Methods in Pharmacology and Toxicology

Springer Protocols

Kunal Roy *Editor*

Multi-Target Drug Design Using Chem-Bioinformatic Approaches

EXTRAS ONLINE

Humana Press

Methods in Pharmacology and Toxicology

Series Editor
Y. James Kang
Department of Pharmacology &
Toxicology, University of Louisville
Louisville, Kentucky, USA

For further volumes:
http://www.springer.com/series/7653

Methods of Pharmacology and Toxicology publishes cutting-edge methods and protocols in all areas of pharmacological and toxicological research. Each book in the series offers time tested laboratory protocols and step by step methods for reproducible lab experiments to aid toxicologists and pharmaceutical scientists in laboratory testing. With an emphasis on the molecular biology of toxicological methods, Methods of Pharmacology and Toxicology focuses on topics with wide ranging implications to human health, such as Immunotoxicology, Drug Metabolism, and Metabolomics to provide investigators with highly useful compendiums of key strategies and approaches to successful research in drug development.

More information about this series at http://www.springer.com/series/7653

Multi-Target Drug Design Using Chem-Bioinformatic Approaches

Edited by

Kunal Roy

Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India

Humana Press

Editor
Kunal Roy
Department of Pharmaceutical Technology
Jadavpur University
Kolkata, West Bengal, India

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 1557-2153 ISSN 1940-6053 (electronic)
Methods in Pharmacology and Toxicology
ISBN 978-1-4939-8732-0 ISBN 978-1-4939-8733-7 (eBook)
https://doi.org/10.1007/978-1-4939-8733-7

Library of Congress Control Number: 2018961559

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Dedication

For Aatreyi, Arpit and Chaitali

Preface

Despite significant development of novel rational drug design strategies and high-throughput screening methods, the cost of drug development has sharply increased, and at the same time, the rate of failures in clinical trials has escalated [1]. The "one drug, one target, one disease" approach has failed to appreciate the complexities of disease pathways and the system-wide effects of drugs [2]. Diseases are often multifactorial involving a combination of constitutive and/or environmental factors, and they result from the breakdown of robust physiological systems due to multiple genetic and/or environmental factors, leading to the establishment of robust disease conditions. Thus, complex disorders are more likely to be healed or alleviated through simultaneous modulation of multiple targets. Until now, there are still not fully effective drugs for treating complex, multifactorial diseases, such as cancer, metabolic diseases, and neurological diseases [1]. Polypharmacology that addresses small-molecule interactions with multiple targets has generated a great interest in drug discovery [3]. This approach allows for studies of off-target activities and the facilitation of drug repositioning. Multi-target drugs expand the number of pharmacologically relevant target molecules by introducing a set of indirect, network-dependent effects [4]. Moreover, low-affinity binding of multi-target drugs eases the constraints of druggability and significantly increases the size of the druggable proteome. Multi-target agents are a promising strategy to face complex, multifactorial disorders and drug resistance issues. Compared to combination therapies, they present several advantages, including more predictable pharmacokinetics, lower probabilities of drug interactions, and higher patient compliance [5]. Several already existing efficient drugs, such as nonsteroidal anti-inflammatory drugs, antidepressants, anti-neurodegenerative agents, and multi-target kinase inhibitors, affect many targets simultaneously [4]. Hybridization of drugs is also a powerful tool to develop better treatments for several human diseases, as this can provide combination therapies in a single multifunctional agent in a more specific and powerful way than conventional treatments [6].

In polypharmacology, one of the most important goals is to rationally design compounds that act on multiple key targets driving the pathogenesis of a given disease. Therefore, targeting multiple proteins simultaneously stands a good chance to increase drug efficacy and decrease the possibility of drug resistance. In order to achieve these goals, it would be necessary to develop state-of-the-art computational techniques for data curation, model development, and quantitative predictions [2]. Computational approaches are capable of predicting the activity profile of ligands to a set of targets, anticipating potential selectivity issues, and discovering desired multi-target activities early in the iterative design and optimization steps typical of a preclinical drug discovery project. These approaches are based on 2D or 3D shape and chemical similarity, pharmacophore mapping, target and binding site similarity assessment, docking experiments, bioinformatics, graph theory and modeling, machine learning algorithms, and chemogenomics [3]. These approaches can be classified into statistical data analysis and bioinformatics, ligand-based, and structure-based approaches, all of which are well-documented in the literature. The structure-based methods include inverse docking, binding site similarity analysis, inverse pharmacophore modeling, molecular dynamics simulation, etc., while the ligand-based methods include similarity ensemble approach, extended-connectivity fingerprint, fragment-based shape similarity search, etc. which can be used in combination with a variety of machine learning methods including deep learning [2]. Systems biology approaches and cellular networks help to understand complex diseases and their mechanisms and offer a lot of possibilities to point out the key elements as potential drug targets and thus suggest new therapeutic treatment strategies. Proteochemometric modeling (PCM) simultaneously considers the bioactivity of multiple ligands against multiple targets and permits exploration of the selectivity and promiscuity of ligands on biomolecular systems of different complexity [7].

Computational modeling including quantitative structure-activity relationship (QSAR), pharmacophore mapping, docking, virtual screening, and other cheminformatics and proteochemometric approaches play a vital role in finding and optimization of leads in any drug discovery program. Computational modeling helps to understand the important molecular features contributing to the binding interactions with the target proteins, thus facilitating design of new potential compounds and prediction of activity of designed compounds which have not yet been tested. These approaches can save time, money, and more importantly animal sacrifice in the complex, long, and costly drug discovery process.

This volume (*Multi-target Drug Design Using Chem-Bioinformatic Approaches*) intends to showcase the recent advances in computational design of multi-target drug candidates involving various ligand- and structure-based strategies. Different chem-bioinformatic modeling strategies that can be applied for design of multi-target drugs have been discussed in this book. Apart from a few literature reviews on the application of chemometric and cheminformatic modeling tools for multi-target drug design, several case studies are also presented. Important databases and web servers in connection with multi-target drug design are also discussed. There are a total of 21 chapters in this book.

The first chapter “Cheminformatics Approaches to Study Drug Polypharmacology” provides a tutorial overview on selected cheminformatics methods useful for assembling, curating/preparing a chemical database, and assessing its diversity and chemical space. This chapter also discusses the methods for evaluating the structure-activity relationships and polypharmacology.

The second chapter “Computational Predictions for Multi-target Drug Design” highlights the current state-of-the-art methodologies used in multi-target identification for therapeutic effects of known drugs or new drug candidates. This chapter emphasizes experimental validation of model-derived predictions.

The third chapter “Computational Multi-target Drug Design” discusses multi-target or polypharmacological drug discovery and several in silico methodologies like quantitative structure-activity relationship (QSAR), pharmacophore modeling, and molecular docking used in the process of discovery of multi-targeted drugs.

The fourth chapter “Multi-target Drug Design for Neurodegenerative Diseases” presents an overview of multi-target computational methods as well as of their successful applications to neurodegenerative diseases. This chapter recommends application of virtual screening encompassing both structure-based and ligand-based techniques for effective multi-target drug design.

The fifth chapter “Molecular Docking Studies in Multi-target Antitubercular Drug Discovery” gives an overview of various targets for antitubercular drug development followed by a literature survey of application of docking studies for the development of multi-target compounds for developing new promising drug candidates against tuberculosis.

The sixth chapter “Advanced Chemometric Modeling Approaches for the Design of Multi-target Drugs Against Neurodegenerative Diseases” discusses the recent advances in chemometric techniques in multi-target anti-neurodegenerative drug design. This chapter recommends the use of proteochemometric modeling for multi-target-directed ligand design.

The seventh chapter "Computational Studies on Natural Products for the Development of Multi-target Drugs" provides an overview of the currently used computational methods in natural product research, with special reference to multi-target drug design. This chapter discusses that pan-assay interference compounds (PAINS) are for the most part not extraordinarily promiscuous and should not be disregarded prematurely.

The eighth chapter "Computational Design of Multi-target Drugs Against Alzheimer's Disease" provides the basic background about the molecular targets implicated in the pathogenesis of Alzheimer's disease. Furthermore, the chapter reviews structure-activity relationships (SAR), 2D and 3D quantitative structure-activity relationships (QSAR), as well as other computational modeling studies performed on multi-target agents for Alzheimer's disease.

The ninth chapter "Design of Multi-target-Directed Ligands as a Modern Approach for the Development of Innovative Drug Candidates for Alzheimer's Disease" reviews some examples of the exploitation of the multi-target-directed ligand approach in the rational design of novel drug candidate prototypes for the treatment of Alzheimer's disease.

The tenth chapter "Virtual Screening for Dual Hsp90/B-Raf Inhibitors" describes a computational strategy leading to the identification of the first dual inhibitors of heat shock protein 90 (Hsp90) and protein kinase B-Raf, both being validated targets for anticancer drug discovery.

The eleventh chapter "Strategies for Multi-target-Directed Ligands: Application in Alzheimer's Disease (AD) Therapeutics" presents several in silico strategies adopted for the development of multi-target anti-Alzheimer compounds followed by a case study leading to their in vitro validation.

The twelfth chapter "Computational Design of Multi-target Kinase Inhibitors" summarizes two effective computational strategies to identify multi-target kinase inhibitors. The first approach involved a combination of merged pharmacophore matching, database screening, and molecular docking to reliably identify potential multi-target kinase inhibitors. The second strategy employed ensemble pharmacophore-based screening (EPS) of a compound database, post-EPS filtration (PEPSF) of the ligand hits, and multiple dockings.

The thirteenth chapter "Proteochemometrics for the Prediction of Peptide Binding to Multiple HLA Class II Proteins" discusses "proteochemometrics" (PCM) as a method for deriving QSAR. This chapter presents a protocol applied to a set of peptides binding to seven polymorphic HLA class II proteins from locus DP.

The fourteenth chapter "Linked Open Data: Ligand-Transporter Interaction Profiling and Beyond" presents a workflow for retrieving and curating information for multiple drug targets from the open domain, provides insights into how the retrieved data can be employed in ligand- and structure-based approaches, and discusses the hurdles to consider with respect to data analysis.

The fifteenth chapter "Design of Novel Dual-Target Hits Against Malaria and Tuberculosis Using Computational Docking" reviews different approaches (knowledge-based and screening-based) for designing multi-target inhibitors. Additionally, a step-by-step guide (protocol) and different computational resources are also discussed in detail to design multi-target hits for malaria and tuberculosis.

The sixteenth chapter "Computational Design of Multi-target Drugs Against Breast Cancer" presents protocols and computational practices for screening of multi-target drug molecules for breast cancer receptors. However, the authors emphasize that validation of the screened molecules is essential in the in vitro and in vivo conditions.

The seventeenth chapter "Computational Methods for Multi-target Drug Designing Against *Mycobacterium tuberculosis*" presents available strategies for computational multi-target drug designing with their advantages and disadvantages. This chapter also discusses an easy, fast, and accurate protocol for multi-target drug designing against the *Mycobacterium tuberculosis.*

The eighteenth chapter "Development of a Web Server for Identification of Common Lead Molecules for Multiple Protein Targets" presents a computational protocol that involves screening, docking, and scaffold-based optimization of hit molecules from a variety of compound libraries against any two specified protein targets. The protocol is made available via a web server named "Multi-target Ligand Design."

The nineteenth chapter "Computational Method for Prediction of Targets for Breast Cancer Using siRNAs Approach" discusses the development and application of a web-based database, BOSS, for selection of potential RNAi based on the sequences that have been used and validated for RNAi-mediated suppression of breast oncogenes. This database includes the latest information regarding used RNAi molecules that can be cost-effective and less time-consuming.

The twentieth chapter "Historeceptomics: Integrating a Drug's Multiple Targets (Polypharmacology) with Their Expression Pattern in Human Tissues" presents "historeceptomics" as a new, integrative informatics approach to describing the mechanism of action of drugs in a holistic, in vivo context. The chapter discusses that this approach may give new insights into drug mechanism of action, drug repurposing, and prediction of adverse effects, including the design and development of multi-target drugs or drug combinations.

The twenty-first chapter "Networking of Smart Drugs: A Chem-Bioinformatic Approach to Cancer Treatment" reviews the existing network of "smart drugs" by using a chem-bioinformatic approach toward cancer treatment. According to the authors, an application of computational tools in smart drug designing for cancer treatment will be path-breaking in the future.

I am sure that this collection of 21 chapters will be useful to the researchers working in the field of drug discovery and development.

Kolkata, India ***Kunal Roy***

References

1. Lu J-J, Pan W, Hu Y-J, Wang Y-T (2012) Multi-target drugs: the trend of drug research and development. PLoS One 7(6):e40262. doi:10.1371/journal.pone.0040262
2. Chaudhari R, Tan Z, Huang B, Zhang S (2017) Computational polypharmacology: a new paradigm for drug discovery. Expert Opin Drug Discov 12(3):279–291, doi: 10.1080/17460441.2017.1280024
3. Rastelli G, Pinzi L (2015) Computational polypharmacology comes of age. Front Pharmacol 6:157. doi: 10.3389/fphar.2015.00157
4. Korcsmáros T, Szalay MS, Böde C, Kovács IA, Csermely P (2007) How to design multi-target drugs. Expert Opin Drug Discov 2:799–808. doi: 10.1517/17460441.2.6.799
5. Talevi A (2015) Multi-target pharmacology: possibilities and limitations of the "skeleton key approach" from a medicinal chemist perspective. Front Pharmacol 6:205. doi: 10.3389/fphar.2015.00205
6. Bérubé G (2016) An overview of molecular hybrids in drug discovery. Expert Opin Drug Discov 11:281–305. doi: 10.1517/17460441.2016.1135125
7. Cortes-Ciriano I, van Westen GJP, Murrell DS, Lenselink EB, Bender A, Malliavin TE (2015) Applications of proteochemometrics—from species extrapolation to cell line sensitivity modeling. BMC Bioinform 16(Suppl 3):A4. doi: 10.1186/1471-2105-16-S3-A4

Contents

Part I Chem-Bioinformatic Tools

Part II Computational Multi-Target Drug Design: Literature Reviews

Part III Case Studies

Part IV Databases and Web Servers

Part V Special Topics

Contributors

Azizeh Abdolmaleki • *Department of Chemistry, Tuyserkan Branch, Islamic Azad University, Tuyserkan, Iran*

Domenico Alberga • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

Andrew Anighoro • *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms- Universität, Bonn, Germany*

Jamal Mohammad Arif • *Department of Bioscience, Integral University, Lucknow, Uttar Pradesh, India*

Jürgen Bajorath • *Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Soumalee Basu • *Department of Microbiology, University of Calcutta, Kolkata, West Bengal, India*

Ruchika Bhat • *Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India; Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India*

Timothy Cardozo • *New York University School of Medicine, NYU Langone Health, New York, NY, USA*

Marco Catto • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

M. Natália D. S. Cordeiro • *LAQV@REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal*

Sucharita Das • *Department of Microbiology, University of Calcutta, Kolkata, West Bengal, India*

Kris Simone Tranches Dias • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil*

Ivan Dimitrov • *Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria*

Irini Doytchinova • *Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria*

Gerhard F. Ecker • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

Darren R. Flower • *School of Life and Health Sciences, Aston University, Birmingham, UK*

Matheus de Freitas Silva • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil; Programa de Pós-Graduação em Química, Federal University of Alfenas, Alfenas, Brazil*

Rajanikant G. K. • *School of Biotechnology, National Institute of Technology Calicut, Calicut, Kerala, India*

Jahan B. Ghasemi • *Drug Design in Silico Lab, Chemistry Faculty, University of Tehran, Tehran, Iran*

Vanessa Silva Gontijo • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil; Programa de Pós-Graduação em Química, Federal University of Alfenas, Alfenas, Brazil*

NEELIMA GUPTA • *Centre of Advanced Study, Department of Chemistry, University of Rajasthan, Jaipur, India*

DIMITRA HADJIPAVLOU-LITINA • *Department of Pharmaceutical Chemistry, School of Pharmacy, Faculty of Health Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece*

AMIT KUMAR HALDER • *LAQV@REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal*

EVA HELLSBERG • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

SANKALP JAIN • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

QAZI MOHAMMAD SAJID JAMAL • *Department of Health Information Management, College of Applied Medical Sciences, Buraydah Colleges, Buraydah, Al Qassim, Saudi Arabia; Novel Global Community Educational Foundation, Sydney, NSW, Australia*

ABHILASH JAYARAJ • *Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India; Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India*

B. JAYARAM • *Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India; Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India; Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, New Delhi, India*

J. JESÚS NAVEJA • *Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico; PECEM, School of Medicine, Universidad Nacional Autónoma de México, Mexico City, Mexico*

SOTIRIOS KATSAMAKAS • *Department of Pharmaceutical Chemistry, School of Pharmacy, Faculty of Health Sciences, Aristotle University of Thessaloniki, Thessaloniki, Greece*

KAVINDRA KUMAR KESARI • *Department of Applied Physics, Aalto University, Helsinki, Finland; Department of Bioproduct and Biosystem, Aalto University, Helsinki, Finland*

STEFANIE KICKINGER • *Department of Pharmaceutical Chemistry, University of Vienna, Vienna, Austria*

MANOJ KUMAR • *Department of Chemistry, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India; Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON, Canada*

GIUSEPPE FELICE MANGIATORDI • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

JOSÉ L. MEDINA-FRANCO • *Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico*

MUKTI N. MISHRA • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

ANA S. MOURA • *LAQV@REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal*

ORAZIO NICOLOTTI • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

JÉSSIKA DE OLIVEIRA VIANA • *Federal University of Paraíba, Health Center, João Pessoa, PB, Brazil*

CINDY JULIET CRISTANCHO ORTIZ • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil; Programa de Pós-Graduação em Química, Federal University of Alfenas, Alfenas, Brazil*

Prateek Pandya • *Amity Institute of Forensic Sciences, Amity University, Noida, India*

Amita Pathak • *Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India; Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India*

Luca Pinzi • *Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy*

Giulio Rastelli • *Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy*

Fernanda I. Saldívar-González • *Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico*

Daniela Schuster • *Institute of Pharmacy/Pharmacognosy and Center for Molecular Biosciences Innsbruck, University of Innsbruck, Innsbruck, Austria; Department of Pharmaceutical and Medicinal Chemistry, Institute of Pharmacy, Paracelsus Medical University Salzburg, Salzburg, Austria*

Luciana Scotti • *Federal University of Paraíba, Health Center, Teaching and Research Management—University Hospital, João Pessoa, PB, Brazil*

Marcus T. Scotti • *Federal University of Paraíba, Health Center, João Pessoa, PB, Brazil*

Anuj Sharma • *Department of Chemistry, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India*

Ashok Sharma • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

Fereshteh Shiri • *Department of Chemistry, University of Zabol, Zabol, Iran*

Mohd. Haris Siddiqui • *Department of Bioengineering, Integral University, Lucknow, Uttar Pradesh, India*

Manpreet Singh • *Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi, India*

Gaurava Srivastava • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

Sinoy Sugunan • *School of Biotechnology, National Institute of Technology Calicut, Calicut, Kerala, India*

Norberto Sánchez-Cruz • *Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico*

Veronika Temml • *Institute of Pharmacy/Pharmacognosy and Center for Molecular Biosciences Innsbruck, University of Innsbruck, Innsbruck, Austria*

Ashish Tiwari • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

Shubhandra Tripathi • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

Daniela Trisciuzzi • *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Bari, Italy*

Atul Tyagi • *Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, Uttar Pradesh, India*

Seema Verma • *Centre of Advanced Study, Department of Chemistry, University of Rajasthan, Jaipur, India*

Flávia Pereira Dias Viegas • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil*

CLAUDIO VIEGAS JR. • *PeQuiM, Laboratory of Research in Medicinal Chemistry, Institute of Chemistry, Federal University of Alfenas, Alfenas, Brazil; Programa de Pós-Graduação em Química, Federal University of Alfenas, Alfenas, Brazil*

VENTSISLAV YORDANOV • *Faculty of Pharmacy, Medical University of Sofia, Sofia, Bulgaria*

Part I

Chem-Bioinformatic Tools

Methods in Pharmacology and Toxicology (2018): 3–25
DOI 10.1007/7653_2018_6
© Springer Science+Business Media New York 2018
Published online: 09 June 2018

Cheminformatics Approaches to Study Drug Polypharmacology

J. Jesús Naveja, Fernanda I. Saldívar-González, Norberto Sánchez-Cruz, and José L. Medina-Franco

This work is dedicated to the loving memory of Nicolás Medina Sandoval.

Abstract

Herein is presented a tutorial overview on selected chemoinformatics methods useful for assembling, curating/preparing a chemical database, and assessing its diversity and chemical space. Methods for evaluating the structure–activity relationships (SAR) and polypharmacology are also included. Usage of open source tools is emphasized. Step-by-step KNIME workflows are used for illustrating the methods. The methods described in this chapter are applied onto a chemical database especially relevant for epi-polypharmacology that is an emerging area in drug discovery. However, the methods described herein could be extended to other therapeutic areas and potentially to other areas of chemistry.

Keywords Chemoinformatics, ChemMaps, Chemical space, Data mining, Epigenetics, Epi-informatics, KNIME, Molecular diversity, Open-access, Polypharmacology, Structure–activity relationships, SmARt

1 Introduction

The rapid growth of chemical information demands efficient and reliable computational algorithms to analyze the accumulated data. Similarly, current trends in drug discovery such as polypharmacology [1, 2] demand the organization and efficient mining of multiple drug–target interactions and study of structure–multiple activity relationships (SMARt) efficiently [3]. Indeed, a plethora of methods and resources for exploiting SMARt and other data relevant to polypharmacology have been published, and many of them are open access [4]. This review includes methodological details for implementing scalable KNIME cheminformatics workflows for:

a. Curating a chemical database;

b. Computing chemical descriptors;

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/7653_2018_6) contains supplementary material, which is available to authorized users.

c. Analyzing and comparing database diversity, and
d. Visualizing their chemical space.

Of note, KNIME is an open-access initiative intended for generating data mining pipelines or workflows, which are capable of integrating multiple tools [5].

Although sufficiently detailed, this review aims at being a quick practical guide. More comprehensive tutorials in chemoinformatics can be found elsewhere [6, 7]. Additionally, web applications for cheminformatics methods that have been developed by our research group are mentioned in the respective subsections. These applications are part of the D-Tools initiative for generating open cheminformatics resources (available at https://www.difacquim.com/d-tools/). The D-Tools usage is further described elsewhere [4, 8–11], and these are not the focus of this review.

2 Methods

2.1 Construction and Curation of a Compound Database

Due to the increase in the amount of chemical information, where it is common to the concept of big data [12], the efficient management of information represents a challenge today. This is of particular importance in polypharmacology where large compound datasets contain information of screening across several biological endpoints. In response to this need, the construction of compound and other databases can be a convenient way to sort information according to the data available and the specific objectives of the study.

In chemoinformatics, construction of databases is a fundamental practice to perform various computational studies like the design of chemical libraries, characterization and comparison of the chemical space, the study of the structure–activity relationships (SAR), and virtual screening studies, among others.

Currently, web pages of large public databases such as DrugBank [13], ChEMBL [14], ZINC [15], and BindingDB [16] allow the user to download their own databases (complete or partial downloads) with information on approved drugs, drugs in the experimental phase, commercially available compounds, molecular targets, etc. However, these databases are not always updated, so they can be enriched with new information published in books or in scientific articles.

Also, in research groups devoted to the synthesis, isolation from natural sources and/or evaluation of new chemical entities can be carried out for the construction of completely new compounds' databases. Such collections are usually referred to as in-house databases.

The process of building and annotating chemical databases is not trivial. Each organization has its own rules, conventions, and procedures. However, the steps that are considered essential are listed below:

1. Identify compounds and resources that contain information required, e.g., journals and databases with chemical information [4, 17].
2. In a spreadsheet, it is recommended that the user has the following information for each compound:
   a. Name of each compound. This can be searched in public databases.
   b. A number that identifies this compound in the database that has been consulted, for example, ChemSpider ID, Substance or Compound ID (SID, CID in PubChem, the CAS registry number, or an internal and consistent code if building an in-house collection).
   c. Structure input. An example of this is the use of Canonical SMILES notation used for encoding molecular structures that can be imported to other molecular editing systems. It is worth noting the relevance of creating a single computational representation. This can be achieved by using various algorithms in a process known as canonicalization.
3. Once this information is collected in the spreadsheet, save the database preferably in *.csv* format (comma delimited). Other database formats with chemical information and compatible with most computer programs as KNIME are *sdf* (structure data file), *mol* (molecular data file), and mol2 (tripos mol2 file).

For the management and analysis of databases, the KNIME Example Server provides access to many explanatory workflows. The example server is accessible via the KNIME Explorer panel within the KNIME workbench and represents a great help when starting a new workflow.

Some of the nodes to start working with files with chemical information are: *Molecule Type Cast*, a node useful for reading chemical data from a .csv file or database, and this node casts any string as a chemical type (i.e., It tells KNIME "This is a smiles string") and *Marvin MolConverter*, a node provided by Chemaxon/Infocom that translates seamlessly between types (smiles ↔ sdf ↔ mrv).

An important aspect to consider when analyzing molecular databases generated by other scientists is that these may contain wrong information or unnecessary information for the intended application or project. Therefore, cleaning or curating the information is highly relevant to enhance the quality of the data and to avoid erroneous results [18].

As in the construction of databases, there is no widely accepted standard protocol for the preparation of small molecules. However, hereunder are described the essential points in the preparation and curation of databases:

1. Normalize the chemical structures. In this step, each chemical structure is checked for valid atom types, valence checks, and functional groups such as nitro groups are converted to a consistent representation. This is followed by a standardization step in which chemical structures are converted to a canonical tautomeric form, aromatic structures are kekulized, placement of stereo bonds is standardized, and all implicit hydrogens are converted to explicit hydrogens [19].
2. Remove duplicates. After the molecules have been properly standardized, it is appropriate to detect duplicates. InChiKeys is a useful method to identify several states of protonation and tautomers of a molecule.
3. Discard inorganic and organometallic atoms or molecules if these are not the object of study. It is worth mentioning that the majority of the chemoinformatics programs currently available are developed to process small organic molecules.
4. Wash the compound database by applying to each molecule a set of rules of "cleaning" such as the elimination of salts and the adjustment of the protonation states. The purpose of this step is to ensure that each chemical structure is in a form suitable for the subsequent modeling.
5. Enumerate tautomers and stereoisomers. This step is important in virtual screening studies, particularly when using search methods such as docking or pharmacophore.
6. Optimize the geometry and minimize the energy if the database will be used to evaluate the potential of each compound to bind to a receptor or enzyme, or to calculate descriptors which depend on the three-dimensional conformation of the molecule. The specific method to optimize the geometry will largely depend on the type, quantity of molecules to optimize, and, most importantly, on the specific application.

In addition, if the quantity of compounds is too large to be examined or tested with the resources available, different strategies can be employed to reduce the number of compounds in a rational and consistent manner. Such strategies include: filtering—essentially imposing secondary search criteria to eliminate compounds, clustering—taking a representative subset of a larger set, and human inspection of the compound structures (with or without extra data) [20].

In several articles, the impact of the use of duplicates and inconsistencies in the molecular structures in prediction models had already been discussed [21]. For this reason, the project CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) has developed a workflow to curate databases [22]. A similar workflow can be found at the link https://github.com/zhu-lab/curation-workflow/blob/master/Structure%20Standardizer2.zip.

Gally et al. also report a workflow designed to prepare molecular databases but focused on studies of virtual screening [23]. In addition to carrying out of the standardization of chemical structures, the workflow of Gally et al. has implemented filters (based on molecular property distribution) to characterize specific subsets of chemical libraries such as drug-like, lead-like, or fragment-like subsets of compounds.

See Workflow 1 in the Supplementary Information for an example in KNIME.

The following analyses use an epigenomics chemical database that has already been curated and published [24].

2.2 Diversity Analysis

In drug discovery projects focused on one single target or multiple targets, it is of high relevance quantifying the structural diversity of compound datasets. For instance, if the goal of a high-throughput screening campaign is to identify hit compounds with a desirable polypharmacological profile, it is desirable to screen a compound collection with high diversity. This will increase the possibilities to find active molecules with a desirable profile. If the goal of the screening campaign is to further develop a focused library (e.g., increase the structure–activity information of a focused region in chemical space [25]), it is desirable to screen a compound dataset with high internal similarity (low diversity).

The diversity in a chemical library can be assessed in multiple ways, mainly depending on the data under scrutiny. In addition to the diversity metric, a key aspect of diversity analysis is molecular representation [26, 27]. The most common ways to represent molecules in chemoinformatic applications are molecular descriptors (including physicochemical properties and molecular fingerprints), and chemical scaffolds [28]. Depending on the type of descriptor and the level of accuracy desired (considering the time of computation and the number of compounds to analyze), the input structures can be in two or three dimensions (the latter requires conformational analysis). The choice of molecular representation depends on the goals of the study.

A more detailed description on how to use molecular descriptors and scaffolds as an input for diversity analysis follows in the next paragraphs. See Workflow 2 in the Supplementary Information for an exemplary diversity analysis in KNIME.

2.2.1 Molecular Descriptors

Molecular descriptors capture information of the whole molecule and are usually straightforward to interpret. Also, whole molecular properties such as physicochemical properties of pharmaceutical interest are usually part of empirical rules for drug likeness that aids to guide drug discovery programs. KNIME includes RDKit, CDK, and Indigo nodes, with which complexity descriptors (e.g., chiral carbons, and fraction of sp^3 carbon atoms), and physicochemical properties of pharmaceutical interest (including molecular weight, number of hydrogen bond donors and acceptors, number of rotatable bonds, logarithm of octanol–water partition coefficient, and topological polar surface area) [28].

Starting with curated databases (discussed in Sect. 2.1), the steps for quantifying diversity with molecular descriptors are:

1. Select the features to be evaluated (usually the six commonest physicochemical properties of pharmaceutical relevance, vide supra).
2. Scale the data using a Z-transformation. This transforms the data to dimensional units. The purpose is to improve the comparability of the variables and give a similar weight to all of them independently of the units with which they were originally measured.
3. Compute pairwise euclidean distance. For a database with n compounds, $n \times (n-1)/2$ pairwise comparisons are to be computed. Euclidean distance is calculated with the formula:

   $$D(\mathrm{A}, \mathrm{B}) = \sqrt{\sum\nolimits_{i=1}^{n} (a_i - b_i)^2},$$

   where $D(\mathrm{A}, \mathrm{B})$ is the euclidean distance between compound A and B, a_i and b_i are the i-th descriptor, and n the total number of descriptors [29]. $D(\mathrm{A}, \mathrm{B})$ can take any positive real number as value.
4. Compute a central tendency statistic (e.g., mean or median) for all the pairwise comparisons. The larger the mean or median, the more diverse the dataset is [30].
5. Finally, for comparison, the statistic can be computed for other reference databases or looked up at the literature if already reported.

2.2.2 Molecular Fingerprints

Many structural features escape the very general information obtained with physicochemical and complexity descriptors. Molecular fingerprints are vectors that aim towards a more comprehensive set of features (usually more than a hundred) to compare molecules. Every feature is encoded as a Boolean variable, where "0" represents absence and "1" represents presence of the feature. Therefore, repeated motifs are not generally acknowledged. For every molecule, a Boolean vector of features is obtained, and these are susceptible of standard set operations [31–33]. However, molecular fingerprints do have limitations, for example, they could be more difficult to interpret intuitively, and therefore pose a greater difficulty for extracting insights relevant for medicinal chemistry.

The steps for computing diversity based on fingerprints are:

1. Select a molecular fingerprint. Although the selection of the "best" fingerprint could be different from case to case, it has been consistently found that MACCS keys 166-bits [34] are useful for quantifying database diversity. In turn, extended connectivity fingerprints of diameter 4 (ECFP4) [32] as well as other circular fingerprints are, overall, better suited for virtual screening, activity landscape modeling, and SAR studies in general.
2. Compute pairwise Tanimoto similarity [27, 35]. For a database with n compounds, $n \times (n - 1)/2$ pairwise comparisons are to be computed. Tanimoto similarity is calculated with the expression:

   $$T(\mathrm{A},\mathrm{B}) = \frac{c}{a + b - c'}$$

   where $T(\mathrm{A},\mathrm{B})$ is Tanimoto similarity with possible values being any real number between 0 and 1, c is the number of features for which both molecules A and B have a "1" value, a is the number of features for which molecule A has a "1" value, and b is the number of features for which molecule B has a "1" value. Dissimilarity matrices implemented in KNIME are quite efficient at these calculations. However, by default they compute values as dissimilarities, the complement of similarities, or distance matrices. Conversion from Tanimoto dissimilarity to similarity is accomplished by just subtracting the value from 1 (Ts = 1 − Td, where Ts is Tanimoto similarity and Td is Tanimoto dissimilarity).
3. Compute a central tendency statistic (e.g., mean or median) for all the pairwise comparisons. Conversely to Euclidean distance (and any distance metric in general), the smaller the mean or median, the more diverse the dataset is [30].
4. Finally, for comparison, the statistic can be computed for other reference databases or looked up at the literature if already reported.

2.2.3 Molecular Scaffolds

KNIME has nodes for finding Murcko scaffolds [36, 37]. By definition, Murcko scaffolds contain all the cyclic systems in a molecule as well as the linkers between them. All other decorations and ramifications are omitted. The greatest benefit of working with scaffolds data is that, unlike molecular fingerprints, they are readily interpreted by medicinal chemists. Nonetheless, the representation is rougher and loses information from the side chains. Also, more advanced methods must be applied to account for the structural relations among the scaffolds.

It is logical and generally accepted that a dataset is more diverse when it has a large number of different scaffolds, and the proportions of compounds with each scaffold are evenly distributed. The procedure for measuring scaffold diversity is as follows:

1. Find Murcko scaffolds for every molecule in the dataset.
2. Compute a frequency table of the scaffolds.
3. From here, there are a number of different methods for assessing the diversity [38]:
   a. Order the scaffolds by their frequency of occurrence and compute the median (i.e., the minimum number of scaffolds in the database that contain at least 50 % of the total entries). Lower values in this statistic mean higher diversity.
   b. Order the scaffolds by their frequency of occurrence. This order would be an index from 1 to n, where n is the total number of different scaffolds in the dataset. Divide all indexes by n, such that the highest index value is 1. Using scaffold indexes in the x-axis and their respective cumulative proportions in the y-axis, compute the area under the curve as a diversity statistic. This statistic admits as value any real number in the domain [0.5, 1.0]. Lower values in this statistic mean higher diversity.
   c. Compute scaled Shannon entropy (SSE) with the formula:

$$\mathrm{SSE} = \frac{\mathrm{SE}}{\log_2 n'}$$

where SE $= \sum_{i=1}^{n} -p_i \log_2 p_i$,
where p_i is the proportion in the dataset of th i-th scaffold (calculated by dividing the occurrence of this i-th scaffold by the total number of entries/molecules), SE is the Shannon entropy, and n is the total number of scaffolds in the dataset. SSE takes as value a real number in the range [0,1]. For this statistic, higher values mean higher scaffold diversity.

4. Finally, the statistic can be computed for other reference databases for comparison.

2.2.4 Consensus Diversity Plots

In the light of numerous variables that can be used to quantify diversity, visual representations have been built in order to summarize multiple of them simultaneously. These are the consensus diversity plots (CDPs). A CDP, as defined by González-Medina et al. [10], renders 2D diversity measured by scaffolds, fingerprints, physicochemical properties, and the number of compounds in the databases. It is also possible to integrate 3D data [24]; however, we will not emphasize on 3D data usage here. The steps required for plotting a CDP from data are:

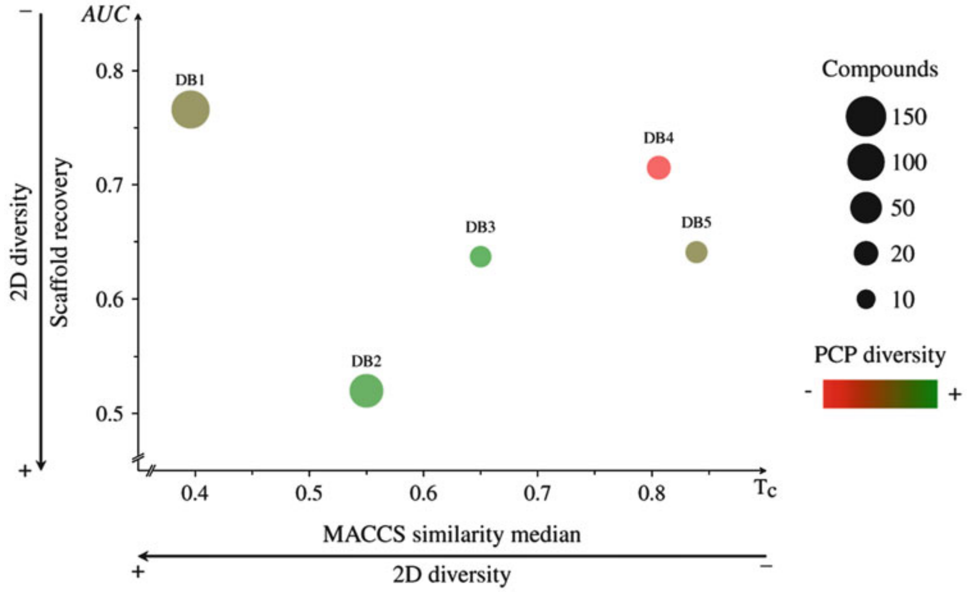


Fig. 1 An exemplary consensus diversity plot (CDP). Each data point represents a compound database. Molecular fingerprints diversity is plotted in the *x*-axis, the scaffold diversity in the *y*-axis, the physicochemical properties diversity in a color continuous scale, and the relative number of compounds in the database as the data point size. *AUC* area under the curve, *PCP* physicochemical properties

1. Curate databases; calculate diversity with physicochemical properties, molecular fingerprints, and scaffolds (see above for details).
2. Plot the molecular fingerprints diversity in the *x*-axis, the scaffold diversity in the *y*-axis, the physicochemical properties in a color continuous scale, and the number of compounds in the database as the data point size. Every data point represents a database. (See Fig. 1 and Supplementary KNIME Workflow 3 for a few examples.)

As an alternative, an online server was developed for generating CDPs and is also available in D-Tools (see Sect. 1). A video tutorial is available at https://youtu.be/lruo1ypKGbE, and detailed written instructions about how to use it can be found at http://132.248.103.152:3838/CDPlots/.

2.3 Structure–Activity Relationship Analysis

A common assumption in virtual screening is that similar molecules are expected to have similar properties, e.g., comparable biological activity. This assumption is called the similarity principle. Although virtual screening is often useful for detecting active compounds, it is reassuring to verify whether the similarity principle is valid for the molecules under scrutiny. Such insights can be obtained through a subtype of SAR analysis, activity landscape modeling. SAR analysis of chemical libraries, for which activity against a biological target is known, can also reveal substructures that are relevant for inhibiting the target in question. The next paragraphs give details onto some useful methods for assessing SAR of single and multiple libraries simultaneously. Workflow 4 in the Supplementary Information illustrates a KNIME implementation of the methods described below.

2.3.1 Structure–Activity Similarity Maps

Structure–activity similarity (SAS) maps are bidimensional activity landscape representations that contrast structural similarity (e.g., measured with Tanimoto coefficient of molecular fingerprints) and activity similarity (for example, as pIC_{50} or p*K*i). Systematic pairwise compound comparisons are included in the plot [39]. Each point in a SAS map represents a pair of compounds and is colored according to the most active compound of the pair. The sequence of steps for generating and ultimately interpreting a SAS map is as follows:

1. Given n compounds in a library, compute the $n \times (n - 1)/2$ paired chemical similarity as described in Sect. 2.2.2.
2. Similarly, for the same paired comparisons calculate the absolute difference in potency. All compounds should have potency in pIC_{50} units. It is calculated from IC_{50} measurements in nanomolar concentration units with the formula (ideally, all compounds should have IC_{50} values measured under the same protocol and assay conditions):

$$\mathrm{pIC}_{50} = -\log_{10}(\mathrm{IC}_{50}\,[\mathrm{nM}]).$$

3. Plot the structural similarity in the x-axis and the potency difference in the y-axis. The color of the data points can also be set to render more information, for example, the maximum potency value in the pair.
4. The resultant plot, illustrated in Fig. 2, can be divided into four quadrants with thresholds defined a priori: (a) smooth (high structural similarity and low activity difference), (b) activity cliffs (high structural similarity but high activity difference), (c) scaffold hops (low structural similarity but low activity difference), and (d) uncertainty (low structural similarity and high activity difference) [40–42]. Typical potency thresholds are 2 for deep activity cliffs and 1 for shallow activity cliffs. In the case of structural similarity, 1 or 2 standard deviations above the mean could be used.

Alternatively, a web application for plotting SAS maps can be found at D-Tool under https://unam-shiny-difacquim.shinyapps.io/ActLSmaps/. A video tutorial is available at https://youtu.be/52jHCcg5mXU.

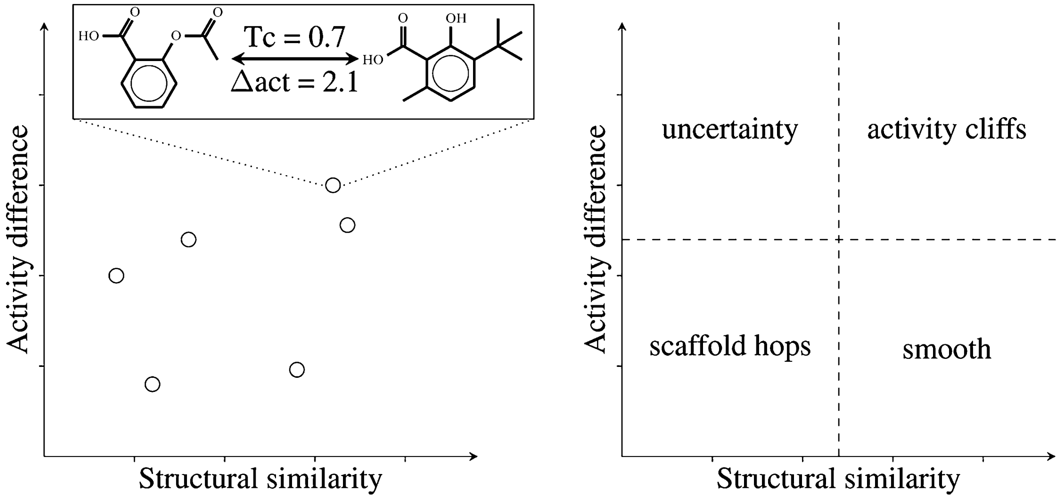


Fig. 2 Structure–activity similarity (SAS) maps. Each data point represents a pair of compounds. The *x*-axis plots the structural similarity, while the *y*-axis plots the activity difference. Four quadrants are formed as described in Sect. 2.3.1. A color scale might be added to represent density of points or the maximum activity value in the pair. *Tc* Tanimoto coefficient

2.3.2 Scaffold Enrichment Factor

SAR can also be explored based on chemical scaffolds. For every dataset with activity annotations against a particular biological target, every scaffold could be considered as a cluster of molecules. At this point, it is interesting to find which clusters have a higher or lesser proportion of active molecules, pointing towards clusters of highly related molecules that tend to be more or less active than the average. This is the basis of the calculation of enrichment factors (EF) for scaffolds, which are obtained as follows:

1. If activity is represented quantitatively in the dataset, a threshold of activity should be set a priori. Often, a pIC_{50} of 5–6 or more is useful for defining a compound as active.
2. Essentially, the EF is an odds ratio, i.e., a ratio of proportions. Specifically, the proportion of active compounds with a given scaffold is divided by the proportion of active compounds in the general dataset. A more formal definition would be that, for every scaffold λ, an EF is calculated using the equation [43]:

$$\mathrm{EF}(C_\lambda) = \frac{\mathrm{Act}(C_\lambda)}{\mathrm{Act}\,(C)}$$

where $\mathrm{Act}(C_\lambda) = \frac{|C_\lambda^+|}{|C_\lambda|}$

and $\mathrm{Act}\,(C) = \frac{|C^+|}{|C|}$,

where, in turn, C is the total number of compounds tested, C^+ the number of compounds active, C_λ the number of total compounds with a scaffold λ tested, and C_λ^+ the number of compounds with a scaffold λ active against the target. Values above 1 imply a positively enriched scaffold (i.e., a scaffold that has a higher proportion of active compounds than the general dataset), while values below 1 have the opposite meaning.

3. EFs are susceptible of hypothesis testing. For finding statistically significant enriched scaffolds, chi-squared tests can be run using a 2 × 2 contingency table for the compounds considering as variables whether they have a given scaffold and whether they are active. Since sometimes values in the cells might be lesser than 5, and this interferes with the analytic calculation of the chi-squared statistic, simulated values can be obtained.
4. After running all p-values for every scaffold, the false discovery rate correction (or other method for correcting for multiple hypothesis testing) should be applied.

2.3.3 Degree of Polypharmacology

The methods for SAR analysis mentioned above are useful for single target studies. However, sometimes inhibition data of multiple targets are available for single compounds. These data could lead to polypharmacology studies. Maggiora and Gokhale recently formalized the notion of polypharmacology and polyspecificity [44]. In practical terms, the degree of polypharmacology of a molecule equals the number of different targets against which the molecule is active, while the analogous degree of polyspecificity of a target equals the number of different molecules that are active against the target.

2.3.4 Multiple Structure–Activity Relationship Analysis

A review addressing SmARt analysis in epigenetics was recently published [3]. Two of the most useful SmARt tools are methodologically explained in the following paragraphs: dual-activity difference (DAD) maps and structure–promiscuity index difference (SPID). Similarly as for other SAR analyses, Workflow 4 in the Supplementary Information contains practical tools for computing them.

Dual-Activity Difference Maps

DAD maps are designed to compare at once the activity of compounds against two biological endpoints, in contrast to SAS maps [45]. However, DAD maps lose structural information, which is accounted for with SAS maps. The procedure for generating a DAD map is straightforward:

1. Select a library of compounds with the activity of each independently measured against two different endpoints.
2. Plot in the x-axis one of the measurements and on the y-axis the other. A general form of a DAD map is presented in Fig. 3.

Structure–Promiscuity Index Difference

Aiming towards a statistic for quantifying the relationship between structural similarity and polypharmacology (or promiscuity), the SPID was created [46]. It is computed with the formula:

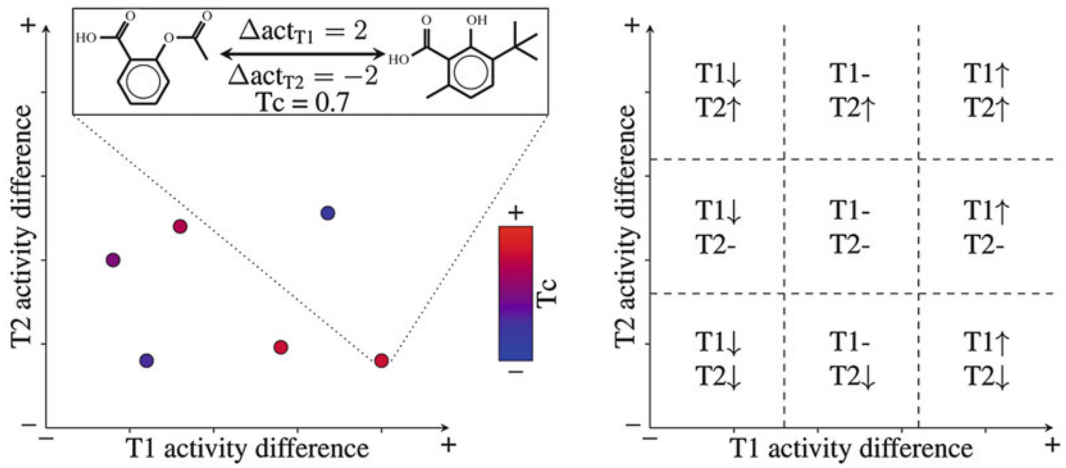


Fig. 3 Dual-activity difference (DAD) maps. Each data point represents a pair of compounds. The *x*-axis plots the activity difference of target 1, while the *y*-axis the activity difference of target 2. A color continuous scale might be added to the plot to represent chemical similarity of each pair of compounds. Up to nine regions can be distinguished depending on whether activity is conserved, increased, or decreased for any of the two targets. *Tc* Tanimoto coefficient, *T1* target 1, *T2* target 2

$$\mathrm{SPID}(\mathrm{A},\mathrm{B}) = \frac{|P_{\mathrm{A}} - P_{\mathrm{B}}|}{1 - T(\mathrm{A},\mathrm{B})}$$

where A and B are chemical compounds, P_{A} and P_{B} are the potencies of compounds A and B, respectively, and T(A, B) is the Tanimoto similarity of compounds A and B computed as in Sect. 2.2.2 using molecular fingerprints.

3 Chemical Space

Visual representations of the relationships of the compounds in a database are often useful for assessing libraries' diversity and SAR. Furthermore, the recent development of database fingerprints (DFPs) (described below) has made easier to chart multiple target-focused libraries in the chemical space, thereby providing polypharmacology insights [24]. Workflow 5 in the Supplementary Information illustrates a KNIME implementation of the methods described in this section.

3.1 Principal Components Analysis for Charting Compounds

There are no universal methods for chemical space representations [47, 48]. A commonly used approach involves calculating similarity matrices, which capture all the pairwise comparisons. These matrices are squared and have n columns and rows, with n equal to the number of compounds in the dataset. Finally, principal components analysis (PCA) as well as other dimensionality reduction methods is useful to compress most of the relevant information in a few variables. This makes possible to obtain visualizations of the chemical space. The concrete steps for creating visualizations of the chemical space using the approach presented above are as follows:

1. Select the set of descriptors with which the similarity or distance will be calculated. Common sets are: physicochemical properties (see Sect. 2.2.1) and molecular fingerprints (see Sect. 2.2.2). Compute the similarity matrix accordingly.
2. Apply PCA to the matrix. Select two or three principal components for plotting. It is useful to consider the percentage of variance captured with each principal component.

This method may become impractical for large datasets (>1000 compounds). See Sect. 3.3 for a chemical space visualization method that is less computationally expensive.

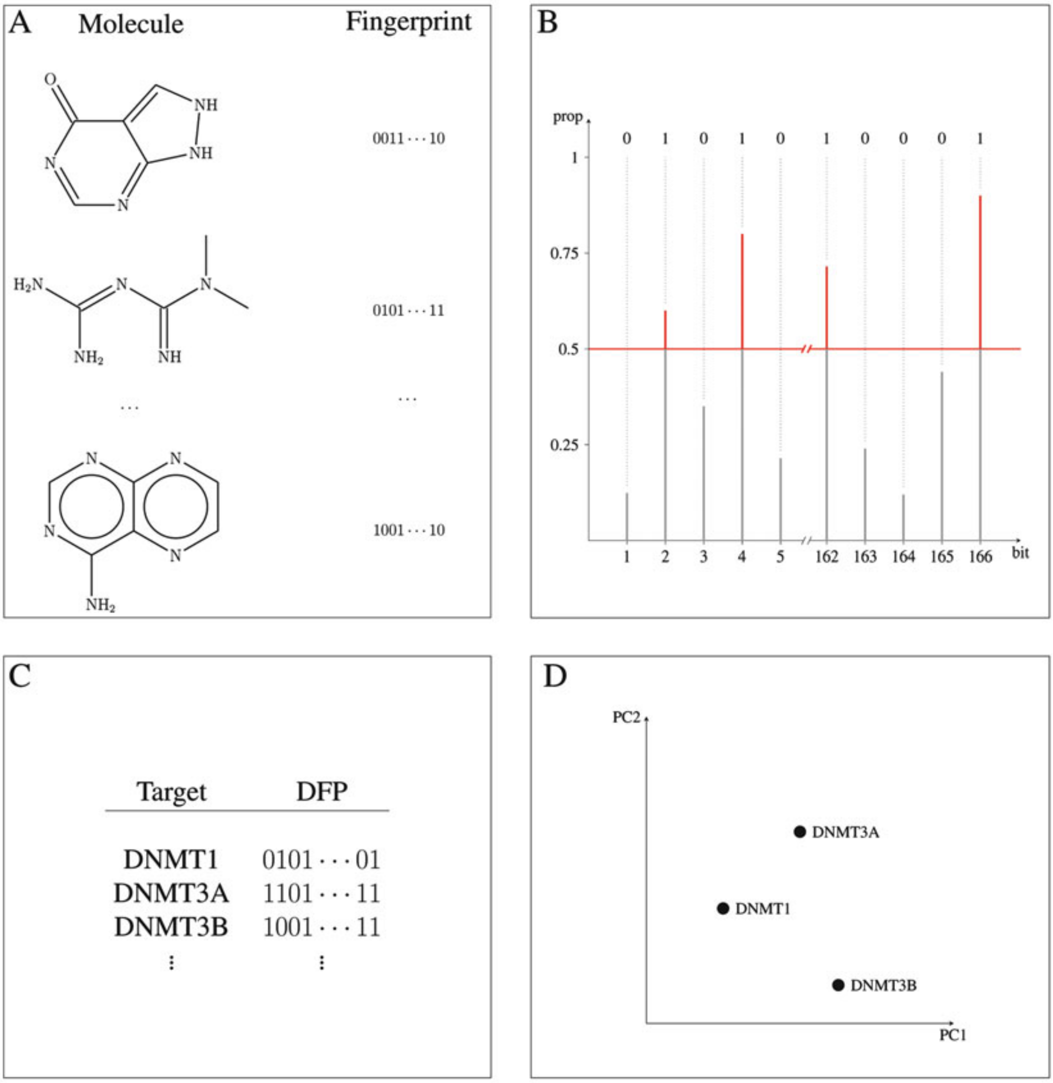
3.2 Comparing Multiple Libraries in the Chemical Space

DFPs are a recently introduced approach to simplify the representation of all compounds in a dataset using a single bit-vector for each database, thereby summarizing every individual fingerprints it contains. DFPs retain the predominant information captured in the molecular fingerprints of the molecules in a given chemical dataset. Briefly, if a given bit had a "1" value in at least 50 % of the compounds in the dataset, it is set to "1" in the DFP, or as "0" otherwise. Further details of the DFPs standardization are described elsewhere [49]. This adds only one step prior to chemical space visualization as commented in Sect. 3.1. If it is intended to include SAR in these plots, libraries could be filtered to include only active compounds. Figure 4 shows schematically the concept of DFPs.

3.3 ChemMaps

Several chemical space visualizations are based upon pairwise similarity measurements. Remarkably, computation of similarity matrices has exponential complexity. Thus, sometimes calculation times make impractical to chart the chemical space of more than 1000 compounds. ChemMaps aim at simplifying the computational task, by adaptively selecting some molecules in the database as comparison references or "satellites." This method reduces up to 30 % of the time needed for generating a visualization of the chemical space, depending on the size and diversity of the database [50]. The method is as follows:

1. Select at random 25 % of the compounds in a library to use as satellites.
2. Compute the pairwise similarity matrix of all the compounds against the satellites.
3. Perform PCA on the matrix and select the first two or three principal components.
4. Using the principal components as descriptors, compute the distance matrix for all the charted compounds or a subset.



Target	DFP
DNMT1	0101···01
DNMT3A	1101···11
DNMT3B	1001···11
⋮	⋮

Fig. 4 Database fingerprint (DFP). (**a**) For every compound in a chemical database, different kind of fingerprints might be obtained. (**b**) Usually, fingerprints store data in bits. If 50 % or more of the compounds in the database have a value of “1” for a given bit, then it is set as “1” in the DFP, otherwise it is set as “0.” (**c**) This procedure could be applied to many target-focused libraries. (**d**) DFPs of multiple libraries can be visualized to represent the chemical space of such libraries. DFPs can also be used for other applications, such as virtual screening. *DFP* database fingerprint

5. Add another 5 % of the database compounds to be used as satellites and repeat steps 2–4.
6. Calculate the correlation between the distances obtained with the PCA as descriptors and repeat step 5 until a correlation of 0.9 or higher is achieved.
7. Plot the chemical space. See Fig. 5.

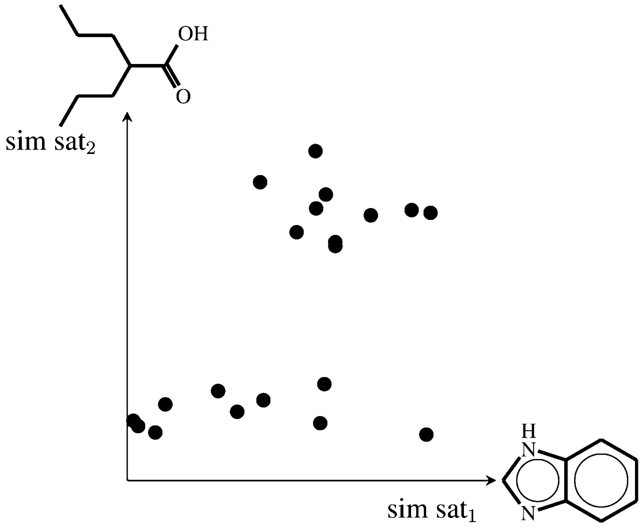


Fig. 5 ChemMaps concept. Chemical space is charted relative to adaptive chemical satellites. Two satellites are used in the example

3.4 Activity Landscape Sweeping

It is common that some structural clusters tend to form when analyzing the chemical space of libraries. Moreover, these clusters may also have different SAR morphologies, with a smoother or rougher application of the similarity principle [11, 51]. The SAR studies and their use for selecting clusters of molecules from a given library are named "activity landscape sweeping." Such approach is useful to characterize discrete regions in the chemical space where predictive methods that heavily rely upon the similarity principle could be applied. The method is quite straightforward:

1. As a baseline, compute the general SAS map for the whole library as described in Sect. 2.3.1.
2. Plot the chemical space as described in either Sect. 3.1 or 3.3.
3. For defining clusters in the chemical space, apply some method for unsupervised clustering, such as *k*-means. *K*-means method could use many principal components for defining the clusters. For selecting a number of principal components to use, a rule of thumb is to plot the contribution of variances of the principal components and select the "elbow" of the curve (i.e., the inflexion point whereupon adding more principal components do not significantly add information). Given that *k*-means also requires to a priori define the number of clusters, a similar procedure as that for selecting the number of principal components could be applied. However, instead of plotting the variances contribution, the within groups sum of squares is used. However, the number of clusters can also be defined visually by manually adjusting it.

4. Once that clusters of compounds are defined, individual SAS maps per cluster are plotted as described in Sect. 2.3.1.
5. The SAS maps and the proportions of activity cliffs are compared, in order to identify regions with smoother SAR.

4 Target Fishing

In polypharmacology, the identification of all the likely targets for a given chemical compound is of utmost importance and has been an active area of research in recent years [52]. This problem is known as reverse virtual screening or "target fishing" [53]. There is a plethora of computational approaches applied in this field. Chemoinformatics methods are mostly based on the principle of SAR [54] which suggests that similar compounds are likely to overlap between the sets of targets that they show activity against [55].

This identification of targets for a given compound can be carried out based on the similarity it presents with other compounds that are known to be active or inactive against some targets. If quantitative and comparable activity values are available, it is possible to build quantitative structure–activity relationships (QSAR) models [21, 56] for every target of interest. If the activity values are not completely reliable, a better alternative is the use of the categorical form of them to build machine-learning models for clustering and classification [57]. Although the general objective of most of these methodologies is the identification of targets for a given compound, the amount and type of biological information available can lead to various applications. This section describes the methodologies implicated in them.

4.1 Target Identification

The most general application of target fishing strategies consists of predicting all the possible targets for a given compound, or at least all of them for which bioactivity data is known. Most of these strategies treat the target fishing problem as a multi-label classification problem, in which every target is a label that a given compound belongs to and for which a predictive model is constructed [52, 58]. The main differences between different approaches are the molecular representation employed and the predictive models used. This work is not intended to provide a detailed description on the construction of these models, which can be found in several other works [59, 60], but of the general strategy for their application.

4.1.1 Multi-label Classifiers

One of the most used alternatives to face the target fishing problem is by building a multi-label classifier. The general steps to build such model are described below:

1. Given a set of targets of interest, a set of compounds, and a defined bipartite activity relation between them, construct and

curate compound databases for each target according to the methods discussed in Sect. 2.1.

2. Build and validate a binary classifier for each database, which allows to distinguish between active and inactive compounds. At this point lies the main difference between distinct models, because the pertinence of a compound to one class or another can be defined according to a priori defined thresholds for a given score. For instance, a similarity coefficient when dealing with similarity searches (discussed in Sect. 2.2), an activity value in the case of QSAR models, or the probability coming from a machine-learning model.
3. Finally, evaluate a compound of interest with all binary models. The targets associated to that compound will be those for which the binary classifiers assign a score higher than the established threshold.

The general scheme of a multi-label classifier is presented in Fig. 6a. The application of these types of strategies in drug design projects is discussed in other works [21, 61, 62] and currently there are several web implementations of these methods [58, 63].

4.1.2 Cluster Analysis

Another methodology to address the multi-label classification problem of target fishing is clustering, which is the task of grouping objects (compounds) such a way that objects belonging to the same group are more similar to each other in comparison to those belonging to other groups. This kind of methodologies only take into account the structure and properties of compounds known to

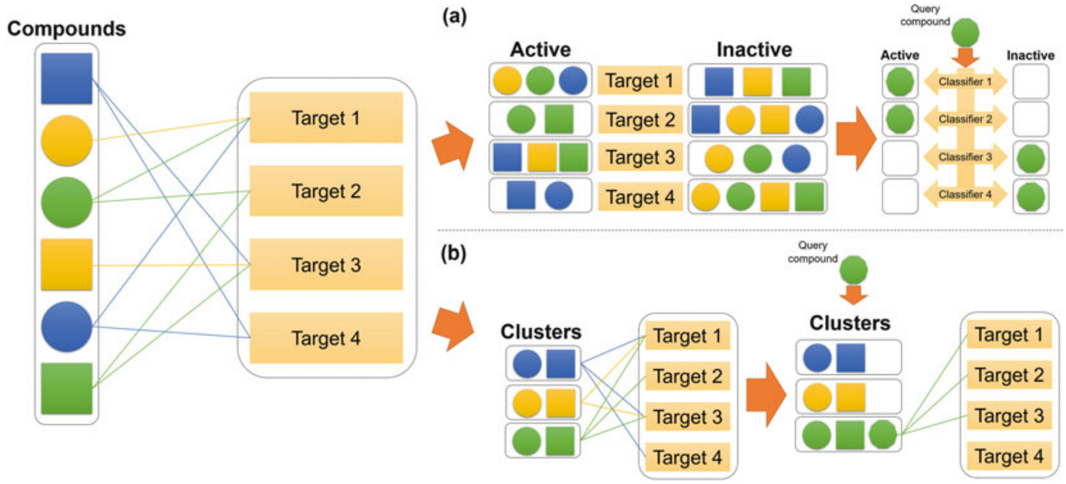


Fig. 6 (**a**) Representation of a multi-label classifier. The targets associated to the query compound are those for which the corresponding classifiers identify them in the active class. (**b**) Representation of a clustering analysis. The targets associated to the query compound are those associated to the cluster in which such compound is grouped