Methods in Molecular Biology 1762

Springer Protocols

Mohini Gore
Umesh B. Jagtap *Editors*

Computational Drug Discovery and Design

EXTRAS ONLINE

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Computational Drug Discovery and Design

Edited by

Mohini Gore

Department of Basic and Applied Sciences, Dayananda Sagar University, Bangalore, KA, India

Umesh B. Jagtap

Department of Biotechnology, Shivaji University, Kolhapur, MH, India;
Department of Botany, Government Vidarbha Institute of Science and Humanities, Amaravati, MH, India

Humana Press

Editors
Mohini Gore
Department of Basic and Applied Sciences
Dayananda Sagar University
Bangalore, KA, India

Umesh B. Jagtap
Department of Biotechnology
Shivaji University
Kolhapur, MH, India

Department of Botany
Government Vidarbha Institute of Science and Humanities
Amaravati, MH, India

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7755-0 ISBN 978-1-4939-7756-7 (eBook)
https://doi.org/10.1007/978-1-4939-7756-7

Library of Congress Control Number: 2018935920

Printed on acid-free paper

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.

Preface

Computer-aided drug design is an indispensable approach for accelerating and economizing the costly and time-consuming process of drug discovery and development. In the recent years, there has been a spurt in the protein and ligand structure data. This has led to a surge in the number of databases and bioinformatics tools to manage and process the available data. Optimal application of the vast array of available computational tools is crucial for the discovery and design of novel drugs.

The aim of this volume on *Computational Drug Discovery and Design* is to provide methods and techniques for identification of drug target, binding sites prediction, high-throughput virtual screening, lead discovery and optimization, and prediction of pharmacokinetic properties using computer-based methodologies. This volume includes an overview of the possible techniques of the available computational tools, developing prediction models for drug target prediction and de novo design of ligands. Structure-based drug designing, fragment-based drug designing, molecular docking, and scoring functions for assessing protein-ligand docking protocols have been outlined with practical examples. Phylogenetic analysis for protein functional site prediction has been described. Virtual screening and microarray studies for identification of potential compounds for drug discovery have been described using examples. The use of molecular dynamics simulation for virtual ligand screening, studying the protein-ligand interaction, estimating ligand binding free energy, and calculating the thermodynamic properties of bound water has been presented with stepwise protocols. In silico screening of pharmacokinetic and toxicity properties of potential drugs has been demonstrated. The currently available algorithms and software for protein-protein docking have been discussed with latest examples. Protocols for quantitative structure-activity relationship have been described. Computational approaches for the prediction of protein dynamics and protein aggregation have been presented with relevant protocols. The important methods of enhanced molecular dynamics have been analyzed with the help of practical procedure description. In silico analysis for inclusion of solvent in docking studies has been described with detailed methodology. We have also included a chapter on data analytics protocol, which is useful to summarize independent studies on drug designing.

There is abundant literature available on bioinformatics. However, there is very limited literature which will provide a step-by-step approach to utilize the various bioinformatics tools. In this volume, we present a stepwise description of the protocols for the use of bioinformatics tools in drug discovery and design. This volume will assist graduate and postgraduate students, researchers, and scholars working in the fields of drug discovery and design, pharmacology, bioinformatics, chemoinformatics, computational biology, medicinal chemistry, molecular biology, and systems biology to effectively utilize computational methodologies in the discovery and design of novel drugs.

We would like to express our heartfelt gratitude to the series editor John Walker for his valuable advice and support during every stage of development of this book. We thank all the authors who contributed to this book in a timely manner and shared their practical knowledge by providing stepwise methodology for the utilization of bioinformatics tools for drug discovery and design. We hope that this volume will be helpful to both novice in the field of bioinformatics and scientists actively engaged in drug discovery research.

Bangalore, KA, India ***Mohini Gore***
Kolhapur, MH, India ***Umesh B. Jagtap***

Contents

Contributors

Amr H. A. Abdallah • *Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, West Lafayette, IN, USA*

Samia Aci-Sèche • *Institut de Chimie Organique et Analytique (ICOA), UMR7311 CNRS-Université d'Orléans, Université d'Orléans, Orléans Cedex 2, France*

Matteo Aldeghi • *Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, Oxford, UK; Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*

Adriano D. Andricopulo • *Laboratório de Química Medicinal e Computacional, Centro de Pesquisa e Inovação em Biodiversidade e Fármacos, Instituto de Física de São Carlos, Universidade de São Paulo (USP), São Carlos, SP, Brazil*

David B. Ascher • *Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, VIC, Australia; Department of Biochemistry, University of Cambridge, Cambridge, UK*

Heval Atas • *Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey; Cancer Systems Biology Laboratory (CanSyL), METU, Ankara, Turkey*

Khaled Barakat • *Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada*

Damian Bartuzi • *Department of Synthesis and Chemical Technology of Pharmaceutical Substances with Computer Modelling Lab, Medical University of Lublin, Lublin, Poland*

Shaherin Basith • *National Leading Research Laboratory (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, Seoul, Republic of Korea*

Philip C. Biggin • *Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, Oxford, UK*

Joseph P. Bluck • *Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, Oxford, UK*

Pascal Bonnet • *Institut de Chimie Organique et Analytique (ICOA), UMR7311 CNRS-Université d'Orléans, Université d'Orléans, Orléans Cedex 2, France*

Abdennour Braka • *Institut de Chimie Organique et Analytique (ICOA), UMR7311 CNRS-Université d'Orléans, Université d'Orléans, Orléans Cedex 2, France; Centre de Biophysique Moléculaire (CBM), CNRS, UPR 4301, Orléans Cedex 2, France*

Yang Cao • *Center of Growth, Metabolism and Aging, Key Lab of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, People's Republic of China*

Ishwar Chandra • *Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India*

Radha Chaube • *Department of Zoology, Institute of Science, Banaras Hindu University, Varanasi, Uttar Pradesh, India*

Sun Choi • *National Leading Research Laboratory (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, Seoul, Republic of Korea*

Benjamin P. Cossins • *Computer-Aided Drug Design and Structural Biology, UCB Pharma, Slough, UK*
Georges Czaplicki • *Institute of Pharmacology and Structural Biology, UMR 5089, University of Toulouse III, Toulouse, France*
Wentao Dai • *Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China*
Julien Diharce • *Institut de Chimie Organique et Analytique (ICOA), UMR7311 CNRS-Université d'Orléans, Université d'Orléans, Orléans Cedex 2, France*
Tunca Doğan • *Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey; Cancer Systems Biology Laboratory (CanSyL), METU, Ankara, Turkey; European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK*
Tianhua Feng • *Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada*
Leonardo G. Ferreira • *Laboratório de Química Medicinal e Computacional, Centro de Pesquisa e Inovação em Biodiversidade e Fármacos, Instituto de Física de São Carlos, Universidade de São Paulo (USP), São Carlos, SP, Brazil*
Francis Gaudreault • *National Research Council Canada, Ottawa, Canada*
Jodi A. Hadden • *Department of Chemistry and Biochemistry, University of Delaware, Newark, DE, USA*
Mar Huertas • *Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA; Department of Biology, Texas State University, San Marcos, TX, USA*
Agnieszka A. Kaczor • *Department of Synthesis and Chemical Technology of Pharmaceutical Substances with Computer Modelling Lab, Medical University of Lublin, Lublin, Poland; School of Pharmacy, University of Eastern Finland, Kuopio, Finland*
Lisa M. Kaminskas • *School of Biomedical Sciences, University of Queensland, St. Lucia, QLD, Australia*
Shashank P. Katiyar • *Department of Biochemical Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, India*
Daisuke Kihara • *Department of Biological Science, Purdue University, West Lafayette, IN, USA; Department of Computer Science, Purdue University, West Lafayette, IN, USA*
Leslie A. Kuhn • *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA; Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA; Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA*
Anjani Kumari • *Department of Biochemical Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, India*
Priyanka Kumari • *Department of Biotechnology, Delhi Technological University, Delhi, India*
Alastair D. G. Lawson • *Computer-Aided Drug Design and Structural Biology, UCB Pharma, Slough, UK*
Yoonji Lee • *National Leading Research Laboratory (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, Seoul, Republic of Korea*

Siu-wai Leung • *State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China; School of Informatics, University of Edinburgh, Edinburgh, UK*
Weiming Li • *Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA*
Markus A. Lill • *Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, West Lafayette, IN, USA*
Vidhi Malik • *Department of Biochemical Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, India*
Dariusz Matosiuk • *Department of Synthesis and Chemical Technology of Pharmaceutical Substances with Computer Modelling Lab, Medical University of Lublin, Lublin, Poland*
Laurent Maveyraud • *Institute of Pharmacology and Structural Biology, UMR 5089, University of Toulouse III, Toulouse, France*
Grégory Menchon • *Laboratory of Biomolecular Research, Paul Scherrer Institute, Villigen PSI, Switzerland*
Zhichao Miao • *European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK; Wellcome Trust Sanger Institute, Cambridge, UK*
Louis-Philippe Morency • *Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada*
Rafael Najmanovich • *Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada*
Abhigyan Nath • *Department of Zoology, Institute of Science, Banaras Hindu University, Varanasi, Uttar Pradesh, India*
Umesh Panwar • *Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India*
Samuel Peña-Díaz • *Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Spain; Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Spain*
Juan R. Perilla • *Department of Chemistry and Biochemistry, University of Delaware, Newark, DE, USA*
Douglas E. V. Pires • *Centro de Pesquisas René Rachou, FIOCRUZ, Belo Horizonte, Brazil*
Jordi Pujols • *Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Spain; Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Spain*
Sebastian Raschka • *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA*
Sergey A. Samsonov • *Laboratory of Molecular Modeling, Department of Theoretical Chemistry, Faculty of Chemistry, University of Gdańsk, Gdansk, Poland*
Ricardo N. dos Santos • *Departamento de Físico-Química, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil*
Anne M. Scott • *Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA*
Jana Selent • *GPCR Drug Discovery Group, Research Programme on Biomedical Informatics (GRIB), Universitat Pompeu Fabra (UPF)-Hospital del Mar Medical Research Institute (IMIM), Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain*
Jiye Shi • *Computer-Aided Drug Design and Structural Biology, UCB Pharma, Slough, UK*

Woong-Hee Shin • *Department of Biological Science, Purdue University, West Lafayette, IN, USA*

Kamya Singh • *Department of Biochemical Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, India*

Sanjeev Kumar Singh • *Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India*

Tomasz Maciej Stępniewski • *GPCR Drug Discovery Group, Research Programme on Biomedical Informatics (GRIB), Universitat Pompeu Fabra (UPF)-Hospital del Mar Medical Research Institute (IMIM), Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain*

Durai Sundar • *Department of Biochemical Engineering and Biotechnology, DBT-AIST International Laboratory for Advanced Biomedicine (DAILAB), Indian Institute of Technology Delhi, New Delhi, India*

Venkatesan Suryanarayanan • *Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India*

Alan Talevi • *Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB), Faculty of Exact Sciences, National University of La Plata (UNLP), Buenos Aires, Argentina; Argentinean National Council of Scientific and Technical Research (CONICET), Buenos Aires, Argentina*

Nurcan Tuncbag • *Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey; Cancer Systems Biology Laboratory (CanSyL), METU, Ankara, Turkey*

Salvador Ventura • *Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Spain; Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, Spain*

Ying Yang • *Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, West Lafayette, IN, USA*

Sze Chung Yuen • *State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China*

Hongmei Zhu • *State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China*

Sonia Ziada • *Institut de Chimie Organique et Analytique (ICOA), UMR7311 CNRS-Université d'Orléans, Université d'Orléans, Orléans Cedex 2, France*

Chapter 1

Computer-Aided Drug Design: An Overview

Alan Talevi

Abstract

The term drug design describes the search of novel compounds with biological activity, on a systematic basis. In its most common form, it involves modification of a known active scaffold or linking known active scaffolds, although de novo drug design (i.e., from scratch) is also possible. Though highly interrelated, identification of active scaffolds should be conceptually separated from drug design. Traditionally, the drug design process has focused on the molecular determinants of the interactions between the drug and its known or intended molecular target. Nevertheless, current drug design also takes into consideration other relevant processes than influence drug efficacy and safety (e.g., bioavailability, metabolic stability, interaction with antitargets).

This chapter provides an overview on possible approaches to identify active scaffolds (including in silico approximations to approach that task) and computational methods to guide the subsequent optimization process. It also discusses in which situations each of the overviewed techniques is more appropriate.

Key words ADMET, Anti-target, Computer-aided drug design, Ligand-based approaches, Molecular optimization, Pharmacophore, QSAR, Structure-based approaches, Target-based approaches, Virtual screening

1 Introduction

The term drug design describes the search of novel compounds with biological activity, on a systematic, rational basis. Basically, it relies on experimental information of the intended molecular target or a similar biomolecule (direct drug design) and/or known binders of such target (indirect drug design). Lately, however, the idea of using direct or indirect structural information on relevant antitargets has gained increasing attention to improve ligand selectivity and reduce off-target interactions, leading to enhanced safety and even improved pharmacokinetic profile [1–4]. In other words, modern drug design not only relies on available molecular information on the proposed molecular targets but also on the information on antitargets.

In its most common form, drug design involves modification of a known active scaffold (molecular optimization) or linking known

Mohini Gore and Umesh B. Jagtap (eds.), *Computational Drug Discovery and Design*, Methods in Molecular Biology, vol. 1762, https://doi.org/10.1007/978-1-4939-7756-7_1,

active scaffolds, although de novo drug design (i.e., from scratch) is of course also possible (e.g., fragment growing approximations). In any case, a starting point or seed is required to build up or optimize the active compound. Computer-aided methods have gained a prominent role in both stages of modern drug discovery: searching for starting points and making rational decisions regarding which chemical modifications are more convenient to introduce to them.

Whereas in silico or virtual screening (VS) (i.e., using computational methods to explore vast collections of chemicals and identify novel active scaffolds) represents a rational way of finding starting points to implement a drug design campaign, it should be conceptually separated from drug design. Drug design is intrinsically and unequivocally related to finding *molecular novelty*, that is, novel chemical entities. Novelty is the key, underlying drug design. In contrast, in silico screening, which can be and usually is coupled with drug design, typically explores the known chemical universe in search of new active motifs. The novelty in virtual screening is not in the chemistry of the emerging hits, but in uncovering an unknown, hidden association between known chemicals and a given biological activity. There are, however, many alternatives to in silico screening to discover such association.

Besides its rationality, an attractive aspect of computer-aided drug design is its accessibility. The technology gap between high- and low-income countries is smaller for computer-aided drug discovery than for any other process or approach in the drug discovery cycle. This is in part because many computational resources and applications have been made publicly available, and many computational tools used in the field run fairly smoothly in any modern personal computer.

It should be emphasized, though, that several constrains operate on the process of drug design. First, synthetic feasibility of the designed compounds should not be neglected [5]. A proposed compound might not be synthetically attainable due to universal technical reasons (lack of a given synthetic route) or to local limitations (e.g., lack of access to required technology and/or reactants, expensive synthesis). Equally important is the fact that drug discovery is a challenging multiobjective problem, where numerous pharmaceutically relevant objectives should be simultaneously addressed [6], a problem further complicated by the fact that, occasionally, some of those objectives might be conflicting, resulting in very complex solution spaces. For example, it is in general accepted that higher selectivity leads to safer medications; however, efficacious treatments for complex disorders might require multitarget therapeutic agents which, by definition, are not exquisitely selective [7]. On the other hand, as implicit in the famous Lipinski's rule of five and similar rules of thumb [8], a certain degree of aqueous solubility is often pursued to assure absorption, but an excessive solubility could be detrimental to absorption and

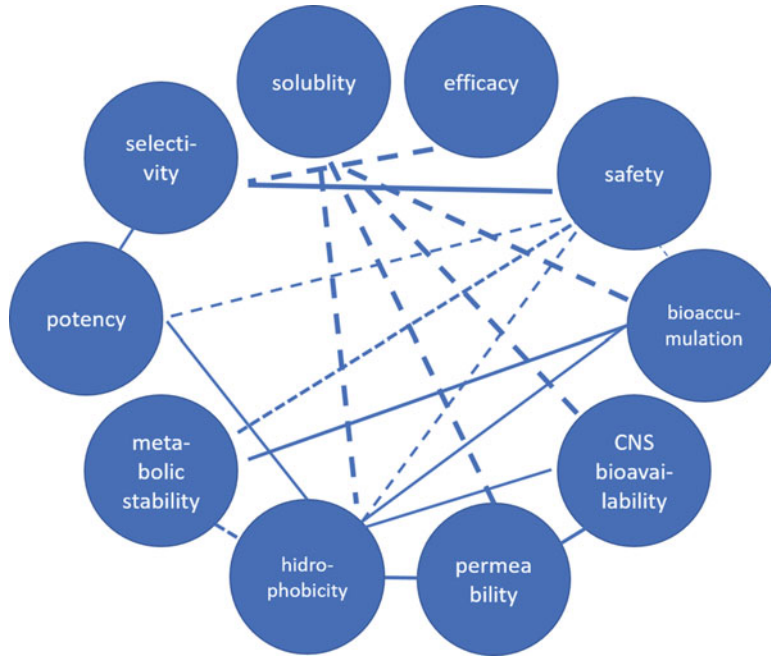

Fig. 1 A complex, conflicting interplay is observed between pharmaceutically relevant properties that are taken into consideration when facing a drug design project. An inverse, possibly conflicting relationship between two properties is indicated by a dashed line. Oppositely, a direct, favorable relationship is shown with a continuous line

biodistribution. Introduction of lipophilic substituents into adequate positions of a ligand often translates into a gain in potency [8], and certain degree of lipophilicity is also desirable in central nervous system medications to achieve brain bioavailability [9]. However, high lipophilicity conspires against both drug dissolution [10] and metabolic stability [11]. The key word in drug design seems to be balance, which explains why multiobjective optimization methods have gained such popularity in the field in the past years [6, 12].

A scheme illustrating the complex interplay between some pharmaceutically important drug properties is shown in Fig. 1. Naturally, the scheme is an oversimplification. The nature of the relationship between two properties might not be linear and many counterexamples to the illustrated relationships can be found, e.g., while it is accepted that lipophilicity has a positive impact on cell permeability, excessively high lipophilic drugs might become sequestered inside the cell, with little improvement on permeability across biological barriers (prominently, endothelial and epithelial tissues) [13], thus determining a parabolic relationship between lipophilicity and permeability. In general, hit identification is potency-driven, preferring ligands with affinities in the nM range. Whereas potent ligands are undoubtedly pursued in some cases

(e.g., to treat anti-infectious diseases), the most potent ligand might not be the first choice if trying to restore sensitive physiologic systems (e.g., the brain or the heart) to its normal functioning, since highly potent ligands will tend to impair normal functioning and produce intolerable side effects.

This chapter overviews the computational approaches that can be used to find novel active chemotypes and guide the subsequent molecular optimization. General principles of rational drug design are also tangentially visited.

2 Where to Start—The Value of Novelty

A critical question when conceiving a drug discovery project is where to start. Obviously, any target-driven drug discovery project today starts by choosing one (single-target agents) or more (tailored multitarget agents) drug targets. What makes a good drug target? First, it must be disease-modifying. Second, it must be druggable, that is, it should be modulated by binding a small molecule or, according to some authors, a biologic [14, 15]. If no ligand is known to bind the potential target, druggability prediction can be performed, which generally involves examining the target surface for binding sites, or checking the existence of similar proteins which have already proven to be druggable [16–18]. Other desirable features include assayability, differential expression throughout the body and a favorable intellectual property situation (no competitors focused working on the same target!) [14].

Second, if we exclude entirely de novo approximations, where forcefully one should begin from a model of the molecular target, any other approach requires a starting (and hopefully novel) active scaffold (ligand) into which chemical modifications are introduced.

Leaving aside serendipitous discoveries (which are of course useful but unsystematic), hints on potential active scaffolds of natural origin can be found in traditional medicine. Alternatively, one might resort to information on the natural ligand of an intended molecular target to start a drug design project.

At this point, it is worth emphasizing that chemical structural novelty is a key factor in the pharmaceutical sector. Novelty is a fundamental requisite to obtain intellectual property rights on an invention (and thus exclusivity). And although recently drug repurposing (finding new medical uses to already known drugs) has raised considerable interest within the health community, it also faces nontrivial intellectual property, regulatory, and commercial challenges [19, 20]. Accordingly, the search of novel active chemotypes remains a priority within the pharmaceutical industry due to their intellectual property potential.

High-throughput screening (HTS) methods are among the most frequent approaches to explore the vast universe of known chemicals in search of novel active scaffolds. It is the modern version of the traditional trial-and-error, "exhaustive" screening. The rationality of HTS lies in the integration of automation and miniaturization to the screening process, which results in efficient exploration of the chemical space [21]. Moreover, the approach has been greatly improved by the design of target-focused libraries [22] and the recognition of privileged scaffolds [23] (molecular frameworks/building blocks that are present in many biologically active ligands against a diverse array of targets). However, it should be mentioned that HTS requires very expensive technological platforms which are not frequently found in the academic sector or low- and middle-income countries.

In contrast, VS requires considerably more accessible technology, with many resources being completely publicly available, from specialized software to online chemical repositories. The term VS refers to the application of a diversity of computational approaches to rank digital chemical collections or libraries in order to establish which compounds are more likely to obtain favorable results when experimentally tested through in vitro and/or animal models. They have been conceived to minimize the volume of experimental testing and optimize the results, thus being advantageous in terms of cost-efficiency, bioethics, and environmental impact.

VS approaches can be essentially classified in two categories: structure-based (or direct or target-based) and ligand-based (or indirect) approximations.

Molecular docking is prominently used for structure-based VS. Starting from an experimental structure of the target (or, at worst, a homolog from other species or another protein belonging to the same family, i.e., comparative or homology modeling), the binding event is simulated and a scoring function is used to predict, for the most likely binding poses, the free energy difference due to the binding of the screened compounds to the target. While rigid (computationally undemanding) or more accurate, flexible (computationally demanding) approximations are possible, docking can be considered a computationally demanding VS approach in comparison with ligand-based methods. A search/sampling algorithm is used to generate a diversity of ligand-binding orientations (rigid-body approximations) or ligand binding orientations and conformations (flexible approximations). A major obstacle for the implementation of structure-based VS approaches comes from the fact that the structures of many validated drug targets have not yet been solved experimentally. Another caveat of docking relates to the empirical nature of scoring functions, which in general, depending on the type of scoring function, include a variable degree of parameterization. This limits the reliability of the method, plagued by a high incidence of false positives [24]. Since the scoring functions

are parameterized/trained against a number of experimentally determined binding affinities or experimental structures, the performance of the docking approach tends to be highly system-dependent and scores are, at best, weakly predictive of affinities [25]. Results are sometimes improved when different scoring functions are combined into a consensus score [25]. A persistent problem of the scoring function is the elusive entropic contribution to free-energy [24, 26] which is ignored in many cases or very approximately estimated in others. The reader should remember that, upon the binding event, the ligand will lose translational, rotational, and conformational freedom, whereas the target will mostly lose conformational freedom. The contributions of desolvation and water molecules mediating ligand–protein interactions (which also impact the initial and final entropy of the system) should not be neglected either [27, 28], but often are. Free energy simulations, which employ molecular dynamics or Monte Carlo simulations, provide a much more rigorous solution to binding free-energy estimation [24, 29, 30]. The emergence of low cost parallel computing is starting to relegate docking to the role of a prescreening tool, in favor of molecular dynamics-based VS [24, 29]. *See* Fig. 2 for a caption of a ligand–protein interaction simulation.

Ligand-based approximations may be applied whenever a model of the target structure is not available or to complement structure-based approximations. Concisely, ligand-based screening methods can be classified into similarity searches, machine learning approaches (prominently, supervised machine learning used in the frame of the Quantitive Structure–Activity Relationhip—QSAR—theory) and superposition approximations [31–33]. These techniques differ in a number of factors, from their requisites to their active enrichment or scaffold hopping.

Similarity search employs molecular fingerprints obtained from 2D or 3D molecular representations, comparing database compounds with one or more reference molecules in a pairwise manner. Remarkably, only one reference molecule (e.g., the physiologic ligand of a target protein) is required to implement a similarity-based VS campaign. Similarity searches are frequently the only option to explore the chemical universe for active compounds when lacking experimental knowledge on the target or related proteins, or when the number of known ligands is too small and impedes using supervised machine learning approaches.

Supervised machine learning approaches operate by building models from example inputs to make data-driven predictions on the database compounds. Machine learning approximations require several learning or calibration examples. The general model development protocol involves dataset compilation and curation (*see* **Note 1**); splitting the dataset into representative training (calibration) and test (validation) sets (whenever the size of the database allows it) (*see* **Note 2**); choosing which molecular descriptors

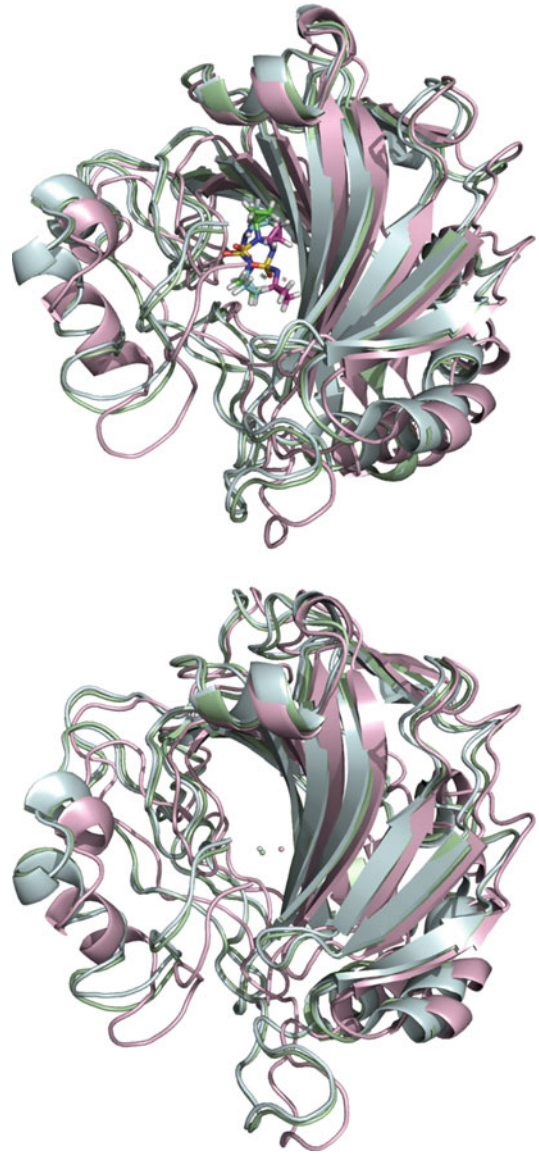

Fig. 2 Snapshots from a molecular dynamics simulation of the interaction between anticonvulsant sulfamides and carbonic anhydrase. Note the significant conformational changes induced by the ligand binding event

should be included into the model (*see* **Note 3**); weighing the contribution of such descriptors to the modeled activity; validating the model internally and externally and; checking the applicability domain of the model whenever a prediction is made [34]. Molecular diversity of the training samples is critical for VS applications of supervised machine learning: the molecular diversity of the calibration examples is directly correlated with a wide applicability domain of the resulting model.

Finally, superposition techniques are conformation-dependent methods that analyze how well a compound superposes onto a reference compound or, more frequently, how well they fit a fuzzy model (pharmacophore) in which functional groups are stripped off their exact chemical nature to become generic chemical properties relevant for the ligand–target interaction (hydrophobic points, H-bond donor, H-bond acceptors, charged groups, etc.). The pharmacophore is thus a geometric, 3D arrangement of generic, abstract features which are essential for the drug–target recognition event. Some approaches that have been used for pharmacophore generation can also include negative features (features that conspire against biological activity) in the model. In contrast with docking, which considers the key features required for drug–target interaction in a direct manner, superposition techniques do the same in an indirect way, by inferring such features from known ligands. Superimposition methods are, by far, the most visual, easy to interpret and physicochemically intuitive ligand-based approaches. The process is facilitated if the modeler counts on an active rigid analog with limited conformational freedom. Usually, though, one may resort to flexible alignment (superimposition) of a set of flexible ligands, either generating a set of low energy conformations and considering each conformer of each ligand in turn or exploring conformational space on the fly, i.e., the conformational search is performed simultaneously to the pattern identification stage (alignment stage) [35, 36]. It should be noted that, when applying pharmacophore-based VS, orientation sampling is probably as important as conformational sampling, since chemical diversity is expected in the screened chemical library and defining an orientation criteria is thus nontrivial. It should also be mentioned that structure-based pharmacophores are also possible [37].

Which in silico screening method should be chosen to start a rational drug discovery project? Of course, as indicated in the preceding paragraphs, the selection is restricted by the available data (structure-based approaches require experimentally solved 3D structure of the target or similar target; supervised machine learning requires a minimum of calibration samples, and so on.). But even if the technical requirements to implement any approach were met. . .is there a single approach that universally, consistently outperforms the remaining ones? Is there a method of first choice?

As a rule, the more complex approximations (structure-based approaches and, then, pharmacophore superposition) are the most advantageous in terms of scaffold hopping (they retrieve more molecular diverse hits), while simpler approaches are computationally more efficient while simultaneously achieving good active enrichment metrics [38]. Furthermore, structure-based approaches and pharmacophores explain, in an explicit or implicit way, respectively, the molecular basis of ligand–target interaction. They are visual and easily interpretable, two points which are not covered

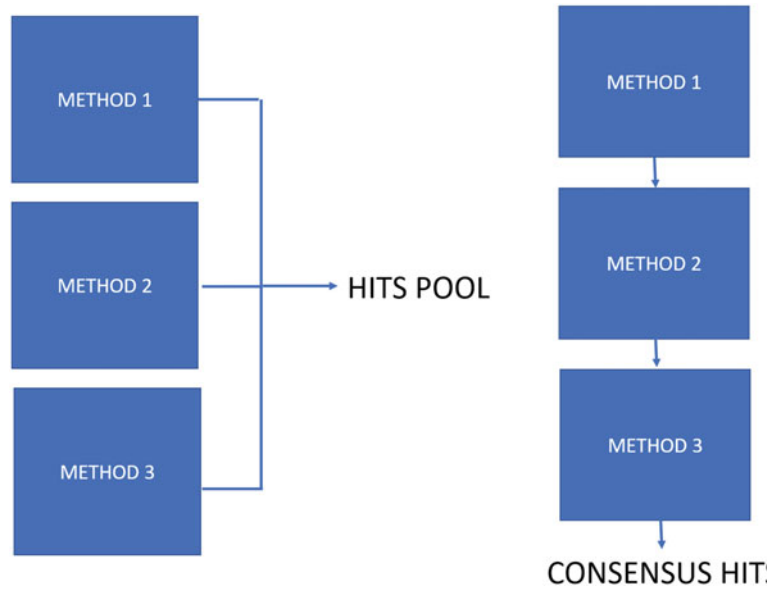

Fig. 3 While in parallel VS hybrid methods result in combination of complementary sets of hits (thus retrieving more chemical diversity), serial hybrid methods tend to produce more robust, consensus hit sets

by other approximations. These advantages should not be underestimated. Not only are they important from an epistemological perspective (they provide results and explanations), they also provide a visual support to their predictions and visual support is extremely important to communicate results to nonspecialized audiences (e.g., scientific collaborators from other fields, investors). Having said so, one should have in mind that the efficacy of a given technique is highly dependent on the chosen molecular target. Regarding VS approaches, a gold standard has not been found yet, a fact that explains the need of rigorous in silico validation before moving to VS and subsequent wet experiments. Some validation approaches are briefly discussed (*see* **Note 4**).

Frequently, different techniques are complementary in nature [39] and the simplest methods have surprisingly good outcomes in some cases. This allows the definition of hybrid protocols combining simple and complex approximations either serially or in parallel [40] (Fig. 3); serial combined approaches tend to provide robust solutions.

A final and important step to prune the hits emerging from systematic screening involves filtering out promiscuous compounds, unspecific inhibitors and reactive compounds, such as PAINS and REOS filters [41, 42].

3 The Actual Design: Hit to Lead and Beyond

Let us assume that one or more hits have emerged from systematic (wet or in silico) screening (or, maybe, that a starting active scaffold has been obtained from natural ligands of the intended target or from traditional medicine or from a serendipitous observation). The actual drug design process starts here, and involves introducing changes to the active scaffold in order to optimize the interaction with the target thus gaining potency, and/or to provide selectivity in relation to nontargeted similar proteins (e.g., nontargeted isoforms). Today, the optimization of other pharmaceutically relevant properties (e.g., chemical and biological stability) is also considered. Hits emerging from VS are usually active in the μM range (or, at best, in the high nM range) [43, 44]. A similar scenario has been observed in HTS campaigns [45]. Molecular optimization will usually decrease the dissociation (affinity) constant in about two orders of magnitude. From the 1990s onward, however, the pharmaceutical sector has understood that potency is not the only property to take into consideration, a realization that was expressed in the adoption of the "fail early, fail cheap" philosophy with the inclusion of in silico in vitro absorption, distribution, metabolism, excretion, and toxicity (ADMET) filters in the early stages of drug discovery [46, 47] and the emergent interest in low affinity ligands within certain therapeutic categories [48]. Classical optimization strategies include extension, ring variations, ring expansion or contraction, bioisosteric replacement and rigidification. In the case of (complex) active compounds of natural origin, simplification is also explored.

With the exception of similarity methods, which are of no use for optimization purposes, all the other approaches described in Subheading 2 of the chapter can be used to guide optimization. If the structure of the intended target has been solved, docking and structure-based pharmacophores are the first choices to guide optimization. They are the only methods that allow exploring, in a rational manner and without the need of trial and error learning, interactions with regions of the target that have not been exploited with previously known ligands. Among ligand-based approximations, pharmacophore superposition is the friendliest approach to molecular optimization. However, the QSAR approach is also suitable for design purposes, guiding the substitutions made onto the active scaffold; moreover, the *inverse QSAR* approach (in which, from molecular descriptors, new molecules having the desired activity could be "recovered") are also suitable for design of de novo molecules [49–51]. It should be noted that, while classification models are useful for VS campaigns, since they can compensate model errors related to data compiled from different laboratories, outlier compounds and mislabeled data points [34], when the

QSAR model is meant for optimization purposes regression modeling can be particularly useful, since the training dataset is usually synthesized inhouse and experimentally tested in the same laboratory. Furthermore, whereas VS applications require chemically diverse datasets, QSAR models used in optimization campaigns would typically display a narrower applicability domain, since they are obtained from a set of compounds with a common scaffold which has been modified to explore the surrounding chemical space.

4 In Silico ADMET Filters and Antitargets

From the 1990s onward, the search of more potent derivatives of an active scaffold has been balanced with early detection of potential bioavailability and toxicity issues. As a result, in silico and in vitro ADME filters are now fully integrated in the early stages of drug discovery and development. Such strategy has resulted in an impressive reduction of project termination rates related to ADME issues [46, 47] though pharmacokinetics and bioavailability still represent a significant cause for attrition at Phase I clinical trials [52–54]. Toxicology failures (both at preclinical and clinical stage) represent one of the key challenges still facing the pharmaceutical industry [52–54].

The earliest ADME filters involved simple rules of thumb derived from distribution analysis of physicochemical properties of drugs having or lacking a desired behavior. Lipinski's rule of five at Pfizer pioneered this kind of analysis [8], which was later followed by other similar rules related to the prediction of drug bioavailability, such as Veber's [55]. This trend was also explored in relation to toxicity, e.g., the "3/75" rule [56]. Later, however, arguments have been raised against rigid implementations of these kinds of rules [57], and the possible advantages of moving beyond the "rule of five" chemical space for difficult targets have been emphasized [58, 59], as well as notable systematic exceptions to this rule (e.g., natural products) [59, 60]. Lipinski himself, when first reporting his famous rule, recognized that acceptable drug absorption depended on the triad "potency–permeability–solubility", and that his computational alert did not factor in drug potency (a point of his analysis that is often overlooked) [8]; he also recognized the potential contribution of drug formulation to oral bioavailability, a contribution that can be addressed today through in silico tools [61].

It has been suggested that control of physicochemical properties is unlikely to have a significant effect on attrition rates; moreover, if a safety issue results from the primary drug target mechanism or from specific off-target interactions (e.g., hERG channel blockade), it is unlikely that physicochemical properties

would be predictive of toxicity [52]. A similar point could be made regarding prediction of bioavailability issues linked to specific interactions with enzymes (e.g., CYP450 enzymes) or transporters (e.g., ABC efflux transporters). In these cases, using previously discussed computational tools (docking, pharmacophores, QSAR models) in connection with the antitarget concept could be more advantageous.

The use of more complex (yet simple) multiparameter algorithms that address the interplay of physicochemical properties could also prove rewarding [12].

5 Final Remarks

We have presented an overview of the most relevant methods used in computer-aided drug design. While human beings (and scientist in particular) are naturally inclined to a way of thinking based on pattern recognition and identification of generalities, successful drug design comprises such a complex interplay between a number of objectives (e.g., efficacy, safety, and desired physicochemical properties) that the drug designer should beware oversimplification and dogmatic principles, which may lead not only to bad decisions, but also to loss of opportunities and novelty.

As the name itself suggests, drug design per se resembles an attentive artisan craftwork. The screening stages and the application of ADMET-related computational alerts, in contrast, involve more automated decisions, compatible with the idea of efficient exploration and fast pruning of a vast chemical universe. Fast pruning usually leads, however, to an over reduced chemical space. Flexible decision rules should be preferred over rigid ones, since they expand the borders of the more frequently explored regions of the chemical universe.

The decision to stop a drug candidate for toxicological or pharmacokinetic reasons involves complex and subtle judgements that should take into consideration cost–benefit analysis and available options to compensate the predicted difficulties (e.g, formulation alternatives, targeted-drug carriers). It is advised to be careful with excessive automation, to favor critical case-by-case decision-making as much as possible and to consider difficulties in a multidisciplinary way, including contributions of different professionals involved in the drug discovery cycle at each stage of the drug project.

6 Notes

1. Compiling and curating a dataset is one of the most important steps in supervised machine learning. The dataset will be used to infer the model and to validate it. The inferred model will

only be as good as the biological or biochemical data from which it is derived: the unknown noise in the training data is one of the factors that influence the generalization error. It is accepted that biochemical data (e.g., dissociation constants) are cleaner than biological data. The activities of all the training instances should be of comparable quality. Ideally, they should have been measured in the same laboratory under the same conditions, so that variability in the measured biological or biochemical activity only (or mostly) reflects treatment variability. This requirement is often accomplished when building models for optimization purposes from a series of inhouse synthesized compounds, but rarely met when building models for VS purposes (in this case, the need of a large and diverse training set frequently leads to compile experimental data from different laboratories).

Data distribution should be studied in order to avoid poorly populated regions within the studied chemical space as well as highly populated narrow intervals: extrapolation is forbidden but intrapolation in regions which are poorly populated by training examples is also risky. The dependent variable should span at least two or three orders of magnitude, from the least to the most active compound, and it should be (if possible) uniformly distributed across the range of activity (rarely achieved). The inclusion of leverage points (outliers, i.e., data exceptions represented by extreme values in the descriptor or response space which is not due to measurement or labelling errors) is discouraged.

Conscientiously curate the dataset: read data sources carefully and remove training examples extracted from inadequate or dubious experimental protocols. There are currently several databases that compile experimental data for small molecules (e.g., ChEMBL); such resources are manually curated from primary scientific literature. ChEMBL developers flag activity values that are outside a range typical for a given activity type, possibly missing data and suspected or confirmed author errors. Classification models can be used to alleviate the influence of data heterogeneity; they are useful for VS applications but less practical for models intended for optimization purposes.

Not only experimental data but also chemical structures should be curated. Do not underestimate the importance of this step: it is quite common that medicinal chemistry papers and chemical databases include structural mistakes. Remove those data points that are usually not handled by conventional cheminfomatic techniques: inorganic and some organometallic compounds, counterions, salts and mixtures (there exist molecular descriptors, however, that can be used to characterize ionic species if the dataset molecules are charged at the biologically