Machine learning methods for ligand-protein molecular docking

Kevin Crampon[a,b,c], Alexis Giorkallos[c], Myrtille Deldossi[c], Stephanie Baud[a], Luiz Angelo Steffenel*[b]

[a] *Université de Reims Champagne Ardenne, CNRS, MEDyC UMR 7369, 51097 Reims, France*
[b] *Université de Reims Champagne Ardenne, LICIIS - LRC CEA DIGIT, Reims, France*
[c] *Atos SE, Center for Excellence in Performance Programming, Echirolles, France*

Abstract

Artificial intelligence (AI) is often presented as a new Industrial Revolution. Many domains use AI, and molecular simulation for drug discovery does not escape that logic. This paper proposes a wide overview of ligand-protein molecular docking and how machine learning (ML) and especially deep learning (DL), a subset of ML is transforming it by tackling its challenges. A list of methods that describes and compares the data representation and approaches is presented, and the paper also provides a compilation of performances and discusses them to allow an informed choice.

Keywords: Molecular docking, Sampling, Scoring, Machine Learning, Deep Learning, Data representation

Teaser: ML is transforming *in silico* ligand-protein molecular docking. Our paper details some methods for binding site detection, scoring, or classification to provide a comprehensive overview of this field.

Introduction

Drug discovery is a highly complex process whose most known steps include *in vitro* tests of putative drugs and *in vivo* validation. Before these steps, however, researchers need to perform an extensive evaluation of candidate molecules, from which perhaps a single drug may become a commercialised product. Testing an extensive database, even *in vitro*, is time-consuming and costly. Indeed, the drug discovery process takes on average 12 years from start to commercialisation and has an average cost of $1.8 billion [1].

Researchers and pharmaceutical industries aim at reducing time and cost. They can use the molecular docking process as a complex filter to keep only the most interesting candidates to perform experiments like drug discovery. However, even if molecular docking is often presented in a drug discovery context, it can also be used to detect potential drug side effects or molecule toxicity. The molecular docking process uses the 3-dimensional (3D) structures of two molecules,the ligand and the target, to predict the preferred orientation of the first with respect to the second when bounded to each other to form a stable complex [2]. Usually, the ligand is the smallest molecule, but the denomination choice is project-dependent. In drug discovery, the ligand is an active principle, and the target is a biological macromolecule (for instance, a protein or DNA). However, the docking covers a wider range of pair possibilities: protein-DNA, protein-RNA, protein-sugar, protein-peptide, and protein-small compounds. We decided to focus this survey on the protein-small compounds (commonly called protein-ligand) molecular docking because it covers an important part of existing docking methods. However, we hereafter present concepts for ligand-protein docking, which are also usable for other docking types.

*Corresponding author: Angelo.Steffenel@univ-reims.fr
Preprint version of https://doi.org/10.1016/j.drudis.2021.09.007

Several experimental methods allow getting 3D structures of a molecule. While X-ray-based methods are by far the most prominent ones, nuclear magnetic resonance (NMR) and electron microscopy (EM) based are also part of the literature. For instance, the Protein Data Bank (PDB), a well-known database of protein's 3D structures, includes almost 90% of the structures solved with X-ray crystallography (and almost 99% if only the structures for which the ligand-protein binding affinity is known, are taken into account)[3] and almost 8% with NMR. Information about methods and statistics are available on the Protein Data Bank [4] website (https://www.rcsb.org/) respectively in its educational portal and its statistics page.

As previously mentioned, drug discovery often needs to test against a comprehensive ligand library on one target. This process is called virtual screening (VS) [5] or sometimes high-throughput virtual screening (HTVS). VS is used to reduce the number of tested ligands with *in vitro* and *in vivo* experiments; the ligand's ranking allows to eliminate candidates displaying very low affinities and thus not interesting from a pharmaceutical perspective. Finally, through the VS process, the most interesting molecules are selected and can be proposed for further *in vitro* and *in vivo* testing. VS can be ligand-based, i.e., only ligand information is used: depending on the method, its structure, its chemical properties, or a combination of both are used to predict the binding. The idea is that similar ligands will bind similar targets. VS can also be structure-based, using the complex molecular structure to determine if ligand will bind the target. Molecular docking can be used to perform a structure-based VS campaign. It should be noted that some papers use the molecular docking expression as a structure-based VS synonym.[6, 7, 8] In this article, we deal only with structure-based methods because it is our direct area of expertise.

Detecting the best-bonded ligand among a database is with no doubt the most important use for molecular docking. However, it can also be the basis of research to find new targets (in the case of drug repositioning) or to characterise potential side-effects such as toxicity [9]. In that scenario, protein and ligand roles are switched (*cf.* Figure 1). The authors call this process: inverse docking [10], reverse docking [11], inverse virtual screening [9], or target screening [12]. Additionally, inverse docking processes are performed using classical docking methods structured in a specific pipeline.

In recent years, ML methods such as DL have been implemented to optimise the docking process. This paper aims at presenting the basis of ligand-protein docking and the associated ML approaches. That is why we present classical molecular docking methods and then those based on ML.

The remainder of this paper is organised as follows: we begin with some fundamentals about ligand-protein molecular docking and its challenges, then move to data with ML basics. Next, we present some ML methods with a particular focus on DL. The following section presents and compiles some performance metrics extracted from the methods' papers. Finally, we conclude this survey by discussing the potential of the methods against the molecular docking challenges.

Ligand-protein molecular docking

From 3D structures, a molecular docking[13] experiment can predict the conformation of a complex and its binding affinity. Molecular docking is a combination of two sub-processes. The first step is sampling, which consists of generating a set of conformations from a rigid 3D ligand. The method is evaluated on its capacity to explore the ligand's conformational space. This space gathers all theoretically possible conformations. The second step is scoring, which evaluates the binding affinity of each formed protein-ligand complex (called a pose). Even if sampling and scoring are introduced separately, they can be significantly correlated since scoring functions (SF) often guides the sampling method.

The main challenges for any molecular docking method are dealing with molecular flexibility and faithfully reflecting real binding, both with a reasonable computing time. The following sections present some classical answers (without ML) to these issues. This part aims to present the basis of molecular docking before an in-depth study of ML methods. Figure 2 shows a simplified vision of the molecular

docking process. The remainder of this section presents the three main challenges previously mentioned. It should be noted that this part does not aim at bringing new information about ligand-protein molecular docking, it summarises information to define the molecular docking problem and its current challenges.

First challenge: Molecule's flexibility

In real conditions, the molecules' flexibility is reflected through the vibrations of bounds, angles, and dihedrals. Even though it is an essential element on molecular docking, many of the pioneer methods considered both molecules as rigid structures and used the principle of *lock-and-key* [14] to solve docking problems. New approaches based on heuristics and the improvement of computing capacity allowed the integration of ligand's flexibility by exploring the ligand's conformational space. These methods are called semi-flexible because only the smallest molecule is considered flexible, while the target is still rigid [15]. Progressively, other methods have been developed and allowed to consider both molecules as flexible. Hence, the target's flexibility can be considered in different ways [16]: the target's conformational space can be assessed with extensive sampling (through molecular dynamic, for instance), and relevant structures can be selected to perform numerous rigid targets docking experiments. Another strategy consists of considering the side-chain flexibility of the residues around the binding site. Since it is hypothesised that the presence of the ligand induces these changes, it is known as *induced-fit* [16].

The ligand conformational space sampling

A molecule can have several degrees of freedom: three to describe its position, three to its orientation, and the last to characterise its intrinsic flexibility regarding rotatable bounds or dihedrals. All of which generate the conformational space. Still, today exploring this space is computationally infeasible even for a small compound. Thus, a wide range of sampling methods exists, each of which attempts to optimise its exploration and to find the best conformations. Sampling methods can be classified into shape matching, systematic, stochastic, and simulation [17, 18]. Table 1 presents several docking software and their associated sampling techniques. This table shows that several docking software have been created since the 1990s and, within these programs, many sampling methods have been implemented. Even today, stochastic methods are the most used and regroup a broad panel of methods.

Shape Matching methods

Shape Matching is a method used by the first docking program, DOCK [19]. Such techniques represent molecules (the ligand and the receptor) with geometrical shapes like spheres or polyhedrons and use the principle of matching or complementary shapes to find new conformations. However, since it does not consider the internal ligand flexibility, a solution is to generate ligand conformations right before the search [16].

Systematic methods

Systematic methods allow for quantitative exploration of the ligand's conformational space. Iterative methods (IM) belong to the first category and attempt to generate all ligand's conformations, starting from a given conformer. All degrees of freedom are explored, and a given increment controls the size of the sampling. The generated conformational space may be huge even for a small ligand, its exploration becoming a brute force exploration.

Database methods are among the second category, where databases of conformers such as Flexibase [20] are used. These databases contain, for each ligand, a set of conformations, and instead of computing all possible geometries, it favours a communication with a database holding precomputed conformations. Thus, the computing time is reduced at the expense of important storage space for the databases.

Finally, fragment-based methods (FB) [21] can be used to search for the best conformation, either through *place-and-join* strategies or *incremental* strategies. *Place-and-join* methods cut the ligand into

fragments and place them around the target site. Then, each fragment is moved to minimise its energy, and finally, all fragments are joined to rebuild the ligand. Instead, *incremental* methods place the first fragment, minimise its energy and then add the next fragment, which is also minimised. The process is repeated until the ligand is fully rebuilt. Note that the ligand cutting may bring uncertainty in the final ligand pose. Indeed, energy minimisation may differ between an isolated piece of ligand and the whole molecule. In the end, the rebuilt poses may sum all ligand pieces' imprecision.

Stochastic methods

Unlike systematic methods, stochastic methods are used to explore only a small part of the ligand's conformational space. These methods use pseudo-random functions to generate conformations and SFs to guide them in their educated exploration of the conformational space. The most used methods are Monte-Carlos (MC) [22], Ant Colony (AN), Genetic Algorithm (GA), and Particle Swarm Optimisation (PSO) [23]. The choice of the hyperparameters strongly influences the stochastic methods and, because of this choice, some relevant areas may be forgotten.

Simulation methods

These methods explore the molecule's conformations thanks to computed simulations like molecular dynamics. Simulation methods use classic physics laws like Newton's law to simulate atomic and molecular motions and generate new conformations. For instance, De Azevebo [24] uses the program GROMACS [25], a molecular dynamic solution. Simulation-based methods notorious drawback is the compute time to explore the conformational space, which is why these methods mainly complement other methods [17].

Second challenge: The binding scoring

Ranking of the ligand's bound conformation is managed, for all software, with a scoring function. Most often, the SF aims to estimate the free energy of binding. Since computing the exact value of this energy is computing-intensive, SFs can be designed to produce a score accurate enough to be used in docking simulation, allowing for many evaluations. Besides, as mentioned previously, SFs can be used to guide sampling algorithms. Among the different classes of SFs, we can cite historical and hierarchical families [17, 26, 27] as follows: physics-based, empirical, knowledge-based, and consensus. But first, let us review the mathematical foundations of scoring functions: the scoring function space. Some software's SF, and standalone SF are presented in Table 1. Moreover, since 2010, a new class of ML-based SFs methods appeared (we detail this class in *Machine learning for ligand-protein molecular docking* part).

The scoring function space

A SF aims to find the ligand's conformation that binds best with a given protein. A protein may be seen as a protein space item, and the first protein space definition deals only with sequence [28]. However, regarding our problem, the best definition is 'a space containing all protein folds, where similar structures are close together' [29]. Hence, the ligand can be considered as a chemical space item that gathers all small compounds [30].

Each complex is a set composed from an item of the protein space and an item of the chemical space. Finally, a third space is called the scoring function space, which contains all possible scoring functions. It assumes that at least one scoring function space item can predict the binding affinity between a protein space's structure and a compound from the chemical space. Computational methods (of all kinds) allow exploring this space looking for the best SF for the considered protein and chemical subsets.

Physics-based scoring functions

The physics-based family was first introduced by Li *et al.* [27] to gather different SF types, whose the most famous is the force field class. This subclass of SFs estimates the free energy with a weighted

sum of several energy terms. Which selection depends on the chosen force field. The most common energy terms are Van der Waals, electrostatic interactions, and hydrogen bonds. A large variety of force fields is available in the molecular modelling community: AMBER [31], GROMOS [32], OPLS [33], and CHARMM [34] are examples. Force field-based SFs can be designed using a single or a combination of different force fields. Force field functions are often used for their accuracy related to the utilisation of atomic distances and separate computing of bound and unbound complex energies, such as implemented in AutoDock4 [35]. The physics-based family is also composed of solvent models and quantum mechanics classes. The former adds solvation/desolvation effect and torsion entropy to classical force-field terms [36].In contrast, the latter mixes quantum and molecular mechanics to improve SF accuracy in a reasonable computing time [37]. In their paper, Li *et al.* found that quantum mechanics-based SFs are the most trend physics-based subclass today [27].

Empirical scoring functions

Similar to force field-based methods, empirical methods estimate the free energy of binding but without massive computing. This estimation is achieved by evaluating a weighted sum of parameters such as the number of hydrogen bonds, hydrophobic/hydrophilic contacts, etc. These parameters are simpler than force-field parameters thus also quicker to compute.

Knowledge-based scoring functions

Knowledge-based SFs rely on the elaboration of a potential of mean force [38]. Based on the statistical analysis of intermolecular interactions within large 3D structural databases of complexes, the score attributed to a new complex considers that intermolecular interactions between certain types of atoms or functional groups are more probable than others.

Consensus scoring functions

Each choice comes with a set of compromise, some SFs perform better on an entire class of complex but poorer on another. Through the combination of different types of SFs, consensus SFs aim at optimising their respective advantages. This association can be operated in different ways such as number-by-number, rank-by-number, vote-by-number, or just through a linear combination [26].

Third challenge: The computing time

The computing time is a key metric for both sampling (huge space to explore) and scoring (invoking occurrences). In both cases, the choice of algorithm and its implementation are critical. Regarding the sampling, a way to reduce computing time consists of docking the ligand on a delimited zone of the protein surface (for instance, a cube with a 20Å edge centred on a specific point of interest is regularly used). The knowledge of the localisation of the interaction site on the target is therefore crucial and is very often related to biological results. The drawback associated with this method lies in the fact that there is no possibility to generalise the results to uncharacterised or different systems easily. Indeed, if known ligands are bound on a particular target's site, there is no insurance that new ligands bind in the same site. Similarly, a localised search could not be transposed to a new target. Some studies use more demanding docking simulations without any *a priori* knowledge and explore the whole target's surface to overcome these limitations: this process is called blind docking. The choice of a delimited area has a big impact on the docking accuracy: if the box does not contain the binding site or just a part of it, then the docking will be erroneous.

Furthermore, some methods, namely binding site detection, allow the use of delimited search on target without *a priori* by predicting some putative binding sites on the target surface. Commonly, this search is done either by a geometrical search like FPocket [39] or by looking for the most interesting zones regarding the free energy of binding, like in Q-SiteFinder, which used a $-CH_3$ probe to detect them [40]. It should be noted that usually, no information about the ligand is necessary. Another way

to reduce time is to use integrative docking methods, which integrate experimental data to drive their model [41].

Finally, another way to accelerate computing is by benefiting from a High-Performance Computing (HPC) environment. Even if this avenue is independent of the docking software, it should not be dismissed. That is why we recently developed a parallel approach called AMIDE [42] (for Automatic Molecular Inverse Docking Engine), initially intended for inverse docking and fitted for a classic approach. AMIDE is based on AutoDock4 and its default SF, and includes a set of scripts allowing parallel execution on HPC environments.

Data

Data play a key role in molecular docking method development and even more for ML-based methods. Data quantity, quality, and how the model represents it significantly impact performance and accuracy. Regarding data volume, the previously presented Protein Data Bank provided an extensive database of molecular complexes.

Data quality

When developing ML models for molecular docking, it is important to train and validate the models over established datasets instead of synthetic or augmented datasets. This guarantees representativeness, exhaustiveness and variety for the training set, and allows for inter-method comparison on objective criteria. . The most common datasets are listed below:

- PDBbind [43] is based on Protein Data Bank and updated each year with new complexes; each new version comes with three sets of different sizes: General (21 382), Refined (4 852), and Core (285) for the 2019 version.
- Directory of Useful Decoys (DUD) [44] and DUD-E (for Enhanced) [45] contains 40 (resp. 102) target molecules and 2 950 (resp. 22 886) active ligands. Each ligand has 36 (resp. 50) decoys that are physically close but topologically different.
- The Maximum Unbiased Validation (MUV) [46] dataset contains 17 targets, ligands (30 per target), and decoys (50 per ligand). It is based on the National Institute of Health (NIH) PubChem.
- The Community Structure-Activity Resource (CSAR) [47] is a docked-complex database.
- SC-PDB [48] is a database based on the PDB, but contrary to the previously mentioned dataset, it also contains information about protein binding sites.

Data representation

Data representation is a central piece of the data science response to a specific problem. Throughout, data has become more detailed and incorporates increasingly complex pieces of information. The choice of representation type has a significant impact on docking performances. Even though 3D coordinates can be directly used as input, methods often use other representations produced from 3D coordinates. We can cite lists of descriptors, molecule fingerprints, or interaction fingerprints, image-based, or graphs. The following sections present these different approaches.

A set of descriptors is surely the easiest way to represent a molecular complex. A descriptor is a hand-engineered feature characterising a variable degree of fidelity a complex or a molecule. Descriptors can also reflect physio-chemical properties such as a list of atoms of some types, the number of atoms pair between the ligand and the target for a given threshold, or an energy term. Descriptors can also be geometrical if they are derived from the molecule's 3D structure. Finally, a combination of several of the previously mentioned descriptors is usually used to represent a complex. This descriptors class

has the advantage to be easily understandable and usable, but the descriptors can only represent the complex as a unique object, limiting the model performances.

Fingerprints are a high-level representation of molecules or complexes. The first category relies on the molecular fingerprint where 3D data is squashed into 1D data, commonly a string of bits, integers, or characters. The chemical formula is not detailed, while the structural formula has more details but may be less suitable from a computing standpoint. Fingerprints can represent 2D structures like the MACCS molecular fingerprint, which accounts for some additional chemical properties [49] or encode the 3D structure like FuzCav, which represents the protein binding site 3D structure combined with some chemical properties [50]. DL can also be used with Molecular Surface Interaction Fingerprinting (MaSIF) method to encode a protein [51]. The second category is based on interactions fingerprinting, the most iconic of which are Structural Interaction Fingerprint (SIFt) [52] and Structural Protein-Ligand Interaction Fingerprint (SPLIF) [53]. Fingerprints bring a better complex description abstraction. Contrary to a set of descriptors that list some chemical/geometrical properties, a fingerprint projects the elements to a latent space more suited for ML. In some ways, they behave like autoencoders, a well-studied class of dimensionality reduction algorithms [54] used to reduce the input dimensions.

The emergence of DL and particularly convolutional neural networks (CNN) have made possible the use of a new kind of data representation in the form of its actual 3D structure. Complexes are first discretised on a 3D grid, where each cell of the lattice is a voxel (volumetric pixel). Atoms are sparsely distributed in the lattice. Additionally, voxels have channels (say RGB for images) that can complement the set of features with properties such as atom type, charge, hybridisation, etc. Image-based data representations better reflect the complexity, including the 3D structure, than a classic fingerprints method. Moreover, even if a lot of information is integrated into this representation, it is still concise. However, the main drawback is that this data representation is very sensitive to noise because a slight rotation (nudge) of the molecule in one direction would result in a completely different data point. Moreover, the discretisation of atoms' coordinates may involve a loss of accuracy on the molecule's conformation. These issues can be partially fixed: with data augmentation for the first problem, and a coarser molecular representation (considering residues rather than atoms) for the latter.

Graph data can circumvent previous representation limitations because they break the absolute character of the frame of reference to a relative, more flexible one. Additionally, considering the unstructured nature of its data, graphs are a natural way to represent molecules. A general formalism allows filling nodes, edges, and global properties with all sorts of attributes. Even though molecule representation as a graph is more instinctive, the authors in [55] use a graph to represent the interaction between a ligand and a target.

Machine learning for ligand-protein molecular docking

ML can bring new strategies for complex's scoring: either by optimising an existing SF (refine empirical function's weights, for instance) or by developing a new SF taking a complex's structure as input, the studied scoring approach hereafter [56]. Moreover, ML is sometimes used for virtual screening (classification mode) and binding site detection. Once the dataset is chosen and the data representation is decided, the ML model can be developed. The use of ML in molecular docking evolve very rapidly, and the last decade saw the emergence of a great number of these methods, all bringing significant improvement.

This section is intended as a comprehensive overview of ML methods used in the context of ligand-protein molecular docking, presenting some functions used for scoring, classification (VS mode), and binding site detection. Some existing works [57, 58, 59, 60] already draw a comprehensive overall picture of the domain, and Tables 2 and 3 recall some of the elements specific to ML and DL methods, respectively. Although the ML renaissance is more than a decade old, ML methods have been introduced in the field of molecular docking only in recent years. Therefore, we have classified the methods according

to their type. Furthermore, we can note that this classification is more or less chronological for the DL methods (since MLPs).

0.1. Linear Regression

The most basic use of ML is linear regression, to find linear equation's weights. For instance, TABA (Tool to Analyze the Binding Affinity)[61] represents a ligand-protein interaction as a set of mass-spring contacts and then uses ML methods to parametrise the complex's affinity equation.

Random Forest (RF) methods

Random Forests (RF) were the first attempt at using ML methods for molecular docking. A RF is an ensemble method that builds upon and smooths the results of an ensemble of decision trees. Each tree is built with nodes representing a split on a single and unique criterion. Additionally, training on different randomised subsets reduces variance, thus improving with overfitting issues, a method called bagging.

The first version [62] of RF-Score takes a set of descriptors as input that describe the number of atom pairs from both molecules involved in the docking. Pairs are conserved if the distance between two atoms is less than a certain cut-off (which is a hyper-parameter), and atoms belong to one of this type: C, N, O, S, P, F, Cl, Br, I, for a total of 36 descriptors. RF-Score was updated twice, and the last revision uses energy terms from AutoDock Vina's SF to improve its complex description [63]. All three versions use a set of 500 trees to perform their models. In 2017, the same set of models had been trained against the DUD-E dataset under the name RF-Score-VS [64] to classify complexes instead of scoring them. More recently, Yasuo *et al.* [65] introduced SIEVE-Score (for Similarity of Interaction Energy VEctor-Score). Contrary to RF-Scores, SIEVE-Score performs the search on a thousand random trees and uses a residues level representation: for each of the targets' residues, three interaction energies with the ligand (Van der Waals, Coulomb, and hydrogen-bond) are computed. A complex is represented by a vector of size $3 * n_{res}$ where n_{res} is the number of residues, called the complex's interaction fingerprint. This method is simple and powerful yet still problematic: variable-length input vectors tend to be limiting for many ML models.

Gradient Boosting trees method

In Gradient Boosting, sub-models are trained sequentially instead of simultaneously and from a residual set of its predecessor. It is a form of knowledge distillation and often shows better results than standard bagging.

In 2019, Nguyen *et al.* proposed the Algebraic Graph Learning Score (AGL-Score) [66], which uses a multi-scale, multi-class weight-coloured sub-graphs data representation. The entire molecule is a graph in which nodes' attributes express the type of a selection of atoms along with spatial positioning, and edges represent non-covalent bonds such as Van-der-Waals or hydrogen bond between the connected atoms. Once the graph is built, a series of descriptive statistics is produced from the adjacency matrix's eigenvalues (or the Laplacian matrix) and used as the input vector to train a Boosted Tree.

Support Vector Machines (SVM) methods

Support Vector Machines (SVMs) were a popular class of ML algorithms before DL took off. First introduced for classification problems, SVMs were then adapted for regression (SVRs). The model intends not only to draw a separation between classes but also to maximise the margin between elements closest to its centre. Combined with kernel methods, they are a tool capable of solving non-linear problems.

In [67], Li *et al.* introduce two models of this kind. The first is based on a knowledge-based pairwise potentials vector (SVR-KB). The other approach takes a set of physicochemical (Van der Waals energy, ratio of ligand buried solvent-accessible surface area, hydrophobic effect) descriptors

as input (SVR-EP). ID-Score [68] is another SVR for scoring. This method is based on the same representation as SVR-EP only with additional descriptors, such as, for instance, metal-ligand bonding interaction or desolvation effect. Finally, PLEIC-SVM is a SVM for specific-target VS that does not rely on descriptors but on an embedding of a fingerprint called the Protein-Ligand Empirical Interaction Components (PLEIC) fingerprint [69]. Three values are computed for each target's residue: Van der Waals interaction, hydrophobic contact, and hydrogen bonding. Then all residue feature vectors are concatenated to produce the complex's feature vector used as input by a SVM.

Multi-Layers Perceptrons (MLPs) methods

Multi-Layers Perceptrons (MLPs) are the first deep neural network topology in history and are inspired by the Perceptron. They consist of stacks of layers made of a series of units, all connected from layer to layer.

NNscore v1 [70] was the first attempt to bring artificial neural networks to molecular docking. It is a simple feed-forward MLP with an input vector of 194 features (including basic pairwise atom binding, energy terms, or the number of rotatable binds), a single 5-units hidden layer, and a classification output layer ('good' or 'poor' binder). A year later, its v2 [70] made use of energy terms from Vina's SF as primary descriptors and added features from BINding ANAlyzer (including v1's descriptors) [71]. In addition, the network is rewritten to deal with regression (one output neuron), having a better capacity (hidden layer pushed to 10 neurons). In 2020, Gentille *et al.* introduced Deep Docking [72]. The labels are produced by performing molecular docking on a subset of the ZINC15 [73] ligand database with a specific set of proteins. As there is no mention of the network topology, a set of physicochemical descriptors are used instead. Deep Docking takes the Morgan fingerprint [74] to represent the ligand's molecular structure. Deep Docking trains its network on the previously mentioned subset of ZINC15 and classifies the other ligands between two classes (binder and non-binder).

Convolutional Neural Networks (CNNs) methods

Convolutional Neural Networks (CNNs) are made of convolutional layers and a tool to catch spatial correlations. Filters weights are learned from sliding across the layer input to build a relevant abstract representation of the original data.

AtomNet©[75] is a commercial molecular docking software and one of the firsts to rely on CNN. It makes use of a 3D grid, where each cell represents some basic structural features (atom types or SPLIF, SIFt fingerprints, for example). The network's input is a vectorised grid of 20Å edge and 1Å spacing, then four convolutional layers, followed by two hidden layers of 1024 neurons. In the end, a logistic regression classifies the input between two classes. For DeepVS, also a CNN, Pereira *et al.* [76] define the initial atoms feature set with a context (atom types, atomic partial charges, amino-acid types, and the distances to neighbours) for each complex's atom. To compensate for variable input size, the network incorporates an embedding in the form of a lookup table. The resulting vector is a fixed-size float array that summarises input data. Then, it is processed by a single 2D convolutional layer to extract abstract information and two classic layers to produce a classification. Ragoza *et al.*, in 2017, introduces a CNN-based SF [77] that works on similar 3D-grid images. The novelty here is that each atom is represented by an uncertainty distribution around the atom's centre instead of a fixed value. The network is a succession of 3 blocks (convolution and pooling) followed by a fully connected (FC) binary classification layer.

Atomic CNN [78] is built from two types of unique operations: atom-type-specialised convolutions of 1×1 filters and radial pooling that filters across the atom neighbours. This approach uses atom coordinates and atom types as inputs, the first allows to build the interatomic distance matrix, and the second is used for the atom types matrix preparation. The first layer (atomic convolution) combines matrices to another, and then the radial pooling layer is used to reduce the matrix's dimension. Finally, an *Atomistic FC* layer flattens the feature volume (signature vector), followed by two FC layers

producing a final regression output. While previous methods focused exclusively on binding scoring or classification, DeepSite [79] aims at finding potential binding sites. The input protein 3D grid is augmented along the channel axis with 8 physicochemical descriptors, and the network is a standard CNN (3D convolution followed by MaxPooling), leading to a regression score of potential. Imrie *et al.* DenseFS [80] combines Ragoza's data representation and a particular kind of skip-connection network called a Densely Connected Convolutional Network (DenseNet) [81]. In 2018 Stepniewska-Dziubinska *et al.* designed Pafnucy [82], a classic CNN built to estimate affinity between a ligand and a target from an initial 4D tensor (3D coordinates discretised on a 3D grid and 19 features). The network is composed of three convolutional layers followed by three FC layers that produce a binding score. DeepAffinity [83] is another unusual network engineered around Recurrent Neural Networks (RNN) for scoring, taking a SMILES representation of the ligand, while the target embedding is a string called Structural Property Sequence. Both terms are then independently milled into a sequence-to-sequence (autoencoder) model, their latent vectors processed each by a 1D convolutional layer, then concatenated before a FC layer produces the affinity score..

In DeepBindRG [84], Zhang *et al.* cleverly flattens the input complex into a projected 2D image and performs Residual Network (ResNet) [85] computations to produce an affinity score. Later in 2019, Zheng *et al.* OnionNet [86] suggested a Multiple-Layer Intermolecular-Contact, where a series of shells is built around a central atom. Inside each onion layer, we find a relevant feature set (depending on its encapsulating atoms). This allows the authors to account for non-local interactions. Eight atom types (leading to 64 pairs) and 60 shells are stacked for a total of 3840 features. The model is three convolutional layers followed by three FC CNN. FRSite for faster *R-CNN site predictor* [87] was developed to predict protein binding sites, it takes a 3D grid with 8 commonly used channels to represent the target. The authors employ a particular 3D CNN adapted from the Faster R-CNN [88]. This network is split into three sub-networks: the first part is a 3D CNN feature extractor, whose output is fed to the second and third parts of the network. The second part is a *Region Proposal Network-3D* allowing to extract putative binding sites. Finally, the outputs of the first and the second parts are given to the third one to classify the resulting sites.

Francoeur *et al.* [89] extended the work from Ragoza *et al.* [77], taking the same input data representation and the same general model architecture but performing a comprehensive hyper-optimisation, ending up with more convolutional layers and average pooling instead of max-pooling. The authors from Pafnucy [82] also worked on binding site detection with the same protein representation used in Pafnucy and proposed Kalasanty [90]. The protein is discretised on a 3D grid, and 18 descriptors are used for each atom. Taking inspiration from semantic image segmentation, Stepniewska-Dziubinska *et al.* used a U-Net [91] to identify potential binding sites. The Kalasanty data representation has inspired the DeepSurf [92] authors who have adapted it. Instead of discretising all the molecule atoms as was made in the original paper, the authors selected only a few points of interest from a *solvent accessible surface* mesh. Then each point neighbourhood is discretised on a 3D grid with the same features as Kalasanty. Finally, the model is Bottleneck 3D-LDS-Resnet, itself an evolution of ResNet [85].

Graph Neural Networks (GNNs) methods

Graph Neural Networks (GNNs) is a variety of neural network that works on graph-formatted data. They have evolved from spectral methods to a much more flexible comprehensive modelling tool. Graph Convolutional Networks (GCNs) are a particular class of GNNs, applying convolution and pooling operations from CNNs to graphs.

The first molecular docking method to use graph data was PotentialNet [93]. Instead of just covalent bonds, authors consider additional bonds with one adjacency matrix for each bond type, concatenated along the channel resulting in a 3D adjacency matrix. Moreover, they use a distance matrix that indicates the distance between each atom pair. The network is a GCN split into three stages: on the first stage, only covalent bonds are used for the propagation, then both covalent and non-covalent bonds

are used for the propagation, and finally, a 'graph gather' step, which gathers matrix's rows by summing, is followed by a FC layer used to produce a binding score. In 2019, Lim *et al.* [94] introduce a Graph Neural Network (GNN) with a Gated-Augmented Attention Layer (GAT). For each node, in addition to regular edges, atoms in a close neighbourhood (5Å) are also connected. This method works on three matrices: the first is the node features matrix, the second is the adjacency for only covalent bonds (in the ligand and the protein), and the third is adjacency for inter-molecular interaction (the second matrix is included in it). At each step of the network, the node feature matrix is on the one hand updated by a GAT, and the second matrix is, on the other hand, updated by another GAT, which uses the third matrix. Then the second updated node features matrix is subtracted from the first. After some steps, all node feature vectors are summed, and a FC layer uses this vector to classify the complex.

Torgn *et al.* propose a VS method that uses two graphs to represent the target and the ligand [95]. On the target side, graph nodes are the residues (restricted to the binding site), edges connect every neighbour in a sphere of 7Å, and the features are extracted from the FEATURE program [96]. The ligand graph is a classic 2D molecular graph. Then training is a two-steps process: the first encodes the binding-site graph (dimensionality reduction). The encoder is kept for the second step, which concatenates its output to a second GCN trained on ligand graphs. The result is fed to the FC layer and a Softmax classifier. Tanebe *et al.* [97] use GNNs to classify good or bad binders. This work represents the ligand by a graph generated from SMILES string where nodes are atoms and edges are bonds. The target is a graph where nodes are residues, and edges types (5 in total) depend only on the distance between C_α of each residue. A GNN then embeds both graphs, and the resulting concatenation is used to classify the complex. In the Tsubaki *et al.* method [98], the authors use the SMILES representation of the ligand to produce a graph and then use a GNN to embed this graph into a vector. For the target, the amino acid sequence is embedded by a CNN. Both are concatenated, and an FC followed by Softmax makes a prediction. Recently Morrone *et al.* proposed a new DL method for the docking problem [55] using GCN. This method uses two graphs as input. The first represents the covalent ligand graph (L). The second graph is a contact graph built by hopping from the protein atoms to the ligand atoms in a 4Å neighbourhood (LP). This modular method can take L, LP, or L+LP as input. In any case, the input is embedded by a GCN and fed to a CNN for prediction.

Networks' architecture comparison

The DL family is divided into three classes: MLPs, CNNs, and GNNs. We presented three MLPs, but only two describe their architectures, which is not enough to perceive an evolution of MLP's architecture. Moreover, both with known architecture are just different versions of the same method.

Regarding CNNs, a large variety of networks (ResNet, UNet...) bring many architectural possibilities. For the CNNs, the main architecture is adapted from 3D grids (3DCNN), like AtomNet or Pafnucy. The topologies are not identical, but they use the common CNN layers. However, some methods propose new original architecture like the Atomic CNN or use well-known architecture like Kalansaty, which is based on UNet, a network historically created for image segmentation. Because of the black-box nature of DL-based models, it is difficult to assert the superiority of a CNN topology with regards to its counterparts: it heavily depends on the chosen data representation.

In the case of GNNs, the limited number of available methods does not allow highlighting a particular architecture. It should be noted that some methods use GNNs as a first step to embed the inputs and then use FC layers or a CNN to produce an output.

Performance measurement

So far, we have been focusing on modelling only, but as mentioned previously, datasets are not only dedicated to training but also to evaluate and assess the methods. Therefore, this section aims to present usual performance metrics for classification methods (VS), SF, and binding site detection.

VS assessment

In addition to datasets, authors also use a series of metrics to compare to other existing contributions. In the case of VS, the model is evaluated on its capacity to distinguish between binding and non-binding ligands. Generally, the Enrichment Factor (EF) or the Area Under the Curve (AUC) of the Receiver Operating Characteristic curve (ROC) are used. EF evaluates if selected ligands are better binders than randomly selected ones and takes only real positive values: a poor classifier has $EF =< 1$, a better-than-random one has $EF > 1$. This metric allows comparing the rate of true binders among the top-ranked ligands (in top $\{1\%, 2\%, 5\%, 10\%\}$) to that rate in a random selection. The ROC curve, on the other hand, is used to visually assess the quality of a classifier as its discrimination threshold varies. The best AUC value is 1.0 and 0.0 in the worst case (random case being 0.5).

Table 4 gathers AUC performances from various methods including those drawn from [58, 59]. This table shows the difficulty to draw a simple conclusion about VS performances. The dataset is the first problem that arises when trying to compare pure performance. Second, most methods are not self-supporting and require the adjunction of other classic sampling software. So, even if two methods are assessed on the same dataset, their performances are heavily impacted by the chosen sampling method. Moreover, even though the sampling method is theoretically the same, they may differ on parameter initialisation, as explained by Shen *et al.* [59]. Consequently, we are condemned to use raw performances given in the same paper to compare some methods. For example, Lim's method is better than AtomNet and Ragoza's method, according to Lim *et al.* [94].

SF assessment

Su *et al.* Comparative Assessment of Scoring Functions (CASF) [99] introduced three criteria to assess a SF method: scoring power, ranking power, docking power. In more details:

- **Scoring** reflects the ability for a SF "to produce binding scores in a linear correlation with experimental binding data". It uses Pearson's correlation coefficient (R_p) (sometimes R_p^2) and the standard deviation in linear regression (SD). Pearson coefficient can take its value between -1 and $+1$. The closest it is to 1, the better is the method assessed. For SD, the smallest value is best. R_p and, to a lesser extent, SD are the most used criterion.

- **Ranking** refers to the ability for a SF "to correctly rank the known ligands of a certain target protein by their binding affinities when the precise binding poses of those ligands are given". The assessment uses Spearman's rank correlation coefficient (ρ), Kendall's rank correlation coefficient (τ) and, the Predictive Index (PI). These criteria have values in $[-1, 1]$, for all $+1$ indicate a perfect ranking and -1 a reverse [99].

- **Docking** represents the ability for a SF "to identify the native ligand binding pose among computer-generated decoys". Assessment uses RSMD (eq 1) to compare top-ranked ligand by the SF method and native ligand pose. A threshold is used to consider docking as a success. The success rate is used to judge the method as a success. Commonly the used cutoff is 2.0Å.

$$RMSD = \sqrt{\frac{\sum_{i=0}^{n}(X_E^i - X_S^i)^2}{n}} \tag{1}$$

  Where E is expected atoms' coordinates, S is simulated atom's coordinates, and n is the number of ligand's atoms.

Note that the CASF dataset is identical to the PDBbind core set for the corresponding year. Table 5 lists some method assessments for scoring power. Contrary to Table 4, Table 5 items are more comparable: to assess a SF, authors used a dataset of docked complexes. Consequently, the sampling

step is unnecessary, and thus assessments differ only by the used dataset. However, a wide dataset variety is available, each with several subsets and versions (PDBbind for instance). If the used datasets are identical, then their respective performance can finally be compared. For example, OnionNet has a better R_p score than Francoeur's method on the PDBbind 2017 core set.

Binding site detection assessment

To assess binding site detection methods, there are two main options. At first, we can use a dataset of already docked ligand-protein complexes (like PDBbind) and predict binding sites for proteins. Then, for each complex, it is possible to consider the method output as a success if at least one predicted protein site is the real binding site. This approach is interesting if the binding site composition is unknown.

However, all previously mentioned methods used the SCpdb [48] dataset for training and assessment. This dataset contains the sites' atomic composition and, once the predicted sites are defined, atomic compositions can be compared. Authors use two metrics:

- **The distance to the centre of the binding site (DCC)** metric measures the distance either between the centre of the real binding site and the closest atom of the predicted site, the distance or between the centre of the real binding site and the centre of the predicted site. In both cases, the site detection is a success regarding a threshold fluctuating between 4Å and 20Å. Better the success ratio is, the better is the method.

- **The discretised volumetric overlap (DVO)** metric aims to assess the overlapping between the predicted and the real binding site. Authors use the Jaccard Index on the sites' convex hull. Both volumes are discretised, and the ratio between the overlapping volume and merged volume is computed. Closest to 1 the Jaccard Index is, the better is the method.

We do not propose a table gathering binding site detection methods' performances because data are often provided as charts without the raw values. However, most recent methods are compared by their authors to the older ones in their respective papers.

Conclusion

This paper aimed to present how ML and particularly DL can help us to tackle molecular docking challenges. We have presented three challenges: sampling, scoring, and computing time. Regarding the sampling challenge, from the best of our knowledge, no ML method attempts to tackle it.

Concerning the scoring challenge, it is, without doubt, the most studied problem. Indeed, ML scoring methods are very interesting regarding the scoring function space exploration. Many ML methods have been developed for it, and most of them outperform classical methods. Thus, ML SF may be regarded as a hybrid evolution between knowledge-based and empirical functions. Indeed, similarly to knowledge-based, ML methods attempt to extract statistics from a comprehensive database to build the most relevant model. On the other hand, ML methods use relatively simple inputs (see Data representation part) and find links between them. It is even more evident for DL methods that aim to optimise the networks' weights, which is very similar to the goal of an empirical function. Although they are not the main focus for this review, a particular class of ML models called physics-informed DL has a lot of potential because it incorporates physical constraints in the learning process.

This survey showed ML methods outperform the classic ones, whether for scoring or classification. Moreover, recently proposed GNN methods have interesting performances and are still today little-explored. Therefore, we think it is an approach that may be more studied in-depth.

The last-mentioned challenge is about time computing. It should be noted that no ML scoring methods are compared with others regarding time computing. That is why it is harder to discuss the capacity of ML towards time reduction. However, as explained, a delimited search may be used to reduce

time, and some ML methods to predict binding sites are presented in our survey, and they outperform classical binding site detection methods, according to their authors. Therefore, we think GNNs may be an interesting approach to propose a better method.

One general drawback is that almost all methods are not proposed and assessed in a complete docking pipeline; it would be interesting to compare classical ones like AutoDock with ML workflow. Moreover, the training and inference times of ML methods are never mentioned by their authors. We believe this information should be included in more works as it provides invaluable insights on the models' complexity.

Our future works aim to propose a complete pipeline including data preparation, binding site detection, and molecular docking, the last two with ML. Our work will especially focus on the Extra-Cellular Matrix, a complex 3D network that supports cells. This tissue is involved in ageing or some diseases like tumours and a main research target from our teams.

Acknowledgements

The authors would like to thank Prof. Manuel Dauchez from the MEDyC laboratory, for his helpful and fruitful discussions.

This work has been supported by the Association Nationale de la Recherche et de la Technologie (ANRT) through a CIFRE grant.

References

[1] Sinha, Sandeep and Vohora, Divya*Chapter 2 - Drug Discovery and Development: An Overview*:19–32. Elsevier 2018.

[2] Lengauer, Thomas and Rarey, MatthiasComputational methods for biomolecular docking *Current Opinion in Structural Biology.* 1996;6:402–406.

[3] Veit-Acosta, Martina and de Azevedo Junior, Walter FilgueiraThe Impact of Crystallographic Data for the Development of Machine Learning Models to Predict Protein-Ligand Binding Affinity *Current Medicinal Chemistry.* 2021;28:1–17. add-1.

[4] Berman, Helen M and Westbrook, John and Feng, Zukang and Gilliland, Gary and Bhat, Talapady N and Weissig, Helge and Shindyalov, Ilya N and Bourne, Philip EThe protein data bank *Nucleic acids research.* 2000;28:235–242.

[5] Muegge, Ingo and Rarey, Matthias*Small Molecule Docking and Scoring*:1–46. John Wiley & Sons, Inc. 2001.

[6] Shan, Wenying and Li, Xuanyi and Hequan, Yao and Lin, KejiangConvolutional Neural Network-based Virtual Screening *Current Medicinal Chemistry.* 2021;28:2033–2047. add-3.

[7] Bitencourt-Ferreira, Gabriela and Duarte da Silva, Amauri and Filgueira, de Azevedo Jr and othersApplication of Machine Learning Techniques to Predict Binding Affinity for Drug Targets: A Study of Cyclin-dependent Kinase 2 *Current medicinal chemistry.* 2021;28:253–265.

[8] Musella, Simona and Verna, Giulio and Fasano, Alessio and Di Micco, SimoneNew Perspectives on Machine Learning in Drug Discovery *Current medicinal chemistry.* 2021.

[9] Xu, Xianjin and Huang, Marshal and Zou, XiaoqinDocking-based inverse virtual screening: methods, applications, and challenges *Biophysics Reports.* 2018;4:1–16.

[10] Chen, Y Z and Zhi, D GLigand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule *Proteins: Structure, Function, and Bioinformatics.* 2001;43:217–226.

[11] Fan, Jiyu and Fu, Ailing and Zhang, LeProgress in molecular docking *Quantitative Biology.* 2019;7:83–89.

[12] Khan, Abbas and Chandra Kaushik, Aman and Ali, Syed Shujait and Ahmad, Nisar and Wei, Dong-QingDeep-learning-based target screening and similarity search for the predicted inhibitors of the pathways in Parkinson's disease *RSC Advances.* 2019;9:10326–10339.

[13] Sulimov, Vladimir B. and Kutov, Danil C. and Sulimov, Alexey VAdvances in Docking *Current Medicinal Chemistry.* 2019;26:7555–7580. add-2.

[14] Fischer, EmilEinfluss der Configuration auf die Wirkung der Enzyme *Berichte der deutschen chemischen Gesellschaft.* 1894;27:2985–2993.

[15] Leach, Andrew R. and Kuntz, Irwin D.Conformational analysis of flexible ligands in macromolecular receptor sites *Journal of Computational Chemistry.* 1992;13:730–748.

[16] Huang, Sheng-You and Zou, XiaoqinAdvances and Challenges in Protein-Ligand Docking *International Journal of Molecular Sciences.* 2010;11:3016–3034.

[17] Sousa, Sérgio Filipe and Fernandes, Pedro Alexandrino and Ramos, Maria JoãoProtein-ligand docking: Current status and future challenges *Proteins: Structure, Function, and Bioinformatics.* 2006;65:15–26.

[18] Novic, Marjana and Tibaut, Tjasa and Anderluh, Marko and Borisek, Jure*The comparison of docking search algorithms and scoring functions: an overview and case studies*:99–127. IGI Global 2016.

[19] Kuntz, Irwin D. and Blaney, Jeffrey M. and Oatley, Stuart J. and Langridge, Robert and Ferrin, Thomas E.A geometric approach to macromolecule-ligand interactions *Journal of Molecular Biology.* 1982;161:269–288.

[20] Kearsley, Simon K. and Underwood, Dennis J. and Sheridan, Robert P. and Miller, Michael D.Flexibases: A way to enhance the use of molecular docking methods *Journal of Computer-Aided Molecular Design.* 1994;8:565–582.

[21] Taylor, R D and Jewsbury, P J and Essex, J WA review of protein-small molecule docking methods *Journal of Computer-Aided Molecular Design.* 2002;16:161–166.

[22] Metropolis, Nicholas and Ulam, S.The Monte Carlo Method *Journal of the American Statistical Association.* 1949;44:335–341.

[23] Yang, Xin-She*Nature-inspired metaheuristic algorithms.* Luniver press 2010.

[24] F. de Azevedo, W.Molecular Dynamics Simulations of Protein Targets Identified in Mycobacterium tuberculosis *Current Medicinal Chemistry.* 2011;18:1353–1366.

[25] Abraham, Mark James and Murtola, Teemu and Schulz, Roland and Páll, Szilárd and Smith, Jeremy C. and Hess, Berk and Lindahl, ErikGROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers *SoftwareX.* 2015;1–2:19–25.

[26] Oda, Akifumi and Tsuchida, Keiichi and Takakura, Tadakazu and Yamaotsu, Noriyuki and Hirono, ShuichiComparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes *Journal of Chemical Information and Modeling.* 2006;46:380–391.

[27] Li, Jin and Fu, Ailing and Zhang, LeAn Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking *Interdisciplinary Sciences: Computational Life Sciences.* 2019;11:320–328.

[28] Smith, John MaynardNatural selection and the concept of a protein space *Nature.* 1970;225:563–564.

[29] Bitencourt-Ferreira, Gabriela and Azevedo, Walter Filgueira*Exploring the Scoring Function Space*;2053 of *Methods in Molecular Biology*:275–281. Springer New York 2019.

[30] Bohacek, Regine S and McMartin, Colin and Guida, Wayne CThe art and practice of structure-based drug design: A molecular modeling perspective *Medicinal research reviews.* 1996;16:3–50.

[31] Weiner, Paul K. and Kollman, Peter A.AMBER: Assisted Model Building with Energy Refinement. A General Program for Modeling Molecules and Their Interactions *Journal of Computational Chemistry.* 1981;2:287–303.

[32] van Gunsteren, Wilfred F. and Berendsen, Herman J. C.Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry *Angewandte Chemie International Edition in English.* 1990;29:992–1023.

[33] Jorgensen, William L. and Tirado-Rives, JulianThe OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin *Journal of the American Chemical Society.* 1988;110:1657–1666.

[34] Brooks, Bernard R. and Bruccoleri, Robert E. and Olafson, Barry D. and States, David J. and Swaminathan, S. and Karplus, MartinCHARMM: A program for macromolecular energy, minimization, and dynamics calculations *Journal of Computational Chemistry.* 1983;4:187–217.

[35] Morris, Garrett M. and Huey, Ruth and Lindstrom, William and Sanner, Michel F. and Belew, Richard K. and Goodsell, David S. and Olson, Arthur J.AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility *Journal of Computational Chemistry.* 2009;30:2785–2791.

[36] Genheden, Samuel and Ryde, UlfThe MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities *Expert Opinion on Drug Discovery.* 2015;10:449–461.

[37] Chaskar, Prasad and Zoete, Vincent and Röhrig, Ute F.Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function *Journal of Chemical Information and Modeling.* 2014;54:3137–3152.

[38] Muegge, Ingo and Martin, Yvonne C.A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach *Journal of Medicinal Chemistry.* 1999;42:791–804.

[39] Le Guilloux, Vincent and Schmidtke, Peter and Tuffery, PierreFpocket: An open source platform for ligand pocket detection *BMC Bioinformatics.* 2009;10:1–11.

[40] Laurie, A. T. R. and Jackson, R. M.Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites *Bioinformatics.* 2005;21:1908–1916.

[41] Karaca, Ezgi and Bonvin, Alexandre M.J.J.Advances in integrative modeling of biomolecular complexes *Methods.* 2013;59:372–381.

[42] Vasseur, Romain and Baud, Stéphanie and Steffenel, Luiz Angelo and Vigouroux, Xavier and Martiny, Laurent and Krajecki, Michaël and Dauchez, ManuelInverse docking method for new proteins targets identification: A parallel approach *Parallel Computing.* 2015;42:48–59.

[43] Wang, Renxiao and Fang, Xueliang and Lu, Yipin and Wang, ShaomengThe PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures *Journal of Medicinal Chemistry.* 2004;47:2977–2980.

[44] Huang, Niu and Shoichet, Brian K. and Irwin, John J.Benchmarking Sets for Molecular Docking *Journal of Medicinal Chemistry.* 2006;49:6789–6801.

[45] Mysinger, Michael M. and Carchia, Michael and Irwin, John. J. and Shoichet, Brian K.Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking *Journal of Medicinal Chemistry.* 2012;55:6582–6594.

[46] Rohrer, Sebastian G. and Baumann, KnutMaximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data *Journal of Chemical Information and Modeling.* 2009;49:169–184.

[47] Smith, Richard D. and Dunbar, James B. and Ung, Peter Man-Un and Esposito, Emilio X. and Yang, Chao-Yie and Wang, Shaomeng and Carlson, Heather A.CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions *Journal of Chemical Information and Modeling.* 2011;51:2115–2131.

[48] Desaphy, Jérémy and Bret, Guillaume and Rognan, Didier and Kellenberger, Esthersc-PDB: a 3D-database of ligandable binding sites—10 years on *Nucleic Acids Research.* 2015;43:D399–D404.

[49] Durant, Joseph L. and Leland, Burton A. and Henry, Douglas R. and Nourse, James G.Reoptimization of MDL Keys for Use in Drug Discovery *Journal of Chemical Information and Computer Sciences.* 2002;42:1273–1280.

[50] Weill, Nathanaël and Rognan, DidierAlignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites *Journal of Chemical Information and Modeling.* 2010;50:123–135.

[51] Gainza, P. and Sverrisson, F. and Monti, F. and Rodolà, E. and Boscaini, D. and Bronstein, M. M. and Correia, B. E.Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning *Nature Methods.* 2020;17:184–192.

[52] Deng, Zhan and Chuaqui, Claudio and Singh, JuswinderStructural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions *Journal of Medicinal Chemistry.* 2004;47:337–344.

[53] Da, C. and Kireev, D.Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study *Journal of Chemical Information and Modeling.* 2014;54:2555–2561.

[54] Goodfellow, Ian and Bengio, Yoshua and Courville, Aaron*Deep Learning.* MIT Press 2016.

[55] Morrone, Joseph A. and Weber, Jeffrey K. and Huynh, Tien and Luo, Heng and Cornell, Wendy D.Combining Docking Pose Rank and Structure with Deep Learning Improves Protein–Ligand Binding Mode Prediction over a Baseline Docking Approach *Journal of Chemical Information and Modeling.* 2020;60:4170–4179.

[56] Heck, Gabriela S. and Pintro, Val O. and Pereira, Richard R. and de Ávila, Mauricio B. and Levin, Nayara M.B. and de Azevedo, Walter F.Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity *Current Medicinal Chemistry.* 2017;24:2459–2470.

[57] Ain, Qurrat Ul and Aleksandrova, Antoniya and Roessler, Florian D. and Ballester, Pedro J.Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening: Machine-learning SFs to improve structure-based binding affinity prediction and virtual screening *Wiley Interdisciplinary Reviews: Computational Molecular Science.* 2015;5:405–424.

[58] Li, Hongjian and Sze, Kam-Heung and Lu, Gang and Ballester, Pedro J.Machine-learning scoring functions for structure-based drug lead optimization *WIREs Computational Molecular Science.* 2020;10:e1465.

[59] Shen, Chao and Ding, Junjie and Wang, Zhe and Cao, Dongsheng and Ding, Xiaoqin and Hou, TingjunFrom machine learning to deep learning: Advances in scoring functions for protein–ligand docking *WIREs Computational Molecular Science.* 2020;10:e1429.

[60] Abbasi, Karim and Razzaghi, Parvin and Poso, Antti and Ghanbari-Ara, Saber and Masoudi-Nejad, AliDeep Learning in Drug Target Interaction Prediction: Current and Future Perspectives *Current Medicinal Chemistry.* 2021;28:2100–2113. add-6.

[61] Duarte da Silva, Amauri and Bitencourt-Ferreira, Gabriela and de Azevedo Jr, Walter FilgueiraTaba: A Tool to Analyze the Binding Affinity *Journal of Computational Chemistry.* 2020;41:69–73. add-7 / Found.

[62] Ballester, Pedro J. and Mitchell, John B. O.A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking *Bioinformatics.* 2010;26:1169–1175.

[63] Li, Hongjian and Leung, Kwong-Sak and Wong, Man-Hon and Ballester, Pedro J.Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets *Molecular Informatics.* 2015;34:115–126.

[64] Wójcikowski, Maciej and Ballester, Pedro J. and Siedlecki, PawelPerformance of machine-learning scoring functions in structure-based virtual screening *Scientific Reports.* 2017;7:1–10.

[65] Yasuo, Nobuaki and Sekijima, MasakazuImproved Method of Structure-Based Virtual Screening via Interaction-Energy-Based Learning *Journal of Chemical Information and Modeling.* 2019;59:1050–1061.

[66] Nguyen, Duc Duy and Wei, Guo-WeiAGL-Score: Algebraic Graph Learning Score for Protein–Ligand Binding Scoring, Ranking, Docking, and Screening *Journal of Chemical Information and Modeling.* 2019;59:3291–3304.

[67] Li, Liwei and Wang, Bo and Meroueh, Samy OSupport Vector Regression Scoring of Receptor-Ligand Complexes for Rank-Ordering and Virtual Screening of Chemical Libraries *Journal of chemical information and modeling.* 2012;51:2132–2138.

[68] Li, Guo-Bo and Yang, Ling-Ling and Wang, Wen-Jing and Li, Lin-Li and Yang, Sheng-YongID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions *J. Chem. Inf. Model.*. 2013;53:592–600.

[69] Yan, Yuna and Wang, Weijun and Sun, Zhaoxi and Zhang, John Z. H. and Ji, ChanggeProtein-Ligand Empirical Interaction Components for Virtual Screening *Journal of Chemical Information and Modeling.* 2017;57:1793–1806.

[70] Durrant, Jacob D. and Friedman, Aaron J. and Rogers, Kathleen E. and McCammon, J. Andrew-Comparing Neural-Network Scoring Functions and the State of the Art: Applications to Common Library Screening *Journal of Chemical Information and Modeling.* 2013;53:1726–1735.

[71] Durrant, Jacob D. and McCammon, J. AndrewBINANA: A novel algorithm for ligand-binding characterization *Journal of Molecular Graphics and Modelling.* 2011;29:888–893.

[72] Gentile, Francesco and Agrawal, Vibudh and Hsing, Michael and Ton, Anh-Tien and Ban, Fuqiang and Norinder, Ulf and Gleave, Martin E. and Cherkasov, ArtemDeep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery *ACS Central Science.* 2020;6:939–949.

[73] Sterling, Teague and Irwin, John J.ZINC 15 – Ligand Discovery for Everyone *Journal of Chemical Information and Modeling.* 2015;55:2324–2337.

[74] Morgan, H. L.The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation.* 1965;5:107–113.

[75] Wallach, Izhar and Dzamba, Michael and Heifets, AbrahamAtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery *arXiv:1510.02855.* 2015.

[76] Pereira, Janaina Cruz and Caffarena, Ernesto Raúl and dosSantos, Cicero NogueiraBoosting Docking-Based Virtual Screening with Deep Learning *Journal of Chemical Information and Modeling.* 2016;56:2495–2506.

[77] Ragoza, Matthew and Hochuli, Joshua and Idrobo, Elisa and Sunseri, Jocelyn and Koes, David RyanProtein–Ligand Scoring with Convolutional Neural Networks *Journal of Chemical Information and Modeling.* 2017;57:942–957.

[78] Gomes, Joseph and Ramsundar, Bharath and Feinberg, Evan N. and Pande, Vijay S.Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity *arXiv:1703.10603.* 2017.

[79] Jiménez, J and Doerr, S and Martínez-Rosell, G and Rose, A S and De Fabritiis, GDeepSite: protein-binding site predictor using 3D-convolutional neural networks *Bioinformatics.* 2017;33:3036–3042.

[80] Imrie, Fergus and Bradley, Anthony R. and Van der Schaar, Mihaela and Deane, Charlotte M.Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data *Journal of Chemical Information and Modeling.* 2018;58:2319–2330.

[81] Huang, Gao and Liu, Zhuang and Van Der Maaten, Laurens and Weinberger, Kilian Q.Densely Connected Convolutional Networks in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:4700–4708 2017.

[82] Stepniewska-Dziubinska, Marta M and Zielenkiewicz, Piotr and Siedlecki, PawelDevelopment and evaluation of a deep learning model for protein–ligand binding affinity prediction *Bioinformatics.* 2018;34:3666–3674.

[83] Karimi, Mostafa and Wu, Di and Wang, ZhangyangDeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks *Bioinformatics.* 2019;35:3329–3338.

[84] Zhang, Haiping and Liao, Linbu and Saravanan, Konda Mani and Yin, Peng and Wei, YanjieDeepBindRG: a deep learning based method for estimating effective protein–ligand affinity *PeerJ.* 2019;7:e7362.

[85] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, JianDeep Residual Learning for Image Recognition in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:770–778 2016.

[86] Zheng, Liangzhen and Fan, Jingrong and Mu, YuguangOnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction *ACS Omega.* 2019;4:15956–15965.

[87] Jiang, Mingjian and Wei, Zhiqiang and Zhang, Shugang and Wang, Shuang and Wang, Xiaofeng and Li, ZhenFRSite: Protein drug binding site prediction based on faster R–CNN *Journal of Molecular Graphics and Modelling.* 2019;93:107454.

[88] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, JianFaster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks *IEEE transactions on pattern analysis and machine intelligence.* 2016;39:1137–1149.

[89] Francoeur, Paul G and Masuda, Tomohide and Sunseri, Jocelyn and Jia, Andrew and Iovanisci, Richard B and Snyder, Ian and Koes, David RThree-Dimensional Convolutional Neural Networks and a Cross- Docked Data Set for Structure-Based Drug Design *J. Chem. Inf. Model..* 2020;60:4200–4215.

[90] Stepniewska-Dziubinska, Marta M. and Zielenkiewicz, Piotr and Siedlecki, PawelImproving detection of protein-ligand binding sites with 3D segmentation *Scientific Reports.* 2020;10:1–9.

[91] Ronneberger, Olaf and Fischer, Philipp and Brox, ThomasU-Net: Convolutional Networks for Biomedical Image Segmentation *International Conference on Medical image computing and computer-assisted intervention.* 2015:234–241.

[92] Mylonas, Stelios K. and Axenopoulos, Apostolos and Daras, PetrosDeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins *arXiv preprint arXiv:2002.05643.* 2020.

[93] Feinberg, Evan N. and Sur, Debnil and Wu, Zhenqin and Husic, Brooke E. and Mai, Huanghao and Li, Yang and Sun, Saisai and Yang, Jianyi and Ramsundar, Bharath and Pande, Vijay S.PotentialNet for Molecular Property Prediction *ACS Central Science.* 2018;4:1520–1530.

[94] Lim, Jaechang and Ryu, Seongok and Park, Kyubyong and Choe, Yo Joong and Ham, Jiyeon and Kim, Woo YounPredicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation *Journal of Chemical Information and Modeling.* 2019;59:3981–3988.

[95] Torng, Wen and Altman, Russ B.Graph Convolutional Neural Networks for Predicting Drug-Target Interactions *Journal of Chemical Information and Modeling.* 2019;59:4131–4149.

[96] Bagley, Steven C. and Altman, Russ B.Characterizing the microenvironment surrounding protein sites *Protein Science.* 1995;4:622–635.

[97] Tanebe, Toshitaka and Ishida, TakashiEnd-to-End Learning Based Compound Activity Prediction Using Binding Pocket Information in *International Conference on Intelligent Computing*:226–234Springer 2019.

[98] Tsubaki, Masashi and Tomii, Kentaro and Sese, JunCompound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences *Bioinformatics.* 2019;35:309–318.

[99] Su, Minyi and Yang, Qifan and Du, Yu and Feng, Guoqin and Liu, Zhihai and Li, Yan and Wang, RenxiaoComparative Assessment of Scoring Functions: The CASF-2016 Update *Journal of Chemical Information and Modeling.* 2019;59:895–913.

[100] Abagyan, Ruben and Totrov, Maxim and Kuznetsov, DmitryICM-A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation *Journal of Computational Chemistry.* 1994;15:488–506.

[101] Jones, Gareth and Willett, Peter and Glen, Robert C and Leach, Andrew R and Taylor, RobinDevelopment and Validation of a Genetic Algorithm for Flexible Docking *Journal of Molecular Biology.* 1997;267:727–748.

[102] Burkhard, P and Taylor, P and Walkinshaw, M.DAn example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å X-ray structure of a Thrombin-Ligand complex *Journal of Molecular Biology.* 1998;277:449–466.

[103] Terp, Gitte Elgaard and Johansen, Bent Nagstrup and Christensen, Inge Thøger and Jørgensen, Flemming SteenA New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein-Ligand Binding Affinities *Journal of Medicinal Chemistry.* 2001;44:2333–2343.

[104] Venkatachalam, C.M. and Jiang, X. and Oldfield, T. and Waldman, M.LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites *Journal of Molecular Graphics and Modelling.* 2003;21:289–307.

[105] Velec, Hans F. G. and Gohlke, Holger and Klebe, GerhardDrugScore - Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction *Journal of Medicinal Chemistry.* 2005;48:6296–6303.

[106] Zhang, Chi and Liu, Song and Zhu, Qianqian and Zhou, YaoqiA Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes *Journal of Medicinal Chemistry.* 2005;48:2325–2335.

[107] Friesner, Richard A. and Murphy, Robert B. and Repasky, Matthew P. and Frye, Leah L. and Greenwood, Jeremy R. and Halgren, Thomas A. and Sanschagrin, Paul C. and Mainz, Daniel T.Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes *Journal of Medicinal Chemistry.* 2006;49:6177–6196.

[108] Korb, Oliver and Stützle, Thomas and Exner, Thomas E.*PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design*;4150:247–258. Springer Berlin Heidelberg 2006.

[109] Chen, Hung-Ming and Liu, Bo-Fu and Huang, Hui-Ling and Hwang, Shiow-Fen and Ho, Shinn-YingSODOCK: Swarm optimization for highly flexible protein–ligand docking *Journal of Computational Chemistry.* 2007;28:612–623.

[110] Zsoldos, Zsolt and Reid, Darryl and Simon, Aniko and Sadjad, Sayyed Bashir and Johnson, A. PetereHiTS: A new fast, exhaustive flexible ligand docking system *Journal of Molecular Graphics and Modelling.* 2007;26:198–212.

[111] Zhao, Xiaoyu and Liu, Xiaofeng and Wang, Yuanyuan and Chen, Zhi and Kang, Ling and Zhang, Hailei and Luo, Xiaomin and Zhu, Weiliang and Chen, Kaixian and Li, Honglin and et al.An Improved PMF Scoring Function for Universally Predicting the Interactions of a Ligand with Protein, DNA, and RNA *Journal of Chemical Information and Modeling.* 2008;48:1438–1447.

[112] Trott, Oleg and Olson, Arthur J.AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading *Journal of Computational Chemistry.* 2009:455–461.

[113] Plewczynski, Dariusz and Łażniewski, Michał and Grotthuss, Marcin Von and Rychlewski, Leszek and Ginalski, KrzysztofVoteDock: Consensus docking method for prediction of protein-ligand interactions *Journal of Computational Chemistry.* 2011;32:568–581.

[114] McGann, MarkFRED Pose Prediction and Virtual Screening Accuracy *Journal of Chemical Information and Modeling.* 2011;51:578–596.

[115] Vaudel, Marc and Breiter, Daniela and Beck, Florian and Zahedi, Rene PD-score: A search engine independent MD-score *Proteomics.* 2013;13:1036–1041.

[116] Gaudreault, Francis and Najmanovich, Rafael J.FlexAID: Revisiting Docking on Non-Native-Complex Structures *Journal of Chemical Information and Modeling.* 2015;55:1323–1336.

[117] Paul, D. Sam and Gautham, N.MOLS 2.0: software package for peptide modeling and protein–ligand docking *Journal of Molecular Modeling.* 2016;22:1–9.

[118] Antunes, Dinler A. and Moll, Mark and Devaurs, Didier and Jackson, Kyle R. and Lizée, Gregory and Kavraki, Lydia E.DINC 2.0: A New Protein–Peptide Docking Webserver Using an Incremental Approach *Cancer Research.* 2017;77:e55–e57.

[119] Ballester, Pedro J. and Schreyer, Adrian and Blundell, Tom L.Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *Journal of Chemical Information and Modeling.* 2014;54:944–955.

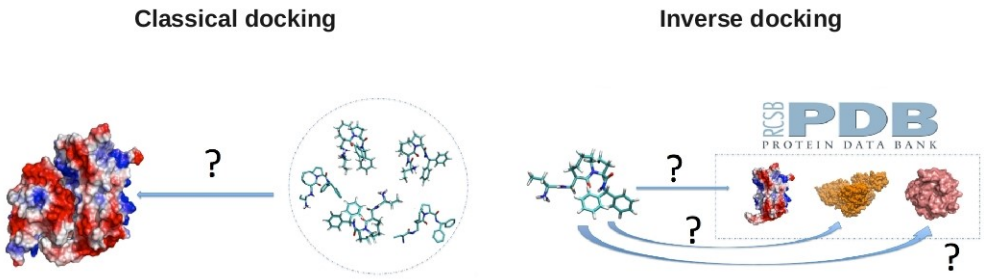

Figure 1: On the left, the illustration of the classical docking process that consists of finding the best ligand for a given protein. On the right, the illustration of the inverse docking process consisting of finding the best target for a given ligand.

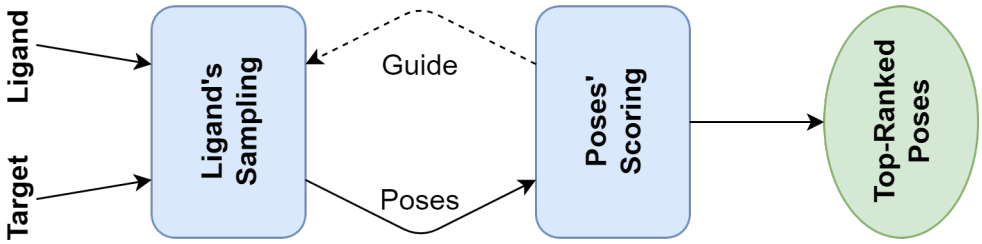

Figure 2: A simplified workflow of the molecular docking process with sampling and scoring sub-processes.

Software	Year	Sampling	Scoring	Ref
ICM	1994	Stochastic (MC)	Force field (ECEPP)	[100]
GOLD	1995	Stochastic (GA)	Force field (AMBER)	[101]
SANDOCK	1998	Shape Matching	Empirical	[102]
MultiScore[a]	2001	-	Consensus	[103]
LigandFit	2003	Stochastic (MC)	Empirical	[104]
DrugScore[a]	2005	-	Knowledge-based	[105]
DFire[a]	2005	-	Knowledge-based	[106]
Glide	2006	Stochastic (MC)	Empirical	[107]
PLANTS	2006	Stochastic (AC)	Empirical	[108]
SODOCK	2007	Stochastic (PSO)	Force field (AMBER)	[109]
eHiTS	2007	Systematic (FB)	Empirical	[110]
KScore[a]	2008	-	Knowledge-based	[111]
AutoDock 4	2009	Stochastic (GA)	Force field (AMBER)	[35]
AutoDock Vina	2010	Stochastic (GA)	Empirical	[112]
VoteScore[a]	2011	-	Consensus	[113]
FRED	2011	Systematic (IM)	Customisable	[114]
D-Score[a]	2013	-	Force field (Tripos)	[115]
FlexAID	2015	Stochastic (GA)	Empirical	[116]
MOLS 2.0	2016	Systematic (IM)	Force field (AMBER)	[117]
DINC 2.0	2017	Systematic (FB)	Empirical	[118]

Table 1: A non-exhaustive list of some molecular docking software and the class of their sampling and scoring methods. [a] are only scoring functions. This list is based on (R. Vasseur, PhD thesis, University of Reims Champagne-Ardenne, 2015)

Name	Year	Input	Usage	Refs
Linear Regression				
TABA	2020	Mass-Spring System	Scoring	[61]
Random Forest (RF)				
RF-Score	2010	Set of descriptors	Scoring	[62]
RF-Score-v2	2014	Set of descriptors	Scoring	[119]
RF-Score-v3	2015	Set of descriptors	Scoring	[63]
RF-Score-VS	2017	Set of descriptors	VS	[64]
SIEVE-Score	2019	Energy vectors	VS	[65]
Gradient Boosting trees				
AGL-Score	2019	Multi-scale weighted coloured sub-graphs	Scoring, VS	[66]
Support Vector Machine (SVM)				
SVR-KB	2011	Pairwise potential vector	Scoring	[67]
SVR-EP	2011	Set of descriptors	Scoring	
ID-Score (SVR)	2013	Set of descriptors	Scoring	[68]
PLEIC-SVM	2017	PLEIC Fingerprint	Target-specific VS[a]	[69]

Table 2: Machine Learning methods for ligand-protein docking, [a]: Target-specific VS means this method was developed for a specific type of target.

Name	Year	Input	Usage	Ref
Multi-Layers Perceptrons (MLPs)				
NNscore	2010	Set of descriptors	VS	[70]
NNscore 2.0	2011	Set of descriptors	Scoring	
Deep Docking	2020	2D fingerprints	VS	[72]
Convolutional Neural Networks (CNNs)				
AtomNet©	2015	3D grid with descriptors	VS	[75]
DeepVS	2016	Set of descriptors	VS	[76]
Ragoza2017	2017	3D grid of 34 channels	VS	[77]
Atomic CNN	2017	3D structure	Scoring	[78]
DeepSite[a]	2017	3D grid of 8 channels	Binding site detection	[79]
DenseFS	2018	3D grid of 34 channels	VS	[80]
Pafnucy	2018	3D grid of 19 channels	Scoring	[82]
DeepAffinity	2019	Fingerprints	Scoring	[83]
DeepBindRG	2019	2D matrix	Scoring	[84]
OnionNet	2019	64 descriptors for each shell	Scoring	[86]
FRSite[a]	2019	3D grid of 8 channels	Binding site detection	[87]
Francoeur2020	2020	3D grid of 28 channels	Scoring	[89]
Kalasanty[a]	2020	3D grid of 18 channels	Binding site detection	[90]
DeepSurf[a]	2020	3D grid of 18 channels	Binding site detection	[92]
Graph Neural Networks (GNNs)				
PotentialNet	2018	Atoms bond graph	Scoring	[93]
Lim2019	2019	Atoms bond graph	VS	[94]
Torgn2019[a]	2019	Target's residues graph + Ligand graph	VS	[95]
Tanebe2019[a]	2019	Target's residues graph + Ligand graph	VS	[97]
Tsubaki2019[a]	2019	Target's sequence + Ligand graph	VS	[98]
Morrone2020	2020	Ligand bond graph + Contact graph	VS	[55]

Table 3: Deep Learning methods for ligand-protein docking. According to CNNs' terminology, which comes from image terminology, channels are used to represent descriptors.[a] indicates end-to-end methods.

Presented method	**Docking engine**	**Assessed method**	1^{st} **Used dataset**	*AUC*	2^{nd} **Used dataset**	*AUC*	**Ref**[a]
PLEIC-SVM	Glide	Gscore	DUD	0.82			[69]
		PLEIC-SVM		**0.93**			
NN-Score methods	Vina	Vina	DUD	0.70			[70]
		NNScore-v1		**0.78**			
		NNScore-v2		**0.76**			
DeepVS	Dock 6.6	Dock 6.6	DUD	0.48			[76]
		DeepVS		**0.74**			
	Vina	Vina	DUD	0.62			
		DeepVS		**0.81**			
RF-Score-VS	Vina	Vina	DUD-E	0.74			[64]
		RF-Score-V3		0.67			
		RF-Score-VS		**0.84**			
	Dock 6.6	Dock 6.6	DUD-E	0.61			
		RF-Score-V3		0.66			
		RF-Score-VS		**0.80**			
AtomNet©	Smina	Smina	DUD-E	0.696			[75]
		AtomNet©		**0.895**			
Ragoza's method	Smina	Smina	DUD-E	0.716	MUV	0.549	[77]
		RF-Score		0.622		0.512	
		NN-Score		0.584		0.441	
		Ragoza's method		**0.868**		**0.522**	
DenseFS	Vina	Vina	DUD-E	0.703	MUV	0.546	[80]
		Ragoza's method		0.862		0.507	
		DenseFS		**0.917**		**0.534**	
Lim's method	Smina	Smina	DUD-E	0.689	MUV	0.533	[94]
		AtomNet©		0.855			
		Ragoza's method		0.868		0.518	
		Lim's method		**0.968**		**0.536**	
Torgn's method		Vina	DUD-E	0.716	MUV	0.538	[95]
		RF-Score		0.622		0.536	
		NNScore		0.584		0.454	
		Ragoza's method		0.868		0.567	
		Torgn's method		**0.886**		**0.621**	
Morrone's methods	Vina	Vina	DUD-E	0.70			[55]
		Morrone L		**0.82**			
		Morrone LP		**0.65**			
		Morrone L+LP		**0.81**			

Table 4: Some assessment of VS methods with DUD, DUD-E, and MUV datasets, [a] listed scores are sometimes in supporting information. A same method can have different performances on the same dataset and with the same docking engine, it is due to the impact of data preparation and docking engine setting.

Presented method	**Dataset**	**Assessed SF**	R_p	SD	**Ref**[a]
RF-Score	PDBbind 2007 Core Set	ChemScore	0.441	2.15	[62]
		GoldScore	0.295	2.29	
		RF-Score	**0.776**	**1.58**	
RF-Score-v2	PDBbind 2007 Core Set	**RF-Score-v2**	**0.803**	**1.54**	[119]
RF-Score-v3	PDBbind 2007 Core Set	**RF-Score-v3**	**0.803**	**1.42**	[63]
SVR-KB and SVR-EP	CSAR-SETI1	**SVR-KB**	**0.59**[b]		[67]
		SVR-EP	**0.55**[b]		
	CSAR-SETI2	**SVR-KB**	**0.67**[b]		
		SVR-EP	**0.50**[b]		
ID-Score	PDBbind 2007 Core Set	**ID-Score**	**0.753**	**1.63**	[68]
Atomic CNN	PDBbind 2015 Core Set	**Atomic CNN**	**0.448**[b]		[78]
	PDBbind 2015 Refined Set	**Atomic CNN**	**0.529**[b]		
Pafnucy	PDBbind 2016 Core Set	**Pafnucy**	**0.78**	**1.37**	[82]
	CASF-2013	**Pafnucy**	**0.70**	**1.61**	
DeepBindRG	CASF-2013	Vina	0.5725		[84]
		DeepBindRG	**0.6394**		
AGL-Score	CASF-2007	ID-Score	0.753		[66]
		Vina	0.554		
		AGL-Score	**0.830**		
	CASF-2013	**AGL-Score**	**0.792**		
OnionNet	PDBbind 2013 Core Set	AutoDock	0.54	1.61	[86]
		Vina	0.54	1.60	
		ChemScore	0.592	1.82	
		Pafnucy	0.70	1.61	
		RF-Score-V3	0.74	1.51	
		AGL-Score	0.792	1.45	
		OnionNet	**0.78**	**1.45**	
	PDBbind 2016 Core Set	Pafnucy	0.78		
		RF-Score-V3	0.80		
		AGL-Score	0.833		
		OnionNet	**0.816**		
Francoeur's method	PDBbind 2016 Core Set	**Francoeur**	**0.733**		[89]

Table 5: Some assessment of SF methods, [a] listed scores are sometimes in supporting information, values with [b] are $(R_p)^2$.