CHAPTER FOUR

Use of molecular docking computational tools in drug discovery

Francesca Stanzione*, Ilenia Giangreco, and Jason C. Cole
Cambridge Crystallographic Data Centre, Cambridge, United Kingdom
*Corresponding author: e-mail address: fstanzione@ccdc.cam.ac.uk

Contents

Keywords: Drug discovery, Molecular target, Homology modelling, Molecular interaction, Virtual screening (VS), Structure-based drug design (SBDD), Scoring functions, Protein databases, Small molecule databases

1. Introduction

Drug discovery is a challenging process and identifying the right lead compound is a determining factor of the overall success of the project. In 2016, the Tufts Center for the Study of Drug Development estimated that the cost associated with developing and bringing a new drug to market has

Progress in Medicinal Chemistry, Volume 60
ISSN 0079-6468
https://doi.org/10.1016/bs.pmch.2021.01.004

increased by almost 145% in the last decade [1]. Furthermore, while the average time to bring a drug to clinical trials has decreased, the success rate of drugs obtaining the US Food and Drug Administration (FDA) approval has dropped to 12% [1]. Computer-aided drug design (CADD) has helped to reduce the costs and the time associated with drug discovery by directing experimental research towards optimal compounds more quickly. Within CADD, techniques such as molecular docking and virtual screening (VS) have provided a valuable complement to the time-consuming and expensive experimental process of high-throughput screening (HTS).

The ability to computationally screen large libraries of compounds that either possess similarity towards known inhibitors (ligand-based) or complementarity towards target structures (structure-based) has proven to be successful at identifying highly focused subsets from which actives can then be experimentally confirmed [2]. To date, generating such 'educated guesses' has provided many examples of lead compounds. Several marketed drugs, such as imatinib [3], zanamivir [4], nelfinavir [5], erdafitinib [6], and several clinical candidates are known to have been discovered or optimised with the aid of computational methodologies [7–9]. We can only speculate as to the number of unpublished examples that reside in corporate collections, but we believe this to be considerable.

In the case of molecular docking, the process of predicting the best position, orientation and conformation of a small molecule (drug candidate) when bound to a protein, provides the additional benefit of simplifying future lead optimisation [10]. Knowing exactly how and where a ligand binds helps to rationally design changes to optimise the protein-ligand interaction, to improve activity, and to avoid changes that could lead to protein-ligand clashed.

2. Molecular docking

Molecular docking is the most common computational structure-based drug design (SBDD) method and has been widely used ever since the early 1980s [11]. It is the tool of choice when the three-dimensional (3D) structure of the protein target is available. Molecular docking popularity has been facilitated by the dramatic growth in availability and power of computers, and the increasing number of and ease of access to small molecule and protein structures.

The main goal of molecular docking is to understand and predict molecular recognition, both structurally (i.e. finding possible binding modes) and energetically (i.e. predicting binding affinity). Molecular docking was originally designed to be performed between a small molecule (ligand) and a target macromolecule (protein) however, in the last decade there has been a growing interest in protein-protein docking, nucleic acid (DNA and RNA)-ligand docking and nucleic acid-protein-ligand docking. In this chapter, we will focus on protein-ligand docking and we will provide an overview on the challenges related to protein-protein docking, protein-peptide docking and nucleic acid-ligand docking.

Molecular docking applications in drug discovery are varied, including structure-activity studies, lead optimisation, finding potential leads by virtual screening, providing binding hypotheses to facilitate predictions for mutagenesis studies and also in assisting X-ray and cryogenic electron microscopy (cryo-EM) crystallography in the fitting of substrates and inhibitors to electron density.

Docking has proved to be extremely successful in SBDD, therefore it has been developed and improved for many years. Over the last 2 decades, more than 60 different docking tools have been developed in academic and commercial settings. Some of these programs will be mentioned in this chapter, but there are several available reviews that detail evaluation and comparison of the different docking programs [12–15].

2.1 Theory of docking

The molecular docking process involves two basic steps: prediction of the ligand (usually a small molecule) conformation as well as its position and orientation within the protein binding site (usually referred to as pose) and assessment of the quality of the pose using a scoring function. Ideally, the sampling algorithm should be able to reproduce the experimental binding mode and the scoring function should also rank it highest among all generated poses.

A further task of the docking procedure would be to score active compounds higher than known inactives (*predictive docking*). However, this level of accuracy is difficult to achieve and it is generally influenced by many factors that are external to the protein. Therefore, primarily a docking algorithm only aims to get the prediction of the ligand pose and assessment of quality of the pose correct (though many scoring functions are developed with active/inactive ranking as a consideration too).

2.2 Searching algorithm

The sampling process is non-trivial. The conformational space involves many degrees of freedom including the rotation and translation of a molecule relative to another, the additional conformational degrees of freedom of both the ligand and the protein and sometimes degrees of freedom related to the solvent [16]. In practice, with current computational resources, it is impossible to explore the search space exhaustively by enumerating all possible conformations and all possible rotational and translational orientations of a single molecule relative to a protein within a second of elapsed time (a time scale that is realistically needed for virtual screening). Hence, efficient sampling of conformational space is still a challenge in molecular docking.

Early approaches to account for these sampling problems treated both the ligand and the protein as rigid bodies thereby reducing the number of degrees of freedom to just six. Such approaches rely on the shape similarities between the ligand and the protein binding site. A very well cited example of a program using this algorithm is DOCK [17].

In a rigid docking approach, the ligand and the protein binding sites are represented as pharmacophore spheres of varying radii and the search algorithm tries to pair the ligand spheres with the protein spheres based on the match of the internal distances of all the ligand's spheres and the internal distances of all the protein binding site's spheres. The ligand is then oriented within the binding site using a least square fit of the atoms to the sphere centres [18,19]. In case of unacceptable orientation (e.g. clashes between the ligand and the protein binding site), the ligand is reoriented until an acceptable orientation is obtained. The orientation is then scored based on the degree of overlap between the ligand and protein pharmacophore spheres (Fig. 1).

Despite its computational efficiency, failing to model molecular flexibility limits effectiveness as the conformation is interlinked with protein-ligand interactions; the optimal binding conformation of a small molecule is a compromise between the best internal geometry of said small molecule and the interactions it forms with the binding site. Rigid docking fails to account for this. This is even more relevant in predictive docking where there is the additional complexity derived from the conformational change of the ligand from its unbound (isolated) conformation and its bioactive (bound) conformation.

To overcome to such limitations, most docking programs reached a trade-off and started to account for the whole conformational space of the ligand while limiting flexibility in the protein to regions of the receptor (e.g. binding

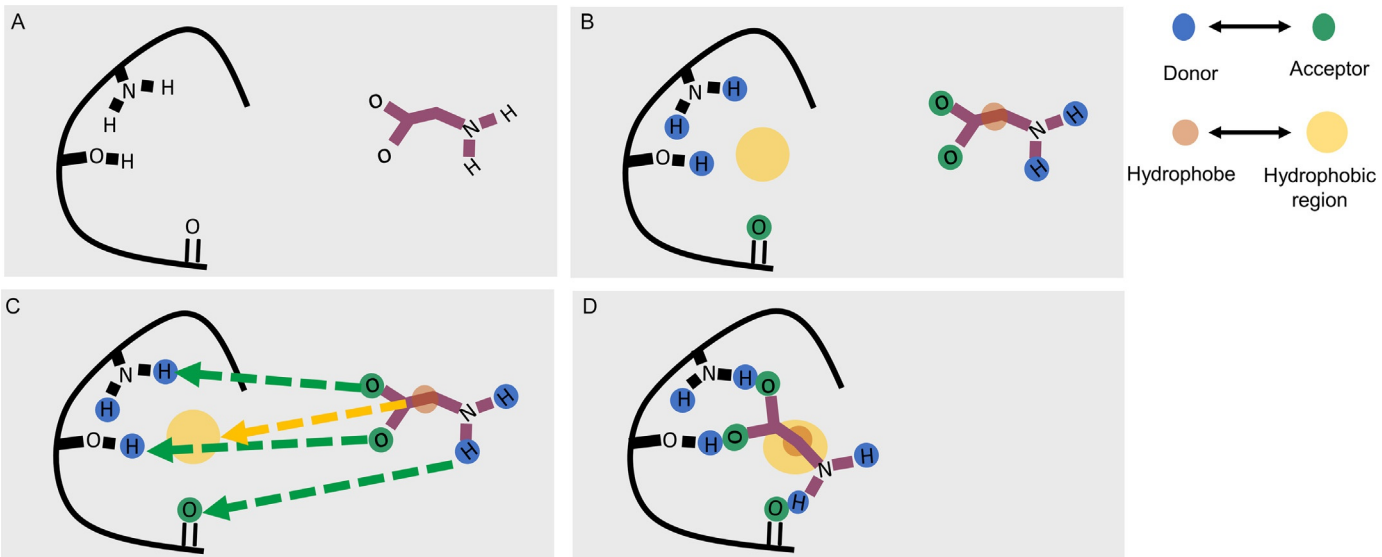

Fig. 1 Rigid docking approach: (A) protein and ligand. (B) The initial pose generation through pharmacophore point matching. Donor–acceptor pharmacophore point and hydrophobic pharmacophore points are added to the protein and the ligand. (C) The searching algorithm tries to match the protein and the ligand fitting points by matching donor with acceptors and hydrophobic atoms with hydrophobic cavities. (D) Different solutions are found.

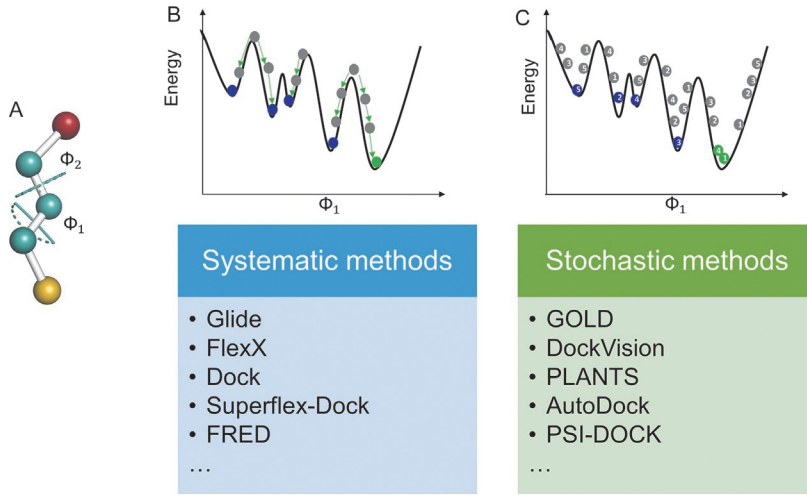

Fig. 2 Small molecule conformational search methods. (A) A molecule containing two bulky groups (red and yellow spheres) has its conformation defined by two internal dihedrals Φ_1 and Φ_2. If we freeze the Φ_2 dihedral, the energy variation due to rotation of Φ_1 can be plotted in a 1D energy landscape. The initial structure (grey spheres) is modified by changing the Φ_1 dihedral. (B) In the systematic search approach, the changes are applied to all structural parameters until local (blue spheres) or global (green sphere) energy minimum is reached; a few examples of docking programs using systematic methods are also listed. (C) Stochastic search approaches explore the conformational space by generating distinct conformations, with an element of randomness, populating a broad range of the energy landscape. A few examples of docking programs using stochastic search algorithms are also listed. *Figure adapted from Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Molecules 2015;20(7):13384–13421.*

site), or keeping the entire protein rigid. The underlying reason is that, since ligands are small molecules, they are likely to undergo larger conformational change; furthermore, given their smaller size, accounting for ligand flexibility is also computationally affordable. The ligand conformations can be sampled with both systematic and stochastic methods (Fig. 2).

2.2.1 Systematic methods

Systematic search methods sample the ligand search space at predefined intervals and are deterministic; they can be classified as exhaustive, fragmentation or conformational ensemble methods. The main difference between them is in the approach they take to deal with the ligand flexibility. In exhaustive search methods, for example, the docking is performed by systematically rotating all possible rotatable bonds in the ligand at a given

interval. Despite its sampling completeness, the number of possible combinations is huge and increases with the number of rotatable bonds in the ligand (Fig. 2). Therefore, this approach is limited to small, relatively flexible ligands and, in most cases, to make the docking 'practical', it is necessary to apply geometrical or chemical constraints to the initial screening of the ligand poses. A well cited example of a program using the exhaustive sampling method is Glide [20,21].

The fragmentation method is an incremental approach where ligands are divided into modular pieces. One of the fragments is anchored to the protein binding site and then the ligand binding conformation is incrementally grown by placing additional fragments one at the time. The anchor is generally chosen to be the fragment which shows maximum interaction complementarity with the receptor surface (i.e. H-bonds), has the minimum number of alternate conformations, and is fairly rigid (for example, a ring system). One docking program that uses a fragmentation sampling approach is FlexX [22]. A variation of this approach is to dock all the fragments into the binding site and then link them covalently.

As for the exhaustive method, the fragmentation method is restricted to medium and smaller sized ligands and it is not feasible for big ligands where the number of fragments would be too large.

In conformational ensemble methods, the ligand flexibility is represented by rigidly docking an ensemble of pre-generated ligand conformations. Using such an approach removes the computational cost due to the exploration of ligand conformational space, however, it involves additional tools to generate the required ensemble of conformations of the ligand. One limitation in this approach is that the ensemble of generated conformers may not include the bioactive conformation of the ligand.

2.2.2 Stochastic methods

In stochastic algorithms, the ligand binding orientations and conformations are sampled by making changes to the ligand that have some dependence on one or more values generated at random at each step (Fig. 2). The change is then accepted or rejected according to an algorithm-dependent criterion.

The advantage of stochastic algorithms is that they can generate large ensembles of molecular conformations and explore a broad range of the energy landscape increasing the probability of finding a global energy minimum. However, this also means that computational costs associated with this procedure represent an important limitation. Genetic algorithm, Monte Carlo, ant colony optimization (ACO) and tabu search methods are a few

examples of stochastic algorithms that differ in the way they generate given moves and the probability criteria of acceptance.

In a genetic algorithm method, a population of potential solutions is set up at random. Each member of the population is encoded as a 'chromosome', which contains information about the mapping of ligand fitting points (e.g. H-bond atoms) onto the complementary protein's fitting points. Each chromosome is assigned a fitness score based on its predicted binding affinity and the chromosomes within the population are ranked according to fitness score. At each step, a point mutation may occur in a chromosome, while the crossover operator exchanges information between two chromosomes of the population. Other operations are also used in some implementations (for example, island migration). GOLD [23] is one of the well cited docking programs that uses a genetic algorithm to explore the ligand conformational space. DockVision [24] is an example of docking program that uses Monte Carlo stochastic method where the probability to accept a random change is calculated by using the Boltzmann probability function. PLANTS [25], instead, is an example of a docking program based on ACO. ACO is inspired by the behaviour of real ants finding the shortest path between their nest and a food source. In the case of protein-ligand docking, an artificial ant colony is employed to find a minimum energy conformation of the ligand in the binding site. These ants mimic the behaviour of real ants and mark low energy ligand conformations with pheromone trails. The artificial pheromone trail information is then modified in subsequent iterations to generate low energy conformations with a higher probability [25].

The tabu search method is a variation of the Monte Carlo approach which maintains a record of the search space of the binding site which has already been visited and thus ensures that the binding site is explored to the maximum. PSI-DOCK [26] is an example of a docking tool that uses a tabu search.

2.2.3 Scoring functions

Scoring functions are fast approximate mathematical methods used to predict the strength of the interaction (or binding affinity) between two or more molecules.

Four aspects should be considered when assessing the reliability of a scoring function [27]: (1) *scoring power*: the ability to produce scores which linearly correlate with experimental binding affinity data, (2) *ranking power*: the ability to correctly rank a given set of ligands that bind to a common target protein by their binding affinities when their binding poses are known,

(3) *docking power*: the ability to identify the native binding pose of a ligand as the one with the best score, and when screening a large set of generated decoy poses, (4) *screening power*: the ability to identify the true binders to a given target protein among a library of random molecules. Ideally, an accurate scoring function would perform equally well on all these four tasks; however, each existing scoring function only perform well on one or two of them at the same time.

Scoring functions can be grouped into four main classes: physics-based, empirical, knowledge-based and machine learning-based scoring functions [28].

The first three types are commonly referred to as 'classical' scoring functions and are based on the assumptions that the change in free energy upon binding of a ligand to its target can be decomposed into a sum of individual energy contributions, and that all these energy contributions are linearly combined. In reality, such linear correlation may not always exist [29]. Two major limitations of classical scoring functions are their minimal description of protein flexibility and the implicit treatment of solvent.

Machine learning-based scoring functions instead use more sophisticated techniques, such as random forests (RF), support vector machines (SVM), and deep learning (DL), to approximate non-linear problems (Fig. 3).

Physics-based or force-field based scoring functions compute the binding energy by summing up the contribution of the bonded interactions (bond stretching, angle bending and torsion angles) and non-bonded interactions (van der Waals and electrostatic interactions) within the protein-ligand complex which accounts for the contribution of enthalpy to energy. Hydrogen bonds are usually considered by adding an additional term to the binding energy. Alternatively, they can be included implicitly in the electrostatic energy term.

Parameters for this type of scoring function are usually derived from both experimental data and ab initio quantum mechanical calculations. The major challenge for physics-based scoring functions is the treatment of the solvent in ligand binding. To overcome this limitation, implicit solvent approaches like Poisson–Boltzmann (PB) or Generalised-Born (GB) continuum solvation models have been widely used [30]. However, more computationally expensive approaches that treat water molecules explicitly are also available (such as free energy perturbation (FEP) and thermodynamic integration (TI) techniques) [31] (Eq. 1).

$$E_{bind} = E_{bond} + E_{no-bond} + [E_{H-bond}] + E_{solv} \quad (1)$$

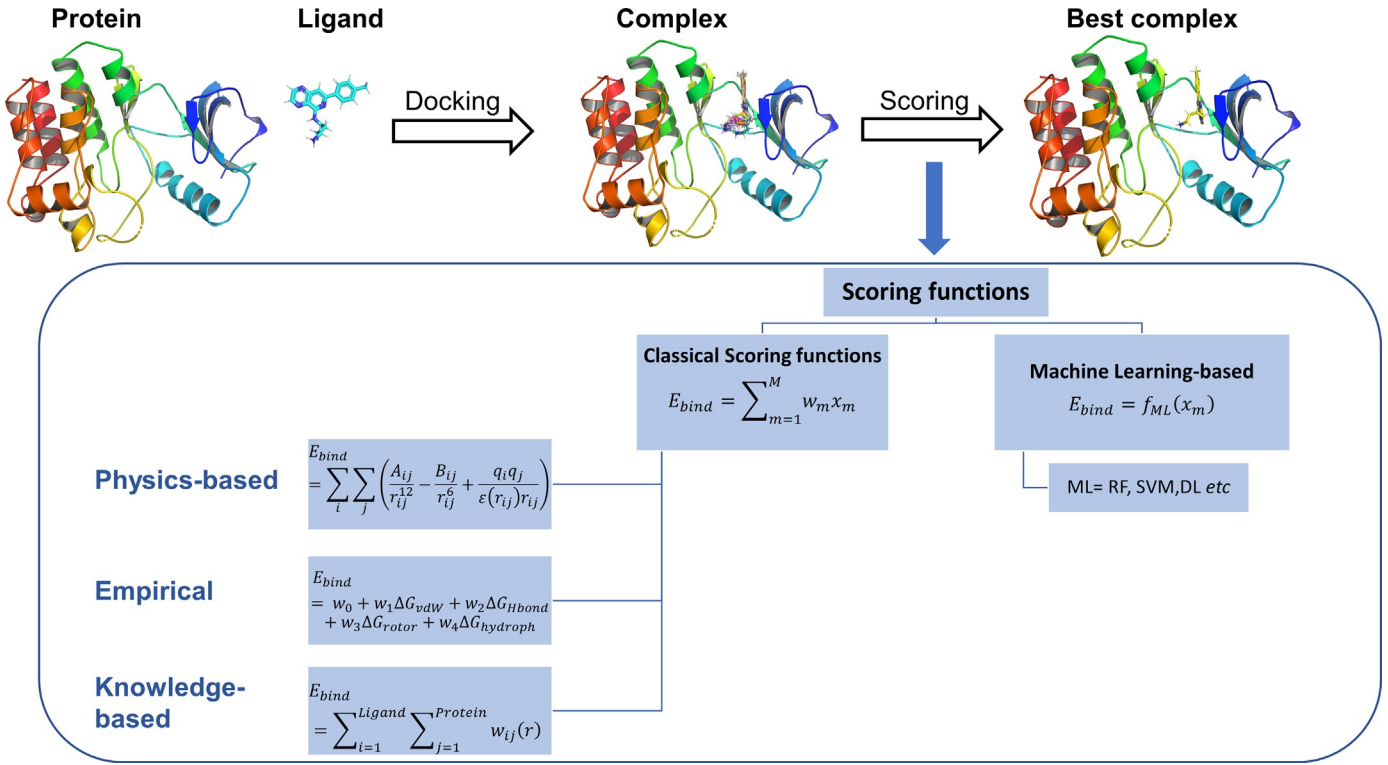

Fig. 3 Scoring functions in docking. Scoring functions can be divided into classical and machine learning scoring functions. The classical scoring functions are physics-based, empirical and knowledge-based. From a mathematical perspective, all classical scoring functions assume an additive functional form; however, they are distinguished by the type of structural descriptors employed. By contrast, non-parametric machine learning scoring functions do not make assumptions about the form of the functional. Instead, the functional form is inferred from training data in an unbiased manner. Machine learning-based scoring functions use more sophisticated machine learning techniques, such as random forests (RF), support vector machines (SVM), and deep learning (DL), to approximate non-linear problems.

In addition to the above, accounting for the entropic effect is an even more severe challenge for physics-based scoring functions. This shortcoming is due to the lack of a reasonable physical model to describe this phenomenon. In principle, the individual terms in Eq. (1) account for the main energetic contributions to the protein-ligand binding however, a separation into individual terms is only possible if the system of interest is divided into mutually independent subsystems [32]. In fact, many of the individual terms are highly correlated with each other and thus can affect the binding affinity in more than one way (i.e. positive or negative contribution) [33]. Moreover, whether the free energy of ligand binding can be decomposed into a linear combination of individual interaction terms without calculating the partition function (ensemble average) also remains in question (the 'non-additive' problem). Despite these approximations, physics-based scoring functions are very appealing as the simplifications result in functions that can be evaluated very rapidly, which is important in a high-throughput docking setting. Another obvious advantage of physics-based methods is that they can ride on the progress of modern force fields, quantum mechanics methods, solvation models, and other developments.

Despite the lucid physical meaning, rigorous physics-based scoring functions are normally computationally expensive. Examples of physics-based scoring functions are GoldScore [34], AutoDock [35], Generalised-Born Volume Integral/Weighted Surface area (GBVI/WSA) [36].

Empirical scoring functions estimate the binding affinity of a protein-ligand complex by summing up different energetic factors involved in the protein-ligand binding, such as hydrogen bonds, hydrophobic effects, protein-ligand clashes, etc. (see Fig. 3). Each factor is multiplied by a coefficient that is obtained from multiple linear regression analyses fitted to a training set of protein-ligand complexes with known binding affinities. Compared to force-field or physics-based scoring functions, empirical scoring functions are much faster in binding score calculations due to their simple treatment of the energy terms. However, the accuracy of empirical scoring functions is directly correlated to the accuracy and the coverage of the protein-ligand training set that is used to develop the model. Examples of empirical scoring functions are ChemScore [37], GlideScore [21] and ChemPLP [25,38].

Knowledge-based scoring functions use statistical analyses to derive the observed interatomic contact frequencies and/or distances in a large database of crystal structures of protein-ligand complexes and employ the Boltzmann law to transform the atom pair preferences into distance-dependent pairwise potentials (Eq. 2).

$$w_{ij}(r) = -k_B T \ln\left(\frac{\rho_{ij}(r)}{\rho^*(r)}\right) \quad (2)$$

In this equation k_B is the Boltzmann constant, T is the absolute temperature of the system, $\rho_{ij}(r)$ is the number density of the protein-ligand atom pair at distance r in the training set, and $\rho^*(r)$ is the pair density in a reference state where the interatomic interactions are zero. The protein-ligand density functions $\rho_{ij}(r)$ are constructed by summing the static densities observed in different proteins rather than by averaging different states of the same protein.

The knowledge-based scoring functions assume that favourable interactions occur with higher frequencies than less favourable interactions therefore, the final score is calculated by favouring preferred contacts and penalising repulsive interactions between each atom in the ligand and protein within a given cut-off. Compared to the physics-based and empirical scoring functions, knowledge-based scoring functions offer a good balance between accuracy and speed, because they do not rely on ab initio calculations (physics-based methods) or reproducing binding affinities (empirical methods). Also, because the training protein-ligand dataset database can be large and diverse, they are insensitive to the training set [39].

The main challenge in deriving knowledge-based scoring functions is represented by the calculation of the reference state ($\rho^*(r)$) [40]. Currently, there are two classical strategies used to determine this: traditional atom-randomised reference state and corrected reference state. Traditional methods approximate the reference state by the random distribution of atomic pairs in the training set. Examples include DrugScore [41,42] and GOLD/ASP [43]. The drawback of the atom-randomisation approximation is the neglect of the effects of excluded volume and interatomic connectivity [40]. To overcome these limitations, later approaches introduced correction terms such as the volume factor correction term for the reference state [44,45]. Nevertheless, the accuracy of the reference state remains a challenge for knowledge-based scoring functions. The problem is more relevant for binding mode predictions and virtual screening, as the pairwise potentials, derived from bound structures, are not sufficiently sensitive to different ligand positions and may give good scores even to bad binding modes. A third approach to solving this problem, is to circumvent the accurate calculation of the reference state using iterative methods [39,46]. The basic idea of this method is to adjust the pair potentials by iteration until the interaction potentials reproduce the experimentally determined pair distribution function in the training set, yielding a set of potentials that can

discriminate the native structures from decoys. During the iteration procedure, the improvement for the potentials is guided through the difference between the predicted and experimentally observed pair distribution functions rather than through accurate calculation of the reference state [47].

Other challenges for knowledge-based scoring functions include extension of the pairwise interactions to many-body interactions to account for hydrogen bonding and other directional interactions [48,49] and the development of an accurate method that includes the contributions from solvation and entropy [30,50,51].

Over the years several studies have assessed and compared different scoring functions; each one has its virtues and drawbacks. None of the scoring functions available outperforms the others in all tasks but each scoring function may perform better than others in a specific task [27,52]. To compensate for drawbacks in individual scoring functions, the simultaneous use of different scoring functions, or even individual terms of multiple scoring functions, has been widely used to obtain a consensus score [53]. One way to perform consensus scoring is to re-evaluate the best docked pose of each compound with other scoring functions. Only the top scored compounds common to each scoring function will be identified as candidates for bioassay. Consensus ranking is thought to increase hit rates, either by reducing the number of false positives or by statistically reducing the errors in the scores/ranks [34,54,55]. On the other hand, in some cases, single-scoring-function ranking has been shown to outperform consensus-ranking methods [56].

With the abundance of experimental biological, biochemical and biophysical data becoming available, a direct link between energetic and structural information of protein-ligand complexes has become accessible, leading to the design of data-oriented scoring functions using machine learning (ML) techniques [57]. These methods introduced quantitative structure-activity relationship (QSAR) analysis into the protein-ligand interaction evaluation. If the properties of the ligand and the protein, such as atom pairs or structural interaction fingerprints, as well as their interaction patterns (electrostatic interactions, hydrogen bonds, or aromatic stacking), geometrical descriptors (surface or shape properties), and conventional ligand-based descriptors (molecular weight, number of rotatable single bonds, etc.) can be encoded, then ML techniques can be applied to derive statistical models that compute protein-ligand binding scores [47].

Rather than the predetermined functional form, ML-based scoring functions can automatically learn both generalised non-linear functional forms and feature information from the training data. Thus, like empirical scoring

functions, ML-based scoring functions also need a training set of protein-ligand complexes with known structures and binding data to derive their final models. Given a set of active and inactive ligands for training, ML-based scoring functions can be trained to distinguish between known ligands by potency with high accuracy. Thus, they have gradually emerged as potential alternatives to classical scoring functions [58,59]. Recently, significant progress has been made on ML-based scoring. In particular a wide range of ML algorithms, such as random forest (RF), support vector machine (SVM), artificial neural network, gradient boosting decision tree (GBDT) and convolutional neural network (CNN), have been applied to the development of new scoring functions [28].

Machine learning scoring functions have been proved to outperform classical scoring functions in such tasks as ranking and screening however, they are rarely directly incorporated into docking software and are mostly used for rescoring [57,58,60]. Examples of machine learning scoring functions are RF-Score [61] and SVM-Score [62].

2.2.3.1 Tailored scoring functions

The intrinsic simplistic nature of the classic scoring functions cannot capture all the features involved in the protein-ligand binding. Some of these scoring functions performs better with one protein target or another based on their structural and chemical features. For example, it was observed that among metalloenzymes, the performance of different scoring functions varied based on the specific metal in the protein, its location in the binding site, type of dominant interactions, exposure to solvent, etc. [63]. Furthermore, the chemical space and properties (such as protonation state, partial charge, number of rotatable bonds) of the small molecules tested in docking has a significant impact on the performance of one scoring function versus another [64].

One can evaluate which scoring function to use for a target protein by experimenting and optimising the choice based on an actives/decoys test set for each specific protein target (Fig. 4).

This strategy is subject to the availability of experimental information on active ligands. Several databases such as BindingDB [65–67], PDBbind [68–70] and ChEMBL [71,72], store experimental binding affinity data of ligands against several protein targets.

This information can be used to select active and inactive ligands for the target of interest that can be used to test each scoring function. Alternatively, services like the DUD-E (database of useful decoys-enhanced) decoys server

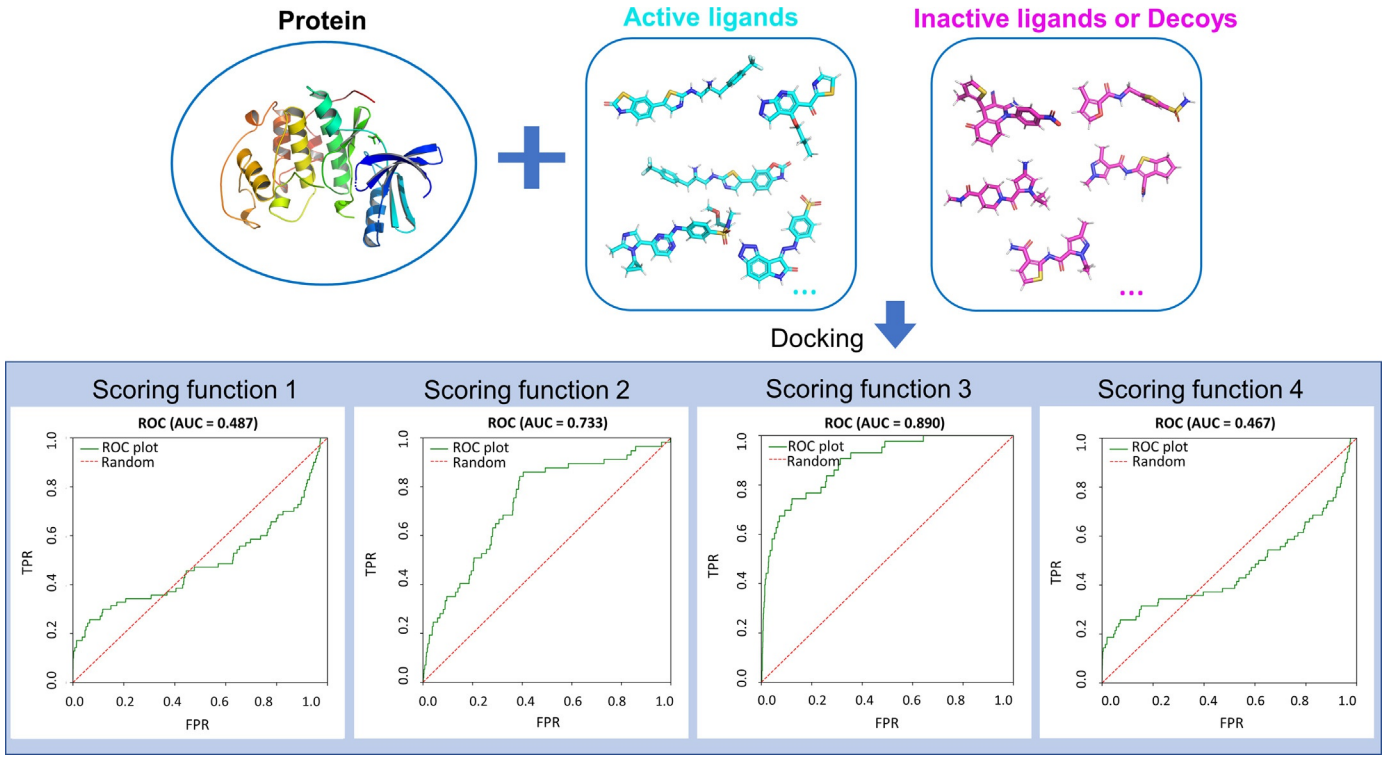

Fig. 4 Approach to evaluate scoring function for a specific protein target. Known active ligands and inactive ligands are docked together to evaluate the performance of different scoring functions. The performance of the docking run can be evaluated by the enrichment of annotated ligands of known binders (True Positive Rate—TPR) from among of non-binding, or decoy molecules (False Positive Rate—FPR). The area under the receiver operating characteristic (ROC) curve is widely used to evaluate its performance, where generally higher is better.

[73] can be adopted to generate decoy molecules with similar physicochemical properties of the actives. The customised actives/decoys test set can therefore be used to assess the performance of the different scoring functions to find the best one available for the specific target for the virtual screening campaign. In the absence of enough data (active ligands), scoring functions are usually selected based on the ability to reproduce the native ligand binding conformation (self-docking). Success of such prediction is defined by the Root Mean Square Deviation (RMSD) value between the top-ranked ligand conformations and the experimentally observed (native) structure. Generally, if the RMSD is 1.5 Å or below, the prediction is considered successful.

Because of its simplicity, the RMSD criterion has been widely used to evaluate the prediction power of scoring functions. However, it may not be reliable for small or symmetrical ligands that are likely to have good RMSD values even when they are randomly placed in a protein binding site. On the other hand, for large flexible ligands, a large RMSD value due to a solvent exposed and/or substituent groups, may hide the correctness in prediction of the overall binding mode. To overcome these limitations, several alternative methods have been proposed for pose evaluations, such as interaction-based accuracy classification (IBAC) [74], real space R-factor (RSR) [75] and Generally Applicable Replacement for RMSD (GARD) [76].

2.3 Practical aspects in molecular docking

Fig. 5 shows the key steps in molecular docking that are common to all protocols. A molecular docking calculation needs 3D structures of the input molecules, both protein and ligand(s). The structure of the protein target is usually determined by experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) or cryogenic electron microscopy (cryo-EM), and can frequently be downloaded from the Protein Data Bank (PDB) (https://www.rcsb.org/) [77]. Other sources of crystallographic structure of proteins and protein-ligands complexes are listed in Table 1.

There are several aspects that should be considered when assessing the quality of the protein target [78].

First is the crystal structure resolution, which is a measure of the degree of measurable diffraction observed in the crystallographic experiment using the protein or nucleic acid. This measure is conveniently represented in a form that expresses the level of details that could be observable when the electron density map is calculated. The crystal resolution depends on the degree of

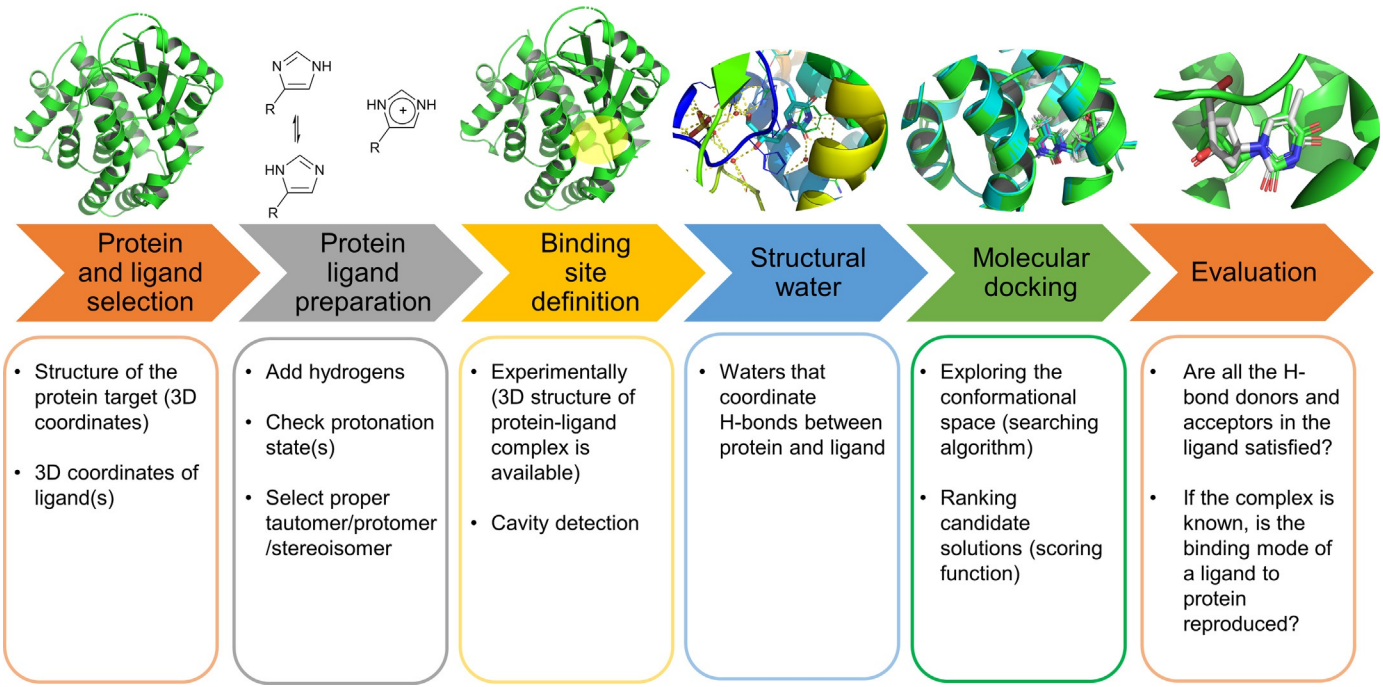

Fig. 5 A typical docking workflow. This flowchart shows the key steps common to all docking protocols. The 3D structures of the target macromolecule and the small molecule must first be chosen, and then each structure must be prepared in accordance with the requirements of the docking method being used. The binding site should be defined using computational tools or using experimental information. Active or structural water should be included as well. Following the docking, the results must be analysed, selecting the binding modes with the best scores, and evaluated.

Table 1 Sources of crystallographic structure of proteins and protein-ligands complexes.

wwPDB: worldwide Protein Data Bank https://www.wwpdb.org/ www.rcsb.org www.pdbe.org www.pdbj.org www.bmrb.wisc.edu	PDB is the central archive of all experimentally determined protein structure data. Today the PDB is maintained by an international consortium known as wwPDB. wwPDB has four members: • Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB) • Protein Data Bank in Europe (PDBe) • Protein Data Bank Japan (PDBj) • Biological Magnetic Resonance Data Bank (BMRB) Among these, rcsb acts as archive keeper to ensure that there is only one version of the data which is identical for all users
BindingDB https://www.bindingdb.org/bind/index.jsp	Rich repository of structural and thermocalorimetric data about ligand-protein interactions, including standard free energy, enthalpy and entropy changes upon binding, $\Delta G°$, $\Delta H°$, and $-T\Delta S°$ for a small but growing number of complexes
BindingMOAD—Mother Of All Databases https://bindingmoad.org/	Subset of the PDB containing every high-quality example of protein crystal with clearly identified biologically relevant ligands annotated with experimentally determined binding data extracted from literature. Ligands may be a peptide of 10 amino acids or less; oligonucleotide of 4 nucleotides or less; small organic molecule, and cofactors. Small molecules like crystallographic additives, salts, metals or solvent are not considered as ligands
PDBbind http://www.pdbbind-cn.org/	Comprehensive collection of measured binding affinity data (Kd, Ki, and IC50) exclusively for the protein-ligand complexes available in the PDB. It thus provides an essential linkage between energetic and structural information of these complexes, which is helpful for various computational and statistical studies on molecular recognition occurred in biological systems

Table 1 Sources of crystallographic structure of proteins and protein-ligands complexes.—cont'd

ModBase: Database of Comparative Protein Structure Models https://modbase.compbio.ucsf.edu/	Database of theoretically calculated protein structure models. The models are derived from an automated modelling pipeline relying on PSI-BLAST and MODELLER. In addition to the protein structure models ModBase contains information about putative ligand binding sites, SNP annotation and protein-protein interactions
PDB-REDO databank https://pdb-redo.eu/	Database of optimised existing PDB entries with electron density maps, a description of model changes, and a wealth of model validation data. It is a good starting point for any structural biology project. All the entries are treated with a consistent protocol that reduces the effects of differences in age, software, and depositors
EBI: The European Bioinformatics Institute https://www.ebi.ac.uk/services/structures	EBI provide access to several data resources including: • EMDB, a public repository for electron microscopy (EM) density maps of macromolecular complexes and subcellular structures • EMPIAR-Electron Microscopy Public Image Archive, a public resource for raw, 2D electron microscopy images. It allows users to upload, and download and reprocess the thousands of raw, 2D images used to build a 3D structure • PDBe, the European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes)

order of the crystal and the intensity and coherence of the diffraction beam used in the experiment. If a crystal is highly ordered and atoms are in well-defined positions throughout the crystal and remain so over time, then they will scatter X-rays the same way, and the diffraction pattern will show the fine details of the crystal.

In contrast, if atoms move over time or the content of one unit cell differs from the other, the diffraction pattern will not contain as much fine information. The highest resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to see every atom in the electron density map. Lower resolution structures, with resolution of 3 Å or higher, show only the basic contours of the protein chain, and the atomic structure must be inferred (Fig. 6). As a rule of thumb, we have more confidence in the location of atoms in structures with resolution values that are lower than 2 Å, called 'high-resolution structures'.

Although the resolution of the data provides information about the theoretical limits on the precision of the model, it does not say anything about the quality or the completeness of the data. Any missing data leads to a deterioration of the model parameters in the same way as reduced resolution does. Therefore, if two data sets are collected at the same resolution, the one with the lower completeness has the poorer data set and will more likely result in a less precise model.

Moreover, the experimental electron density data can be used to calculate several values such as the *R*-factor and R_{free} that can be used to address the quality of the data instead [79].

R-factor is a measure of the difference between measured data and data predicted from the model. A totally random set of atoms will give an

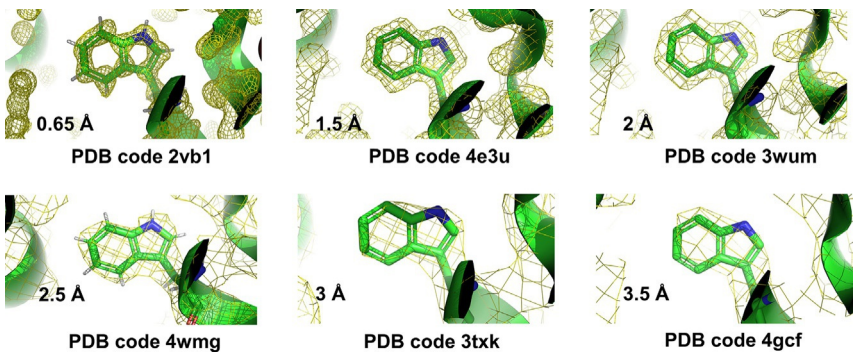

Fig. 6 Electron density maps for structures with a range of resolutions. The three structures on the top panel show tryptophan 28 from Lysozyme C, from PDB entries 2vb1 (0.65 Å resolution), 4e3u (1.5 Å resolution), and 3wum (2 Å resolution), respectively. The three structures on the bottom panel show tryptophan 28 from Lysozyme C, from PDB entries 4wmg (2.5 Å), 3txk (3.0 Å resolution) and 4gcf (3.5 Å resolution), respectively. The protein backbone is displayed as a ribbon cartoon and coloured in green, the tryptophan residue is rendered as green sticks and the electron density map surrounding regions of high electron density is coloured in yellow mesh.

R-factor of about 0.63, whereas a perfect fit would have a value of zero. A small difference indicates a more consistent model and therefore if two structures have similar resolutions the one with lower *R*-factor is the model that best fits the experimental data.

One potential problem with using *R*-factor to assess the quality of a structure is that it reflects the fit of the calculated and observed reflection data, the very thing that the refinement process uses to improve the atomic model. In effect, using the *R*-factor means we are evaluating the refinement model against the data used to train that model and so its value can be artificially low.

A less biased approach is to remove some of the experimental data (usually about 10%) before starting the refinement process, in this way only the remaining data (90%) are used for the refinement while the 10% is used for cross-validation during the refinement process to avoid over-fitting to the data. The R_{free} value represents how well the model predicts the 10% of data that were not used in refinement. However, over-fitting may still occur because there are usually insufficient data to uniquely determine all atom coordinates. For an ideal model that is not over-fitting the data, the R_{free} will be similar to the *R*-factor with a value of about 0.26 [80]. Another aspect to consider when evaluating the quality of a protein structure is the Debye Waller factors (DWF), or B-factors which are used to describe the relative vibrational motion of different parts of the protein. Atoms with low B-factors belong to well-ordered regions of the protein target while atoms with large B-factors belong to regions of the protein that are very flexible. It is important to ensure that atoms included in the binding site of the protein structure have a low B-factor, as high values imply that their coordinates are less reliable, and the docking experiment could then be affected. Finally, atomic occupancies should be considered. In some structures, regions are so labile that it is impossible to identify atomic positions at all. If prior to refinement a crystallographer knows that a given residue is of a given type from sequence information, they may choose to include a set of model coordinates for it with occupancies set to zero. These atoms will not contribute to the fit and so are, in effect, being placed by the refinement program in an arbitrary single position that may not be accurate.

When the 3D structure of the target protein is not available, protein structure prediction techniques such as homology modelling and protein threading are commonly used to obtain a 3D model of the target protein [81]. Homology modelling or comparative modelling relies on the correlation between the sequence of the target protein and its homologous protein

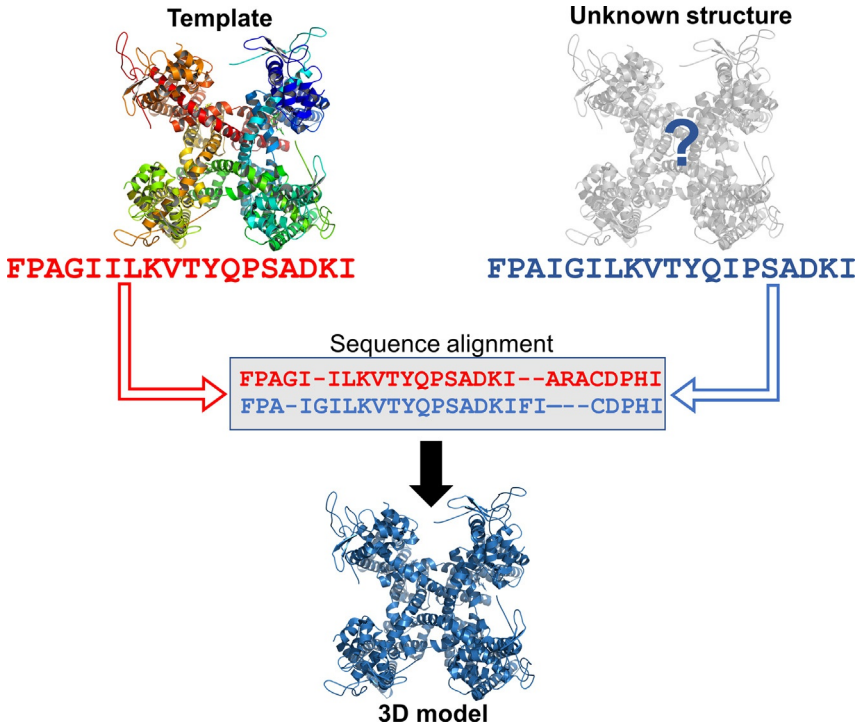

Fig. 7 Outline of the homology modelling process. Given the sequence of a protein with unknown structure, the first step is the identification of a related protein with known 3D structure that serves as a template. An alignment of the target and template sequences is necessary to assign the correspondence between target and template residues. A model is then built for the target based on the alignment and structure of the template, and further refined and validated.

structures (template) available in the PDB (Fig. 7). Such approaches can also be used to generate a variety of receptor conformations using either single-template or multiple-template structures enhancing the understanding of selectivity. Homology modelling can produce high-quality structural models when the protein target and template are closely related. Above 50% sequence identity, structural models tend to be reliable, with only minor errors in side chain packing. In the 30–50% identity range, errors can be more severe and are often located in flexible regions of the proteins (i.e. loop regions). For structures with lower sequence identities (<30%) homology modelling can be extremely difficult and other modelling techniques such as protein threading are then recommended. In protein threading or fold recognition techniques, the protein modelling is not based on the sequence

similarities, but on structural fit of the target sequence against different known folds. Protein threading is based on two basic observations. Firstly, the number of different folds in nature is quite small (approximately 1000). Secondly, 90% of the new crystallographic structures added to the PDB have similar structural folds to ones already deposited [82].

2.3.1 Protein preparation

Once the 3D structure of the protein target has been obtained (either downloaded from the PDB or generated using protein prediction methods), there are several protein and ligand preparation steps that should be followed before starting a docking run (Fig. 5). Here we will discuss the protein preparation procedure while the procedure for ligand preparation will be covered later in the chapter (see Section 2.3.2).

Due to insufficient resolution, most of the entries in the PDB only contain coordinates of non-hydrogen atoms. To work with these entries, the most common protein preparation task is the placement of the missing hydrogen atoms. This is not trivial, as it should account for the important ambiguities of protein structures, such as rotatable hydrogens, tautomers and protonation states of particular amino acids, alternative water orientations, and terminal side chain flips.

In addition, during protein preparation it is important to ensure that missing side chains are added, missing bonds are detected and fixed, bond orders are assigned, and where alternate locations are present, the atoms with highest frequencies are selected.

Other, more complex, procedures in protein preparation include prediction of protonation states and identification of which water molecules (if any) should be retained in the protein target structure.

The following subsections will look at the methods used to predict the protonation/tautomer states and at how to identify structural water molecules that are known to be vital in mediating hydrogen-bonding interactions, even in some cases key for facilitating tight binding, and hence should be considered part of the protein target structure.

2.3.1.1 Protonation state

Assigning the aqueous protonation state to protein residues is a key preprocessing step when working with crystal structures of protein-ligand complexes, and plays an important role in the prediction of the correct binding mode or binding affinity [64,83,84] of a ligand. This is even more relevant in

virtual screening where an incorrect predicted binding mode could lead to the identification of false positives or miss potential true binders.

It is worth mentioning that not all the scoring functions are equally affected by an incorrect protonation; in general, the physics-based scoring functions are likely to be more susceptible in comparison to knowledge-based or empirical scoring functions [85].

Protonation states of ionisable amino acids residues are constantly changing due to the dynamic nature of a protein. This effect is particularly relevant in protein-ligand complexes, where the ionisation states of residues in the binding site could vary, affecting the protein-ligand interactions. To accurately predict the conformation of a ligand in a protein binding site, the protonation state of the protein must be relevant to the bound conformation and in accordance with the pH of the experimental conditions [85].

Inspection of crystal structures and known, experimentally identified active ligands can yield a wealth of knowledge on the protonation state, steric clashes, and hydrogen-bonding networks between ligand and receptor [86]. Assigning protonation states to aspartic acid (Asp), glutamic acid (Glu), arginine (Arg) and lysine (Lys) during the protein preparation is generally straightforward, with deprotonated acids (Asp and Glu) and protonated bases (Arg and Lys) [86]. Histidine (His), however, provides a unique challenge in terms of protonation, as it can be protonated in three different ways. The imidazole ring of the His side chain can be protonated in a neutral form at either the ε-nitrogen or the δ-nitrogen, or in a charged (+1) form where both the ε- and δ-nitrogens are protonated. To further complicate choosing the correct form of the imidazole side chain ring, ambiguities in crystal structures, due usually to poor resolution, often switch the carbon and nitrogen, creating an additional three rotameric conformations, termed 'flipped' [86,87]. In addition, His is a weak base ($pK_a \sim 6.0$) and its protonation state is highly affected by the surrounding environment. To determine the correct protonation state of histidine residues, it is good practice to look at each His in the binding site individually; analysis of hydrogen-bonding networks is likely to yield the most detail about the correct side chain protonation [88]. Glutamine (Glu) and asparagine (Asp) too can be problematic: in poorly resolved crystal structures, the side chain terminal amides can, on occasion, be flipped so that the oxygen and NH_2 groups are misplaced.

Most of the available docking tools provide a set of protocols to prepare the protein target however, because most of these are automatic, it is always good practice to check particularly challenging residues like histidine, asparagine and glutamine.

2.3.1.2 Binding site definition and cavity detection

When targeting a protein, one wants to bind potential small molecule therapeutics to locations that induce a therapeutic effect (Fig. 5). A protein will typically have a key binding site where a given substrate will bind and will possibly have other sites of an allosteric nature. Binding to these sites is desirable, but first such sites need to be identified. They are usually determined experimentally, so that the structure of the protein-ligand complex is available, however, the process of identification of the active site in proteins becomes critical when the bound ligand is absent in the crystal structure.

With protein-ligand docking, a user typically provides guidance to the software to ensure that the predictions generated are placed in the correct regions of the protein. (We note that there is active research into 'blind' docking where no cavity information is assumed up front, which will be discussed later in the chapter.)

To identify likely binding regions, users will typically use cavity detection methods prior to docking. In some cases, the binding sites generated will be further validated experimentally: one can use single point site mutagenesis, for example, to identify whether given residues are essential for binding of known substrates. If a given residue is indeed critical, one can infer that the residue is likely to be in a therapeutically active region of a protein and so researchers will tend to examine the region containing said residue more closely with a view to designing new therapeutic agents.

Programmatic cavity detection is a well-researched field [89]. Many methods using a variety of different algorithmic approaches exist. For example, the classical algorithm LIGSITE [90] uses grid-based ray tracing to identify pockets and GHECOM [91] uses recursive sphere-based detection whereas newer methods, such as CavVis [92], use methods inspired from computer graphics [93]. The best methodology will be open to debate, however, for most docking algorithms the more critical component is identifying the set of atoms to be treated as active in docking.

If there are pre-existing known binding ligands, one can use them as a basis for the docking cavity. A common problem in the evaluation of docking is how to define the cavity in this case; if a cavity definition only encapsulates the close residues to a known binder, it is likely that the performance of the software will be over-represented when redocking the known binder; conversely if the cavity is very large, the search space for the algorithm is made far larger. Experience suggests that a reasonable compromise appears to be to include all the residues within 7 Å of the atoms of a known binder, though a user should visually inspect this cavity: in some cases one can have a larger cavity, particularly if one is starting from a fragment bound to a protein

structure. We would recommend surveying the protein using a cavity detection algorithm, even if there are structures of known binders.

2.3.1.3 Protein flexibility

As mentioned earlier in the chapter, due to the large computational resources required for fully sampling both protein and ligand conformations, the standard approach in docking and particularly in structure-based virtual screening is to dock fully flexible ligands in a rigid protein. Moreover, separating the contribution of protein effects on ligand conformation and simultaneous ligand effects on protein conformation, while developing the algorithms to handle such complex interactions, is still a significant challenge in molecular docking.

While conformational variability in different *apo* structures (i.e. without bound ligands) of the same protein suggests intrinsic disorder, variability in different ligand-bound states seems to be related to the presence of the ligand [94]. Ligands with diverse chemical scaffolds and of different sizes can induce different types, or different extents, of changes in protein conformation. Several studies have proven that the incorporation of protein flexibility in automated docking algorithms enables more accurate binding pose prediction and better virtual screening enrichments [95–97], in addition to providing a more realistic description of the physics of the protein-ligand binding interaction. The drawback is the required computational cost; docking with flexible ligands and a flexible protein usually requires the use of supercomputers to be achieved in a reasonable timeframe.

A statistical analysis of the PDB revealed that 85% of the proteins contain only one to three flexible residues in the active site and that single rigid protein dockings predict an incorrect binding pose for 50–70% of all ligands [94,95,98]. These flexible residues are subject to conformational changes ranging from simple side chain movements (rearrangement) to backbone-loop movements, to major domain rearrangements. Combinatorial approaches making use of side chain rotamer libraries and soft potentials are considered very efficient to account for small movement of the protein side chains, while the ensemble of rigid protein structures (ensemble docking) has been proven successful to account for larger changes in the protein [95,96,99,100].

2.3.1.3.1 Soft docking and side chain rotamer libraries Soft docking attempts to treat flexibility by allowing a small degree of overlap between the protein and the ligand. This implicitly models protein accommodation

by loosening the criterion for steric fit. This is achieved by reducing the steepness of the repulsion term in the Lennard-Jones potential function [101].

Soft docking has the advantage of being computationally efficient (only the scoring function parameters need to be changed), however, it can take care of only minor side chain movements. Applications of such an approach have been used for both protein-protein docking [102] and protein-ligand docking [103,104].

Side chain rotamer exploration offers an alternative way to model receptor flexibility and is similar in spirit to the approach of exploring multiple ligand conformations for simulating ligand flexibility. Side chain rotamer libraries are usually knowledge-based and generated from experimentally determined structures [105]. Such libraries also tabulate the probability with which different rotamers are observed, therefore in principle, rotamer libraries can also help to decrease conformational search space with the use of rotamers as representative conformations of energetically favourable states [106]. Side chain rotamer exploration can be applied to several amino acid side chains in the binding site. However, based on the conformational space to explore, a computational cost must be added to the docking run. Generally, it is good practice to limit the side chain rotamer library only to those residues that have a large conformational change when comparing unbound and bound forms of the protein structures. One should also note that adding protein flexibility introduces additional scoring terms, extending the necessary approximations made to the scoring functions further than in rigid protein docking.

2.3.1.3.2 Ensemble docking An ensemble docking methodology aims to address protein flexibility by docking flexible ligands against multiple conformations of the target protein rather than just the single rigid protein structure used in standard docking (Fig. 8).

The ensemble of protein conformations mimics the conformational equilibrium which characterises the native state of the target protein and provides a structural degree of freedom by which the conformation of the protein may be matched to fit any particular ligand [107]. Ensemble docking is a two-step process including generation of an ensemble of protein conformations and the actual docking to the selected protein structures.

The ensemble of protein structures can be generated by using available crystal structures of the same protein target that have been isolated and/or co-crystallised with various ligands. Such an approach accounts for induced fit changes occurring upon ligand binding.

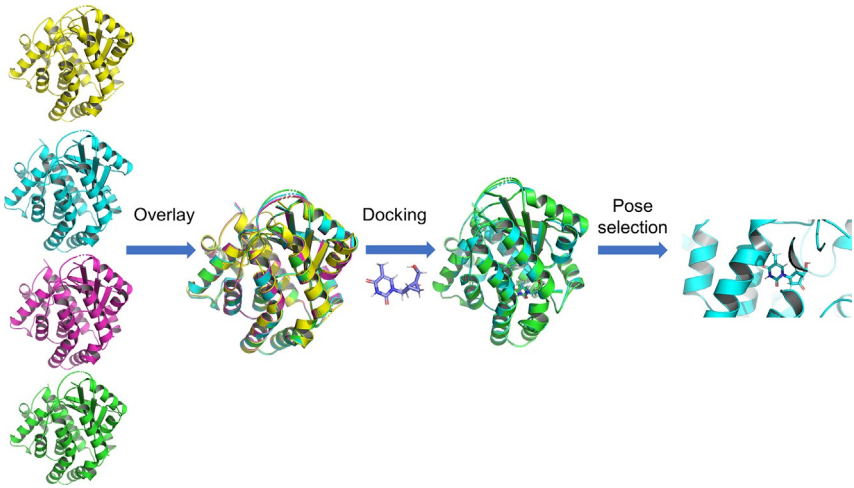

Fig. 8 Ensemble docking workflow for the TK_{HSV1} protein. TK_{HSV1} has been determined in both *apo* form (PDB entry 1e2h displayed in yellow cartoon), and in complex with different nucleoside prodrug ligands (PDB entries 1e2i, 1of1 and 4ivq shown as cyan, magenta, and green cartoon, respectively). Ensemble docking is a two-step process including generation of an ensemble of protein conformations (overlay) and the actual docking to the selected protein structures.

The protein ensemble can also be generated by molecular dynamics (MD) simulations, where different snapshots are isolated from single or multiple trajectories, thus helping to explore possible protein conformation changes across time [100].

To balance the computational efficacy and prediction accuracy of ensemble docking, a relatively small set of representative structures should be selected. Several studies have explored the best way to select experimental protein structures and the best number of structures for ensemble docking, finding in general that the use of many receptor conformations may not necessarily improve the docking performance because a large number of false positives may reduce the enrichment rate in virtual screening [97,107,108]. These studies tend to focus on specific systems however, to our knowledge, there is no general automated solution to the problem of optimal ensemble selection; a key part in ensemble docking [99,109].

2.3.1.4 Structural water molecules

The presence of structural or labile water molecules in the protein binding site plays an important role in protein-ligand binding and therefore can affect the accuracy of protein-ligand docking predictions (Fig. 4). Labile waters can either stabilise a protein-ligand complex by mediating hydrogen bonding

between the ligand and the protein, or they can be displaced by the binding ligand. They can also contribute significantly to entropic and enthalpic changes in the protein-ligand complex, where the entropic loss associated with transferring a water molecule from the bulk solvent to protein-ligand binding is compensated for by the enthalpic contribution of the additional hydrogen bonds.

The importance of water molecules is well recognised in structure-based drug design, where displacing, mimicking, and/or targeting of bound water molecules can improve the binding affinity of ligand molecules [110]. For this reason, understanding the role of active water molecules has drawn extensive attention in the past decade, and many tools and strategies have been developed to predict the locations and thermodynamic profile of water molecules in the binding site [111–114].

Most of the commercially-available docking programs are able to deal with active water molecules [115–117]. When water molecules are implicitly or explicitly included in a docking run, the docking algorithm will try to find the best docking pose with a binding site occupied by the water molecule. Among them, GOLD has the additional feature to treat active water molecules as switchable (i.e. the water can be bound or displaced by the ligand) and flexible (i.e. the water can rotate and translate within a given distance to optimise hydrogen bonding). The position of water molecules within an active site can be highly variable. Treating them as static in nature can bias towards ligands that complement the specific orientation and prejudice those that would physiologically replace the given water molecule, leading to an increase in false negatives [86,118]. In GOLD, a simple entropy penalty term has been introduced to account for the unfavourable loss of rotational and translational entropy that accompanies the tight binding of a water molecule to a protein surface; only water molecules with binding affinity that outweighs the loss of rigid-body entropy on binding are considered to be bound [115].

Several studies have proven that docking and virtual screening with active water molecules can improve ligand docking poses and greatly increase the ligand enrichment, however, particularly in a virtual screening context, the addition of water is frequently neglected as the additional computational cost required affects the rapid screening of a large library [119–121]. It is therefore important to determine which water molecules must be kept and exclude those water molecules that are not essential.

2.3.1.4.1 How to recognise active water? One approach to assess which water molecule(s) should be kept for docking is to attempt to replicate the binding mode of experimental ligand structures in the absence of explicit

waters. If the accuracy is diminished by the absence of waters in the binding site, then it is important to identify those water molecules which are crucial for binding [86].

In general, waters that are not hydrogen bonded to the protein, and those that are located outside the binding site, have little or no effect on ligand binding and can be removed [86,121]. However, waters that form hydrogen bonds with the receptor, or those with low B-factors, are likely to be highly stable within the protein binding site and should be included in docking studies as they are likely to stabilise the protein and they will be difficult to displace on ligand binding [110]. Waters that form hydrogen bond bridges between the ligand and protein are also likely to be important in ligand binding. However, a hydrogen bond network may be ligand-specific and its importance in virtual screening, where a diverse set of compounds are under study, should be thoughtfully assessed in advance. For this reason, when active water molecules are included in docking, and more importantly in virtual screening, they should, ideally be treated as flexible [121]. As an alternative, several computational methods have been developed to predict the locations of the water molecules in the protein binding site [122–125].

2.3.2 Ligand preparation

Ligand preparation consists of generation, optimisation, and validation of its 3D structure.

3D structures of ligands can be obtained experimentally, for example from protein-ligand co-crystal complexes, or they can be generated using software able to convert 1D and 2D structures (e.g. SMILES, SMARTS, InChi) into 3D molecular structures (Fig. 9).

The 3D structure of the ligand must have realistic bond lengths and angles as these will not usually change during docking. Optimisation of the starting ligand geometry is sometimes required for particularly complex molecules.

Several programs exist to generate and optimise the 3D structure of a ligand (e.g. CSD Conformer Generator [126,127], Omega [128,129], Confab [130], Confect [131], RDKit [132]). They differ in the algorithm used; some systems use force fields to infer intramolecular geometries, whereas others including the CSD Conformer Generator, rely directly on crystal structure data derived from the Cambridge Structural Database (CSD) [133] to produce realistic ensembles of high probability ligand structures.

As in the protein, hydrogens and formal charges must be added to the 3D structure of the ligand. The protonation state should be set according to the physiological pH or the pH of the simulation, and tautomeric states of the

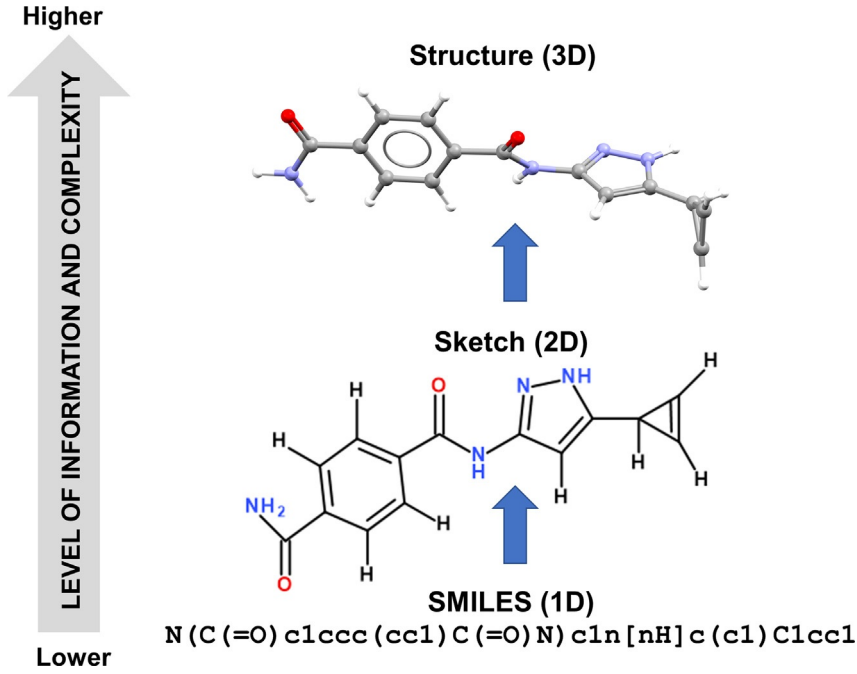

Fig. 9 1D to 3D workflow for a small molecule.

ligand should also be defined. In some cases, it may be worth generating multiple possible protonation states for a given ligand, with a view to docking all forms (in particular, if the pK_a of a given proton dissociation is close in numerical value to the physiological pH).

Tautomers are isomers differing only in the positions of hydrogen atoms and electrons, therefore even a simple molecule can have several different tautomeric forms. Moreover, the protonation and deprotonation of the ionisable sites in the molecule produces additional forms called protomers. Tautomers and protomers differ in shape, functional groups, surface, and hydrogen bonding. Therefore, tautomerism and protonation may result in alternative binding modes that can affect the efficiency of docking and virtual screening [64,84]. For chiral compounds, proper enumeration of the relevant stereoisomers is also necessary to effectively use docking and virtual screening as a drug discovery tool. Since enantiomers of a chiral compound might have different binding affinities for a given receptor, it is important that both enantiomers are used for docking. The docking software should be able to produce different docking scores for enantiomers that bind to the protein receptor. Failure to include appropriate diastereomers and/or

enantiomers of compounds could significantly increase the number of false negatives and adversely affect enrichments in virtual screening [134].

2.4 Small molecule databases

Several databases of lead and drug-like small molecules are available for virtual screening purposes. Although a substantial overlap is found among some of the collections, each database has unique features that may make it better than others for a particular virtual screening project. A relevant number of unique compounds is found within each database that makes it worth considering more than one library if possible. Many chemical databases are freely available and may be designed to possess desirable characteristics such as 'drug-likeness', dictated by the 'Lipinski's rule of five' which states that drug-like compounds should have molecular weight lower than 500, lipophilicity (logP) lower than 5, less than 5 hydrogen bond donors, and less than 10 hydrogen bond acceptors [135]. However, over the years, an increasing number of compounds violating some of these rules have been approved as drugs and entered the market (e.g. many natural product drugs as well as 50% of marketed drugs do not comply with the rule of five) [136]. A strict compliance to this rule can strongly limit the variety of chemotypes, indicating that it is advisable to follow it with a certain degree of flexibility [136]. Other available databases contain chemical structures from natural products or approved drugs therefore include compounds that break the rule of five.

Here we are going to provide a brief overview of some of the most common chemical databases that are used for virtual screening. A more exhaustive list is provided in Table 2.

2.4.1 ZINC database

ZINC is a free database of commercially-available compounds developed in the Department of Pharmaceutical Chemistry at the University of California, San Francisco [137]. It contains a constantly growing number of 3D structures ready-to-dock from catalogues of several vendors with annotated relevant information about protonation and tautomeric states, and properties such as size, calculated logP, number of rotatable bonds, etc. Each molecule in the database also contains purchasability and vendor information, making this ZINC's focus on docking and availability the main distinctive characteristic from other databases.

In its latest version, ZINC20 [138] comprises over 736 million lead-like compounds (molecular weight less than 400 g/mol, calculated logP less than 4 and rotatable bonds less than 7), 509 million of these compounds are available for download in 3D ready for docking, together with information

Table 2 List of available small molecule databases.

Database	
ZINC https://zinc.docking.org/	Free database of commercially-available and annotated compounds for virtual screening. ZINC contains over 736 million lead-like compounds of which 509 million are in ready-to-dock, 3D formats. ZINC also contains over 1.3 billion purchasable compounds which you can rapidly search for analogues
ENAMINE https://enamine.net/11-databases	The world's largest collections of novel building blocks (225,000+) and screening compound libraries (2,740,000+)
NCI Open Database https://cactus.nci.nih.gov/download/nci/	NCI database has more than 275,000 small molecules structures, a very useful resource for researchers working in the fields of cancer and AIDS
ChEMBL https://www.ebi.ac.uk/chembl/	ChEMBL provides comprehensive information about 1 million bioactive compounds (small drug-like molecules) with 8200 drug targets
DrugBank https://www.drugbank.com/	The database combines detailed drug data with comprehensive drug target information. It contains 6,712 drug entries including 1,448 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5,080 experimental drugs
ASINEX Database http://www.asinex.com/	The ASINEX database is a commercial collection of compounds which contains more than 600,000 screening compounds, 27,000 macrocycles, 20,000 fragments and over 22,000 building blocks
Cambridge Structural Database (CSD) https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/	The Cambridge Structural Database (CSD) is a repository for small molecule organic and metal-organic crystal structures. With over 1 million structures from X-ray and neutron diffraction analyses, the CSD includes several subsets, such as the CSD Drug Subset with entries that feature in the approved drug list provided by DrugBank and a CSD COVID-19 subset that includes structures of interest in the fight against COVID-19

Continued

Table 2 List of available small molecule databases.—cont'd

Database	
PubChem https://pubchem.ncbi.nlm.nih.gov/	Database of chemical molecules which maintains three types of information namely, substance, compound and BioAssays
SPECS Database https://www.specs.net/	SPECS database (>240,000 compounds) is composed of novel drug-like small molecules obtained from academia and research institutes. SPECS contains available compounds that can be purchased upon request. Every molecule in the collection must fulfil structural characteristics of a biologically active compound and meet ADMET requirements
MAYBRIDGE Database https:/maybridge.com/	Maybridge Screening Hit Discovery collection (over 53,000 compounds) is a commercial library of small hit-like and lead-like organic compounds that covers ~87% of the 400,000 theoretical drug pharmacophores complying with the rule of five and of good ADMET properties. Maybridge also offers a fragment library (30,000 fragments) and a hit-to-lead building block collection
LIFE CHEMICALS Database https://lifechemicals.com/	Contains a commercial compound collection for HTS of 1,213,000 lead-like and drug-like new diverse chemical entities that follow Lipinski's rules. Furthermore, it offers several different diversity libraries on demand: i.e. building blocks, fragment- and scaffold-based libraries, natural product-like compounds, covalent inhibitors, etc.
CHEMBRIDGE Database https://www.chembridge.com/	Contains 1 million drug-like and lead-like compounds in 2 non-overlapping collections of respectively 460,000 and 620,000 compounds, that cover different chemical spaces and that can be customised to create diversity libraries, targeted libraries (KINASet, CNS-Set, and IONSet libraries) and fragment libraries, which can be purchased upon request

regarding target and biological activity, related scaffolds and bioactive and biogenic compounds (with a Tanimoto likeness index of 0.6). ZINC also offers other features such as the possibility to define target-focused libraries and to download subsets of a physical property space (fragment-like, lead-like and drug-like subsets) [139].

2.4.2 ENAMINE database

ENAMINE provides several different commercial collections of compounds for screening. The screening collection currently contains over 2 million low molecular weight organic compounds. The HTS collection (>2,115,000 compounds) represents a highly diverse set of chemotypes developed from in-house research and partner academic organisations, while the Advanced Collection (>493,000 compounds) is intended for lead discovery. This library has been designed according to lead-like properties and/or valuable pharmacophores such as carboxylic, primary amino and amide groups. The Premium collection (>44,500 compounds), instead, contains compounds with the most favourable physicochemical properties (high Fsp^3, low logP and MW).

Enamine also provides a pharmacologically diverse set (10,240 drug-like compounds clustered by activities from biologically relevant chemical space), a 3D diversity set (50,240 compounds from conformational analysis and shape clustering of the HTS collection) and a covalent screening library (10,480 compounds of well validated covalent binders) as well as targeted libraries (e.g. central nervous system (CNS), antibacterial, ion channel, coronavirus, kinase and lipid G protein-coupled receptor (GPCR) libraries) and fragment libraries (e.g. covalent, sp3-rich, Protein-Protein Interactions (PPI), fluorinated and brominated fragments).

The largest Enamine database is called the *REAL* database and is a virtual collection of over 1.36 billion molecules that can be used to find new hit molecules using large-scale virtual screening and to search for analogues of hit compounds. Each molecule in the *REAL* database complies with the rule of five and also the Veber criteria (rotatable bonds ≤ 10, and TPSA $\leq$ 140) [140].

The structure data files of these various collections are regularly updated and can either be directly downloaded from the ENAMINE webpage in MDL SD (.sdf) or MDL ISIS (.db) formats or obtained by request. Along with SMILES and catalogue IDs, the entry for each molecule in the database lists important physicochemical parameters (MW, sLogP, HBA, HBD, etc.), structural alerts (PAINS, Brenk, and Eli Lilly medchem rules), type of chemistry and the difficulty of synthesis ('s', simple chemistry, standard effort, 'm', advanced chemistry, higher effort).

2.4.3 NCI open database

The NCI Open Database is a freely accessible database developed by the Developmental Therapeutics Program of the National Cancer Institute

(NCI). This database contains a set of compounds that have been collected by the NCI, since 1955 for testing in anticancer, and from the 1980s for anti-AIDS screens, that are not covered by confidentiality agreements. The database currently contains more than 260,000 molecules both from organic synthesis and natural source extracts that can be downloaded in .sdf format [141]. The compounds in the database are annotated with information fields including release, structure source and evaluation, calculated/predicted logP, biological activity and commercial availability in addition to 3D atom coordinates, added hydrogens, number of rotatable bonds, stereocentres and bond stereocentres [141].

2.4.4 ChEMBL

ChEMBL is a manually curated database of bioactive molecules with drug-like properties [71,72]. The ChEMBL database contains more than 1.8 million compounds and over 15 million records of their effects on biological systems. It contains information about how small molecules interact with their targets, how the compounds affect cells and the whole organism, and information on absorption, distribution, metabolism, excretion and toxicity (ADMET). Additional data on clinical progress of compounds has been integrated into ChEMBL (ChEMBL Drugs). This highly curated dataset includes marketed compounds and compounds that are or have previously been in clinical development and are annotated with information about their known therapeutic targets and associated indications.

The data in ChEMBL are extracted and curated from the primary medicinal chemistry and pharmacology literature and cover a significant fraction of the SAR and discovery of modern drugs. Additionally, the ChEMBL database contains data deposited by researchers and data extracted from other public databases.

ChEMBL includes 2D structures with calculated molecular properties (e.g. logP, molecular weight, Lipinski parameters) and bioactivity data (such as binding constants, pharmacology and ADMET) with the bioactivity data tagged to show links between molecular targets and published assays.

2.4.5 DrugBank

DrugBank is a web-enabled curated database containing comprehensive molecular information about FDA-approved drugs as well as experimental drugs going through FDA approval [142]. As both a bioinformatics and a cheminformatics resource, DrugBank combines detailed drug data (i.e. chemical, pharmacological and pharmaceutical) with comprehensive drug

target (i.e. sequence, structure, and pathway) information. All data in DrugBank is non-proprietary or is derived from a non-proprietary source. It is freely accessible and available to anyone and nearly all data are fully traceable and explicitly referenced to their original source.

To date, DrugBank contains over 13,700 drug entries including approved small molecule drugs, approved biologics (proteins, peptides, vaccines, and allergenics), nutraceuticals and over 6,000 experimental (discovery-phase) drugs. Additionally, over 5,000 non-redundant protein (i.e. drug target, carrier, transporter and enzyme) sequences are linked to the drug entries together with drug-drug and drug-food interactions [143]. All the chemical structures in DrugBank are accessible in canonical SMILES, sdf, .mol, .pdb, InChI and InChIKey formats.

2.4.6 ASINEX database

The ASINEX database is a commercial collection of compounds which contains more than 600,000 screening compounds, 27,000 macrocycles, 20,000 fragments and over 22,000 building blocks [144]. The screening compounds are organised in different libraries that cover different chemical characteristics and try to address different steps in the drug discovery process. The Gold & Platinum Collections of over 260,000 compounds includes diverse and cost-effective coverage of drug-like chemical space. Most compounds have a high degree of drug-likeness, in accordance with Lipinski's rule of five. Other libraries are focused on lead-like compounds are intended for the early stages of drug discovery. For example the ASINEX Synergy and Elite Library of more than 91,000 compounds has been screened against a panel of early ADMET tests to make sure screening hits do not have potential ADMET problems and are amenable for rapid hit-to-lead optimisation.

A subset of over 170,000 compounds (BioDesign) incorporates key structural features of known pharmacologically relevant natural products (e.g. alkaloids and other secondary metabolites) into synthetically feasible medicinal chemistry scaffolds. ASINEX also offers targeted libraries including those focused on CNS disorders, immuno-oncology, PPI, GPCRs, peptide-mimetics, nucleoside-mimetics, etc. All the libraries can be downloaded in SDF format and can be directly purchased.

2.4.7 Cambridge structural database (CSD)

The Cambridge Structural Database (CSD) is the world's leading repository for small molecule organic and metal-organic crystal structures. It contains over 1 million structures from X-ray and neutron diffraction analyses

forming a unique database of accurate 3D structures [133,145,146]. Every entry is manually curated and each structure in the CSD is enriched with chemical representations, as well as bibliographic, chemical, and physical property information.

The CSD comes with several subsets including CSD entries that feature in the approved drug list provided by DrugBank. This subset has a wide scope; it contains any solvates, co-crystals or hydrated forms, and currently provides a set of 12,277 entries to help users gather insights into drug-like compound. A single-component CSD drug subset is also available and includes 1,989 CSD entries where a drug molecule is the only modelled component in the crystal structure. Recent additions to the CSD subsets are: the CSD COVID-19 subset with 121 structures of interest in the fight against COVID-19 and the CSD Pesticides subset with 972 entries.

2.4.8 PubChem

PubChem is an open chemistry database managed by the National Institutes of Health (NHI). It contains mostly small molecules, but also larger molecules such as lipids, carbohydrates, nucleotides, peptides and other chemically modified macromolecules. The data in PubChem are organised into three interlinked databases: Substance (as of writing more than 286 million substance descriptions), Compound (over 111 million unique chemical structures) and BioAssay (1.2 million biological assays covering more than 10,000 target protein sequences).

Most of the structures in PubChem are drug-like compounds that satisfy Lipinski's rule of five. Among them more than 10 million are fragment-like compounds which satisfy Congreve's rule of three (molecular weight of a fragment is <300, the cLogP is ≤ 3, the number of hydrogen bond donors is ≤ 3 and the number of hydrogen bond acceptors is ≤ 3) [147]. In addition to bioactivity data, PubChem contains compound information that can be useful for virtual screening. Because of the data integration with DrugBank, PubChem includes comprehensive information on FDA-approved and investigational drugs, including their drug indications, mechanisms of action, target macromolecules, interactions with proteins and genes, ADMET properties and many others. In addition, PubChem provides links to crystal structures available from the CSD. PubChem also offers links between about 6 million patent documents and more than 16 million unique chemical structures, with over 336 million chemical substance-patent links covering the USA.

3. Fragment-based screening

Fragment-based screening aims to identify small chemical fragments which bind weakly to the binding site of a target protein. Although weak, given the limited numbers of interactions those fragments can make with the protein, these are high-quality interactions as they must overcome a substantial entropic barrier to binding [148]. This approach is useful for identifying interaction hot spots for the protein-ligand binding. The key principle in fragment screening is the ability to sample the chemical space more efficiently than when using larger complex molecules. Thus, in theory, this allows sampling of a broader and more diverse chemical space than in currently practical to achieve via standard screening methods. Moreover, smaller fragments are less likely to contain interfering moieties that could block an otherwise favourable ligand–protein interaction, thus optimal binding spots are less likely to be hidden by non-binding elements (Fig. 10). In principle it is possible to use molecular docking for virtual fragment-based screening, however, by definition, fragments are relatively small compared to lead-like or drug-like compounds, which makes molecular docking more challenging. Fragments only form few, key interactions with the binding site, and this usually results in a low docking score, with the risk of missing potential fragment hits if they have weak interactions with the protein. Free energy differences between different binding modes of a fragment are much smaller than those of larger compounds therefore, given the inaccuracies inherent in current scoring functions, they make it more difficult to distinguish the correct/incorrect binding mode of a fragment.

Additionally, due to the relatively small size, docking poses for fragments can be promiscuous, with fragments binding to multiple sub-regions showing similar physicochemical properties inside a binding pocket. In these cases docking results are hard to interpret and post-processing the results can be time-consuming [150].

As for a standard docking screening, to perform a fragment screening one would need access to the 3D structure of the target protein and to the fragment library to be screened. The protein target could be obtained from an experimental or theoretical model and should be prepared as discussed in Section 2.3.1.

Fragment libraries tend to be small because the fragment space is smaller than the chemical space and can be more effectively probed with a relatively

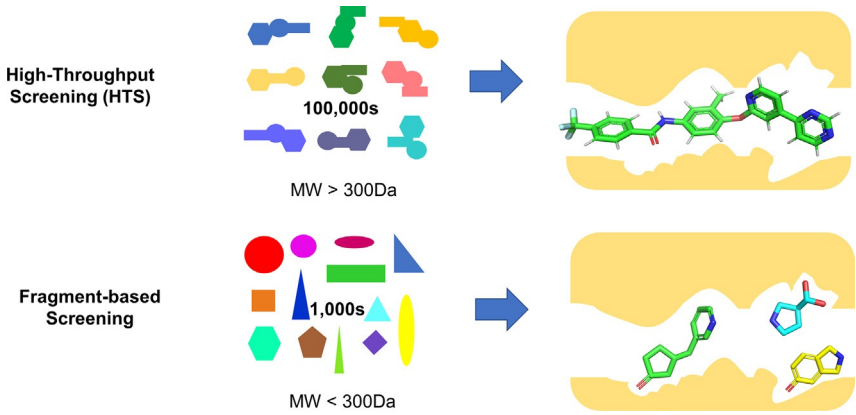

Fig. 10 High-throughput screening (HTS) and fragment-based screening overview. HTS enables testing of a large number of diverse drug-like and lead-like compounds (MW > 300 Da) against protein targets. In order to efficiently cover all the chemical space, the database should contain more than 100,000 compounds. In a fragment screening approach, a database of simple molecules or fragments with MW <300 Da are screened against targets. Due to the small size of fragments, they often have low potency, but high-quality interactions and can be further optimised into potent leads by linkage. Fragment libraries usually are of a small size (~1,000 s) because the fragment space is smaller than chemical space and can be more effectively probed with a relatively small library [149]. *Adapted from https://figshare.com/articles/SSIEM_presentation_Fragment_screening_for_drug_discovery_in_primary_hyperoxaluria_type_1/7057826.*

small library. Screening a few thousand fragments would search a greater fraction of the chemical space than could be achieved by screening over a million full-sized compounds (Fig. 10) [149,151].

Fragment libraries should be chemically diverse, synthetically expandable and should represent a wide range of physicochemical properties, aqueous solubility, molecular diversity, and drug-likeness with medicinal chemistry scaffolds. They contain fragments that comply with the rule of three [147]. Although the validity of this rule is a debatable topic [152], it still remains the preferred reference for fragment selection and fragment library generation. A growing number of commercial companies are now offering well-defined fragment libraries for screening [153]. For instance, the Diamond-SGC Poised Library (DSPL) [154] is a library containing about 760 fragments with at least one functional group that is open to rapid follow-up synthesis whilst maximising chemical diversity. While the majority of fragment libraries are chemically diverse, they still lack shape diversity. To overcome this limitation, the UK 3D Fragments Consortium is working on generating a library focused on fragments that incorporate 3D structures [155].

Once the fragment identification step is completed, these fragments can be grown, linked, or merged to develop the potential lead compounds. Fragment growing is the most used strategy in fragment-based lead design. As the name suggests, the core fragment is modified to increase its size to improve its properties and affinity for the target [150,156]. Fragment growing strategies have been successfully applied to various targets [150,157–160] including Alzheimer's Disease target BACE1 [161] and matrix metalloproteinases [162].

Fragment merging can be used in cases where two distinct fragments partially occupy the same region, or when two binding sites have regions in common and therefore their ligands are partially competitive with respect to the site. In such cases the overlapping parts can be fused into a single molecule [156]. Fragment merging is not only used in novel compound design but also remains a tool for chemical modification and derivative generation [150]. While less common than the growing methodology, a few successful examples of fragment merging have been discussed in the literature with targets including Hsp90 [163], PIγ kinase [164] and mycobacterial transcriptional repressor EthR [165].

Fragment linking describes the process of joining two non-competitive fragments (i.e. fragments that bind in two different sub-pockets of the binding site) with a chemical linker or spacer. Suitable linkers should respect the original conformational constraints of the two initial fragments while making favourable interactions within the protein binding site. Linking strategies have been successfully applied to different targets [160,166,167] including protein kinase ck2 [168] and thrombin [169].

4. Protein-protein docking

Protein-protein docking is an emerging research topic, due to its potential for predicting protein-protein interactions (PPIs) and identifying hot spot residues at the protein-protein interface.

Although protein-protein docking shares the principles of protein-small molecule docking, sampling of the conformational space in protein-protein docking is extremely challenging. Even for relatively rigid proteins, it is difficult to explore the rotational-conformational space of mutual orientations potentially sampled by a pair of proteins as they interact. Due to the huge number of degrees of freedom the computational cost of the search algorithms is considerable for protein-protein docking.

Because of the large size of binding sites in PPIs, the orientational search algorithm often requires strategies for protein-protein docking that aredifferent from those for protein-ligand docking. Additionally, contact surfaces where

two proteins interact with each other are significantly different from ligand-protein binding cavities. Protein-protein binding sites are often relatively flat interfaces with no single large and well-defined pocket [170]. Predicting the association of proteins is further complicated by their flexibility. Proteins are dynamic; they constantly interconvert between conformers of varying energies and capturing this flexibility is still a challenge in molecular docking.

Despite the complexity of the problem, a variety of docking methods is currently available for predicting the structures of protein-protein complexes. The choice of the method to use depends on the nature of the docking problem [171]. Rigid docking procedures are generally used when the crystallographic structures of the proteins to dock (or of their very close homologues) and their complex are available. Knowledge of the protein-protein complex makes the prediction of related protein complexes feasible for template-based and homology modelling methods, even when the structures of component proteins are not available. The general idea is to separate the two proteins from the complex structure and use a rigid docking procedure to try to reproduce a near-native approximation of the complex (bound docking) as shown in Fig. 11.

Bound docking is the easiest docking case, because, by definition, it does not involve conformational change and existing rigid-body procedures

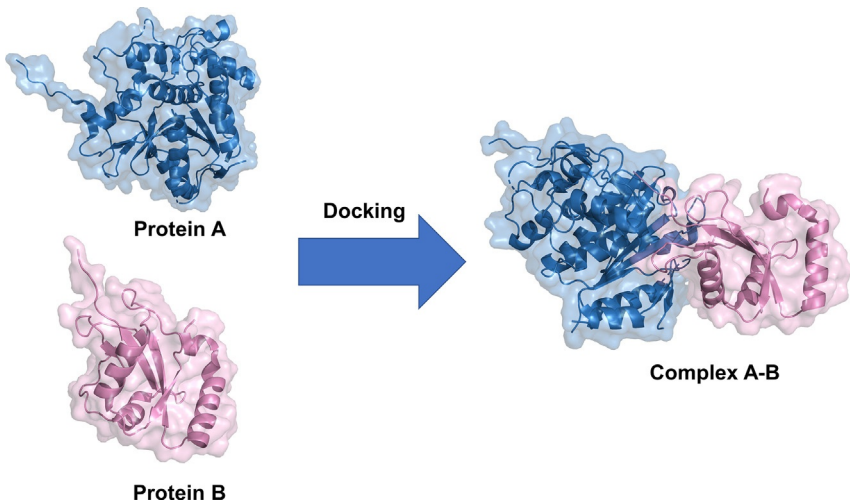

Fig. 11 Example of bound docking procedure. Protein A, in cartoon with the surface coloured in pink, represents the Human Atg4B (HsAtg4B; a mammalian orthologue of yeast Atg4) and protein B, in cartoon with the surface coloured in blue represents the LC3 protein (a mammalian orthologue of yeast Atg8). The complex A–B is HsAtg4B–LC3 complex (PDB code 2z0d).

generally result in good (i.e. near-native) structures of the complex among the top pose predictions.

When the crystallographic structure of the complex is not available, the docking algorithm should be able to predict the complex from the separately determined protein structures (unbound docking). In such cases, the docking algorithm has to deal with the conformational difference between the unbound and the bound structures [172] thus, the landscape of docking solutions could also include many false positives with good surface complementarity that differ considerably from the native complex.

In the case of unbound docking, one approach is to combine a rigid-body search with flexible rescoring and refinement of the docked conformations [173,174]. Other strategies include rigid-body docking of ensembles of structures [175,176], scoring functions that feature soft potentials to allow minor molecular overlaps [177,178], simplified protein representations (coarse-grained), where at such low levels of structural resolution the difference between unbound and bound conformations is less significant [179], and relaxing the interface of docking poses using techniques such as molecular dynamics (MD), Monte Carlo, or simulated annealing [177,179,180]. Additional docking strategies include scoring functions that use a combination of terms to describe physical interactions and penalise models that do not recapitulate the available experimental data. This method allows side chain flexibility at the PPI interface, which is a key aspect to improve the accuracy of the results. Examples of such approaches are HADDOCK [181] and IMP [182].

Protein-protein docking methods have improved substantially over the past few years (as demonstrated by the results of the Critical Assessment of PRedicted Interactions (CAPRI) [183]) but there are still many scenarios that are particularly challenging and pose limitations to current docking approaches. Examples of unsolved challenges are: large movements upon binding; weak or transient binding; and unavailability of 3D structure for one or both subunits. The protein-protein docking field needs to improve not only in optimising and/or developing new docking methods to overcome these challenges but also in identifying those problematic cases and evaluating the reliability of the predictions.

5. Protein-peptide docking

Peptide drugs are gaining attention in drug discovery as a solution for targeting 'undruggable' intracellular protein-protein interfaces characterised by large and relatively featureless interfaces. Peptides can bind large protein

interfaces with high potency and great selectivity, which translates into fewer off-target side effects and less potential for toxicity than small molecule drugs [184].

Computational methods, including docking, have proven to be successful in the discovery and design of small molecule drugs as well as in the field of peptide therapeutics.

Protein-peptide docking involves computation steps such as conformational sampling, structural refinement and scoring, similar to traditional protein-small molecule docking techniques. However, peptides are more flexible than small molecules and tend to adopt numerous conformations. Thus, modelling protein-peptide interactions is a challenging and time-consuming task.

Over the past decades a wide range of docking methods has been developed that can be used directly or indirectly for docking peptides on a protein with various degree of success. These methods are described in Fig. 12 and include protein-small molecule docking, protein-protein docking and protein-peptide docking. Due to the limitation of the search algorithm, protein-small molecule docking approaches are currently limited to short peptides (up to 15 amino acids) with a well-structured conformation. However, modelling long peptides can be overcome by docking peptide fragments followed by their merging [185].

Protein-protein docking methods, such as ZDOCK [186], and Hex [187] have also been used to dock peptides onto a protein. However, compared to proteins, peptide molecules are much more flexible and less stable. Protein partners usually have well-defined 3D structures before forming protein-protein complexes while peptides usually do not. Furthermore, peptide-mediated interactions are often transient and weaker than protein-protein interactions due to the smaller interface between peptides and their protein partners.

Therefore, protein docking approaches need to address these challenges before they can routinely be used to predict peptide–protein binding.

The dramatic increase of peptide–protein structures available in the PDB has facilitated the development of more powerful docking and refinement methods for predicting peptide–protein interactions. Peptide-docking approaches can be classified as template-based docking or template-free docking, according to the amount of required input data. Template-based docking methods, or comparative methods, use known structures (templates) as a scaffold to generate a model of the desired complex [188]. These methods are favoured when the template is similar to the investigated

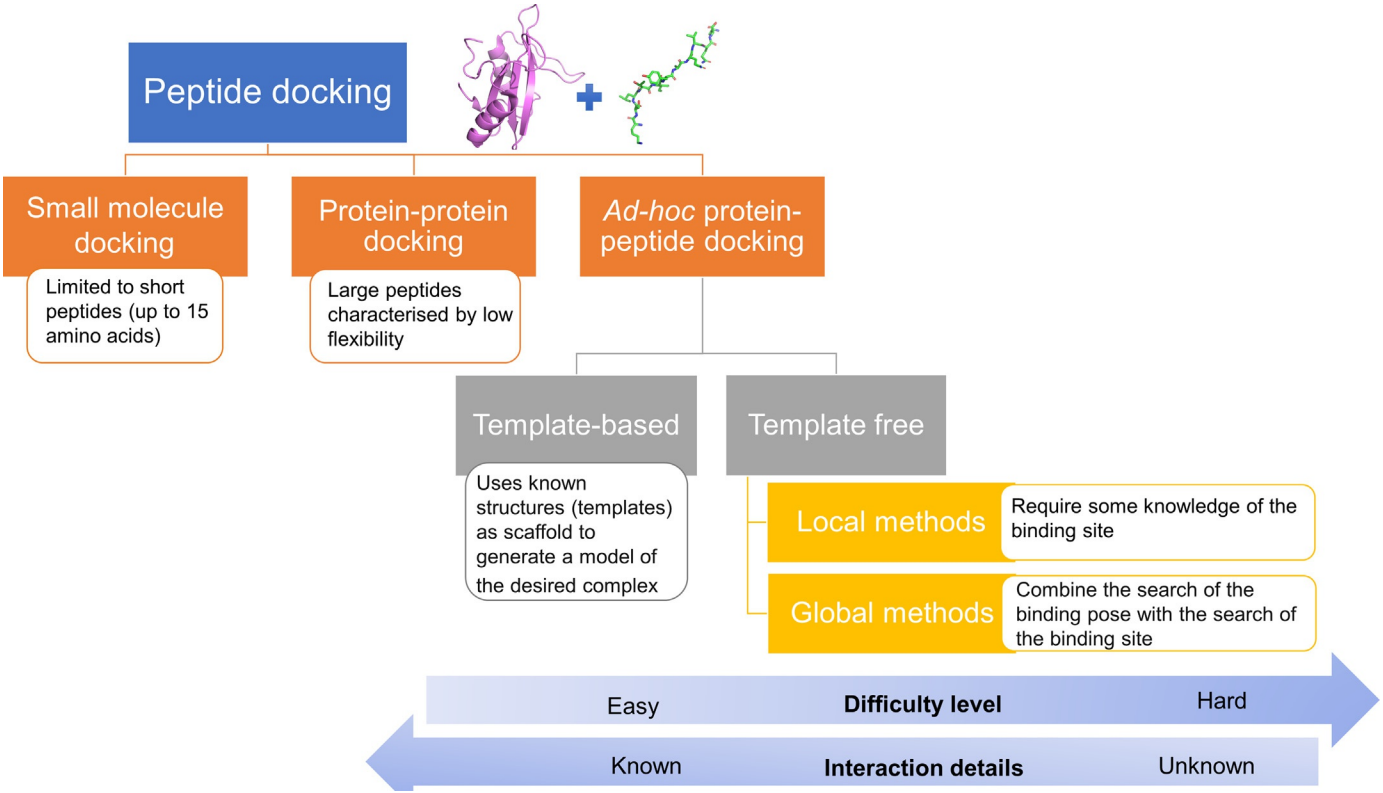

Fig. 12 Different methods in peptide docking. These methods include protein-small molecule docking, protein-protein docking and ad hoc protein-peptide docking. Due to the limitation of the searching algorithm, protein-small molecule docking approaches are limited to short peptides (up to 15 amino acids). Protein-protein approaches instead are limited to large peptides characterised by low flexibility. Ad hoc protein-peptide docking can be divided into two categories: template-based methods that utilise knowledge about the structure of similar complexes (templates) and template-free methods that can be implemented when no templates are available. Template-free methods are further divided in two categories according to the amount of required input data: local docking methods that require some knowledge about the binding site; and global docking methods that assume no knowledge about the peptide beyond its sequence. Different approaches have different level of difficulty and offer different levels of prediction accuracy, often determined by the amount of interaction information provided as input.

complex. Although several successful examples have been published, the main limitation of template-based docking methods is the number of known templates available, thus resulting in an innate limitation for general application [189]. An example of template-based method is GalaxyPepDock, a popular server that performs similarity-based docking by using experimentally determined protein-peptide structures to generate high-resolution complexes [190,191]. A recently developed machine learning-based method, PBRpredict, uses models trained from peptide binding residues of diverse types of domains to build models that robustly predict interacting residues in peptide binding [192].

Alternatively, users can use template-free methods. These methods are further classified based on the amount of information available for the protein and the protein-peptide binding site. Local docking methods require some knowledge about the binding site while global docking methods assume no prior knowledge about the peptide beyond its sequence [193].

Local docking methods are the mostly commonly used strategies and require an initial model of the complex prepared by the user to perform a search for a peptide binding pose in the proximity of a user-defined binding site. DynaDock [194], Rosetta FlexPepDock [195] and PepCrawler [196] are the most popular methods and provide different approaches to defining peptide binding sites.

However, when backbone conformational information of the query peptide is not available, sampling methods that allow acquisition of near-native peptide conformations are essential prior to performing local docking [188]. Rosetta FlexPepDock ab initio, for example, combines ab initio peptide folding with local docking by placing the query peptide into a user-defined binding site from any arbitrary backbone conformation [188,197].

Global docking methods combine the search of the binding pose with the search of the binding site. This makes global docking the method of choice when no prior information is available on the protein binding sites. The simplest approach in such docking is to treat the protein and the peptide input as rigid and to perform exhaustive rigid-body docking. More sophisticated approaches automatically predict the peptide conformation using a sequence provided by the user [188]. Alternatively, global docking can be combined with predictions of the binding site such as in in AnchorDock, which automatically identifies potential binding sites in the target protein and docks a flexible peptide in the proximity of these spots [198].

Although several docking programs have been developed, there is still a lack of a systematic evaluation to reveal the advantages and limitations of

these docking programs for protein-peptide systems. A recent effort to systematically assess the performance of docking programs was published by Weng et al. [199] using a large benchmark called PepSet composed of 185 protein-peptide complexes with peptide lengths ranging from 5 to 20 residues. Despite the advances in computational modelling of peptide–protein structures, authors show that major challenges still remain. First among these is flexibility; simultaneously modelling the backbone and side chain conformational change of both peptide and its target protein. Another is the integration of experimental data; combining all available experimental data to solve data ambiguity. For example, NMR experiments to identify native contacts, small-angle X-ray scattering (SAXS) or high-resolution cryo-EM to provide the shape of the bound complex. A third challenge is scoring; many lower-ranked poses were found to be of higher quality in the docking results than the top-ranked poses and vice versa. Interestingly, CAPRI experiments have revealed that hybrid approaches, using energy-based scoring as well as other methods such as mutagenesis, co-evolutionary information, sequence- or structural-clustering function can improve the docking performances and generate accurate peptide–protein docking results that are closer to native models [188].

6. Nucleic acid docking

Nucleic acids (NAs) play important roles in a large number of cellular processes, including cellular reproduction and protein synthesis. Thus DNA binders could interfere with the DNA replication process which affects cell proliferation or regulate the transcription process, and may result in the inhibition of gene expression. Similarly, RNA binders could interfere with the transcription and translation processes. Consequently NAs are potential drug targets for a number of diseases particularly in the area of anticancer, antibacterial and antivirus therapy (Fig. 13) [200].

Small molecules interact with NAs using different mechanisms: intercalation, cross-linkage, strand-cleavage, and reading-molecules. At the time of writing, there are 523 RNA-ligand co-crystalised structures and 730 DNA-ligand co-crystalised structures in the PDB, and the number is increasing year-on-year. These structural data provide not only the opportunity for investigating the molecular interaction between NAs and ligands but they enable structure-based computational methods for the design of nucleic acid-targeting ligands for specific diseases.

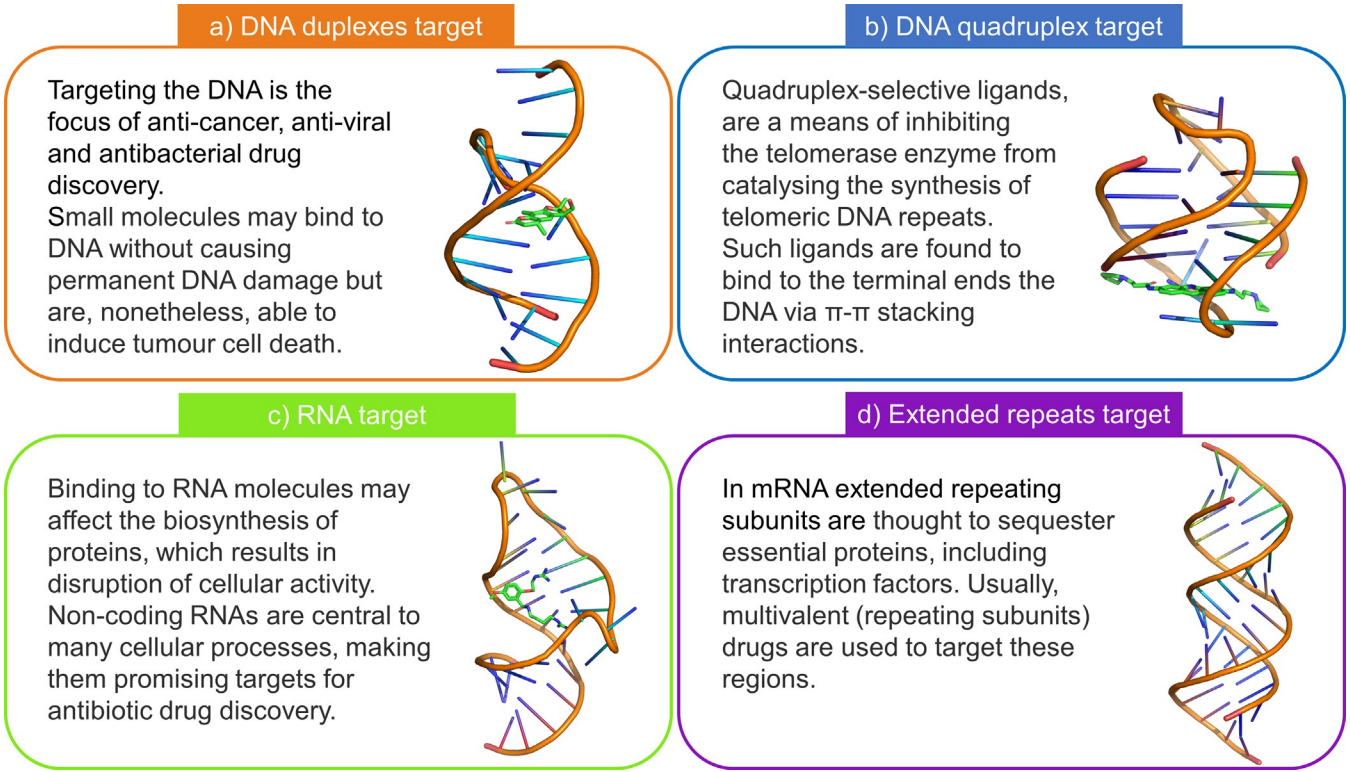

Fig. 13 Example of nucleic acid targets: (A) DNA duplex target; 3D structure of DNA duplex bound by psoralen molecule (PDB code 1fhz). (B) DNA quadruplex targets; 3D structure of DNA quadruplex bound by BSU6039 molecule (PDB code 1l1h). (C) RNA targets; 3D structure of HIV-1 TAR RNA bound to an inhibitor small molecule (PDB code 1uud). (D) Extended repeats target; 3D structure of the RNA duplex containing the CUG repeats associated with several neurodegenerative diseases, including myotonic dystrophy type 1 (PDB code 3gm7).

Generally, most of the protein-ligand docking programs can be used to dock small molecules into NAs, as they follow similar physicochemical binding principles. However, these programs often fail because of incomplete sampling of the conformational space and/or deficiencies of their internal scoring functions, which were not designed for docking of NAs (Fig. 14) [183,201].

While small molecule targeted proteins usually contain a well defined, generally hydrophobic binding site, NAs are characterised by more solvent exposed binding pockets with a high charge density and polarity. These differences require modifications to existing protein-ligand docking programs before they can be used for docking small molecules to NAs. Moreover, most of the current docking methods ignore flexibility and treat NAs as rigid receptors. However, NAs, and RNA in particular, are not rigid; they are usually characterised by induced fit movements and conformational changes in response to ligand binding (e.g. riboswitch) [202,203]. To handle flexibility, specific docking tools such as MORDOR [204] have been developed to allow induced fit binding of small molecule ligands with RNA targets via flexible receptors. MORDOR allows flexibility on both ligand and NAs by applying molecular mechanics minimisation with a restrained conformational search based on the X-ray or NMR experimental structure [205]. Despite the high accuracy, the energy minimisation step is time-consuming and makes this method less satisfactory for screening large libraries [204]. A more computationally efficient approach to deal with NA flexibility is to use ensemble docking, using a set of predetermined NA conformations that can be obtained from different X-ray crystal structures, NMR models, or normal-mode analysis of an MD simulation.

Additionally, scoring functions have been specifically designed for predicting NA-ligand affinities to address the high polarity of NAs compared to proteins and the interactions that require more attention in NAs (e.g. electrostatic interaction, and solvation). RiboDock [206] for example, includes an empirical scoring function with a number of extensions to capture important NA-ligand interaction motifs, such as the interaction between positively charged carbons (e.g. guanidinium) and negatively charged groups (e.g. carbonyl), and an energetic term for parallel π-systems (stacking) to account for electrostatic interactions that characterise NA-ligand complexes.

Two novel knowledge-based scoring functions are DrugScore^{RNA} [207], which was trained from statistical analyses of a set of NA-ligand complexes, and LigandRNA [208] which, in contrast to DrugScore^{RNA} considers the directionality of hydrogen bonds.

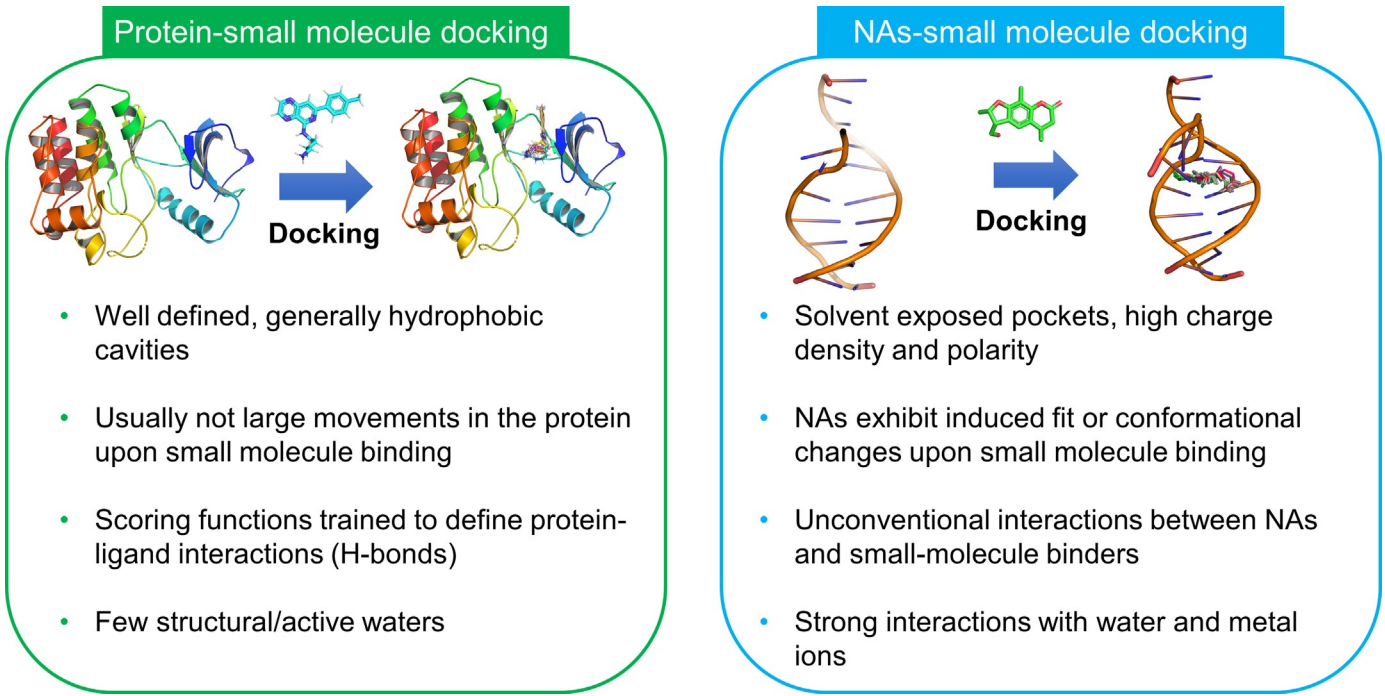

Fig. 14 Main differences between protein-small molecule docking and NAs-small molecule docking.

Another challenge in docking into NAs is the presence of water molecules and metal ions. The highly charged nature of NAs leads to strong interactions with water molecules and metal ions that are not only essential for stabilising and shaping NA structures, but are often necessary for optimal ligand binding. Therefore, docking tools that allow explicit treatment of waters are highly recommended when docking small molecules into NAs [209,210].

Inclusion of water molecules in the docking run has shown better success rate with redocking (or self-docking procedures) [14]. However, caution should be used with cross-docking which may require a different set of water molecules for different ligands. To address this issue, virtually 'displaceable' molecules and a GB/SA model (Generalised-Born model with solvent accessible surface area (SA) term) has been implemented in the solvation module of DOCK [209] to reproduce the electrostatic contribution of waters and ions. The combination with explicit water and counterions in DOCK reaches accuracy values comparable to DrugScore^{RNA} when applied to RNA. Metal ions such as Mg^{2+} and Mn^{2+} are often found at the binding site of NAs, acting as a 'metal bridge' coordinating interactions between ligand and NA residues. When metal ions are included as part of the NA target, it results in improved pose prediction of several targets including TPP riboswitch [211].

Despite the limitations, docking-based screens have already helped medicinal chemists identifying novel NA binders of DNA quadruplex [212,213], triple helix DNA [214] and RNA [215]. These prospective investigations reveal the potential of docking programs in NA binders discovery campaigns [201].

7. Current challenges

Docking is currently in a mature stage of development, but it is still far from perfect.

Most docking programs available are normally able to predict known binding poses with averaged accuracies of about 1.5–2 Å with reported success rates in the range of 70–80% [12]. However, the calculation of accurate binding energies is one of the major limitations in molecular docking, directly correlated with all the approximations made during a docking run (e.g. the treatment of solvent and the flexibility of the macromolecular system).

The lack of a suitable scoring function and searching algorithm, able to efficiently combine both accuracy and speed, are perhaps the most detrimental weakness of docking and have been widely discussed elsewhere in the chapter.

Therefore, despite its invaluable contribution to understanding target-ligand interactions in support of drug discovery projects, the results of a docking experiment should not be taken as the end result, but rather as a good starting point or as part of a workflow for a deeper and more accurate analysis.

In Section 7.1, we will focus on some of the areas where the above-mentioned limitations are likely to impact.

7.1 Blind docking

Blind docking refers to docking a ligand to the whole surface of a protein without any prior knowledge of the target pocket. In blind docking, the entire protein is considered as a region where a ligand might bind leading to a much larger search space with a corresponding increase in the running time. Moreover, the complexity of blind docking grows exponentially according to the number of possible binding sites, which severely limits its use in practice. However, blind docking methods which can predict bound conformations with no a priori knowledge of binding site locations will be needed for fully automated computational approaches for in silico drug design.

The enormous search space is the principal problem blind docking must tackle. Two possible ways to mitigate this are either to reduce the search complexity and split the docking box into multiple boxes, sacrificing the flexibility of some parts of the ligand, or to repeat the search several times using different seeds and then merge the results together, as opposed to one larger blind docking run that covers the complete protein structure.

An alternative approach to decrease the computational complexity of blind docking is to combine binding site prediction tools to identify putative ligand binding sites, with the docking experiment followed by fine-tuning and ranking of the initial solutions using scoring functions and optimisation methods. An example of such approach is BSP-SLIM, an integrated tool in which algorithms for the template-based ligand binding site prediction are incorporated with the SLIM docking method [216].

A more advanced approach for sampling flexibility in blind docking is represented by the Protein Energy Landscape Exploration (PELE), a

Monte Carlo-based technique combined with a protein structure prediction algorithm. Three main steps are performed. Firstly, the ligand and protein perturbation (using a rotamer library), secondly, side chain sampling (via algorithms), and lastly minimisation and acceptance using the Metropolis acceptance criteria. Although such an approach is computationally expensive, it is still lower than for MD simulations [217].

7.2 Covalent docking

Historically, drug discovery mainly focuses on non-covalent drugs due to potential off-target effects and toxicity issues of irreversible covalent drugs. However, in recent years and with the outbreak of Covid-19, we have witnessed the resurgence of covalent drugs [218–221]. Compared to non-covalent drugs, covalent drugs might have extra advantages, including better efficacy, since they are more competitive than non-covalent endogenous substrates. They also offer a lower patient burden and less drug resistance due to lower and less frequent dosing and improved target specificity by careful designs that target specific protein residues [219].

The rational design of covalent ligands is still faced with particular challenges, mostly related to the fact that covalent bond formation, bond breaking and bond rearrangements are quantum mechanical (QM) phenomena which cannot adequately be handled by the force fields or empirical approaches typically used for non-covalent protein-ligand interactions [222].

Historically, to overcome such limitations, many manual interventions and ad hoc solutions have been used to adapt existing docking tools for application with covalent ligands [222]. This has changed more recently, and dedicated workflows and protocols for handling covalent ligands and covalent docking in virtual screening have started to emerge. Even though contributions from QM methods are increasingly incorporated into docking applications, a full QM treatment is still unfeasible in routine applications, given the size of the molecular systems and the number of configurations and compounds to consider. However, the need for QM calculations in covalent docking could be circumvented with faster and simpler modelling approaches and in many cases the QM treatment of docking process may not be required.

Is the binding site known? Is the targeted amino acid and its reactivity known? Is the type of electrophilic warhead of the ligands known? Depending on the answer to these questions different scenarios and different requirements apply. In the simplest case where the target site is well known,

particular nucleophilic amino acids featuring in the binding site can be targeted; frequently, the nucleophilic amino acid to target is also known in advance (i.e. from previous experiments). In these cases, the primary task for docking is just to elucidate the binding modes and/or to rank the compounds of interest according to their suitability to fit into the active site after covalent linking to the protein. Therefore, a reactivity assessment is not really required, and classical scoring functions developed on non-covalent interactions are sufficient for the task as are most of the docking programs. For advanced tasks such as the design of covalent ligands for systems without much prior knowledge, more sophisticated approaches are required. A QM-based scoring function has been developed to describe covalent binding of ligands by providing firstly a description of covalent bond breakage using hybrid-QM/semiempirical-QM restrained optimisations starting from the covalent complex (crystal structure or a modelled structure derived thereof) and, second, the addition of a new term ($\Delta G'_{cov}$), to the non-covalent score. This term describes the 'free' energy difference between the covalent and non-covalent complex [223]. Such an approach helps to overcome the common neglection of the energy contributions from covalent bond formation however, it is computationally demanding and more suitable as a post-docking rescoring procedure for selected candidates.

Despite recent advances in QM-based methods [222], an approach in which the covalent docking process itself is driven by QM calculations is not yet on the horizon. Similarly, blind docking remains impossible for covalent ligands because for all approaches available the binding site and the targeted reactive residue must be known in advance.

7.3 Reverse docking

Reverse docking (RD) or inverse docking, as the name suggests, involves docking of a set of one or a few ligands against an array of protein families with the aim of identifying a potential target, their binding affinity or polypharmacology profile (Fig. 15). Additionally, RD can provide a valuable contribution in drug repositioning or drug repurposing and drug rescue, it may reveal targets of drugs with so far unknown mechanism, and contribute to rationally designing less toxic or multitarget drugs [224]. Therefore, clinically approved drugs could be repurposed for other diseases, different from the one they were originally designed for [225,226]; a well-known example is sildenafil [227] a phosphodiesterase-5 (PDE5) inhibitor, used to treat erectile dysfunction but which was first developed for the treatment of angina. Another

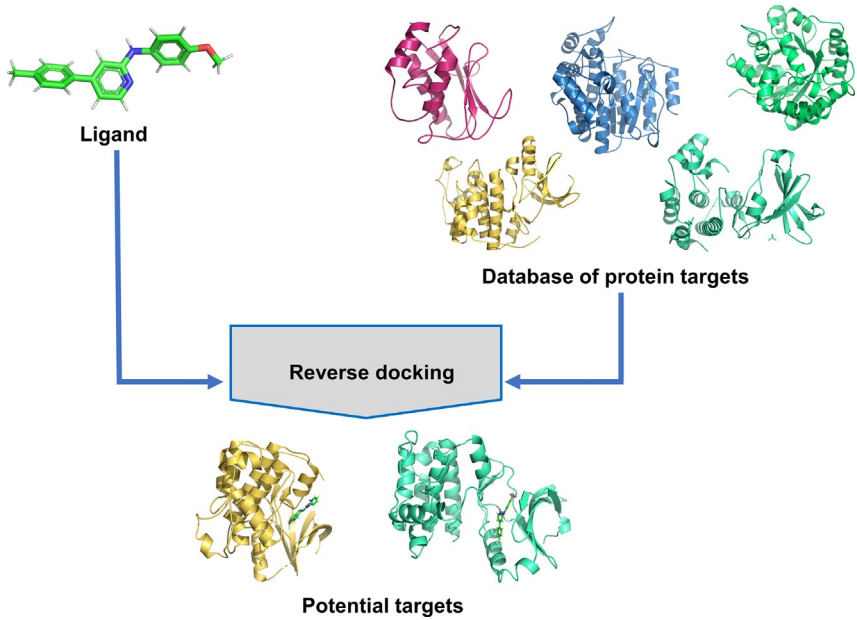

Fig. 15 A flowchart of reverse docking. In a reverse docking a ligand (or multiple ligands) is docked against an array of protein families with the aim of identifying potential targets.

example of a profitable drug repurposing is Minoxidil, which was originally developed for hypertension but was later repositioned to treat male hair loss [228]. While RD was not employed in the discovery of these materials they spurred the application of computational approaches for repurposing [229] that are for example being utilised in the search for treatments for infectious diseases, including Covid-19 [230].

With the development of computing resources, RD has drawn more and more attention in recent years and has had some success in target identification with the predicted results verified by bioassays and crystallographic studies [231,232]. These successful cases show that reverse docking is playing an important role in protein target predictions of small molecules. RD approaches have also proved successful for identifying adverse effects of drugs. For example, RD screenings of a series of anti-HIV drugs highlighted several proteins whose modulation has been associated with adverse reactions in literature [233].

With the increasing amount of crystallographic data available, several databases of protein targets are available to help preform RD screens, providing information about protein structures, diseases, biological functions,

and drugs [234]. Moreover, tailored libraries of targets can also be manually built upon publicly available databases of crystal structures and binding pockets, such as PDB [77], sc-PDB [235] and the Therapeutic Target Database (TTD) [236] or can be additionally extended through homology modelling techniques. Because these libraries were not specifically designed for RD, each structure in the databases needs to be properly prepared for the docking calculations. Depending on the library and the number of targets included this step can be time-consuming. These databases are far from exhaustive. Using the data in DrugPort (http://www.ebi.ac.uk/thornton-srv/databases/drugport/) as an example, there are a total of 1664 known druggable protein targets in the database, but only about half of them have 3D structures in the PDB [237]. Furthermore, targets with known structures are not evenly distributed among different superfamilies, due to experimental limitations. For example, the superfamily of membrane proteins, the GPCR, is one of the most important targets in drug design; they account for over a quarter of the known drug targets and about half of the drugs on the market target GPCRs specifically. However, only a fraction of the GPCRs have experimental structures because the structural resolution of membrane proteins like GPCRs is much more complicated and difficult to elucidate than globular proteins such as enzymes [237].

The main limitation of RD approaches is primarily related to the accuracy of current scoring functions to distinguish true target from non-target proteins. Many (if not all) contemporary scoring functions are designed for docking or screening small molecules in protein binding sites without special optimisation for RD, which certainly would be influenced by the properties of protein pockets, resulting in scoring bias to the proteins with particular properties. This bias would produce large number of false positives, interfering with the identification of true targets. In recent years, several attempts have been made to improve the accuracy of docking scores in reverse docking [238–240]. Thus, the docking score or the scoring functions of current docking programs should be rationalised to suit the reverse docking.

Very recently, integration of a docking approach with more sophisticated ML-based methods has also been explored for target prediction [241–243]. Although such methods have achieved notable prediction performances, both in terms of target ranking and on multitarget selectivity predictions [241] they are time-consuming and computationally demanding; moreover, they also need a large amount of bioactivity data to train the models, which will not always be available for some of the targets under study [244].

8. Looking forward

The modalities by which docking is used to assist the different tasks of drug discovery have changed over time. Although it was initially developed and used as a standalone method, docking is now mostly employed in combination with other computational approaches within integrated workflows, to overcome some of the most relevant intrinsic limitations characterising molecular docking.

Applications of combined workflows which include docking, have been explored to assist different tasks of drug discovery. For example, docking has been used in tandem with ligand-based, binding free energy calculations, and AI/ML approaches to improve the prediction performances in de novo virtual screening, as well as to assist target fishing, adverse drug reactions (ADR) prediction and drug repurposing [244]. Likewise, different approaches can also be applied at different phases of the screening workflow to improve docking predictions. For example, MD could be combined with AI-based methods to identify suitable receptor conformations for docking. Then, ligand-based approaches could be applied for rescoring the predicted docking poses.

Considering the number of in silico tools and techniques currently available, there are still countless opportunities for docking to be explored in integrated workflows. Moreover, their integration has also been facilitated by the continuous improvements in hardware and software engineering. For example, parallelisation of molecular docking has enabled the in silico screening of millions of compounds in affordable time by processing the most computationally consuming task (the energy calculation phase) on multiple CPUs using distributed computing infrastructures (DCIs).

In the context of blind docking this computationally expensive task has to be multiplied by the number of binding sites which can also be very large, thus it is of great interest to find different ways to speed-up the whole docking process. A DCI can be a cloud computing resource. Cloud computing is now available on-demand and users are typically charged on a pay-per-use basis. This can make scientific applications, such as VS, more accessible for scientists around the world, lowering the cost of using complex computing infrastructure.

Over the years, there has been rapid progress in developing faster architecture based on graphics processing unit (GPU) clusters, more adequate algorithms for optimised computational analysis, and the tracking of ligand-receptor binding expressed in scoring functions.

Though the initial applications of GPUs were exclusively for creation of computer graphics, GPUs have now matured to a state where they can be successfully applied for non-graphical purposes. For example, GPUs are widely used in drug discovery to deal with demanding computational tasks such as MD simulation or QM calculations. In the context of docking and VS, GPU calculations are particularly appealing to enable one to explore the large conformational landscape potentially accessible to proteins, in shorter times compared to CPUs. GPU-optimised docking and cloud solutions may become the standard soon. Lastly, GPU computing made big-data driven computation tasks more generally accessible, and it is expected to play a prominent role, not only in docking but in future in silico drug design in general.

Another important factor to consider in the further development and optimisation of the docking methodologies is the availability of 3D data. In recent decades, the number of crystallographic data has increased exponentially and new technologies such as cryo-EM and NMR have filled several gaps left by the limitation of X-ray crystallography. Of the three, cryo-EM has emerged as successful for determining the structures of large and dynamic complexes that have proved difficult to obtain by other approaches.

Despite the fact that cryo-EM is a decades-old technique, it has garnered increasing interest since around 2013 due to a series of technological and algorithmic advances that together drove a striking improvement in the resolution obtainable by this technique. Recently, Yip et al. [245] and Nakane et al. [246] reported the sharpest images yet obtained, with a resolution of 1.2 Å, by using a method termed single-particle cryo-EM, enabling the location of individual atoms in a protein to be determined for the first time. Ultimately, these developments will help researchers gain a better understanding, at unprecedented resolution, of how proteins work in health and disease, with the potential to aid the design of better therapeutics. Advancements in cryo-EM have also increased the rate at which structures are solved, such that many general AI-based methods for protein-ligand docking will be applied to modelling protein-ligand recognition in the future. The improved quality of molecular structures has the additional advantage of improving ML and AI-based tools where the main limitation for their application in drug discovery is the lack of large, annotated, unbiased, high-quality data. As a data mining technology, the amount of available data directly affects the performance of the related deep learning models since the successful training of deep neural networks highly relies on a large amount of data.

Developments in computing power coupled to imaging techniques, AI approaches and protein fold prediction tools (i.e. AlphaFold [247]) point to a golden age for the application of computational approaches to drug discovery, where understanding the protein's shape and knowing its role within the cell, provides better protein structures for molecular docking, contributing to the development of specific and selective new therapeutics, while also reducing the costs associated with experimentation.

References

[1] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 2016;47:20–33.
[2] Ripphausen P, Nisius B, Peltason L, Bajorath J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. J Med Chem 2010;53(24):8461–7.
[3] Druker BJ, Lydon NB. Lessons learned from the development of an Abl tyrosine kinase inhibitor. J Clin Invest 2000;105(1):3–7.
[4] Von Itzstein M, Wu WY, Kok GB, Pegg MS, Dyason JC, Jin B, et al. Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature 1993;363 (6428):418–23.
[5] Kaldor SW, Kalish VJ, Davies JF, Shetty BV, Fritz JE, Appelt K, et al. Viracept (nelfinavir mesylate, AG 1343): a potent, orally bioavailable inhibitor of HIV-1 protease. J Med Chem 1997;40(24):3979–85.
[6] Squires M, Ward G, Saxty G, Berdini V, Cleasby A, King P, et al. Potent, selective inhibitors of fibroblast growth factor receptor define fibroblast growth factor dependence in preclinical cancer models. Mol Cancer Ther 2011;10(9):1542–52.
[7] Talele T, Khedkar S, Rigby A. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. Curr Top Med Chem 2010;10 (1):127–41.
[8] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3 (11):935–49.
[9] Clark DE. What has computer-aided molecular design ever done for drug discovery? Expert Opin Drug Discov 2006;1(2):103–10.
[10] Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C. Lead optimization via high-throughput molecular docking. Curr Opin Drug Discov Devel 2007; 10(3):264–74.
[11] Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161(2):269–88.
[12] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev 2017;9(2):91–102.
[13] Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys 2016;18 (18):12964–75.
[14] Li Y, Shen J, Sun X, Li W, Liu G, Tang Y. Accuracy assessment of protein-based docking programs against RNA targets. J Chem Inf Model 2010;50(6):1134–46.
[15] Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, et al. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 2009;49(6):1455–74.
[16] Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Molecules 2015;20(7):13384–421.

[17] Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161(2):269–88.
[18] Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr Sect A 1976;32(5):922–3.
[19] Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr Sect A 1978;34(5):827–8.
[20] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47(7):1739–49.
[21] Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 2004;47(7):1750–9.
[22] Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261(3):470–89.
[23] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking 1. J Mol Biol 1997;267(3):727–48.
[24] Hart TN, Read RJ. A multiple-start Monte Carlo docking method. Proteins 1992;13 (3):206–22.
[25] Korb O, Stützle T, Exner TE. PLANTS: application of ant colony optimization to structure-based drug design. In: Dorigo M, Gambardella LM, Birattari M, Martinoli A, Poli R, Stützle T, editors. Ant colony optimization and swarm intelligence. Berlin Heidelberg: Springer; 2006. p. 247–58.
[26] Pei J, Wang Q, Liu Z, Li Q, Yang K, Lai L. PSI-DOCK: towards highly efficient and accurate flexible ligand docking. Proteins 2006;62(4):934–46.
[27] Li Y, Han L, Liu Z, Wang R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. J Chem Inf Model 2014;54(6):1717–36.
[28] Li J, Fu A, Zhang L. An overview of scoring functions used for protein–ligand interactions in molecular docking. Interdiscip Sci Comput Life Sci 2019;11(2):320–8.
[29] Ain QU. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci 2015;5:405–24.
[30] Zou X, Yaxiong, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. J Am Chem Soc 1999;121(35):8033–43.
[31] Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. J Chem Inf Model 2017;57 (12):2911–37.
[32] Mark AE, van Gunsteren WF. Decomposition of the free energy of a system in terms of specific interactions: implications for theoretical and experimental studies. J Mol Biol 1994;240(2):167–76.
[33] Williams DH, Maguire AJ, Tsuzuki W, Westwell MS. An analysis of the origins of a cooperative binding energy of dimerization. Science 1998;280(5364):711–4.
[34] Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52(4):609–23.
[35] Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 1998;19(14):1639–62.
[36] Corbeil CR, Williams CI, Labute P. Variability in docking success rates due to dataset preparation. J Comput Aided Mol Des 2012;26(6):775–86.
[37] Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11(5):425–45.

[38] Korb O, Stützle T, Exner TE. Empirical scoring functions for advanced protein–ligand docking with PLANTS. J Chem Inf Model 2009;49(1):84–96.
[39] Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem 2006;27(15):1876–82.
[40] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257(2):457–69.
[41] Velec HFG, Gohlke H, Klebe G. Drug score CSD-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem 2005;48(20):6296–303.
[42] Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295(2):337–56.
[43] Mooij WTM, Verdonk ML. General and targeted statistical potentials for protein–ligand interactions. Proteins Struct Funct Bioinf 2005;61(2):272–87.
[44] Muegge I, Martin YC. A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 1999;42(5):791–804.
[45] Muegge I. PMF scoring revisited. J Med Chem 2006;49(20):5895–902.
[46] Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. J Comput Chem 2006;27(15):1866–75.
[47] Huang S-Y, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Phys Chem Chem Phys 2010;12(40):12899–908.
[48] Verdonk ML, Ludlow RF, Giangreco I, Rathi PC. Protein–ligand informatics force field (PLIff): toward a fully knowledge driven "force field" for biomolecular interactions. J Med Chem 2016;59(14):6891–902.
[49] Zheng M, Xiong B, Luo C, Li S, Liu X, Shen Q, et al. Knowledge-based scoring functions in drug design: 3. A two-dimensional knowledge-based hydrogen-bonding potential for the prediction of protein–ligand interactions. J Chem Inf Model 2011; 51(11):2994–3004.
[50] Neudert G, Klebe G. DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. J Chem Inf Model 2011;51(10):2731–45.
[51] Iruela-Arispe ML, Liska DJ, Sage EH, Bornstein P. Differential expression of thrombospondin 1, 2, and 3 during murine development. Dev Dyn 1993;197(1):40–56.
[52] Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model 2009;49(4):1079–93.
[53] Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42(25):5100–9.
[54] O'Boyle NM, Liebeschuetz JW, Cole JC. Testing assumptions and hypotheses for rescoring success in protein − ligand docking. J Chem Inf Model 2009;49(8): 1871–8.
[55] Palacio-Rodríguez K, Lans I, Cavasotto CN, Cossio P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. Sci Rep 2019;9 (1):5142.
[56] Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, et al. Virtual screening using protein − ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44(3):793–806.
[57] Khamis MA, Gomaa W, Ahmed WF. Machine learning in computational docking. Artif Intell Med 2015;63(3):135–52.
[58] Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 2017;7(1):46710.

[59] Shen C, Hu Y, Wang Z, Zhang X, Zhong H, Wang G, et al. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. Brief Bioinform 2021;22(1):497–514.
[60] Li H, Sze KH, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdiscip Rev Comput Mol Sci 2020;10:1–20. e1465.
[61] Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26 (9):1169–75.
[62] Li L, Wang B, Meroueh SO. Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. J Chem Inf Model 2011;51(9):2132–8.
[63] Çınaroğlu SS, Timuçin E. Comparative assessment of seven docking programs on a nonredundant metalloprotein subset of the PDBbind refined. J Chem Inf Model 2019;59(9):3846–59.
[64] ten Brink T, Exner TE. Influence of protonation, tautomeric, and stereoisomeric states on protein − ligand docking results. J Chem Inf Model 2009;49(6):1535–46.
[65] Chen X, Lin Y, Liu M, Gilson MK. The binding database: data management and interface design. Bioinformatics 2002;18(1):130–9.
[66] Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. Binding DB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 2015;44(D1):D1045–53.
[67] Chen X, Liu M, Gilson MK. Binding DB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen 2001;4(8):719–25.
[68] Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind database: methodologies and updates. J Med Chem 2005;48(12):4111–9.
[69] Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. J Med Chem 2004;47(12):2977–80.
[70] Grudinin S, Popov P, Neveu E, Cheremovskiy G. Predicting binding poses and affinities in the CSAR 2013–2014 docking exercises using the knowledge-based convex-PL potential. J Chem Inf Model 2015;56(6):1053–62.
[71] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014;42(Database issue):D1083–90.
[72] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1): D1100–7.
[73] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 2012;55 (14):6582–94.
[74] Kroemer RT, Vulpetti A, McDonald JJ, Rohrer DC, Trosset J-Y, Giordanetto F, et al. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. J Chem Inf Comput Sci 2004;44(3):871–81.
[75] Yusuf D, Davis AM, Kleywegt GJ, Schmitt S. An alternative method for the evaluation of docking performance: RSR vs RMSD. J Chem Inf Model 2008;48(7):1411–22.
[76] Baber JC, Thompson DC, Cross JB, Humblet C. GARD: a generally applicable replacement for RMSD. J Chem Inf Model 2009;49(8):1889–900. Aug 24.
[77] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[78] Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein–ligand structures in drug discovery. Drug Discov Today 2012;17(23–24): 1270–81.

[79] Wlodawer A, Minor W, Dauter Z, Jaskolski M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. FEBS J 2008;275(1):1–21.
[80] Read RJ, Adams PD, Arendall WB, Brunger AT, Emsley P, Joosten RP, et al. A new generation of crystallographic validation tools for the protein data bank. Structure 2011;19(10):1395–412.
[81] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 2019;20(11):681–97.
[82] Mishra S, Demo G, Koča J, Wimmerová M. In silico engineering of proteins that recognize small molecules. InTech; 2012.
[83] Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 2013;27(3):221–34.
[84] Kalliokoski T, Salo HS, Lahtela-Kakkonen M, Poso A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. J Chem Inf Model 2009;49(12):2742–8.
[85] Onufriev AV, Alexov E. Protonation and pK changes in protein-ligand binding. Q Rev Biophys 2013;46(2):181–209.
[86] Berry M, Fielding B, Gamieldien J. Practical considerations in virtual screening and molecular docking. Emerg Trends Comput Biol Bioinf Syst Biol; 2015. p. 487–502.
[87] Li S, Hong M. Protonation, tautomerization, and rotameric structure of histidine: a comprehensive study by magic-angle-spinning solid-state NMR. J Am Chem Soc 2011;133(5):1534–44.
[88] Kim MO, Nichols SE, Wang Y, McCammon JA. Effects of histidine protonation and rotameric states on virtual screening of M. tuberculosis RmlC. J Comput Aided Mol Des 2013;27(3):235–46.
[89] Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J 2020;18:417–26.
[90] Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15(6):359–63.
[91] Kawabata T. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 2010;78(5):1195–211.
[92] Simões TMC, Gomes AJP. CavVis—a field-of-view geometric algorithm for protein cavity detection. J Chem Inf Model 2019;59(2):786–96.
[93] Coleman RG, Sharp KA. Protein pockets: inventory, shape, and comparison. J Chem Inf Model 2010;50(4):589–603.
[94] B-Rao C, Subramanian J, Sharma S. Managing protein flexibility in docking and its applications. Drug Discov Today 2009;14:394–400.
[95] Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. J Med Chem 2004;47(21):5076–84.
[96] Cosconati S, Marinelli L, Di Leva FS, La Pietra V, De Simone A, Mancini F, et al. Protein flexibility in virtual screening: the BACE-1 case study. J Chem Inf Model 2012;52(10):2697–704.
[97] Rueda M, Bottegoni G, Abagyan R. Recipes for the selection of experimental protein conformations for virtual screening. J Chem Inf Model 2010;50(1):186–93.
[98] Najmanovich R, Kuttner J, Sobolev V, Edelman M. Side-chain flexibility in proteins upon ligand binding. Proteins Struct Funct Genet 2000;39(3):261–8.
[99] Korb O, Olsson TSG, Bowden SJ, Hall RJ, Verdonk ML, Liebeschuetz JW, et al. Potential and limitations of ensemble docking. J Chem Inf Model 2012;52(5):1262–74.
[100] Campbell AJ, Lamb ML, Joseph-McCarthy D. Ensemble-based docking using biased molecular dynamics. J Chem Inf Model 2014;54(7):2127–38.

[101] Carlson HA, Masukawa KM, McCammon JA. Method for including the dynamic fluctuations of a protein in computer-aided drug design. J Phys Chem A 1999;103 (49):10213–9.
[102] Fernández-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci 2002;11(2):280–91.
[103] Vieth M, Hirst JD, Kolinski A, Brooks III CL. Assessing energy functions for flexible docking. J Comput Chem 1998;19(14):1612–22.
[104] Hou X, Li K, Yu X, Sun J, Fang H. Protein flexibility in docking-based virtual screening: discovery of novel lymphoid-specific tyrosine phosphatase inhibitors using multiple crystal structures. J Chem Inf Model 2015;55(9):1973–83.
[105] Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins Struct Funct Bioinf 2000;40(3):389–408.
[106] Gaudreault F, Chartier M, Najmanovich R. Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. Bioinformatics 2012;28(18):i423–30.
[107] Bottegoni G, Rocchia W, Rueda M, Abagyan R, Cavalli A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. PLoS One 2011;6(5): e18845.
[108] Tian S, Sun H, Pan P, Li D, Zhen X, Li Y, et al. Assessing an ensemble docking-based virtual screening strategy for kinase targets by considering protein flexibility. J Chem Inf Model 2014;54(10):2664–79.
[109] Sander B, Korb O, Cole J, Essex JW. How to pick a winning team: approaches towards the selection of computationally derived protein structures for ensemble-based virtual screening. J Cheminform 2013;5(Suppl. 1):O7.
[110] Roberts BC, Mancera RL. Ligand–protein docking with water molecules. J Chem Inf Model 2008;48(2):397–408.
[111] Abel R, Young T, Farid R, Berne BJ, Friesner RA. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. J Am Chem Soc 2008;130 (9):2817–31.
[112] Young T, Abel R, Kim B, Berne BJ, Friesner RA. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. Proc Natl Acad Sci 2007;104(3):808–13.
[113] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79 (2):926–35.
[114] Imai T, Hiraoka R, Kovalenko A, Hirata F. Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. Proteins Struct Funct Bioinf 2007;66(4):804–13.
[115] Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, et al. Modeling water molecules in protein–ligand docking using GOLD. J Med Chem 2005;48(20):6504–15.
[116] Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins 2002;46(1):34–40.
[117] Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J Med Chem 2006;49(21):6177–96.
[118] Kroemer RT. Structure-based drug design: docking and scoring. Curr Protein Pept Sci 2007;8(4):312–28.
[119] Zhong H, Wang Z, Wang X, Liu H, Li D, Liu H, et al. Importance of a crystalline water network in docking-based virtual screening: a case study of BRD4. Phys Chem Chem Phys 2019;21(45):25276–89.

[120] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. AAPS J 2012;14(1):133–41.
[121] Huang N, Shoichet BK. Exploiting ordered waters in molecular docking. J Med Chem 2008;51(16):4862–5.
[122] Hu B, Lill MA. WATsite: hydration site prediction program with PyMOL interface. J Comput Chem 2014;35(16):1255–60.
[123] Verdonk ML, Cole JC, Taylor R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. J Mol Biol 1999;289(4):1093–108.
[124] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28 (7):849–57.
[125] Sridhar A, Ross GA, Biggin PC. Waterdock 2.0: water placement prediction for Holo-structures with a pymol plugin. PLoS One 2017;12(2):1–17.
[126] Cole JC, Korb O, McCabe P, Read MG, Taylor R. Knowledge-based conformer generation using the Cambridge structural database. J Chem Inf Model 2018;58 (3):615–29.
[127] Taylor R, Cole J, Korb O, McCabe P. Knowledge-based libraries for predicting the geometric preferences of drug like molecules. J Chem Inf Model 2014;54(9): 2500–14.
[128] Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model 2010;50 (4):572–84.
[129] Ebejer J-P, Morris GM, Deane CM. Freely available conformer generation methods: how good are they? J Chem Inf Model 2012;52(5):1146–58.
[130] O'Boyle NM, Vandermeersch T, Flynn CJ, Maguire AR, Hutchison GR. Confab-systematic generation of diverse low-energy conformers. J Cheminform 2011;3(1):8.
[131] Schärfer C, Schulz-Gasch T, Hert J, Heinzerling L, Schulz B, Inhester T, et al. CONFECT: conformations from an expert collection of torsion patterns. ChemMed Chem 2013;8(10):1690–700.
[132] Landrum G, et al. RDKit: Open-source cheminformatics, http://www.rdkit.org/. 2018; 2018.
[133] Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. Acta Crystallogr Sect B Struct Sci Cryst Eng Mater 2016;72(2):171–9.
[134] Brooks WH, Daniel KG, Sung S-S, Guida WC. Computational validation of the importance of absolute stereochemistry in virtual screening. J Chem Inf Model 2008;48(3):639–45.
[135] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46(1–3):3–26.
[136] Shultz MD. Two decades under the influence of the rule of five and the changing properties of approved oral drugs. J Med Chem 2019;62(4):1701–14.
[137] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 2012;52(7):1757–68.
[138] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20—a free Ultralarge-scale chemical database for ligand discovery. J Chem Inf Model 2020;60(12):6065–73.
[139] Sterling T, Irwin JJ. ZINC 15—ligand discovery for everyone. J Chem Inf Model 2015;55(11):2324–37.
[140] Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the Oral bioavailability of drug candidates. J Med Chem 2002;45(12):2615–23.

[141] Anon. Downloadable Structure Files of NCI Open Database Compounds. Available from https://cactus.nci.nih.gov/download/nci/.
[142] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006;34(Database issue):D668–72.
[143] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46(D1): D1074–82.
[144] ASINEX. Screening libraries. Available from: http://www.asinex.com/libraries-html/.
[145] Cole JC, Wiggin S, Stanzione F. New insights and innovation from a million crystal structures in the Cambridge structural database. Struct Dyn 2019;6(5):054301.
[146] Taylor R, Wood PA. A million crystal structures: the whole is greater than the sum of its parts. Chem Rev 2019;119(16):9427–77.
[147] Congreve M, Carr R, Murray C, Jhoti H. A 'rule of three' for fragment-based lead discovery? Drug Discov Today 2003;8(19):876–7.
[148] Verdonk ML, Giangreco I, Hall RJ, Korb O, Mortenson PN, Murray CW. Docking performance of fragments and druglike compounds. J Med Chem 2011;54(15): 5422–31.
[149] Leach AR, Hann MM, Burrows JN, Griffen EJ. Fragment screening: an introduction. Mol Biosyst 2006;2(9):429–46.
[150] Bian Y, Xie X-QS. Computational fragment-based drug design: current trends, strategies, and applications. AAPS J 2018;20(3):59.
[151] Fink T, Reymond J-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. J Chem Inf Model 2007;47(2):342–53.
[152] Jhoti H, Williams G, Rees DC, Murray CW. The "rule of three" for fragment-based drug discovery: where are we now? Nat Rev Drug Discov 2013;12(8):644.
[153] Anon. Fragment collections. Available from https://www.cambridgemedchemconsulting.com/resources/hit_identification/fragment_collections.html.
[154] Cox OB, Krojer T, Collins P, Monteiro O, Talon R, Bradley A, et al. A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2){,} an atypical bromodomain. Chem Sci 2016;7(3):2322–30.
[155] Morley AD, Pugliese A, Birchall K, Bower J, Brennan P, Brown N, et al. Fragment-based hit identification: thinking in 3D. Drug Discov Today 2013;18 (23):1221–7.
[156] de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, et al. In silico strategies to support fragment-to-lead optimization in drug discovery. Front Chem 2020;8:93.
[157] Kumar A, Voet A, Zhang KYJ. Fragment based drug design: from experimental to computational approaches. Curr Med Chem 2012;19(30):5128–47.
[158] Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent developments in fragment-based drug discovery. J Med Chem 2008;51(13):3661–80.
[159] Chessari G, Woodhead AJ. From fragment to clinical candidate—a historical perspective. Drug Discov Today 2009;14(13):668–75.
[160] Hung AW, Silvestre HL, Wen S, Ciulli A, Blundell TL, Abell C. Application of fragment growing and fragment linking to the discovery of inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase. Angew Chemie Int Ed 2009;48(45):8452–6.
[161] Cheng Y, Judd TC, Bartberger MD, Brown J, Chen K, Fremeau RT, et al. From fragment screening to in vivo efficacy: optimization of a series of 2-aminoquinolines as potent inhibitors of beta-site amyloid precursor protein cleaving enzyme 1 (BACE1). J Med Chem 2011;54(16):5836–57.

[162] Taylor SJ, Abeywardane A, Liang S, Muegge I, Padyana AK, Xiong Z, et al. Fragment-based discovery of indole inhibitors of matrix metalloproteinase-13. J Med Chem 2011;54(23):8174–87.
[163] Brough PA, Barril X, Borgognoni J, Chene P, Davies NGM, Davis B, et al. Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. J Med Chem 2009;52(15):4794–809.
[164] Hughes SJ, Millan DS, Kilty IC, Lewthwaite RA, Mathias JP, O'Reilly MA, et al. Fragment based discovery of a novel and selective PI3 kinase inhibitor. Bioorg Med Chem Lett 2011;21(21):6586–90.
[165] Villemagne B, Flipo M, Blondiaux N, Crauste C, Malaquin S, Leroux F, et al. Ligand efficiency driven design of new Inhibitors of *Mycobacterium tuberculosis* transcriptional repressor EthR using fragment growing, merging, and linking approaches. J Med Chem 2014;57(11):4876–88.
[166] Istyastono EP, Kooistra AJ, Vischer HF, Kuijer M, Roumen L, Nijmeijer S, et al. Structure-based virtual screening for fragment-like ligands of the G protein-coupled histamine H4 receptor. Med Chem Commun 2015;6(6):1003–17.
[167] Friberg A, Vigil D, Zhao B, Daniels RN, Burke JP, Garcia-Barrantes PM, et al. Discovery of potent myeloid cell leukemia 1 (Mcl-1) inhibitors using fragment-based methods and structure-based design. J Med Chem 2013;56(1):15–30.
[168] De Fusco C, Brear P, Iegre J, Georgiou KH, Sore HF, Hyvönen M, et al. A fragment-based approach leading to the discovery of a novel binding site and the selective CK2 inhibitor CAM4066. Bioorg Med Chem 2017;25(13):3471–82.
[169] Howard N, Abell C, Blakemore W, Chessari G, Congreve M, Howard S, et al. Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. J Med Chem 2006;49(4):1346–55.
[170] Huang S-Y. Search strategies and evaluation in protein–protein docking: principles, advances and challenges. Drug Discov Today 2014;19(8):1081–96.
[171] Porter KA, Desta I, Kozakov D, Vajda S. What method to use for protein–protein docking? Curr Opin Struct Biol 2019;55:1–7.
[172] Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA. Side-chain conformational changes upon protein-protein association. J Mol Biol 2011;408(2):356–65.
[173] Jackson RM, Gabb HA, Sternberg MJ. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276(1):265–85.
[174] Król M, Tournier AL, Bates PA. Flexible relaxation of rigid-body docking solutions. Proteins 2007;68(1):159–69.
[175] Król M, Chaleil RAG, Tournier AL, Bates PA. Implicit flexibility in protein docking: cross-docking and local refinement. Proteins 2007;69(4):750–7.
[176] Grünberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure 2004;12(12):2125–36.
[177] Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331(1):281–99.
[178] Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 2003;12(6):1271–82.
[179] Zacharias M. ATTRACT: protein-protein docking in CAPRI using a reduced protein model. Proteins 2005;60(2):252–6.
[180] Fernández-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins 2003;52(1):113–7.
[181] Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein − protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125(7):1731–7.

[182] Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol 2012;10(1):e1001244.
[183] Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJE, Vajda S, et al. CAPRI: a critical assessment of PRedicted interactions. Proteins 2003;52(1):2–9.
[184] Tsomaia N. Peptide therapeutics: targeting the undruggable space. Eur J Med Chem 2015;94:459–70.
[185] Liao J, Wang Y-T, Lin CS. A fragment-based docking simulation for investigating peptide–protein bindings. Phys Chem Chem Phys 2017;19(16):10436–42.
[186] Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 2014;30(12):1771–3.
[187] Macindoe G, Mavridis L, Venkatraman V, Devignes M-D, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 2010;38(suppl_2):W445–9.
[188] Lee AC-L, Harris JL, Khanna KK, Hong J-H. A comprehensive review on current advances in peptide drug development and design. Int J Mol Sci 2019;20(10):2383.
[189] Agrawal P, Singh H, Srivastava HK, Singh S, Kishore G, Raghava GPS. Benchmarking of different molecular docking methods for protein-peptide docking. BMC Bioinf 2019;19(13):426.
[190] Shin W-H, Seok C. GalaxyDock: protein–ligand docking with flexible protein side-chains. J Chem Inf Model 2012;52(12):3225–32.
[191] Lee H, Heo L, Lee MS, Seok C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. Nucleic Acids Res 2015;43 (W1):W431–5.
[192] Iqbal S, Hoque MT. PBRpredict-suite: a suite of models to predict peptide-recognition domain residues from protein sequence. Bioinformatics 2018;34(19):3289–99.
[193] Ciemny M, Kurcinski M, Kamel K, Kolinski A, Alam N, Schueler-Furman O, et al. Protein–peptide docking: opportunities and challenges. Drug Discov Today 2018;23 (8):1530–7.
[194] Antes I. DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 2010;78(5):1084–104.
[195] Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 2010;78(9):2029–40.
[196] Donsky E, Wolfson HJ. PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics 2011;27(20):2836–42.
[197] Raveh B, London N, Zimmerman L, Schueler-Furman O. Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS One 2011;6(4):e18934.
[198] Ben-Shimon A, Niv MY. AnchorDock: blind and flexible anchor-driven peptide docking. Structure 2015;23(5):929–40.
[199] Weng G, Gao J, Wang Z, Wang E, Hu X, Yao X, et al. Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. J Chem Theory Comput 2020;16(6):3959–69.
[200] Wang M, Yu Y, Liang C, Lu A, Zhang G. Recent advances in developing small molecules targeting nucleic acid. Int J Mol Sci 2016;17(6):779.
[201] Luo J, Wei W, Waldispühl J, Moitessier N. Challenges and current status of computational methods for docking small molecules to nucleic acids. Eur J Med Chem 2019;168:414–25.
[202] Bao L, Zhang X, Jin L, Tan Z-J. Flexibility of nucleic acids: from {DNA} to {RNA}. Chin Phys B 2016;25(1):18703.

[203] Stagno JR, Liu Y, Bhandari YR, Conrad CE, Panja S, Swain M, et al. Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography. Nature 2017;541(7636):242–6.
[204] Guilbert C, James TL. Docking to RNA via root-mean-square-deviation-driven energy minimization with flexible ligands and flexible targets. J Chem Inf Model 2008;48(6):1257–68.
[205] Tessaro F, Scapozza L. How "protein-docking" translates into the new emerging field of docking small molecules to nucleic acids? Molecules 2020;25(12):2749.
[206] Morley SD, Afshar M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®. J Comput Aided Mol Des 2004;18(3):189–208.
[207] Pfeffer P, Gohlke H. DrugScoreRNA—knowledge-based scoring function to predict RNA-ligand interactions. J Chem Inf Model 2007;47(5):1868–76.
[208] Philips A, Milanowska K, Lach G, Bujnicki JM. LigandRNA: computational predictor of RNA-ligand interactions. RNA 2013;19(12):1605–16.
[209] Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 2009;15 (6):1219–30.
[210] Moitessier N, Westhof E, Hanessian S. Docking of aminoglycosides to hydrated and flexible RNA. J Med Chem 2006;49(3):1023–33.
[211] Chen L, Calin GA, Zhang S. Novel insights of structure-based modeling for RNA-targeted drug discovery. J Chem Inf Model 2012;52(10):2741–53.
[212] Chan DS-H, Yang H, Kwan MH-T, Cheng Z, Lee P, Bai L-P, et al. Structure-based optimization of FDA-approved drug methylene blue as a c-myc G-quadruplex DNA stabilizer. Biochimie 2011;93(6):1055–64.
[213] Kaserer T, Rigo R, Schuster P, Alcaro S, Sissi C, Schuster D. Optimized virtual screening workflow for the identification of novel G-quadruplex ligands. J Chem Inf Model 2016;56(3):484–500.
[214] Holt PA, Ragazzon P, Strekowski L, Chaires JB, Trent JO. Discovery of novel triple helical DNA intercalators by an integrated virtual and actual screening platform. Nucleic Acids Res 2009;37(4):1280–7.
[215] Daldrop P, Reyes FE, Robinson DA, Hammond CM, Lilley DM, Batey RT, et al. Novel ligands for a purine riboswitch discovered by RNA-ligand docking. Chem Biol 2011;18(3):324–35.
[216] Lee HS, Zhang Y. BSP-SLIM: a blind low-resolution ligand-protein docking approach using predicted protein structures. Proteins 2012;80(1):93–110.
[217] Grebner C, Iegre J, Ulander J, Edman K, Hogner A, Tyrchan C. Binding mode and induced fit predictions for prospective computational drug design. J Chem Inf Model 2016;56(4):774–87.
[218] Singh J, Petter RC, Baillie TA, Whitty A. The resurgence of covalent drugs. Nat Rev Drug Discov 2011;10(4):307–17.
[219] Li Q, Wang Z, Zheng Q, Liu S. Potential clinical drugs as covalent inhibitors of the priming proteases of the spike protein of SARS-CoV-2. Comput Struct Biotechnol J 2020;18:2200–8.
[220] Liu S, Zheng Q, Wang Z. Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus. Bioinformatics 2020;36(11):3295–8.
[221] Hoffman RL, Kania RS, Brothers MA, Davies JF, Ferre RA, Gajiwala KS, et al. Discovery of ketone-based covalent inhibitors of coronavirus 3CL proteases for the potential therapeutic treatment of COVID-19. J Med Chem 2020;63:12725–47.
[222] Sotriffer C. Docking of covalent ligands: challenges and approaches. Mol Inf 2018;37 (9–10):1800062.
[223] Fanfrlík J, Brahmkshatriya PS, Řezáč J, Jílková A, Horn M, Mareš M, et al. Quantum mechanics-based scoring rationalizes the irreversible inactivation of parasitic *Schistosoma*

mansoni cysteine peptidase by vinyl sulfone inhibitors. J Phys Chem B 2013;117 (48):14973–82.
[224] Kharkar PS, Warrier S, Gaud RS. Reverse docking: a powerful tool for drug repositioning and drug rescue. Future Med Chem 2014;6(3):333–42.
[225] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 2004;3(8):673–83.
[226] Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462(7270):175–81.
[227] Terrett NK, Bell AS, Brown D, Ellis P. Sildenafil (VIAGRATM), a potent and selective inhibitor of type 5 cGMP phosphodiesterase with utility for the treatment of male erectile dysfunction. Bioorg Med Chem Lett 1996;6(15):1819–24.
[228] Zappacosta AR. Reversal of baldness in patient receiving minoxidil for hypertension. N Engl J Med 1980;303:1480–1.
[229] Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. J Cheminform 2020;12(1):46.
[230] Mohamed K, Yazdanpanah N, Saghazadeh A, Rezaei N. Computational drug discovery and repurposing for the treatment of COVID-19: a systematic review. Bioorg Chem 2020;106:104490.
[231] Chen X, Ung CY, Chen Y. Can an in silico drug-target search method be used to probe potential mechanisms of medicinal plant ingredients? Nat Prod Rep 2003;20 (4):432–44.
[232] Erić S, Ke S, Barata T, Solmajer T, Antić Stanković J, Juranić Z, et al. Target fishing and docking studies of the novel derivatives of aryl-aminopyridines with potential anticancer activity. Bioorg Med Chem 2012;20(17):5220–8.
[233] Ji ZL, Wang Y, Yu L, Han LY, Zheng CJ, Chen YZ. In silico search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction. Toxicol Lett 2006;164(2):104–12.
[234] Tanoli Z, Seemab U, Scherer A, Wennerberg K, Tang J, Vähä-Koskela M. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. Brief Bioinform 2020;1–23.
[235] Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D. Sc-PDB: an annotated database of druggable binding sites from the protein data bank. J Chem Inf Model 2006;46(2):717–27.
[236] Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic Acids Res 2002;30(1):412–5.
[237] Xu X, Huang M, Zou X. Docking-based inverse virtual screening: methods, applications, and challenges. Biophys Rep 2018;4(1):1–16.
[238] Schomburg KT, Bietz S, Briem H, Henzler AM, Urbaczek S, Rarey M. Facing the challenges of structure-based target prediction by inverse virtual screening. J Chem Inf Model 2014;54(6):1676–86.
[239] Kellenberger E, Foata N, Rognan D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. J Chem Inf Model 2008;48(5):1014–25.
[240] Luo Q, Zhao L, Hu J, Jin H, Liu Z, Zhang L. The scoring bias in reverse docking and the score normalization strategy to improve success rate of target fishing. PLoS One 2017;12(2):e0171433.
[241] Nogueira MS, Koch O. The development of target-specific machine learning models as scoring functions for docking-based target prediction. J Chem Inf Model 2019;59 (3):1238–52.
[242] Luo H, Fokoue-Nkoutche A, Singh N, Yang L, Hu J, Zhang P. Molecular docking for prediction and interpretation of adverse drug reactions. Comb Chem High Throughput Screen 2018;21(5):314–22.

[243] Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform 2020;22(1):247–69.
[244] Pinzi L, Rastelli G. Molecular docking: shifting paradigms in drug discovery. Int J Mol Sci 2019;20(18):4331.
[245] Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-EM. Nature 2020;587(7832):157–61.
[246] Nakane T, Kotecha A, Sente A, McMullan G, Masiulis S, Brown PMGE, et al. Single-particle cryo-EM at atomic resolution. Nature 2020;587(7832):152–6.
[247] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577(7792):706–10.