Cloud-Based High Throughput Virtual Screening in Novel Drug Discovery

Abdurrahman Olğaç[1,2(✉)], Aslı Türe[3], Simla Olğaç[4], and Steffen Möller[5]

[1] Laboratory of Molecular Modeling, Evias Pharmaceutical R&D Ltd., 06830 Ankara, Turkey
aolgac@evias.com.tr
[2] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Gazi University, 06330 Ankara, Turkey
abdurrahman.olgac@gazi.edu.tr
[3] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Marmara University, 34668 Istanbul, Turkey
asli.demirci@marmara.edu.tr
[4] Department of Biochemistry, Faculty of Pharmacy, Gazi University, 06330 Ankara, Turkey
csimla@gazi.edu.tr
[5] Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA), Rostock University Medical Center, 18057 Rostock, Germany
steffen.moeller@uni-rostock.de

Abstract. Drug discovery and development requires the integration of multiple scientific and technological disciplines in chemistry, biology and extensive use of information technology. Computer Aided Drug Discovery (CADD) methods are being used in this work area with several different workflows. Virtual screening (VS) is one of the most often applied CADD methods used in rational drug design, which may be applied in early stages of drug discovery pipeline. The increasing number of modular and scalable cloud-based computational platforms can assist the needs in VS studies. Such platforms are being developed to try to help researchers with various types of applications to prepare and guide the drug discovery and development pipeline. They are designed to perform VS efficiently, aimed to identify commercially available lead-like and drug-like compounds to be acquired and tested. Chemical datasets can be built, libraries can be analyzed, and structure-based or ligand-based VS studies can be performed with cloud technologies. Such platforms could also be adapted to be included in different stages of the pharmaceutical R&D process to rationalize the needs, e.g. to repurpose drugs, with various computational scalability options. This chapter introduces basic concepts and tools by outlining the general workflows of VS, and their integration to the cloud platforms. This may be a seed for further inter-disciplinary development of VS to be applied by drug hunters.

Keywords: Drug discovery · Virtual screening · High performance computing · Cloud computing

J. Kołodziej and H. González-Vélez (Eds.): cHiPSet, LNCS 11400, pp. 250–278, 2019.
https://doi.org/10.1007/978-3-030-16272-6_9

1 Introduction

Pharmaceutical drug discovery is a long-lasting and costly process, spanning over 12 to 15 years and costing about 1–2 billion US Dollars [1]. The process to identify new active pharmaceutical ingredients (API) [2] starts with target identification and validation steps and follows hit identification, lead discovery and lead optimization to acquire safe and effective new drug molecules at the preclinical stage [3]. Biological screening is used to identify possible target of a hit molecule as a developable drug-candidate as the first step in drug discovery. Advances in systematic biological screening have generated automated parallel biological screening technologies, called high throughput screening (HTS) [4]. Virtual screening (VS) is a widely applied computational approach, which is performed as a hit identification method in early stages of drug discovery pipeline. VS protocols involve searching chemical libraries to identify hit compounds with a putative affinity for a specific biological target (enzyme or a receptor) for further development.

CADD methods in conjunction with VS studies emerged as valuable tools to speed up this long process and limit the cost expansion of R&D. Such studies demand a strong combination of computational resources and skills, biochemical understanding and medicinal motivation.

Effects of drugs in the human body are investigated with two main parameters, namely pharmacokinetics and pharmacodynamics [5]. While pharmacokinetic studies investigate the fate of drug substances during absorption, distribution, metabolism and elimination (ADME) processes, pharmacodynamics determines the required concentration of drug to be delivered at the site of action and the biochemical and physiological effect, which may be responsible for the targeted biological response.

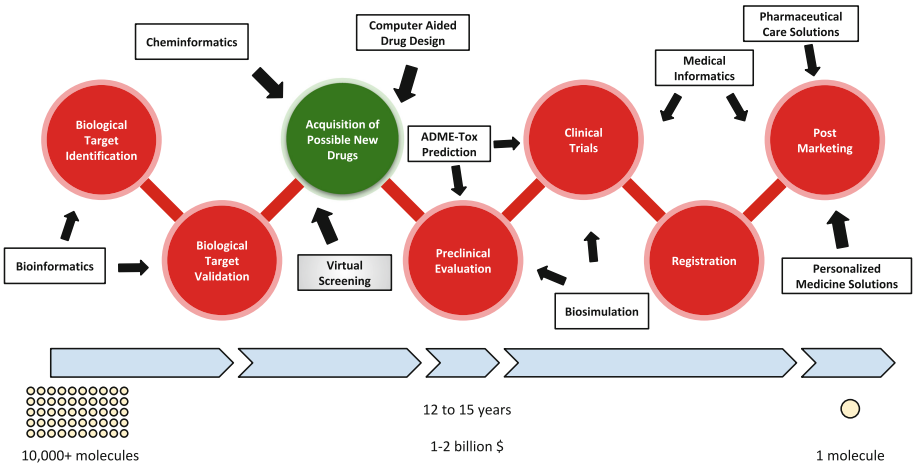

Fig. 1. Drug discovery pipeline presented together with some of the computational approaches which are used to rationalize the process.

CADD implementations deal with computational calculations of both pharmacokinetics and pharmacodynamics parameters. Therefore, ADME and even toxicity (Tox) properties of a given compound can be predicted with computational chemistry programs prior to any experimental studies. Furthermore, *in silico* approaches can also be applied to determine the putative interactions between a ligand and a receptor (Fig. 1).

In this chapter, we briefly address challenges and applications of biochemical - and computational-drug discovery and development approaches, and their transformation in VS to be applied in cloud platforms.

2 Background

2.1 Drug Discovery and Development

Until the 19th century, the drug discovery and development process was based on the trial and error learning approach for diagnosis and curing the diseases. The therapeutic effect was completely produced with natural products (NPs) [6]. These drugs were obtained from the whole or a fraction of the NPs that contain the active pharmaceutical ingredient [7]. Natural compounds are an important source for drugs, helped with improved sensitivity and better means for their biochemical separation [8].

Starting with the first successful *in vitro* organic synthesis in laboratory by Wöhler [9], it was clear that organic compounds could be produced out of the bodies of the living organisms. However, to synthesize chemically different organic compounds, structural information had to be explained in a more efficient way. In 1858, Kekulé proposed structural theories which successfully followed by different theories from the different scientists [10] leading to the discoveries of new findings [11].

Research Paradigms in Classical Terms of Drug Discovery

Mendelian Inheritance. Gregor Mendel stated that at least one dominant or two recessive gene pair are required for mutations causing phenotype-observable properties (e.g. diseases) [12] (Fig. 2).

Fig. 2. Mendelian genetics presumed phenotypic traits to depend on single genetic variation.

Magic Bullet. The term of magic bullet was created by Paul Ehrlich, who started the first systematic pharmaceutical screening and won the Nobel Prize in Physiology or Medicine in 1908. He found out that chemical compounds can directly be delivered and bound only to its biological target, called 'chemoreceptors' [13]. In light of this information, biological targets are determined to play role in the first step of the drug discovery depending on the developing technologies and knowledge [14] (Fig. 3).

Fig. 3. Summary of the biological understanding after the "magic bullet" approach.

Research Paradigms in Modern Terms of Drug Discovery

Polygenicity. It has been shown that many common diseases can occur as a result of mutations on multiple genes, such as diabetes, asthma and heart disease. Such diseases may develop as a result of an interplay of genetical inheritance, mutations on multiple genes or contribution of environmental factors [15] (Fig. 4).

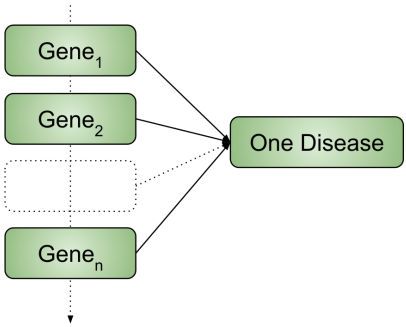

Fig. 4. Common diseases are polygenic, and this is one of the biggest challenges in the field.

Magic Shotgun. The term of magic shotgun has emerged as a result of the detection of secondary effect properties, such as adverse effects associated with the ability of the drug to affect multiple targets. This concept is related to the development of drugs with high selectivity, and to predict the potential adverse effects prior to clinical trials [16] (Fig. 5).

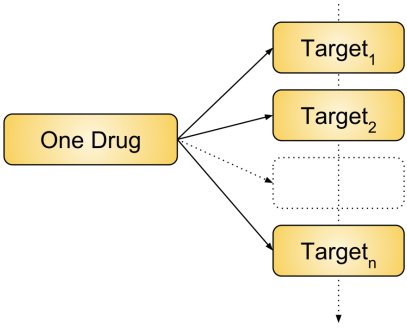

Fig. 5. Single compound may affect multiple targets. Such side-effects may be positive or undesired.

Several *in silico* approaches have been developed to predict side-effects and to repurpose drugs that are already on the market.

Individualized Medicine. The genetic polymorphisms of drug targets, metabolizing enzymes, or transporters may explain differences in the molecular pathophysiology of patients, even of those who are assigned with the same diagnosis [17]. The aim is to find drugs to compensate genetic deficiencies, i.e. to fight the disease at its roots. As of today, with the transcriptome and the genetic parameters obtained from a patient's tissue or blood, one can be informed about their contributions to disorders [18]. For instance, pharmacogenetics investigates how a genetic variation affects the binding site of a drug. That may suggest a higher dosage because of a disturbed molecular recognition.

Target Fishing. It can be regarded as the inverse screening wherein the ligand is profiled against a wide array of biological targets to elucidate its molecular mechanism of action by experimental or computational means.

Drug Repositioning. An approach to identify new medicinal applications for approved drugs to treat other diseases because the drugs may bind to other receptors.

Polypharmacology. Special ligand design approach to exert an effect on multiple disease-associated biological targets.

2.2 Molecular Recognition Theories

Key and Lock. In 1894, Emil Fischer proposed the model of key and lock for the molecules bearing potential effects on biological systems. According to this model, it is assumed that a ligand binds to the active site of its target which behaves like a key that fits its lock [19, 20] (Fig. 6).

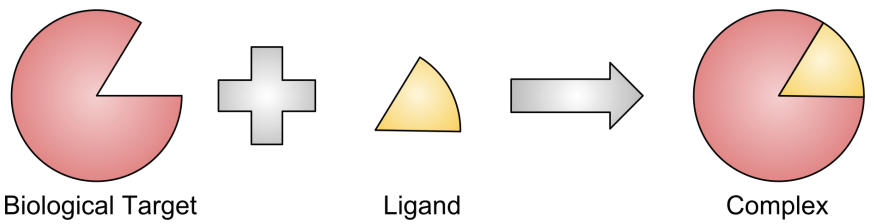

Fig. 6. Key and lock model.

Induced Fit. In 1958, it was proposed by Daniel Koshland to consider the conformational changes during the interaction between the ligand and its biological target, which wasn't considered in the lock-and-key theory of Fischer. In this model, optimum-compatible binding is occurred by small but expected conformational changes during the interaction between the ligand and the biological target [36, 37]. Later, this specific conformation of the ligand has been referenced as the bioactive conformation (Fig. 7).

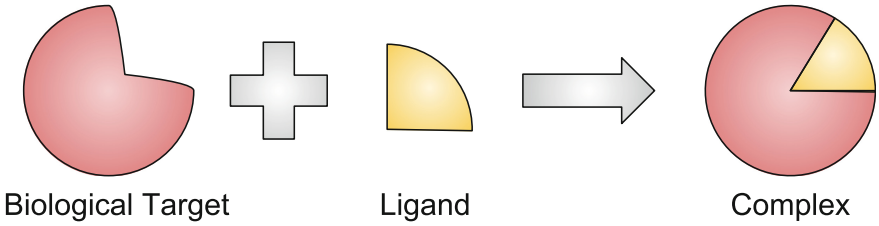

Fig. 7. Induced fit model.

Thermodynamics in Ligand-Protein Binding. During the binding process of ligands to their biological targets, it is known that conformational arrangements and desolvation occur, with the help of specific physicochemical interactions [38] (Fig. 8).

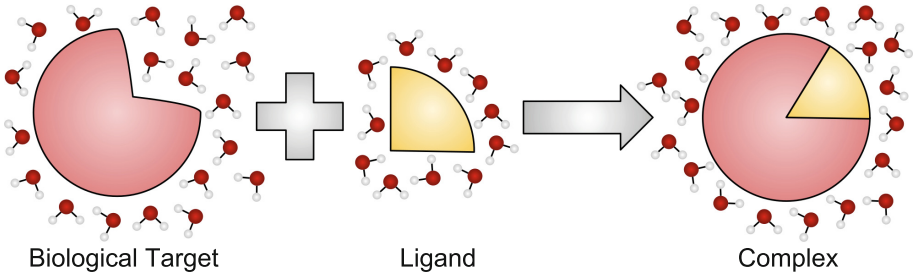

Fig. 8. Desolvation effect and thermodynamics in ligand-protein binding.

The most important strategy in the process of developing novel drug candidates is to increase the affinity of binding of the ligand to its target. The energy change after the binding process is quantitatively expressed by logarithm of K_d or K_i values. These values are related to Gibbs free energy of binding [21].

$$\Delta G_{binding} = RTlnK_d \sim RTlnK_i \tag{1}$$

In this equation, R is the gas constant, T is the absolute temperature, K_d is the equilibrium constant and K_i is the inhibitory constant that exist in the Cheng-Prusoff equation [40].

In non-covalent thermodynamic binding process, ligand-protein binding energy change is occurred and two thermodynamic quantities, the enthalpy change (ΔH) and entropy change (-TΔS), determine the sign and magnitude of the binding free energy [33, 39].

$$\Delta G_{binding} = \Delta H_{binding} - T\Delta S_{binding} \tag{2}$$

2.3 Chemical Space

In 1996, Regine Bohacek, generated predictions about possible chemical compound types that might be chemically accessed. Her estimation was pointing to 10^{60} chemical compounds making up "chemical space" that was virtually identified by using carbon, oxygen or nitrogen atoms, and by considering linear molecules with up to 30 atoms. While making the predictions, Bohacek regarded chemical groups to be stable and chemical branches, ring structures and stereochemical possibilities were taken into account [22].

In later studies, "the limits of the chemical space" was drawn to be between 10^{18}–10^{200}, according to the results of the analyses by different methods and descriptors. Although there are many reports about this range to be accepted as Bohacek defined. But it is expected that this number will continuously increase by the discovery of new chemical skeletons [23–26]. Additionally, the number of organic compounds accessed experimentally is 10^8, according to CAS and Beilstein databases which contain records obtained from the scientific papers, those have been published by the scientific community, since 1771 [27, 28].

2.4 Rational Drug Design

The increasing scientific knowledge for novel drug discovery has opened new horizons and generated useful technological developments for the researchers in the field. When these new tools are wisely brought together with recent knowledge, they would provide many advantages in drug design and development studies. Moreover, the available theoretical and experimental knowledge about drug safety and the idea of being appropriate for human use generate extra difficulties for drug design and development [29]. It is known that not all candidate molecules with high potency can reach to a drug status due to several reasons such as inefficient systemic exposure, unwanted side effects and off-target effects. Also, a drug may not be right for every patient due to the genetic variations and off-target binding. This also effects drugs that are already on the market (Table 1).

However, molecular reasoning may give second chances for drugs that once failed in late clinical studies (at great expense) or that have been retracted from clinical use. With an improved molecular understanding, and with hindsight from the now feasible pharmacogenetics and –genomics, these compounds have a chance to find their niche for a reentry.

Conventional drug design and development approaches are associated with high affinity binding of a ligand to its target. In rational drug discovery studies, evaluating the ligands by only trial-and-error approach has lost its validity and assessing their binding to the target related diseases is insufficient [30]. By taking control of some parameters like off-target binding, ADME-Tox and bioavailability properties of the molecules that have appropriate binding properties should also be discovered for the drug candidates [49]. Therefore, revising the classical strategies with "rational" approaches has become substantial. This process is called *reverse pharmacology* or *target-based drug discovery*. In this recent approach, the questions about "Which chemotypes will be worked on?" and "Why?" should also be considered during drug

design and development studies, before directly evaluating the target binding properties of the molecules [29, 30].

Table 1. Definition of the key terms related with informatic approaches used in early stage drug discovery.

Bioinformatics. The discipline responsible for biological data recording and interpretation using information technologies. Some of its applications on drug discovery include the detection of protein binding pocket, prediction of protein-protein interactions, occurrence of mutations, analysis of the biological sequences of macromolecules (e.g. similarity searches and fingerprint matches), estimation of 3D structures of biological macromolecules [31].
Cheminformatics. The discipline which accumulates and processes the chemistry related data, using information technologies. Some of its applications on drug discovery include construction of the 2D and 3D structures, storing and searching the chemical data, building chemical compound databases and datasets, QSAR / QSPR, estimation of ADME-Tox properties [31]. It can also be used to map the chemical space of compound libraries.
Pharmacoinformatics. The discipline which combines the cheminformatics and bioinformatics tools for pharma-related processes [32, 33].

3 Computer Aided Drug Design (CADD)

Development of mathematical formulas to calculate the potential and kinetic energies of biomolecular systems has made possible the implementation of such complex calculations with computers [34]. CADD is applicable for hit molecule discovery for new different chemotypes and for designing new derivatives.

CADD processes may be divided into molecular mechanical methods and quantum mechanical methods. In both techniques, the results are obtained through energy-based calculations. Molecular mechanics deals with the calculations at the molecular level that can be performed on an atomic basis, while quantum mechanics involves electron related complex calculations performed at the quantum level [34].

During existence of the obscurity in drug discovery studies, it is hard to reach the desired target. But, the physicochemical parameter as a factor can be useful about this topic by measuring the ADME properties [30, 35]. Also, the drug-candidate should be bound to its target with high affinity [30]. In relation to that, drug design processes are carried out within the framework of selected strategies, with the acquisition of three-dimensional bioactive conformation of the ligands. CADD is used for identifying and designing biologically active compounds and this field can be synergistically integrated with all other medicinal chemistry related fields like pharmacology, biology and chemistry [36].

3.1 Ligand-Based Drug Discovery

It is possible to store the chemical and biological information in special databases to preserve the molecular features and their biological effects obtained from a series of

assays. If there is no molecular level structural data about the biological target, previously obtained structure-activity data can be used in CADD studies to find and analyze the trend between the known compounds and their activity retrospectively, then to design and discover new chemical entities prospectively [37]. A main workflow of ligand-based studies addresses the generation of a model by training on a series of compounds and subsequent testing stage of the model with test series of compounds. Later, the generated model(s) can be validated with an external series of compounds which were not used during model generation. Then the model can be used to virtually screen chemical databases within the applicability domain of the model (Fig. 9).

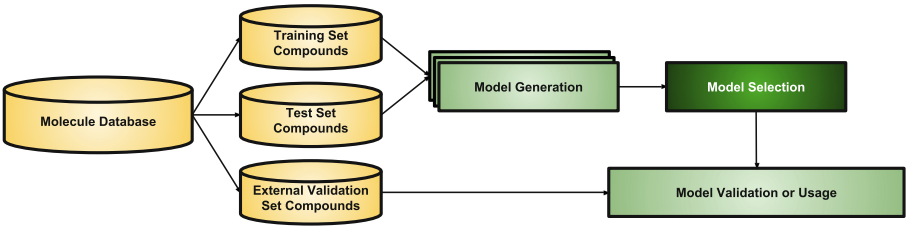

Fig. 9. Basic workflow of ligand-based modeling applications.

Quantitative Structure Activity/Property Relationships (QSAR/QSPR). In 1863, an approach about the relationship between chemical structure and biological activity was proposed by Cros [38]. After that, in 1868 Crum-Brown and Fraser evaluated this correlation in molecular level and described following equation to computationally predict the biological response (Φ) to be identified as the function of the chemical compound (C) [38, 39].

$$\Phi = f(C) \tag{3}$$

In the course of the history, many researchers have conducted studies to relate physicochemical properties and biological effects of compounds with different approaches. Through these studies, the necessity to consider molecular substitutions has emerged to try to explain biological effect or physicochemical property of a chemical structure. In 1937, Louis Hammett compared ionization rates of benzene derivatives substituted in various positions [40]. The first quantitative parameter was determined as sigma (σ), the potential electronic contribution value, which is defined by calculating electronic substituent constant values. This was the first identified quantitative parameter [38]. Then the first steric parameter for ester hydrolysis was determined by Taft as E_s constant [41]. Many other two-dimensional parameters have been continued to be developed for being used in QSAR studies [38]. Later, a multiparametric formula was presented by Corwin Hansch that brought these parameters together. The formula was designed to calculate minimal concentration needed to mathematically formulate the biological activity, as logarithm of concentration and was measured with several independent factors in different cases; such as partition coefficient (log P), aromatic substituent constant (π), electronic substituent constant (σ),

steric parameter (E_s). Free-Wilson [42], Fujita-Ban [43], A Mixed Approach - Based on Hansch and Free-Wilson Analysis [44], and Kubinyi Bilinear Analysis Method [45] are some of the first generated models used in QSAR analysis.

When all the steps are evaluated, it can be observed that QSAR has different applications. Structural activity or physicochemical property studies performed with one independent variable are named 0D-QSAR calculations and the studies with multivariate equations are called 1D-QSAR. In such equations, log P and Hammett constant can be used as independent variables. Studies that take into account the molecular descriptors and fingerprints containing information about structural and bonding characteristics resulting from the two-dimensional molecular presentation are called 2D-QSAR, if extra structural information (e.g. chirality) is included in studies these studies are named as 2.5D QSAR [46].

The efforts to explain the binding orientations of the ligands have emerged 3D-QSAR, together with the help of the developments in 3D molecular understanding through the experimental methods, force-field methods and improved molecular visualization technologies. Implementation of this approach is accomplished by forming equally divided grids in the 3D coordinate system and the surface interaction areas of the chemical compounds are determined by the following analysis methods [47] (Table 2).

Table 2. Summary of some of the 3D-QSAR approaches.

Comparative Molecular Field Analysis (CoMFA). CoMFA is a subtype of the 3D-QSAR study, which includes steric and electronic fields (Lennard-Jones and Coulomb potentials) generated on the bioactive conformation of the compounds in the training set. In CoMFA studies, it is only possible to analyze the effects of enthalpic properties on ligand-protein binding [47]**.**
Comparative Molecular Similarity Index Analysis (CoMSIA). To explain the structure activity relationships by considering major contributions obtained by steric, electrostatic, hydrophobic, and hydrogen bond donors (HBD) or hydrogen bond acceptors (HBA) properties are added into CoMFA and this generated CoMSIA analysis [48].
Receptor-based QSAR. In this study model, which is carried out using a combination of ligand-based and structure-based approaches, bioactive conformation of the ligands is obtained by structure-based methods. Afterwards, 3D-QSAR studies are conducted through this bioactive conformation [47].

Pharmacophore Modeling. A pharmacophore aggregates functional groups of chemical compounds which are responsible for the biological response and which exhibit appropriate interaction with biological target. The term pharmacophore modeling refers to the identification and 3D display of important pharmacophore groups to illustrate the basic interactions between the ligand and the receptor. Pharmacophore modeling is generally applied to determine common structural features within a series of similar or diverse molecules by subtracting 3D maps. Once determined, the generated pharmacophore hypotheses may be used to virtually screen and to predict the biological activity

of other molecules [37]. The model is generally obtained from the information belonging to the ligands. However, it is also possible to generate a pharmacophore model from the receptor itself or with a combined approach as well.

After the formation of possible bioactive conformers of the compounds, pharmacophore model can be generated in 3D by aligning the structures and mapping the 3D binding properties. More than one pharmacophore hypothesis can be generated and the most suitable one(s) can be identified by enrichment factor within their applicability domain. While generating the model(s), the optimum volume of each pharmacophore property takes the major interest.

The aim of pharmacophore modeling is to determine the optimum volume for the identified properties. If identified volumes are larger than required, the selectivity of active compounds by this model decreases and active and inactive compounds can be found together by using this model. Conversely, if the fields are smaller than they need to be, the active compounds cannot be identified by pharmacophore screening. While creating a pharmacophore model; HBA and HBD features, hydrophobic (aliphatic or aromatic) properties and negative/positive charges can be used. Besides, it is possible to identify the desirable/undesirable regions or features without any specificity [37]. Most pharmacophore modeling programs create these properties according to optimized ligand-receptor interactions. The ideas behind pharmacophores are also applied to specify new compounds that aggregate as many pharmacophores, e.g. from a series of *in silico* ligand docking experiments or after a first *in vitro* validation. The assembly of new compounds can be iterative, i.e. growing from existing binders, or starting a *de novo* design of novel drugs.

Machine Learning (ML). Relations between ligand/receptor structural properties and biochemical effects are found by statistical models. With the advent of computational chemistry and advances in structural biology, an increasing number of features may be identified *in silico* for any given ligand and receptor, which may affect virtual screening. The integration of all these data sources and tools allows for the formulation of many models.

ML covers the computer-aided development of models (or rules) and their evaluation. The use of ML techniques for drug discovery has been increasing recently both in ligand-based and structure-based studies to find rules from a set of molecular features and to predict a specific property [49]. Estimation of ADME-Tox properties by the use of physicochemical descriptors, generation of hit or lead molecules with the studies on prediction of biological activity, development of homology models, determination of bioactive conformation by the help of docking studies or pharmacophore modeling are some of the examples of ML applications in drug discovery [49].

ML can be applied as regression or classification models [49]. In the regression model, quantitative models are formed automatically. Statistically the most appropriate model is selected from the generated ones. Classifiers utilize such models to cluster the data. The learning process is carried out by known characteristics of the molecules to predict their activities/properties. In the classification model, the branches are formed on a classification tree. Biological and chemical data are placed on the leaves of the

branches of the tree. It can be generated and used for various statistical decision-making scenarios. Artificial neural networks, support vector machines, decision trees and random forest techniques are some of the most applied ML techniques used in drug discovery studies [49].

3.2 Structure-Based Drug Design

Experimental studies on elucidation of the 3D structures of biological macromolecules are generally performed by x-ray crystallography and NMR methods [50] (Fig. 10). Obtained structural information about the macromolecules is stored in regional protein databases. There are three major regional databases (Protein Data Bank, PDB) for storing 3D the crystal structures of these macromolecules such as RCSB PDB [51–54] (USA), PDBe [54, 55] (Europe), PDBj [54, 56] (Japan). Likewise, the 3D structures of relatively small biological macromolecules obtained by NMR techniques are stored in the Biological Magnetic Resonance Data Bank [54, 57] (BMRB). Records stored in these databases are synchronized via the Worldwide PDB [54] (wwPDB) organization.

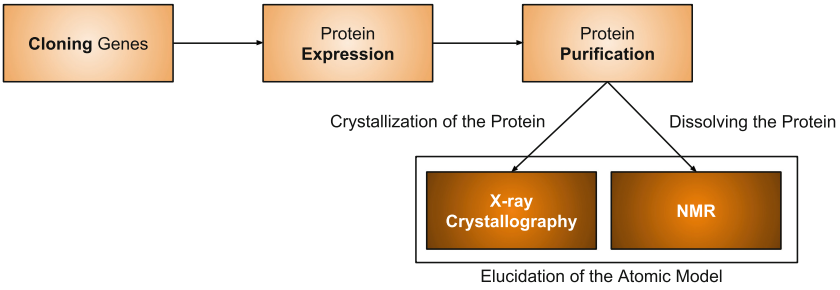

Fig. 10. Obtaining the 3D protein structure.

Homology Modeling. Homology modeling techniques are used to predict 3D representations of biomolecular structures. The model generation is done with the sequence of monomers (nucleotides or amino acids). The applied algorithm transfers spatial arrangements from high-resolution crystal structures of other phylogenetically sequence-similar biological structures [58].

Protein sequence information is stored in reference databases such as UniProt [59], where the information on the protein sequence and its functions are collected in one place. The UniProt database offers other web services such as BLAST [60, 61] and Clustal Omega [62] where sequence similarity and alignment can be retrieved. Consensus annotation across protein families across species is aggregated in PFAM [63] with functional annotation (Fig. 11).

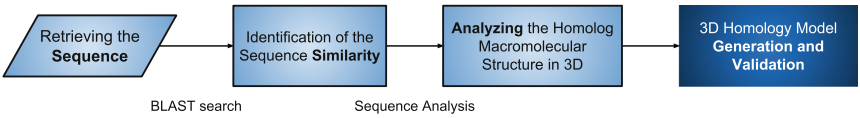

Fig. 11. Basic workflow for homology model generation.

Docking. A structure-based simulation method which is used to predict the 3D binding modes and molecular interactions between ligands and their biological targets (Table 3). This technique is based on the thermodynamic calculations by following the determination of the possible binding orientation of the ligands and the proteins with appropriate interactions in the active site. The docking scores are determined based on these interactions and conformational rearrangement costs which help evaluating the generated results [58, 64, 65] (Fig. 12).

Table 3. Docking concepts, and their limitations.

Technology	Limitations
Rigid Docking. Receptor structure is treated as a rigid body and the ligands are prepared with conformational sample(s), then fit into the active site of the protein	The overall protein structure and its active site residues are flexible, this affects the binding orientations of the ligands in docking results
Induced-fit Docking. Flexibility of the active site residues is considered for the protein and these flexible active site residues flexibly adapt to accommodate the ligand	It is a computationally expensive approach that requires careful arrangement of the docking parameters and cuts off the values prior to docking simulation
Covalent Docking. In this method, binding region of the receptor that ligands bind covalently is identified and the docking procedure is performed in this specific condition	This method may generate chemically wrong binding pattern and incorrect binding pose but lowers the computational search costs of docking
Peptide Docking. Peptide docking method is used to determine binding modes of peptide structures in the active site of its biological target	Peptides, or fractions of proteins, are large and flexible ligands and hence difficult to parameterize and computationally expensive compared to the small molecules
Protein-Protein Docking. It is the general name of the docking method that is used to predict protein-protein or protein-DNA interactions those are taking place biologically	This is the most computationally demanding approach to simulate the interactions, due to the size and complexity of the macromolecules
Reverse Docking. It can be applied for target fishing the drugs or active substances, or to become aware of side-effects	Conceptionally, this approach requires a catalog of configurations for screening many targets by docking [66]

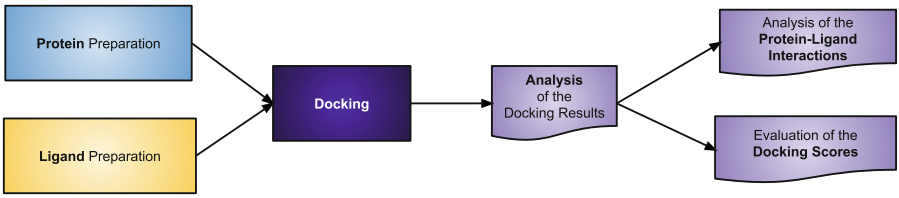

Fig. 12. Basic workflow for docking simulations.

Molecular Dynamics (MD). MD is a technique to simulate the dynamics of connected atoms in a joint molecular system of ligands and their biological targets by considering environmental factors (temperature, solvents, ions, membrane, pressure, temperature, etc.). The simulation may be run within a specified time interval, all depending on the ambient conditions set by the force-field [64, 67]. The most preferred binding orientation of the conformational samples that emerge during the simulation can be obtained by analysing the energy of the whole or a portion of the simulation and by visualizing and analyzing the inter/intra-molecular interactions (Fig. 13).

Fig. 13. General workflow for MD simulations.

4 Application of Virtual Screening

VS is a chemical search to match ligands with their biological targets. The process can be applied by either ligand-based or structure-based approaches. Related to the principles of the method used to find a hit molecule, VS covers various concepts. Regardless of the CADD approach, the process starts with a chemical database to be prepared for each molecular entry. This can be followed with *in silico* predictions of the pharmacokinetic and pharmacodynamic profiles of the molecules and a filtering of molecules within the database to obtain statistically significant and manageable subclusters of the chemical space [6] (Fig. 14).

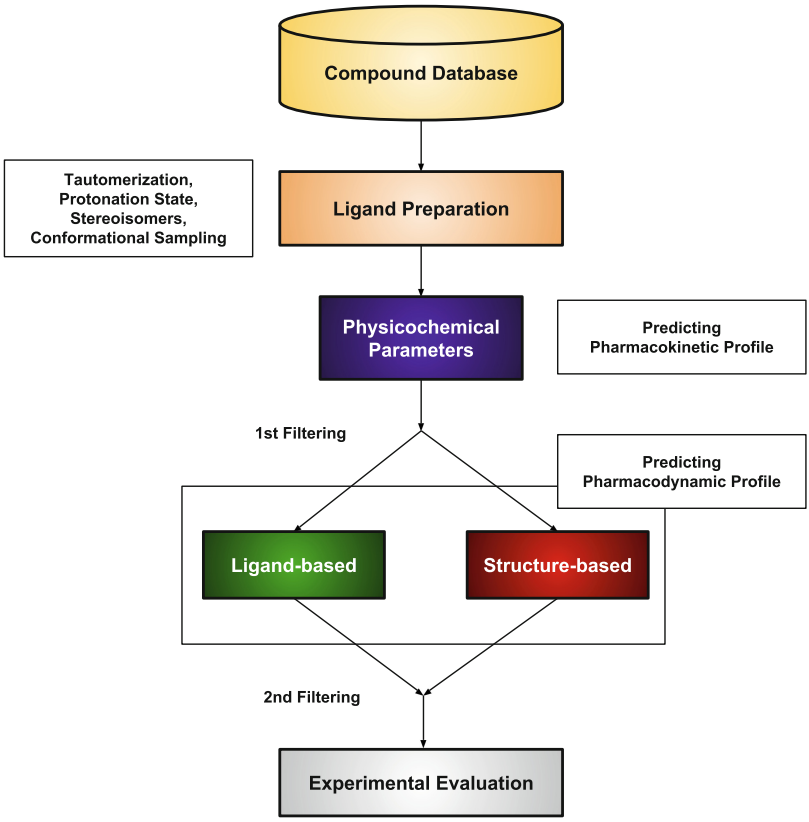

Fig. 14. General VS workflow.

4.1 Accessing Compound Databases

Chemical compound databases represent significant chemical catalogues for the discovery of hit molecules with biological activity, which compile large numbers of chemically different entities from either synthetic or natural origin [68]. They can be generated with in-house data, a combination of compound databases (previously tested or untested screening compound series) or *in silico* synthesis with building blocks. Compound databases of curated biological activity data, commercial sources of the molecules and pharmaceutical reference databases of the marketed drugs can be seen in Table 4.

Table 4. Compound databases.

Database/URL	Comment
BindingDB [69] www.bindingdb.org	Around 1.5 million binding affinity data points are available which are obtained from patents and scientific articles
chEMBL [70] www.ebi.ac.uk/chembl	More than 15 million biological activity results are available curated from scientific papers and patents
ChemSpider [71] www.chemspider.com	Contains over 67 million structural data, obtained from different data sources
DrugBank [72–76] www.drugbank.ca	Database of drug entries linked to their biological targets
DrugPort www.ebi.ac.uk/thornton-srv/databases/drugport	Structural representation of drugs with their receptors with data from DrugBank and PDB
eMolecules www.emolecules.com	Database comprises over 7 million screening compounds and 1.5 million building blocks, which are ready to order
Molport www.molport.com	Comprising around 20 million ready to be synthetized molecules, over 7 million screening compounds and around 0.5 million building blocks, which are ready to order
MMsINC [77] mms.dsfarm.unipd.it/MMsINC/search	Helps to search subsets of the database, containing over 4 million unique compounds with tautomeric and ionic states at physiological conditions as well as possible stable conformer for each molecular entry
Open PHACTS [78] www.openphacts.org	Large knowledge-base of data integration for compounds, their receptors, and pathways
PubChem [79] pubchem.ncbi.nlm.nih.gov	Contains around 100 million compounds and 250 million bioactivity data records
ZINC [80] zinc.docking.org	Contains around 740 million purchasable molecules in 2D and 220 million purchasable lead-like molecules in 3D

4.2 Preparing Ligands for Virtual Screening

It is important to assign the correct stereoisomeric or tautomeric forms and protonation states of each ligand at a specified pH to avoid changing physicochemical and conformational behavior of the molecules. Parameterization during this stage must be done very carefully for a VS study, because the molecules are generally stored in 1D or 2D within the databases. For example, a chiral entry within the database may be in a single enantiomer or a racemic mixture or a molecule may not be neutral as it is stored in the database and may be at different protonation states at different pH values.

Conformational flexibility of ligands is also important and computationally more expensive. Every additional rotation increases the number of conformations that needs to be generated and this results with a computationally more expensive VS process.

4.3 Compound Filtering

The main idea of compound filtering is to exclude the molecules which are not carrying suitable pharmacokinetic or pharmacodynamic properties. The aim is to prevent possible risks in preclinical or clinical phases of the studies. This also helps to keep down computational costs in VS studies.

Clustering by Chemical Diversity. It is possible to estimate the chemical similarities of the compounds by computer-based calculations. This lets identification of similar or diverse subsets of the libraries. Many different techniques are used for this purpose. The basic approach is to use similarity algorithms based on mathematical parameters obtained from chemical descriptors [81, 82]. Such calculations are generally done by calculating molecular fingerprints, which are a way of encoding the structure of a molecule, usually as binary digits used to symbolize the presence or absence of a chemical substructure within the molecule allowing for a similarity search in a chemical database. The most common similarity and classification algorithm is Tanimoto's similarity estimation and classification practice that regulate mathematical classification studies [82].

Filtering with Physicochemical Parameters. One of the major examples of filtering by pharmacokinetics related approaches is based on the scientific analysis of the drugs. Most drugs in the market can be used orally and it is found that there is a generally applicable rule to determine whether a biologically active compound has physicochemical properties that would make it an orally administrable drug in humans. In 1997, it is formulated by Christopher Lipinski [83, 84] and called Lipinski's rule of five (RO5). It is designed to evaluate drug-likeness of a compound, to be used in drug discovery and development studies. The properties of RO5 depend on; a molecular weight less than 500 Dalton, an octanol-water partition coefficient (log P) less than 5, equal or less than 5 HBD groups, equal or less than 10 HBA. Drug candidates which have properties of the RO5 tend to have more success in clinical trials, so have more chance to reach to the market. Such approaches let the identification of drug-like chemical space to fall within the biologically relevant subspace with a better pharmacokinetic profile of a vast unbounded chemical space.

Filtering with Ligand-Based or Receptor-Based Screening of Chemical Libraries. Ligand-based approaches focus on model generation with pre-established binding affinity data of small molecules against biological targets. These approaches are used with calculated descriptors to predict the characteristics within the molecular database. Though, structure-based approaches don't necessarily rely on existing data and try to place the molecules in the binding sites of the target and evaluate their potential affinity by binding scores within the complex biomolecular structures. Such computationally expensive calculations are preferred to run on distributed compute platforms to speed up the *in silico* screening process.

Ties with Pre-clinical Research. The search for new biologically relevant compounds is not possible without a laboratory result in which theories are confirmed, refined or disproved. Considerable time and effort are needed to be invested into the development of such tests. Eventually, it is needed to transfer the findings to human tissues.

VS may support the selection of the ligands for testing once a drug target is identified. This process is long and demands deep insights into the molecular understanding of the pathology of a disease. The selection of a target is likely to be supported from genetic studies, i.e. an observed association of DNA variation with disease phenotypes, hints the gene to target. Knock-down/out experiments are performed in the laboratories to confirm the molecular mechanism that trigger the disease.

Many researchers on rare diseases do not have industrial partners, and no drugs may be marketed for such "orphan" diseases. There are many rare diseases and consequentially many patients. To find new drugs, the availability of *in silico* approaches may be a significant guidance for the researchers, e.g. for repositioning drugs already on the market. The same approach can be applied to identify the biological targets of traditional medicines obtained from NPs [6].

Ties with Individualized Medicine. Individualized medicine identifies genetic variations that are causative for a disorder for every patient and adjusts therapy accordingly. Once a compound binds to its receptor, this yields a detailed molecular understanding how the drug works.

Such insights are of interest whenever a genetic variation may affect the binding of a drug to its receptor. Some drugs address cells with a high mutation rate, like a virus or a tumor cell. Mutations appear at random, but when the pathogen/cell benefits from it because of a lower binding affinity to the drug, it will continue to divide and pass that genetic information on. Thus, the pathogen or tumor evades the therapy. For that reason, The FightAIDS@Home project has searched for compounds with strong binding affinities to many sequence variations of the viral protein target [85].

Today, cancer patients are monitored by DNA sequencing to learn the changes in relative frequencies of genetic variations of the disease-tissue [86]. The adoption of next-generation sequencing technologies is a recent development. It will indicate the frequency of "escape routes" that may be closed with new drugs for the same target, possibly in the same binding pocket. It is with the same modeling technology that is at the root of *in silico* screening that this data can be interpreted. Without a molecular interpretation, clinicians would need to wait for insights from observations of how nuclear variation and subsequent therapy decisions affect the clinical outcome. A molecular reasoning, in contrast, is possible for every patient individually.

Automation. Several tools are available to prepare receptors and ligands. Glue language can be used to integrate different processes of such calculations [87]. In addition, there are workflow environments for CADD in particular like OPAL [88] or KNIME [89]. The users can easily develop any integration of CADD workflow that needs to be automated. Recent works integrate formal descriptions in databases (like bio.tools) [90] for content-dependent semantics-driven execution of the tools [91].

4.4 Experimental Evaluation and Validation

For biological research, *in silico* produced findings are needed to be evaluated with *in vitro* or *in vivo* biological assays. It is preferable to know the statistical strength of the model to evaluate the functionality of the model with retrospective studies, prior to biological testing. That test is likely to be performed in several stages and can be

exemplified as given below. It should be noted that this stage of the work strictly depends on the nature of the target:

(a) Test of binding affinities (calorimetry, SPR)
(b) Induced structural changes (co-crystallization, NMR)
(c) Gene/Protein-function inhibition/induction (cell line, tissue sample)
(d) Effect on disease models (animal trials)

These tests are expensive, may cost the lives of animals. If the wrong compounds are chosen for testing, then this delays the treatment of humans. Testing diverse sets of compounds is preferred with respect to their binding mode and chemical structure.

5 Infrastructures and Computational Needs

Virtual screening is a data parallel process. The throughput of ligands tested in the computer almost linearly scales with the number of processors contributing. A second concern is the number of applications contributing – to the screening itself. The availability of high-performance computing (HPC) infrastructure is needed to speed up the *in silico* screening studies. In addition, accessing stronger HPC facilities has direct impact on covering bigger chemical and conformational space of compounds in VS studies or testing higher numbers of druggable targets to identify biomolecular targets of the compounds in target fishing studies.

When it is applied on thousands or millions of a series of chemical compounds, the total VS simulation time exceeds the limits of single workstations. For example, a molecular docking-based VS run, covering chemical space of millions of compounds, may take years of computation time on a single workstation. The overall compute time is fixed, also with multiple machines contributing. But by distributing the load, the effective wall-clock time can be lowered. The same VS run can be taken to months-scale by accessing hundreds of central processing unit (CPU) cores with an in-house small or medium scale HPC infrastructure or hours/days-scale by accessing thousands of CPU cores with supercomputers or on-demand cloud resources. Recent technical advancements to employ graphical processing unit (GPU) for parallel computing have fostered the concept to apply on data, i.e. deep learning or molecular dynamics.

We here discuss alternative sources for computational infrastructures and single out unique properties of the cloud environments.

5.1 Local Compute Cluster

Many computers are integrated in a local network and combined with a grid software to distribute compute jobs. The concept of a batch system is familiar to IT specialists who maintain such hardware. Larger clusters with special hardware to allow for fast I/O are referred to as HPC environments. The VS is mostly compute-intensive, data parallel computation with individual compute nodes which do not need to communicate with each other during the computations.

5.2 Volunteer Computing

Volunteer computing refers to the concept of having a program running in the background addressing a scientific problem. If several thousand individuals contribute, then one has a compute power or storage resource that can compete with the large computational clusters. It should be noted that one needs to interact with the community to keep the momentum.

Several *in silico* screening runs have been performed voluntarily. One of the most well-known was the search for an HIV protease inhibitor in FightAIDS@Home [85] with the underlying BOINC technology [92]. Because of the enormous compute power acquired by devices that are not operated by humans, the decentralization of computing is a strategically important route. Tapping into these resources for computing and to generally support the interaction of devices at the periphery is called as *edge computing*.

5.3 Cloud Computing

Cloud technologies are a means to integrate different workflows and tools on dynamically allocated resources. These instances share respective services in their specification of CADD studies, and cloud computing is an emerging solution for VS needs. The infancy of cloud solutions was in "IaaS", i.e. dynamically configured and started machines that were paid on use. Today, IaaS services are provided by many compute centers at very competitive price levels. Setups are easy to use for individuals familiar with remote access with command line interface (CLI) or graphical user interfaces (GUI).

Researchers in both academia and industry are likely to have access to a local compute cluster. But this is costly when the expenses of hardware upgrades, electricity and maintenance are added. Remote sources are highly competitive in pricing. This is meant for IT-savvy individuals who could use local resources if they had them. There is technology allowing these remote resources to behave much like an extension of a local cluster.

The most known cloud service providers are Amazon Web Services [93] and Google Cloud [94] but there are increasing numbers of open source cloud solutions that may be of interest. The OpenStack middleware [95] has evolved into a de facto standard to set up public or private cloud environment. When a software is accessed via a web interface, the server-side implementation becomes hidden. A look to the side, e.g. how payments are performed online with an interplay of many service providers, tell what clouds also mean: An interplay of services that scale.

Anticipating variants of VS principles to be employed in different and barely predictable ways are outlined in the next section. One may expect a cloud-based glue-layer of molecular modeling and genomics to emerge. But there are no CADD-Microservices/Functions as a Service, yet (Table 5).

Table 5. Classification of cloud services.

IaaS: Infrastructure as a Service is a service model which to provide remote access on demand to client-specified compute environments. Besides the typical virtualized setups with multiple users sharing a single processor, for HPC also "bare metal" setups are available. CADD tools may be readily usable deployed as a cloud image or be installed posteriori to a running image.
PaaS: Platform as a Service is a technical concept for an aggregation of services to facilitate software running in support of IaaS to scale in multi-user, multi-project environments.
SaaS: Software as a Service runs remotely, and stars copies of itself on multiple nodes in an IaaS environment to scale to whatever size of a problem or to as many users as there may be. The software is typically presented as a web interface, but the concept has found broad adoption from business applications to the gaming industry.
FaaS: Function as a Service. Element of a SaaS that may be invoked without context. Offered by all major cloud vendors, an open source implementation is Open Whisk [96]. This technology offers a route for VS-associated services to integrate into larger workflows.

5.4 Cloud-Based Solutions

Recent reviews [97, 98] and the "Click2Drug" catalog [99] give a vivid expression on describing many different sources for many partial solutions – available as semi-automated web services or as instances that may be started at one's independent disposal. How exactly a transfer of information between these tools should be performed, or how these tools should be run to generate redundant findings with then higher confidence – all this is yet to be clarified – both semantically and technically.

In silico approaches have been integrated in drug discovery pipeline and big pharmaceutical companies are likely to perform these often. Cloud environments are used for their advantages to reduce setup costs for professionals. Also, for creative new approaches in tackling a disease, the information on which selected target is already crucial not to lose competitive advantage. This may not be acceptable to be computed in extra mural facilities. However, for common targets, the interception of individual VS results, e.g. from a volunteer computing project, may not be considered critical.

Table 6 shows remote services that may use cloud services internally or that are offered for integration with a self-implemented workflow. Most services are agnostic with respect to the source of a request. However, some may fall back to the cloud provider's infrastructure for monetary compensation. Thus, the cloud becomes an integral part of the VS tool's deployment.

Table 6. Cloud-based solutions for virtual screening and target fishing. The table lists services that describe features of ligands or their receptors, receptor-based and ligand-based services. Services integrating multiple tools are tagged as a workflow. F stands for "Feature", R for "Receptor-based", L for "Ligand-based" and W for "Workflows".

Service	Properties				URL	Comment
	F	R	L	W		
3decision	X	X		X	3decision.discngine.com	Collaboration environment for researchers to exchange opinions on ligand-receptor interactions
AceCloud [100]		X			www.acellera.com/products/acecloud-molecular-dynamics-cloud-computing	Cloud image to run MD simulations with CLI on Amazon Cloud, can be used for high-throughput MD [101]
Achilles [102]		X		X	bio-hpc.ucam.edu/achilles	Blind docking server with web interface
BindScope [103]	X	X	X	X	playmolecule.com/BindScope	Structure-based binding prediction tool
DINC WebServer [104]		X			dinc.kavrakilab.org	A meta-docking web service for large ligands
DOCK Blaster [105]		X			blaster.docking.org	Automated molecular docking-based VS web service
DockingServer [106]		X			www.dockingserver.com	Molecular docking and VS service
Evias Cloud Virtual Screening Tool [107]		X		X	www.evias.com.tr/vst	Integrated with chemical library management system as a scalable HPC platform for structure-based VS service as a web service
HDOCK [108]		X			hdock.phys.hust.edu.cn	Protein-protein and protein-nucleic acid docking server
HADDOCK Web Server [109]		X			milou.science.uu.nl/services/HADDOCK2.2/haddock.php	Web-based biomolecular structure docking service
idock [110]		X			istar.cse.cuhk.edu.hk/idock	Flexible docking-based VS tool with web interface
iScreen [111]		X		X	iscreen.cmu.edu.tw/intro.php	A web server which started with docking of traditional Chinese medicine
LigandScout Remote [112]	X	X	X	X	www.inteligand.com	A desktop application, letting access to the cloud-based HPC for VS studies
mCule [113]	X	X	X	X	mcule.com	Integrates structure-based VS tools with purchasable chemical space

(*continued*)

Table 6. (*continued*)

Service	Properties				URL	Comment
	F	R	L	W		
MEDock [114]	X	X			medock.ee.ncku.edu.tw	A web server for predicting binding site and generating docking calculations
MTiOpenScreen [115]		X		X	bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen	Integration of AutoDock and MTiOpenScreen in bioinformatics Mobyle environment
ParDOCK [116]		X			www.scfbio-iitd.res.in/dock/pardock.jsp	Fully automated rigid docking server, based on Monte Carlo search technique
PatchDock [117, 118]		X			bioinfo3d.cs.tau.ac.il/PatchDock	A web-server for generating docking simulations with geometry and shape complementarity principles
Polypharmacology Browser 2 (PPB2) [119]			X		ppb2.gdb.tools	A web server letting target prediction for the ligands
ProBiS [120]	X	X	X		probis.cmm.ki.si	A web-based analysis tool for binding site identification
ProBiS-CHARMMing [121]					probis.nih.gov	In addition to ProBis, it is possible to do energy minimization on ligand-protein complexes
Py-CoMFA			X	X	www.3d-qsar.com	Web-based platform that allows generation and validation of 3D-QSAR models
SwissDock [122]		X		X	www.swissdock.ch	Web-based docking service to predict ligand-protein binding
USR-VS [123]		X			usr.marseille.inserm.fr	Ligand-based VS web server using shape recognition techniques
ZincPharmer [124]	X		X		zincpharmer.csb.pitt.edu	Online pharmacophore-based VS tool for screening the purchasable subset of the ZINC or Molport databases

6 Conclusions

The goal of VS is to identify novel hit molecules within the vast chemical space to ameliorate symptoms of a disease. Hereto, the compound interacts with a target protein and changes its action in a pathological condition. None of the *in silico* molecular modeling techniques can generate perfect models for all kinds of biochemical processes. However, a large variety of tools is provided by the scientific community.

Alone or in combination, there are available VS technologies which suite the problem at hand. With clouds as HPC resources, complete workflows have been established to directly address the needs of medicinal chemists. The same technology also supports collaborative efforts with computational chemists to adjust workflows to emerging preclinical demands.

Another problem to address is the demand for HPC facilities for VS projects. These are indirect costs when using workflows already prepared as a cloud service. Clouds are attractive both for scaling with computational demands and for the enormous diversity of tools one can integrate. To minimize the costs and to complete the computation as quickly as possible, workflows should be well considered to select which tool to use while performing VS studies.

VS projects demand the investment of considerable time prior to the computation for their preparation and afterwards for *in vitro* analyses. For any isolated project it is likely to be more cost effective to share the compute facilities and expertise by using cloud-based solutions.

The authors of this text hope to have helped with an initial team building for interdisciplinary projects: There is a lot of good work to be done, algorithmically or on pre-clinical data at hand, but few groups can develop and use these techniques all on their own.

Acknowledgement. This work was partially supported by the Scientific and Technological Research Council of Turkey (Technology and Innovation Funding Programmes Directorate Grant No. 7141231 and Academic Research Funding Program Directorate Grant No. 112S596) and EU financial support, received through the cHiPSet COST Action IC1406.

References

1. U.S. Food and Drug Administration (FDA) - The Drug Development Process. https://www.fda.gov/ForPatients/Approvals/Drugs. Accessed 31 Dec 2018
2. Kopp, S.: Definition of active pharmaceutical ingredient revised, pp. 1–4 (2011)
3. Hughes, J.P., Rees, S.S., Kalindjian, S.B., Philpott, K.L.: Principles of early drug discovery. Br. J. Pharmacol. **162**, 1239–1249 (2011)
4. Cronk, D.: High-throughput screening. In: Drug Discovery and Development, pp. 95–117. Elsevier, Amsterdam (2013)
5. Introduction to pharmacokinetics and pharmacodynamics. In: Concepts in Clinical Pharmacokinetics, pp. 1–18. ASHP (2014)
6. Olğaç, A., Orhan, I.E., Banoglu, E.: The potential role of in silico approaches to identify novel bioactive molecules from natural resources. Future Med. Chem. **9**, 1663–1684 (2017)
7. Kinghorn, A.D., Pan, L., Fletcher, J.N., Chai, H.: The relevance of higher plants in lead compound discovery. J. Nat. Prod. **74**, 1539–1555 (2011)
8. Taylor, P., Mueller-Kuhrt, L.: Successful, but often unconventional: the continued and long-term contribution of natural products to healthcare (2015)
9. Wöhler, F.: Ueber künstliche Bildung des Harnstoffs. Ann. Phys. **88**, 253–256 (1828)
10. Hafner, V.K.: Gmewue 91. Angewandte Chemie **91**, 685–695 (1979)
11. Hoffmann, R.W.: Natural product synthesis: changes over time. Angewandte Chemie - Int. Ed. **52**, 123–130 (2013)
12. Mendel, G.: Versuche über Plflanzenhybriden. Verhand-lungen des naturforschenden Vereines Brünn, Bd. IV für das Jahr 1865, Abhandlungen, pp. 3–47 (1866)

13. Kaufmann, S.H.E.: Paul Ehrlich: founder of chemotherapy. Nat. Rev. Drug Discov. **7**, 373 (2008)
14. Strebhardt, K., Ullrich, A.: Paul Ehrlich's magic bullet concept: 100 years of progress. Nat. Rev. Cancer **8**, 473–480 (2008)
15. Burghes, A.H.M., Vaessin, H.E.F., de la Chapelle, A.: The land between mendelian and multifactorial inheritance. Science **293**, 2213–2214 (2001)
16. Roth, B.L., Sheggler, D.J., Kroeze, W.K.: Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat. Rev. Drug Discov. **3**, 353–359 (2004)
17. Ma, Q., Lu, A.Y.H.: Pharmacogenetics, pharmacogenomics, and individualized medicine. Pharmacol. Rev. **63**, 437–459 (2011)
18. Cariaso, M., Lennon, G.: SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res. **40**, 1308–1312 (2012)
19. Fischer, E.: Ueber die optischen Isomeren des Traubenzuckers, der Gluconsaure und der Zuckersaure. Berichte der Dtsch. Chem. Gesellschaft. **23**, 2611–2624 (1890)
20. Fischer, E.: Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der Dtsch. Chem. Gesellschaft. **27**, 2985–2993 (1894)
21. Ferenczy, G.G., Keseru, G.M.: Thermodynamics guided lead discovery and optimization. Drug Discov. Today **15**(21–22), 919–932 (2010)
22. Bohacek, R.S., McMartin, C., Guida, W.C.: The art and practice of structure-based drug design: a molecular modeling perspective. Med. Res. Rev. **16**, 3–50 (1996)
23. Kirkpatrick, P.: New horizons in chemical space. Nat. Rev. Drug Discov. **3**, 375 (2004)
24. Kirkpatrick, P., Ellis, C.: Chemical space. Nature **432**, 823 (2004)
25. Dobson, C.M.: Chemical space and biology. Nature **432**, 824–828 (2004)
26. Lipinski, C., Hopkins, A.: Navigating chemical space for biology and medicine. Nature **432**, 855–861 (2004)
27. Chemical Abstracts Service (CAS) - CAS Databases. https://www.cas.org/content/cas-databases
28. Heller, S.R.: The Beilstein online database. In: ACS Symposium Series, vol. 436, pp. 1–9 (1990)
29. Huggins, D.J., Sherman, W., Tidor, B.: Rational approaches to improve selectivity in drug design. J. Med. Chem. **55**, 1424–1444 (2012)
30. Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D.: Is there a difference between leads and drugs? A historical perspective. J. Chem. Inf. Comput. Sci. **41**, 1308–1315 (2001)
31. Villoutreix, B.O., Lagorce, D., Labbe, C.M., Sperandio, O., Miteva, M.A.: One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. Drug Discov. Today **18**, 1081–1089 (2013)
32. Jelliffe, R.W., Tahani, B.: Pharmacoinformatics: equations for serum drug assay error patterns; implications for therapeutic drug monitoring and dosage. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 517–521 (1993)
33. Olgaç, A., Carotti, A.: Pharmacoinformatics in drug R&D process. In: GPSS, p. 45 (2015)
34. Levitt, M.: The birth of computational structural biology. Nat. Struct. Biol. **8**, 392–393 (2001)
35. Hopkins, A.L., Groom, C.R., Alex, A.: Ligand efficiency: a useful metric for lead selection. Drug Discov. Today **9**, 430–431 (2004)
36. Guedes, R., Serra, P., Salvador, J., Guedes, R.: Computational approaches for the discovery of human proteasome inhibitors: an overview. Molecules **21**, 1–27 (2016)
37. Petterson, I., Balle, T., Liljefors, T.: Ligand based drug design. In: Textbook of Drug Design and Discovery, pp. 43–57 (2010)
38. Aki-Yalcin, E., Yalcin, I.: Kantitatif Yapı-Etki İlişkileri Analizleri (QSAR). Ankara Üniversitesi Eczacılık Fakültesi Yayınları (2003)

39. Crum-Brown, A., Fraser, T.R.: On the connection between chemical constitution and physiological action. Part II - on the physiological action of the ammonium bases derived from Atropia and Conia. Trans. R. Soc. Edinburgh **25**, 693 (1869)
40. Hammett, L.P.: Effect of structure upon the reactions of organic compounds. Benzene derivatives. J. Am. Chem. Soc. **59**, 96–103 (1937)
41. Taft, R.W.: Steric Effects in Organic Chemistry (1956)
42. Free, S.M., Wilson, J.W.: A mathematical contribution to structure-activity studies. J. Med. Chem. **7**, 395–399 (1964)
43. Fujita, T., Ban, T.: Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. J. Med. Chem. **14**, 148–152 (1971)
44. Kubinyi, H.: Quantitative structure-activity relationships. 2. A mixed approach, based on Hansch and Free-Wilson analysis. J. Med. Chem. **19**, 587–600 (1976)
45. Kubinyi, H.: Quantitative structure-activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. J. Med. Chem. **20**, 625–629 (1977)
46. Polanski, J.: Receptor dependent multidimensional QSAR for modeling drug - receptor interactions. Curr. Med. Chem. **16**, 3243–3257 (2009)
47. Sippl, W.: 3D-QSAR - Applications, recent advances, and limitations. In: Recent Advances in QSAR Studies, pp. 103–126 (2010)
48. Klebe, G., Abraham, U., Mietzner, T.: Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem. **37**, 4130–4146 (1994)
49. Lima, A.N., Philot, E.A., Trossini, G.H.G., Scott, L.P.B., Maltarollo, V.G., Honorio, K.M.: Use of machine learning approaches for novel drug discovery. Expert Opin. Drug Discov. **11**, 225–239 (2016)
50. van der Kamp, M.W., Shaw, K.E., Woods, C.J., Mulholland, A.J.: Biomolecular simulation and modelling: status, progress and prospects. J. R. Soc. Interface **5**(Suppl. 3), S173–S190 (2008)
51. Berman, H.M., et al.: The Protein Data Bank. Nucleic Acids Res. **28**, 235–242 (2000)
52. Rose, P.W., et al.: The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. **43**, D345–D356 (2015)
53. RCSB Protein Data Bank. http://www.rcsb.org. Accessed 31 Dec 2018
54. Berman, H., Henrick, K., Nakamura, H., Markley, J.L.: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. **35**, 2006–2008 (2007)
55. Protein Data Bank in Europe. http://www.ebi.ac.uk/pdbe. Accessed 31 Dec 2018
56. Protein Data Bank Japan. https://pdbj.org. Accessed 31 Dec 2018
57. Biological Magnetic Resonance Data Bank. http://www.bmrb.wisc.edu
58. Jorgensen, F.S., Kastrup, J.S.: Biostructure based modeling. In: Textbook of Drug Design and Discovery, pp. 29–42 (2010)
59. Bateman, A., et al.: UniProt: A hub for protein information. Nucleic Acids Res. **43**, D204–D212 (2015)
60. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., Gish, W.: WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res. **31**, 3795–3798 (2003)
61. Altschul, S.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
62. Goujon, M., et al.: A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. **38**, 695–699 (2010)
63. Finn, R.D., et al.: Pfam: the protein families database. Nucleic Acids Res. **42**, 222–230 (2014)

64. Fenu, L.A., Lewis, R.A., Good, A.C., Bodkin, M., Essex, J.W.: Scoring functions. In: Jhoti, H., Leach, A.R. (eds.) Structure-Based Drug Discovery, pp. 223–245. Springer, Dordrecht (2007). https://doi.org/10.1007/1-4020-4407-0_9
65. Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S., Olson, A.J.: Computational protein–ligand docking and virtual drug screening with the AutoDock suite. Nat. Protoc. **11**, 905–919 (2016)
66. Schomburg, K.T., Bietz, S., Briem, H., Henzler, A.M., Urbaczek, S., Rarey, M.: Facing the challenges of structure-based target prediction by inverse virtual screening. J. Chem. Inf. Model. **54**, 1676–1686 (2014)
67. Rognan, D.: Structure-based approaches to target fishing and ligand profiling. Mol. Inform. **29**, 176–187 (2010)
68. Moura Barbosa, A.J., Del Rio, A.: Freely accessible databases of commercial compounds for high-throughput virtual screenings. Curr. Top. Med. Chem. **12**, 866–877 (2012)
69. Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K.: BindingDB: a web-accessible database of experimentally determined protein – ligand binding affinities. Nucleic Acids Res. **35**, 198–201 (2007)
70. Gaulton, A., et al.: The ChEMBL database in 2017. Nucleic Acids Res. **45**, 1–10 (2016)
71. Pence, H.E., Williams, A.: ChemSpider: an online chemical information resource. J. Chem. Educ. **87**, 1123–1124 (2010)
72. Wishart, D.S., et al.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. **34**, D668–D672 (2006)
73. Wishart, D.S., et al.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. **36**, 901–906 (2008)
74. Knox, C., et al.: DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. Nucleic Acids Res. **39**, 1035–1041 (2011)
75. Law, V., et al.: DrugBank 40: shedding new light on drug metabolism. Nucleic Acids Res. **42**, 1091–1097 (2014)
76. Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., Shi, T.: TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res. **41**, 1089–1095 (2013)
77. Masciocchi, J., et al.: MMsINC: a large-scale chemoinformatics database. Nucleic Acids Res. **37**, 284–290 (2009)
78. Williams, A.J., et al.: Open PHACTS: Semantic interoperability for drug discovery. Drug Discov. Today. **17**, 1188–1198 (2012)
79. Kim, S., et al.: PubChem substance and compound databases. Nucleic Acids Res. **44**, D1202–D1213 (2016)
80. Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S., Coleman, R.G.: ZINC: a free tool to discover chemistry for biology. J. Chem. Inf. Model. **52**, 1757–1768 (2012)
81. Koutsoukas, A., et al.: From in silico target prediction to multi-target drug design: current databases, methods and applications (2011)
82. Tanimoto, T.T.: An elementary mathematical theory of classification and prediction. International Business Machines Corporation (1958)
83. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. **23**, 3–25 (1997)
84. Lipinski, C.A.: Lead- and drug-like compounds: the rule-of-five revolution (2004)
85. Chang, M.W., Lindstrom, W., Olson, A.J., Belew, R.K.: Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. J. Chem. Inf. Model. **47**, 1258–1262 (2007)
86. Xue, Y., Wilcox, W.R.: Changing paradigm of cancer therapy: precision medicine by next-generation sequencing. Cancer Biol. Med. **13**, 12–18 (2016)

87. Möller, S., et al.: Robust cross-platform workflows: how technical and scientific communities collaborate to develop, test and share best practices for data analysis. Data Sci. Eng. **2**, 232–244 (2017)
88. Ren, J., Williams, N., Clementi, L., Krishnan, S., Li, W.W.: Opal web services for biomedical applications. Nucleic Acids Res. **38**, 724–731 (2010)
89. Berthold, M.R., et al.: KNIME - the Konstanz information miner. SIGKDD Explor. **11**, 26–31 (2009)
90. Ison, J., et al.: Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. **44**, 38–47 (2016)
91. Palmblad, M., Lamprecht, A.-L., Ison, J., Schwämmle, V.: Automated workflow composition in mass spectrometry based proteomics. Bioinformatics **35**(4), 656–664 (2019). https://www.ncbi.nlm.nih.gov/pubmed/30060113
92. Balan, D.M., Malinauskas, T., Prins, P., Möller, S.: High-throughput molecular docking now in reach for a wider biochemical community. In: 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, pp. 617–621 (2012)
93. Amazon Web Services. https://aws.amazon.com. Accessed 31 Dec 2018
94. Google Cloud. https://cloud.google.com. Accessed 31 Dec 2018
95. Open Stack. https://www.openstack.org/. Accessed 31 Dec 2018
96. Open Whisk. https://openwhisk.apache.org/. Accessed 31 Dec 2018
97. Banegas-Luna, A.J., et al.: Advances in distributed computing with modern drug discovery. Expert Opin. Drug Discov. **14**, 9–22 (2019)
98. Potemkin, V., Grishina, M., Potemkin, A.: Internet resources for drug discovery and design. Curr. Top. Med. Chem. **18**(22), 1955–1975 (2018). https://www.ncbi.nlm.nih.gov/pubmed/30499394
99. Click2Drug Catalog. https://www.click2drug.org/directory_Docking.html. Accessed 31 Dec 2018
100. Harvey, M.J., De Fabritiis, G.: AceCloud: molecular dynamics simulations in the cloud. J. Chem. Inf. Model. **55**, 909–914 (2015)
101. Doerr, S., Harvey, M.J., Noé, F., De Fabritiis, G.: HTMD: High-throughput molecular dynamics for molecular discovery. J. Chem. Theory Comput. **12**, 1845–1852 (2016)
102. Sánchez-linares, I., Pérez-sánchez, H., Cecilia, J.M., García, J.M.: High-throughput parallel blind virtual screening using BINDSURF. BMC Bioinform. **13**, S13 (2012)
103. Skalic, M., Martinez-Rosell, G., Jimenez, J., De Fabritiis, G.: PlayMolecule BindScope: large scale CNN-based virtual screening on the web. Bioinformatics 1–2 (2018)
104. Antunes, D.A., Moll, M., Devaurs, D., Jackson, K.R., Kavraki, L.E., Liz, G.: DINC 2.0: a new protein – peptide docking webserver using an incremental approach, pp. 2017–2020 (2017)
105. Irwin, J.J., et al.: Automated docking screens: a feasibility study, 5712–5720 (2009)
106. Bikadi, Z., Hazai, E.: Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. J. Cheminform. **16**, 1–16 (2009)
107. Olgac, A., Budak, G., Cobanoglu, S., Nuti, R., Carotti, A., Banoglu, E.: Evias web services: cloud-based drug discovery platform. In: EuroQSAR 2016, p. 79 (2016)
108. Yan, Y., Zhang, D., Zhou, P., Li, B., Huang, S.Y.: HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. **45**, W365–W373 (2017)
109. Zundert, G.C.P., et al.: The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J. Mol. Biol. **128**, 720–725 (2016)
110. Li, H., Leung, K.S., Wong, M.H.: idock: A multithreaded virtual screening tool for flexible ligand docking. In: 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2012, pp. 77–84 (2012)

111. Tsai, T.-Y., Chang, K.-W., Chen, C.Y.-C.: iScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. J. Comput. Aided Mol. Des. **25**, 525–531 (2011)
112. Kainrad, T., Hunold, S., Seidel, T., Langer, T.: LigandScout remote : a new user-friendly interface for HPC and cloud resources, 1–9 (2018)
113. Kiss, R., Sandor, M., Szalai, F.A.: http://Mcule.com: a public web service for drug discovery. J. Cheminform. **4**, P17 (2012)
114. Chang, D.T.H., Oyang, Y.J., Lin, J.H.: MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. Nucleic Acids Res. **33**, 233–238 (2005)
115. Labbe, C.M., et al.: MTiOpenScreen: a web server for structure-based virtual screening. Nucleic Acids Res. **43**, 448–454 (2015)
116. Gupta, A., Gandhimathi, A., Sharma, P., Jayaram, B.: ParDOCK: an all atom energy based Monte Carlo docking protocol for protein-ligand complexes. Protein Pept. Lett. **14**, 632–646 (2007)
117. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J.: PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. **33**, 363–367 (2005)
118. Duhovny, D., Nussinov, R., Wolfson, H.J.: Efficient unbound docking of rigid molecules, 185–200 (2002)
119. Awale, M., Reymond, J.: The polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. J. Chem. Inf. Model. **59**(1), 10–17 (2019). https://www.ncbi.nlm.nih.gov/pubmed/30558418
120. Konc, J., Janezic, D.: ProBiS: a web server for detection of structurally similar protein binding sites. Nucleic Acids Res. **38**, 436–440 (2010)
121. Konc, J., et al.: ProBiS-CHARMMing: web interface for prediction and optimization of ligands in protein binding sites. J. Chem. Inf. Model. **55**, 2308–2314 (2015)
122. Grosdidier, A., Zoete, V., Michielin, O.: SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. **39**, 270–277 (2011)
123. Li, H., Leung, K.-S., Wong, M.-H., Ballester, P.J.: USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. Nucleic Acids Res. **44**, W436–W441 (2016)
124. Koes, D.R., Camacho, C.J.: ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Res. **40**, 409–414 (2012)