, calculated fields can be associated with project data and assigned security attributes. Both annotation tables and calculated fields are full members of the overall integrated database and can be searched, sorted, and combined with any other data in the database.

14.3.2 Metaview

As mentioned earlier, the data sources that underlie Mobius were created at different times by different groups for differing purposes using different data modeling conventions and each presents its own view of the world. To bridge the differences between the sources, Mobius creates an overarching metaview that makes the sources appear as a unified whole from the user's perspective that can be queried and manipulated. The Mobius metaview provides a hybrid relational/hierarchical model of data. Relational in the sense that each data source is viewed as being composed of a set of simple tables and hierarchical in that the tables are hierarchically interrelated and queries return hierarchical result sets rather than flat relational result sets. Hierarchical result sets avoid the "Cartesian product" problem that can occur when standard relational joins produce undesirable duplicate results.

The metaview is defined by a collection of metadata which describes the specific mapping from each underlying data source to the metaview. There are two fundamental types of metadata that are employed; metatables and the metatree.

14.3.2.1 Metatables The metatable defines the mapping from each data source to a set of simple tables (i.e., metatables). The metatable column data types consist of the common database types along with extended types for chemical structures, qualified numbers, and images (e.g., concentration response curves). A qualified number type consists of a basic number along with an optional qualifier (e.g., >) and statistics for summarized values. Results are often reported with qualifiers and/or as summarized values. The use of the qualified number data type can reduce the number of columns that are present in the user interface by up to 75%, thus reducing clutter.

The data for a single data source may be presented in multiple views, each of which can be usefully employed depending on the users' needs. For example, in vitro data are commonly stored in an unpivoted "tall-skinny" form. Within Mobius, these data are presented both in its native unpivoted form and in a pivoted form where a

metatable is available for each assay with an appropriate set of result types displayed based on the results that the assay reports.

14.3.2.2 Metatree The metatree organizes the metatables for the set of structure collections and research projects into a single hierarchical tree that is a representation of the users' natural models of how the data interrelate. The hierarchical relationships may be explicit in the underlying data sources but are often implicit. For example, the Mobius metatree organizes in vitro bioassay data in three alternative views:

- By therapeutic area, project, and assay
- By gene family, target, and assay
- By pathway, target, and assay

Included in the tree and associated with the proper project are any user-defined annotation tables, calculated fields, saved queries, and saved registry number lists. In this way, all of the basic project data, user additions to the data via annotation tables, and calculated fields, queries, and lists are kept together in one location. A user selects their main project as their default location and the tree is opened to that project each time Mobius starts.

The Mobius metatree is relatively large containing 17,000 distinct metatables (most of which are in vitro bioassays), 4500 annotation tables, 5800 calculated fields, and 19,000 saved queries in a single view; however, its organization and associated search tools make it straight forward for users to navigate to the data that they need.

14.3.3 Query Engine

The query engine has the job of decomposing queries on the metaview into native SQL subqueries that are then executed against each data source with the returned results recomposed into hierarchical results conforming to the metaview. Queries are passed to the query engine in a hierarchical dialect of SQL called Mobius Query Language (MQL). MQL is essentially SQL without the join criteria included. These join criteria are implicitly supplied by the metadata and factored into the native SQL that is produced for execution. This approach allows efficient queries to be executed across the full set of data sources including annotation tables and calculated fields. New data sources can be integrated by hand coding the metatable/metatree XML or by supplying a "metafactory" class that performs this task for the more complex sources. The internal operation of the query engine, although straight forward, is complex and beyond the scope of this chapter.

14.3.4 Ad Hoc Query Interface

A major goal in the design of the Mobius interface was to make querying as easy as possible for the user. To help achieve this goal, two types of queries can be defined and executed from the initial screen; a simple Quick Search or a normal query consisting of one

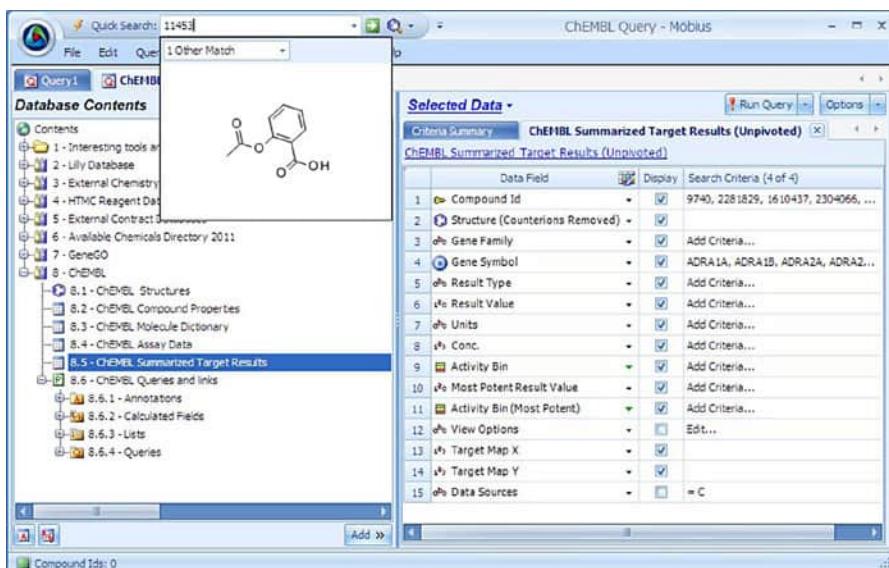


FIGURE 14.4 Mobius Ad hoc Query Interface. Example of the ad hoc query interface in Mobius. For color details, please see color plate section.

or more metatables with criteria. The main Mobius screen containing the Quick Search text box, metatree (Contents Tree) and query builder areas, is illustrated in Figure 14.4.

For a Quick Search, the text box shown in the upper left corner of the figure is used. Its operation is somewhat similar to the Google search line. As the user types, Mobius continuously monitors the entered text and performs two types of searches; one over the metatree and the other over the registry numbers of the databases that appear in the metatree. If one or more matches are found in the metatree, Mobius displays a dropdown box of the matches in Google fashion where they can be selected and operated on. If a registry number match is found, the structure corresponding to the registry number is displayed. If the enter key is pressed while a structure is displayed, Mobius builds and executes a query to retrieve all available data for the specified compound and displays it in a pivoted-by-assay format.

Normal query building, beyond what is possible with Quick Search, is performed in an ad hoc fashion by simply dragging individual metatables from the metatree on the left, or the Quick Search dropdown box, to the query table list on the right. The columns present in the currently selected query table are displayed and can be selected for retrieval, have criteria defined, sorting specified, and/or formatting configured. In the interest of simplicity, a traditional custom forms-based interface is not used.

When Run Query is clicked, the list of tables, columns, criteria, and sorting information is converted into an MQL statement, which is then executed by the query engine and the results, such as those shown in Figure 14.5, are displayed. In addition to the common tabular displays of results, a collection of useful graphical visualization formats are available. Also, alerts can be defined on queries such that

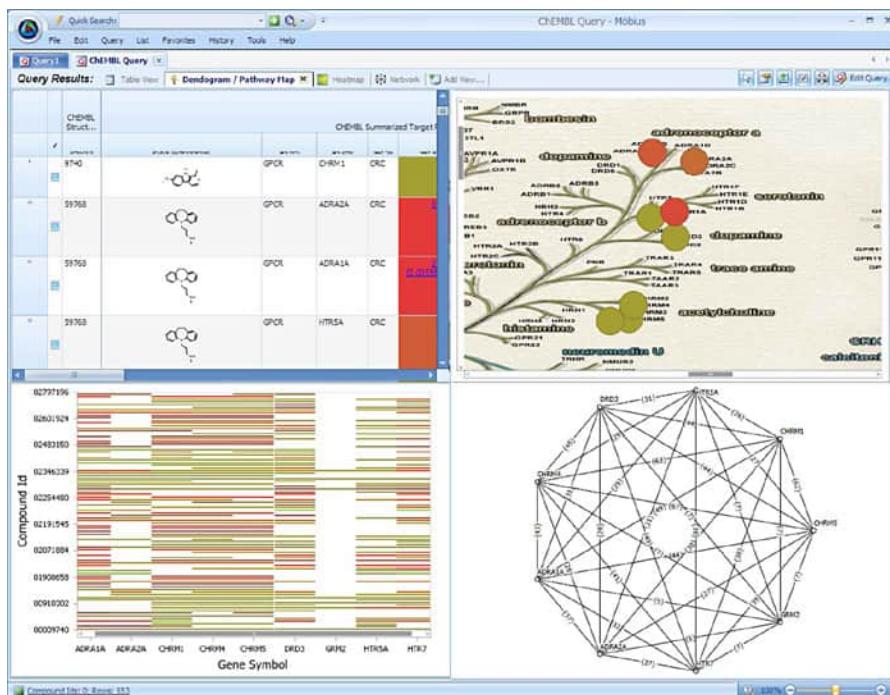


FIGURE 14.5 Query results display. Example results from Mobius queries. For color details, please see color plate section.

the user is notified via email of any changes to the query results as a consequence of database updates. Alerts can be configured to export the latest query results to external programs or shared file areas as well.

The bulk of Mobius use involves the building and execution of queries and the analysis of the retrieved results for decision support. In cases where further analysis of data is called for, Mobius can export retrieved data to several standard formats and applications (e.g., Excel, Spotfire). Mobius itself includes a menu of specialized analysis tools (e.g., R-Group decomposition, SAR landscape analysis [10], and an API for defining additional tools).

14.3.5 Software Components

In the current configuration of Mobius, the metaview manager, query builder, query engine, presentation framework, and specialized tools are custom in-house code (261k lines). Everything else is commercial/open source software including: Oracle (data storage and searching), Accelrys Draw, Direct and Cheshire (structure editing, rendering, searching, and transformation), DevExpress (user interface and data visualization controls), and NodeXL (network diagram visualization control). The Mobius application is written in Microsoft C#.Net WinForms/WPF/WCF and runs under Microsoft Windows as a Client/Services application.

14.4 LESSONS LEARNED: THE BOTTOM LINE

Over the years, we have learned many lessons and have described several that came out of specific experiences in earlier sections. In this section, we will attempt to summarize some of these lessons (not in any particular order) and drive home what we feel are the most important.

14.4.1 Quality Software

To build a great house, one needs a solid foundation. In the case of a CIDBS, that foundation consists of the overall architecture and the code base that implements it. This foundation must be solid and evolve as software technologies change. This solidity translates into stability which is important since a CIDBS needs to be up and running and returning accurate data around the clock for a world-wide discovery organization.

14.4.2 Data, Data, Data

Over the years we have seen an increasing need for diverse data sources and their integration. We do not see this trend ending anytime soon. However, it is sometimes a challenge to get access to data due to the protective instincts present in large organizations. It is our experience that it is more important to start building a CIDBS that is able to access the most important sources and add others as time and resources allow. It is often best to show value as soon as possible as resources in a pharmaceutical environment tend to get reassigned if the value proposition is not quickly realized.

We feel that the addition of new data sources often gives a synergistic benefit, in that the value of having all the data integrated together is greater than the sum of the values of the individual sources. This benefit ultimately acts as a significant force that promotes the inclusion of new data sources and reduces objections to eventual integration.

In the case of Mobius, we have data integrated from over 100 different database schemas. Having all these data available through a single interface provides a tremendous advantage and opportunity for discovery.

14.4.3 Commitment to the User

It seems clear to us that part of the reason for our success in developing well-regarded informatics systems is that we have been directly involved for a long time and that we have a customer-focused approach. This commitment to the user is reflected in the design of the user interface and how we interact with the user community. Our approach has always been to provide tools that allow users to meet their needs, not to enforce particular behaviors or provide tools that management thinks would be useful. Thus, it is essential to be able to interact with users as much as possible. This includes problem resolution, project support, training, and hand-holding.

14.4.4 Continuity

Continuity is important and has several dimensions including continuity in the people developing and supporting the CIDBS and continuity in the CIDBS design and implementation. Continuity in the developer team leads to consistency in the underlying architecture of the system, consistency in the techniques used to implement specific features, and efficiency in implementing new features and in fixing any defects. Continuity in an established and proven user interface provides additional benefits, the most important of which is consistency in the interaction with the system from the user perspective. A consistent style of interaction maintains user familiarity with a system and directly translates into usability. We believe that continuous incremental development rather than extended periods of software stagnation periodically punctuated by major interface disruptions provides many valuable benefits ranging from a reduction in necessary training to being able to keep abreast of the evolving science of Drug Discovery.

14.4.5 Importance of Developers That Understand the Science

Many systems have been developed over the years and have failed the test of time. Hundreds of millions of dollars have been spent on systems that have provided little or no benefit. It may surprise some that Cousin, ChemLink, or Mobius never had a formal set of specified requirements. How can that be? It is our feeling that having a set of disjointed and inevitably incomplete “requirements” obtained by interviewing the “experts” is not a prescription for *guaranteed* success. This is because the requirements often do not necessarily fit together to form a basis for a coherent workflow.

Having people that can develop great code, understand the science of drug discovery, and that care about the ultimate consumer of the CIDBS is key to understanding what makes sense to deliver and how to deliver it. It is well known that people will tend to ask for what it is they think can be done and what suits what they are doing today. They do not tend to think creatively about what innovations might help them do their jobs better tomorrow. Thus, by having a good knowledge of the science, where the unmet informatics needs are, and what is possible with current technology a development team is in a position to formulate and implement new innovative solutions to informatics problems. This provides the opportunity and means to form a true partnership with scientists and develop a CIDBS that is truly “ahead of the curve.”

14.4.6 Speed Kills

Speed kills or, rather, lack of speed kills. A system that cannot deliver the necessary data in what is perceived to be a reasonable amount of time is doomed. The interesting thing is, however, that speed is relative. One needs to have a real connection with the user community to understand this expectation.

14.4.7 Bottom-Up Beats Top-Down

The primary value of a system like Mobius is the delivery of the information and analytics to scientists that help them to drive a Discovery project. At most companies, the people making the compounds, running the assays, and doing many of the Mobius searches are not the senior leaders. However, when many systems get built, these are precisely the people that get asked for what it is they think is needed. While senior leadership can often give excellent high-level guidance, the people that are likely to give you the best feedback and guidance will be the people actually doing the work.

14.4.8 Use Off-the-Shelf Whenever Possible

Off-the-shelf software such as Word, Excel, PowerPoint, Spotfire, and so on are well known and well understood tools. By simply dovetailing these types of applications into a CIDBS, one can get many benefits for very little cost. At one point, Cousin did not use any commercial or open source chemistry software components. Eventually, the decision was made to switch to using commercial software for basic structure drawing, storage, and searching functions. This had a number of advantages from a development point of view since the team no longer had to maintain the internal code and automatically benefited from any improvements in new versions of the commercial software. In addition, resources were freed to pursue new problems. Of course, commercial and open source software has disadvantages as well but since a small team cannot and should not try to implement everything, the benefits of commercial/open source components clearly outweigh the disadvantages. The key challenge is in finding the proper balance between internal, commercial, and open source components.

14.4.9 Rollout Systems Using the Apostle Approach

In our experience, the best way to release totally new software or significant enhancements is to use the apostle approach. In this approach, one selects (anoints) individuals that are influential, catch on quickly to new tools, are inquisitive, energetic, and come from diverse areas of the community you intend to serve. Once identified, these individuals get individual and group training—whatever is necessary to ensure they “get it.” Then hold regular discussion/debriefing sessions to identify problems, issues, bugs, and anything that detracts from or can lead to a good user experience. It is critical here to make rapid changes and updates to fix any problems identified.

What tends to happen using this approach is that these people act as apostles and then spread the word, showing others the great new tools that they know how to use. This approach results in a groundswell of interest that culminates in people coming to you that are highly motivated to learn. However, for this approach to work, the CIDBS must be of sufficient quality to maintain the interest.

14.4.10 Training

Our approach has always been to provide a CIDBS that was easy to navigate by novice users while still providing sophisticated tools for experts. This is opposed to providing a complicated interface that is very powerful but hard to use. Consequently, we have never done a lot of training. Most of the users of all our systems were trained using a group lecture/demo approach. In this approach, the leader is at a podium and operates a real demo and goes over whatever it is to be covered. We have found that detailed hands-on training is not really worth the effort even though some people really seem to want this. This seems to be more of a comfort issue than a necessity for users. On the whole, we feel it is best to show people (repeatedly perhaps) how to do things. We encourage people to NOT WRITE step-by-step procedures down but to pay attention to the interaction. Once they understand the philosophy of the interaction they then can then successfully navigate the software and get answers to questions they care about.

14.4.11 Support/Maintenance

Our strong opinion is that for scientific software like Mobius, support needs to be managed by people well versed in the science and the application—generally one of the developers. Even though this seems to be a waste of a senior scientist’s time, there are great benefits to it. First, you build an incredible degree of trust and goodwill in the user community. Second, you understand precisely what problems people are facing—maybe it is a training issue or interface issue? You can then quickly resolve it. Thirdly, people may actually be trying to use the software in a way that does not currently make sense but that can turn into an idea for an enhancement.

14.4.12 Succession Planning

The authors feel incredibly lucky and fortunate to have worked together on a project that has presented a never-ending series of challenges and opportunities for over 30 years. It seems clear to us that if/when a company invests in the in-house development of software like Mobius that there is a risk of, what we like to call, the “hit by a bus” scenario. To guard against this, a company needs, in our opinion, to invest in the recruitment of qualified personnel to back up the key individuals to make for a smooth transition if and when it becomes necessary. It makes good business sense to have a backup plan if something unexpected occurs. However, in spite of our frequent recommendations, this is often not the case. No company that we have worked at has yet planned for this. Consequently, the full impact and ultimate consequences of an unplanned disruption in our experience is a story yet to be told.

14.5 BUILD VERSUS BUY VERSUS OPEN SOURCE

When the Cousin project was started in the mid-1970s, there was no existing chemical database software available that could do the job; so the only option was to build everything from the structure drawing program to the database search software to the

presentation code ourselves. In spite of the advent and evolution of commercial chemical database software since then, Upjohn, Pharmacia, Pfizer, and Lilly made the decision to maintain a core of in-house developed software based on the value that it added over commercial offerings. Other companies have made the same decision; for example, Johnson & Johnson's ABCD system [11] and Vertex's ASAP (unpublished).

We believe that this added value is largely the result of factors that we have discussed including the integration of a broad range of discovery data, the tailoring of the system to the needs of the company, the ability to quickly extend the system to meet new data and/or analysis needs, and high-quality science-oriented support of the user community. The best evidence we have as to the effectiveness of this approach are the usage statistics we continue to see (Figure 14.3) and the positive formal user surveys and anecdotal user feedback that we receive. With Mobius, we typically see 20,000 user sessions and 30,000 query executions by 900 of our total of 1200 active users per month. While these numbers are not impressive from an internet web application point of view, we consider this to be a high level of use given the size of our internal user community.

Although the design principals of Cousin/ChemLink/Mobius have remained similar for over 30 years, the underlying hardware, the system software, and the application code have changed many times. The general approach for the application code has been to replace custom in-house software with commercial/open source components whenever the commercial/open source software was judged to be as good as or better than the custom code. As mentioned earlier, the metadata model, query builder, query engine, presentation framework, and specialized tools are currently custom in-house code. Everything else is commercial/open source components.

While the hybrid in-house/commercial/open source development approach used in Cousin/ChemLink/Mobius has worked well for many years it is not ideal. In particular, it is dependent on maintaining the continuity of a small, expert, and dedicated in-house team of development/support staff and a companies' ongoing commitment to such a group. For most companies and in the currently in-flux state of the pharmaceutical industry, this is often difficult to achieve. One potential solution to this dilemma is for a company to integrate and/or donate proprietary code to an open source project. Many open source chemistry efforts [12], including database-oriented projects, are currently thriving. In some cases, these projects include industry-submitted source code. This approach can potentially relieve a company of the burden of maintaining proprietary software while still deriving benefits from its use.

14.6 CONCLUSIONS AND SUMMARY

To get the most out of the large investment in informatics systems, it is not sufficient to just have great software. There is a great deal more to it than that. So, thinking of an informatics System like Cousin, ChemLink, or Mobius in primarily technological terms is a mistake. An informatics system at its most fundamental level provides a way for scientists to get the data and information they need to make better decisions.

This is less about the technology and more about providing effective solutions. This is a customer-centric philosophy that the authors have embraced and we feel has lead to the success of Cousin, ChemLink, and Mobius.

Integrated information systems play an essential role in Pharmaceutical companies. The unfortunate lack of viable commercial solutions has driven many companies to develop their own. It is hoped by relating some of the authors' experiences and observations over the years that we have provided some insight as to how to be successful for those that wish to embark on that journey. It is extremely important to recognize that it is truly a journey and not a destination for the work is never finished. There will always be more data, new ideas, and new innovation.

REFERENCES

1. Corey EJ, Wipke WT, Cramer RD, et al. J Am Chem Soc 1972;94:421–431.
2. Howe WJ, Hagadone TR. J Chem Inf Comput Sci 1982;22:8–15.
3. Blasgen MW, Astrahan MM, Chamberlin DD, et al. IBM Syst J 1981;20(1):41–55.
4. Hagadone TR, Howe WJ. J Chem Inf Comput Sci 1982;22:182–186.
5. Hagadone TR, Lajiness MS. Tetrahedron Comput Methodol 1988;1:219–230.
6. Maggiora GM, Johnson MA, Lajiness MS, et al. Math Comput Model 1988; 11:626–629.
7. Hagadone TR. J Chem Inf Comput Sci 1992;32:515–521.
8. Hagadone TR, Schulz MW. J Chem Inf Comput Sci 1995;35:879–884.
9. Lajiness MS. Applications of molecular similarity/dissimilarity in drug research. In: van de Waterbeemd H, editor. *Structure-Property Correlation's in Drug Research*. Austin: R.G. Landes Company; 1996. p 179–205.
10. Guha R, Van Drie JH. J Chem Inf Model 2008;48(3):646–658.
11. Agrafiotis DK, Alex S, Dai H, et al. J Chem Inf Model 2007;47:1999–2014.
12. O'Boyle NM, Guha R, Willighagen EL, et al. Murray-Rust J Cheminform 2011, 3–37.

CHAPTER 15

MOLECULAR SIMILARITY ANALYSIS

JOSÉ L. MEDINA-FRANCO and GERALD M. MAGGIORA

Similarity like pornography is difficult to define, but you know it when you see it.¹

15.1 INTRODUCTION

While the notion of molecular similarity, which is the foundation of molecular similarity analysis (MSA), has been around for many years [1], it is only within the last few decades that it has become recognized as an important tool in many phases of chemical research, especially those associated with drug research. Much of this work has been driven by the explosive growth in databases that contain information on chemical structures and associated physicochemical and biological properties [2–4]. During the 1970s, computer systems were primitive, and consequently efforts to develop computerized chemical information systems were focused primarily on representing chemical structures in machine readable form so they could be retrieved by computer. Substructure searching represented an important advance that broadened search capabilities by allowing searches for sets of compounds that contain a given substructure or substructures. Nevertheless, it still represented too narrow a search criterion, since compounds either contained or did not contain the query substructure. A more powerful criterion was needed.

Molecular similarity seemed to be ideally suited for the task, except for one important thing. Unlike the certainty of substructure searches that identified compounds that either did or did not contain the query substructure, similarity is a subjective notion and thus is more difficult to exploit since there is no “right” answer. For example, in similarity-based searching, knowing a compound’s structure is not sufficient; the key concern is what molecular features do the two molecules have in common. This begs a number of questions. What molecular features should be considered? How is their relative importance assessed? How is this information

represented and combined to produce a computable measure of similarity? The choices made have a profound effect on the answers obtained. All of these are exacerbated by the fact that, because of its subjective character, there is no absolute standard of similarity that can be used to validate a given approach rigorously.

Beyond the issues of how it is determined, molecular similarity plays a substantive role in many areas of chemical informatics and drug research. This is illustrated by the work described in the chapters of this volume, which clearly shows that the molecular similarity landscape that was once only populated with methods designed primarily to handle small molecules has been transformed. Not only does it now include improved methods for treating small molecules, but also includes macromolecules (or parts thereof), especially those that are putative targets for drugs and small molecule bioprobes.

Section 15.2 of this chapter will provide a brief introduction to the history of similarity in the natural sciences from the time of the Greek philosophers to the present, with special emphasis on the development of molecular similarity over the last three decades. As similarity is a subjective concept it draws heavily, both explicitly and implicitly, on cognitive psychology. Thus, Section 15.3 provides a discussion of the psychological elements underlying molecular similarity that touches on some fundamental aspects of similarity in human cognition. It also deals with issues that are more pertinent to the cognitive aspects of molecular similarity such as “Is the similarity scale uniform?” For example, does a molecular similarity value of, say, 0.4 have as much “meaning” to a chemist as a value of, say, 0.9? Another related question is, do the size and complexity of objects influence our perceptions of similarity? This appears to be the case in many human activities, although our ability to abstract features from complex objects significantly complicates attempts to understand its many ramifications. Mathematically, the situation is more straightforward as it is well-known that some similarity measures exhibit size-dependent behavior [5]. Finally, the crucial issues of representation in human cognition and molecular similarity are touched upon. At this point, suffice it to say that the representations needed in cognitive psychology are much more varied and complex than those typically required in molecular similarity. Since cognitive similarity values need not be explicitly computed from the “mental representations,” their complexity and vagueness do not present practical limitations in cognitive psychology.

Section 15.4 provides a discussion of similarity measures, which depend on three factors: (1) the representation used to encode the desired molecular and chemical information, (2) whether and how much information is weighted, and (3) the similarity function (sometimes called the similarity coefficient) that maps the set of ordered pairs of representations onto the unit interval of the real line. Each of these factors is discussed in separate subsections. Section 15.5 presents a discussion of a number of questions that address significant issues associated with MSA: Does asymmetric similarity have a role to play? Do two-dimensional (2D) similarity methods perform better than three-dimensional (3D) methods? Do data fusion and consensus similarity methods exhibit improved results? Are different similarity measures statistically independent? How do we compare similarity methods? Can similarity measures be validated? Section 15.6 provides a discussion of activity landscapes

and activity cliffs and their emerging roles in chemical informatics and medicinal chemistry. Section 15.7 concludes with an overall summary and draws several conclusions regarding how and where MSA can impact medicinal chemistry and drug research.

15.2 A BRIEF HISTORY OF MOLECULAR SIMILARITY ANALYSIS

The concept of similarity is as old as human history. Not surprisingly, Greek philosophers were perhaps the first to explicitly apply the notion of similarity to matter, which in the fifth century B.C. was considered to be made up of the four basic elements—air, earth, fire, and water [6]. The views of Aristotle and Plato held sway until the early nineteenth century [7]. Plato postulated an analogy that was based on the concept of Platonic solids that served as models of the four basic elements.

The first major change toward chemistry as we know it today came in the early 1800s when Dalton proposed his atomic theory of molecules [8]. In it atoms were classified by their masses; all atoms with the same mass were taken to be equivalent. Not only did his theory describe the atomic nature of matter but it also described how the atoms could be combined to form compounds through “chemical reactions.” Dalton’s work undoubtedly set the stage for the eventual development of the entire field of chemistry.

Building on Dalton’s atomic theory, the concept of chemical structure continued to be developed and refined leading, in 1874, to the concept of stereoisomerism developed by van’t Hoff and Le Bel [9]. Perhaps, the most important achievement during the latter half of the nineteenth century was the development of the Periodic Table of the elements by Mendeleev [10] and Meyer [11], both of whom were students of Robert Bunsen. The form of Mendeleev’s classification, which is the progenitor of today’s modern Periodic Table, is based on atomic mass. However, Meyer’s work, which was based on valency, also contributed to its development.

All of this work led to a reasonably well-developed notion of chemical structure, even though it predated the development of quantum mechanics [12]. Importantly, it set the stage for the rapid growth of many fields of chemistry during the twentieth century. Throughout this period, the concept of similarity was applied to many aspects of chemistry. Table 15.1, which is based on Table 15.4 of Rouvray [1], provides a summary of these applications. In the present work, the focus is mainly chemical informatics and drug design, as these areas form the bulk of its usage in chemistry today.

Early applications of molecular similarity can be traced back to the groundbreaking work of Adamson and Bush [13, 14], which predated other efforts by more than 5 years [15–18]. Following these developments, Peter Willett and his colleagues made major contributions to the development and evaluation of a host of different similarity methods [19–21]. Also during this early era, a number of edited books were published summarizing various aspects of the field [19, 22, 23].

In the mid to late 1990s, the power of MSA became increasingly manifest, driven in large measure by the need to support functions of a growing number of chemical

TABLE 15.1 Chemical Applications of Similarity Concepts in the Twentieth Century

Methodology	Year	Authors
Concept of isotopy	1913	Soddy ^a
Concept of isosterism	1919	Langmuir ^b
Franck-Condon Principle	1925	Franck ^c
Principle of minimal structural change	1934	Hückel ^d
Rice-Teller principle	1938	Rice and Teller ^e
Molecular topology descriptors	1947	Weiner ^f
Shuler principle	1953	Shuler ^g
Molecular sequence comparisons in evolution studies	1953	Fox ^h
Sequence similarity comparisons in biomacromolecules	1955	Fox and Homeyer ⁱ
Hammond transition state postulate	1955	Hammond ^j
Structure-activity correlations	1964	Hansch and Fujita ^k
Woodward-Hoffman rules	1969	Woodward and Hoffman ^l
Similarity matrices for amino acid sequences	1970	Needleman and Wunsch ^m
Product stability principle	1971	Hine ⁿ
Principle of least nuclear motion	1977	Hine ^o
Principle of minimum chemical distance	1980	Jochum, Gasteiger, and Ugi ^p
Molecular charge similarity	1980	Carbo, Leyda, and Arnau ^q

Adapted from [1].

^aF. Soddy, *Nature*, 92, 399 (1913).

^bI. Langmuir, *J. Amer. Chem. Soc.*, 41, 1543 (1919).

^cJ. Franck, *Trans. Faraday Soc.*, 536 (1925).

^dW. Hückel, *Theoretische Grundlagen der Organische Chemie*, 2nd Edition, Akad. Verlag, Leipzig, p. 139, 1934.

^eF.O. Rice and E. Teller, *J. Chem. Phys.*, 6, 489 (1938).

^fH. Weiner, *J. Am. Chem. Soc.*, 69, 17 (1947).

^gK.E. Shuler, *J. Chem. Phys.*, 21, 624 (1953).

^hS.W. Fox, *Amer. Naturalist*, 87, 253 (1953).

ⁱS.W. Fox and P.G. Homeyer, *Amer. Naturalist*, 89, 163 (1955).

^jG.S. Hammond, *J. Amer. Chem. Soc.*, 77, 334 (1955).

^kC. Hansch and T. Fujita, *J. Amer. Chem. Soc.*, 86, 1616 (1964).

^lR.B. Woodward and R. Hoffmann, *Angew. Chemie Int. Ed. Engl.*, 8, 781 (1969).

^mS.B. Needleman and C.D. Wunsch, *J. Mol. Biol.*, 48, 443 (1970).

ⁿJ. Hine, *J. Amer. Chem. Soc.*, 93, 3701 (1971).

^oJ. Hine, *Adv. Phys. Org. Chem.*, 15, 1 (1977).

^pC. Jochum, J. Gasteiger, and I. Ugi, *Angew. Chemie Int. Ed. Engl.*, 19, 495 (1980).

^qR. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.*, 17, 1185 (1980).

information systems. Today, there are many examples of applications of MSA. Typical ones, some of which are touched upon in other chapters of this volume, include (in no particular order) compound acquisition [24], diversity analysis [25–27], comparative analysis of sets of compound [28], library design [29], subset selection [30, 31], ligand-based virtual screening (LBVS) [32–36], and analysis of screening data [37–39].

In recent years, the concept of *chemical space* has gained currency as it provides an effective framework for assessing structure and property (including biological activity) relationships among molecules, which are part of the fundamental underpinning of medicinal chemistry and drug discovery research (see Section 15.6). MSA plays a crucial role since it provides a practical computational means for assessing these relationships.

Augmenting chemical space, which is typically of high dimension, with an additional dimension associated with biological activity leads to the concept of activity landscapes. Many features found in these landscapes are analogous to those found in ordinary geographical landscapes and provide powerful metaphors for understanding many aspects of structure–activity relationships (SARs) in a more systematic and unified manner. For example, activity cliffs arise when two structurally similar molecules have significantly different biological activities. Also, regions of activity landscapes resembling gently undulating hills correspond to groups of structurally similar compounds with slowly varying biological activities. These and other features of chemical spaces and activity landscapes will be discussed further in Section 15.6.

While molecular similarity has no doubt played and will continue to play an important role in drug research and to a lesser extent other chemical-oriented fields, it is not without its issues. Primary among them is the sensitivity of similarity measures to the representation and similarity function used—as is well known, different measures tend to yield different similarity values. Such differences can radically transform chemical spaces so that nearest neighbors in one space are no longer nearest neighbors in another space [40, 41]. This behavior has significant consequences for many of the research activities such as LBVS that typically employs similarity methods. To counter this deficiency, a variety of methods have been developed based on combining, in some fashion, the results from multiple similarity procedures. These methods, typically called fusion or consensus methods, have led to some improvements and are discussed in Section 15.5.3.

15.3 COGNITIVE ASPECTS OF SIMILARITY

Similarity is a subjective concept, and as such it is associated with aspects of human cognition. In psychology, similarity is related to the “psychological proximity” of two “mental representations.”² Although there is some controversy among cognitive psychologists as to the exact role of similarity in the human thinking process [42, 43], few people would dispute the fact that it pervades our everyday lives. It is a powerful concept that allows us to classify, categorize, characterize, and reason about all manner of objective, subjective, and vaguely defined things (objects, concepts, ideas, etc.). But this power comes at a cost since it is difficult to describe in rigorous terms exactly how we employ similarity to accomplish these tasks, which begs the question as to how humans perceive similarity. Do we perceive similarity in the same way for all things? For example, do we perceive concrete objects that we can see, such as buildings, automobiles, and tigers, in the same way as we perceive ideas and concepts? Most likely not, since the latter two are abstractions. But are our perceptions of

concrete entities also subject to some level of abstraction? Moreover, the description of all things, regardless of their nature, is in many instances vague or uncertain. This undoubtedly applies in cases where judgments of similarity are made such as “Your home is similar to mine” or “The weather today is similar to yesterday’s.” In the case of molecular similarity, the complexity and types of the representations employed, while not trivial as will be seen in the sequel, are generally simpler than those required in cognitive psychology.

Cognitive psychologists having been trying to mathematically codify the notion of similarity and a number of approaches have been pursued. Because the notion of proximity underlies similarity, Shepard [44] developed a theory based upon distances between points in an abstract “mental space,” where points in the space represent concepts. Points located in close proximity are considered to be more conceptually similar than those located a greater distance apart. An advantage of this approach is that there exist methods such as multidimensional scaling for constructing spaces from distance values [44]. However, since distances are symmetric, that is, the distance from *A* to *B* is the same as the distance from *B* to *A*, conceptual similarities that do not exhibit symmetry cannot be treated using Shepard’s method.

Tversky [45] addressed this problem using a feature-based approach, which admits the concept of asymmetric similarities (cf. the discussion of asymmetric similarities related to Tversky’s work [46]). As will be seen in the sequel, asymmetric similarities also have a role to play when comparing molecules, albeit a relatively small one currently. In a manner that is quite like that used in molecular applications of similarity, (*vide infra*) lists of features are used to characterize the mental representations being compared. Some type of relationship is then used to assess the similarity based on the number of features that are common and different between two representations and their relative importance. The form of the functions investigated by Tversky [45] are closely related to some of those used in molecular similarity studies (see also Sections 15.4.2–15.4.4 and Table 15.3 and Table 15.4).

Feature-based approaches assume that the subsets of common and different features are independent of each other, which is not likely to be true in the case of cognitive aspects of similarity. This fact gave rise to the structural approach developed by Gentner and Markman [47]. Their approach is based on the dividing features of two mental representations into two sets containing alignable and nonalignable pairs of features, respectively. In the former, there are two subsets, one containing pairs of aligned features that are the same and other containing pairs of aligned features that are different. Psychologically, differences between aligned features strongly influence human perception of the difference between mental representations. For example, humans and bats are both mammals, but bats have wings and humans have arms; similarly, humans and tables both have legs, but humans have two and tables have three or more legs that are not made of flesh and bone. On the other hand, differences between nonaligned features play a less important role in assessing the difference between mental representations.

Aligned differences may also come into play in molecular similarity for representations based on weighted molecular fingerprints. In this case, the count of (typically substructural) features may differ between two molecules for particular features such

as methyl or amine groups. Other approaches to similarity have also been developed in cognitive psychology, but their form is not as close to methods used to treat molecular similarity, so they are omitted from further discussion here.

There are additional aspects of similarity associated with human cognition that are relevant to applications of molecular similarity. For example, is the scale by which humans assess similarity uniform? The following argument, which is related to the *assumed* complementary behavior between similarity and dissimilarity, shows that the answer to this question is no. In most mathematical applications, similarity values are taken to lie on the unit interval [0,1]. Dissimilarity is considered to be its “1’s complement,” that is, $\text{Dissimilarity} = 1 - \text{Similarity}$. Thus, the less similar two entities are to each other, the more dissimilar they are considered to be, and vice versa. However, from the point of view of cognition, humans can distinguish the degree of similarity of two similar objects far better than they can distinguish degree of dissimilarity between two highly dissimilar objects. As the objects become more dissimilar, a point is reached where it is tough for humans to assess how similar or dissimilar they are to each other. From a computational point of view, whether similarity is high or low is immaterial (*vide supra*). However, this may not accord with chemists’ perceptions of similarity—cf. [48], which could pose a problem in assessing the “chemical meaningfulness” of low similarity pairs of compounds.

Discerning the similarity of two objects also depends on the complexity of the objects being compared. Assessing the similarity of large complex objects is more difficult than assessing the similarity of small simple objects, a factor in assessing similarity that applies to molecules as well. In addition, assessing the similarity of a complex object with respect to a smaller and simpler object tends to downgrade the overall similarity value, a trend also observed in the case of molecules. This size-dependent behavior is found in some molecular similarity functions such as the popular Tanimoto similarity coefficient [5], behavior that can sometimes be corrected using alternative similarity functions such as those associated with asymmetric similarity measures [41].

The most important element in any determination of similarity is how the entities being compared are represented—what properties or features are, or should be, considered. Representation determines what we can know about the relationship of things to each other, be they objects, concepts, ideas, and so on, or molecules. The demands in cognitive psychology are especially high because of the multifaceted nature of mental representations required. However, unlike the case in molecular similarity, mental representations need not be explicitly computable as they are employed mostly as hypothetical constructs in theories of cognitive psychology.

Following on with this thought, since explicit similarity values are not necessarily needed in cognitive psychology, there is no corresponding need to assess the quality of the values so far as they represent true measures of similarity of the mental representations being compared. Thus, the subjectivity of the concept of similarity is not a significant problem. However, the situation is different in molecular similarity studies, since explicit values are computed for similarities, but since similarity of molecules or anything else is somewhat subjective, no absolute standard exists by which to measure the degree to which a given similarity measure faithfully captures

the similarity of two molecules. This is the essential problem associated with any attempt to compute the “true” similarity values. This issue will be discussed in Section 15.5.4 where the question of validating computed similarity values is discussed.

15.4 MOLECULAR SIMILARITY MEASURES

15.4.1 Mathematical Description of Molecular Similarity

Mathematically, molecular similarity measures are defined as the ordered triple $\mathcal{S} = (M, R, S)$, where

- M is a set of molecules, $M = \{M_1, M_2, \dots, M_n\}$.
- R is the set of molecular representations associated with M and is based on a common scheme (e.g., MACCS-key fingerprints [49]), $R = \{R_1, R_2, \dots, R_n\}$.
- $S(i, j)$ is a *computable similarity function (relation)* that maps ordered pairs (R_i, R_j) onto the unit interval of the real line, $S(i, j) : (R_i, R_j) \rightarrow [0, 1]$ for all $R_i, R_j \in R$

Because they map onto $[0, 1]$, similarity relations are fuzzy relations [50], which differ from classical relations that map pairs of elements onto the set of binary values $\{0, 1\}$. Similarity relations satisfy two mathematical properties, namely, they are reflexive, $S(i, i) = 1$ if $m_i = m_j$, and generally are symmetric, $S(i, j) = S(j, i)$ for $1 \leq i, j \leq n$, but they are generally intransitive. Asymmetric similarities, which will be discussed in Section 15.5.1, have been employed in MSA, but the number of applications is relatively small to date.

Distance measures are not considered in this chapter, although they are related to similarity measures [51]. From a cognitive perspective, however, similarity measures tend to be more effective because they are generally based on features in common between molecules, while distance measures are typically based on features that are different. This is a manifestation of the point discussed in Section 15.3 that humans can recognize entities that are highly similar much more easily than entities that are highly dissimilar. This follows because in the former case the presence of common features dominates our perception of similarity. However, in the latter case the number of common features is small, while the number of features that are not common is large, a situation that is more difficult for humans to grasp.

15.4.2 Representing Molecular and Chemical Information

A key element in computing molecular similarity is the choice of a representation that captures the chemical information appropriate to the desired application. Because similarity is a subjective concept this is not an easy task, the choices made can have a profound effect on the results obtained. For example, consider the set of bulk physicochemical properties of the two aromatic ring compounds, benzene and thiophene, shown in Figure 15.1. It is clear from the figure that both molecules share property



Benzene



Thiophene

MW = 78.11 Da
mp = 5.5°C
 bp = 80.15°C
 $\text{Log } P = 2.13$
 MR = 26.4
 $d = 0.879 \text{ g/cm}^3$

MW = 84.14 Da
mp = -38°C
 bp = 84°C
 $\text{Log } P = 1.81$
 MR = 25.0 cm^3
 $d = 1.057 \text{ g/cm}^3$

FIGURE 15.1 Comparison of a set of physicochemical properties associated with the aromatic compounds benzene and thiophene. The properties include molecular weight (MW), melting point (mp), boiling point (bp), octanol/water partition coefficient ($\text{Log } P$), molecular refractivity (MR), and density (d).

values that are quite close except for their freezing points; benzenes being considerably higher at 5.5°C than thiophenes at -38°C. Thus, if freezing points were the standard of comparison, one would conclude that benzene and thiophene are not very similar. If, on the other hand, molecular weights or boiling points were the standard, one would conclude that benzene and thiophene are quite similar. If all of the properties given in Figure 15.1 were considered together, one would conclude, as is in fact generally considered to be the case, that benzene and thiophene are reasonably similar with respect to their bulk properties.

In actual applications of MSA, many different types of representations are utilized to compute molecular similarities [41, 52–54]. Johnson [55] has provided a detailed discussion of the manifold types of mathematical spaces and their associated representations. The information contained in the representations is usually in the form of molecular or chemical features called descriptors that are derived from the structural and chemical properties of molecules. Descriptors are nominally classified as 1D (one-dimensional), 2D, or 3D. 1D descriptors are usually associated with whole molecule properties such as molecular weight, $\log P$, solubility, number of hydrogen bond donors, number of rotatable bonds, and so on. 2D descriptors are associated with the topological structure of molecules as typically depicted in chemists' drawings. Such depictions show the atoms, the bonds connecting them, and in some cases include stereochemical features, but they do not explicitly depict the 3D structures of molecules. 3D descriptors, as their name implies, are associated with the 3D structures of molecules. Todeschini and Consonni [56] have compiled an extensive reference containing many of the descriptors used in chemical informatics applications.

Since it is highly unlikely that a single representation and set of descriptors will capture all of the many different aspects of molecular and chemical information [57] needed for specific MSAs, the use of multiple representations have been proposed in different applications such as similarity searching [58, 59], diversity analysis [25–27, 60], and activity landscape modeling (see Sections 15.6.2 and 15.6.3).

Four different mathematical structures are the basis for essentially all representations used in MSA: sets, graphs, vectors, and functions.

15.4.2.1 Set-Based Representations Currently, the most prevalent entities in MSA are finite sets with ordered binary-valued elements, generally called (molecular or structural) fingerprints, bit (binary) vectors, or bit strings. Equation 15.4.1 provides an example of a molecular fingerprint

$$\mathbf{m}(i) = \{m_1(i), m_2(i), \dots, m_p(i)\} \quad (15.4.1)$$

where “*i*” designates the molecule being represented ($i \Leftrightarrow M_i$) and each $m_k(i)$ is typically related to a structural descriptor.³ If the *k*th structural descriptor of M_i is present *at least once*, $m_k(i)=1$, if it is not, $m_k(i)=0$. The binary function $m_k(i)$ is the *characteristic function* of classical set theory (see Section 15.4.5.1 and the Appendix of [41]). Thus, molecular fingerprints satisfy all of the algebraic relationships of finite sets. Instead of the brackets used in Equation 15.4.1, many formulations of molecular fingerprints employ parenthesis, which are more indicative of vectors (*p*-tuples), and parallel vector-based formulations have also been described [61].

Fingerprints come in two basic forms, namely, with a fixed number or a variable number of descriptor elements. In fixed-length fingerprints, descriptors are preassigned, and, hence, their number is fixed. Such descriptors are usually in the form of substructural fragments [49]. For variable-length fingerprints, on the other hand, the descriptors depend on the structure of the molecule and, thus, their number is molecule dependent. Examples include atom pairs [17], topological torsions [62], extended connectivity [63], and the atom environment descriptors (MOLPRINT 2D) [64]. Sastry et al. [65] have recently carried out an extensive analysis of a variety of 2D fingerprints.

Since most of the descriptors are derived from 2D structural data, such fingerprints are often times called 2D fingerprints, although fingerprints representing 3D structural data have also been constructed. Examples include the 3- and 4-point pharmacophore-based⁴ fingerprints developed by Mason and coworkers [66, 67].

The order of the elements is arbitrary but fixed for all molecules represented by a given type of fingerprint. As the number of elements can be quite large and variable, some fingerprints are hashed and sometimes “folded” to reduce their size. These have the benefit of reducing storage requirements and increasing computational speed, but the identity of the elements is lost so that molecular “interpretations” are no longer possible (see also the discussion given in footnote 3).

Not accounting for multiple occurrences of features causes a loss of chemical information that may in some cases lead to degeneracies when two molecules with different structures have a similarity of unity, making them “appear” identical with respect to the particular type of fingerprint used. Fingerprints whose elements are integer-valued fall into the class of sets called multisets [68]. Similarity functions analogous to those used for binary sets can be suitably generalized to handle multisets as well [69]. Although proper stereochemistry is important to the biological activity of many compounds, it is not incorporated in a “natural” way into many of the fingerprints in use today with the possible exception of 3D fingerprints such as the 3- and 4-point pharmacophores developed by Mason [66, 67]. Lastly, set-based

representations employing discretized, ordered variables have also been employed [70], but rather than employing multisets, expanded binary fingerprints were used (see [41] for additional discussion).

In an alternate view, binary fingerprints are treated as binary vectors whose components take on values of “1” or “0.” Fortunately, in the binary case at least, many of the properties of fingerprints (i.e., classical sets) are reproduced for binary vectors. For example, the intersection of two binary fingerprints produces a result that is equal in value to that obtained by taking the inner product of the two binary vectors. Similarly, the cardinality or size of a binary fingerprint is equal to the value of the inner product of the binary vector with itself. Both of these *pseudo equivalences* obtain because the product terms corresponding to the presence of some feature in both vectors satisfy $1 \times 1 = 1$. Thus, vector-based formulas for the different types of molecular similarity given in Table 15.4 will produce identical results to those based on the binary sets given in Table 15.3 (*vide infra*).

In the case of weighted features, however, the individual vector components may have integer values that are greater than “1” or may be fractions (i.e., rational numbers). In such cases, the values obtained from vector-based similarity functions do not strictly reproduce their corresponding set-based versions, although the former are used routinely. Multisets are the proper mathematical framework for treating weighted fingerprints whose components have integer values, while fuzzy sets are the proper framework for treating weighted fingerprints with components with fractional values that lie on the unit interval of the real line. Further discussion of these points is given in Section 15.4.3 on weighted fingerprints, in Section 15.4.4.1 on set-based similarity functions, and in Section 15.4.4.2.

15.4.2.2 Chemical Graph-Based Representations In terms of general chemical applications, chemical graphs [71] are a natural representation to use for assessing molecular similarity. They are typically defined mathematically as an ordered triple of sets, $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{L})$, where \mathcal{G} is the graph, \mathbf{V} is the set of its t vertices (“atoms”),

$$\mathbf{V} = \{V_1, V_2, \dots, V_t\} \quad (15.4.2)$$

\mathbf{E} is the corresponding set of u edges (chemical bonds) connecting the vertices,

$$\mathbf{E} = \{E_1, E_2, \dots, E_u\} \quad (15.4.3)$$

and \mathbf{L} is the set of w symbols that label each atom and/or bond

$$\mathbf{L} = \{L_1, L_2, \dots, L_w\}. \quad (15.4.4)$$

Typical atom labels include “C” (carbon), “N” (nitrogen), and “O” (oxygen); typical bond labels include “s” (single), “d” (double), “t” (triple), and “ar” (aromatic). Graph-based representations used in most cheminformatics applications generally employ some variant of hydrogen-suppressed graphs, which are chemical graphs

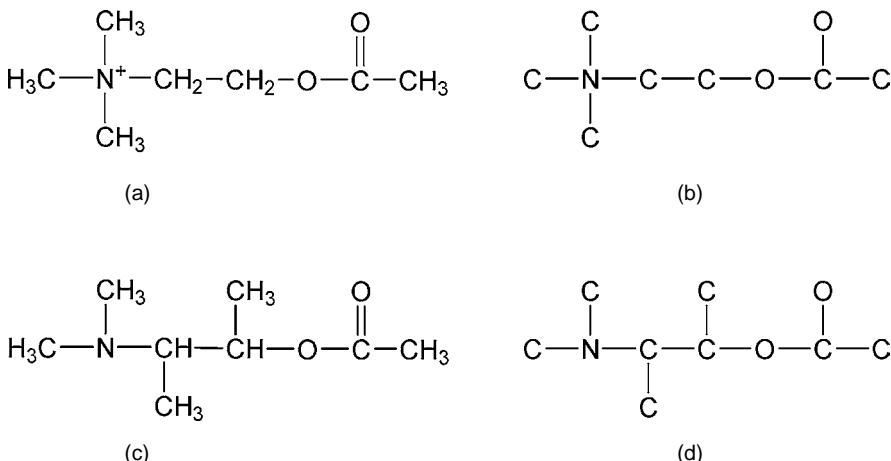


FIGURE 15.2 (a) Chemist’s 2D structure of the biomolecule acetylcholine. (b) Chemical graph of acetylcholine. (c) Chemist’s 2D structure of an analog of acetylcholine, 3-(dimethylamino)butan-2-yl acetate. (d) Chemical graph of acetylcholine analog given in (c). Note that the bond types in the chemical graphs are not explicitly labeled, but the atoms are labeled for clarity.

with their hydrogen atoms removed.⁵ To simplify notation, the chemical graph of M_i , $\mathcal{G}(m_i)$, is abbreviated as $\mathcal{G}(i)$, which is consistent with the notation used for sets and vectors.

Chemical graphs contain information on the topology of molecules (i.e., the connectivity of their atoms), and as such represent essentially 2D structural information. 3D structural information has also been represented in chemical graphs [72], but their use has been much more limited.

Figure 15.2 depicts the 2D structure of the biomolecule acetylcholine, one of its analogs, and their corresponding hydrogen-suppressed chemical graphs. Note that the bond labels are not explicitly included in the diagrams but the atoms are included for clarity. As will be seen in Section 15.4.5.2, the chemical graph-based similarities described in this chapter employ a bond metric that simply involves a count of the number of bonds, excluding bonds to hydrogen atoms, irrespective of their bond types—a single, double, or aromatic bond are counted as one bond each. While there are other possible metrics [72, 73], the bond metric is quite reasonable for many cheminformatics applications.

15.4.2.3 Vector- and Function-Based Representations The discussion of vector-based representations in this section will focus on vectors with continuous-valued components (cf. the discussion in Section 15.4.2.1 on binary vectors). While some of the specifics of continuous functions differ from those of vectors, the general forms of their similarity functions are the same (*vide infra*). This is because in most cases both the vectors and functions belong to vector spaces, and thus, they satisfy the axioms of these spaces. For example, adding two

vectors (functions) or multiplying a vector (function) by a scalar should yield another vector (function) that lies in the same vector space; many functions also satisfy these axioms.

Many of the vectors employed in chemical informatic applications must be viewed as geometric rather than as algebraic objects since they do not satisfy the vector space axioms. Since the component values of molecular vectors are, except in rare cases, positive their associated vectors will lie in the positive hyper-quadrant. Subtracting two such vectors, an operation that is allowed in vector spaces, may yield a vector that does not lie within the positive hyper-quadrant, and thus, does not correspond to any molecule that can be represented by that form of representation. For example, vectors whose components are BCUT descriptors [74, 75] do not satisfy the vector-space axioms, since adding two such vectors may produce a vector that lies outside of BCUT chemical space. This does not, however, pose a practical problem since such vectors can be thought of as entities that describe points in a geometric space not vectors in a vector space.

Using similar notation to that of Equation 15.4.1 for sets, a vector description of m_i is given by the p -tuple

$$\mathbf{v}(i) = (v_1(i), v_2(i), \dots, v_p(i)) \quad (15.4.5)$$

Vectors described by p -tuples are the same as the column vectors extensively employed in matrix manipulations. Since there is no need for such manipulations in this work, the p -tuple representation is used throughout. It should also be noted, albeit not explicitly stated, that the space in which these vectors reside generally is assumed to be Euclidean, and hence, the coordinate system is orthogonal. Although, as pointed out by Gower [76], even vectors satisfying metric distance properties are not necessarily embeddable in Euclidean spaces.

The elements of the vectors are continuous variables that are usually associated with physicochemical properties, chemical graph-based topological invariants such as branching or shape indices, quantum mechanical properties such as molecular orbital or ionization energies, or macroscopic chemical properties such as solubility, $\log P$, or heat of vaporization. In most but not all cases, the values of the components are positive real numbers. This is not a severe restriction since the components generally are associated with some type of molecular property whose values are usually positive, except for entities that are associated with energy quantities such as molecular orbital energies, ionization potentials, and Gibb's free energies, which are usually negative. When the vector components all are positive in value, the corresponding vectors lie within the positive hyper-quadrant of the Euclidean space in which the vectors are embedded. Real vectors with negative component values can lead to similarity values that are also negative, and thus fall outside of the usual [0,1] range of most similarities (see Section 15.4.4.3 for further discussion).

In contrast to finite dimensional vectors, continuous functions play a somewhat different role in MSA as they usually provide an approximate representation of the fields associated with molecules. These include electron density, molecular electrostatic potential, and lipophilicity fields. Although the latter is not a true field

in the quantum mechanical sense, it nevertheless represents, to some degree at least, regions of molecules that exhibit lipophilic behavior. Good and Richards [77] have described many aspects of 3D similarity measures associated with continuous functions.

Typically, linear combinations of radially symmetric 3D Gaussian functions are used to approximate the fields [77, 78], although ellipsoidal Gaussians have been employed with success in some recent work on molecular shape-based similarity searches [79]. Alternatively, properties can be assigned to individual Gaussians [80, 81]. Both spherical and ellipsoidal Gaussians have the desirable property that the integral of the product of Gaussians can be written in closed form, thus significantly speeding up calculations. Aside from their computational speed they are much easier to deal with, having fewer pathologies than discrete, field-based methods, which rely on the calculation of field values at a large set of fixed grid points distributed in the space surrounding the molecules. Grid methods are computationally much slower and less robust than function-based methods.

15.4.3 Weighted Representations

The various components that are part of any type of molecular representation can be weighted in some fashion to emphasize or deemphasize their contributions to the overall similarity value. Willett and Winterman [82] carried out the earliest studies based on molecular fingerprints. Although they investigated several different weighting schemes, the size of their dataset was too small to draw any meaningful conclusions. Recently, a much more ambitious study was carried out by Arif et al. [83] based on the frequency of occurrence of different features in molecular fingerprints. Numerous computational experiments were carried out based on two large chemical databases, namely, MDL Drug Data Report (MDDR) and the World of Molecular Bioactivity (WOMBAT), using a variety of fingerprints to encode the structural information of the molecules in these databases. Their studies showed that fingerprints based on occurrence rather than incidence of structural features consistently outperformed the latter. Moreover, they showed that by standardizing the raw occurrence data by taking the square root of their frequencies led to further improvements.

The weighted fingerprints used in both of these studies are treated as vectors with integer-valued components. Thus, the appropriate similarity functions are those given in Table 15.4 for vector-based representations, not those given in Table 15.3 for set-based representations. Although the vector-based approach seems to be a natural one, multisets are the proper mathematical objects for treating multioccurrence-based fingerprints, as noted at the end of Section 15.4.2.1. For example, consider the summation formula for multiset-based Tanimoto similarity given in Table 15.3. The summed terms in the numerator, which correspond to set intersection for classical, fuzzy, or multisets, employ the “Min function” that takes the fingerprint element with the smallest value for each of the pairs of elements in the entire fingerprint and then sums the results. This will typically be smaller than the value obtained from the corresponding vector-based formula given in Table 15.4, which multiplies these

values together before summing them. This may explain why using square roots of occurrences in the fingerprints yields improved results.

Wang and Bajorath [84] have developed a position-dependent scheme that weights individual positions in a vector-based similarity function but not the positions of the individual fingerprints. The weights are derived for specific compound classes associated with various biological activities. In addition to their use in vector-based methods, weighting factors have also been employed in methods based on continuous functions [80, 81].

15.4.4 Molecular Similarity Functions or Coefficients

Once the type of representation has been selected, the similarity function or coefficient must be evaluated with respect to the chosen representation. Most of the similarity functions in use today are computed as ratios. The numerator is generally associated with some measure of the common features of the molecules being compared, while the denominator is associated with some measure of all of the features of the molecules. The mathematical forms of the similarity functions are quite similar, although evaluation of the expressions may require different types of mathematical operations. More specifically, the expressions set- and graph-based representations can be grouped together as can vector- and function-based representations.

15.4.4.1 Set-Based Similarity Functions Similarity functions for all set-based representations (i.e., classical sets, fuzzy sets, and multisets) utilize mathematical functions of the same form. The expression for molecular fingerprints given in Equation 15.4.1 applies to all three classes of sets.

The possible set of values v that an element of a fingerprint m_k can attain is given in Table 15.2. From the table, it is clear that m_k is a function that maps the elements of the set $\mathbf{m}(i)$, the fingerprint of the i th molecule, into the appropriate value set. The values of the characteristic function of classical sets specify whether an element is in the set (“1”) or not (“0”) (see also Section 15.4.2.1). In contrast, the values of the membership function of fuzzy sets, which lie on the unit interval of the real line, specify the degree of membership of an element in the set. Thus, in this case, the value set is continuous and lies on the real line. Lastly, the value set of multisets is the set of nonnegative integers and, thus, the values correspond to the number of occurrences in the set of each of its basic elements. Since the value set associated with multisets has no upper bound, the operation of complementation, $\bar{m}_k = 1 - m_k$, is not available. Thus, the set algebra

TABLE 15.2 Set Functions and Corresponding Value Sets of the Classical Sets, Fuzzy Sets, and Multisets Associated with Molecular Fingerprints

Set Type	Value Set, v	Set Function, $m_k(i)$
Classical set (CS)	{0,1}	Characteristic
Fuzzy set (FS)	[0,1]	Membership
Multiset (MS)	\mathbb{N}	Count

of multisets does not contain any expressions involving complementation such as De Morgan's Laws. This limitation, however, does not restrict the use of multisets in the similarity functions typically used in cheminformatics.

Categorical variables with ordered value sets can also be handled using multisets, although such applications have, to the best of our knowledge, not been carried out in cheminformatics. Typically, sets with integer or categorical value sets are handled by transforming the multiset into an "equivalent" binary-valued set [70] by hashing the fingerprint components [85, 86], or by treating the multisets as vectors with integer coefficients [61].

The *cardinality* ("size" or measure) of \mathbf{m} is given by the same expression for all three types of sets

$$|\mathbf{m}(i)| = \sum_{k=1}^n m_k(i) \quad (15.4.7)$$

For classical sets and multisets, cardinality is just the total number of elements in the set, while for fuzzy sets no such simple interpretation is possible. Nonetheless, it is clear that in all three cases, cardinality provides a measure of the size of the sets.

In order to evaluate set-based similarity functions it is necessary to consider the cardinality of the respective *intersection* and *union* sets of $\mathbf{m}(i)$ and $\mathbf{m}(j)$

$$|\mathbf{m}(i) \cap \mathbf{m}(j)| = \sum_{k=1}^n \min[m_k(i), m_k(j)] \Leftrightarrow |\mathbf{m}(i \cap j)| \quad (15.4.8)$$

$$|\mathbf{m}(i) \cup \mathbf{m}(j)| = \sum_{k=1}^n \max[m_k(i), m_k(j)] \Leftrightarrow |\mathbf{m}(i \cup j)| \quad (15.4.9)$$

where the abbreviated notations $|\mathbf{m}(i \cap j)|$ and $|\mathbf{m}(i \cup j)|$ are used to simplify the mathematical expressions. Graph-based similarities can also be determined by equations of similar form (*vide infra*). Interestingly, an identical set of equations has been developed by Baldi and coworkers [87] without directly appealing to the general set-based properties of cardinality, intersection, and union (cf. the work of Chen and Reynolds [88]).

Willett, Barnard, and Downs provide an extensive listing of many types of similarity functions [61]. Five similarity functions will be considered here to illustrate how the current formulation provides a unified description: Tanimoto ("Tan"), Cosine ("Cos"), Dice, and the related pair Max and Min. Table 15.3 summarizes the various forms. *Importantly, the formulas apply to all three types of sets—classical, fuzzy, and multisets—providing a unity to the set-based approach.*

Gower [76] has described a number of functional relationships among the various set-based similarity functions. For example, simple algebra shows that the values of S_{Tan} are less than and are monotonically increasing with respect to those of S_{Dice} .⁶ As will be discussed further in Section 15.5.1, S_{Min} and S_{Max} are closely related to asymmetric similarity functions originally described by Tversky [45]. Although it is not described here, a number of set-based similarity functions can be written as various means of S_{Min} and S_{Max} (cf. [46]). As seen in Equations 15.4.10 and 15.4.11, S_{Min} and S_{Max} may provide a potentially useful interpretation of molecular similarity (*vide infra*):

TABLE 15.3 Five Similarity Functions (Coefficients) Commonly Used with Molecular Fingerprints

Similarity Function	Set-Based Formulas	Summation-Based Formulas
Tanimoto (Jacard)	$S_{\text{Tan}}(i, j) = \frac{ \mathbf{m}(i \cap j) }{ \mathbf{m}i \cup j }$	$S_{\text{Tan}}(i, j) = \frac{\sum_{k=1}^n \min[m_k(i), m_k(j)]}{\sum_{k=1}^n \max[m_k(i), m_k(j)]}$
Dice	$S_{\text{Dice}}(i, j) = \frac{ \mathbf{m}(i \cap j) }{\frac{1}{2}(\mathbf{m}(i) + \mathbf{m}(j))}$	$S_{\text{Dice}}(i, j) = \frac{\sum_{k=1}^n \min[m_k(i), m_k(j)]}{\frac{1}{2} \left[\sum_{k=1}^n m_k(i) + \sum_{k=1}^n m_k(j) \right]}$
Cosine	$S_{\text{Cos}}(i, j) = \frac{ \mathbf{m}(i \cap j) }{\sqrt{ \mathbf{m}(i) \cdot \mathbf{m}(j) }}$	$S_{\text{Cos}}(i, j) = \frac{\sum_{k=1}^n \min[m_k(i), m_k(j)]}{\sqrt{\sum_{k=1}^n m_k(i) \cdot \sum_{k=1}^n m_k(j)}}$
Min	$S_{\text{Max}}(i, j) = \frac{ \mathbf{m}(i \cap j) }{\min[\mathbf{m}(i) , \mathbf{m}(j)]}$	$S_{\text{Max}}(i, j) = \frac{\sum_{k=1}^n \min[m_k(i), m_k(j)]}{\min \left[\sum_{k=1}^n m_k(i), \sum_{k=1}^n m_k(j) \right]}$
Max	$S_{\text{Min}}(i, j) = \frac{ \mathbf{m}(i \cap j) }{\max[\mathbf{m}(i) , \mathbf{m}(j)]}$	$S_{\text{Min}}(i, j) = \frac{\sum_{k=1}^n \min[m_k(i), m_k(j)]}{\max \left[\sum_{k=1}^n m_k(i), \sum_{k=1}^n m_k(j) \right]}$

All of the set-based similarity functions are symmetric

$$S_{\text{Min}}(i, j) = \frac{|\mathbf{m}(i \cap j)|}{\max[|\mathbf{m}(i)|, |\mathbf{m}(j)|]} \quad (15.4.10)$$

$$S_{\text{Max}}(i, j) = \frac{|\mathbf{m}(i \cap j)|}{\min[|\mathbf{m}(i)|, |\mathbf{m}(j)|]} \quad (15.4.11)$$

Assume without loss of generality that the i th molecule is “smaller” than the j th molecule, that is,

$$|\mathbf{m}(i)| = \rho \cdot |\mathbf{m}(j)| \quad (15.4.12)$$

where the “size ratio” $\rho \in [0,1]$. Under this assumption, the two equations then become

$$S_{\text{Min}}(i, j) = \frac{|\mathbf{m}(i \cap j)|}{|\mathbf{m}(j)|} = S_*(j, i) \quad (15.4.13)$$

$$S_{\text{Max}}(i, j) = \frac{|\mathbf{m}(i \cap j)|}{|\mathbf{m}(i)|} = S_*(i, j) \quad (15.4.14)$$

which are the asymmetric similarity functions defined by Tversky [45], that is, $S_*(i, j) \neq S_*(j, i)$, and discussed further in Section 15.5.1. In an earlier work, Holliday and coworkers [5] described identical similarity functions to those in Equations 15.4.13 and 15.4.14 that they defined as modified versions of the Forbes and Russell-Rao similarity coefficients, respectively.

$S_*(j, i)$ can be interpreted as the fraction of the j th molecule that is similar to the i th molecule, and $S_*(i, j)$ as the fraction of the i th molecule that is similar to the j th molecule. This is analogous to the work of Kosko [89] on the degree of subsethood of fuzzy sets as well as the work of Miyamoto [51] in the area of information retrieval.

Because the numerators of $S_*(i, j)$ and $S_*(j, i)$ are identical, the two asymmetric similarities are related as in Equation 15.5.3

$$\begin{aligned} S_*(j, i) &= \frac{|\mathbf{m}(i)|}{|\mathbf{m}(j)|} \cdot S_*(i, j), \\ &= \rho \cdot S_*(i, j) \end{aligned} \quad (15.4.15)$$

which clearly shows that as the number of features encoded in the fingerprints of the two molecules, that is, their respective “sizes,” approach one another the two asymmetric similarity values become approximately equal. In contrast, as the size ratio grows smaller, that is, as the disparity between the sizes of the two molecules increases so does the disparity in the values of the two asymmetric similarity functions.

All of the set-based similarity functions in Table 15.3 are symmetric, have identical numerators, and are bounded by 0 and 1. Except for the Tanimoto similarity function, the denominators of S_{Min} , S_{Dice} , S_{Cos} , and S_{Max} are aggregation functions [50, 90] that correspond to the max, arithmetic mean, geometric mean, and min functions. Since the values of these functions are ordered, the magnitudes of the corresponding similarity functions are also be ordered (cf. [91]). It can also be shown that the Tanimoto similarity function is a lower bound to the four other functions so that

$$0 < S_{\text{Tan}} \leq S_{\text{Min}} \leq S_{\text{Dice}} \leq S_{\text{Cos}} \leq S_{\text{Max}} \leq 1. \quad (15.4.16)$$

As is well known, the Tanimoto similarity coefficient, which is the most widely used similarity measure, exhibits size-dependent behavior [5, 92–95] that can significantly influence the results of similarity searches. A significant part of the problem can be traced to the terms in the denominator of the Tanimoto function that counts the number of elements that are common to both molecular fingerprints. Thus, when molecules of widely varying sizes are treated, the number of elements in fingerprint

of the largest molecule dominates the count (i.e., the number of elements in the smaller molecule can be neglected). This tends to artificially lower the computed similarity values. Similar arguments as those for S_{Tan} can be made for S_{Dice} and S_{Cos} as is seen by comparing the respective denominators of these three functions in Table 15.3. Such size-dependent effects can also be traced back to factors associated with the complexity of molecules, which is not entirely surprising given that molecular size, and thus fingerprint bit density, tend to be linearly correlated with molecular complexity (see Section 15.5.1 for additional discussion).

A number of workers have dealt with various aspects of the size dependency issue (*vide supra*). Bajorath and coworkers have addressed it from the stand point of the Tversky asymmetric similarity function [84, 96, 97]. Mestres and Maggiora [46] have discussed size dependency in the context of asymmetric similarity functions associated with vector- and function-based representations.

15.4.4.2 Chemical Graph-Based Similarity Functions There are three key aspects to the computation of graph-based similarity measures: (1) subgraph relationships, (2) maximum common substructures (MCSs), and (3) graph cardinalities. For clarity in the following discussion, explicit account of elements of the label set L will be omitted.

The chemical graph $\mathcal{G}(i)$ is said to be a subgraph of $\mathcal{G}(j)$, written $\mathcal{G}(i) \subseteq \mathcal{G}(j)$, if their corresponding vertex (atom) and edge (bond) sets satisfy $\mathbf{V}(i) \subseteq \mathbf{V}(j)$ and $\mathbf{E}(i) \subseteq \mathbf{E}(j)$, respectively. Common substructures (CS) and MCS play a crucial role in graph-based MSA. Although there are several possible forms that MCSs can take [72, 73], a simple description that focuses solely on their edges will be described here. This yields maximum common edge structures that are close to what chemists perceive as “chemically meaningful” substructures [98]—this is indicated by the subscript “E,” so that MCS becomes MCS_E . Thus, the cardinality of a chemical graph in the edge representation is given by $|\mathcal{G}(i)|_E = |\mathbf{E}(i)|_E$, where again the subscript “E” is included to emphasize that the set cardinality is based solely on the edge count of the graph.

A common edge substructure of two graphs is given by

$$\text{CS}_E(\mathcal{G}(i), \mathcal{G}(j))_{k,\ell} = \mathbf{E}_k(i) \cap \mathbf{E}_\ell(j) = \mathbf{E}_k(i) = \mathbf{E}_\ell(j), \quad (15.4.17)$$

where $\mathbf{E}_k(i)$ and $\mathbf{E}_\ell(j)$ are subsets of their respective edge sets. Since there are many $\text{CS}_E(\mathcal{G}(i), \mathcal{G}(j))_{k,\ell}$, finding $\text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j))$ is equivalent to finding $\max |\text{CS}_E(\mathcal{G}(i), \mathcal{G}(j))_{k,\ell}|$. The intersection of two chemical graphs is *equivalent* to their maximum common edge substructure,

$$\mathcal{G}(i) \cap \mathcal{G}(j) \equiv \text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j)) \quad (15.4.18)$$

Thus, finding the intersection of two graphs is a graph-matching, discrete optimization problem that is similar in spirit to the structural alignments of molecules of some 3D similarity methods (see Section 15.4.4.3). Determining the cardinality of the intersection graph from Equation 15.4.7 is straightforward

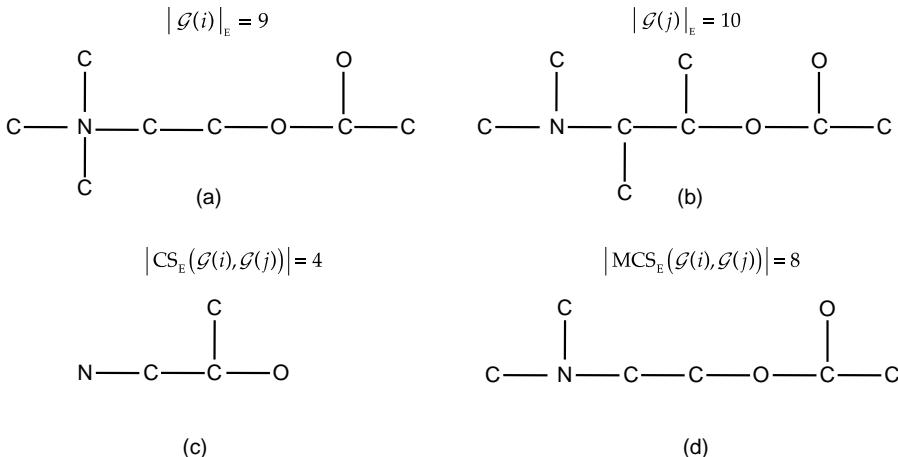


FIGURE 15.3 Example of a chemical graph-based calculation of molecular similarity. (a) Chemical graph of acetylcholine. (b) Chemical graph of an acetylcholine analog. (c) A common structure (CS) of (a) and (b). (d) The maximum common substructure (MCS) of (a) and (b).

$$|\mathcal{G}(i) \cap \mathcal{G}(j)|_E \equiv |\text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j))|. \quad (15.4.19)$$

The corresponding union is given by

$$\left| \mathcal{G}(i) \cup \mathcal{G}(j) \right|_E = \left| \mathcal{G}(i) \right|_E + \left| \mathcal{G}(j) \right|_E - \left| \text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j)) \right|. \quad (15.4.20)$$

so that Tanimoto similarity (see Table 15.3) is given by

$$S_{\text{Tan}}(\mathcal{G}(i), \mathcal{G}(j)) = \frac{|\mathcal{G}(i) \cap \mathcal{G}(j)|_E}{|\mathcal{G}(i) \cup \mathcal{G}(j)|_E} = \frac{|\text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j))|}{|\mathcal{G}(i)|_E + |\mathcal{G}(j)|_E - |\text{MCS}_E(\mathcal{G}(i), \mathcal{G}(j))|} \quad (15.4.21)$$

Other chemical graph-based similarities (e.g., Dice, Cosine, and Asymmetric) can be constructed using the analogous set-based formulas given in Table 15.3. As noted earlier, the main impediment to wider adoption of graph-based methods is their excessive computational demands.

Figure 15.3 provides an example of acetylcholine (Figure 15.2a) and one of its derivatives (Figure 15.2c). From the cardinalities of the corresponding chemical graphs given in Figure 15.3a and 15.3b (see also Figure 15.2b and 15.2d) and the cardinality of the MCS given in Figure 15.3d, it is possible to compute a chemical graph-based analog of the Tanimoto similarity function

$$\begin{aligned}
 S_{\text{Tan}}(\mathcal{G}_i, \mathcal{G}_j) &= \frac{|\text{MCS}_{\text{E}}(\mathcal{G}_i, \mathcal{G}_j)|}{|\mathcal{G}_i|_{\text{E}} + |\mathcal{G}_j|_{\text{E}} - |\text{MCS}_{\text{E}}(\mathcal{G}_i, \mathcal{G}_j)|} \\
 &= \frac{8}{9+10-8} \\
 &= \mathbf{0.73}
 \end{aligned} \tag{15.4.22}$$

15.4.4.3 Vector- and Function-Based Similarity Functions Several important geometric properties are required to implement vector- and function-based similarity computations. The properties include the norm (“length”) of a vector and the distance and cosine of the angle between two vectors, which are given in Equations 15.4.23, 15.4.24, and 15.4.25, respectively:

$$\|\mathbf{v}(i)\| = \sqrt{\langle \mathbf{v}(i), \mathbf{v}(i) \rangle} = \sqrt{\sum_{k=1}^p v_k(i)^2} \tag{15.4.23}$$

$$\begin{aligned}
 \text{dist}(\mathbf{v}(i), \mathbf{v}(j)) &= \|\mathbf{v}(i) - \mathbf{v}(j)\| = \sqrt{\langle \mathbf{v}(i) - \mathbf{v}(j), \mathbf{v}(i) - \mathbf{v}(j) \rangle} \\
 &= \sqrt{\sum_{k=1}^p (v_k(i) - v_k(j))^2}
 \end{aligned} \tag{15.4.24}$$

$$\begin{aligned}
 \cos(\mathbf{v}(i), \mathbf{v}(j)) &= \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\|\mathbf{v}(i)\| \cdot \|\mathbf{v}(j)\|} \\
 &= \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\sqrt{\sum_{k=1}^p v_k(i)^2 \cdot \sum_{\ell=1}^p v_{\ell}(j)^2}}
 \end{aligned} \tag{15.4.25}$$

The terms in angular brackets, $\langle \mathbf{v}(i), \mathbf{v}(j) \rangle$, are inner products.⁷ It is important to note that these formulas only pertain to vectors and represented by rectilinear orthogonal coordinate systems such as those in Euclidean spaces. Also note that other norms and distance functions can be defined, but these are less important in cheminformatics and will not be discussed in this chapter.

Table 15.4 provides a summary of formulas for several vector-based similarity functions, all of which are symmetric. Comparable formulas can also be defined for continuous functions, the only difference being that summations over a finite number of elements as shown in Equations 15.4.23, 15.4.24, and 15.4.25 are replaced by integrations over the ranges of the continuous variables (see [41] for additional details).

Vector- and function-based similarity functions also have ordering properties similar to those given in Equation 15.4.16 for set-based functions [99], although Tanimoto similarity can be an outlier in some cases. There is, however, a caveat,

TABLE 15.4 Five Similarity Functions (Coefficients) Commonly Used with Vector- and Function-Based Representations

Similarity Function	Vector-Based Formulas	Summation-Based Formulas
Tanimoto	$S_{\text{Tan}}(i, j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\ \mathbf{v}(i)\ ^2 + \ \mathbf{v}(j)\ ^2 - \langle \mathbf{v}(i), \mathbf{v}(j) \rangle}$	$S_{\text{Tan}}(i, j) = \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\sum_{k=1}^p v_k(i)^2 + \sum_{k=1}^p v_k(j)^2 - \sum_{k=1}^p v_k(i) \cdot v_k(j)}$
Hodgkin–Richards (Dice)	$S_{\text{H-R}}(i, j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\frac{1}{2}(\ \mathbf{v}(i)\ ^2 + \ \mathbf{v}(j)\ ^2)}$	$S_{\text{H-R}}(i, j) = \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\frac{1}{2}\left(\sum_{k=1}^p v_k(i)^2 + \sum_{k=1}^p v_k(j)^2\right)}$
Carbo (Cosine)	$S_{\text{Car}}(i, j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\sqrt{\ \mathbf{v}(i)\ ^2 \cdot \ \mathbf{v}(j)\ ^2}}$	$S_{\text{Car}}(i, j) = \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\sqrt{\sum_{k=1}^p v_k(i)^2 \cdot \sum_{k=1}^p v_k(j)^2}}$
Min	$S_{\text{Max}}(i, j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\min[\ \mathbf{v}(i)\ ^2, \ \mathbf{v}(j)\ ^2]}$	$S_{\text{Max}}(i, j) = \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\min\left[\sum_{k=1}^p v_k(i)^2, \sum_{k=1}^p v_k(j)^2\right]}$
Max	$S_{\text{Min}}(i, j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\max[\ \mathbf{v}(i)\ ^2, \ \mathbf{v}(j)\ ^2]}$	$S_{\text{Min}}(i, j) = \frac{\sum_{k=1}^p v_k(i) \cdot v_k(j)}{\max\left[\sum_{k=1}^p v_k(i)^2, \sum_{k=1}^p v_k(j)^2\right]}$

The names in parentheses are those commonly used for molecular-fingerprint-based similarity functions. All of the similarity functions are symmetric. Similarity functions for function-based representations are identical in form to those for vector-based representations except that the summations are replaced by integrations over the corresponding function spaces [41].

namely that the components of the vectors and the function values must all be positive, although vectors with negative components and functions with negative values (e.g., molecular electrostatic potentials) can in many cases be handled [99].

Vector-based methods can use combinations of 1D, 2D, and 3D descriptors (*vide supra*) and are considerably faster computationally than function-based methods (cf. [100]) for two reasons. First, summations over vector components are generally faster than integrations over whole molecules. In one case, fast 3D similarity searches were carried out using a vector-like representation whose components were derived from a set of molecular shape-based criteria [100].

Second, function-based methods require optimization with respect to the alignment of the fields,⁸ a procedure that can involve many evaluations of the similarity function. This has been somewhat ameliorated by the development of new search methods that employ Fourier transforms that expedite the search process.

Since the similarity function is nonlinear, multiple maxima may occur. Thus, as with most nonlinear optimizations, locating the global maximum can be difficult. Even in cases where the global maximum is attained, it is no guarantee that the alignment obtained is the correct one. This issue and its possible solution, which involves evaluating the similarity function for multiple molecules simultaneously, has been discussed in several publications [41, 101]. Lastly, because the alignments depend on the 3D structure of molecules, molecules with rotatable bonds are problematic. Most methods typically employ single conformations (presumably the lowest energy conformation). Some methods for multiple conformations have also been developed [81, 102] (also see the discussion in [41]), although multiconformer procedures have not seen wide application.

As shown by Maggiora et al. [99], the denominators of field-based similarity functions correspond to different types of averages, a feature that also applies to similarity functions associated with vectors with continuous valued components. This enabled the authors to show that the different similarity functions are ordered in an identical manner to that given in Equation 15.4.10 and discussed in Section 15.4.4.1 for set-based similarity functions.

15.5 SOME ISSUES IN MOLECULAR SIMILARITY ANALYSIS

Now that the essential features of molecular similarity measures have been described, a number of issues pertaining to their behavior will be discussed:

- What is the role (if any) of asymmetric similarity?
- Are 2D similarity methods better than 3D methods?
- Do data fusion and/or consensus-based similarity measures yield improved results?
- Since similarity is subjective, how can we validate similarity measures?
- How do we compare similarity measures?

TABLE 15.5 Parameter Values That Convert the Tversky Similarity Function, Equation 15.5.1, into the Other Well-Known Set-Based Similarity Functions Listed in Table 15.3

Parameter Values ^a	Similarity Functions
$\alpha = \beta = 1$	$S_{\text{Tve}}(i,j 1,1) = S_{\text{Tan}}(i,j)$
$\alpha = \beta = \frac{1}{2}$	$S_{\text{Tve}}\left(i,j\left \frac{1}{2},\frac{1}{2}\right.\right) = S_{\text{Dice}}(i,j)$
$\alpha = 1, \beta = 0, \rho$	$S_{\text{Tve}}(i,j 1,0) = \rho^{-1} \cdot S_{\text{Cos}}(i,j)$
$\alpha = 0, \beta = 1, \rho$	$S_{\text{Tve}}(i,j 0,1) = \rho \cdot S_{\text{Cos}}(i,j)$
$\alpha = 1, \beta = 0$	$S_{\text{Tve}}(i,j 1,0) = S_s(i,j)$
$\alpha = 0, \beta = 1$	$S_{\text{Tve}}(i,j 0,1) = S_s(j,i)$

^aThe parameter ρ is the proportionality constant that relates $|\mathbf{m}(j)|$ to that of $|\mathbf{m}(i)|$, that is, $|\mathbf{m}(i)| = \rho \cdot |\mathbf{m}(j)|$ (see Equation 15.5.2 and related discussion).

15.5.1 Asymmetric Similarity

Similarity is generally considered to be symmetric: If A is similar to B, then B must be equally similar to A. Tversky has challenged this notion and provided a number of general examples of asymmetric similarity [45]. Although it has been slow in coming, the emerging role of asymmetric similarity in chemical informatics is beginning to be recognized [41, 96, 97, 103–108]. Essentially, all cheminformatics applications of asymmetric similarity are based on the expression first described by Tversky [45],

$$S_{\text{Tve}}(i,j|\alpha,\beta) = \frac{|\mathbf{m}(i \cap j)|}{\alpha|\mathbf{m}(i-j)| + \beta|\mathbf{m}(j-i)| + |\mathbf{m}(i \cap j)|} \quad (15.5.1)$$

where $0 \leq S_{\text{Tve}}(i,j) \leq 1$ for all $\alpha, \beta \geq 0$.^{9, 10} Moreover, as shown in Table 15.5, by selecting different values of α and β , many of the other popular similarity functions can be generated.

Clearly, Equation 15.5.1 only strictly applies to set-based representations, although closely related asymmetric similarity functions can also be defined for graph-, vector-, and function-based representations (see Table 15.3 and Table 15.4 and the associated discussions in Sections 15.4.2 and 15.4.4). Because most applications employ set-based similarity functions and binary molecular fingerprints, the discussion in this section focuses on this category of MSA. Equivalent analyses can, however, be carried out with respect to other similarity measures (see e.g., [46]).

Recent work by Chen and Brown [108] employed a slightly more restrictive form of Equation 15.5.1, where $\beta = 1 - \alpha$,

$$S_{\text{Tve}}(i,j|\alpha) = \frac{|\mathbf{m}(i \cap j)|}{\alpha|\mathbf{m}(i-j)| + (1-\alpha)|\mathbf{m}(j-i)| + |\mathbf{m}(i \cup j)|} \quad (15.5.2)$$

where $|\mathbf{m}(i-j)|$ is the numbers of bits unique to the i th molecule—nominally the *reference (probe)* molecule—and $|\mathbf{m}(j-i)|$ is the number of bits unique to the j th molecule—nominally the *target (database)* molecule. Although Equation 15.5.2 cannot represent the Tanimoto similarity function, which requires that $\alpha=\beta=1$ in Equation 15.5.1, this is not a problem since, as shown by Gower [76] (see also footnote 6 in this work), S_{Tan} is monotonically related to S_{Dice} , which can be obtained from Equation 15.5.2, when $\alpha=0.5$. Equation 15.5.2 can be written in a simpler and more revealing form by combining the contribution of $|\mathbf{m}(i \cap j)|$ with those of $|\mathbf{m}(i)|$ and $|\mathbf{m}(j)|$ giving

$$S_{\text{Tve}}(i,j|\alpha) = \frac{|\mathbf{m}(i \cap j)|}{\alpha|\mathbf{m}(i)| + (1-\alpha)|\mathbf{m}(j)|} \quad (15.5.3)$$

This form of Equation 15.5.2 clarifies the role of α in modulating the contributions of $|\mathbf{m}(i)|$ and $|\mathbf{m}(j)|$ to $S_{\text{Tve}}(i,j|\alpha)$, as discussed further in the following paragraph.

Chen and Brown used three different molecular fingerprints to study the hit rates of compounds in the NCI anti-AIDS and Johnson & Johnson Corporate databases for values of α that varied from $0 \rightarrow 1$ in increments of 0.1. Hit rates were calculated as the ratio of actives n to the m top nearest neighbors. From Equation 15.5.3, it is clear that increasing α weights the similarity function toward the reference (i th) molecule. The authors showed that higher hit rates were generally favored by higher α values. However, as will be seen in subsequent discussion, this observation depends somewhat on the size (complexity) of the reference molecule and the mean size (complexity) of the molecules in the database being searched. As $\alpha \rightarrow 1$, the weight for the probe molecule becomes much greater than that for the target (j th) molecule, and Equation 15.5.3 approaches the asymmetric similarity function $S_*(i,j)$. It is, however, important to note that the value of $S_*(i,j)$ depends on the relative “sizes” of the probe and target molecules i and j , respectively—smaller probes tend to lead to larger asymmetric similarity values and vice versa. Thus, the correlation of higher hit rates with higher α values indicates that a number of factors, some hidden, may be at work here (*vide infra*).

An interesting observation made by Chen and Brown is that

to retrieve more remotely similar active compounds, a more asymmetric similarity measure should be used, that is, more weight should be put on the probe compound. On the other hand, to retrieve more highly similar active compounds, approximately equal weights should be put on both the probe and target compounds, leading to a measure close to symmetric.

However, the arguments associated with Equations 15.4.10–15.4.16 and discussed in Section 15.4.1 suggest that in the case of large probe molecules the asymmetric

similarity function will tend to have diminished values for all target molecules except for other comparably large molecules. And there are other subtle issues as well. Thus, a more complete interpretation of the interesting work of Chen and Brown awaits the results of future studies.

Wang and Bajorath [97] have carried out an extensive study based on their earlier work [96]. Both studies used the Tversky similarity function given in Equation 15.5.2, to assess how molecular complexity (“size”) and bit density influence the results of similarity searches based on molecular fingerprints. Generally, but not always, molecular complexity and bit density are closely related, that is, more complex molecules tend to have greater bit densities than less complex molecules. A key element of their study is the construction of bit-density invariant similarity functions that account for the distribution of both 1-bits and 0-bits. The functions are based on *weighted* combinations of terms of the form given in Equation 15.5.2 or 15.5.3

$$S_{\text{Tve}}(i,j|\alpha,\beta) = \beta \cdot \frac{|\mathbf{m}(i \cap j)|_1}{\alpha |\mathbf{m}(i)|_1 + (1-\alpha) |\mathbf{m}(j)|_1}, \\ + (1-\beta) \cdot \frac{|\mathbf{m}(i \cap j)|_0}{\alpha |\mathbf{m}(i)|_0 + (1-\alpha) |\mathbf{m}(j)|_0} \quad (15.5.4)$$

where $0 \leq \beta \leq 1$ and the subscripts “1” and “0” correspond to whether the terms are associated with 1-bits or 0-bits. Varying α affects the relative contributions of reference-probe and target-database molecules to the value of the 1-bit and 0-bit asymmetric Tversky similarity functions, respectively; varying β accounts for the influence of molecular complexity and associated bit-density factors on the combined similarity value. Using this strategy, Wang and Bajorath achieved improved results in similarity searches over a number of compound sets corresponding to different activity classes.

Combining asymmetric similarity measures, regardless of how it is accomplished, results in a loss of information. A simple example from statistics illustrates this point. Taking an average (weighted or otherwise) of a set of numbers is not a reversible process since the number of sets that can yield the same average is infinite. Thus, knowing only the average tells you nothing about which set produced the observed average. This argument can be applied to similarity searching using pairs of asymmetric similarity functions such as those of Tversky [45]. Although combining them yields a single similarity value that is easier to manipulate using “traditional” similarity search algorithms, the loss of information may be too great?

15.5.1.1 2D Asymmetric Similarity Searching In what follows, a brief description of an alternative approach will be outlined that does not combine the similarity values of the two asymmetric similarity functions [109], leading to what might be called 2D asymmetric similarity searches. Since there are now two asymmetric similarity functions rather than a single symmetric function, will the added information lead to more effective similarity searches?

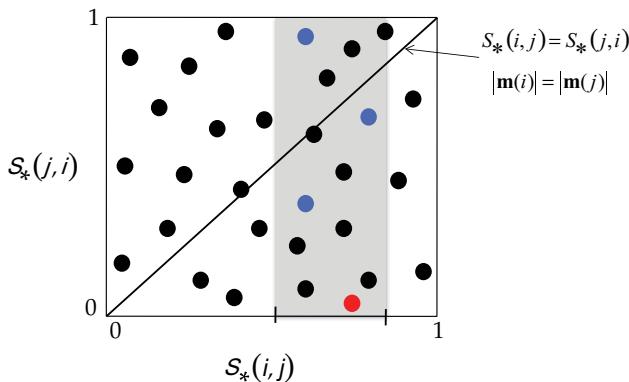


FIGURE 15.4 2D plot of asymmetric similarity functions. The red dot corresponds to a pair of molecules where the i th molecule, which is taken to be active, is “smaller” than the j th molecule with which it is paired. The blue dots located within the gray shaded region of the plot are also associated with the i th molecule but the molecules with which it is paired decrease in size (relative to the i th molecule) as one moves vertically up the gray shaded region. Dots located near the diagonal correspond to cases where the asymmetric similarities are approximately equal in value. For color details, please see color plate section.

Consider the pair of asymmetric similarity functions given in Equations 15.4.12 and 15.4.13 and in Table 15.3 as the *ordered pair* (or 2D vector)

$$(S_*(i,j), S_*(j,i)) \quad (15.5.5)$$

that can be mapped onto a plot such as that given in Figure 15.4 for a hypothetical set of molecules. As mentioned in Section 15.4.4.1, similarity functions identical to these were described by Holliday et al. [5]. Equation 15.4.15 shows that the two asymmetric similarity functions are related by a proportionality constant that is approximately equal to the ratio of the respective sizes of the molecules. Thus, molecules associated with ordered pairs that lie on the diagonal of the plot are approximately equal in size, those that lie below the diagonal satisfy $|m(i)| < |m(j)|$, and those that lie above the diagonal satisfy $|m(i)| > |m(j)|$. For molecules corresponding to points in the lower right corner $|m(i)| << |m(j)|$, while those in the upper left corner satisfy $|m(i)| >> |m(j)|$. Hence, the further a point is from the diagonal, the greater the relative disparity in the sizes of the pair of molecules associated with that point.

Now consider the molecule pairs corresponding to points in the lower left and upper right corners of the diagram. In the former case, $|m(i \cap j)| \approx 0$, while in the latter case, $|m(i \cap j)| \approx |m(i)| \approx |m(j)|$. Lastly, consider the cases where $m(i) \subset m(j)$. In such cases, $|m(i \cap j)| \cap |m(i)|$ and $S_*(i,j) \approx 1$ and $S_*(j,i) \approx |m(i)|/|m(j)|$. These relationships are summarized in Figure 15.5.

How can this information be used effectively in similarity searches? As an example, consider the red point in the vicinity of the lower right corner of the diagram in Figure 15.4; thus, $|m(i)| << |m(j)|$. Suppose the i th molecule, which is much

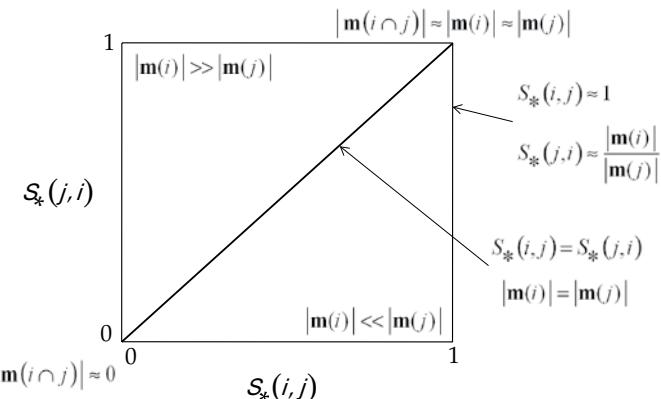


FIGURE 15.5 2D plot of asymmetric similarity functions showing a number of relationships of the asymmetric similarity functions and the cardinalities of the molecular fingerprints used to represent the *i*th and *j*th molecules.

smaller than the *j*th molecule, is active. There are several possible strategies for identifying other potentially active molecules. One possibility is to identify all pairs containing the *i*th molecule, which are indicated in blue, whose similarity values lie within a given range such as that indicated by the gray shading in Figure 15.4. Confining the search to molecules within the gray region ensures that a sufficient fraction of the active *i*th molecule is similar to the *j*th molecule with which it is paired. By moving up the rectangular region, the size of the molecules paired decreases compare to that of the active reference (*i*th) molecule. Thus, relative size information, which may be quite beneficial in such activities as “scaffold hopping” (cf. [109]), is captured directly. A similar argument can be made for the case when the reference-probe molecule is large.

This approach can be implemented as a real-time process in which investigators make decisions in a sequential fashion as to which subsets of molecules should be screened in follow-up experiments. Alternatively, algorithms could be developed that process the information without investigator intervention. Or a hybrid approach could be developed that uses some combination of the aforementioned strategies. A preliminary report of this work was presented recently [109], and a manuscript is currently in preparation.

15.5.2 2D and 3D Similarity Methods

A distinct advantage of 2D over 3D approaches to MSA is computational speed, since most 2D methods, except those utilizing chemical graphs, do not require costly structure alignments. In contrast, many but not all 3D methods require such alignments [110]. In addition, 3D methods, regardless of whether or not they require structure alignment (cf. [100]), bear an added burden due to the conformational flexibility of molecules, which gives rise to multiple conformers close in energy to

that of the most stable conformation (i.e., global minimum conformational energy). To date, most 3D methods choose a single low-energy conformer, typically the global minimum, which appears to be a strategy that produces reasonable results. However, new approaches for selecting single or representative sets of conformers for similarity searching and other applications are actively being pursued [111]. A major limitation of 3D methods is the difficulty of finding an adequate representation of the “biologically significant” conformation(s). In many practical applications, the bioactive conformations of the compounds are unknown, and the single low-energy conformation frequently used might not be truly representative. This approximation may contribute in some cases to the lower performance of 3D compared to 2D methods (*vide infra*).

In contrast to essentially all 2D approaches, most but not all 3D methods have a distinct advantage, in that they account for stereochemistry in a natural way (e.g., 3-point pharmacophore-based methods cannot describe the stereochemistry about asymmetric carbon atoms). Although important, incorporation of information on protonation and tautomeric states is at best sporadically treated in similarity methods due primarily to the difficulty of reliably estimating these states. The papers by Nettles et al. [112] and Oellien et al. [113] provide a nice overview of many of the issues confronting both 2D- and 3D-based methods, and the book by Martin [114] provides discussions and additional references on the role of tautomeric structures in chemical informatics and modeling.

The performance of 2D and 3D similarity approaches has been compared directly in a number of applications including virtual screening [115–118], diversity analysis [119], and scaffold hopping [120–122], and prediction of biological targets [113]. Because 3D methods tend to incorporate more of the “true” molecular features than 2D methods, it intuitively seems that results obtained from 3D methods should be more reliable than those obtained by 2D methods. However, in many case studies 2D methods have outperformed 3D approaches, although it has been noted that this superiority is somewhat case dependent [123]. It is certainly surprising that 2D methods in many cases are competitive with 3D ones, especially considering the generally greater amount of molecular detail incorporated into 3D methods, although the conformational issues can negatively influence the results produced by the latter class of methods. The overall conclusion based on the studies comparing different similarity methods in different applications is that 2D and 3D similarity approaches are complementary and should be used in combination.

15.5.3 Data Fusion and Consensus Methods

Since different molecular representations capture different structural, chemical, and biological aspects of molecules [117, 124], there is not always a clear answer to what molecular descriptions perform “best” in similarity searching and related similarity-based activities. In addition, different similarity measures tend to exhibit different performance characteristics in different application domains. Thus, it is unlikely that a single measure will be sufficient to effectively treat all regions of chemical space uniformly. As an alternative to using a single search method, it has been proposed

that combining the outputs of multiple similarity methods into a single output measure [125–127], commonly referred to in the literature as “data fusion” or “consensus scoring” in the case of docking and structure-based virtual screening applications [128], will result in improved performance. A recent comprehensive study by Holliday et al. [129] provides strong evidence that suggests fusion-based approaches to similarity searching yield improved results over single-search-based similarity methods. Earlier, but much less extensive, studies provide additional evidence [130, 131].

Data fusion has been around for a number of years, primarily in the engineering community [132, 133]. Typically, outputs from multiple sensors were combined (fused) into a single number that in some way reflected the multiple input values obtained from the sensors. The first applications of data fusion in MSA were published in the mid-late 1990s [134, 135].

There are two major approaches, nominally called *similarity fusion* and *group fusion*. Application of these procedures is exemplified by similarity searches carried out in LBVS studies [58, 59]. In the case of similarity fusion, a set of similarity values is computed with respect to a *single* active reference (query) molecule, i_{REF} , using a number of different similarity measures, $S_K(i_{\text{REF}}, j_{\text{DB}})$, where $K=1, 2, \dots, N_{\text{SIM}}$ corresponds to the set of similarity measures and $j_{\text{DB}}=1, 2, \dots, N_{\text{DB}}$ corresponds to the set of database compounds being searched. A list of similarity values is generated for each of the N_{SIM} different similarity measures. The values are then combined using a simple mathematical function to determine their maximum (max), minimum (“min”), sum, mean, or median fusion scores (cf. [136])

$$\hat{S}_{\text{SF}}(i_{\text{REF}}, j_{\text{DB}}) = f(S_1(i_{\text{REF}}, j_{\text{DB}}), S_2(i_{\text{REF}}, j_{\text{DB}}), \dots, S_{N_{\text{REF}}}(i_{\text{REF}}, j_{\text{DB}})), \quad (15.5.6)$$

$$j_{\text{DB}} = 1, 2, \dots, N_{\text{DB}}$$

and a new list ordered from largest to smallest similarity is generated from which typically the top 100–500 molecules are selected for screening. Alternatively, the compounds in each list could be ranked and their rankings combined in a similar manner to that given in Equation 15.5.6, that is,

$$\hat{R}_{\text{SF}}(i_{\text{REF}}, j_{\text{DB}}) = h(R_1(i_{\text{REF}}, j_{\text{DB}}), R_2(i_{\text{REF}}, j_{\text{DB}}), \dots, R_{N_{\text{REF}}}(i_{\text{REF}}, j_{\text{DB}})). \quad (15.5.7)$$

$$j_{\text{DB}} = 1, 2, \dots, N_{\text{DB}}$$

The ranking procedure appears to be the best approach for similarity fusion [126, 127]. An interesting recent paper by Chen, Holliday, and Bradshaw [137] describes a machine learning approach to the development of a weighting scheme for combining the results obtained from a set of different similarity measures.

Instead of a single reference molecule, i_{REF} , group fusion uses a set of $i_{\text{REF}}=1, 2, \dots, N_{\text{REF}}$ active reference molecules, but only a single similarity measure, $S(i_{\text{REF}}, j_{\text{DB}})$. Each reference molecule is then searched against the entire database and a set of lists corresponding to each of the reference compounds is generated. As was

the case for similarity fusion, group fusion values can be calculated in a similar manner using similarity values

$$\hat{S}_{\text{GF}}(i_{\text{REF}}, j_{\text{DB}}) = f(S(1, j_{\text{DB}}), S(2, j_{\text{DB}}), \dots, S_{N_{\text{REF}}}(N_{\text{SIM}}, j_{\text{DB}})) \quad (15.5.8)$$

$$j_{\text{DB}} = 1, 2, \dots, N_{\text{DB}}$$

or rankings

$$\hat{R}_{\text{SF}}(i_{\text{REF}}, j_{\text{DB}}) = h(R_1(1, j_{\text{DB}}), R_2(2, j_{\text{DB}}), \dots, R_{N_{\text{REF}}}(N_{\text{SIM}}, j_{\text{DB}})) \quad (15.5.9)$$

$$j_{\text{DB}} = 1, 2, \dots, N_{\text{DB}}$$

and a new fusion-based list of the top scoring molecules can be generated. Numerous studies by Willett and coworkers [54, 58, 59, 127, 128, 130, 139] suggest that the group fusion approach is generally superior to similarity fusion.

A variant of group fusion called turbo similarity is employed when a single reference structure is available [138]. It is an iterative procedure that takes a subset of the retrieved compounds with high similarity to the active reference compound and uses these “hits,” whether or not they are active, as reference compounds in the next iteration—thus, the correspondence of this procedure with group fusion. The process, which can be continued if desired, is also reminiscent of document retrieval methods that take the “hits” from a given query as queries for subsequent retrievals [51].

The combination of fusion rules (e.g., maximum and mean-fusion) gave rise to the multi-fusion similarity (MFS) maps that were developed for the visual characterization and comparison of compound databases [139]. This approach has been employed to explore SARs of compound datasets [140] and to compare combinatorial libraries [141]. Consensus approaches are also applied in diversity analysis of compound collections; complementary 2D and 3D representations are used to obtain a comprehensive characterization of the diversity of large compound databases [60, 142].

The concept of data fusion has also been extended to activity landscape modeling.¹¹ It is well known that activity landscapes will be largely influenced by the choice of the molecular representation that is used to define the chemical space. In an effort to address this issue, multiple structural representations are combined using data fusion to derive a consensus model of activity landscapes and identify *consensus activity cliffs* [143–145]. Consensus models are designed to prioritize the SAR analysis of activity cliffs and other consistent regions in the activity landscape that are captured by several structure representations. They are not meant to be a means for eliminating data by disregarding, for example, “true” activity cliffs that are not identified by some structure representations.

15.5.4 Statistical Independence of Similarity Measures

An issue that comes into both similarity fusion and consensus methods is the independence of the different similarity measures employed in a given procedure. This issue also occurs in activity landscape modeling (see Section 15.6.1) when fusion or related methods are used to estimate molecular similarities [144, 146]. In such cases, the question typically arises when two methods produce highly correlated values as to whether the set of values from one of them should be removed from the fusion process. The answer seems to be yes, although most of the time in cheminformatics it is ignored. But should it be?

There is an alternative way to think about the problem. Consider first how certain types of sporting events, such as ice-skating, gymnastics, and diving, are scored. A set of judges, all presumably with experience in judging the event and acting (supposedly) independently, produce a set of scores. Since it is assumed that the judges are acting independently and are experienced, the fact that two or more judges have the identical or nearly identical scores is considered as evidence that the scores are an accurate reflection of the athlete's performance.

If the relationship of the relative scores of the same judges is repeated for numerous athletes in the competition, the scores of the judges in question most likely will be linearly correlated. In such cases, should the scores of some of the judges that are statistically correlated with the scores of some of the other judges be eliminated? The answer would surely be no, which seems to violate the observed linear correlation of the judge's scores. *This illustrates an extremely important point regarding statistical correlations, namely, that they do not necessarily imply an underlying mechanistic or causal relationship.* It appears that the case of sports judging exemplifies this point.

How does this example apply to the use of multiple similarity methods? Each of the similarity methods can be considered to be equivalent to an independent judge, since none of the values produced by the other methods have an explicit impact on the value produced by a given method. This may not always be the case, for example, if two methods use MACCS key fingerprints, but one uses the Tanimoto (Jacard) and the other a closely related similarity function (see Table 15.3). As shown by Gower [76], some molecular similarity functions are monotonically related. Thus, comparisons of these functions based on the same molecular representation will produce linear correlations of the values computed by the two functionally similar functions. Hence, only one of the functions should be used.

However, the question remains as to how the independence of two similarity measures can be established in analogy to the judging example described earlier. One possible approach is to determine the relationship of the representations to each, but it is not immediately obvious how to accomplish this since the mathematical forms of different representations can be quite varied (from molecular fingerprints to property vectors to 3D electron density or related distribution functions). While it may not be possible to solve this problem in general, it may be possible to solve a more limited subproblem by, for example, assessing the relationship of different representations within the same general class such as molecular fingerprints.

All of this suggests that when aggregating similarity scores by data fusion or related methods, the scores produced by all of the methods should be considered unless the methods generating the scores are too closely related with respect to the representations and similarity functions being used.

15.5.5 Comparing Similarity Measures

There are a number of ways to compare similarity measures, most of which involve linear correlations in some fashion. This situation is to a large extent different from that described in Section 15.5.4, although when comparing the results of similarity measures one should take care not to use closely related similarity measures (*vide supra*).

As will be discussed in Section 15.6, similarity measures have a significant effect on the nature of the chemical spaces they induce. As noted in earlier sections, nearest-neighbors in one chemical space may not be nearest-neighbors in another. Methods for comparing different similarity measure provide a means for quantitatively assessing how closely related they are to each other. Three approaches are described here (see [40] for additional discussion).

In the first approach, which is also related to material covered in Section 15.5.6, consider a specific reference (probe) molecule, m_R , that may be active in some assay. The issue now is to identify the, say 250, molecules in a large compound database that are most similar to m_R with respect to a given similarity measure. This creates a list, L_1 , of molecules that can be ordered from smallest to largest similarity value. The process is now repeated $R - 1$ times using other similarity measures that are not functionally related in a mathematical sense (see Gower [76] and the discussion in Section 15.5.3 for further discussion). This yields a set of R ordered lists $L = \{L_1, L_2, \dots, L_R\}$ that can be compared in a pairwise fashion using statistical correlation methods in the sequel.

There are two issues with this approach, correlation methods for continuous variables are not very robust, and, more importantly, the retrieved molecules in the pair of lists being compared may not be the same—some molecules present in one list may be missing from the other, and vice versa. Such data are called partially ranked. The issue of robustness can be handled to some degree using methods based on ranked correlations. The second problem, however, remains. Fortunately, methods also exist for handling partially ranked datasets [146]. Although the subject is challenging to master, requiring some sophisticated mathematics, Critchlow's book provides a reasonably clear introduction to it including a set of computer programs for evaluating the appropriate correlation coefficients.

Correlation methods for partially ranked data are useful in applications of ligand-based and structure-based virtual screenings. In both cases, ranked lists of compounds are produced by any number of methods, which ones are not important. The important point is that lists produced in this manner may not contain the same set of molecules (*vide supra*). As noted earlier, correlation methods for partially ranked data proved a means of evaluating how the two types of methods performed. This approach can also be applied to pairwise comparisons of compound lists generated by ligand- and

structure-based virtual screens as well as experimental screens of biological activity carried out in the laboratory. In addition to their robustness, pairwise comparisons of ranked lists have the advantage that the numbers in the list need not refer to the same entities. For example, one list could be ordered with respect to similarity values and the other with respect to biological activities.

An issue raised in Section 15.5.3 is relevant here. Suppose two similarity measures are shown to be highly rank correlated. Basically, this means that both methods are producing compound rankings that are quite similar. If both measures are not directly related mathematically (*vide supra*), this implies that both are acting like the independent judges described in Section 15.5.3. Thus, the fact that they produce similar rankings should reinforce, at least to some degree, our confidence in the reliability of the methods. The more methods that are correlated in this way, the more confidence one should have. This, of course, is a rather intangible approach to validating similarity measures, an important subject that will be addressed in more detail in Section 15.5.6.

The method described so far is very limited since it is confined to similarity searches only within the neighborhood of a single active reference molecule. The process can be easily extended, however, to cover multiple active reference molecules. The results can then be combined using any one of a number of statistical or aggregation methods [90] if a single value is desired. Although conceptually similar, obtaining a global comparison is a bit more computationally intensive. Basically, a large set of reference molecules is generated by randomly sampling the compound collection(s) under investigation. Then, as in the earlier case, the rank correlations can be combined to yield a global measure of the correlation of the two (or more) similarity measures.

The second, and perhaps most correct, way to compare similarity measures are directly through their similarity matrices. The method, pioneered by Mantel [147], has been applied in a number of areas of biological research [148, 149]. The issue raised by Mantel was the lack of statistical independence among the similarity-based elements of a similarity matrix. Consider, for example, three molecules i, j , and k . Suppose $S(i, j)$ and $S(j, k)$ are both large. Then the value of $S(i, k)$ will be constrained by the values of the first two similarities, and, thus, is dependent to some degree on the other similarity values. Hence, the similarity value of the ordered pair (i, k) is not statistically independent.

Mantel dealt with this using what he called the z_M statistic,

$$z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^n S(i,j) \cdot S'(i,j). \quad (15.5.10)$$

where the summations are only over the elements in the upper triangle of the similarity matrices $S(i, j)$ and $S'(i, j)$ because of their symmetry. In some cases, a standardized Mantel statistic, r_M , is used, the only difference between them being that the similarity values are standardized to zero means and unit standard deviations and a multiplicative factor of $\frac{1}{2} n(n-1) - 1$ that accounts for the degrees of freedom is added.

Implementing the Mantel statistic requires an iterative procedure that is as follows: (1) permute the rows and columns of the similarity matrix \mathbf{S} (*N.B.* that either similarity matrix can be permuted with the same result), (2) compute either z_M or r_M , and (3) repeat the first two steps. This process is carried out a number of times in order to estimate the sampling distribution of the Mantel statistic under the assumed (null) hypothesis H_0 that the similarities in \mathbf{S} are not linearly correlated with the corresponding similarities in \mathbf{S}' . The Mantel statistic derived directly from the similarity matrices is then compared with its distribution derived under H_0 . Based on the distribution, if the Mantel statistic is likely to have been obtained, then the null hypothesis is accepted. Otherwise the alternative hypothesis, H_1 , that the similarities between the two matrices are correlated is accepted.

Thus, there exist a number of ways that similarities can be compared, two of which have been described in the current chapter.

15.5.6 Validating Similarity Measures

Model validation is a requirement in the development of almost all cheminformatics and modeling methods. However, there are cases where it might not be totally appropriate. One such case would appear to be MSA. Due to the subjective nature of similarity, it is not possible to come up with well-defined values upon which to test predictions. Attempts to address this problem make use of the similarity-property principle that “similar compounds (tend to) have similar properties and activities” [22]. The basic approach involves determining the recovery rates (or some other related measure) of active compounds from databases known to contain other actives. This is accomplished using a single or small set of active molecules to query the databases for similar molecules that are also known to be active. Importantly, untested compounds, which may be active, are *assumed* to be inactive.

Although this approach is widespread, it is unsatisfactory for a number of reasons. For one, recovery rates tend not to sufficiently account for “early enrichments” in sets of retrieved compounds. This deficiency can be partially eliminated using cumulative recall curves, which plot the fraction of actives against the number of compounds retrieved [58, 150]. These curves are similar to the receiver operating characteristic (ROC) curves that have become a popular in compound retrieval studies. Truchon and Bayly [151] have carried out a comprehensive analysis ROC curves and related metrics and have developed an optimal, statistically more robust index, the *BEDROC* (Boltzmann-enhanced discrimination of ROC) metric.

A serious limitation of all similarity-based methods is that chemical spaces are not invariant to the representation (or similarity function) used. Thus, nearest-neighbors in one space may not be nearest-neighbors in another [41], and what is similar to an active compound in one space (hence also potentially active) may not be in another, begging the question of which space should be used. In addition, the presence of activity cliffs [152–156], which occur when small changes in structure are accompanied by large changes in activity, can further confound the application of similarity methods since they violate the similarity-property principle [22] leading to what Stahura and Bajorath [157] have termed the “similarity paradox.”

As was discussed in detail in Section 15.5.3 (see also the related discussion in Section 15.6), fusion- and consensus-based methods provide possible means for improving similarity methods by combining the results from multiple methods, and thus providing a more comprehensive and robust account of the molecular features that are the foundation of all similarity methods. Results from numerous computation experiments appear to bear this out [54, 58, 59, 126–128, 130, 139].

The inherently subjective character of similarity (see Section 15.3 for a discussion of the cognitive aspects of similarity in general) suggests that regardless of the similarity measures used, it will never be possible to account for the similarities among any set of entities or concepts in some absolute sense. Hence, the best we can hope for in the realm of MSA is to obtain similarity measures that capture the structural, chemical, and/or biological features of molecules in a way that provides a *useful* account of their relationship to one another.

Even with the many caveats noted before, MSA has been and will continue to be a very useful tool in cheminformatics and chemical biology.

15.6 AN APPLICATION OF MOLECULAR SIMILARITY ANALYSIS TO CHEMICAL SPACE AND ACTIVITY LANDSCAPES

15.6.1 Representing Chemical Spaces

Over the last decade, the concept of *chemical space* has begun to play an increasingly important role in cheminformatics and medicinal chemistry until it now pervades many aspects of drug research. Chemical space provides a means for systematically describing the “chemical universe,” which is immense but finite. Most chemical spaces are *coordinate-based*, where each molecule is represented by a point in the space whose position is determined by a set of coordinate values that measure the displacement of the point along each of the coordinate axes defining the space. Since chemical spaces are usually considered to be Euclidean, their corresponding coordinates axes are mutually orthogonal. Typically, the coordinates are associated with the various types of descriptors that are related to molecular, chemical, topological properties, or to the occurrence of substructural fragments (see Section 15.4.2 for further discussion). Figure 15.6a presents a simple 3D example of a coordinate-based chemical space. The red dot represents a biologically active molecule, which is the i th molecule in the set of molecules displayed in the figure, whose position is given by the 3D vector defined by the suitably normalized set of descriptor-based coordinates $(D_1(i), D_2(i), D_3(i))$. The green dot represents the molecule nearest to the red active, that is, its nearest neighbor, and the blue dots represent other molecules in the neighborhood of the active.

Since most molecular representations are actually of high dimension (cf. [74]), their corresponding chemical spaces are intrinsically of high dimension as well. The spatial properties of high-dimensional spaces can, in some cases, give rise to surprising problems since they tend to behave in a manner that is uncharacteristic of low-dimensional spaces [158, 159]. It is possible, however, to construct lower dimension representations of chemical spaces by computing the similarity of or

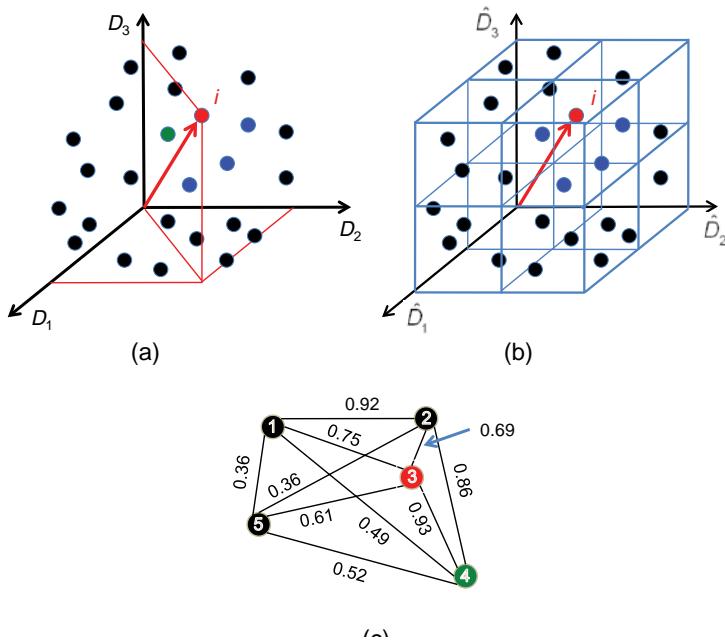


FIGURE 15.6 Simplified schematic diagrams of chemical spaces: (a) coordinated-based, (b) cell-based, and (c) graph-based. The red dot in each figure denotes a biologically active molecule; the green dots designate the nearest neighbor; the blue dots in (a) and (b) represent molecules in the neighborhood of the active (see text for further discussion). For color details, please see color plate section.

distance between molecules with respect to a given representation, and then using these new entities as a basis for constructing new Euclidean spaces [41], although as Gower has pointed out [76], it may not always be possible to construct such Euclidean spaces.

Most chemical spaces are considered to be Euclidean. However, non-Euclidean chemical spaces, which generally better represent the underlying manifold structure of the data, have been investigated by Agrafiotis [160–162]. Although these spaces are of lower dimension than equivalent Euclidean spaces, they are generally more difficult to work with. A major reason such nonlinear spaces are not in mainstream use today, however, is their lack of availability in commercial software.

Alternatively, cell-based chemical spaces can be constructed. Such spaces represent a coarse-grain approximation of coordinate-based spaces, since the latter are needed in the construction of cell-based spaces. All molecules within the same cell are now considered as having identical “coordinates,” which are usually given as integers that describe the partition along each axis. Figure 15.6b portrays the same chemical space depicted in Figure 15.6a as a comparable cell-based chemical space. As shown in the figure, the red dot again corresponds to the active molecule (*i*), which is now described by the integer vector $(\hat{D}_1(i), \hat{D}_2(i), \hat{D}_3(i)) = (2, 2, 2)$ as it

resides in the upper right most cell in the space. Since all of the molecules represented by the blue dots lie within the same cell of the cell-based chemical space as the active, they can all be considered to lie in its neighborhood. In the case of cell-based spaces, the notion of nearest neighbor is somewhat diminished since all of the molecules with a given cell have the same cell coordinates and, in a sense, are nearest neighbors.

It should be noted that a coordinate-based representation of chemical space is not absolutely required. It is also possible to construct coordinated-free similarity- or distance-based chemical spaces that are represented as massive networks of interconnected molecules. Mathematically, the networks are described by undirected, labeled, or fuzzy graphs [71, 50], where each molecule is represented as a vertex of the graph and the edges connecting each pair of vertices (i.e., molecules) are labeled with appropriate similarity or distance values. Figure 15.6c provides a very simple example of a graph-based representation of chemical space, where each edge is labeled with the similarity value associated with the pair of molecules linked by the edge. The red-filled circle, which corresponds to the third molecule in the set, is taken to be a biologically active molecule; its nearest neighbor is the fourth molecule in the set. If molecules that are at least 0.85 similar to the active are considered in a turbo similarity search (see Section 15.5.3 and [139]), the fourth molecule would be selected on the first iteration and the second molecule would be selected on the second iteration since it has a similarity that is greater than 0.85 with respect to the fourth molecule, which is the reference molecule for the second iteration of similarity searching. The figure clearly shows that such graph is complete, since all of its vertices are connected. Because of this, depicting graph-based chemical spaces is not possible for compound collections of realistic size.

Recently, networks typically represented by unlabeled, undirected graphs have been used to depict many relationships between entities of chemical and biological interest [163–166], some of which are detailed in other chapters of this volume.

Different representations typically yield chemical spaces of different dimension, regardless of whether the underlying structure of the spaces is linear or nonlinear, which raises the question “Is there a true dimension to chemical space?” Whether a “true” dimension exists, it is quite likely that they possess fractal geometries. If this is the case, how does one determine its dimension? Using the so-called box counting procedure, Tominaga [167] analyzed the fractal dimensions of a number of compound datasets. Agrafiotis and Rassokhin tackled a related problem, namely, estimating the bin size of cell-based chemical spaces for diversity estimation [168]. Fractal dimension can also be estimated by the “cluster growing” method [169], which is based on a labeled-graph representation of chemical space similar to that given in Figure 15.6c. A distinct advantage of this method is that it does not require a coordinate-based representation of chemical space, as does the box counting method. To our knowledge, such an approach has not been applied specifically to compound datasets. More work in this area is definitely desirable.

As noted earlier, different similarity measures induce different chemical spaces. Figure 15.7 depicts the projection onto three dimensions¹² of four chemical spaces representing the same set of 2250 molecules made up of nine datasets of 250

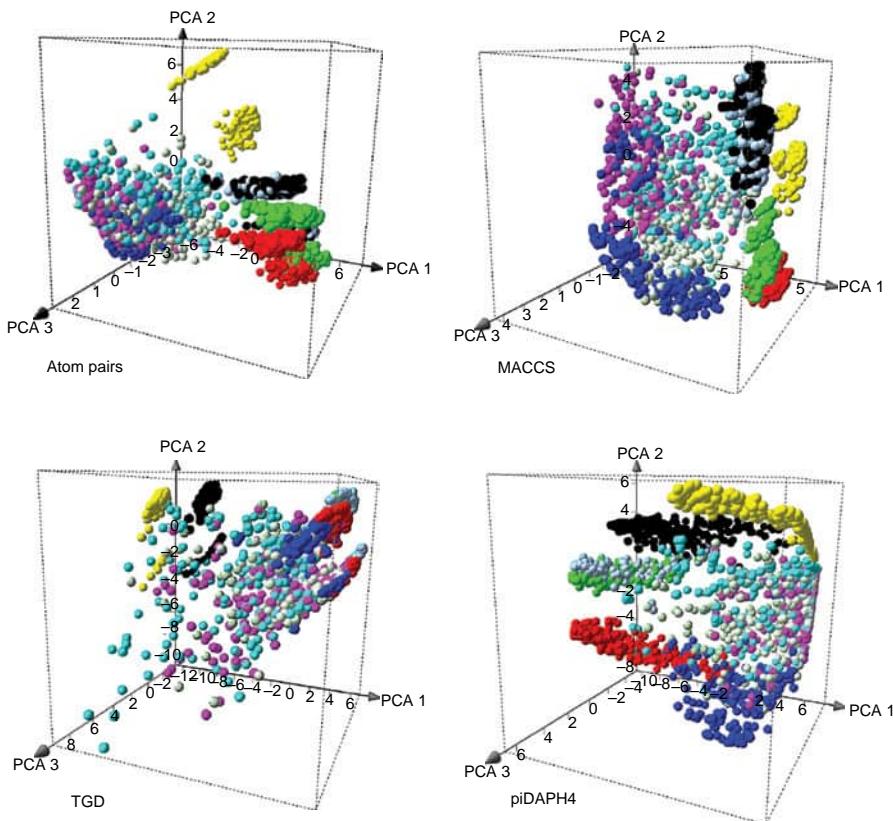


FIGURE 15.7 3D projections of PCA-based chemical spaces generated from a set of 2250 compounds obtained from nine datasets of 250 compounds each using four different molecular fingerprints (Atom pairs, MACCS keys, TGD, and piDAPH4) and the Tanimoto similarity function (see text for further details). For color details, please see color plate section.

molecules each—approved drugs (cyan), natural products (light green), a general screening collection from two compound vendors (magenta), a set of compounds targeted to adenosine receptors (blue), and five in-house combinatorial libraries (red, yellow, green, black, and light blue). Each of the chemical spaces was obtained by principal component analysis (PCA) of the similarity matrix computed using the Tanimoto similarity function and one of four different set-based molecular fingerprints: atom pairs [17], MACCS keys [49], TGD, and piDAPH4 [170]. TGD fingerprints are closely related to atom pair fingerprints, and piDAPH4 are related to 4-point pharmacophore fingerprints. Thus, the latter can contain some structural and stereochemical information. The first three principal components (PCs) account for 80.8% (atom pairs), 85.9% (MACCS), 90.3% (TGD), and 73.0% (piDAPH4) of the variance, respectively. Thus, even though only three PCs are depicted in each of the plots, the percent of the total variance they each contain is significant enough to allow an analysis of the spatial interrelationships of the libraries to each other.

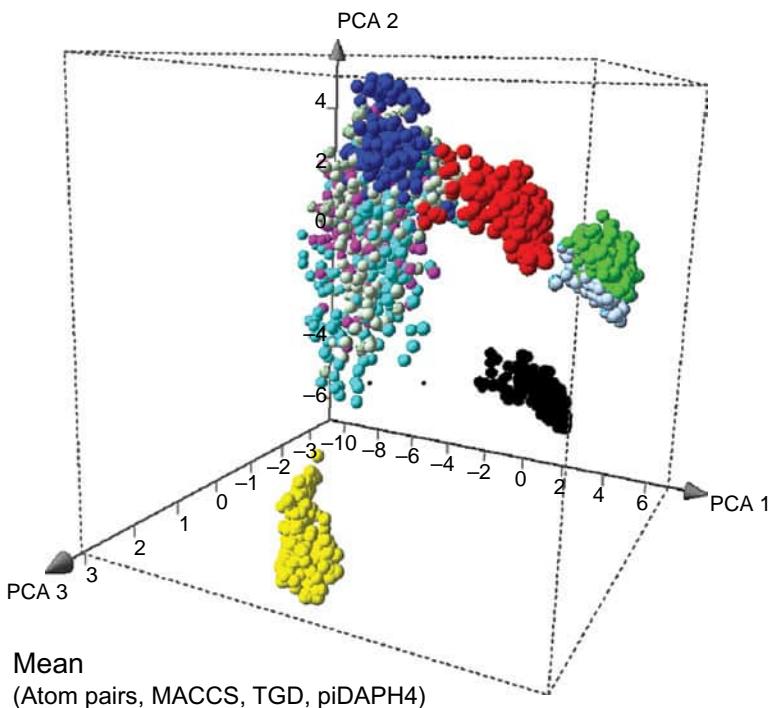


FIGURE 15.8 3D projection of a PCA-based chemical space based on the same set of 2250 compounds in Figure 15.5 and the same four molecular fingerprints (Atom pairs, MACCS keys, TGD, and piDAPH4) but using mean similarity fusion of the individual Tanimoto similarities (see text for further details). For color details, please see color plate section.

From Figure 15.7 it is clear there are substantial changes in the distributions of compounds in each of the libraries. This suggests that even within specific libraries different results would be obtained in similarity searches based on different similarity measures, although, not unexpectedly, this is less pronounced for the five in-house combinatorial libraries. Figure 15.8 depicts a single chemical space for the same set of compounds based on mean fusion of the four different sets of similarity values shown in Figure 15.7. In this case, the three PCs account for about 82% of the total variance. Section 15.5.3 provides a detailed account of fusion methods including “similarity fusion,” the method used in this example. The distribution of compounds in Figure 15.8 differs significantly from those depicted in Figure 15.7, although as was generally true in the latter figure compounds in the in-house libraries which tend to be somewhat more clustered.

All of this suggests (albeit graphically) that compound neighborhoods, except for those associated with several of the in-house libraries, will not remain invariant to changes in similarity measure, a fact that is well recognized within the chemical informatics community. Nevertheless, the graphic portrayal of chemical spaces in Figure 15.7 and Figure 15.8 nicely illustrate this point and suggests that a combination

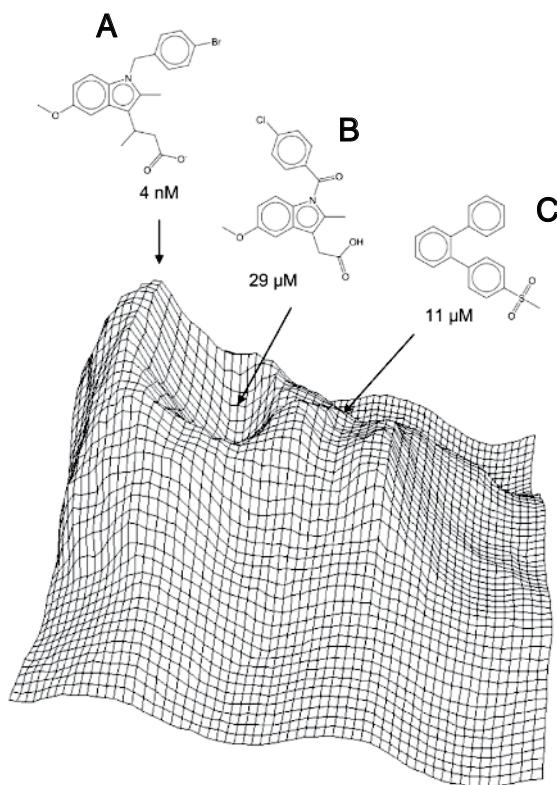


FIGURE 15.9 Activity landscape constructed from a data set of cyclooxygenase-2 inhibitors. The chemical space depicted is the 2D projection of the original, multidimensional space. (Original figure provided courtesy of Prof. Dr. Jürgen Bajorath; see [156]).

of graphical and computation methods may provide some helpful insights on this daunting problem.

15.6.2 Activity Landscapes and Activity Cliffs

Including information on biological activities can augment chemical spaces. Such augmented spaces give rise to *activity landscapes* that exhibit topographical features that resemble geographical landscapes and are closely associated with SARs [153, 156, 171]. Figure 15.9 depicts an example from the work of Bajorath et al. [171] of an activity landscape projected onto 2D chemical space.

An important feature of these landscapes is the presence of *activity cliffs* that arise when large differences in biological activity are accompanied by relatively small structural changes [155], as is nicely illustrated by the three cyclooxygenase-2 inhibitors (A, B, and C) in Figure 15.9. From the figure, it is clear that compounds A and B give rise to a reasonably large activity cliff, while compounds B and C, although

similar in structure, are nearly equipotent. Not surprisingly, activity cliffs violate the Similarity-Property Principle [22]. Such dramatic changes are important because they pinpoint specific structural features associated with corresponding changes in activities that lead to well-characterized SARs. On the other hand, such features can work against the effective use of many quantitative structure–activity relationship (QSAR) methods [155, 172–175].

Because activity landscapes and their associated cliffs are generally of high dimension, a number of attempts have been made to characterize the information inherently contained within them in some low-dimensional form. Alternative methods for representing activity cliffs include the directed-graph representation developed by Guha and Van Drie [152], which is associated with their structure–activity landscape index (SALI), and the network-like similarity graphs (NSGs) developed by Bajorath and coworkers, which is associated with their structure–activity relationship index (SARI) [173, 175]. The former methods provide a local description of activity cliffs in chemical space, while the latter provide a more global view. Thus, combining both approaches provides a fairly comprehensive picture of activity cliffs that can be extended further by considering specificity cliffs [176].

Since similarity measures are not invariant to the representation or similarity function employed, a method that does not use similarity directly could prove useful in determining the location and magnitude of activity cliffs (*vide infra*). Indeed, such a method, which uses the concept of matched molecular pairs (MMPs), exists and has been utilized to identify activity and specificity cliffs [177–180]. One possible limitation of the method is that a numerical measure of the structural difference of a pair of molecules analogous to molecular similarity does not exist. If such was the case, however, the same noninvariance issues that plague molecular similarity measures would limit the utility of the MMP approach. Hence, the lack of such a measure could be counted as a plus. Moreover, the notion of MMPs is much more “chemically intuitive” than many similarity measures and for this reason could be more useful to medicinal chemists in compound optimization.

The concept of activity landscapes can also be extended to *property landscapes* where any set of measurable molecular properties can be added as another dimension to the chemical space of a compound dataset [146]. In line with the concept of activity cliffs, “odor cliffs” [181] and “flavor cliffs” [182] have been recently described.

As will be discussed in more detail in Section 15.6.3, structure–activity similarity (SAS) maps also afford a means not only for characterizing activity cliffs but also the entire activity landscapes within which they are embedded.

15.6.3 Structure–Activity Similarity and Related Maps

Because chemical spaces tend to be multidimensional, they cannot be represented graphically. This has led to the development of simple 2D SAS [183] and related maps [140–142, 184–187]. SAS maps typically represent structural similarity values along the abscissa of the map and activity similarity values, given by

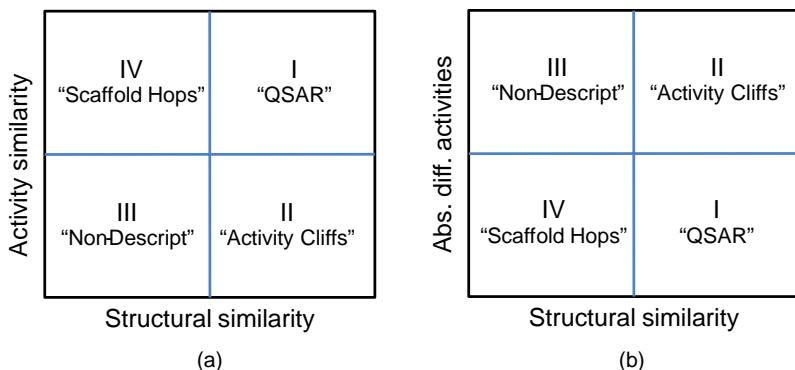


FIGURE 15.10 Schematic diagram of a structure–activity similarity (SAS) map. The ordinate can be described by either activity similarity or the absolute difference of activities (Abs. Diff. Activities) usually given by a negative logarithmic scale (e.g., pK_i or pIC_{50}). SAS maps are roughly divided into four regions I, II, III, IV nominally labeled “QSAR”, “Activity Cliffs”, “Non-Descript”, and “Scaffold Hops” (See text for details).

$$\begin{aligned} S_{\text{Act}}(i, j) &= 1 - \frac{|\text{Act}(i) - \text{Act}(j)|}{\text{Act}_{\max} - \text{Act}_{\min}} \\ &= 1 - \frac{|\Delta\text{Act}(i, j)|}{\Delta\text{Act}_{\max}}. \end{aligned} \quad (15.6.1)$$

along the ordinate. Activity similarity satisfies $0 \leq S_{\text{Act}}(i,j) \leq 1$. $\text{Act}(i)$ and $\text{Act}(j)$ represent the activities of the i th and j th molecules, respectively; Act_{\max} and Act_{\min} represent the maximum and minimum activity values of the entire set of molecules being considered. All of the activity values are given in appropriate units such as pK_i or pIC_{50} . Because $S_{\text{Act}}(i,j)$ maps differences of activity values onto $[0,1]$, this can produce a problem when a wide range of activities are being considered. Thus, in some instances the absolute difference in activities, $|\Delta\text{Act}(i,j)|$, is used directly. In such cases, the ordering along the ordinate is reversed from that when activity similarity is used. Hence, large values of $S_{\text{Act}}(i,j)$ now correspond to small values of $|\Delta\text{Act}(i,j)|$ and vice versa, as illustrated in Figure 15.10, which provides a schematic representation of a prototypical SAS map.

As seen in the figure, SAS maps are roughly divided into four regions. Since each point on a map represents a pairwise comparison, a set of n molecules creates a map containing $n(n-1)/2$ data points. Even a relatively small library of 1000 compounds will generate a map of approximately half a million data points, a number far greater than can be viewed directly on the map. In this regard, it is important to note that datasets this large or greater can nonetheless be analyzed computationally just not visually. Hence, while the size of a compound library or collection can represent a visual impediment, it does not represent a serious impediment in actual applications.

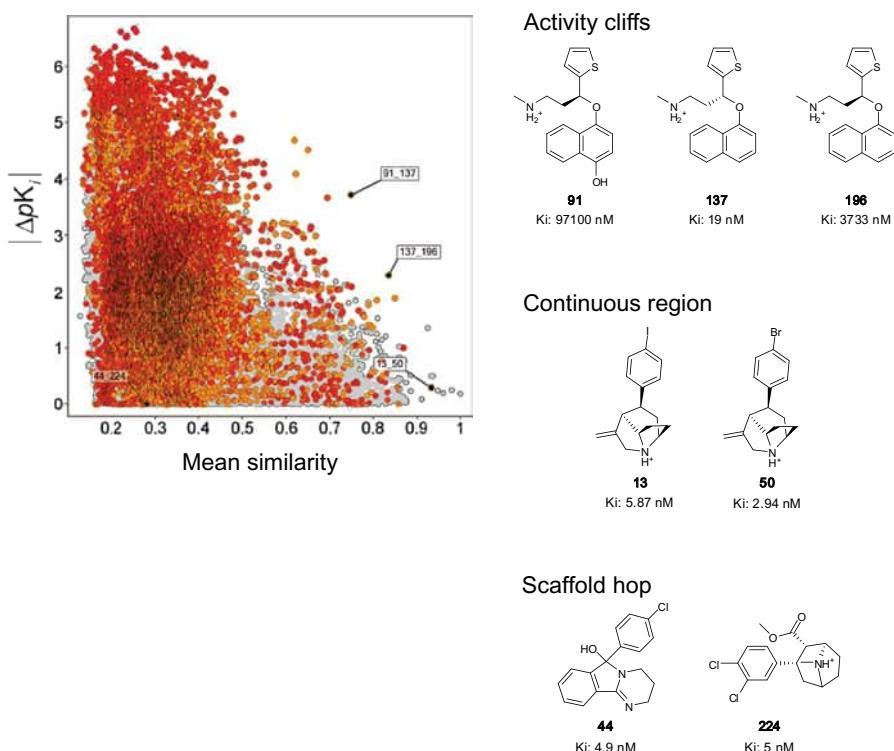


FIGURE 15.11 SAS map of 299 norepinephrine transporter inhibitors (44,551 data points); data points with at least one active compound ($pK_i \geq 7$ nM) in the pair are color-coded by the activity of the most active compound on a continuous scale from orange (least active) to red (most active). The ordinate of the plot gives the absolute potency difference and the abscissa gives the mean similarity obtained by fusing five 2D and 3D molecular fingerprints: radial, atom pairs, MACCS keys, TGD, and piDAPH3. The black dots represent pairs of compounds depicted on the RHS of the figure. For color details, please see color plate section.

As is clear from Figure 15.10, SAS maps are roughly divided into four regions. Points lying within Region I correspond to compound pairs with high structural and activity similarity. Compounds in this region tend to be structurally similar and to exhibit small changes in activity. Thus, compounds in this region are generally appropriate for QSAR studies [41]. Points lying in Region II correspond to compound pairs associated with activity cliffs since compounds in this region are structurally similar but exhibit comparatively large differences in activity. Points lying in Region III are associated with pairs of compounds that show little structural or activity similarity and thus are associated with nondescript SARs. Lastly, points located in Region IV correspond to pairs of compounds that differ significantly in structure but nonetheless possess comparable activities. Such pairs correspond to scaffold hopping [121–123].

Figure 15.11 depicts an SAS map for a set of 299 norepinephrine transporter inhibitors (44,551 data points). In the figure (cf. Figure 15.10b), the ordinate indicates

the absolute potency difference, $|\Delta pK_i|$, and the abscissa indicates the structural similarity computed using the mean fusion of Tanimoto similarity values based on five 2D and 3D molecular fingerprints: radial, atom pairs, MACCS keys, TGD, and piDAPH3 [189] (see also Sections 15.4.2 and 15.4.2.1). Data points with at least one active compound ($pK_i \geq 7$ nM) in the pair are color coded by the activity of the most active compound of the pair based on a continuous scale from orange (least active) to red (most active). The remaining pairs are displayed in light gray. Selected pairs are marked in black and labeled with the compound numbers associated with the molecular structures and K_i 's on the right-hand-side (RHS) of the figure.

Compound **50** located on the RHS of Figure 15.11 is the most active of the seven compounds explicitly depicted in the figure. Its pairing with compound **13** corresponds to the ordered-pair [50, 13],¹³ which lies in the continuous, QSAR region of the map. Its location is supported by the significant structural similarity of the two compounds (> 0.9) and their comparably high activities. Moving up the figure, two points located in the activity cliff region of the map are associated with the ordered-pairs [137, 91] and **(137, 196)**, as is clear from their structural similarities (numerical and graphical) and their disparate activities. Lastly, the point located in the lower left-hand corner of the map, which is associated with the pair **(44, 224)**, corresponds to a scaffold hop since the compounds are clearly dissimilar but have almost identical activities.

SAS maps have also been used to represent consensus models of activity landscapes by means of fusing similarity measures obtained from different 2D and 3D molecule representations and similarity functions [145, 189, 191]. In addition, SAS maps were recently adapted to model multitarget activity landscapes by representing in one axis the activity similarity of compound datasets screened across multiple biological endpoints [192]. SAS maps have been extended to characterize property landscapes other than activity landscapes. For example, structure–flavor similarity maps have been proposed to systematically characterize structure–flavor associations of a comprehensive flavor database [182]. A number of additional types of 2D MFS maps that characterize a different fusion-based similarity on each axis have also been developed and are described in Section 15.5.3.

15.7 FINAL THOUGHTS

Similarity is a fundamental notion that has been used, either implicitly or explicitly, throughout human history. While its exact beginnings are difficult to trace, the Greek philosophers most likely were responsible for early attempts to describe it in more formal terms. It was these same philosophers that also laid down some of the earliest, albeit simplistic, ideas on the nature of matter that formed the basis for the subsequent development many years hence of related theories in physics and chemistry. However, it was not until the mid-nineteenth century that theories describing the atomistic nature of matter began to be developed in earnest. By the early part of the twentieth century many of the concepts of molecular structure and properties were well in hand and many novel applications of molecular similarity were developed. As computers

became more common in the 1970s, the basic components of MSA began to emerge beginning with simple methods for encoding molecular structures in machine-readable form. More sophisticated and comprehensive representations followed, leading to the development of a wide variety computable measures of molecular similarity, many of which remain in use today.

The concept of similarity is one of the most ubiquitous concepts in all of science, although in many cases its use is implicit rather than explicit. It plays an especially significant role in many aspects of chemical informatics because it provides an effective means for dealing with the burgeoning amount chemical and biological information that is connected to the structures of small and macromolecules.

Because of its fundamentally subjective nature, molecular similarity methods provide more powerful means for accessing a broad spectrum of structure and property-based information than more “objective” approaches such as substructure searching. This power, however, comes at a cost because its very subjectivity makes the development and testing of molecular similarity methods problematic at best. In addition, as noted in the quote at the beginning of this chapter, “Similarity like pornography is difficult to define, but you know it when you see it,” but even this phrase does not capture all of the subtleties of similarity be they molecular or otherwise.

Although the subject is vast, a description of several aspects of the theory of human cognition and their relationship to notions of similarity are briefly touched upon in this chapter, raising a number of interesting questions. For example, are similarity and dissimilarity entirely complementary concepts? From a computational point of view they most certainly are. But do humans see similarity and dissimilarity as strict complements of one another? As discussed in this chapter, the answer, in general, is no. As another example, is the similarity of two objects dependent on their size and/or complexity? Again, human perceptions and computable similarity measures are not always consistent with one another. Once a similarity measure of any kind is defined, a process that involves human perception and cognition, its determination for specific objects or their conceptual or graphical representations, involves computation but no further perceptive or cognitive inputs. This is quite different from the way in which humans, who are continually conditioned by their perceptions, experiences, and environments in ways that may or may not influence their assessments of similarity, but are these issues relevant to MSA? The answer is that they are relevant in cases where humans are materially involved in an MSA.

Consider, for example, the process of LBVS where four active compounds have been identified in a preliminary screen. Similarity searches are then typically carried out with respect to each of the reference actives, generating four ranked lists of, say, 500 compounds, where each list is ranked according to the similarities of its compounds with respect to the corresponding reference active. Suppose 20% of the compounds in each list are chosen for subsequent screening with no additional evaluation by chemists or biologists of the “suitability” of the compounds. In such cases, conflicts of human perceptions with computed similarity values do not exist. However, this is no longer the case once humans intervene in the process, especially as human perception of similarity is strongly conditioned by the experiences and training of the

perceiver. For example, one would expect medicinal chemists to have a different view of organic molecules than a physical chemist or a molecular biologist, but perceptions can vary dramatically even among chemists within the same field [48].

What does this all mean? One interpretation suggests that new similarity measures more “compatible” with human perception might be helpful, but how are we to develop them? Since similarity methods provide comparative measures of similarity that are inherently subjective, is it even possible to design new methods that are materially better than those in use today? Regardless of the answers to these and related questions, and despite the fact that the results provided by MSAs are highly subjective in character, it is clear that the concept of molecular similarity will continue to play an important role in many aspects of chemical, biological, and pharmaceutical researches as the material in the following chapters of this book clearly demonstrate.

ACKNOWLEDGMENTS

The authors would like to acknowledge the helpful discussions provided by Prof. Dr. Jürgen Bajorath, Dr. Dimitris Agrafiotis, Dr. Rajarshi Guha, Dr. Karina Martinez-Mayorga, Dr. Fabian López-Vallejo, and Jacob Waddell. J.L.M-F. wishes to acknowledge funds provided by the State of Florida in partial support of this work.

NOTES

1. Fearlessly plagiarized from a quote by Supreme Court Justice Potter Stewart.
2. The notion of a “mental representation” is a theoretical construct of cognitive science. It is a basic concept of the Computational Theory of Mind, according to which cognitive states and processes are constituted by the occurrence, transformation and storage (in the mind/brain) of information-bearing structures (representations) of one kind or another. Paraphrased from the Stanford Encyclopedia of Philosophy <http://plato.stanford.edu/entries/mental-representation/> Accessed June 23, 2011.
3. In the case of hashed fingerprints, the relationship of a given bit position to a specific structural feature is confounded by the fact that hashing mixes the structural features in a way that does not preserve the relationship. Thus, identifying specific structural features is no longer possible with hashed fingerprints (see additional discussion in Section 15.4.2.1).
4. Use of the word pharmacophore is a misnomer in this case, since the fingerprint moieties are not necessarily associated with the biological activities of the molecule—a more appropriate name might be chemophore. Also note that the distances employed in 3D pharmacophore-based fingerprints are not continuous but rather discretized.
5. In some cases, hydrogen atoms are retained in order to distinguish particular functional groups. For example, retaining the hydrogen atom distinguishes the oxygen atoms of hydroxide groups from those of carbonyl groups.
6. $S_{\text{Tan}}(i,j) = S_{\text{Dice}}(i,j)/[2 - S_{\text{Dice}}(i,j)]$.
7. Technically, inner products are only defined with respect to vector spaces. However, the inner products needed here can be defined entirely in terms of their geometric spaces, which can be defined as the summations given in Equations 15.4.22, 15.4.23, and 15.4.24

- for all “vectors” including those in geometric spaces (see discussion on BCUTs given earlier in this Section 15.4.2.3).
8. This process is the continuous analog to that used to determine MCSs in the computation of chemical graph-based similarity measures, as discussed in Section 15.4.5.2.
 9. The Tversky similarity function $S_{\text{Tve}}(i, j | \alpha, \beta)$ is written in a form reminiscent of conditional probabilities to explicitly indicate that the similarity function is conditioned on the value of the parameters α and β .
 10. Swamindass and Baldi [87] have given another interesting form of this expression that facilitates analysis.
 11. See Section 15.6 for a more detailed discussion of activity landscapes and activity cliffs.
 12. The chemical space plots are constructed by carrying out principle component analysis (PCA) on the 2250×2250 -dimensional similarity matrix associated with the complete compound dataset. Although, similarity matrices are not equivalent to the data matrices typically used in PCA, they can be used for this purpose [41]. Each data point (i.e., molecule) is then plotted with respect to the three PCs with the greatest variance.
 13. The first position of the ordered-pair corresponds is taken to be the most active compound of the pair.

REFERENCES

1. Rouvray DH. The evolution of the concept of molecular similarity. In: Johnson MA, Maggiora GM, editors. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 1990.
2. PubChem: <http://pubchem.ncbi.nlm.nih.gov/>. Accessed 2013 Jan 6.
3. ZINC – A free database for virtual screening: <http://zinc.docking.org/>. Accessed 2013 Jan 6.
4. Scior JT, Bernard P, Medina-Franco JL, et al. Large compound databases for structure-activity relationships studies in drug discovery, *Mini-Rev Med Chem* 2007; 7:851–860.
5. Holliday JD, Salim N, Whittle M, et al. Analysis and display of the size dependence of chemical similarity coefficients. *J Chem Inf Comput Sci* 2003;43:819–828.
6. Partington JR. The beginnings of chemistry. In: *A Short History of Chemistry*. London: Macmillan; 1937.
7. Rouvray DH. The changing role of the symbol in the evolution of chemical notation. *Endeavor* 1977;1:23–31.
8. Wikipedia (2011) Discussion of the life of John Dalton. http://en.wikipedia.org/wiki/John_Dalton. Accessed 2013 Jan 6.
9. Grossman RB. Van’t Hoff, Le Bel, and the development of stereochemistry: A reassessment. *J Chem Educ* 1989;66:30–33.
10. Mendeleev D. On the relationship of the properties of the elements to their atomic weights. *J Russ Phys Chem Soc* 1869;1:60.
11. Meyer L. The nature of the chemical elements as a function of their atomic weights. *Ann Suppl* 1870;7:354.
12. Kleppner D, Jackiw R. One hundred years of quantum physics. http://www.4physics.com/phy_demo/QM_Article/article.html. Accessed 2013 Jan 6.
13. Adamson GW, Bush JA. A method for the automatic classification of chemical structures. *Inf Stor Retrieval* 1973;9:561–568.

14. Adamson GW, Bush JA. A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J Chem Inf Comput Sci* 1975;15:55–58.
15. Willett P. A fast procedure for the calculation of similarity coefficients in automatic classification. *Inf Process Manage* 1981;17:53–60.
16. Harding AF, Willett P. Indexing exhaustivity and the computation of similarity matrices. *J Am Soc Inf Sci* 1980;31:298–300.
17. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure activity studies – Definition and applications. *J Chem Inf Comput Sci* 1985;25:64–73.
18. Willett P, Winterman V, Bawden D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J Chem Inf Comput Sci* 1986;26:36–41.
19. Willett P. *Similarity and Clustering in Chemical Information Systems*. Letchworth: Research Studies Press; 1987.
20. Willett P. Algorithms for the calculation of similarity in chemical structure databases. In: Johnson MA, Maggiora GM, editors. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 1990, Chapter 3.
21. Bawden D. Application of two-dimensional chemical similarity measures to database analysis and querying. In: Johnson MA, Maggiora GM, editors. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 1990, Chapter 4.
22. Johnson MA, Maggiora GM, editors. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons; 1990.
23. Dean PM, editor. *Molecular Similarity in Drug Design*. Glasgow: Chapman & Hall; 1994.
24. Maggiora GM, Shanmugasundaram V, Lajiness MS, et al. A practical strategy for directed compound acquisition. In: Oprea TI editor. *Chemoinformatics in Drug Discovery*. Weinheim: Wiley-VCH Verlag GmbH & Co KGaA; 2004. p 317–332.
25. Willett P. Molecular diversity techniques for chemical databases. *Information Research* 2. <http://informationr.net/ir/2-3/paper19.html>. Accessed 2013 Jan 6.
26. Agraftiotis DK. Stochastic algorithms for maximizing molecular diversity. *J Chem Inf Comput Sci* 1997;37:841–851.
27. Agraftiotis DK. Diversity of chemical libraries. In: von Schleyer PR, Allinger NL, Clark T, et al., editors. *The Encyclopedia of Computational Chemistry*, Vol. 1. Chichester: John Wiley and Sons; 1998. p 742–761.
28. Clemons PA, Bodycombe NE, Carrinski HA, et al. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc Natl Acad Sci USA* 2010;107:18787–18792.
29. Ghose A, Viswanadhan V. *Combinatorial Library Design and Evaluation: Principles, Software, Tools, and Applications in Drug Discovery*. New York: Marcel Dekker Inc.; 2001.
30. Lajiness MS. Dissimilarity-based compound selection techniques. *Perspect Drug Discov Des* 1997;7–8:65–84.
31. Gillet VJ. Diversity selection algorithms. *Wiley Interdiscip Rev Comput Mol Sci* 2011;1:580–589.
32. Oprea TI, Matter H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* 2004;8:349–358.
33. Jain AN. Ligand-based structural hypotheses for virtual screening. *J Med Chem* 2004;47:947–961.

34. Jahn A, Hinselman G, Fechner N, et al. Optimal assignment methods for ligand-based virtual screening. *J Cheminform.* <http://www.jcheminf.com/content/1/1/14>. Accessed 2013 Jan 6.
35. Rathke F, Hansen K, Brefeld U, et al. StructRank: A new approach for ligand-based virtual screening. *J Chem Inf Model* 2011;51:83–92.
36. Swann SL, Brown SP, Muchmore SW, et al. A unified, probabilistic framework for structure- and ligand-based virtual screening. *J Med Chem* 2011;54:1223–1232.
37. Baringhaus K-H, Hessler G. Fast similarity searching and screening hit analysis. *Drug Discov Today Technol* 2004;1:197–202.
38. Rippenhausen P, Nisius B, Peltason L, et al. Quo Vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* 2010;53:8461–8467.
39. Schneider G. Virtual screening: An endless staircase? *Nat Rev. Drug Discov* 2010;9:273–276.
40. Cheng C, Maggiora GM, Lajiness MS, et al. Four association coefficients for relating molecular similarity measures. *J Chem Inf Comput Sci* 1996;36:909–915.
41. Maggiora GM, Shanmugasundaram V. Molecular similarity measures. In: Bajorath J, editor. *Chemoinformatics and Computational Chemical Biology*. New York: Humana Press/Springer; 2011. p 39–100.
42. Vosniadou S, Ortony A, editors. *Similarity and Analogical Reasoning*. New York: Cambridge University Press; 1989.
43. Sloman SA, Rips LJ. Similarity as an explanatory construct. In: Sloman SA, Rips LJ, editors. *Similarity and Symbols in Human Thinking*. Cambridge: The MIT Press; 1998, Chapter 1.
44. Shepard RN. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika* 1962;27:25–140.
45. Tversky A. Features of similarity. *Psychol Rev* 1977;84:327–352.
46. Mestres J, Maggiora GM. Putting molecular similarity into context: Asymmetric indices for field-based similarity measures. *J Math Chem* 2005;39:107–118.
47. Gentner D, Markman AB. Structural alignment in analogy and similarity. *Am Psychol* 1997;52:45–560.
48. Lajiness MS, Maggiora GM, Shanmugasundaram V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J Med Chem* 2004;47:4891–4896.
49. Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.
50. Klir GJ, Yuan B. *Fuzzy Sets and Fuzzy Logic – Theory and Applications*. Upper Saddle River: Prentice Hall; 1995.
51. Miyamoto S. *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht: Kluwer Academic Publishers; 1990.
52. Bender A, Glen RC. Molecular similarity: A key technique in molecular informatics. *Org Biomol Chem* 2004;2:3204–3218.
53. Leach AR, Gillet VJ. *An Introduction to Chemoinformatics*. Dordrecht: Kluwer Academic Publishers; 2003.
54. Willett P. Similarity searching using 2D structural fingerprints. In: Bajorath J, editor. *Chemoinformatics and Computational Chemical Biology*. New York: Springer Science+Business Media, New York; 2011.
55. Johnson MA. A review and examination of the mathematical spaces underlying molecular similarity analysis. *J Math Chem* 1989;3:117–145.

56. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics*. 2nd ed. New York: Wiley-VCH; 2009.
57. Bender A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin Drug Discov* 2010;5:1141–1151.
58. Edgar SJ, Holliday JD, Willett P. Effectiveness of retrieval in similarity searches of chemical databases: A review of performance measures. *J Mol Graphics Model* 2000;18:343–357.
59. Chen B, Mueller C, Willett P. Combination rules for group fusion in similarity-based virtual screening. *Mol Inf* 2010;29:533–541.
60. López-Vallejo F, Nefzi A, Bender A, et al. Increased diversity of libraries from libraries: Chemoinformatic analysis of bis-diazacyclic libraries. *Chem Biol Drug Des* 2011;77:328–342.
61. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci* 1998;38:983–996.
62. Nilakantan R, Bauman N, Dixon JS, et al. Topological torsions: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J Chem Inf Comput Sci* 1987;27:82–85.
63. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;50: 742–754.
64. Bender A, Mussa HY, Glen RC. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J Chem Inf Comput Sci* 2004;44:1708–1718.
65. Sastry M, Lowrie JF, Dixon SL, et al. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J Chem Inf Model* 2010;50:771–784.
66. Pickett SD, Mason JS, McLay IM. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J Chem Inf Comput Sci* 1996;36: 1214–1223.
67. Mason JS, Morize I, Menard PR, et al. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J Med Chem* 1999;42:3251–3264.
68. Kanera P. *Sparse Distributed Memory*. Cambridge: MIT Press; 1990. p 26–27.
69. Miyamoto S. Two generalizations of multisets. In: Inuiguchi M, Hirano S, Tsumoto S, editors. *Rough Set Theory and Granular Computing*. New York: Springer; 2003. p 59–68.
70. Xue L, Godden JW, Bajorath J. Database searching for compounds with similar biological activity using short bit string representations of molecules. *J Chem Inf Comput Sci* 1999;39:881–886.
71. Trinajstic N. *Chemical Graph Theory*. 2nd ed. Boca Raton: CRC Press; 1992.
72. Raymond JW, Willett P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des* 2002;16:521–533.
73. Kvasnicka V, Posspichal J. Chemical and reaction metrics for graph-theoretical model of organic chemistry. *J Mol Struct Theochem* 1991;227:17–42.
74. Pearlman RS, Smith KM. Novel software tools for chemical diversity. *Perspect Drug Discov Des* 1998;9–11:339–353.
75. Stanton DT. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J Chem Inf Model* 1999;39:11–20.

76. Gower JC. Measures of similarity, dissimilarity, and distance. In: Kotz S, Johnson NL, Read CB, editors. *Encyclopedia of Statistics*. New York: Wiley; 1982. p 397–405.
77. Good AC, Richards WG. Explicit calculation of 3D molecular similarity. Perspect Drug Discov Des 1998;9–11:321–338.
78. Mestres J, Rohrer DC, Maggiore GM. MIMIC: A molecular field matching program. Exploiting applicability of molecular similarity approaches. J Comput Chem 1997;18:934–954.
79. Grant JA, Gallardo GA, Pickup JT. A fast method of molecular shape descriptors. A simple application of a Gaussian description of molecular shape. J Comput Chem 1996;17: 1653–1666.
80. Kearsley SK, Smith GM. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. Tetrahedron Comput Methods 1990;3:615–633.
81. Labute P, Williams C, Feher M, et al. Flexible alignment of small molecules. J Med Chem 2001;44:1483–1490.
82. Willett P, Winterman V. A comparison of some measurements for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. Quant Struct Act Relat 1986;5:18–25.
83. Arif SM, Holiday JD, Willett P. Analysis and use of fragment-occurrence data in similarity-based virtual screening. J Comput Aided Mol Des 2009;23:6655–668.
84. Wang Y, Bajorath J. Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching. Chem Inf Model 2009;49:1369–1376.
85. Heritage TW, Lewis DR. Molecular hologram QSAR. Rational Drug Design, ACS Symposium Series, Vol. 719. Washington, DC; 1999. p 212–225.
86. Fechner U, Paetz J, Schneider G. Comparison of three holographic fingerprint descriptors and their binary counterparts. QSAR Comb Sci 2005;24:961–967.
87. Swamidass SJ, Baldi P. Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. J Chem Inf Model 2007;47:302–317.
88. Chen X, Reynolds CH. Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients. J Chem Inf Comput Sci 2002;42:1407–1414.
89. Kosko B. *Neural Networks and Fuzzy Systems*, Englewood Cliffs: Prentice-Hall, Inc.; 1992.
90. Beliakov G, Pradera A, Calvo T. *Aggregation Functions: A Guide for Practitioners*. New York: Springer; 2010.
91. Maggiore GM, Petke JD, Mestres J. A general analysis of field-based molecular similarity indices. J Math Chem 2002;31:251–270.
92. Dixon SL, Koehler RT. The hidden component of size in two-dimensional fragment descriptors: Side effects of sampling in bioactive libraries. J Med Chem 1999;42:2887–2900.
93. Godden JW, Xue L, Bajorath J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J Chem Inf Comput Sci 2000;40:163–166.
94. Fligner MA, Verducci JS, Blower PE. A modification of the Jacard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. Technometrics 2002;44:110–119.

95. Flower DR. On the properties of bit string based measures of chemical similarity. *J Chem Inf Comput Sci* 1988;38:379–386.
96. Wang Y, Eckert H, Bajorath J. Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem* 2007;2:1–7.
97. Wang Y, Bajorath J. Balancing the influence of molecular complexity on fingerprint similarity searching. *Chem Inf Model* 2008;48:75–84.
98. McGregor J, Willett P. Use of a maximum common subgraph algorithm in the automatic identification of the ostensible bond changes occurring in chemical reactions. *J Chem Inf Comput Sci* 1998;21:137–140.
99. Maggiora GM, Petke JD, Mestres J. A general analysis of field-based molecular similarity indices. *J Math Chem* 2002;31:251–270.
100. Ballester PJ, Richards WG. Ultrafast shape recognition for similarity search in molecular databases. *Proc R Soc A* 2007;463:1307–1321.
101. Mestres J, Rohrer DC, Maggiora GM. A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. *J Comput Aided Mol Des* 1999; 13:79–93.
102. Blinn JR, Rohrer DC, Maggiora GM. Field-based similarity forcing in energy minimization and molecular matching. In Altman RB, et al., editors. Pacific Symposium on Biocomputing '99. Singapore: World Scientific; 1998. p 415–424.
103. Hagadone TR. Molecular substructure similarity searching; Efficient retrieval in two-dimensional structure databases. *J Chem Inf Comput Sci* 1992;32:515–521.
104. Bradshaw J. Introduction to Tversky similarity measure. Eleventh Daylight User Group Meeting, 1997. <http://www.daylight.com>. Accessed 2013 Jan 6 (To find the link choose “Archives” in the box at the upper right corner of the webpage, and click on MU'97: Tversky Similarity, John Bradshaw on page 15).
105. Maggiora GM, Mestres J, Hagadone TR, et al. Asymmetric similarity and molecular diversity. In: *213th ACS national meeting*; 1997 Apr 13–17; San Francisco.
106. Blankley CJ, Wild DJ. Asymmetric similarity in action. In: *221st ACS National Meeting*; 2001 Apr 1–5; San Diego.
107. MacCuish JD, MacCuish NE. Asymmetric clustering of chemical datasets: An investigation. In: *224th ACS national meeting*; 2002 Aug 18–22; Boston.
108. Chen X, Brown FK. Asymmetry of chemical similarity. *ChemMedChem* 2007;2: 180–182.
109. Senger S. Using Tversky similarity searches for core hopping: Finding the needles in the haystack. *J Chem Inf Model* 2009;49:1514–1524.
110. Lemmen C, Lengauer T. Computational methods for the structural alignment of molecules. *J Comput Aided Mol Des* 2000;14:215–232.
111. Yongye AB, Bender A, Martinez-Mayorga K. Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble. *J Comput Aided Mol Des* 2010;24:675–686.
112. Nettles JH, Jenkins JL, Bender A, et al. Bridging chemical and biological space: “Target Fishing” using 2D and 3D molecular descriptors. *J Med Chem* 2006;49:6802–6810.
113. Oellien F, Cramer J, Beyer C, et al. The impact of tautomer forms on pharmacophore-based virtual screening. *J Chem Inf Model* 2006;46:2342–2354.

114. Martin YC. *Quantitative Drug Design: A Critical Introduction*. 2nd ed. Boca Raton: CRC Press; 2010.
115. Brown RD, Martin YC. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J Chem Inf Model* 1997;37:1–9.
116. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? *Drug Discov Today* 2002;7:903–911.
117. McGaughey GB, Sheridan RP, Bayly CI, et al. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* 2007;47:1504–1519.
118. Ebalunode JO, Zheng W. Unconventional 2D shape similarity method affords comparable enrichment as a 3D shape method in virtual screening experiments. *J Chem Inf Model* 2009;49:1313–1320.
119. Matter H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem* 1997;40:1219–1229.
120. Schneider G, Neidhart W, Giller T, et al. Scaffold-hopping by topological pharmacophore search: A contribution to virtual screening. *Angew Chem Int Ed Engl* 1999;38:2894–2896.
121. Renner S, Schneider G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* 2006;1:181–185.
122. Brown N, Jacoby E. On scaffolds and hopping in medicinal chemistry. *MinRev Med Chem* 2006;6:1217–1229.
123. Zhang Q, Muegge I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *J Med Chem* 2006;49: 1536–1548.
124. Bender A, Jenkins JL, Scheiber J, et al. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J Chem Inf Model* 2009; 49:108–119.
125. Ginn CMR, Willett P, Bradshaw J. Combination of molecular similarity measures using data fusion. *Perspect Drug Discov Des* 2000;20:1–16.
126. Salim N, Holliday J, Willett P. Combination of fingerprint-based similarity coefficients using data fusion. *J Chem Inf Comput Sci* 2003;43:435–442.
127. Willett P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb Sci* 2006;25:1143–1152.
128. Feher M. Consensus scoring for protein-ligand interactions. *Drug Discov Today* 2006; 11:421–428.
129. Holliday JD, Kanoulas E, Malim N, et al. Multiple search methods for similarity-based virtual screening: An analysis of search overlap and precision. *J Cheminform* 2011; 3:29–43.
130. Fechner U, Schneider G. Evaluation of distance metrics for ligand-based similarity searching. *ChemBioChem* 2004;5:538–540.
131. Shanmugasundaram V, Maggiore GM, Lajiness MS. Hit-directed nearest neighbor searching. *J Med Chem* 2005;48:240–248.
132. Hall DL, McMullen SAH. *Mathematical Techniques in Multisensory Data Fusion*. 2nd ed. Boston: Artech House; 2004.
133. Klein LA. *Sensor and Data Fusion*. Bellingham: SPIE Press; 2004.

134. Kearsley SK, Sallamack S, Fluder EM, et al. Chemical similarity using physiochemical property descriptors. *J Chem Inf Comput Sci* 1996;36:118–127.
135. Ginn CMR, Turner DB, Willett P, et al. Similarity searching in files of three-dimensional chemical structures: Evaluation of the eva descriptor and combination of rankings using data fusion. *J Chem Inf Comput Sci* 1997;37:23–37.
136. Torra V, Narukawa Y. *Modeling Decisions—Information Fusion and Aggregation Operators*. New York: Springer; 2007.
137. Chen J, Holliday J, Bradshaw J. A machine learning approach to weighting schemes in the data fusion of similarity coefficients. *J Chem Inf Model* 2009;49:185–194.
138. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* 2006;11:1046–1053.
139. Medina-Franco JL, Maggiora GM, Julianotti MA, et al. A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem Biol Drug Des* 2007;70:393–412.
140. Martínez-Mayorga K, Medina-Franco JL, Julianotti MA, et al. Conformation-opioid activity relationships of bicyclic guanidines from 3D similarity analysis. *Bioorg Med Chem* 2008;16:5932–5938.
141. Gerard B, Duval JR, Lowe JT, et al. Synthesis of a stereochemically diverse library of medium-sized lactams and sultams via S(N)Ar cycloetherification. *ACS Comb Sci* 2011;13:365–374.
142. Singh N, Guha R, Julianotti MA, et al. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model* 2009;49:1010–1024.
143. Medina-Franco JL, Martínez-Mayorga K, Bender A, et al. Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *J Chem Inf Model* 2009;49:477–491.
144. Pérez-Villanueva J, Santos R, Hernández-Campos A, et al. Towards a systematic characterization of the antiprotozoal activity landscape of benzimidazole derivatives. *Bioorg Med Chem* 2010;18:7380–7391.
145. Yongye A, Byler K, Santos R, et al. Consensus models of activity landscapes with multiple chemical, conformer and property representations. *J Chem Inf Model* 2011;51:1259–1270.
146. Critchlow DE. *Metric Methods for Analyzing Partially-Ranked Data*. Berlin: Springer-Verlag; 1985.
147. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967;27:209–220.
148. Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd ed. London: Chapman-Hall; 1997.
149. Legendre P, Legendre L. *Numerical Ecology*. 2nd ed. Amsterdam: Elsevier; 1998.
150. Willett P. Evaluation of molecular similarity and molecular diversity using biological activity data. In: Bajorath J, editor. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Totowa: Humana Press; 2004. p 51–63.
151. Truchon J-F, Bayly CI. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *J Chem Inf Model* 2007;47:488–508.
152. Guha R, Van Drie JH. Structure-activity landscape index: Identifying and quantifying activity cliffs. *J Chem Inf Model* 2008;48:646–658.

153. Guha R. The ups and downs of structure-activity landscapes. In: Bajorath J, editor. *Chemoinformatics and Computational Chemical Biology*. New York: Humana Press/Springer; 2011. p 101–117.
154. Peltason L, Bajorath J. Computational aspects of activity and selectivity cliffs. In: Bajorath J, editor. *Chemoinformatics and Computational Chemical Biology*. New York: Humana Press/Springer; 2011. p 119–132.
155. Maggiora GM. On outliers and activity cliffs—Why QSAR often disappoints (Editorial). *J Chem Inf Model* 2006;46:1535.
156. Bajorath J, Peltason L, Wawer M, et al. Navigating structure-activity landscapes. *Drug Discov Today* 2009;14:698–705.
157. Stahura FL, Bajorath J. Bio- and chemo-informatics beyond data management: Crucial challenges and future opportunities. *Drug Discov Today* 2002;7:S41–S47.
158. Raghavendra AS, Maggiora GM. Molecular basis sets—A general similarity-based approach for representing chemical spaces. *J Chem Inf Model* 2007;47:1328–1340.
159. Rupp M, Schneider P, Schneider G. Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. *J Comput Chem* 2009;30:2285–2296.
160. Agrafiotis DK, Xu H. A self-organizing principle for learning non-linear manifolds. *Proc Natl Acad Sci USA* 2002;99:15869–15872.
161. Agrafiotis DK. Stochastic proximity embedding. *J Comput Chem* 2003;24:1215–1221.
162. Agrafiotis DK, Xu H. A geodesic framework for analyzing molecular similarities. *J Chem Inf Comput Sci* 2003;43:475–484.
163. Araujo RP, Liotta LA, Petricoin EF. Proteins, drug targets and the mechanisms they control: The simple truth about complex networks. *Nat Rev Drug Discov* 2007;6:871–880.
164. Yildirim MA, Goh K-I, Cusick ME, et al. Drug-target network. *Nat Biotech* 2007;25:1119–1126.
165. Paolini GV, Shapland RHB, van Hoorn WP, et al. Global mapping of pharmacological space. *Nat Biotechnol* 2006;24:805–815.
166. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;25:197–206.
167. Tominaga Y. Data structure comparison using box counting analysis. *J Chem Inf Comput Sci* 1998;38:867–875.
168. Agrafiotis DK, Rassokhin DN. A fractal approach for selecting an appropriate bin size for cell-based diversity estimation. *J Chem Inf Comput Sci* 2002;42:117–122.
169. Cohen R, Havlin S. Scaling properties of complex networks and spanning trees. In: Bollobas B, Kozma R, Miklos D, editors. *Handbook of Large-Scale Random Networks*. New York: Springer; 2009. p 143–169.
170. MOE (Molecular Operating Environment), Chemical Computing Group, 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7 (<http://www.chemcomp.com>). Accessed 2011 Dec 24).
171. Wassermann AM, Wawer M, Bajorath J. Activity landscape representations for structure-activity relationship analysis. *J Med Chem* 2010;53:8209–8223.

172. Doweyko AM. QSAR: Dead or alive? *J Comput Aided Mol Des* 2008;22:81–89.
173. Johnson S. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J Chem Inf Model* 2008;48:25–26.
174. Guha R, Van Drie JH. Assessing how well a modeling protocol captures a structure-activity landscape. *J Chem Inf Model* 2008;48:1716–1728.
175. Peltason L, Bajorath J. SAR index: Quantifying the nature of structure-activity relationships. *J Med Chem* 2007;50:5571–5578.
176. Peltason L, Hu Y, Bajorath J. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* 2009;4:1864–1873.
177. Kenny PW, Sadowski J. Structure modification in chemical databases. In: Oprea TI, editor. *Chemoinformatics in Drug Discovery*. Weinheim: Wiley-VCH; 2004. p 271–285.
178. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large datasets. *J Chem Inf Model* 2010;50:339–348.
179. Hu Y, Bajorath J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J Chem Inf Model* 2010;50:500–510.
180. Wassermann AM, Bajorath J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J Chem Inf Model* 2010;50:1248–1256.
181. Martínez-Mayorga K, Medina-Franco JL. Chemoinformatics – Applications in food chemistry. In: Taylor S, editor. *Advances in Food and Nutrition Research*, Vol. 58. Burlington: Academic Press; 2009. p 33–56.
182. Martínez-Mayorga K, Peppard TL, Yongye AB, et al. Characterization of a comprehensive flavor database. *J Chemomet* 2011;25:550–560.
183. Shanmugasundaram V, Maggiora GM. Characterizing activity landscapes using an information-theoretic approach. In: *222nd ACS National Meeting*; 2001 Aug 26–30; Chicago.
184. Medina-Franco JL, Yongye AB, Pérez-Villanueva J, et al. Multitarget structure-activity relationships characterized by activity-difference maps and consensus similarity measure. *J Chem Inf Model* 2011;51:2427–2439.
185. Patterson DE, Cramer RD, Ferguson AM, et al. Neighborhood behavior: A useful concept for validation of “molecular diversity” descriptors. *J Med Chem* 1996;39: 3049–3059.
186. Pérez-Villanueva J, Santos R, Hernández-Campos A, et al. Structure-activity relationships of benzimidazole derivatives as antiparasitic agents: Dual activity-difference (DAD) maps. *Med Chem Commun* 2011;2:44–49.
187. Waddell J, Medina-Franco JL. Bioactivity landscape modeling: Chemoinformatic characterization of structure-activity relationships of compounds tested across multiple targets. *Bioorg Med Chem* 2012;20:5443–5452.

INDEX

- Ab initio* predictions, lead optimization, overview, 149–154
- Absorption, distribution, metabolism and excretion (ADME):
- Chemical Property Structure Planning System (BICEPS), 295–297
 - chemical array analysis, 179–181
 - global model development and applications:
 - ligand-based models, 247–248
 - metabolic lability, 248–262
 - data set characteristics, 251
 - descriptor sets, 251–256
 - DGAT1 inhibitor optimization, 256–262
 - machine learning and molecular descriptors, 249–251
 - structure-based models, 246–247
 - naïve Bayesian models, target identification, 137
- Accuracy of classification, rough set theory, 56–61
- Acetylcholinesterase (ACT), virtual screening, 122–123
- Activity cliffs/landscapes:
- case study, 308–309
 - chemical arrays, 185–186
 - lead optimization, 156–157
 - molecular similarity applications, 383–384
 - structure–activity relationships:
 - antitarget activity hotspots, 228–230
 - classical chemical studies, 214–216
 - outlier investigations, 209–210
 - pharmaceutical applications, 223
 - quantification, 210–211
- Activity prediction models:
- naïve Bayesian models, phenotype/target comparison, 141–143
- three-dimensional QSAR, 45–47
- ActWiki, chemical arrays, 202
- Adenosine di/triphosphate (ADP/ATP), ecto-5'-nucleotidase virtual screening, 125–128
- ADME assays, regression models, 6–7
- Adverse drug reactions (ADRs):
- naïve Bayesian models, 137
 - predictive secondary pharmacology, 274–276
- Alignment rules:
- molecular similarity, 348–350
 - three-dimensional-ligand-based models, 225–227
- Aliphatic side chain substitutions, Topliss Tree, 150–152
- Ames QSAR model, toxicity warning systems, 276–278
- Aminergic GPCR inhibitory activity, partial least squares analysis, 86–87
- Aminomethylcyclohexane (AMC), volumetric/surface-density functions, molecular similarity models, 41–47
- 2-Aminothiazoles, metabolic lability model, DGAT1 inhibitors, 256–262
- Annotation:
- chemical arrays, 194–195
 - controlled substances, 314–315
- Antitarget activity:
- in silico* screening, ligand-based ADMET models, 247–248
 - structure–activity relationships, 227–232
 - activity hotspot identification and application, 228–230
 - hERG and CYP3A4 inhibition, 230–231
 - lead optimization integration, 232
 - transfer mechanisms, 227–228

- Apostle approach:
 chemoinformatics database rollout, 338
 Mobius CIDB system, 329
- Approximation set, rough set theory, 55–56
- AQUASOL database, predictive modeling, 3–6
- Aqueous solubility, predictive model comparisons, 16–17
- Area under the curve (AUC) metrics, fingerprint method comparisons, 104–110
- Area under the receiver operating characteristic curve (ROCAUC), naïve Bayesian models:
 chemist's preferences, 144–145
 potency binning, 140–141
- Aromatic ring substituents, Craig plot, 150–151
- Association rules, rough set theory, 63–64
- Asymmetric similarity:
 mathematical description, 350
 molecular similarity analysis, 366–370
 two-dimensional searching, 368–370
- Atomic colorings:
 L-shaped partial least squares analysis:
 molecular modeling, 91–93
 regression coefficient matrix, 90–91
- AurSCOPE database, antitarget activity analysis, 229–230
- Automated annotation workflow, controlled substances, 314–315
- Automatic chemotype detector, chemical arrays, 195–198
- AZOrange software, safety modeling, 283–285
- Bayesian analysis:
 L-shaped partial least squares analysis, atomic colorings, regression coefficient matrix, 90–91
 toxicity studies, 279
- Bayesian weights, naïve Bayesian models, class comparisons, 143
- BCUT descriptors, molecular similarity, 356
- Beacon Projects, 328
- Benchmark studies, virtual screening techniques, 115–116
- Benzamidine, volumetric/surface-density functions, molecular similarity models, 41–47
- Binary fingerprints, molecular similarity, set-based representation, 353
- Bioisosterism:
 molecular fingerprinting techniques, 99–100
 rescuffolding, 223–227
- Biological activity:
 naïve Bayesian models, 139–140
 toxicity studies, 279–283
- Biological response surface, activity cliffs, outlier investigations, 210
- BioProfile system, 302–307
- BioSim system, toxicity analysis, 279–283
- Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules (BICLAIM) database:
 development of, 293
 search strategies, 309–314
- Boehringer Ingelheim Mining and Exploration of Screening Hit (BIMESH), 299–301
- Boehringer Ingelheim Split Substructure Search (BISCUBE), BICLAIM-space searches, 311–314
- Boehringer Ingelheim Chemical Property Structure Planning System (BICEPS), 295–297
- BIMESH HTS data analysis, 300–301
- BioProfile system, 302–307
- Chemoinformatics Database (CIDB), 293–294
- controlled substances, automatic annotation, 314–315
- Database of Virtual Combinatorial Libraries, 297–299
 search applications, 309–314
- Boltzmann-Enhanced Discrimination of ROC (BEDROC) metric, molecular similarity validation, 377–378
- Box counting procedures, molecular similarity, chemical space representation, 380–383
- Boxplots, datasets, 4–6
- Build vs. buy, chemoinformatics database systems, 339–340
- Bump hunting methods, multiparameter lead optimization, 168–169
- Candidate marker genes, drug-induced gene expression, phospholipidosis:
 identification, 66–68
 ranking, 75–77
- Cardiac liability risk reduction, QSAR safety filter, 271–272
- Cardinality, set-based similarity functions, molecular similarity, 357–359
- Cavity-based alignment, aminergic G protein-coupled receptors, 87–88
- Cell-based chemical space representation, molecular similarity, 378–383
- Central compound database (CDB) system, 292–293
- ChEMBL target families, naïve Bayesian models, 140–141
- Chemical arrays, lead optimization:
 annotation, 194–195

- archive information extraction, 194–201
- automatic chemotype detector, 195–198
- chemistry and property space coverage, 182–186
- data analysis techniques, 192–194
- seed compound detection, 196, 198–201
- self-avoiding random walk, 191–192
- temporal analysis, 186–191
- Chemical graph-based representations, molecular similarity, 353–354
- applications, 378–383
- similarity functions, 361–363
- Chemically advanced template search (CATS): metabolic lability model, 250–251
- rescaffolding, 224–225
- Chemical similarity: metabolic lability model, 253–256
- molecular fingerprinting techniques, 98–100
- Chemical space representation: activity cliffs, heatmaps, 215–216
- machine learning, 138
- molecular similarity, 350–356
 - graph-based representations, 353–354
 - set-based representation, 352–353
 - vector-and function-based representations, 354–356
- multiparameter optimization, hit to candidate, 174–175
- Chemical structure analysis. *See also* Molecular similarity analysis (MSA)
- toxicity studies, 272–278
- Chemiluminescent nitrogen detection (CLND), PubChem, 4–6
- ChemistryConnect warehouse, safety modeling, predictive secondary pharmacology, 275–276
- Chemistry space plots, chemical arrays, 182–186
- ChemLink CIDB, 325–327
- Chemogenomics, quantitative structure-activity relationship, 86
- Chemogenomics studies, partial least squares method, 85–86
 - aminergic GPCR inhibitory activity data, 86–87
 - atomic colorings:
 - molecular modeling, 91–93
 - regression coefficient matrix, 90–91
 - LPLS analysis, 91
 - LPLS ligand and protein descriptors, 87–88
 - L-shaped PLS architecture, 89–90
- Chemoinformatics databases (CIDBs): apostle-based rollout systems, 338
- Beacon Projects, 328
- bottom-up *vs.* top-down, 338
- build *vs.* buy *vs.* open source, 339–340
- ChemLink, 325–327
- continuity, 337
- Cousin system, 323–325
- data sources and characteristics, 336
- Mobius CIDB system, 328–335
- off-the-shelf software, 338
- quality software requirements, 336
- RGate 2003+ CIDB, 327
- support/maintenance requirements, 339
- training requirements, 339
- user needs assessment, 336
- Chemometrics, naïve Bayesian models: data types and quality, 139–141
- enriched features, mining and interpretation, 143–145
- molecular representations and machine learning, 138
- target and phenotype comparisons, 141–142
- virtual screening, 134–137
- Circular fingerprints, two-dimensional analysis, 100–101, 110–111
- Classification models, predictive applications, 6–7
- Classification of accuracy, rough set theory, 56–57
- Cluster analysis, 301
- CODD system, chemical arrays, 202
- Cognitive science, molecular similarity and, 347–350
- Combination of evidence, toxicity studies, 278–279
- Combinatorial chemistry, BioProfile system, 302–307
- CoMFA model:
 - activity cliff quantification, 210–211
 - rescaffolding, 225
 - visualization in SAR analysis, 217–222
- Committee models, metabolic lability, 252–256
- Compass 3D-QSAR approach, molecular similarity models, 40–47
- Competing objectives, lead optimization, 171–173
- Compound structure: metabolic lability model, 253–256
- naïve Bayesian models, 139
 - class comparisons, Bayesian weights, 143
 - enriched feature interpretation, 143–145
- Computable similarity function, molecular similarity, 350
- Computational techniques: drug design, 35–47
- naïve Bayesian models, off-target predictions, 137
- toxicity analysis, current research trends, 267–269
- virtual screening, 115–116
- Computational Theory of Mind, 389

- Condition attributes:
 reducts, 61–63
 rough-set theory, 52
- Confidence intervals, predictive model comparisons, 15–17
- Confusion matrix, metabolic lability model, 252–256
- Consensus activity cliffs, 216
- Coordinate-based chemical spaces, molecular similarity, 378–383
- Correlation coefficient, predictive model comparisons, 15–17
- Correlation methods, molecular similarity measurements, 375–376
- Cousin CIDB system, 323–326
- Craig plot, lead optimization, 150–154
- C*-reducts, 61–63
- Cubist software, metabolic lability model, 249–251
- Customer interface, safety profiles, 283–285
- Cyclopentyl ethers, metabolic lability model, DGAT1 inhibitors, 262
- CYP3A4 inhibitor:
 antitarget activity, 228–230
 atom colorings, Bayesian analysis, 90–91
 metabolic lability model, 248–256
 DGAT1 inhibitors, 257–262
 structure-activity relationships, inhibition mechanisms, 230–231
 structure-based ADMET models, 246–247
- Cytochrome P450s (CYPs):
 metabolic lability model, 248–256
 structure-based ADMET models, 247
- Cytohesin inhibitors, virtual screening models, 124–125
- Dabigatran, BICLAIM-space searches, 312–314
- Data-driven modeling, safety analysis, 269–283
- Data fusion methods, molecular similarity analysis, 371–373
- Data mining, privileged substructures, 209
- Dataset characteristics:
 chemoinformatics databases, 336
 drug-induced gene expression, phospholipidosis, 68–71
 lead optimization, retrospective analyses, 171–173
 matched molecular pairs (MMPs), 212–214
 metabolic lability model, 251
 Mobius CIDB system, 330–332
 naïve Bayesian models, 139–141
 structural-activity relationships, pharmaceutical applications, 222–223
- Daylight similarity:
 molecular fingerprinting techniques, 98–100
- molecular fingerprint method comparisons, 103–104
- Decision analysis techniques, lead optimization, 175–176
- Decision attributes, rough-set theory, 52
- Decision table (DT):
 drug-induced gene expression, phospholipidosis, 68, 71–72
 genomics research, 51–52
 rough set theory, 53–56
- Decision trees, metabolic lability model, 249–251
- DEGAS system, chemical arrays, 202
- De Morgan’s Laws, set-based similarity functions, molecular similarity, 357–359
- Dempster-Shafer theory (DST), toxicity studies, 279
- De novo* drug design:
 naïve Bayesian model limitations and, 146–147
 rescaffolding, chemically advanced template search, 224–225
 safety QSAR modeling, 270–271
- Desirability index, multiparameter lead optimization, 160, 162–164
- Diacylglycerol-acyl-transferase 1 (DGAT-1) inhibitors:
 metabolic lability model, 256–262
 three-dimensional ligand-based relationships, 226–227
- Digital PDP-11 graphics system, 323
- Dimensionality reduction, visualization in SAR, 216–222
- Discernibility classes, rough set theory, 59–61
- Discretization analysis, drug-induced gene expression, phospholipidosis, 68–71
- Dissimilarity plots:
 chemical arrays, 182–186
 molecular similarity and, 348–350
- Distance measurements, molecular similarity, 350
- Distance-to-model methods, safety QSAR modeling, 270–271
- Distill program:
 molecular scaffolds, 208
 visualization in SAR and, 216–222
- Downregulated scores, drug-induced gene expression, phospholipidosis, 68–71
- DRAGON descriptors, 249–256
- D*-reducts, 61–63
 drug-induced gene expression, phospholipidosis, 68, 71–72
- Drug design:
 incrementalism and serendipity, 33–35
 physical processes and computational methods, 35–47
 molecular similarity, 39–45

- protein structure-based methods, 37–39
- three-dimensional QSAR, 45–47
- toxicity warning systems, 276–278
- virtual screening, 121–123
- Drug-induced gene expression, phospholipidosis, 65–77
 - data set characteristics, 68–71
 - D*-reduct determination, 68, 71–72
- Ecto-5'-nucleotidase, virtual screening, 125–128
- Edge effects, classification models, 6–7
- Electronic lab notebook (eLNB), chemical array annotation, 194–195
 - seed compounds, 199–201
- Electronic laboratory network (ELN), controlled substances annotation, 315
- Electrostatic similarity, molecular similarity models, 44–45
- Endpoint analysis, *in vitro* studies, safety QSAR modeling, 269–270
- EON* system, three-dimensional ligand-based relationships, 226–227
- Exclusive-or (XOR) problem, drug design and, 35–47
- Expectation values, Tanimoto similarity, 102
- Extended-connectivity fingerprints of depth 4 (ECFP_4), naïve Bayesian models, 133–134
 - virtual screening and, 134–137
- Extended-connectivity fingerprints of depth 6 (ECFP_6):
 - aminergic GPCR inhibitory activity, 86–87
 - L-shaped partial least squares analysis, atomic colorings, regression coefficient matrix, 91
- Extreme programming, safety modeling, 283–285
- Factorial design, lead optimization, 151–152
- Factor Xa inhibitors:
 - model diagnosis plots, 210–211
 - R-group plots, 214–215
 - visualization in SAR analysis of, 217–222
- FCFP4 fingerprints, metabolic lability model, 254–256
- Feature-based fingerprints:
 - cognitive science and, 348–350
 - molecular similarity applications, 98–100
 - naïve Bayesian models, enrichment and interpretation, 143–145
- Feature trees, rescuffolding, 224–225
- Filter techniques, multiparameter lead optimization, 159–160
- First-in-class inhibitors, virtual screening, 125–128
- Fisher's *z*' distribution, predictive model comparisons, 15–17
- Fixed-length fingerprints, molecular similarity, set-based representation, 352–353
- Fraction of sp₃ carbons, multiparameter lead optimization, 158–159
- Fragility limitations, molecular fingerprinting techniques, 98–100
 - circular fingerprints, 100–101
- Frequent hitter analysis, 304–307
- Fuzzy relations, molecular similarity, 350
 - chemical space representation, 380–383
 - set-based representation, 353
 - set-based similarity functions, 357–361
- Gaussian reward functions:
 - molecular similarity models, 40–45
 - structure-activity relationships, three-dimensional ligand-based relationships, 225–227
- Genetic algorithms:
 - lead optimization, 172–173
 - metabolic lability model, 249–256
- Genetox Warning System (GWS), 276–278
- GOLD molecular docking technique, protein-ligandbinding, 38
- G protein-coupled receptor (GPCR) inhibitory data:
 - L-shaped partial least squares analysis:
 - atomic colorings:
 - molecular modeling, 92–93
 - regression coefficient matrix, 90–91
 - ligand and protein descriptors, 87–88
 - naïve Bayesian models, 145–146
 - quantitative structure-activity relationship, 86
 - aminergic GPCR inhibitory activity data, 86–87
 - safety modeling, predictive secondary pharmacology, 275–276
 - Graph-based algorithms, maximum common substructure, 208
 - Graphic-user interface (GUI), chemoinformatics databases, 326
 - GRASP visualization program, molecular fingerprinting, 111
 - Group fusion, molecular similarity analysis, 372
 - H₂ antagonists, multiobjective lead optimization, 154–157
 - Hansch values, lead optimization, 151–154
 - Heatmap representation, activity cliff analysis, 215–216

- HierS* scaffold clustering system, molecular scaffolds, 208
- High-throughput screening (HTS):
 activity cliffs, outlier investigations, 209–210
 lead optimization, retrospective analyses, 171–173
 molecular fingerprinting techniques, 99–100
 molecular scaffolds, 208
 multiobjective lead optimization, 155–157
- Histamine 3 receptors, safety modeling, predictive secondary pharmacology, 275–276
- Historical data, multiparameter lead optimization, 168–169
- Hit identification. *See also* Lead optimization
 multiparameter optimization, hit to candidate, 174–175
 virtual screening:
 applications, 116–117, 123–128
 first-in-class inhibitor,
 ecto-5'-nucleotidase, 125–128
 multifunctional protein inhibitors, 124–125
 definition and classification, 113–114
 future research issues, 128
 high-throughput screening *vs.*, 119–121
 molecular drug targets, 121–123
 performance evaluation, 117–119
- Human ether-à-go-go related gene (hERG):
 chemical arrays, 202
 chemotype detection, 196, 198
 off-target activity models, 1, 137
 oxadiazole topology and, 213
 regression models, 7
 scoring profile, 165
 structure-activity relationships:
 antitarget activity hotspots, 228–230
 inhibition mechanisms, 230–231
 temporal analysis, 191
 three-dimensional QSAR model, 246–247
 toxicity analysis:
 cardiac liability risk reduction, 271–272
 current research trends, 268–269
- Human hepatoma HepG2 cells, drug-induced gene expression, phospholipidosis, 65–77
- Hungarian algorithm, molecular fingerprinting, 111
- Huuskonen dataset:
 predictive modeling and, 3–6
 applicability, 13–15
 random forest model construction, 9–11
- Hydrogen bond donors and acceptors multiparameter lead optimization, 158–159
- desirability functions, 163–164
- Hydrophobic ligands, protein-ligand binding, 37–39
- Icaris system, chemoinformatics databases, 328
- IF-THEN rules:
 drug-induced gene expression,
 phospholipidosis, preliminary rules generation, 72–75
- Imprecision, similarity analysis, 98–100
- Incremental modification, drug design, 34–35
- Indiscernibility classes:
 drug-induced gene expression,
 phospholipidosis, *D*-reducts, 71–72
- Innovative Medicines Initiative (IMI) eTOX project, toxicity analysis, 282–283
- In silico* predictions:
 ADMET global model development and applications:
 future research issues, 262
 ligand-based models, 247–248
 metabolic lability, 248–262
 dataset characteristics, 251
 descriptor set results, 251–256
 DGAT1 inhibitor optimization, 256–262
 machine learning and molecular descriptors, 249–251
 structure-based models, 246–247
 multiparameter optimization, hit to candidate, 174–175
 safety modeling, 268–269
 preclinical data, 282–283
 predictive secondary pharmacology, 275–276
 three-dimensional ligand-based relationships, 226–227
- Integrated chemoinformatics systems:
 apostle-based rollout systems, 338
 Beacon Projects, 328
 bottom-up *vs.* top-down, 338
 build *vs.* buy *vs.* open source, 339–340
 ChemLink, 325–327
 Cousin system, 323–325
 data sources and characteristics, 336
 Mobius CIDB system, 328–335
 off-the-shelf software, 338
 quality software requirements, 336
 RGate 2003+ CIDB, 327
 succession planning, 339
 support/maintenance requirements, 339
 training requirements, 339
 user needs assessment, 336
- Interquartile range (IQR), datasets, 4–6
- Intersection sets, set-based similarity functions, molecular similarity, 357–359
- In vitro* analysis:
 ADMET global model development and applications, 245–248

- multiparameter optimization, hit to candidate, 174–175
- naïve Bayesian models, 143–145
- toxicity studies:
 - biological data, 281–283
 - chemical structure links, 269–272
 - preclinical data, 282–283
 - safety QSAR endpoint modeling, 269–270
 - warning systems, 276–278
- In vivo* analysis:
 - multiparameter optimization, hit to candidate, 174–175
 - toxicity studies:
 - in vitro* profile data, 281–283
 - warning systems, 277–278
- ISIS components, chemoinformatics databases, 326, 328
- Isostere generation, molecular fingerprinting techniques:
 - basic principles, 99–100
 - WABE programming, 101–102
- IUPAC International Chemical Identifier (InChI), 139
- Kendall’s tau, predictive model performance evaluation, 7–8
- molecular similarity, 14–15
- Kernel loadings matrices, L-shaped partial least squares, 90
- k*-nearest neighbor (kNN) QSAR:
 - metabolic liability model, 249–256
 - Similarity Principle, 207
- KNIME workflow tool:
 - BioProfile system 304–307
 - Boehringer Ingelheim Mining and Exploration of Screening Hit (BIMESH) system, 300–301
 - Boehringer Ingelheim project data marts, 297
 - Boehringer Ingelheim workflow system, 294–295
- Kohonen networks, visualization in SAR and, 219–222
- Lead-hopping, molecular fingerprint method comparisons, 103–104
- Lead optimization:
 - antitarget activity, 232
 - chemical arrays:
 - annotation, 194–195
 - archive information extraction, 194–201
 - automatic chemotype detector, 195–198
 - chemistry and property space coverage, 182–186
 - data analysis techniques, 192–194
 - seed compound detection, 196, 198–201
 - self-avoiding random walk, 191–192
 - temporal analysis, 186–191
- chemoinformatics:
 - overview, 149–154
 - multiobjective methods, 158–169
 - basic rules, 158–159
 - desirability functions, 160–164
 - filters, 159–160
 - probabilistic scoring, 164–165
 - property profile, 165–169
 - multiparameter optimization, hit to candidate, 174
 - retrospective analyses, 169–173
- structure–activity relationships (SARs):
 - activity cliffs:
 - exploration, 214–216
 - outlier investigation, 209–210
 - quantification, 210–211
 - antitarget activity, 227–232
 - activity hotspot identification and application, 228–230
 - hERG and CYP3A4 inhibition, 230–231
 - transfer mechanisms, 227–228
 - matched molecular pairs, 211–214
 - molecular scaffolds, 207–208
 - overview, 205–206
 - pharmaceutical industry applications, 222–223
 - privileged substructures, 208–209
 - rescaffolding, 223–227
 - three-dimensional-ligand-based approaches, 225–227
 - three-dimensional-protein-based approaches, 227
 - two-dimensional approaches, 224–225
 - similarity principle, 207
 - visualization support, 216–222
- LeadScope program, visualization in SAR analysis, 216–222
- Learned ideal distance, molecular similarity models, 40–45
- Lexicographic fingerprints, molecular similarity applications, 98–100
- LHASA project, 323
- Ligand-based ADMET models, properties of, 247–248
- Ligand-based virtual screening (LBVS):
 - applications, 123–128
 - correlation methods, molecular similarity and, 375–376
 - cytohesin inhibitors, 124–125
 - definition and classification, 113–114
 - performance evaluations, 117–119
 - practical applications, 116–117

- Ligand descriptors, L-shaped partial least squares analysis, 87–88
- Ligand-efficiency (LE) analysis, 308–309
- Linear cascade approach, temporal analysis, lead optimization, 187–191
- LINGO fingerprint method:
molecular similarity applications, 98–100
two-dimensional analysis, 101
- Linguistic rules, rough set theory, 55–56
- Lipinski's Rule of Five:
ADMET global model development and applications, 246–248
multiparameter lead optimization, 158–159
- LiSARD* interactive graphics, visualization in SAR and, 219–222
- Local neighborhood plots, visualization in SAR and, 219–222
- Log *P*:
multiparameter lead optimization, 158–159
temporal analysis, lead optimization, 186–191
- Log ratios, naïve Bayesian models, 133–134
- Log *S*, random forest model construction, 9–11
- L-shaped partial least squares (LPLS):
chemogenomics studies:
 aminergic GPCR inhibitory activity data, 86–87
 atomic colorings:
 molecular modeling, 91–93
 regression coefficient matrix, 90–91
 LPLS analysis, 91
 LPLS ligand and protein descriptors, 87–88
 L-shaped PLS architecture, 89–90
- MACCS keys fingerprint:
metabolic lability model, 250–251
molecular similarity applications, 98–100
 chemical space representation, 380–383
 statistical independence, 374–375
two-dimensional analysis, 100–101
- Machine learning:
drug design and, 36–47
metabolic lability model, 249–256
naïve Bayesian models, comparisons with, 138–139
safety QSAR modeling, 270–271
- Magic methyl phenomenon, lead optimization, 156–157
- Maintenance systems, chemoinformatics database systems, 339
- Mantel statistic, molecular similarity techniques, 376–377
- MARS program:
antitarget activity analysis, 230
- three-dimensional ligand-based relationships, 226–227
- Matched molecular pairs (MMPs):
activity cliff analysis, 216
molecular similarity, 384
antitarget activity, 229–230
- Mathematical descriptions, molecular similarity, 350
- Maximum common substructure (MCS):
antitarget activity, 229–230
molecular scaffolds, 207–208
molecular similarity, chemical graph-based similarity functions, 361–363
- Mechanism cliffs, 216
- Metabolic lability model, development and applications, 248–262
- dataset characteristics, 251
- descriptor set results, 251–256
- DGAT1 inhibitor optimization, 256–262
- machine learning and molecular descriptors, 249–251
- MetaPrint 2D software, toxicity warning systems, 277–278
- MetaSite database:
CYP enzymes, 247
metabolic lability model, DGAT1 inhibitors, 257–262
- Metatables, Mobius CIDB system, 332–333
- Metatree, Mobius CIDB system, 333
- Metaview, Mobius CIDB system, 332–333
- Methadone, molecular similarity model, 45
- Methotrexate, WABE molecular fingerprinting, 104–110
- Microsomal lability, metabolic lability model, DGAT1 inhibitors, 260–262
- Minimum distance measurements, molecular similarity models, 40–45
- Mobius CIDB system, 328–335
 ad hoc query interface, 333–335
 data sources, 330–332
 metaview, 332–333
 query engine, 333
 software components, 335
- Mobius Query Language (MQL), Mobius CIDB system, 333
- Model applicability domains, metabolic lability model, DGAT1 inhibitors, 257–262
- Model diagnosis plots, activity cliff quantification, 210–211
- Mode of action (MOA), naïve Bayesian models:
 feature enrichment and interpretation, 143–145
 phenotypic screening, 135–137
- Molecular operating environment (MOE) program, 216–222, 250–252

- Molecular Data Explorer program, structure similarity maps, 219–222
- Molecular descriptors:
- ligand-based ADMET models, 247–248
 - metabolic lability model, 249–256
 - molecular similarity, 350–356
 - graph-based representations, 353–354
 - set-based representation, 352–353
 - vector and function-based representations, 354–356
 - naïve Bayesian models, 138–139
 - safety QSAR modeling, 270–271
 - structure–activity relationships, activity cliffs, outlier investigations, 210
- Molecular docking techniques:
- protein-ligand-binding and, 37–39
 - structure-based ADMET models, 247
- Molecular equivalence number structural classification system (Meqnum), molecular scaffolds, 208
- Molecular fingerprinting:
- chemical arrays, 182–186
 - CYP enzymes, 247
 - metabolic lability model, 253–256
 - molecular similarity:
 - set-based representation, 352–353
 - set-based similarity functions, 357–361
 - weighted representations, 356–357
 - naïve Bayesian models, 138–139
 - safety modeling, predictive secondary pharmacology, 275–276
 - stability analysis:
 - isostere generation, WABE program, 101–102
 - Tanimoto similarity, 102
 - two-dimensional methods, 100–101
 - toxicity studies, 281–283
- Molecular scaffolds, structure–activity relationships, 207–208
- Molecular similarity analysis (MSA):
- activity landscapes and cliffs, 383–384
 - asymmetric similarity, 366–368
 - chemical space representation, 350–356, 378–383
 - graph-based representations, 353–354
 - set-based representation, 352–353
 - vector and function-based representations, 354–356
 - cognitive aspects, 347–350
 - comparison of measurement methods, 375–377
 - data fusion and consensus methods, 371–375
 - functions/coefficients, 357–365
 - chemical graph-based similarity functions, 361–363
 - set-based similarity functions, 357–361
 - vector/function-based similarity functions, 363–365
- metabolic lability model, 253–256
- molecular fingerprinting:
- isostere generation, WABE program, 101–102
 - Tanimoto similarity, 102
- molecular representation, 350–356
- graph-based representations, 353–354
 - set-based representation, 352–353
 - vector and function-based representations, 354–356
- SAR and QSAR techniques, 97–100
- statistical independence, 374–375
 - structure–activity similarity and related maps, 384–387
 - two-dimensional asymmetric similarity searching, 368–370
 - validation of measurements, 377–378
 - weighted representations, 356–357
- Molecular weight, multiparameter lead optimization, 158–159
- desirability functions, 163–164
- MolPrint 2D fingerprints, chemical arrays, 182–186
- MPP12 inhibitors, lead optimization, 172–174
- mRNA scores, drug-induced gene expression, phospholipidosis, 68–71, 77–79
- Muchmore–Martin lead-hop sets, molecular fingerprint method comparisons, 103–104
- Multiclass naïve Bayesian models, potency binning, 140–141
- Multidimensional scaling, visualization in SAR analysis, 219–222
- Multifunctional protein inhibitors, virtual screening models, 124–125
- Multi-fusion similarity, 373
- Multiojective methods:
- lead optimization, process overview, 154–157
 - temporal analysis, lead optimization, 186–191
- Multiparameter optimization (MPO), lead optimization, 158–169
- basic rules, 158–159
 - desirability functions, 160, 162–164
 - filters, 159–161
 - hit to candidate process, 174–175
 - probabilistic scoring, 164–167
 - property profile, 165–169
- Multiple feature tree model (MTree), rescaffolding, 224–225
- Multiplicative methods, lead optimization, desirability functions, 163–164

- Multiset fingerprints, molecular similarity, set-based representation, 352–353
- Multitargeted assays, 132–134, 216
- Muscarinic antagonists, molecular similarity models, 44–45
- Naïve Bayesian models (NBMs):
chemometric applications:
 data types and quality, 139–141
 enriched features, mining and interpretation, 143–145
 target and phenotype comparisons, 141–142
 virtual screening, 134–137
safety modeling, predictive secondary pharmacology, 275–276
- Nearest neighbor techniques:
 molecular similarity, chemical space representation, 378–383
 safety QSAR modeling, 270–271
- Neighborhood plots, activity cliff quantification, 210–211
- Network-like similarity graphs (NSGs):
 activity cliff analysis, 215–216
 virtual screening, 121–128
 visualization in SAR, 221–222
- NIPALS algorithm, L-shaped partial least squares, 89–90
- Nondeterministic rules, rough-set theory, 64
- Nonlinear structure–activity relations:
 multiobjective lead optimization,
 H2-antagonists, 154–157
 visualization techniques, 219–222
- Norepinephrine transporter inhibitors,
 structure–activity similarity (SAS) maps, 385–387
- Off-target candidates:
 drug design, 35
 naïve Bayesian models, 137
 predictive secondary pharmacology, 274–276
- Off-the-shelf software, chemoinformatics databases, 338
- Open source platforms, chemoinformatics database systems, 339–340
- Over-robustness, molecular fingerprinting techniques, 99–100
- Oxadiazole derivatives, matched molecular pairs (MMPs), 212–214
- Oxidative metabolic clearance, metabolic lability model, 248–256
- PARASURF system, metabolic lability model, 250–252
- Pareto optimization:
 multiparameter lead optimization, 166, 168–170
 temporal analysis, lead optimization, 186–191
- Partial least squares (PLS) method:
 activity cliff quantification, 210–211
 chemogenomics studies:
 LPLS analysis, 91
 LPLS ligand and protein descriptors, 87–88
 L-shaped PLS architecture, 89–90
 visualization in SAR, 217–222
- Path-based fingerprints:
 comparison with other methods, 105–110
 molecular similarity applications, 98–100
 two-dimensional analysis, 100–101
- PDE5 inhibitors, protein-ligand-binding, 38–39
- Pearson's *r*:
 molecular fingerprint method comparisons, 103–110
 predictive model performance evaluation, 7
 confidence intervals, 15–17
 molecular similarity, 14–15
 random forest model construction, 10–11
- Perceptrons, 36–47
- PFAKT system, chemical arrays, 202
- Pharmacophore models, ligand-based ADMET models, 247–248
- Pharmacophoric triplets, structure activity relationships, 225–227
- Phenotypic screening, naïve Bayesian models:
 target comparison, chemical and biological activity space, 141–143
 target identification, 135–137
- Phospholipidosis:
 drug-induced gene expression, 65–77
 dataset characteristics, 68–71
 D-reduct determination, 68, 71–72
 preliminary rule generation, 72–75
 rule simplification-attribute value reduction, 75–77
- Physical reality models:
 drug design and, 39–45
 three-dimensional QSAR, activity prediction, 45–47
- Pipeline Pilot:
 automatic chemotype detection, 195
 BICLAIM-space searches, 311–314
 Boehringer Ingelheim project data marts, 297
 Boehringer Ingelheim workflow system, 294–295
 chemical array systems, 192
 molecular fingerprinting and, 86–87
 pairwise structural dissimilarity analysis, 182–183
 seed detection workflow, 201

- Plato system, usage facilitation and drug safety analysis, 285
- Polar moieties, molecular similarity models, 44–45
- Polar surface area (PSA), multiparameter lead optimization, 158–159
- Polypharmacology, toxicity studies, 281–283
- Potassium channel, hERG inhibition, 230–231
ligand-based ADMET models, 247–248
- Potency binning, multiclass naïve Bayesian models, 140–141
- Potential genotoxic impurity (PGI), current research trends, 267–269
- Preclinical data, toxicity studies, 282–283
- Prediction error analysis, metabolic liability model, 253–256
- Predictive models:
applicability, 12–15
experimental error and performance, 11–12
molecular descriptors, 8–9
performance evaluation, 7–8
random forest model example, 9–11
safety QSAR modeling, 270–271
predictive secondary pharmacology, 274–276
source code listings, 20–30
- Predictive secondary pharmacology (PSP),
toxicity analysis, 274–276
in vitro profile data, 281–283
- Pregnane X receptor (PXR), structure-based ADMET models, 246–247
- Primary target pairing, 34–35
- Principal component analysis (PCA):
molecular similarity, chemical space representation, 380–383
visualization in SAR, 217–222
- Privileged substructures, structure-activity relationships, 208–209
- Probabilistic rules:
multiparameter lead optimization, 164–167
- Project coding, chemical array annotation, 194–195
- Project data marts, Boehringer Ingelheim CDB, 297
- Property landscapes:
activity cliff quantification, molecular similarity, 384
molecular similarity, 384
- Property profiles:
matched molecular pairs (MMPs), 212–214
multiparameter lead optimization, 165, 168–169
- Property space plots, chemical arrays, 182–186
- Protein descriptors:
L-shaped partial least squares analysis, 87–88
structure-activity relationships, rescaffolding, 227
- Protein kinases, privileged substructures, 209
- Protein-ligand binding interactions, 36–47
- Protein pocket conformation, protein-ligandbinding, 38–39
- Protein structure-based methods, 37–39
- Pseudoequivalences, molecular similarity, set-based representation, 353
- PubChem, 4–6
molecular similarity, 14–15
- Python programming language, 2–31
- Quantitative annotation, naïve Bayesian models, 139–140
- Quantitative estimate of drug-likeness (QED):
lead optimization, desirability functions, 163–164
multiparameter lead optimization, property profiles, 168–169
- Quantitative structure-activity relationship (QSAR):
activity cliff quantification, 210–211
chemical arrays, 181–182
lead optimization, multiobjective processes, 156–157
metabolic liability model, 253–256
DGAT1 inhibitors, 258–262
omics applications, 86
partial least square method, 85–86
aminergic GPCR inhibitory activity data, 86–87
atomic colorings:
molecular modeling, 91–93
regression coefficient matrix, 90–91
- LPLS analysis, 91
- LPLS ligand and protein descriptors, 87–88
- L-shaped PLS architecture, 89–90
- safety modeling:
cardiac liability risk reduction, 271–272
machine learning, 270–271
predictive secondary pharmacology vs., 274–276
software technology, 283–285
in vitro endpoints, 269–270
- similarity principle, 207
- Quantitative structure-property relationship (QSPR), metabolic liability model, DGAT1 inhibitors, 257–262
- Query compounds:
safety QSAR modeling, 270–271
visualization in SAR, 221–222
- Query engine, Mobius CIDB system, 333
- Random forest model, building and testing, 9–11
- Random walk models, lead optimization, 191–194

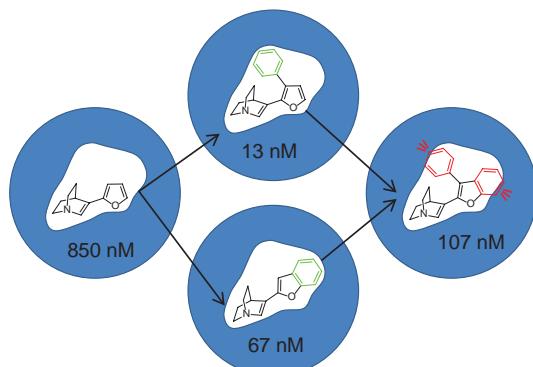
- Rapid Overlay of Chemical Structures (ROCS) program:
 antitarget activity analysis, 230
 molecular similarity models, 39–45
 structure activity relationships, three-dimensional ligand-based relationships, 225–227
- RDKit chemoinformatics programming library, predictive modeling and, 2–31
- Reactive Metabolite Warning System (RMWS), toxicity studies, 277–278
- Read-across analysis, toxicity studies, 274
- RECAP methodology:
 maximum common substructure, 208
 seed compounds, chemical arrays, 199–201
- Receiver operating characteristic (ROC), 377–378
- Recombinant protein assays, multiobjective lead optimization, 156–157
- Reduced graphs, molecular scaffolds, 208
- Reference (probe) molecule:
 asymmetric similarity, 366–370
 molecular similarity measurements, 375–376
- Regression coefficient matrix, L-shaped partial least squares analysis, atomic colorings, 90–91
- Regression models, L endo-and exo-LPLS, chemogenomics analysis, 89–90
- Rescaffolding, structure-activity relationships transfer, 223–227
 three-dimensional ligand-based approaches, 225–227
 three-dimensional protein-based approaches, 227
- Retrospective analyses, lead optimization, 169, 171–173
- Reverse virtual screening, 135–137
- RGate 2003+ CIDB, 327
- R-groups:
 lead optimization:
 activity cliffs exploration, 214–216
 matched molecular pairs (MMPs), 212–214
 seed compound detection, 199
 self-avoiding random walks, 191–192
 visualization techniques, 216–222
- Ring assemblies:
 BICLAIM-space searches, 311–314
 chemotype detection, chemical arrays, 196–198
- ROCK system, chemical arrays, 202
- Rollout systems, chemoinformatics databases, apostle approach, 338
- Root-mean-square deviation (RMSD):
 predictive model performance evaluation, 8
 random forest model construction, 10–11
- Rotatable bonds (RotB), multiparameter lead optimization, 158–159
 desirability functions, 163–164
- Rotated factorial design, lead optimization, 151–152
- Rough set theory (RST):
 chemically induced gene expression data, 57–61
 classification accuracy and quality, 56–57
 drug-induced gene expression, phospholipidosis, 65–77
 dataset characteristics, 68–71
D-reduct determination, 68, 71–72
 preliminary rule generation, 72–75
 rule simplification–attribute value reduction, 75–77
 rule generation, 63–64
- R statistics program:
 predictive modeling, 2–31
 molecular similarity applications, 12–15
 random forest model construction, 9–11
- Rule-based methods, 51–52
- Safety profiles:
 biological data, 279–283
 chemical structure and available data, 272–278
 combination of evidence, 278–279
 data-driven modeling, 269–283
 future research issues, 285–286
 preclinical data, 282–283
 predictive secondary pharmacology, 274–276
 quantitative structure activity relationships:
 cardiac liability risk reduction, 271–272
 machine learning, 270–271
 in vitro endpoints, 269–270
 read-across analysis, 274
 warning systems in, 276–278
- SaliExplorer program, activity cliff analysis, 215–216
- SAR Map technique, visualization in SAR analysis, 216–222
- Scaffold compounds. *See also* Rescaffolding chemotype detection, 196–198
 structure-activity relationship, 207–208
- Scaffold hopping potential, virtual screening performance, 118–119
- Scaffold trees, molecular scaffolds, 208
- Scores matrix, three-dimensional ligand-based relationships, 226–227
- Scoring profiles, multiparameter lead optimization, probabilistic scores, 164–165
- SCRUM, safety modeling, 283–285
- Secins, cytohesin inhibitors, 124–125

- Seed-array scatter plots, chemical arrays, 182–186
 Seed compound detection, chemical arrays, 196, 198–201
 Selectivity cliffs, 216
 self-avoiding random walk (SAW), lead optimization, 191–194
 Self-organizing maps (SOMs), visualization in SAR and, 219–222
 Semi-naïve Bayesian models (SNBMs), 144–145
 Sequel language, 323
 Sequential screening, virtual screening, 121
 Serendipity, drug design, 35–36
 Side effect pairing, 34
 Sildenafil, protein-ligandbinding, 38–39
 Similarity ensemble approach (SEA):
 naïve Bayesian models, class comparisons, 143
 safety modeling, predictive secondary pharmacology, 275–276
 Similarity functions, molecular similarity:
 chemical graph-based similarity functions, 361–363
 set-based similarity functions, 357–361
 vector-based and function-based functions, 363–365
 Similarity fusion, molecular similarity analysis, 372
 Similarity matrices, molecular similarity techniques, 376–377
 Similarity principle. *See also* Molecular similarity analysis (MSA)
 Simplex optimization, lead optimization, 153–154
 Simplified molecular input line entry system (SMILES):
 chemotype detection, chemical arrays, 195–198
 matched molecular pairs (MMPs), 212–214
 molecular similarity applications, 98–100
 two-dimensional analysis, 101
 Singular value decomposition, chemogenomics analysis, 89–90
 Site of metabolism (SOM) predictions:
 CYP enzymes, 247
 metabolic lability model, DGAT1 inhibitors, 257–262
 Small molecule drug development, molecular similarity and, 39–45
 SMARTS notation, matched molecular pairs (MMPs), 212–214
 Software technology:
 chemoinformatics databases, quality controls, 336
 Mobius CIDB system, 335
 off-the-shelf software, 338
 safety modeling, 283–285
 Solvated states, protein-ligand-binding and, 37
 Source code listings, predictive models, 20–30
 Speed hazards, chemoinformatics databases, 337
 Spiral view techniques, visualization in SAR, 221–222
 Splitting operations, BICLAIM-space searches, 312–314
 Spotfire Decision Site, 328
 Stability analysis, molecular fingerprinting:
 future research issues, 110–111
 isotere generation, WABE program, 101–102
 overview, 97–100
 results, 103–110
 Tanimoto similarity, 102
 two-dimensional methods, 100–101
 Standard deviations, Tanimoto similarity, 102
 Statistical independence, molecular similarity, 374–375
 Stepwise analysis, metabolic lability model, 251–256
 Stochastic neighborhood embedding (SNE), visualization in SAR and, 219–222
 Structural alerts, toxicity analysis, 273–274
 Structural complexity (SC), chemical arrays, 195–198
 Structure-activity landscape analysis, 326
 Structure-activity landscape index (SALI):
 activity cliff quantification, 210–211, 215–216
 molecular similarity, 383–384
 antitarget activity analysis, 229–230
 Structure-activity relationship index (SARI),
 activity cliff quantification, molecular similarity, 384
 Structure-activity relationships (SARs). *See also* Quantitative structure-activity relationship (QSAR)
 ADMET models, 246–247
 Boehringer Ingelheim chemoinformatics case study, 305–309
 chemical arrays, 180–181, 201–203
 drug design, 35–47
 drug-induced gene expression, phospholipidosis, 66
 lead optimization:
 activity cliffs:
 exploration, 214–216
 outlier investigation, 209–210
 quantification, 210–211
 antitarget activity, 227–232
 activity hotspot identification and application, 228–230
 hERG and CYP3A4 inhibition, 230–231
 project integration, 232
 transfer mechanisms, 227–228
 future research issues, 232–233
 matched molecular pairs, 211–214

- Structure-activity relationships (SARs). (*Continued*)
 molecular scaffolds, 207–208
 overview, 205–206
 pharmaceutical industry applications, 222–223
 privileged substructures, 208–209
 rescuffolding, 223–227
 three-dimensional-ligand-based approaches, 225–227
 three-dimensional-protein-based approaches, 227
 two-dimensional approaches, 224–225
 similarity principle, 207
 visualization support, 216–222
 molecular similarity and, 97–100
 multiobjective lead optimization, 154–157
 future research issues, 176–177
 naïve Bayesian models, 132–134
 molecular similarity combined with, 138–139
 virtual screening, 121–123
- Structure-activity similarity (SAS) maps:
 molecular similarity applications, 384–387
 visualization in SAR and, 221–222
- Structure-based ADMET models, development and applications, 246–247
- Structure-based virtual screening (SBVS):
 correlation methods, molecular similarity and, 375–376
 definition and classification, 113–114
 ecto-5'-nucleotidase first-in-class inhibitor, 125–128
 performance evaluations, 117–119
 practical applications, 116–117
- Structured Query Language (SQL), 323
 Mobius CIDB system, 333
- Structure-property relationships (SPRs),
 Boheringer Ingelheim chemoinformatics case study, 308–309
- Structure similarity maps, visualization in SAR analysis, 219–222
- Substructures:
 BICLEAM-space searches, 311–314
 privileged substructures, 208–209
 structural alerts, toxicity studies, 273–274
- Succession planning, chemoinformatics database systems, 339
- Superfluous attributes:
 drug-induced gene expression, phospholipidosis, reduction of, 75–77
 rough set theory, 61–63
- Support systems, chemoinformatics database systems, 339
- Support vector machine (SVM) modeling, 124–125
- Surface interactions, molecular similarity models, 39–45
- Surflex-Dock molecular docking, protein-ligand-binding and, 38–39
- Surflex-QMOP approach, three-dimensional QSAR, activity prediction, 45–47
- Surflex-Sim approach, molecular similarity models, 39–45
- Tadalafil, protein-ligandbinding, 38–39
- Tanimoto similarity:
 chemical arrays, 182–186
 fingerprint method comparisons with, 104–110
 Huuskonen training set, 13–15
 molecular fingerprinting techniques:
 expected values, 110–111
 two-dimensional applications, 102
 set-based similarity functions, 360–361
- Target fishing:
 asymmetric similarity, 350
 naïve Bayesian models:
 phenotype comparison, chemical and biological activity space, 141–143
 reverse virtual screening, 135–137
- Temporal analysis:
 lead optimization, chemical arrays, 186–191
 metabolic lability model, 255–256
- Temporal partitioning, molecular similarity model, 45–47
- Test-Driven Development (TDD), safety modeling, 284
- Testicular Toxicity Warning System (TTWS), 278
- Three-dimensional activity landscape models, activity cliff analysis, 215–216
- Three-dimensional fingerprints, molecular similarity, set-based representation, 352–353
- Three-dimensional ligand-based models:
 ADMET models, 246–247
 structure-activity relationships, 225–227
- Three-dimensional protein-based techniques, structure-activity relationships, 227
- Three-dimensional QSAR:
 ADMET models, 246–247
 fingerprint techniques, 98–100
 physically realistic activity prediction, 45–47
- Three-dimensional similarity, computational techniques, 370–371
- Three-point pharmacophore fingerprints, structure-activity relationships, 225–227
- Top-down processing, chemoinformatics databases, 338

- Topliss Tree, lead optimization, 150–154
 Topoisomer searching, rescuffolding, 224–225
 Topological frameworks:
 molecular scaffolds, 208
 rescuffolding, 224–225
 Toxicity studies:
 biological data, 279–283
 chemical structure and available data, 272–278
 customer interface, 283–285
 data-driven modeling, 269–283
 preclinical data, 282–283
 predictive secondary pharmacology, 274–276
 read-across analysis, 274
 structural alerts in, 273–274
 in vitro analysis, chemical structure links, 269–272
 warning systems in, 276–278
 Traffic light visualization, multiparameter lead optimization, 160–161
 probabilistic scores, 164–165
 Training programs, chemoinformatics database systems, 339
 Training set molecules, predictive model performance *vs.*, 12–15
 Tree-based fingerprints, two-dimensional analysis, 101
 $t-u$ plots:
 activity cliff quantification, 210–211
 visualization in SAR, 217–222
 Tversky similarity, asymmetric similarity, 367–370
 Two-dimensional representation:
 chemical graph-based representations, 353–354
 computational techniques, 370–371
 set-based representation, 352–353
 naïve Bayesian models, comparisons with, 138–139
 rescuffolding, non-fingerprint techniques, 224–225
 Unbound states, protein-ligand binding, 37
 Union sets, set-based similarity functions, molecular similarity, 357–359
 Unity fingerprints, 254–256
 Upregulated scores, drug-induced gene expression, phospholipidosis, 68–71
 Urotensin II receptor (UTR), virtual screening, 122–123
 Usage facilitation, safety profiling and, 284–285
 User interface, chemoinformatics databases, 336
 Validation, molecular similarity techniques, 377–378
 Van der Waals interactions, protein-ligand binding, 37–39
 Variable-length fingerprints, molecular similarity, set-based representation, 352–353
 Vector-based representation, molecular similarity, 354–356
 similarity functions, 363–365
 Virtual combinatorial libraries, 297–299
 Virtual screening (VS):
 applications, 116–117, 123–128
 first-in-class inhibitor, ecto-5'-nucleotidase, 125–128
 multifunctional protein inhibitors, 124–125
 definition and classification, 113–114
 high-throughput screening *vs.*, 119–121
 molecular drug targets, 121–123
 naïve Bayesian models, 134–137
 performance evaluation, 117–119
 Visualization techniques, structure-activity relationships, 216–222
 Volumetric functions, molecular similarity models, 41–45
 WABE program, isostere generation, molecular fingerprinting applications, 101–102
 Warning systems, toxicology analysis and, 276–278
 Water partition coefficient, multiparameter lead optimization, 158–159
 Weighted desirability scoring, temporal analysis, lead optimization, 186–191
 Weighted representations, molecular similarity, 356–357
 Weight of evidence (WoE), toxicity analysis, 279–280
 Workflow systems, 294–295, 315–317
 z -scales, L-shaped partial least squares analysis:
 heat mapping, regression coefficient matrix, 92–93
 ligand and protein descriptors, 87–88
 z -score:
 fingerprint method comparisons, 106–110
 Tanimoto similarity, 102

Nonadditivity in SAR



The XOR problem

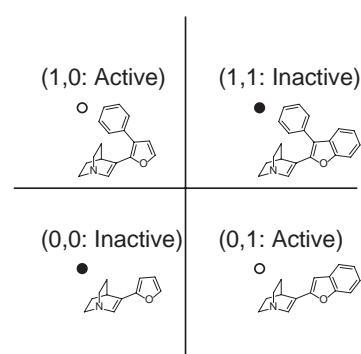


FIGURE 2.3 Four muscarinic antagonists, all on a quinuclidinene-furan scaffold, exhibit a pattern of potency variation that is both highly nonadditive and is isomorphic to the classic XOR problem.

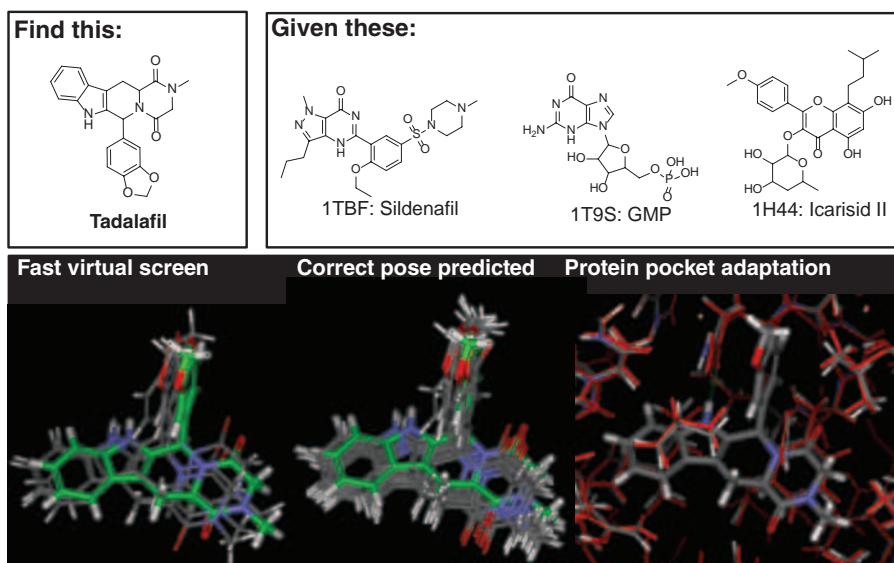


FIGURE 2.5 It is possible to make very substantial structural leaps with docking. However, it can be necessary to make use of multiple protein conformations and to flexibly adapt the pocket in order to identify a ligand and its bound pose correctly.

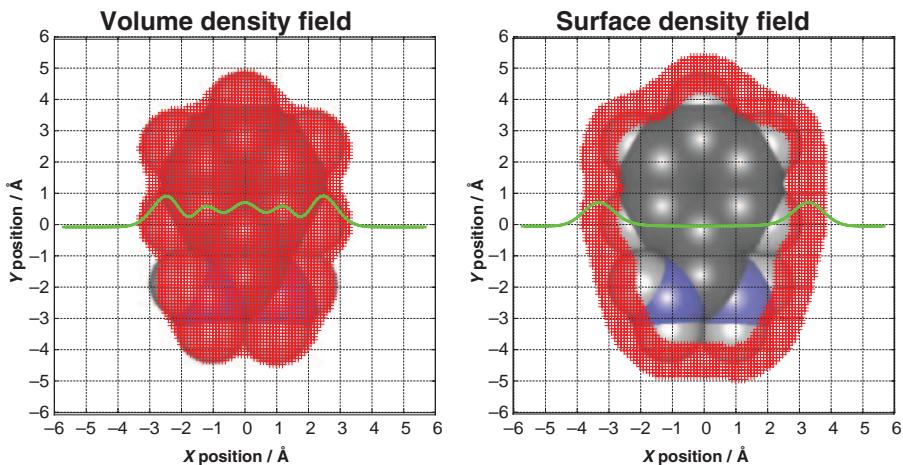


FIGURE 2.6 Volumetric and surface-based molecular density functions for benzamidine.

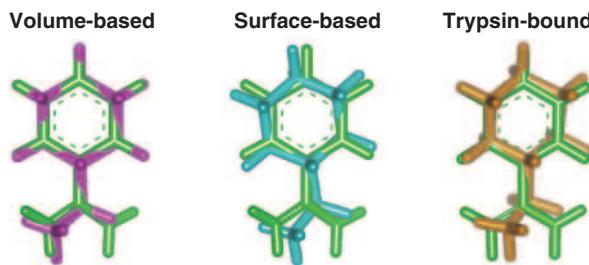


FIGURE 2.7 Volume- and surface-based alignment of aminomethylcyclohexane to benzamidine.

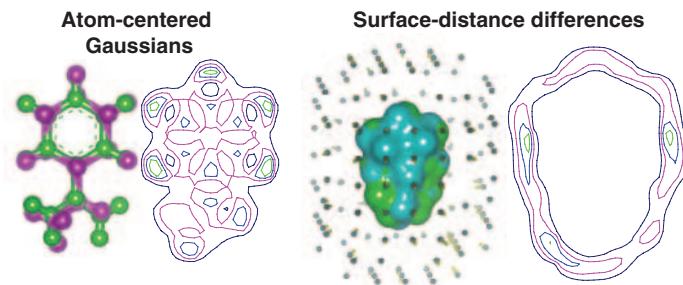


FIGURE 2.8 Relationship of molecular alignments to underlying similarity functions.

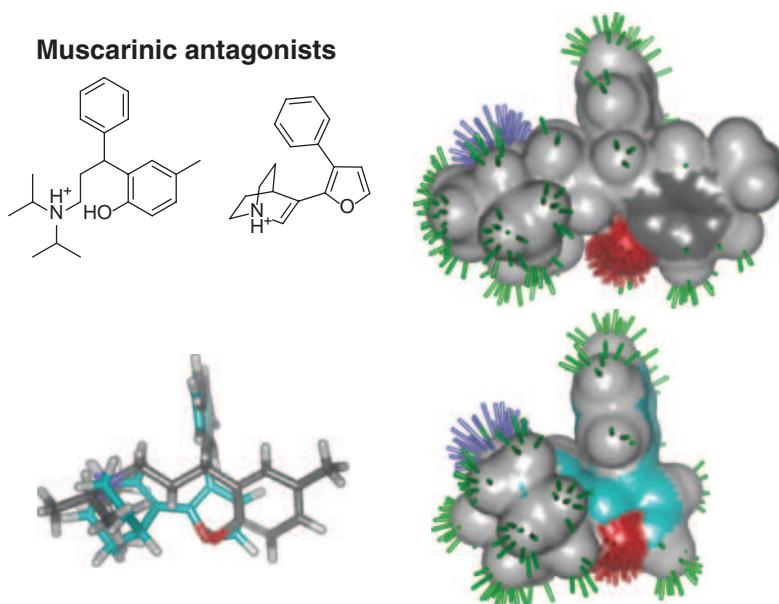
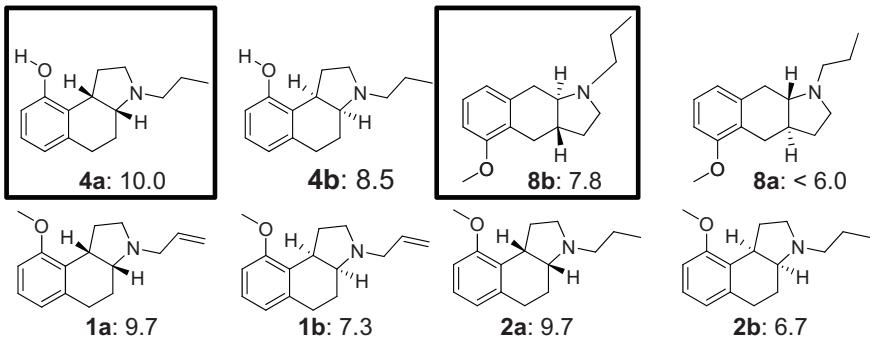
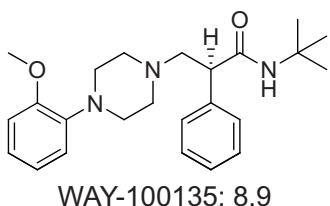


FIGURE 2.9 Optimal alignment of two muscarinic antagonists using surface shape and polarity.

Structure activity data



New ligand:
actual $pK_d = 8.9$



Induced physical pocket:
predicted $pK_d = 8.4$

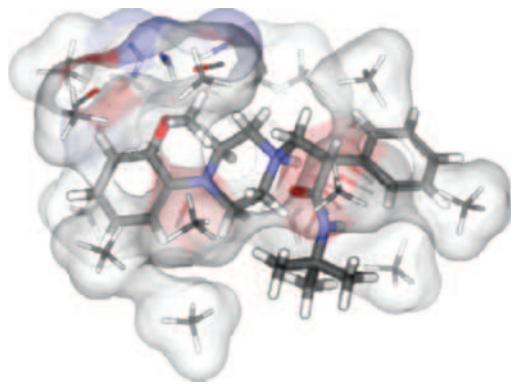


FIGURE 2.10 It is possible to construct a physically realistic model given structure-activity data. Such models can make accurate predictions even when the subject ligand is very different from any ligand on which the model was constructed.

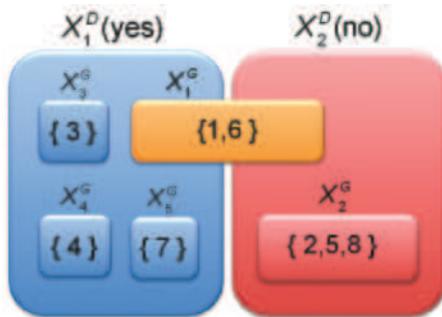


FIGURE 3.2 Schematic depiction of the indiscernibility classes of the example presented in Sections 3.2.3–3.2.5. The objects in the rectangles shaded in blue correspond to molecules that definitely induce phospholipidosis, while those rectangles shaded in blue correspond to molecules that definitely do not induce phospholipidosis. The two molecules within the orange rectangle correspond to molecules in the boundary set that may or may not induce phospholipidosis. Rules involving molecules within the indiscernibility classes depicted by the red and blue-shaded rectangles are deterministic, while those within the yellow-shaded rectangle are nondeterministic (probabilistic).

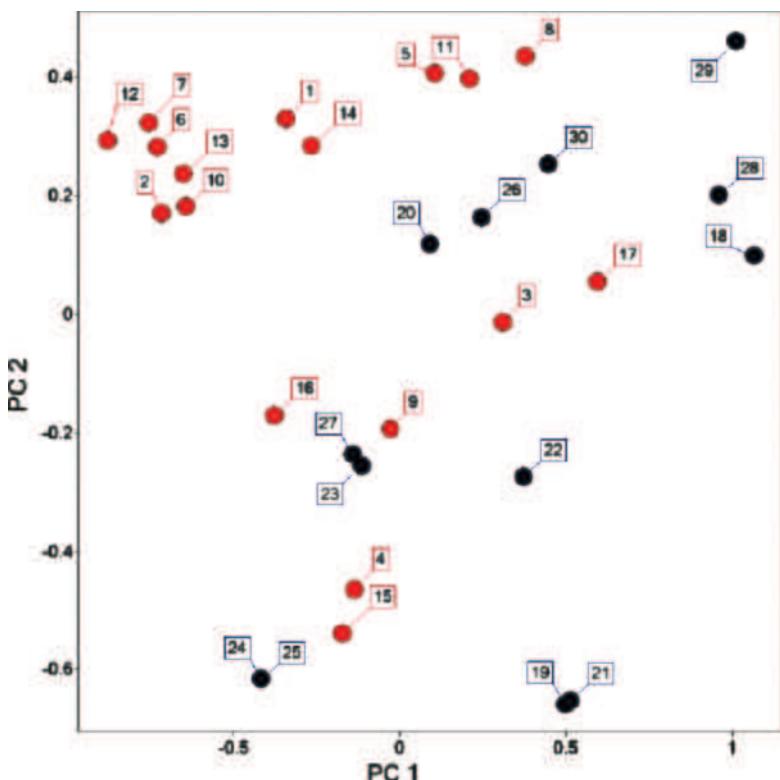


FIGURE 3.3 Two-dimensional representation of the chemical space distribution of the 30 drug molecules tested for phospholipidosis activity. The red filled circles indicate drugs that induce phospholipidosis; the blue filled circles represent drugs that do not induce phospholipidosis. The numbers indicate which specific drug (see Table 3.6) is associated with each data point.

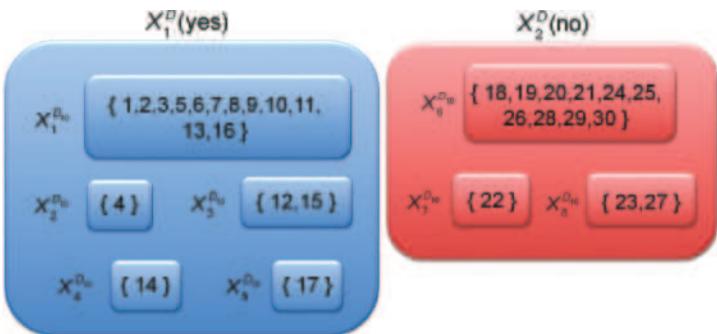


FIGURE 3.5 Schematic depiction of the indiscernibility classes induced by the D -reduct $\{gene-6, gene-10, gene-17\}$ obtained from the hepatoma HepG2 cell gene expression data set presented in Section 3.3.1. The objects in the rectangles shaded in blue correspond to molecules that definitely induce phospholipidosis, while those rectangles shaded in red correspond to molecules that definitely do not induce phospholipidosis. In this case the boundary set is null, so all of the C -indiscernibility classes are subsets of the indiscernibility classes induced by the decision attribute that indicates the presence or absence of phospholipidosis in the hepatoma HepG2 cells. Hence, all of the rules generated from this partitioning are deterministic.

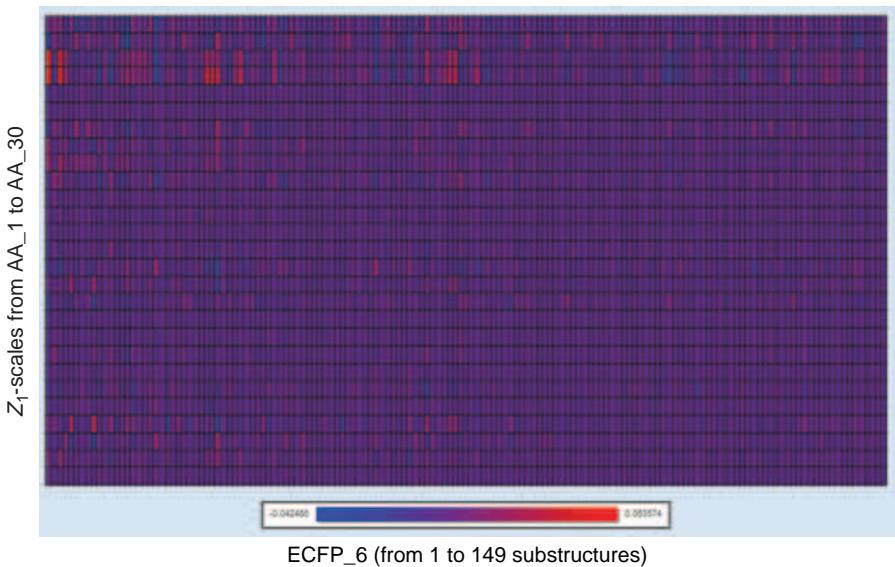


FIGURE 4.2 Heat map of z_1 -scales derived from the regression coefficient matrix.

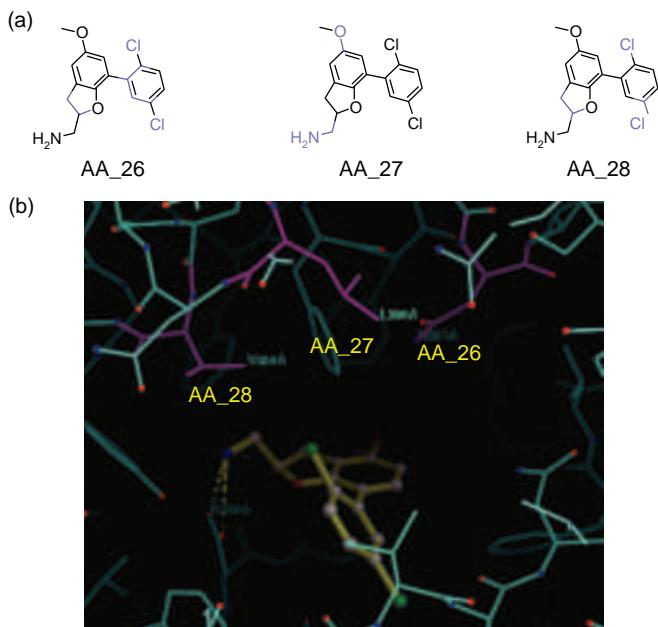


FIGURE 4.3 (a) (Left to right) Atom colorings of AA_26 (Asn), AA_27 (Leu), and AA_28 (Val) of GPCR inhibitor 3090943 docked into 5HT_{2C} GPCR. Strong hydrophobic interactions between each amino acid residue and the GPCR are highlighted in blue. (b) Putative docking mode of 3090943 in the homology model of 5HT_{2C} GPCR. Reprinted from Ref. [19].

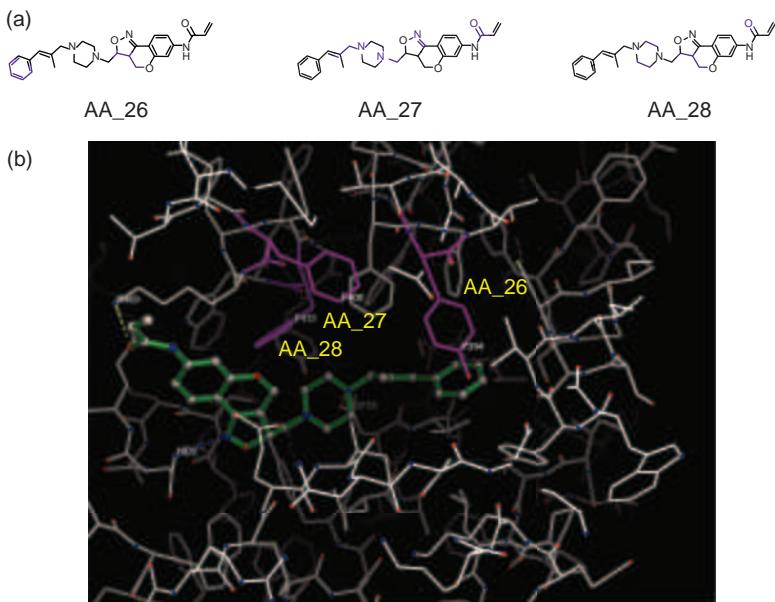


FIGURE 4.4 (a) (Left to right) Atom colorings of AA_26 (Tyr), AA_27 (Phe), and AA_28 (Phe) of GPCR inhibitor 3844318 docked into A_{2A} GPCR. Strong hydrophobic interactions between each amino acid residue and the GPCR are highlighted in blue. (b) Putative docking mode of 3844318 in the X-ray crystal structure of A_{2A} GPCR.

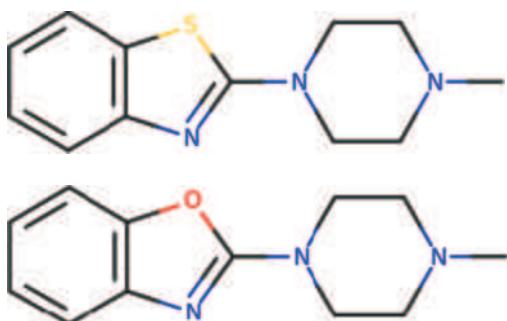


FIGURE 5.1 The structure on the left (2-(4-methylpiperazin-1-yl)-1,3-benzoxazole) is sterically and electrostatically similar to the structure on the right (2-(4-methylpiperazin-1-yl)-1,3-benzothiazole), yet the Daylight Tanimoto between the two structures is only 0.44, well below the threshold typically expected for molecules likely to have similar activity. As explored in the chapter, atoms that are contained in many paths lead to sensitivity to small changes, more so for smaller molecules.

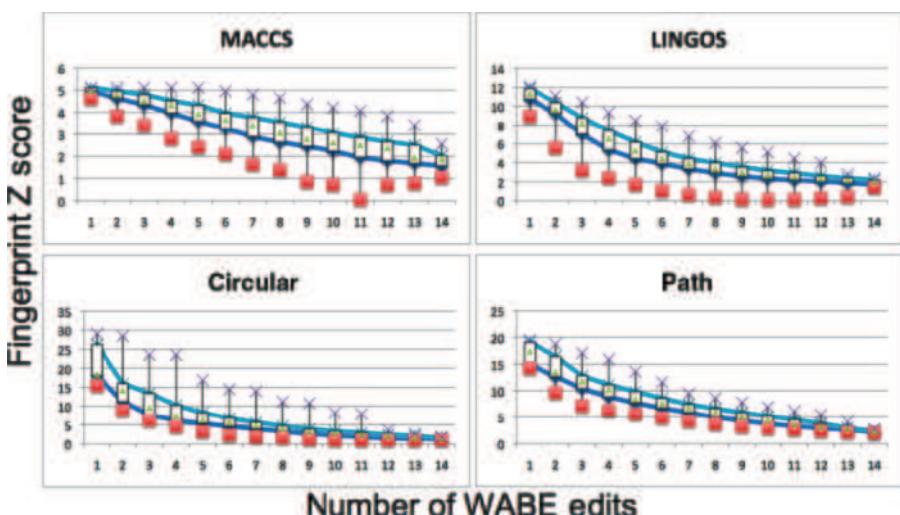


FIGURE 5.4 Graphs of the median, 95% range about the median, and maximum and minimum z -score similarity of the five methods as a function of the number of WABE edits.

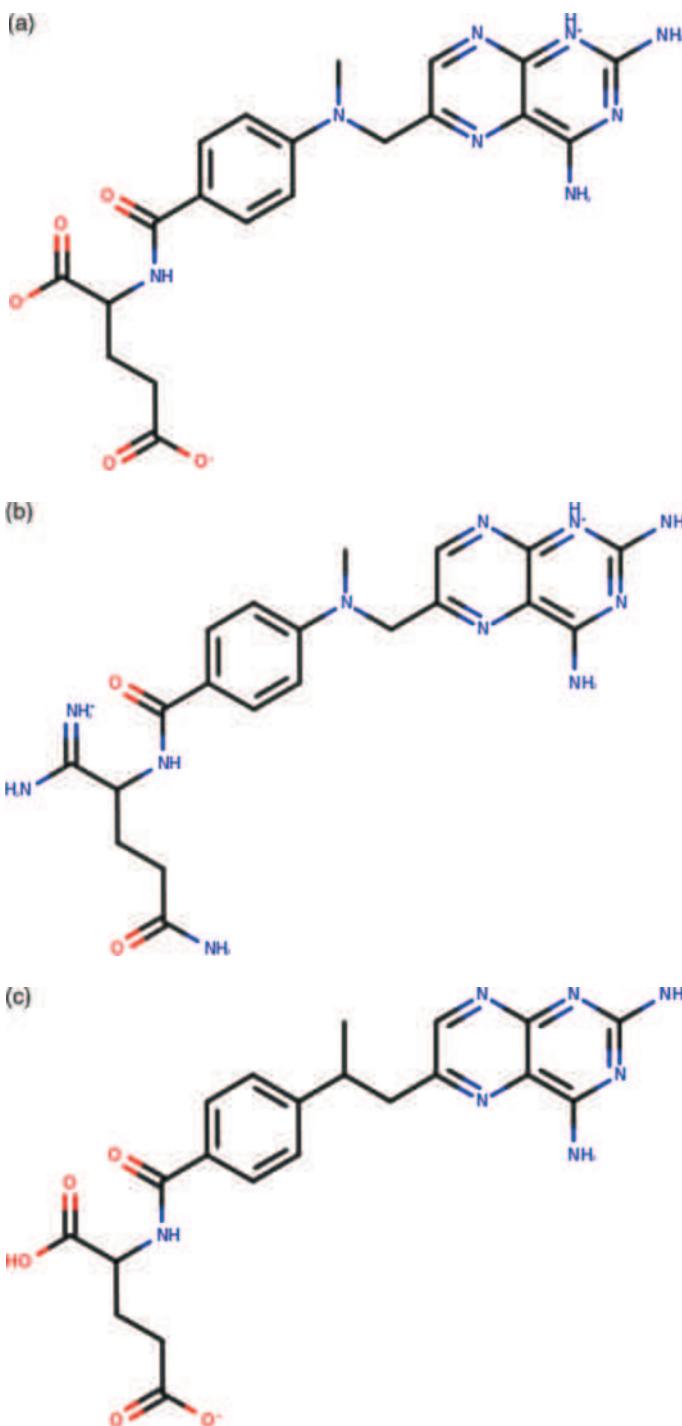


FIGURE 5.5 Structures of methotrexate (a) and the WABE variants most similar (b) and most dissimilar (c) by circular fingerprints when there are three WABE changes. Structure (c) appears very different because circular fingerprints are particularly sensitive to changes in linkers ($C \rightarrow N$) and rings ($N \rightarrow N^+$ in the pteridine ring).

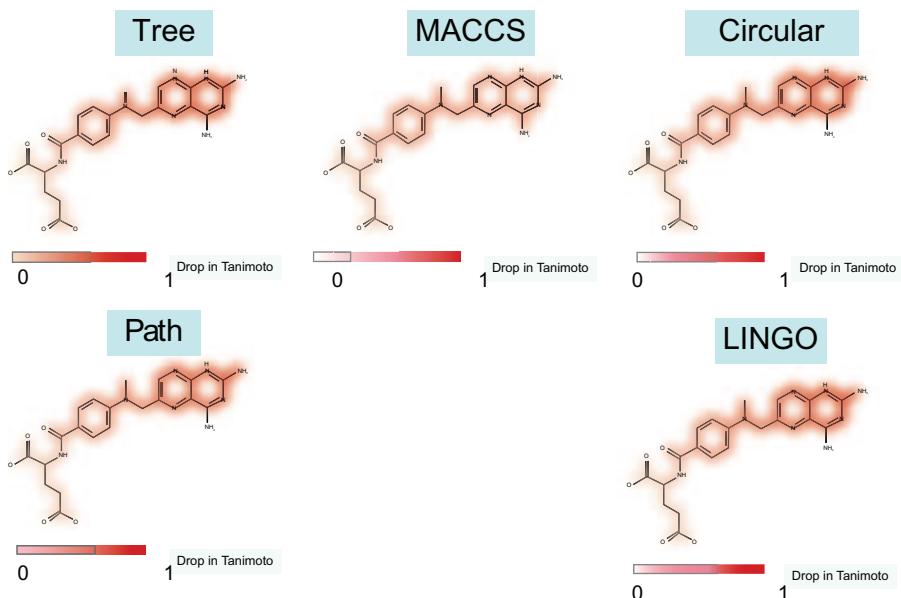


FIGURE 5.6 A graphical illustration of the similarity of methotrexate to a structure where a single atom has been changed (to a “dummy” atom). Atoms are colored more heavily the less the similarity to methotrexate. This figure illustrates many of the conclusions found in this chapter as to the sensitivity (or lack) of different fingerprint methods.

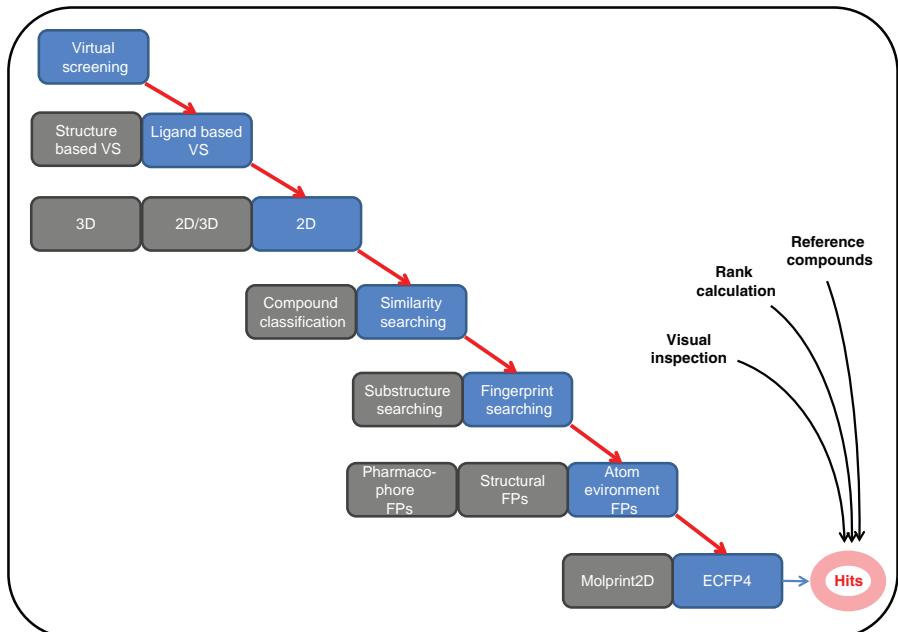


FIGURE 6.1 Exemplary workflow of a practical VS project. A possible VS application is illustrated. Different methodological choices can be made. Additional factors such as visual inspection, rank calculation, and the choice of reference compounds also strongly influence final compound selections.

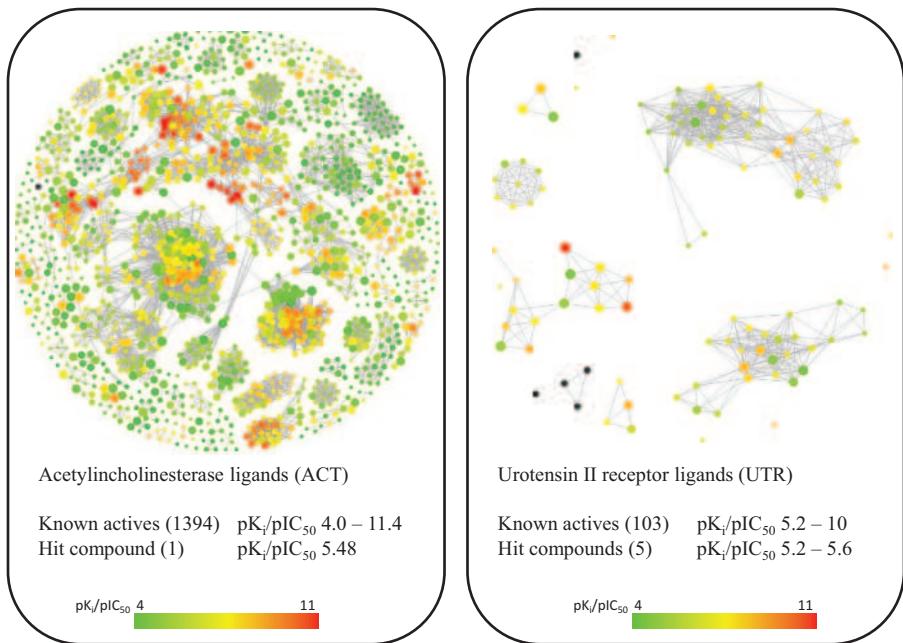


FIGURE 6.3 Network-like similarity graphs for known actives of two exemplary targets. Hits from two VS projects targeting either acetylcholinesterase (ACT) or the urotensin II receptor (UTR) are shown together with all active compounds known prior to VS. In network-like similarity graphs (NSGs), nodes represent compounds and edges pairwise similarity relationships (here fingerprint Tanimoto similarity above a given threshold value). Nodes are color-coded according to potency values using a continuous color spectrum from green (low potency) over yellow to red (high potency) and scaled in size according to compound SAR discontinuity scores. Black nodes represent hits.

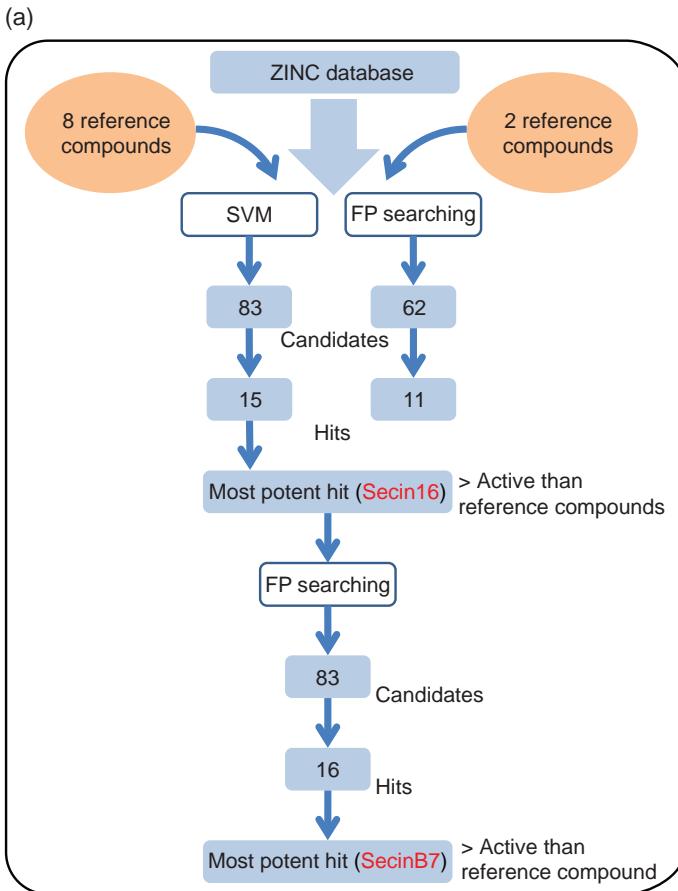


FIGURE 6.4 Virtual screening for cytohesin inhibitors. (a) VS workflow. Detailed description of the workflow leading to selected candidate compounds and experimentally confirmed hits.

(c)

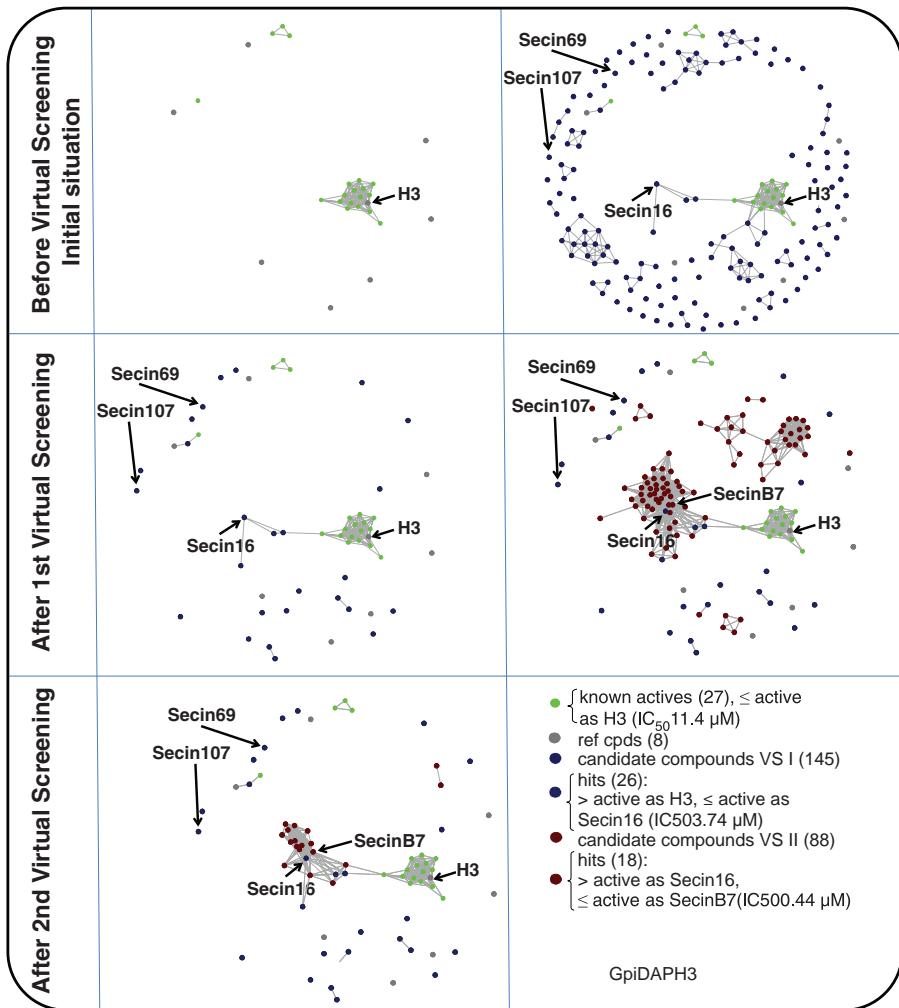


FIGURE 6.4 (c) NSGs capturing different stages of the cytohesin inhibitor VS projects. NSG representations are shown to capture active compounds and their similarity relationships prior VS and following the first and second round of VS. Similarity calculations were carried out using a 2D pharmacophore-type fingerprint termed GpiDAPH3 (Molecular Operating Environment, Chemical Computing Group, Montreal, Canada).

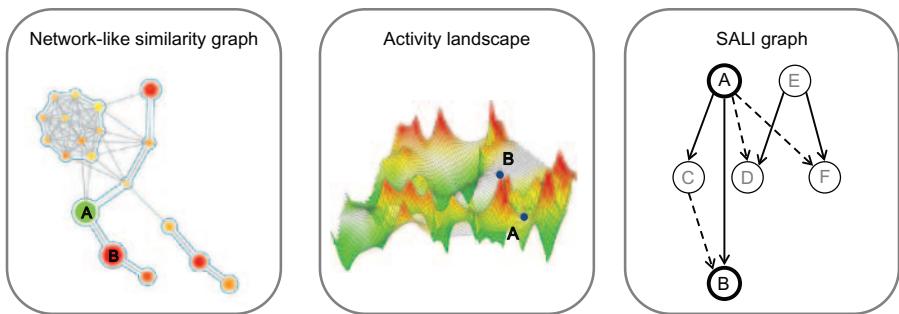


FIGURE 8.8 Three different visualizations of “activity cliffs” in a series of close analogs A–E. Compound A is mildly potent while B is highly potent, leading to discontinuity in local SAR. This is expressed by colors (red=high potency, green=low potency) and for the network-like similarity graph, size is used to show where activity cliffs occur. (Reprinted with permission from D. Stumpfe and J. Bajorath, *J. Med. Chem.* 2012, 55, 2932–2942, Figure 1).

Name	Structure	pIC ₅₀	Selectivity (log)	Solubility (µM)	HLM (%loss @ 40 min)	RLM (%loss @ 40 min)
1 XXX572		9.88	1.05	158	95.5	81.8
2 XXX518		5.76	0.67	148	4.83	38
3 XXX582		6.01	1.07	132	95.1	79.9
4 XXX295		6.25	0.99	146	83	77
5 XXX321		8	0.87	183	55.8	71.8
6 XXX509		6.16	1.13	197	25.6	64.8
7 XXX292		6.28	1.22	192	84	64
8 XXX313		6.8	1.14	880	71.4	53.5
9 XXX274		5.81	0.89	124	41.9	49.2
10 XXX025		5.89	0.71	138	54.2	77.8

FIGURE 8.9 Example of a simple “traffic light” display. Here, good property values are colored green, bad values are colored red, and intermediate values are yellow. But which compound is best?

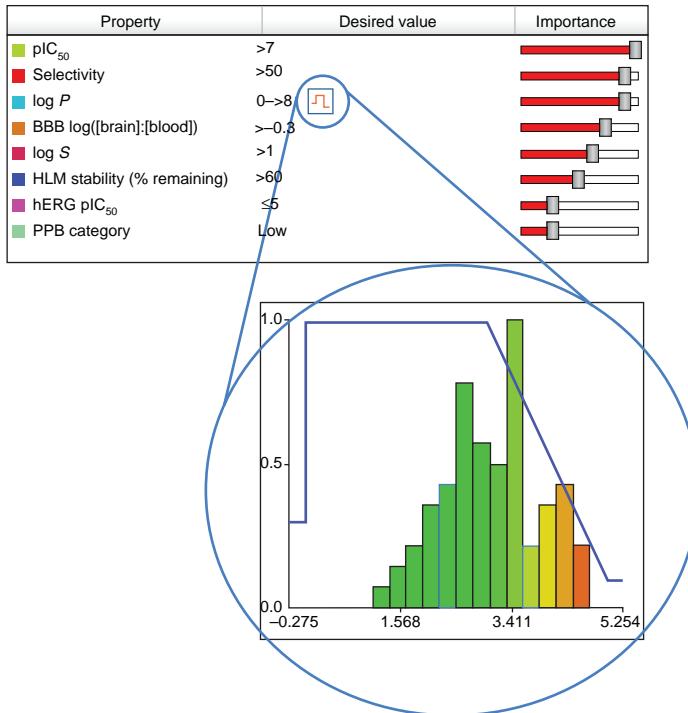


FIGURE 8.11 Example of a scoring profile. This defines the properties of interest, the ideal desired values, and relative importance of each property. Underlying each criterion is a desirability function, an example of which is shown for $\log P$.

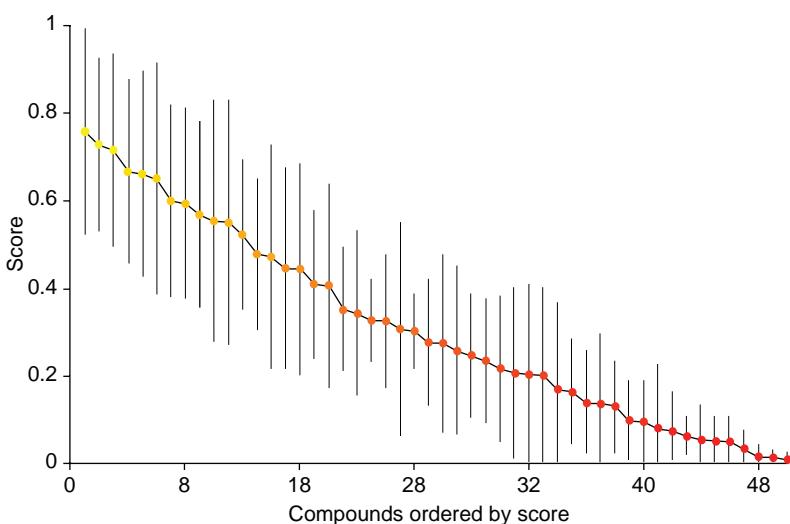


FIGURE 8.12 An example output from probabilistic scoring for 50 compounds. The compounds are ordered from left to right along the x -axis in order of their score and the overall score for each compound is plotted on the y -axis. The overall uncertainty in each score (1 standard deviation), due to the uncertainty in the underlying data, is shown by error bars around the corresponding point. In this case, the error bars of approximately the top 20 compounds overlap indicating qualitatively that these compounds cannot be confidently distinguished by the available data.

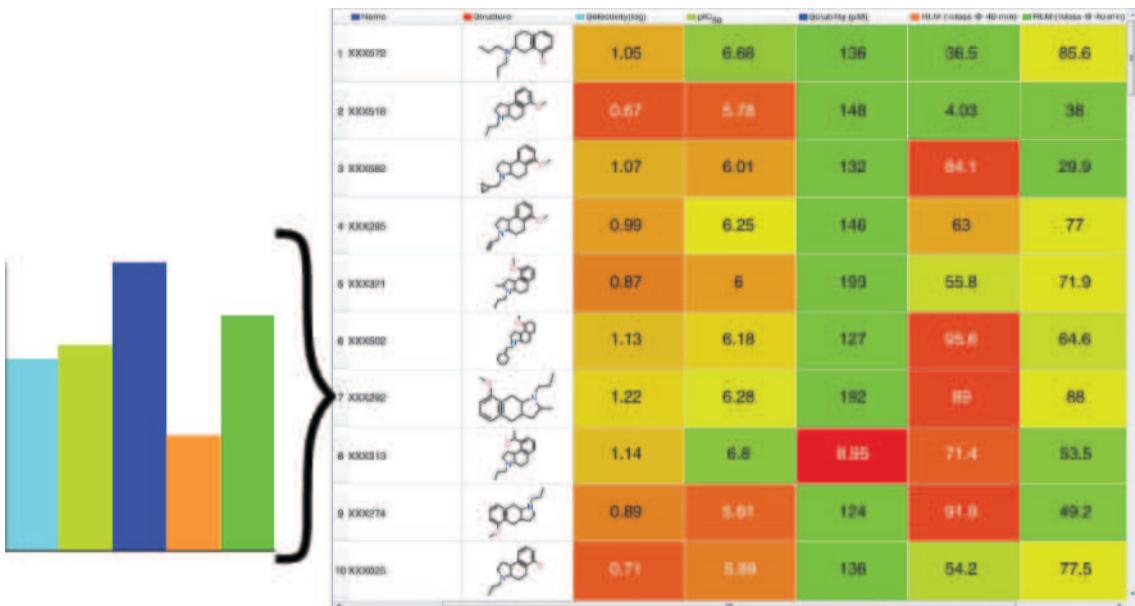


FIGURE 8.13 Examples of approaches to visualize the results from probabilistic scoring. The colors on the heat map on the right reflect not only the values of the data relative to the success criteria but also the importance of the property and the confidence of the outcome. A red cell indicated a poor result for an important property with high confidence. A green cell represents a good outcome for a property with high confidence. For comparison, this is the same data as shown in Figure 8.9. The contributions of each property to the score for a single compound can also be represented as a histogram, as illustrated to the left. Here the height of the bar reflects the impact of the property and, similarly, a low bar indicates that a significant issue has been identified with confidence; in this case, the orange bar indicates an issue which requires attention, corresponding to human liver microsomal stability (HLM).

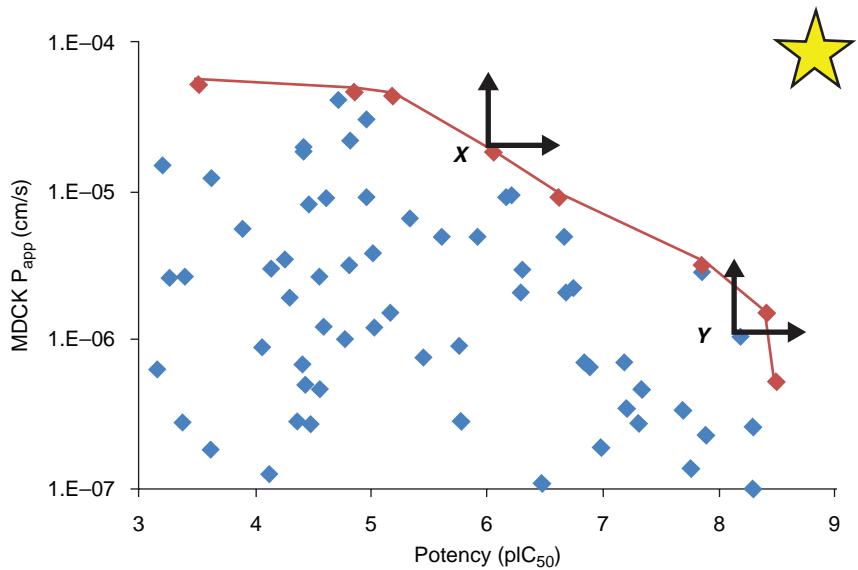


FIGURE 8.14 Illustration of Pareto optimal compounds for two-dimensional optimization of potency (pIC_{50}) and permeability ($\text{MDCK P}_{\text{app}}$). Each point represents the potency and permeability of a compound. The “ideal” compound would have both high potency and permeability as represented by the gold star. The red points, for example the point labeled X, are Pareto optimal, that is, there are no compounds better in both properties. The point labeled Y is not Pareto optimal, the point discussed earlier and to the right is better in both properties. The Pareto optimal compounds define the Pareto front, shown by the red line.

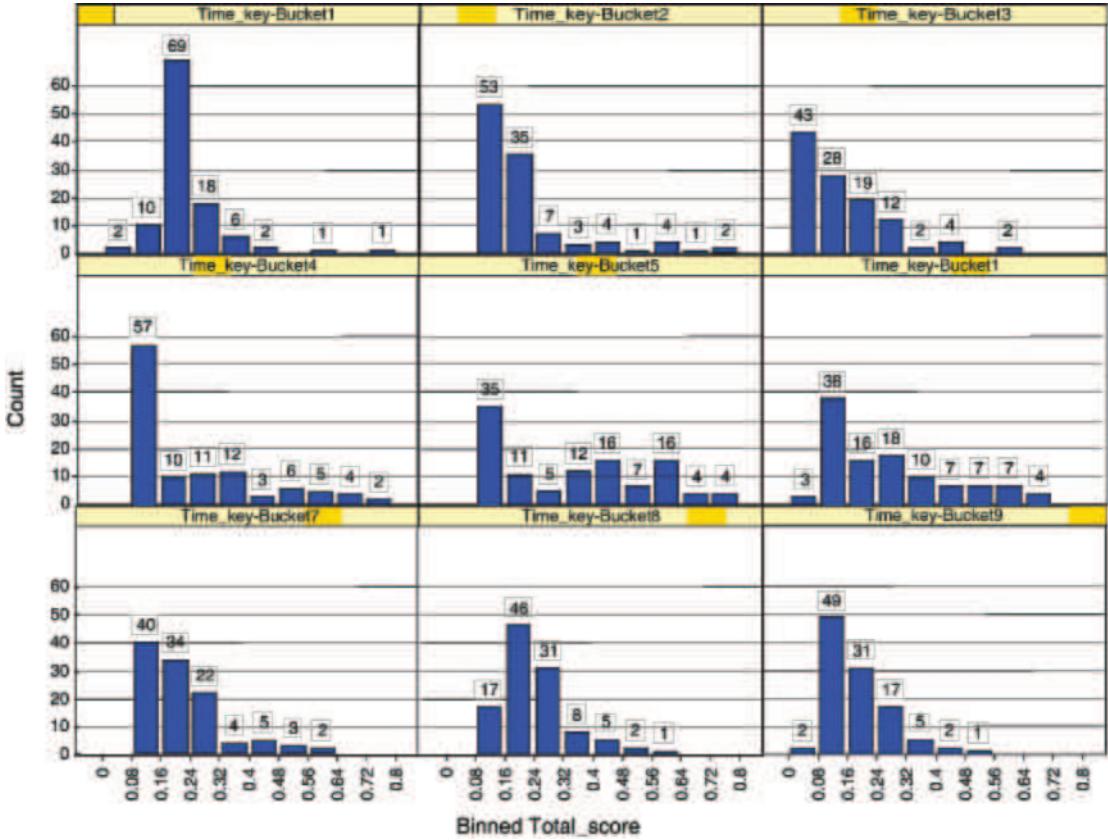


FIGURE 8.15 An example of the change in Pareto front during a lead optimization project. At each time step, the nondominated solutions are in Bin 0. It can be seen that after Time Bucket3 the project ceased to produce new molecules which were true improvements over those previously discovered. In this case, the injudicious application of a potency threshold caused the optimization to stall.

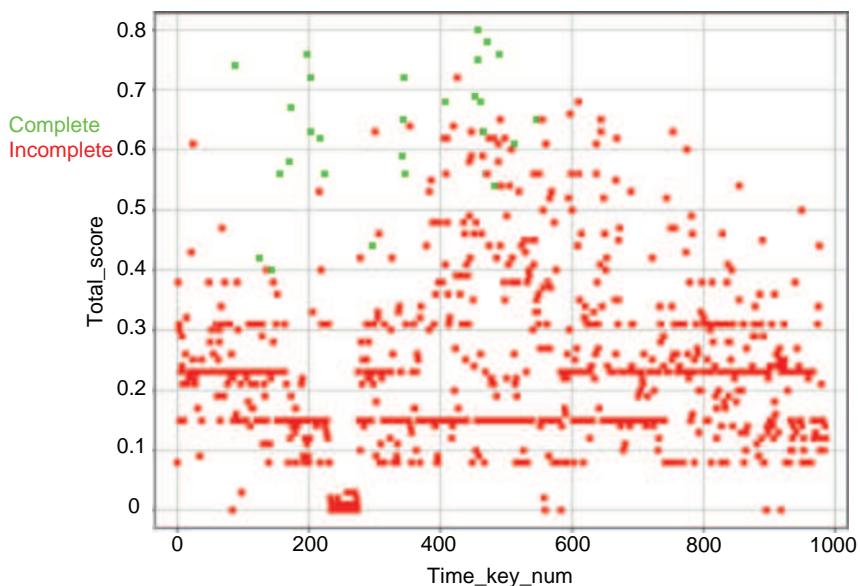


FIGURE 8.16 This figure illustrates the effect of setting too high a potency threshold for progression to other assays (e.g., ADMET evaluation). About halfway through the project, no further molecules generated a full molecular profile as they did not meet a high potency threshold. These molecules could have been superior in every other criteria but potency, and the project would never discover this key information.

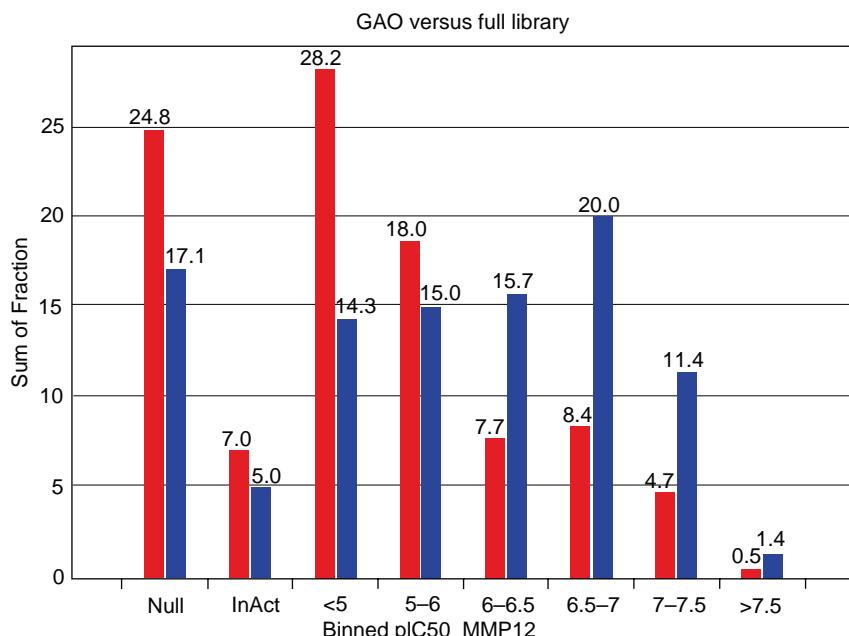


FIGURE 8.19 Proportion of compounds sampled by the GAO algorithm (blue) compared to the full dataset (red), binned by the primary assay. It can be seen that the algorithm samples compounds from all parts of the activity spectrum, but is very efficient at sampling the most potent compounds. Reprinted with permission from Pickett et al. [40], © 2011 American Chemical Society.

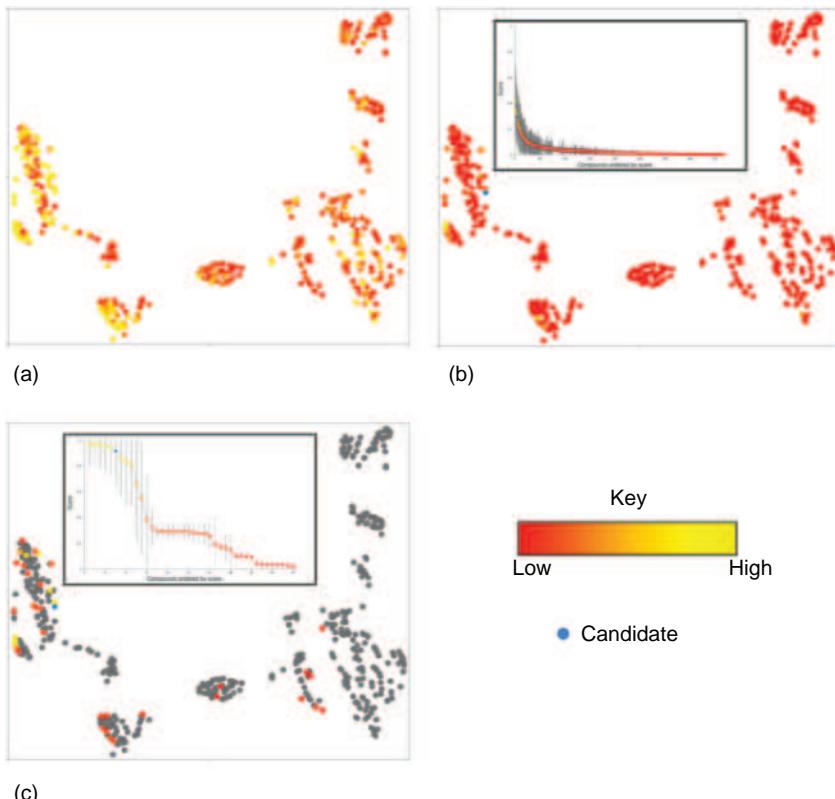


FIGURE 8.20 Chemical space plots illustrating the diversity of the library compounds screened for activity against the project’s therapeutic target. Each point represents a compound and the distance between points represents the structural similarity (defined by Tanimoto similarity of 2D fingerprints). The points in (a) are colored by activity (% inhibition) of each compound against the target, showing that active compounds were identified for a wide diversity of chemistry. The colors in (b) show the score of each compound against the profile shown in Figure 8.21a chosen to identify compounds with a good balance of experimental potency and predicted ADME properties. A plot of the scores is shown inset, indicating that only a small number of compounds, representing a small number of similar chemistries, are likely to achieve this desired profile. Finally, (c) shows the compounds selected for initial *in vitro* ADME studies, focusing on the chemistries most likely to have a good balance of properties. The compound scores, based on the *in vitro* data, against the profile shown in Figure 8.21b, are indicated by the colors and plotted in the inset graph. The compound ultimately selected as the development candidate is highlighted in blue in plots (b) and (c).

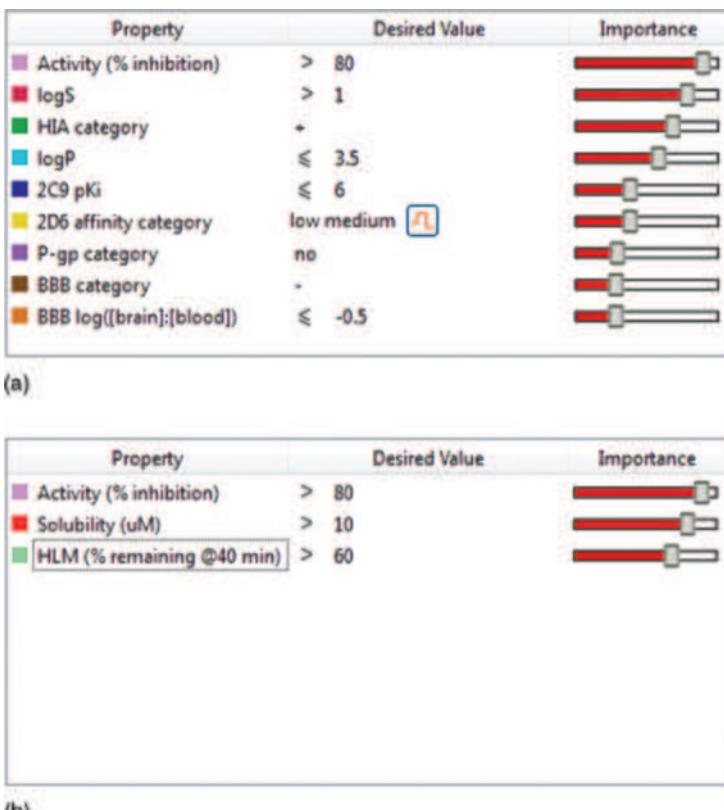


FIGURE 8.21 Scoring profiles used for prioritization of compounds intended for oral dosing against a cardiovascular target. The profile shown in (a) combines the experimentally measured target activity (as a percentage inhibition) with *in silico* predictions of solubility ($\log \mu\text{M}$), human intestinal absorption (HIA), $\log P$, inhibition of cytochrome P450s CYP2D6 and CYP2C9, active transport by P-gp, and blood–brain–barrier penetration (BBB). The profile shown in (b) combines the experimental activity with the primary *in vitro* ADME assay results for solubility in micrometer and human liver microsomal (HLM) stability measured as percentage remaining after a 40 min incubation.

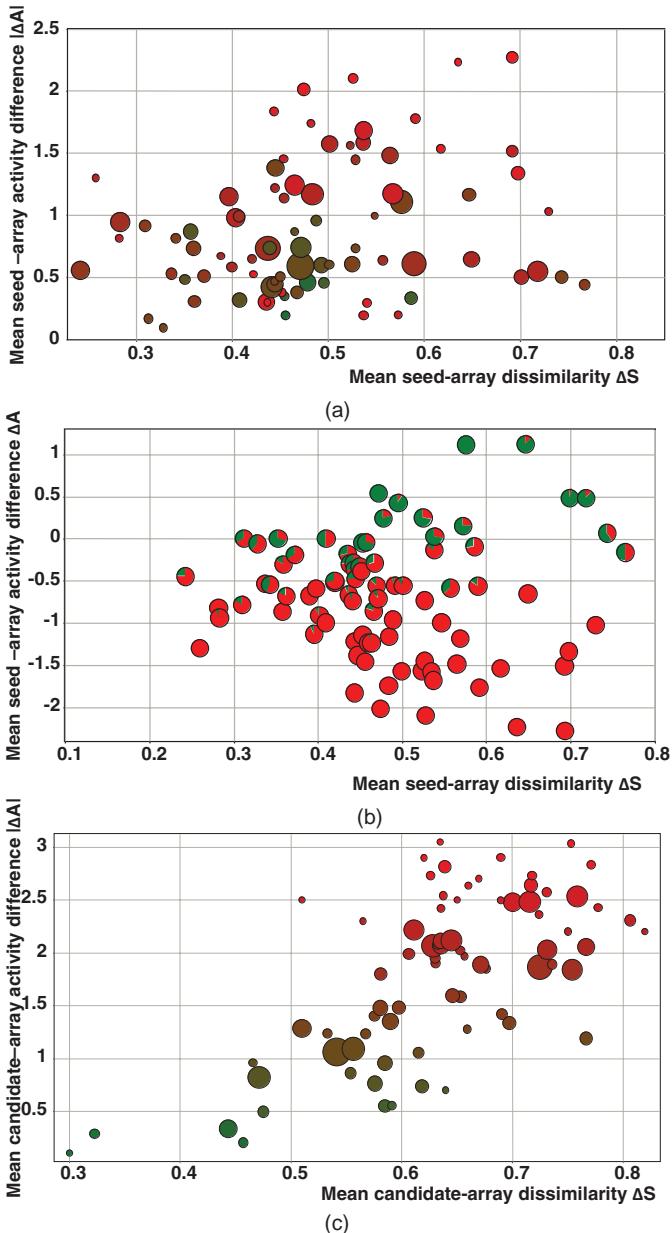
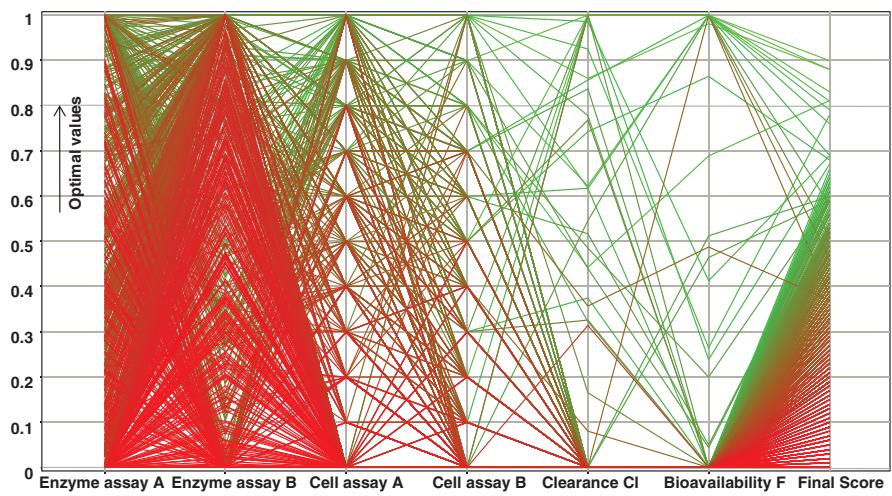
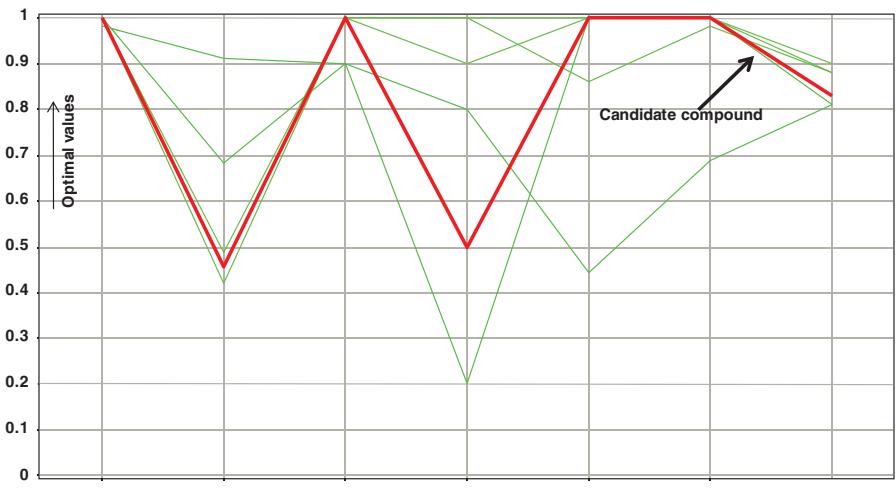


FIGURE 9.3 Each data point represents an array. In (a) and (c), the circles are color-coded by potency pIC_{50} value, ranging from red ($pIC_{50}=5$) to green ($pIC_{50}=8.6$) and the size of the circles corresponds to the size of the array (3–45 members). In (a) and (b), the x - and y -axes denote the average dissimilarity ΔS and absolute average property distance ΔA between the seed and each array member, respectively. In (b), the pies illustrate in green the proportion of the array molecules that have better potency than the seed. In (c), the x - and y -axes denote the average dissimilarity ΔS and absolute average property distance ΔA between Project A's proposed drug candidate ($pIC_{50}=8.2$) and each array member, respectively.

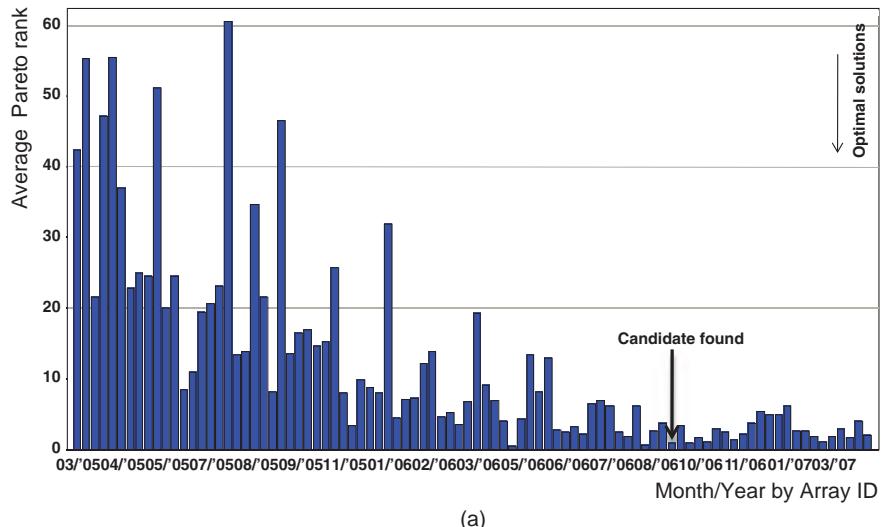


(a)

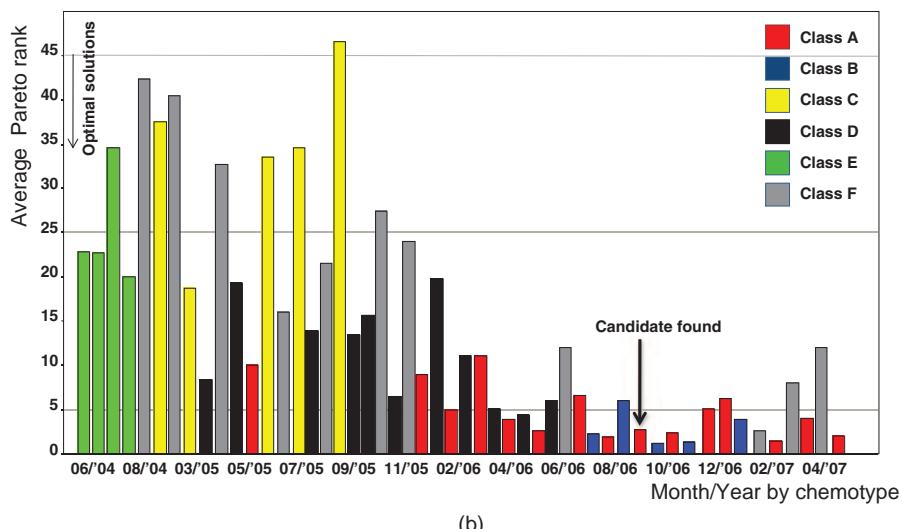


(b)

FIGURE 9.4 (a) Parallel coordinates plot of Project A compounds' six most important experimental properties (normalized from 0 to 1) and weighted desirability score. Higher scores are depicted in green. (b) Parallel coordinates plot of the six compounds having a weighted desirability score above 0.8. The actual candidate of Project A is highlighted in red and has a score of 0.83.

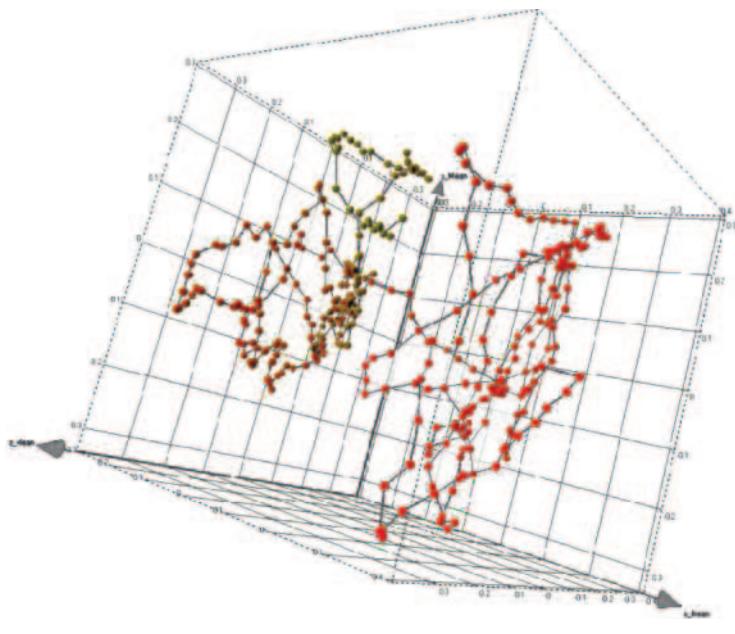


(a)

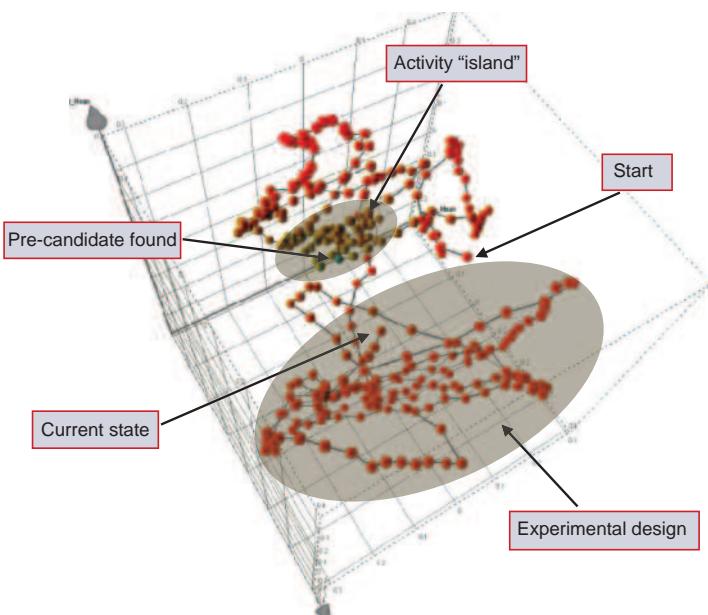


(b)

FIGURE 9.6 (a) Average Pareto rank binned by chronologically ordered chemical arrays. (b) Average Pareto rank binned by date and color-coded by chemotype.



(a)



(b)

FIGURE 9.7 3D SAW-like trajectories revealing the chronological progress of a lead optimization project in chemical space. Every point is the average of the x , y , and z coordinates of 10 chronologically ordered molecules. Chronologically successive points are connected by a line to reveal the trajectory in 3D chemical space. In (a) points are color-coded by their corresponding average potency value. Red indicates weak activity, whereas green indicates strong activity against the project’s primary assay. In (b) the plot is annotated by project-specific milestones such as the start, finish, pre-candidate discovery (shown in light green color), and change in array chemistry approach (gray shaded). An identified activity “island” is also highlighted in gray.

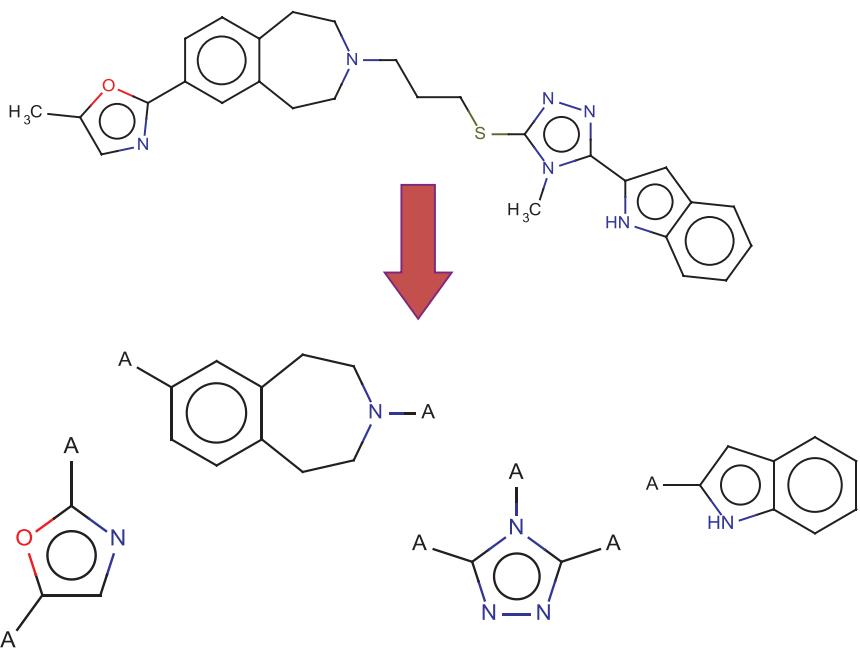


FIGURE 9.8 The input molecule is split in ring assemblies (i.e., contiguous ring systems). Side chains are marked by “A.”

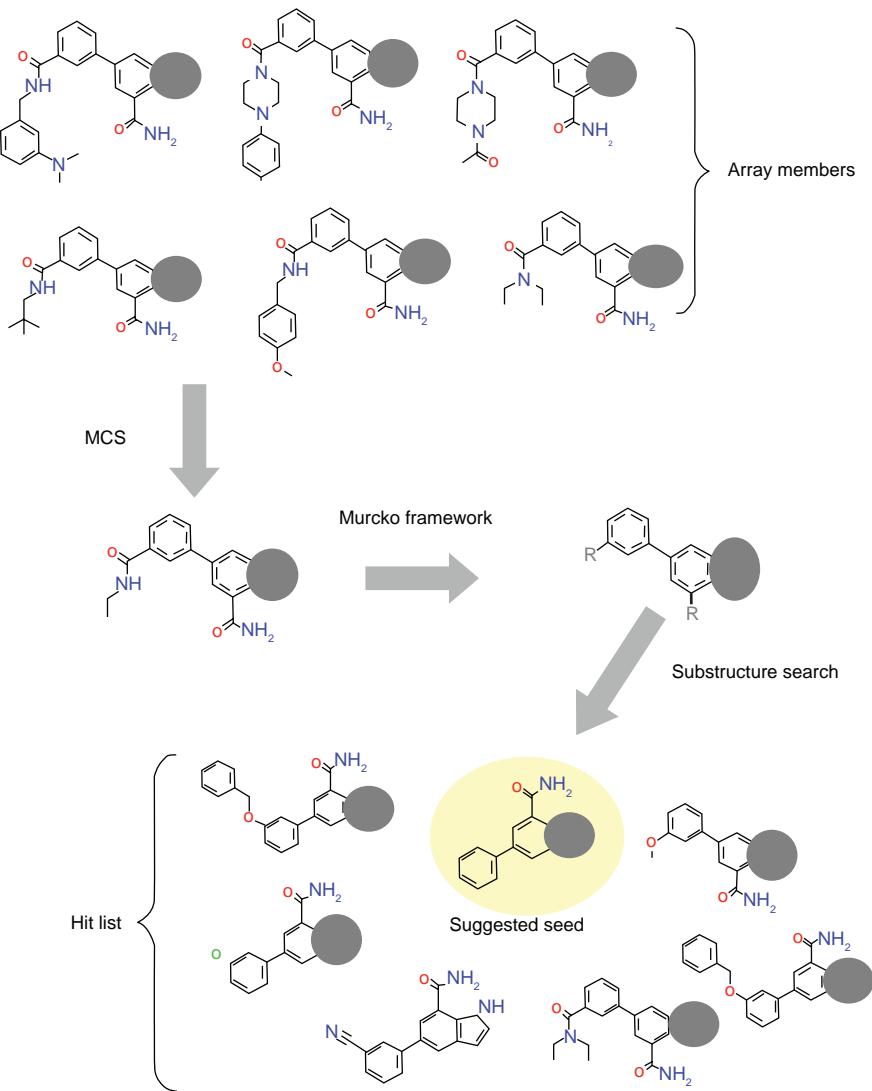


FIGURE 9.9 The seed detector workflow applied to real lead optimization data. The algorithm seeks to identify the seed of each array by applying the seed-likeness criteria to all the relevant molecules. For legal reasons, the molecules depicted in this example have a part of their structure covered.

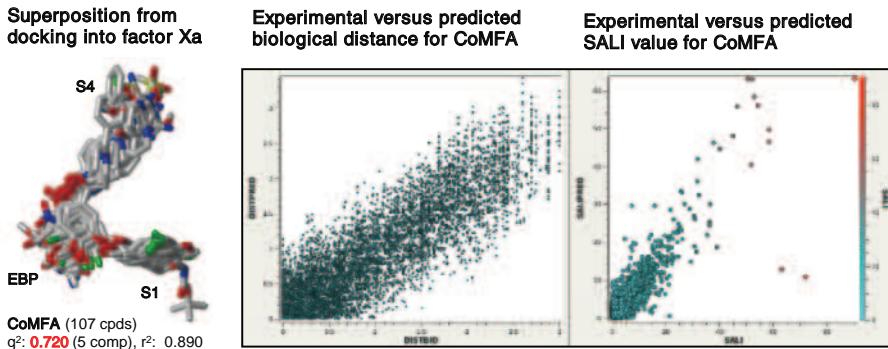


FIGURE 10.1 Model diagnosis plots for 107 3-oxybenzamides as factor Xa inhibitors [58, 59]. Left: Alignment from docking into factor Xa binding site. Middle: Experimental versus predicted biological differences (ΔpK_i) for molecular pairs above a similarity threshold for a CoMFA model (q^2 0.720, r^2 of 0.890, 5 components). Right: Experimental versus predicted SALI values for CoMFA model predictions; informative pairs with high SALI values indicated in red.

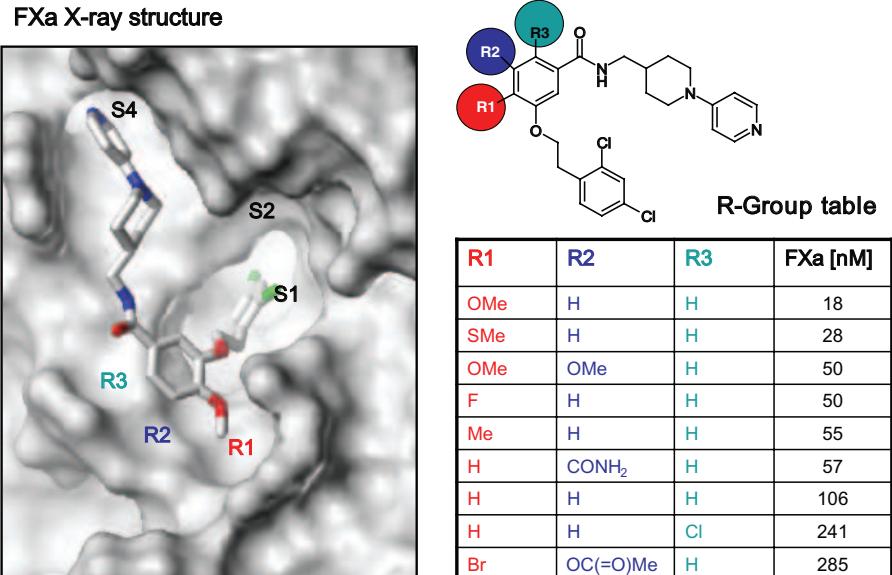


FIGURE 10.3 R-group plot for 3-oxybenzamides as factor Xa inhibitors [58, 59] (right) with binding pose from X-ray crystallography (PDB 2BMG, resolution 2.7 Å).

SaliExplorer Heatmap

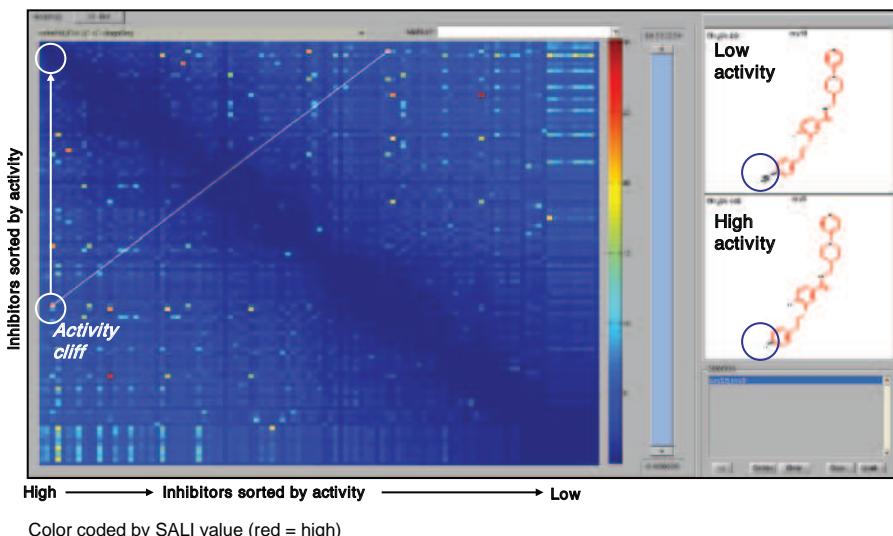
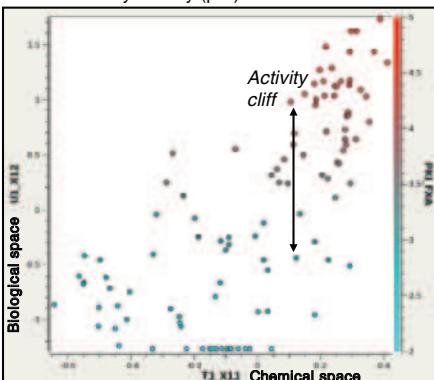


FIGURE 10.4 SaliExplorer heatmap color coded by experimental SALI value (red = high, blue = low) for 107 3-oxybenzamides as factor Xa inhibitors [58, 59]. Inhibitors are sorted by activity on both x - and y -axes; clicking on a point displays the molecular pair.

PLS t-u plot

Points represent single molecules
Color coded by activity (pKi)



Neighborhood plot

Points represent pairs of molecules
Color coded by SALI value

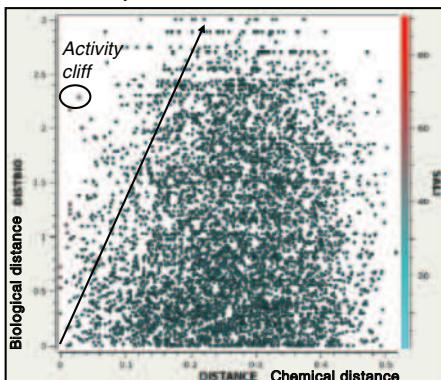


FIGURE 10.5 Left: PLS t - u plot of the first PLS component for the 3-oxybenzamide-based CoMFA model; chemical similarity is plotted on the x -axis versus biological similarity on the y -axis. The vertical arrow indicates an activity cliff with similar coordinates for the chemical space and different coordinates for the biological space. Right: Neighborhood plot for all $n \times (n - 1)/2$ pairs of 107 3-oxybenzamides as factor Xa inhibitors [58, 59] color coded by SALI values (red = high); chemical similarity on the x -axis with most similar compounds on the left is correlated to biological differences on the y -axis. Compounds with highest SALI values are located in the upper left triangle with respect to the arrow defining the neighborhood radius.

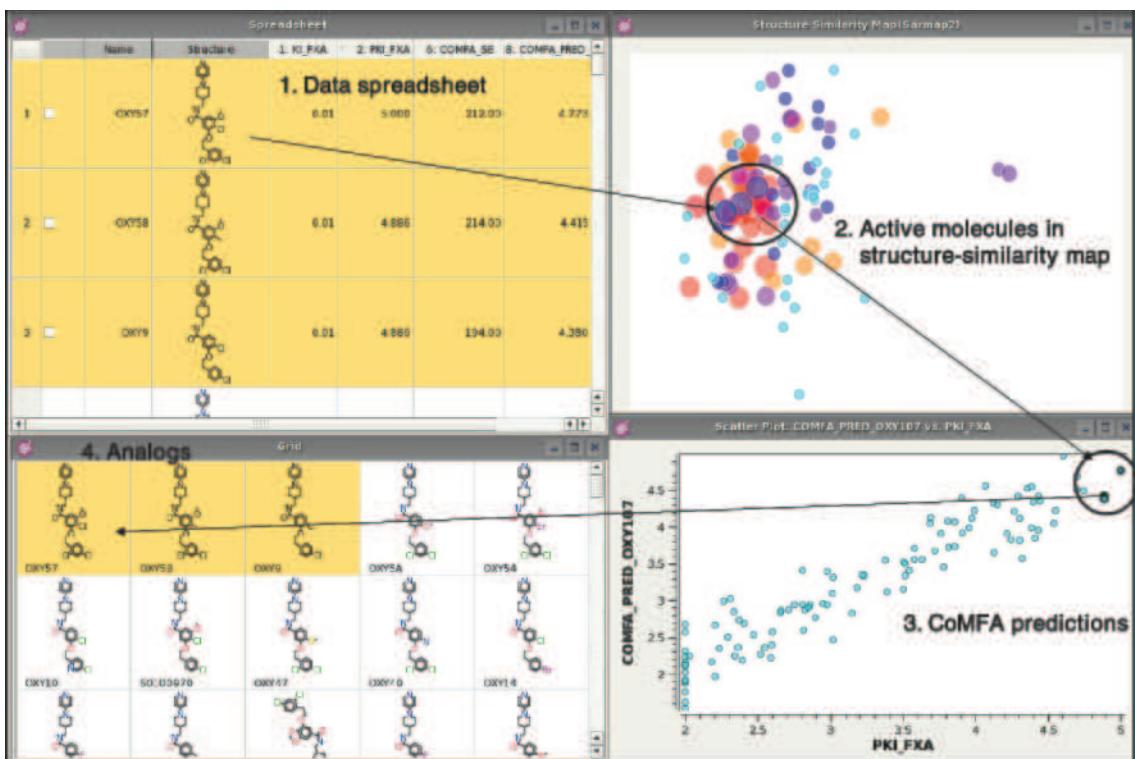


FIGURE 10.6 Interactive exploration of structure-similarity map for factor Xa inhibitors [58, 59]. From the data table (upper left) a structure-similarity map colored and sized by activity (large and red = high pKi) is derived. Interesting regions can be interactively explored and highlighted in a scatter plot with experimental versus predicted activities for a CoMFA model of 107 3-oxybenzamides (lower right), which also allows to display analogs in a 2D-grid plot (lower left).

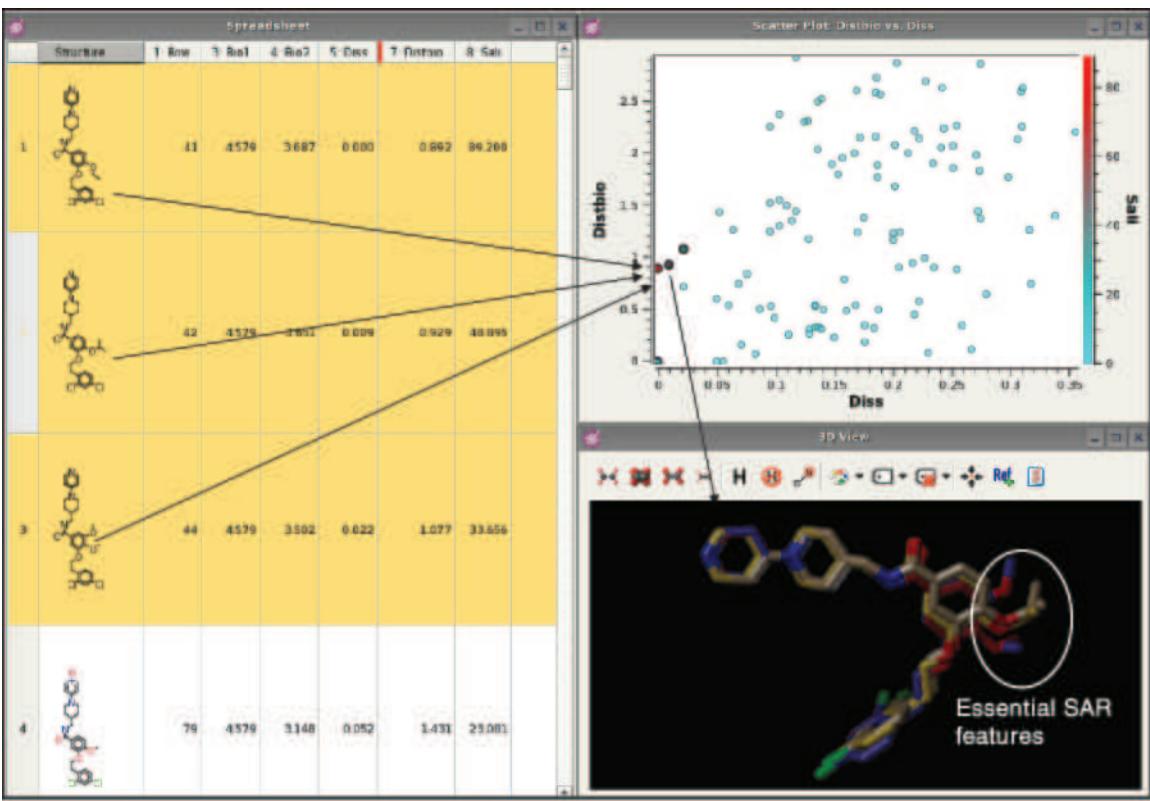


FIGURE 10.7 Local neighborhood plot (upper right) for 3-oxybenzamides as factor Xa inhibitors [58, 59] and the compound in table row 1 (left) as query. Pairs related to the query are displayed with chemical similarity given on the x-axis versus biological difference on the y-axis (upper right), color coded by SALI values (high=red). Interactive exploration by a 2D-table view (left) and a 3D-window (lower right) allows focusing on essential SAR features for activity.

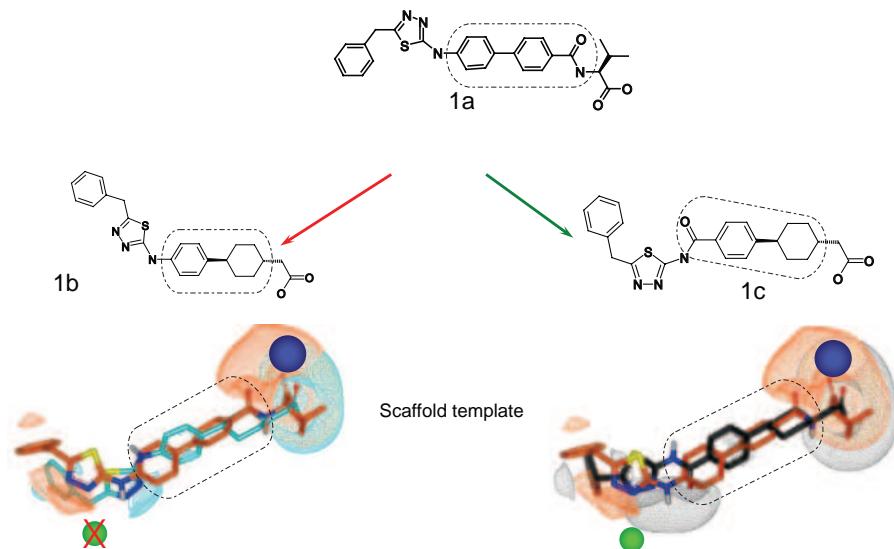


FIGURE 10.8 Rescaffolding of a thiadiazole series of DGAT1 inhibitors using shape-based alignment. Scaffold template elements are illustrated by a dash-dotted line. Replacement of the biphenylbenzoylamide scaffold **1a** by benzoylcyclohexyl **1c** leads to a bioactive compound with full pharmacophoric match, including both the negative (blue) and the acceptor features (green). The replacement by phenylcyclohexyl **1b** leads to a non-bioactive compound with only partial pharmacophoric match of the thiadiazole nitrogen acceptor (green).

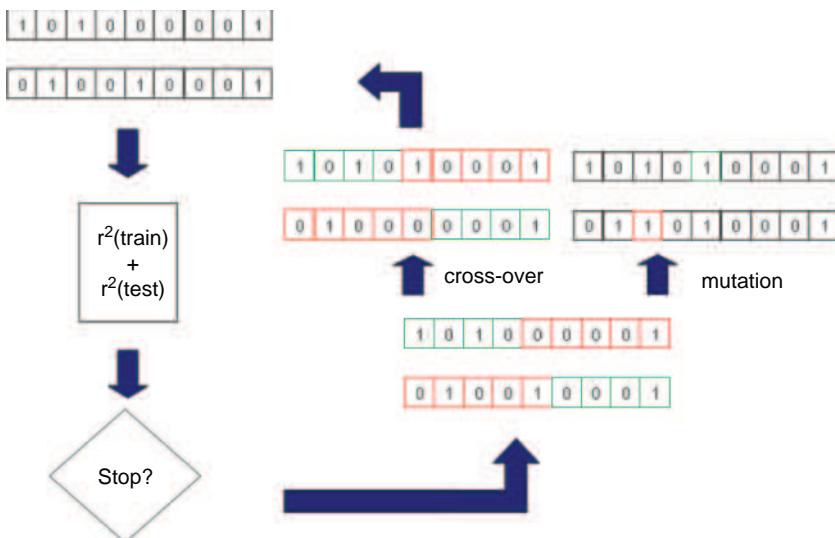


FIGURE 11.1 Workflow of the genetic algorithm used for feature selection to reduce the number of molecular descriptors for model building. The bit string encodes whether a descriptor is used in model building. The stop criterion is derived from the goodness of the fit between predicted and experimental values for the model.

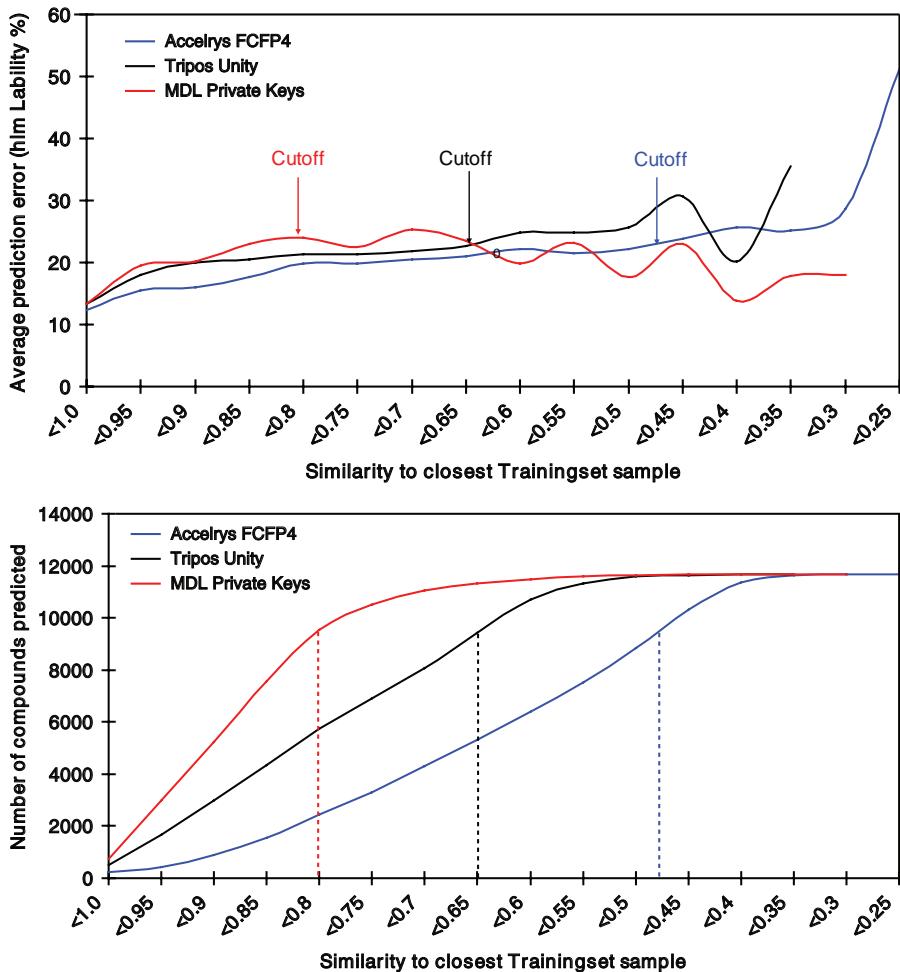


FIGURE 11.2 Analysis of the prediction error depending on the similarity of predicted compound to the closest compound in the training set. The top panel shows the prediction error, the bottom panel plots the number of compounds, which can be predicted at the respective similarity threshold.

(b)

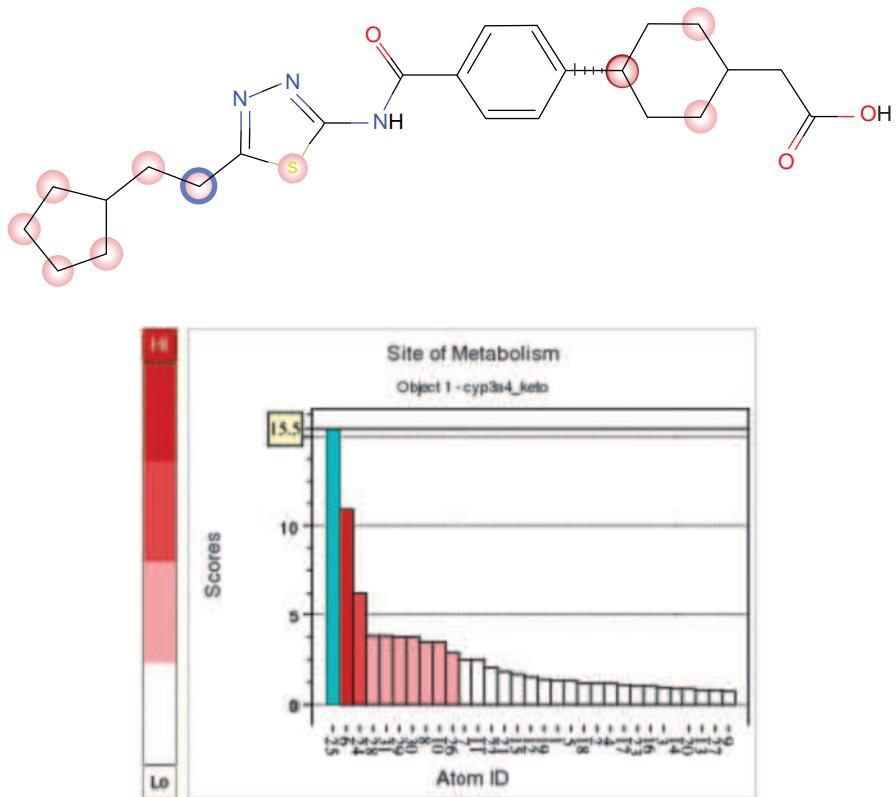


FIGURE 11.5 (b) Analysis of compound 4 with MetaSite3 [99]. The proposed main metabolite is the aliphatic hydroxylation product at the SOM. The scores plot ranks the thiadiazole alpha position on top of all substrate sites (blue bar). Most probable SOM are color coded in a structure diagram (blue circle).

(c)

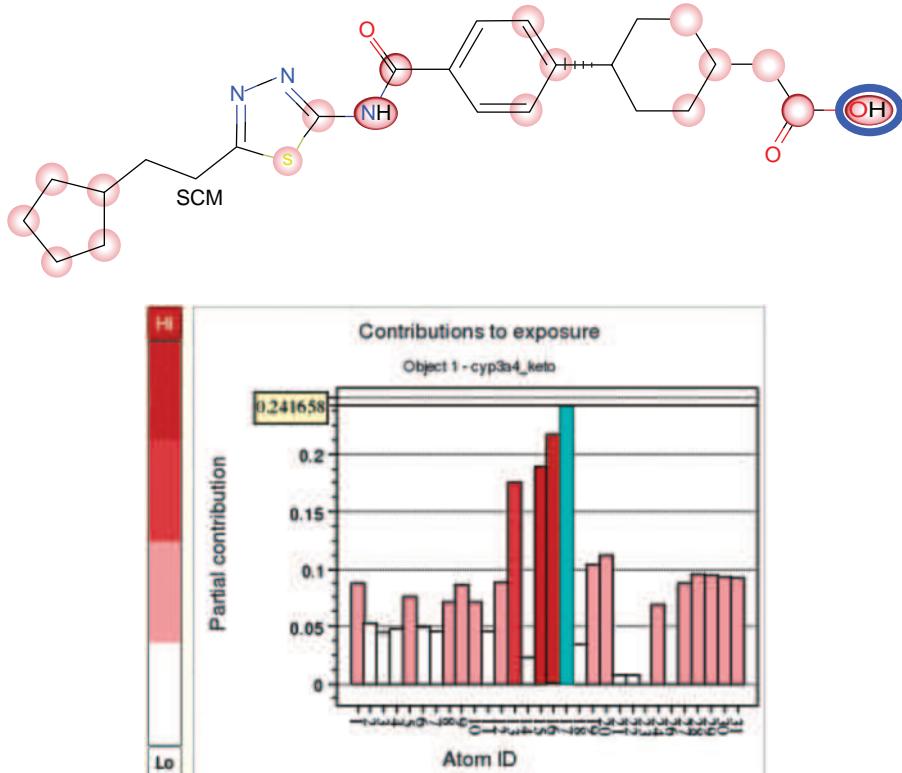


FIGURE 11.5 (c) Depending on the SOM, atomic contributions to the substrate recognition are analyzed in a second graph. A dominant contribution to the orientation of compound 4 is made by the top-ranked carboxylic acid function (blue circle) and the second-ranked amide function.

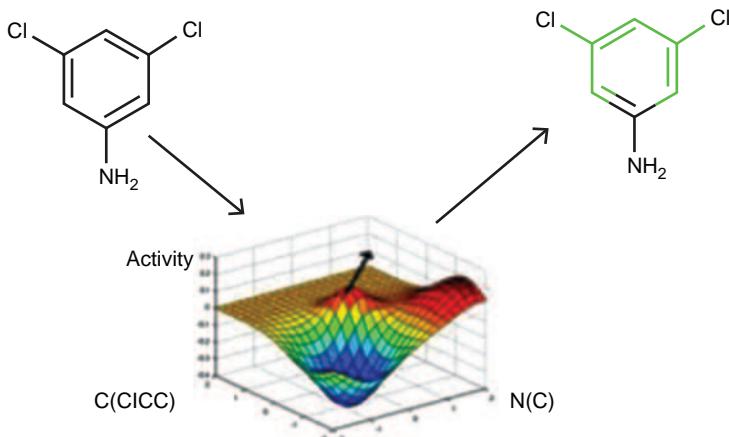


FIGURE 12.1 Schematic representation of how the descriptor influence on each prediction can be extracted from the models response surface. The most significant descriptor fragment can then easily be highlighted in the structure for visual inspection.

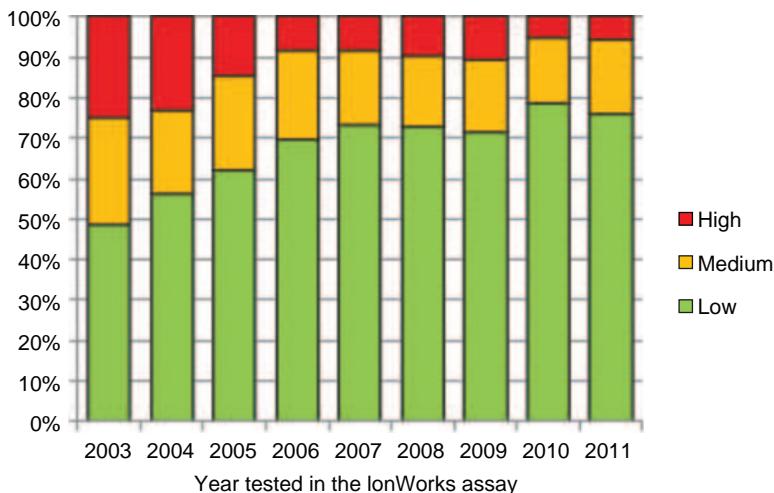


FIGURE 12.2 Histogram of compounds tested for hERG activity in the IonWorks assay from 2003 to 2011, categorized as having “low” $>30\text{ }\mu\text{M}$ (green), “medium” 3–30 μM (amber), or “high” $<3\text{ }\mu\text{M}$ (red) activity.

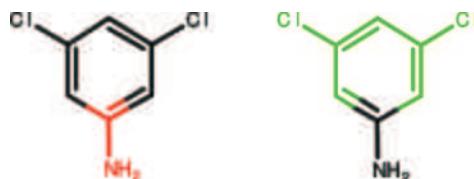


FIGURE 12.3 A general aromatic amine alert highlighted in red on the left. The amine is in general correlated with positive activity in the Ames test. However, when combined with the chlorine substituents in the meta positions, which are correlated with inactivity in the Ames test, the overall assessment is that the compound would be inactive.

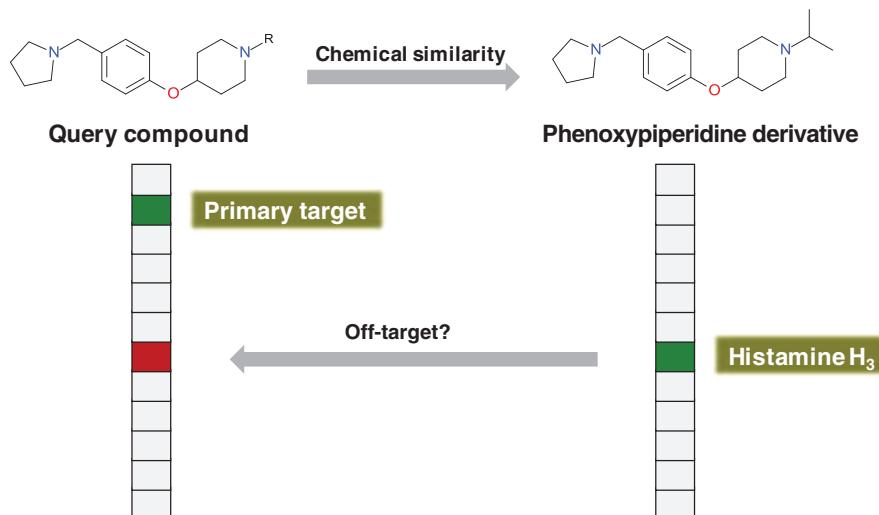


FIGURE 12.4 Concept of predictive secondary pharmacology exemplified with a query compound in the upper left corner. The phenoxy piperidine to the right was one compound identified by the similarity search. The vertical bars denote the compound specific target profiles.

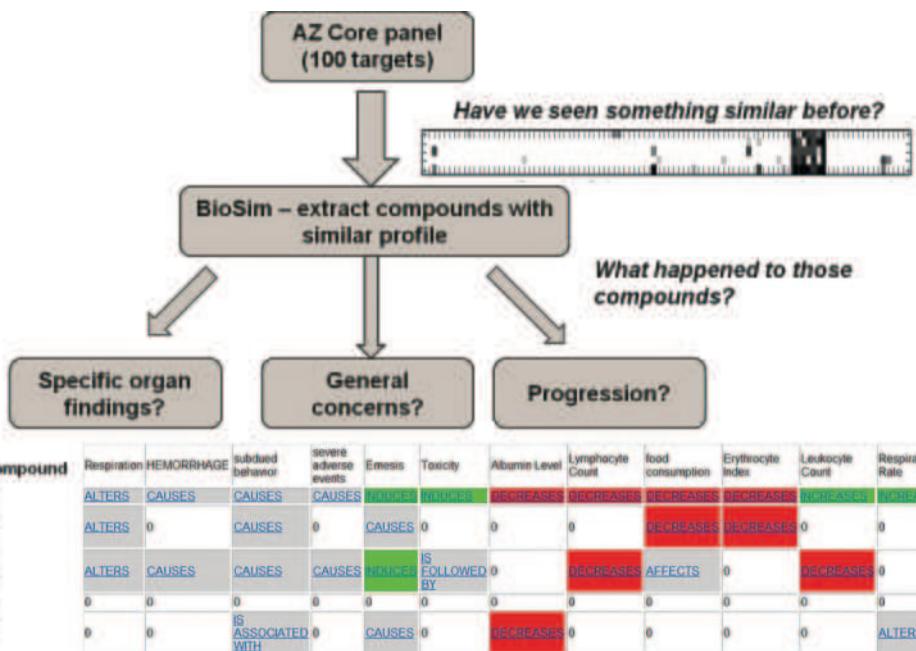


FIGURE 12.5 Schematic overview of the BioSim application. After a compound (A) has been profiled in a broad pharmacology, its profile is compared to all other tested compounds to retrieve the most similar profiles (originating from compound B–E). Among other information, shared *in vivo* findings are highlighted (with links to original documents).

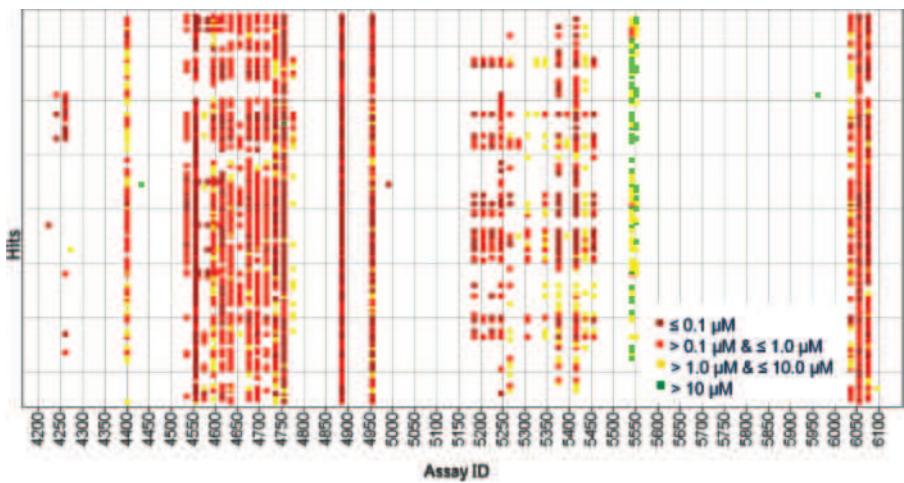


FIGURE 13.5 Analysis of the DR data for a compound class or complete hit set. It can be easily analyzed which other assays overlap with the given hits. The color coding is from green (inactive) to dark red (very active) compounds. Reprinted from *Bioorganic & Medicinal Chemistry*, Vol. 20, Bernd Beck, BioProfile—Extract knowledge from corporate databases to assess cross-reactivities of compounds, 5428–5435, 2012, with permission from Elsevier.

Thrombin SAR Update - 29.10.2012

Class 3A - Series 7

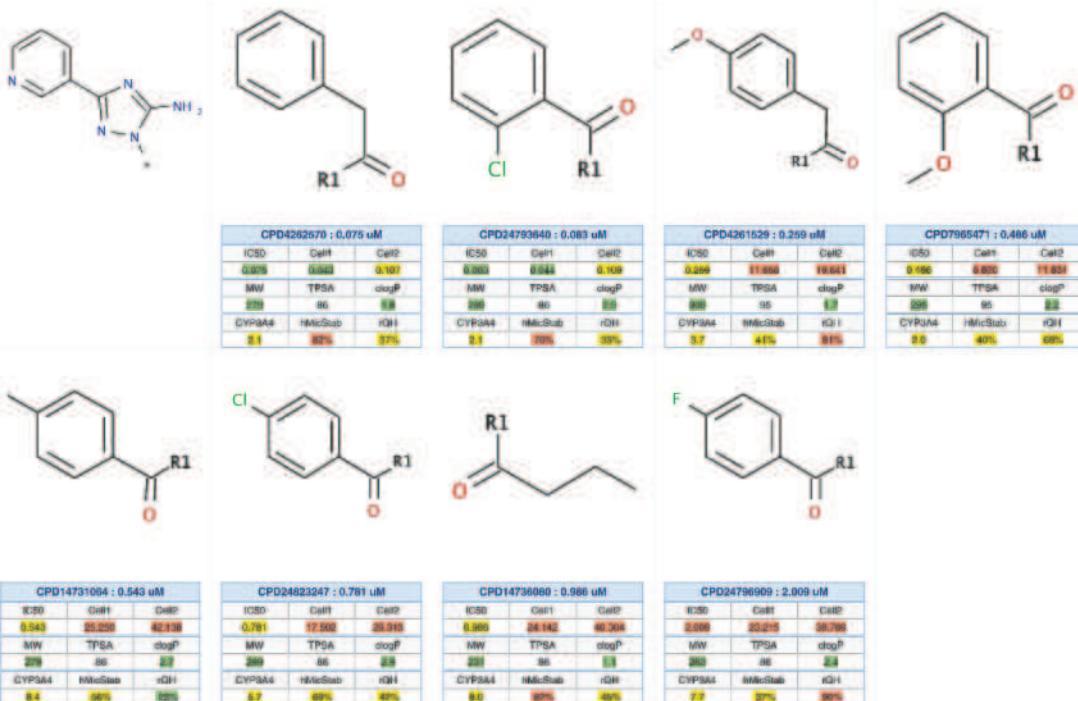


FIGURE 13.6 SAR report based on a matched molecular pairs series. The MMP algorithm efficiently identifies closely related molecules with a conserved constant region (upper left). Molecules within one MMP series can be sorted with respect to the property of interest, for example, IC50. The example shows an MMP series from a public thrombin dataset (PubChem [56], AID 1215), containing IC50 values [μM].

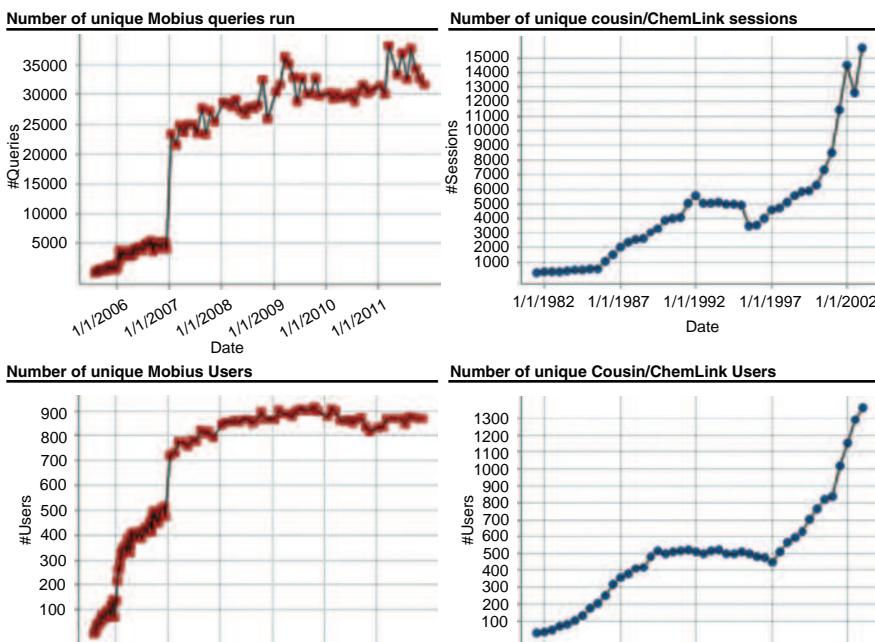


FIGURE 14.3 Usage Statistics for Mobius and Cousin/ChemLink. Usage statistics from Cousin/ChemLink and Mobius. Number of sessions are reported for Cousin/ChemLink versus number of queries for Mobius. On average in our experience there are approximately 2.3 queries run per session.

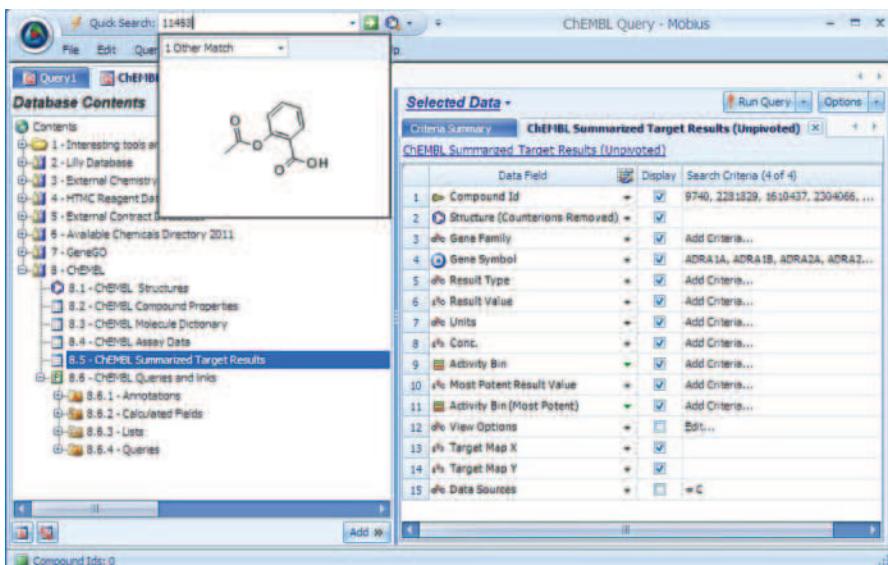


FIGURE 14.4 Mobius Ad hoc Query Interface. Example of the ad hoc query interface in Mobius.

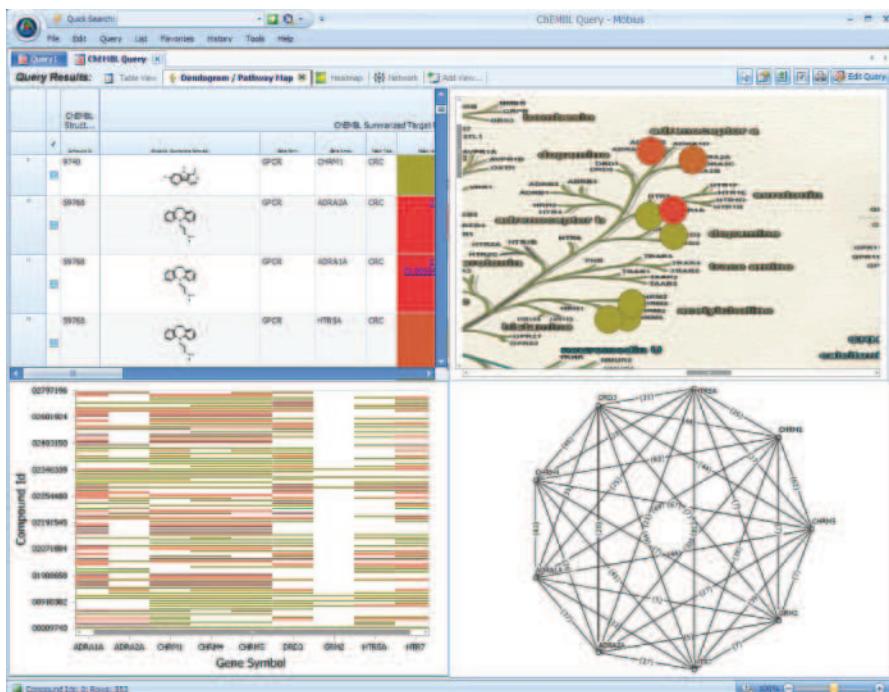


FIGURE 14.5 Query results display. Example results from Mobius queries.

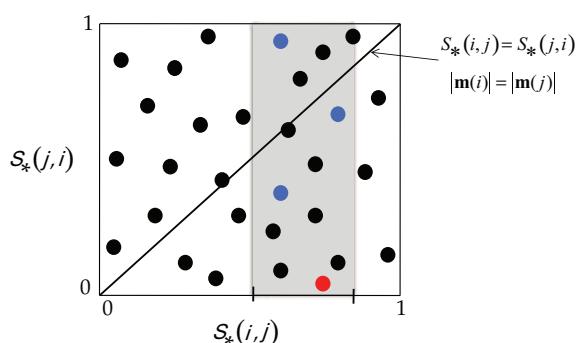


FIGURE 15.4 2D plot of asymmetric similarity functions. The red dot corresponds to a pair of molecules where the i th molecule, which is taken to be active, is “smaller” than the j th molecule with which it is paired. The blue dots located within the gray shaded region of the plot are also associated with the i th molecule but the molecules with which it is paired decrease in size (relative to the i th molecule) as one moves vertically up the gray shaded region. Dots located near the diagonal correspond to cases where the asymmetric similarities are approximately equal in value.

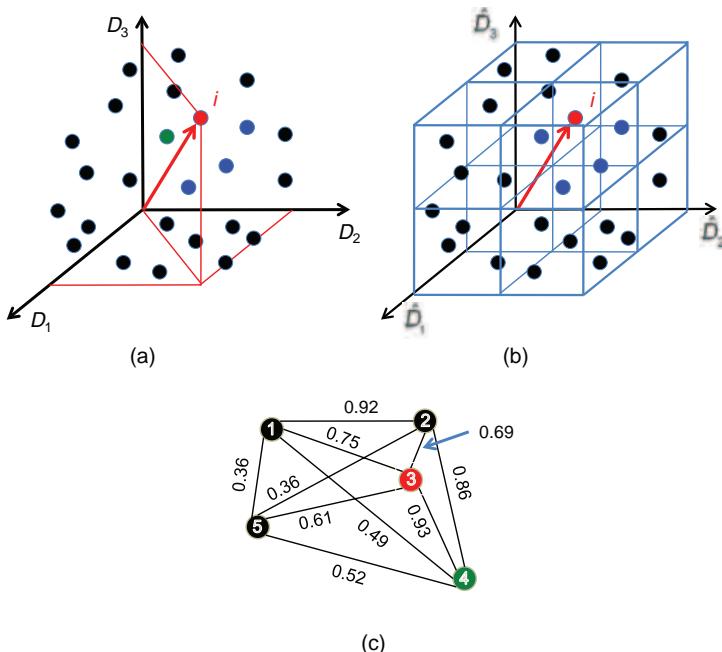


FIGURE 15.6 Simplified schematic diagrams of chemical spaces: (a) coordinated-based, (b) cell-based, and (c) graph-based. The red dot in each figure denotes a biologically active molecule; the green dots designate the nearest neighbor; the blue dots in (a) and (b) represent molecules in the neighborhood of the active (see text for further discussion).

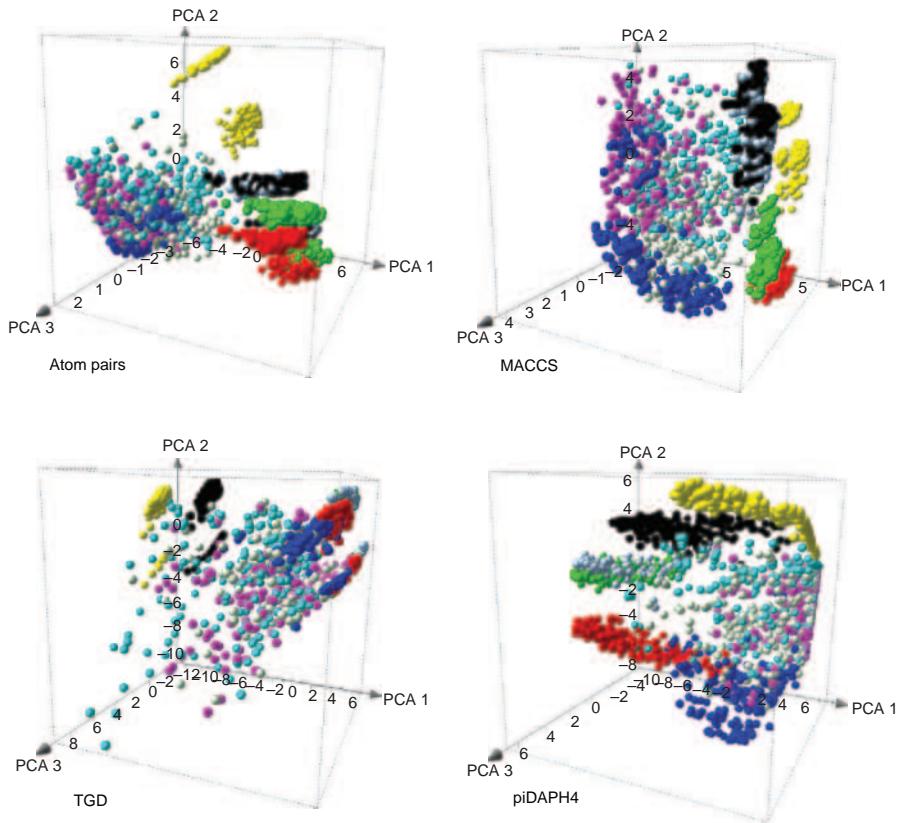


FIGURE 15.7 3D projections of PCA-based chemical spaces generated from a set of 2250 compounds obtained from nine datasets of 250 compounds each using four different molecular fingerprints (Atom pairs, MACCS keys, TGD, and piDAPH4) and the Tanimoto similarity function (see text for further details).

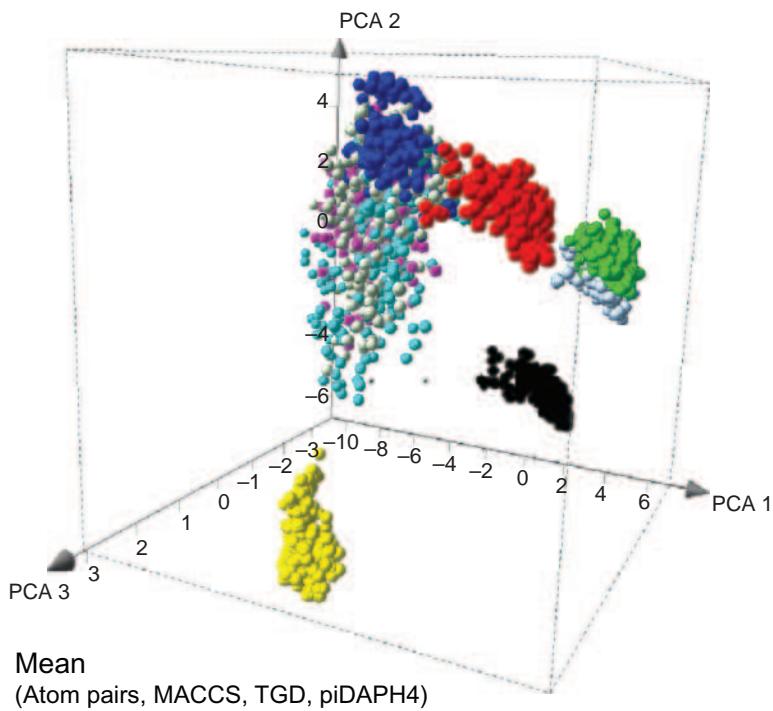


FIGURE 15.8 3D projection of a PCA-based chemical space based on the same set of 2250 compounds in Figure 15.5 and the same four molecular fingerprints (Atom pairs, MACCS keys, TGD, and piDAPH4) but using mean similarity fusion of the individual Tanimoto similarities (see text for further details).

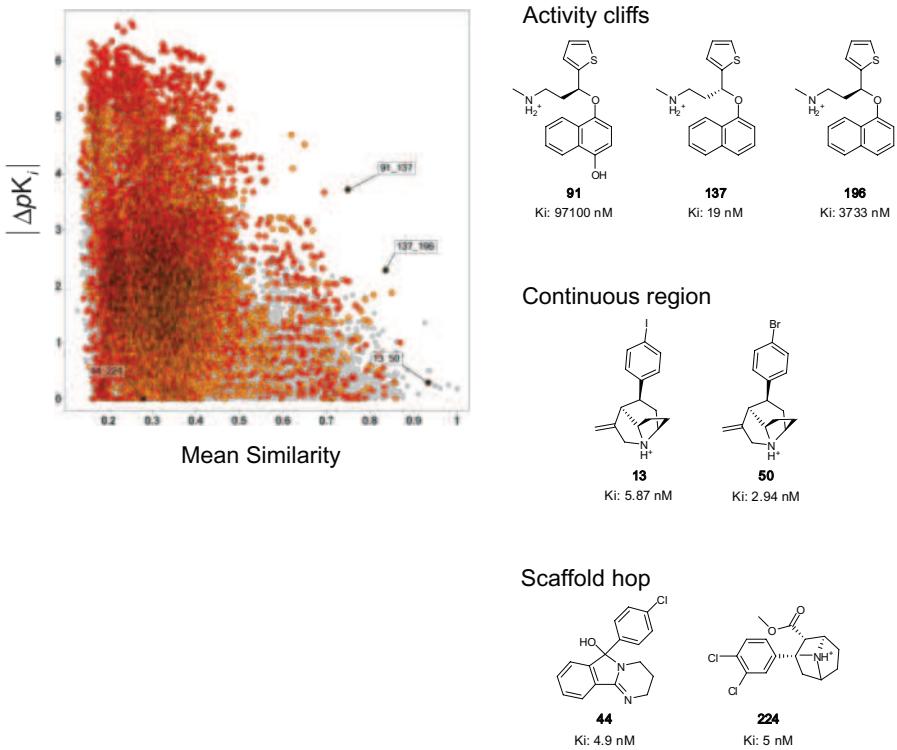


FIGURE 15.11 SAS map of 299 norepinephrine transporter inhibitors (44,551 data points); data points with at least one active compound ($pK_i \geq 7$ nM) in the pair are color-coded by the activity of the most active compound on a continuous scale from orange (least active) to red (most active). The ordinate of the plot gives the absolute potency difference and the abscissa gives the mean similarity obtained by fusing five 2D and 3D molecular fingerprints: radial, atom pairs, MACCS keys, TGD, and piDAPH3. The black dots represent pairs of compounds depicted on the RHS of the figure.