Current Medicinal Chemistry, 2021, *28*, 7006-7022

REVIEW ARTICLE

The Impact of Crystallographic Data for the Development of Machine Learning Models to Predict Protein-Ligand Binding Affinity

Martina Veit-Acosta[1,*] and Walter Filgueira de Azevedo Junior[2,3,*]

[1]Western Michigan University, 1903 Western, Michigan Ave, Kalamazoo, MI49008, United States; [2]Pontifical Catholic University of Rio Grande do Sul (PUCRS); Av. Ipiranga, 6681 Porto Alegre/RS 90619-900 Brazil; [3]Specialization Program in Bioinformatics; Pontifical Catholic University of Rio Grande do Sul (PUCRS); Av. Ipiranga, 6681 Porto Alegre/RS 90619-900, Brazil

ARTICLE HISTORY

Received: November 12, 2020
Revised: December 29, 2020
Accepted: January 02, 2021

DOI:
10.2174/0929867328666210210121320

CrossMark

Abstract: ***Background***: One of the main challenges in the early stages of drug discovery is the computational assessment of protein-ligand binding affinity. Machine learning techniques can contribute to predicting this type of interaction. We may apply these techniques following two approaches. Firstly, using the experimental structures for which affinity data is available. Secondly, using protein-ligand docking simulations.

Objective: In this review, we describe recently published machine learning models based on crystal structures, for which binding affinity and thermodynamic data are available.

Method: We used experimental structures available at the protein data bank and binding affinity and thermodynamic data was accessed through BindingDB, Binding MOAD, and PDBbind databases. We reviewed machine learning models to predict binding created using open source programs, such as SAnDReS and Taba.

Results: Analysis of machine learning models trained against datasets, composed of crystal structure complexes indicated the high predictive performance of these models when compared with classical scoring functions.

Conclusion: The rapid increase in the number of crystal structures of protein-ligand complexes created a favorable scenario for developing machine learning models to predict binding affinity. These models rely on experimental data from two sources, the structural and the affinity data. The combination of experimental data generates computational models that outperform the classical scoring functions.

Keywords: Crystal structures, machine learning, scoring function space, binding affinity, SAnDReS, Taba.

1. INTRODUCTION

The protein data bank (PDB) is the largest data repository of three-dimensional structures of biological macromolecules [1, 2]. The PDB has recently surpassed 170,000 entries in its database (a search carried on November 10, 2020). The structural information at the PDB covers a wide range of biomolecules, such as peptides, proteins, proteins with nucleic acids, enzymes in complexes with inhibitors, isolated nucleic acids, ribosomes, and viruses. Considering the source of data, we have solved the structures using the following techniques: X-ray diffraction crystallography [3], neutron diffraction [4], cryogenic electron microscopy (cryo-EM) [5], and nuclear magnetic resonance (NMR) spectroscopy [6]. For more details of recent developments in the PDB, we recommend the interested readers to the following reviews listed in the references [7-18].

Among the structures available at the PDB, we see the prevalence of X-ray diffraction crystallography. Considering a survey conducted in 2017 about the data available for proteins complexed with ligands at the PDB, we had over 94% of the data generated by X-ray

*Address correspondence to these authors at Western Michigan University, 1903 Western, Michigan Ave, Kalamazoo, MI 49008, United States (MVC) and Pontifical Catholic University of Rio Grande do Sul (PUCRS); Av. Ipiranga, 6681 Porto Alegre/RS 90619-900, Brazil (WFA); Tel/Fax: +55- 51-3320-3545;
E-mails: martina.veitacosta@wmich.edu
and walter@azevedolab.net, walter.junior@pucrs.br

1875-533X/21 $65.00+.00

diffraction crystallography [19]. It is worth noting that the use of cryo-EM has grown in the last three years [5, 20-28].

In the early stages of drug discovery and development, the application of structure-based drug design (SBDD) can facilitate the drug design by studying the structural features responsible for binding affinity. For review papers, please see [29-31]. One of the most successful applications of such an approach is the study of HIV-1 protease (EC 3.4.23.16) inhibitors and their subsequent use as drugs to treat HIV infection. For an interesting review about this protein target, the authors suggest the study by Lawal *et al*. [32]. Considering the studies that used SBDD focused on enzyme targets, the prevalence of X-ray diffraction data is overwhelming [8, 12, 19, 33].

Besides the success of SBDD in the study of inhibitors of HIV-1 protease, we have recently witnessed a crescent number of machine learning models focused on this enzyme [34-37]. These works indicated the potential of combining the crystal data with machine learning techniques to generate targeted scoring functions for the prediction of binding affinity [35].

Applications of machine learning techniques to construct computational models for predicting binding affinity based on the atomic coordinates of receptor-ligand complexes go beyond HIV-1 protease. There have been recent reports (2017 - 2020) of targeted-machine learning models to predict affinity of ligands against the spike protein of SARS-CoV-2 [38, 39], COVID-19 main proteinase (EC 3.4.22.69) [40], cyclin-dependent kinase 2 (CDK2) (EC 2.7.11.22) [41-43], cyclin-dependent kinases (CDKs) [44, 45], 5-lipoxygenase (EC 1.13.11.34) [46], and 3-dehydroquinate dehydratase (DHQD) (EC 4.2.1.10) [47].

Among the machine learning models, most of them are targeted scoring functions that predict inhibition constant (K_i) in an expression where the dependent variable is the $\log(K_i)$. But there are computational models that predict the half-maximal inhibitory concentration (IC_{50}) with a response variable using $\log(IC_{50})$ for CDK [43, 45]. Targeted scoring functions predict thermodynamic parameters, such as variation of Gibbs free energy of binding (ΔG), which are rare, mostly due to the scarcity of crystal structures with this type of data for a specific protein target. On the other hand, there is a model that predicts the ΔG based on an ensemble of high-resolution crystallographic structures with different enzymes in the training set [42, 48]. This computational approach intends to build a general scoring function that can predict the ΔG for any protein-ligand complex.

The availability of structural and functional data (binding affinity and ΔG) made the development of robust computational models to predict binding affinity based on the atomic coordinates of protein-ligand complexes, possible [49-56]. These computational models outperform classical scoring functions implemented in docking programs, such as AutoDock4 (AD4) [57, 58], AutoDock Vina (Vina) [59], and Molegro Virtual Docker (MVD) [60-65].

The development of targeted scoring functions paved the way to establish the theoretical framework to address the binding affinity of receptor-ligand complexes. We may address this problem by employing the concept of scoring function space (SFS) [19, 66]. This space composed of infinite scoring functions focuses on the relationship of the protein [67] and chemical spaces [68-73], where computational approaches scan the SFS to find an adequate model to predict the affinity of an element of the protein space and a sub-space of chemical space composed of binders to this protein.

In this review, we describe the PDB and highlight how to recover data from this database. We explain storing of the structural data of protein-ligand complexes at the PDB. We also show how PDB handles the information about binding affinity and thermodynamic data. The PDB accesses this data through links to three additional databases, that are BindingDB [74, 75], Binding MOAD [76-78], and PDBbind [79, 80]. We also describe how machine learning programs integrate structural and binding data to generate targeted scoring functions, highlighting the importance of crystal data for these approaches. Finally, we update the analysis of the techniques used to solve the structure of protein-ligand complexes.

2. METHODS

2.1. Machine Learning Approaches

Considering recent machine learning approaches for the calculation of binding affinity or thermodynamic data from the atomic coordinates of receptor-ligand complexes, we may highlight the following programs: Statistical Analysis of Docking Results and Scoring Functions (SAnDReS) [81, 82], Pafnucy [83], Tool to Analyze the Binding Affinity (Taba) [44], property-encoded shape distributions together with standard support vector machine (PESD-SVM) [84], Neural-Network-Based Scoring function (NNScore series) [85-87], and Random Forest Score (RF-Score series) [88-92].

Programs to develop scoring functions based on the atomic coordinates of protein-ligand complexes share

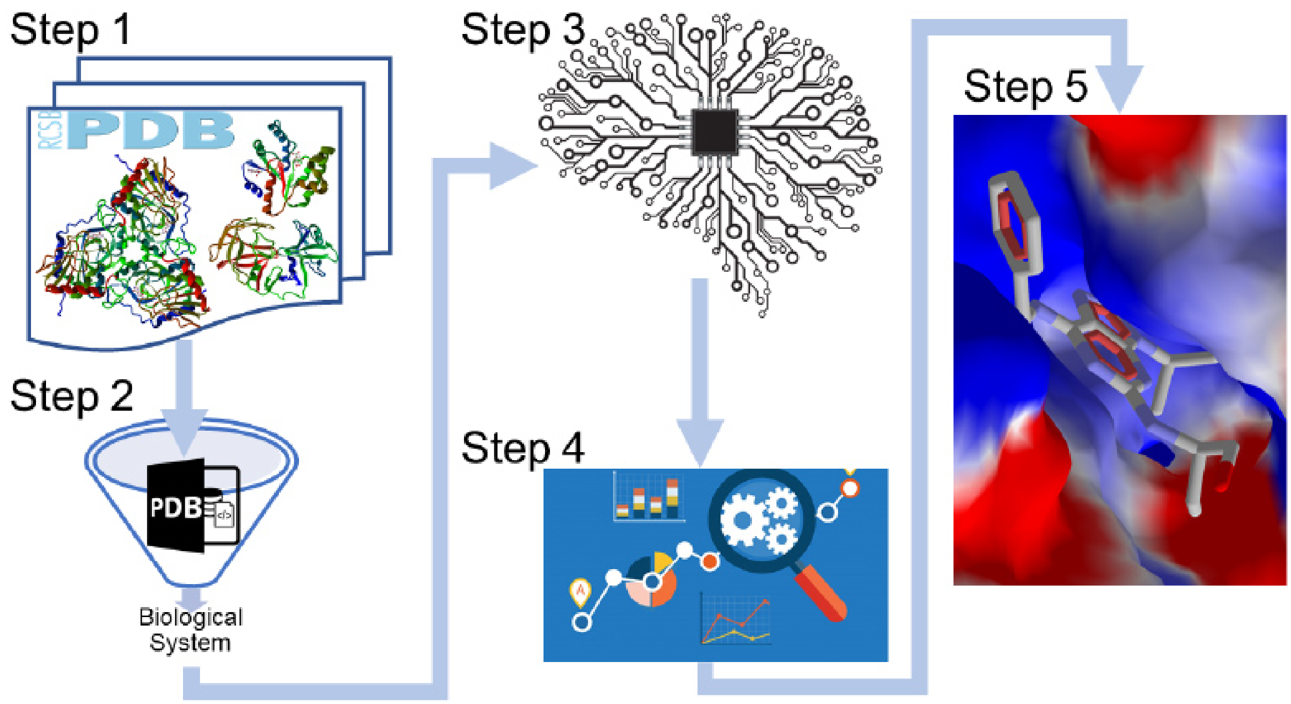

Fig. (1). Roadmap to the development of machine learning models to predict protein-ligand binding affinity; **Step 1** (Definition of the Biological System): In this part, we select the biological system defining the structures and binding affinity data to download from the PDB. **Step 2** (Filtering): Then,we filter our data to eliminate repeated ligands and check for the inconsistencies, such as missing ligands in the dataset. **Step 3** (Machine Learning): In this step, we generate machine learning models using the structures in the training set. **Step 4** (Statistical Analysis): In this phase, we carry out the statistical analysis of the predictive performance. We use the structures in the test set. **Step 5** (Final Model): We select the best machine learning model and save it. We employ this model to calculate binding affinity using the atomic coordinates of protein-ligand complexes. We used the program MVD [60] to generate images of the protein structures. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

the same overall approach, as highlighted in Fig. (**1**). Briefly, in step 1, we define the biological system. The biological system is a protein, or a set of proteins, for which we will generate a machine learning model to predict the binding affinity. In the sequence, we select the PDB access codes of our biological system. Next, we download the structures and the affinity data from the PDB. Programs such as SAnDReS and Taba automatically download these data directly from the PDB. In step 2, we filter the crude data from the PDB. These data may be needed in the sequence to be filtered for the elimination of the structures with repeated ligands. Following this, we separate the dataset into training and test sets. In step 3, we use the data in the training set to develop the machine learning models.

We usually build machine learning models using approximately 70% of data as the training set and ~30% of the dataset as a test set, as recommended in a study [93]. In this step, we may apply different machine learning techniques. There are programs that focus on a specific machine learning technique, such as Random Forest Score (RF-Score series) [88-92]. There are other programs where we may test several machine learning methods to generate models with different predictive performances [44, 81, 82]. In step 4, we assess the predictive performance focused on the test set. If we have more than one scoring function, we select the one with the best overall performance using the test set. In step 5, we have our machine learning model to predict the binding affinity for any ligand.

2.2. Statistical Analysis

As we previously highlighted, to assess the predictive power of the classical scoring functions and targeted models, the machine learning programs calculate the correlation coefficients and p-values [93].

2.3. Protein Data Bank

We can retrieve functional and structural data directly from the PDB. Recent developments in the PDB [7] integrated into the advanced search tool of the PDB (available at https://www.rcsb.org/search/advanced) gave the possibility to carry out searches by combining different sources of data. Especially for those interested in machine learning modeling using the structures for which affinity data is known, it is possible to search for the deposited data with K_i. In doing so, the PDB returns all entries with this binding affinity data. The same type of search can also focus on structures with different binding affinity or thermodynamic parameters, such as dissociation constant (K_d), IC_{50}, and ΔG [81, 94-98].

The PDB stores the atomic coordinates of protein-ligand complexes in three major formats: PDB, mmCIF (macromolecular crystallographic information file), and PDBML/XML (Protein Data Bank Markup Language) [7]. The most used format is PDB. Archaic, protein-ligand docking and machine learning programs rely heavily on the PDB format. Typically, the atomic coordinates have a rigid format followed by all programs that read PDB formats. Machine learning programs used to generate scoring functions, use the atomic coordinates to assess protein-ligand interactions and create energy terms, such as van der Waals [99], electrostatic potential [100], hydrogen bonding [101], and entropy [102]. The calculation of energy terms implemented in the scoring functions use the interatomic distances (r_{ij}) between an atom in the protein (index i) and another in the ligand (index j). The atomic coordinates in the PDB are in the three-dimensional cartesian space and expressed in Å (1Å = 10^{-10} m). The interatomic distance has the following expression,

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{1}$$

Where x_i, y_i, z_i are used for the coordinates of protein atoms and x_j, y_j, z_j for the ligand atoms. The common electrostatic potential energy term ($U_{Electrostatic}$) has the following equation,

$$U_{Electrostatic} = \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \tag{2}$$

where we have the atomic partial charges (q_i and q_j), the permittivity function $\varepsilon(r_{ij})$, and the interatomic distance (r_{ij}), which was calculated using equation (1), taking atomic coordinates as the input [100]. Most of the energy terms used in the scoring functions need atomic coordinates for their calculations [99-101]. Many classical scoring function expressions employ equation (2) for the assessment of electrostatic energies. In equation (2), we calculate the interatomic distances using equation (1). These expressions allow us a fast determination of the electrostatic interactions [58]. These scoring functions facilitate the assessment of poses during docking simulations [57]. Also, we can easily incorporate this energy term in machine learning approaches to develop targeted scoring functions.

Fig. (**2**) shows the fields designed for each type of information stored in a line of atomic coordinates in a PDB file [103-105]. PDB assigns the first six columns to identify the type of information stored in each line. The keyword “ATOM ” in the first six columns indicates that we have the atomic coordinates for the protein structure. On the other hand, the keyword “HETATM” indicates other types of atoms. This keyword could be used to store the atomic coordinates of a ligand. The following field indicates the number of the atom. All remaining fields are defined in Fig. (**2**). It is necessary to obtain further information about the occupancy factor and the B_{factor}.

The occupancy factor is a fraction of the atom at the given atomic coordinates. It is related to a characteristic of protein-ligand complexes in the crystal state [3]. Since the atomic coordinates in a crystal structure are an average of all unit cells found in the crystal, flexible parts of a protein can have two or more positions for the same atom. For these multiple positions of an atom, crystallographers refine the atomic positions assigning occupancy factors below 1.0. For instance, a lysine residue may have two positions for the nitrogen (NZ). In this case, we have two atomic coordinates for the same NZ with occupancy factors proportional to the electron density of the NZ at each position.

For the B_{factor}, we have to keep in mind that atoms oscillate, and the B_{factor} reflects the mean amplitude of this oscillation. In the simple case in which the components of the oscillation are same in all aspects, it is named isotropic oscillation [106]. In the following equation, we have the expression of the mean square amplitude of atomic vibration ($\langle u^2 \rangle$).

$$\langle u^2 \rangle = \frac{B_{factor}}{8\pi^2} \tag{3}$$

Calculation of the B_{factor} for atoms of main-chain and side-chain usually indicate higher values for those in the side chain. It is due to the intrinsic flexibility of these atoms compared to the main chain [106].

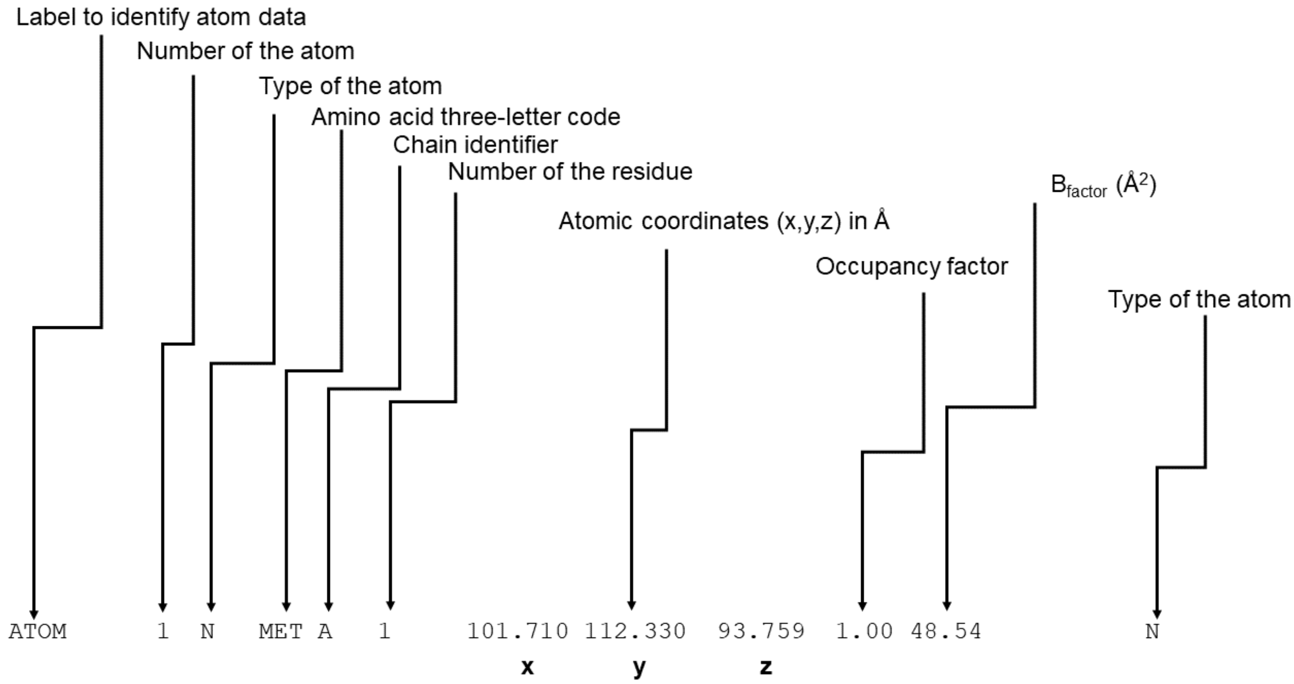

Fig. (2). Fields in lines of atomic coordinates in a PDB file. Keywords for the identification of atoms of a protein (“ATOM“) or ligands (“HETATM”) use columns from 1 to 6. PDB reserves columns ranging from 7 to 11 for the atom order. For the identification of the type of atom, PDB uses columns ranging from 14 to 15. PDB reserves columns ranging from 18 to 20 for the protein residue, ligand, the base of nucleic acid, crystallization co-factors, or water molecules. PDB assigns column 22 to chain identification. PDB takes in account the columns ranging from 23 to 26 or residue/ligand number. PDB stores the atomic coordinates in columns ranging from 31 to 54, columns 31 to 38 for x, 39 to 46 for y, and 47 to 54 for z. We express atomic coordinates in Å. PDB assigns columns ranging from 57 to 60 to occupancy factors. B_{factor} uses columns ranging from 62 to 65. It is expressed in $Å^2$. Column 77 is used for the chemical element.

2.4. Datasets

To highlight the importance of the structural and binding data available at the PDB for machine learning modeling, we describe the previously published modeling of eight different biological systems listed in (Table **1**). For all these systems, we have one machine learning model (developed using SAnDReS or Taba) and the binding affinity was calculated using at least two classical scoring functions (AD4, Vina, MolDock Score (MDS), and PLANTS Score (PLS)). The data used to develop these models are available at https://github.com/azevedolab/sandres and https://github.com/azevedolab/taba.

3. RESULTS AND DISCUSSION

3.1. Biological Systems

The application of machine learning approaches to generate the novel generation of scoring functions have caught the attention of researchers interested in computational models to predict protein-ligand binding affinity [88-92]. Considering recent applications of SAnDReS [81] and Taba [44] for machine learning modeling, we have eight different biological systems highlighted in (Table **1**). In this table, we also show the predictive performances of the classical scoring functions.

The Spearman rank correlation coefficient (ρ) of classical scoring functions implemented in the docking programs AD4, MVD, and Vina for these biological systems range from -0.199 to 0.629 (training set) and from -0.943 to 0.764 (test set). By comparing the classical scoring functions for the test sets, we have the calculated MDS using MVD as the highest correlation for half of the biological systems. Nevertheless, all scoring functions created using machine learning approaches outperform these classical scoring functions.

Analysis of the machine learning models indicate a variation of ρ from 0.390 to 0.721 (training set) and from 0.328 to 0.943 (test set). Considering the test set,

Table 1. Predictive performance of classical scoring functions and machine learning models.

Biological Systems	Reference	Number of Structures	Binding Data	Scoring Function	ρ(training set)	p-value (training set)	ρ (test set)	p-value (test set)
Coagulation factor Xa	[81]	57 (25)	K_i	AD4	0.267	$1.005 \cdot 10^{-01}$	0.325	$6.210 \cdot 10^{-02}$
				MDS	0.160	$2.335 \cdot 10^{-01}$	0.396	$4.995 \cdot 10^{-02}$
				PLS	0.150	$2.578 \cdot 10^{-01}$	0.333	$5.061 \cdot 10^{-02}$
				SAnDReS	0.560	$5.920 \cdot 10^{-06}$	0.435	$2.975 \cdot 10^{-02}$
				Vina	0.245	$1.732 \cdot 10^{-01}$	0.297	$7.848 \cdot 10^{-02}$
$HRIC_{50}$	[43]	118 (55)	IC_{50}	AD4	-0.099	$5.981 \cdot 10^{-01}$	0.142	$1.001 \cdot 10^{-01}$
				MDS	0.284	$1.939 \cdot 10^{-03}$	0.224	$9.678 \cdot 10^{-02}$
				PLS	0.298	$1.625 \cdot 10^{-03}$	0.314	$4.012 \cdot 10^{-02}$
				SAnDReS	0.401	$7.243 \cdot 10^{-06}$	0.328	$1.363 \cdot 10^{-02}$
				Vina	0.190	$9.078 \cdot 10^{-02}$	0.277	$8.022 \cdot 10^{-02}$
$CDK2IC_{50}$	[43]	118* (11)	IC_{50}	AD4	0.099	$5.981 \cdot 10^{-01}$	0.445	$1.697 \cdot 10^{-01}$
				MDS	0.284	$1.939 \cdot 10^{-03}$	0.391	$2.345 \cdot 10^{-01}$
				PLS	0.298	$1.625 \cdot 10^{-03}$	0.682	$2.084 \cdot 10^{-02}$
				SAnDReS	0.401	$7.243 \cdot 10^{-06}$	0.845	$1.045 \cdot 10^{-03}$
				Vina	0.190	$9.078 \cdot 10^{-02}$	0.418	$2.006 \cdot 10^{-01}$
HIV-1 PR	[36]	51 (20)	K_i	MDS	0.218	$1.247 \cdot 10^{-01}$	0.086	$7.193 \cdot 10^{-01}$
				PLS	0.264	$6.162 \cdot 10^{-02}$	0.010	$9.674 \cdot 10^{-01}$
				SAnDReS	0.525	$7.707 \cdot 10^{-05}$	0.368	$1.106 \cdot 10^{-01}$
CDK	[45]	122 (54)	IC_{50}	AD4	0.190	$3.890 \cdot 10^{-02}$	0.213	$1.082 \cdot 10^{-01}$
				MDS	0.059	$5.179 \cdot 10^{-01}$	-0.291	$3.265 \cdot 10^{-02}$
				PLS	-0.162	$7.515 \cdot 10^{-02}$	-0.132	$3.405 \cdot 10^{-01}$
				SAnDReS	0.390	$9.065 \cdot 10^{-06}$	0.346	$1.044 \cdot 10^{-02}$
				Vina	0.339	$1.495 \cdot 10^{-04}$	0.207	$1.267 \cdot 10^{-01}$
DHQD	[47]	16 (6)	K_i	AD4	0.219	$4.140 \cdot 10^{-01}$	0.714	$1.110 \cdot 10^{-01}$
				MDS	-0.199	$4.600 \cdot 10^{-01}$	-0.943	$4.800 \cdot 10^{-03}$
				PLS	0.629	$9.060 \cdot 10^{-03}$	0.314	$5.440 \cdot 10^{-01}$
				SAnDReS	0.675	$4.160 \cdot 10^{-03}$	0.943	$4.810 \cdot 10^{-03}$
				Vina	0.307	$1.055 \cdot 10^{-01}$	0.602	$1.904 \cdot 10^{-01}$
ΔG	[48]	36 (12)	ΔG	AD4	0.284	$9.326 \cdot 10^{-02}$	0.340	$2.799.10^{-01}$
				MDS	0.599	$1.148 \cdot 10^{-04}$	0.764	$3.850 \cdot 10^{-03}$
				PLS	0.534	$7.918 \cdot 10^{-04}$	0.641	$2.470 \cdot 10^{-02}$
				SAnDReS	0.721	$6.975 \cdot 10^{-07}$	0.886	$1.240 \cdot 10^{-04}$
				Vina	0.454	$5.416 \cdot 10^{-03}$	0.746	$3.329 \cdot 10^{-03}$
$CDKK_i$	[44]	22 (9)	K_i	AD4	0.358	$1.018 \cdot 10^{-01}$	-0.133	$7.324 \cdot 10^{-01}$
				MDS	0.299	$1.759 \cdot 10^{-01}$	0.217	$5.755 \cdot 10^{-01}$
				PLS	0.351	$1.095 \cdot 10^{-01}$	0.183	$6.368 \cdot 10^{-01}$
				Taba	0.558	$6.300 \cdot 10^{-01}$	0.783	$1.252 \cdot 10^{-02}$
				Vina	0.267	$2.304 \cdot 10^{-01}$	-0.067	$8.647 \cdot 10^{-01}$

*The $HRIC_{50}$ training set was used to develop a general machine learning model tested against a dataset of 11 CDK2 structures.

Table 2. Structures available in the PDB for each type of binding affinity/Thermodynamic data.

Binding Affinity/thermodynamic Data	Total[1]	X-ray[2]	NMR[3]	Neutron[4]	EM[5]
K_i	6681	6641	29	6	9
K_d	6077	5998	77	4	2
K_a	157	157	0	0	0
IC_{50}	8993	8952	28	2	12
EC_{50}	841	836	1	2	3
ΔG	140	138	1	1	0
ΔH	137	135	1	1	0

[1]Total number of structures for which binding affinity/thermodynamic data is available; The numbers indicate entries available for each type of data. We may count the same complex more than once if it has more than one experimentally determined type of binding affinity/thermodynamic data.
[2]Structures solved by X-ray crystallography for which binding affinity/thermodynamic data is available.
[3]Structures solved by nuclear magnetic resonance (NMR) for which binding affinity/thermodynamic data is available.
[4]Structures solved by neutron crystallography for which binding affinity/thermodynamic data is available.
[5]Structures solved by electron micrography (EM) for which binding affinity/thermodynamic data is available.

we observe the highest ρ for the machine learning model generated to predict the log(K_i) for DHQD. We see the lowest correlation for the $HRIC_{50}$ biological system. This system has 173 crystal structures of different enzymes.

The striking difference was observed in the predictive performances using the same computational approach (SAnDReS); in the case of the first seven biological systems, as mentioned in (Table **1**) the difference may be due to some intrinsic features of the datasets used to train the machine learning models. For instance, we could attribute the worst predictive performance for the $HRIC_{50}$ system to the data heterogeneity, However, we do not have structural information for one specific protein. On the other hand, the model developed for DHQD focuses on one enzyme [19].

In (Table **1**), we highlighted the predictive performances of classical scoring functions and machine learning models. We did not intend to have a complete evaluation of the performances, exploring all available classical scoring functions. Our goal is to emphasize that, at least for these classical scoring functions (AD4, Vina, MVD, and PLS), previously published machine learning models generated with SAnDReS and Taba showed superior performance.

Taken together, we may say that we observe the higher predictive performance for machine learning models developed for a specific protein system that use as binding affinity data the K_i ($CDKK_i$ and DHQD biological systems). On the other hand, general machine learning models with IC_{50} data show a low correlation with the experimental data ($HRIC_{50}$ biological system).

3.2. Structural Information

In 2017, we conducted a survey of the contribution of different techniques employed to generate three-dimensional protein-ligand structures available at PDB [19]. We filtered our data focusing on protein-ligand complexes for which thermodynamic parameters and binding affinity data were available. At that time, we had approximately 120,000 structures deposited at the PDB. We now have 170,597 entries (a search carried out on November 10, 2020).

The PDB advanced tools allow one to filter information considering association constant (K_a), ΔG, enthalpy (ΔH), half-maximal effective concentration (EC_{50}), K_d, K_i, and IC_{50}. Such a combination of data is a promising scenario for the generation of targeted scoring functions developed using machine learning techniques. Employing the same methodology previously reported in a study [19] to quantify the contribution of the methods used to solve complex structures, we still have the X-ray diffraction crystallography as the top experimental approach to solve protein-ligand complexes [19]. This technique contributed 99.3% of the total, calculated using the data available in Table **2**.

We witnessed a rise in the number of deposits related to structures solved using cryo-EM [5, 6], once the contribution of this technique for the number of entries of protein-ligand complexes was analyzed it was found that its participation is low, with 0.113% of the total.

It is crystal clear from the data presented in (Table **2**) that X-ray diffraction crystallography is the dominant experimental approach used to determine the three-dimensional structures of protein-ligand complexes. Although we have information using other techniques, such as NMR spectroscopy and cryo-EM, the overwhelming presence of X-ray diffraction crystallographic data strongly suggests that we can most comfortably rely on this type of data for machine learning modeling.

There are a few possible reasons to explain this prevalence of X-ray diffraction crystallography information of protein-ligand complexes. X-ray diffraction crystallography is the oldest technique to solve biomolecules. The first protein-ligand structures for which binding affinity data were available were published in 1982 [107, 108]. Another aspect that contributes to the prevalence of crystal structures is related to the determination of protein-ligand complexes. We may use the conditions to crystallize the apo form to generate crystals of the complexes. Also, we could use the crystals of the apo structure for soaking experiments. Soaking allows the diffusion of a ligand solution into a crystal of the unliganded protein [109-112].

3.3. Scoring Function Space

The application of the concept of SFS furnishes a robust theoretical framework to analyze machine learning models for the prediction of the binding affinity [19, 66]. Considering the performance variation in the machine learning models [88-92, 113-127], it is clear that the scoring functions developed for a specific protein target outperform general scoring functions. Considering the SFS, we see that focusing on one protein and a subspace of the chemical space has a higher probability of finding an adequate predictive model. It is more likely to generate a model with a low correlation with the experimental data if we take many proteins. Since what we have is an average predictive model extracted from the SFS. In this scenario, the PDB has pivotal importance in providing the integration of the crystallographic structures and binding affinity data to be used for the training of the machine learning models.

3.4. Comparison of Experimental Methods to Elucidate Protein-Ligand Structures

We previously highlighted that X-ray diffraction crystallography is the leading experimental method to assess the three-dimensional structures of protein-ligand complexes, considering those for which we have binding affinity data. On the other hand, alternative methods, such as cryo-EM [5] and NMR spectroscopy [128], have advantages that may change this trend in the future. Considering NMR spectroscopy [129], for instance, to determine the three-dimensional structure using this tool, we do not need to crystallize the protein. The bottleneck of X-ray diffraction crystallography is needed to have crystals of the protein-ligand complexes. There are no guarantees that we may achieve the crystallization of a protein for which we want to determine the structure [3]. Also, considering that we have crystals of the unliganded protein, the complex formation may not be achievable. The addition of the ligand to the protein sample may affect the crystallization process. Therefore, new screenings (co-crystallization) should be necessary to generate X-ray diffracting crystals of the protein-ligand complexes [109]. It is also possible to soak the ligand into preformed protein crystals [3]. This soaking approach also has technical challenges. For instance, we may only generate a soluble ligand solution in a condition that will damage the preformed protein crystal. The crystallization requirement is not present in studies using NMR spectroscopy [6]. There is also an eternal debate between those who defend NMR spectroscopy against crystallography. In physiological conditions, we do not have crystals, and the packing of the protein may affect the conformation of the structure [3, 128, 129].

Yet another technique is the cryo-EM. This experimental tool to determine three-dimensional structures has gained crescent attention in the last years [5, 20-28]. In cryo-EM, we do not need crystals to generate the three-dimensional data. This technique has no limitation on the molecular size of the protein, as in NMR.

In summary, when we consider the problem of determining the three-dimensional structure of a protein, we must consider that these three experimental methods have their pros and cons. The weakest link in the crystallography chain is the need for crystals [3], whereas the NMR and cryo-EM do not need them [128, 129]. It is also possible to combine two techniques, such as cryo-EM and NMR [129]. One problem of the cryo-EM is the resolution of the data [129]. A search on PDB for all protein structures solved using cryo-EM returned 6,434 entries (search carried out on December 24, 2020). Amongst these structures, only seven showed the resolution between 1.0 - 1.5 Å. There are no structures solved with data better than 1.0 Å. Most of the entries determined using cryo-EM have data worse than 3.0 Å resolution (5,821 out of 6,434). The same search with a focus on X-ray diffraction crystallography returned 150,528 entries. We have data up to < 0.5 Å limit. Considering data better than 1.5 Å, we have 9.7% of the entries determined using X-ray diffraction crystallography against 1.09% for the cryo-EM. Since the resolution is fundamental, at the moment, we have the superior performance of the diffraction technique compared to the cryo-EM. Finally, for NMR spectroscopy, the limitations are the size of the biological systems and the need for isotopic enrichment of the ligand to obtain data [128]. We do not face these challenges with X-ray diffraction crystallography.

Besides the impact of X-ray diffraction crystallography, cryo-EM, and NMR spectroscopy, more recently, new techniques have shown promising results. Among them, we may highlight electron tomography (ET) [130]. This method makes it possible to assess an image of biological structures *in situ*. The main issue with this approach is also related to the resolution. We can improve the resolution using subtomogram averaging (STA) [131].

Considering all points highlighted above, we may say that we do not have an ideal experimental method to obtain three-dimensional structures. But, when we take the available protein-ligand data, the chief technology is X-ray diffraction crystallography.

3.5. Computational Modeling of Protein-Ligand Structures

As we highlighted above, the bottleneck of X-ray diffraction crystallography is the need for the crystals. And the two other experimental methods also face challenges, such as the size limitation for NMR spectroscopy and the resolution issues of cryo-EM. And the final challenge, the availability of the protein material for the structural studies, irrespective of any method [3]. Even while facing all these limitations, we may have a three-dimensional structure of a protein of interest.

In the absence of the experimental structural data for a protein, we may generate a three-dimensional model based on the homology, which is achieved through the satisfaction of the spatial restraints implemented in the program, MODELER [132-135]. In this computational technique, we use a previously solved structure with a sequence identity of 30% or higher, as an initial model, named as a template. We take the sequence alignment of the template and the protein we want to model. Then we carry out the modifications in the amino acids whereever necessary. These approaches must satisfy the spatial restraints present in the template structure. Besides MODELLER, we have computational methods such as I-TASSER [136-138], ROSETTA [139-141], and RaptorX [142-144]. Alternatively, we may use deep learning methods, such as the one implemented in the program AlphaFold [145-147], to generate molecular models for the protein where we do not have experimental three-dimensional data. This deep-learning approach recently showed superior predictive performance when modeling the structures available at CASP (Critical Assessment of protein Structure Prediction) [145, 147].

We may generate three-dimensional structures of protein-ligand complexes through the docking simulations [148]. To do so, we use the atomic coordinates of the protein structures obtained through modeling. We have several protein-ligand docking programs, such as AutoDock4 [57, 58] and AutoDock Vina [59]. We may add machine learning to predict the binding affinity of the ligands [41-45]. Then, we use molecular dynamics simulations to confirm the binding of the ligands [149-153]. We may also investigate the dynamics of protein-ligand interactions [153].

3.6. Methods for the Prediction of Binding Affinity

Recent progress in the scoring functions using machine learning methods [19, 35, 41-44] made the superior predictive performance of these approaches clear compared to the classical scoring functions [50-55]. Analysis of the impact of the size of testing and training sets indicated improved the overall performance for larger datasets [154]. The increasing number of protein-ligand structures for which the binding affinity data is available comprises crude data for machine learning models with superior performance. We expect that this trend will continue, which will generate better computational models for the binding calculation.

Other developments in the study of the computational methods to assess intermolecular interactions are related to the following methods: free energy perturbation, thermodynamic integration, molecular mechanics/Poisson–Boltzmann surface area (MM-PBAS), and linear interaction energy. All these computational approaches contributed to generate models for the assessment of the protein-ligand interactions. For the literature describing applications of these methods [150-153, 155-157]. Taken together, we may say that the wide range of the available computational methods made it clear that we may address the SFS from different perspectives [66]. Some studies employed physics-based methods [41, 44, 150-153, 155-157], and the others focused on targeted scoring functions [19].

CONCLUSION

In this review, we highlighted the role of X-ray diffraction crystallography in providing data for protein-ligand complexes. This technique is responsible for over 99% of data about protein-ligand complexes. We considered only those entries for which binding affinity data is available. The integration of structural and functional information provide crude data that make the generation of machine learning models targeted at specific protein systems possible. Machine learning methods targeted to a single protein create scoring functions with superior predictive performance compared to the multi-protein models. By taking several proteins, an average predictive model extracted from the SFS can be generated. We expect that proteins are subjected to evolution and inserted in a complex chemical environment, as found in the biological systems. Moreover, the application of a targeted machine-learning model is an adequate computational approach to build machine learning models to predict binding affinity.

LIST OF ABBREVIATIONS

AD4	= AutoDock4
CDK	= Cyclin-dependent Kinase
CDK2	= Cyclin-dependent Kinase 2
CDK2IC$_{50}$	= CDK2 Structures with IC$_{50}$ data
CDKK$_i$	= CDK Structures with K$_i$ data
COVID-19	= Coronavirus Disease of 2019
Cryo-EM	= Cryogenic Electron Microscopy
ΔG	= Variation of Gibbs Free Energy of Binding
ΔH	= Enthalpy
DHQD	= Dehydroquinate Dehydratase
EC	= Enzyme Classification Number
EC$_{50}$	= Half-maximal Effective Concentration
EM	= Electron Microscopy
ET	= Electron Tomography
HIV-1 PR	= HIV-1 Protease Structures with K$_i$ data
HRIC$_{50}$	= High-resolution Structures with IC$_{50}$ Data
IC$_{50}$	= Half-maximal Inhibitory Concentration
K_a	= Association Constant
K_d	= Dissociation Constant
K_i	= Inhibition Constant
MOAD	= Mother of All Databases
MDS	= MolDock Score
mmCIF	= Macromolecular Crystallographic Information File
MM-PBAS	= Molecular Mechanics/Poisson-Boltzmann Surface Area
MVD	= Molegro Virtual Docker
NMR	= Nuclear Magnetic Resonance
NNScore	= Neural-network-based Scoring Function
PDB	= Protein Data Bank
PDBML/XML	= Protein Data Bank Markup Language
PESD-SVM	= Property-encoded Shape Distributions Together with Standard Support Vector Machine
PLS	= PLANTS Score
ρ	= Spearman Rank Correlation Coefficient
RF-Score	= Random Forest Score
SAnDReS	= Statistical Analysis of Docking Results and Scoring Functions
SARS-CoV-2	= Severe Acute Respiratory Syndrome Coronavirus 2
SBDD	= Structure-based Drug Design
SFS	= Scoring Function Space
STA	= Subtomogram Averaging
Taba	= Tool to Analyze the Binding Affinity
Vina	= AutoDock Vina

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

Walter Filgueira de Azevedo Junior is a researcher for CNPq (Brazil) (Process Number: 309029/2018-0). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001. Martina Veit-Acosta acknowledges the receipt of the Dieter H. Haenicke Scholarship (Haenicke Institute for Global Education).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We acknowledge the assistance of the reviewers of this work, who helped us in many ways through their enlightening comments and valuable suggestions. Without their contributions, the completion of this manuscript would not be possible.

REFERENCES

[1] Vincenzi, M.; Mercurio, F.A.; Leone, M. Protein interaction domains and post-translational modifications: structural features and drug discovery applications. *Curr. Med. Chem.,* **2020**, *27*(37), 6306-6355. http://dx.doi.org/10.2174/0929867326666190620101637 PMID: 31250750

[2] Vincenzi, M.; Mercurio, F.A.; Leone, M. Protein interaction domains: structural features and drug discovery applications (part 2). *Curr. Med. Chem.,* **2021**, *28*(5), 854-892. http://dx.doi.org/10.2174/0929867327666200114114142 PMID: 31942846

[3] Canduri, F.; de Azevedo, W.F. Protein crystallography in drug discovery. *Curr. Drug Targets,* **2008**, *9*(12), 1048-1053. http://dx.doi.org/10.2174/138945008786949423 PMID: 19128214

[4] Coates, L.; Myles, D.A. Prospects for atomic resolution and neutron crystallography in drug design. *Curr. Drug Targets,* **2004**, *5*(2), 173-178. http://dx.doi.org/10.2174/1389450043490613 PMID: 15011950

[5] Van Drie, J.H.; Tong, L. Cryo-EM as a powerful tool for drug discovery. *Bioorg. Med. Chem. Lett.,* **2020**, *30*(22), 127524. http://dx.doi.org/10.1016/j.bmcl.2020.127524 PMID: 32890683

[6] Shimada, I.; Ueda, T.; Kofuku, Y.; Eddy, M.T.; Wüthrich, K. GPCR drug discovery: integrating solution NMR data with crystal and cryo-EM structures. *Nat. Rev. Drug Discov.,* **2019**, *18*(1), 59-82. http://dx.doi.org/10.1038/nrd.2018.180 PMID: 30410121

[7] Berman, H.M.; Vallat, B.; Lawson, C.L. The data universe of structural biology. *IUCrJ,* **2020**, *7*(Pt 4), 630-638. http://dx.doi.org/10.1107/S205225252000562X PMID: 32695409

[8] Westbrook, J.D.; Soskind, R.; Hudson, B.P.; Burley, S.K. Impact of the protein data bank on antineoplastic approvals. *Drug Discov. Today,* **2020**, *25*(5), 837-850. http://dx.doi.org/10.1016/j.drudis.2020.02.002 PMID: 32068073

[9] Ionescu, M.I. An overview of the crystallized structures of the SARS-CoV-2. *Protein J.,* **2020**, *39*(6), 600-618. http://dx.doi.org/10.1007/s10930-020-09933-w PMID: 33098476

[10] Goodsell, D.S.; Burley, S.K. RCSB protein data bank tools for 3D structure-guided cancer research: human papillomavirus (HPV) case study. *Oncogene,* **2020**, *39*(43), 6623-6632. http://dx.doi.org/10.1038/s41388-020-01461-2 PMID: 32939013

[11] Di Costanzo, L.; Geremia, S. Atomic details of carbon-based nanomolecules interacting with proteins. *Molecules,* **2020**, *25*(15), 3555. http://dx.doi.org/10.3390/molecules25153555 PMID: 32759758

[12] Wang, J.; Yazdani, S.; Han, A.; Schapira, M. Structure-based view of the druggable genome. *Drug Discov. Today,* **2020**, *25*(3), 561-567. http://dx.doi.org/10.1016/j.drudis.2020.02.006 PMID: 32084498

[13] Copoiu, L.; Malhotra, S. The current structural glycome landscape and emerging technologies. *Curr. Opin. Struct. Biol.,* **2020**, *62*, 132-139. http://dx.doi.org/10.1016/j.sbi.2019.12.020 PMID: 32006784

[14] Haas, D.J. The early history of cryo-cooling for macromolecular crystallography. *IUCrJ,* **2020**, *7*(Pt 2), 148-157. http://dx.doi.org/10.1107/S2052252519016993 PMID: 32148843

[15] Bascos, N.A.D.; Landry, S.J. A history of molecular chaperone structures in the protein data bank. *Int. J. Mol. Sci.,* **2019**, *20*(24), 6195. http://dx.doi.org/10.3390/ijms20246195 PMID: 31817979

[16] Weber, P.; Pissis, C.; Navaza, R.; Mechaly, A.E.; Saul, F.; Alzari, P.M.; Haouz, A. High-throughput crystallization pipeline at the crystallography core facility of the institut pasteur. *Molecules,* **2019**, *24*(24), 4451. http://dx.doi.org/10.3390/molecules24244451 PMID: 31817305

[17] Liu, B.; He, H.; Luo, H.; Zhang, T.; Jiang, J. Artificial intelligence and big data facilitated targeted drug discovery. *Stroke Vasc. Neurol.,* **2019**, *4*(4), 206-213. http://dx.doi.org/10.1136/svn-2019-000290 PMID: 32030204

[18] Martinez, X.; Krone, M.; Alharbi, N.; Rose, A.S.; Laramee, R.S.; O'Donoghue, S.; Baaden, M.; Chavent, M. Molecular graphics: bridging structural biologists and computer scientists. *Structure,* **2019**, *27*(11), 1617-1623. http://dx.doi.org/10.1016/j.str.2019.09.001 PMID: 31564470

[19] Heck, G.S.; Pintro, V.O.; Pereira, R.R.; de Ávila, M.B.; Levin, N.M.B.; de Azevedo, W.F. Supervised machine learning methods applied to predict ligand- binding affinity. *Curr. Med. Chem.,* **2017**, *24*(23), 2459-2470. http://dx.doi.org/10.2174/0929867324666170623092503 PMID: 28641555

[20] Poitevin, F.; Kushner, A.; Li, X.; Dao Duc, K. Structural heterogeneities of the ribosome: new frontiers and opportunities for cryo-EM. *Molecules,* **2020**, *25*(18), 4262. http://dx.doi.org/10.3390/molecules25184262 PMID: 32957592

[21] Wu, M.; Lander, G.C. Present and emerging methodologies in cryo-EM single-particle analysis. *Biophys. J.,* **2020**, *119*(7), 1281-1289. http://dx.doi.org/10.1016/j.bpj.2020.08.027 PMID: 32919493

[22] Gisriel, C.J.; Wang, J.; Brudvig, G.W.; Bryant, D.A. Opportunities and challenges for assigning cofactors in cryo-EM density maps of chlorophyll-containing proteins. *Commun. Biol.,* **2020**, *3*(1), 408. http://dx.doi.org/10.1038/s42003-020-01139-1 PMID: 32733087

[23] Fica, S.M. Cryo-EM snapshots of the human spliceosome reveal structural adaptions for splicing regulation. *Curr. Opin. Struct. Biol.,* **2020**, *65*, 139-148. http://dx.doi.org/10.1016/j.sbi.2020.06.018 PMID: 32717639

[24] Nygaard, R.; Kim, J.; Mancia, F. Cryo-electron microscopy analysis of small membrane proteins. *Curr. Opin.*

Struct. Biol., **2020**, *64*, 26-33.
http://dx.doi.org/10.1016/j.sbi.2020.05.009 PMID: 32603877

[25] Scheres, S.H.; Zhang, W.; Falcon, B.; Goedert, M. Cryo-EM structures of tau filaments. *Curr. Opin. Struct. Biol.,* **2020**, *64*, 17-25.
http://dx.doi.org/10.1016/j.sbi.2020.05.011 PMID: 32603876

[26] Wu, M.; Lander, G.C. How low can we go? Structure determination of small biological complexes using single-particle cryo-EM. *Curr. Opin. Struct. Biol.,* **2020**, *64*, 9-16.
http://dx.doi.org/10.1016/j.sbi.2020.05.007 PMID: 32599507

[27] Oshima, A. Structural insights into gap junction channels boosted by cryo-EM. *Curr. Opin. Struct. Biol.,* **2020**, *63*, 42-48.
http://dx.doi.org/10.1016/j.sbi.2020.03.008 PMID: 32339861

[28] Luque, D.; Castón, J.R. Cryo-electron microscopy for the study of virus assembly. *Nat. Chem. Biol.,* **2020**, *16*(3), 231-239.
http://dx.doi.org/10.1038/s41589-020-0477-1 PMID: 32080621

[29] Bockman, M.R.; Mishra, N.; Aldrich, C.C. The biotin biosynthetic pathway in *Mycobacterium tuberculosis* is a validated target for the development of antibacterial agents. *Curr. Med. Chem.,* **2020**, *27*(25), 4194-4232.
http://dx.doi.org/10.2174/0929867326666190119161551 PMID: 30663561

[30] Xia, J.; Feng, B.; Wen, G.; Xue, W.; Ma, G.; Zhang, H.; Wu, S. Bacterial lipoprotein biosynthetic pathway as a potential target for structure-based design of antibacterial agents *Curr. Med. Chem.,* **2020**, *27*(7), 1132-1150.
http://dx.doi.org/10.2174/0929867325666181008143411 PMID: 30360704

[31] Xue, W.; Fu, T.; Zheng, G.; Tu, G.; Zhang, Y.; Yang, F.; Tao, L.; Yao, L.; Zhu, F. Recent advances and challenges of the drugs acting on monoamine transporters. *Curr. Med. Chem.,* **2020**, *27*(23), 3830-3876.
http://dx.doi.org/10.2174/0929867325666181009123218 PMID: 30306851

[32] Lawal, M.M.; Sanusi, Z.K.; Govender, T.; Maguire, G.E.M.; Honarparvar, B.; Kruger, H.G. From recognition to reaction mechanism: an overview on the interactions between HIV-1 protease and its natural targets. *Curr. Med. Chem.,* **2020**, *27*(15), 2514-2549.
http://dx.doi.org/10.2174/0929867325666181113122900 PMID: 30421668

[33] Mazanetz, M.P.; Goode, C.H.F.; Chudyk, E.I. Ligand- and structure-based drug design and optimization using KNIME. *Curr. Med. Chem.,* **2020**, *27*(38), 6458-6479.
http://dx.doi.org/10.2174/0929867326666190409141016 PMID: 30963962

[34] Leidner, F.; Kurt Yilmaz, N.; Schiffer, C.A. Target-specific prediction of ligand affinity with structure-based interaction fingerprints. *J. Chem. Inf. Model.,* **2019**, *59*(9), 3679-3691.
http://dx.doi.org/10.1021/acs.jcim.9b00457 PMID: 31381335

[35] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Machine learning to predict binding affinity. *Methods Mol. Biol.,* **2019**, *2053*, 251-273.
http://dx.doi.org/10.1007/978-1-4939-9752-7_16 PMID: 31452110

[36] Pintro, V.O.; de Azevedo, W.F. Optimized virtual screening workflow: towards target-based polynomial scoring functions for HIV-1 protease. *Comb. Chem. High Throughput Screen.,* **2017**, *20*(9), 820-827.
http://dx.doi.org/10.2174/1386207320666171121110019 PMID: 29165067

[37] Agniswamy, J.; Louis, J.M.; Roche, J.; Harrison, R.W.; Weber, I.T. Structural studies of a rationally selected multi-drug resistant HIV-1 protease reveal synergistic effect of distal mutations on flap dynamics. *PLoS One,* **2016**, *11*(12), e0168616.
http://dx.doi.org/10.1371/journal.pone.0168616 PMID: 27992544

[38] Song, Y.; Song, J.; Wei, X.; Huang, M.; Sun, M.; Zhu, L.; Lin, B.; Shen, H.; Zhu, Z.; Yang, C. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. *Anal. Chem.,* **2020**, *92*(14), 9895-9900.
http://dx.doi.org/10.1021/acs.analchem.0c01394 PMID: 32551560

[39] Batra, R.; Chan, H.; Kamath, G.; Ramprasad, R.; Cherukara, M.J.; Sankaranarayanan, S.K.R.S. Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking studies. *J. Phys. Chem. Lett.,* **2020**, *11*(17), 7058-7065.
http://dx.doi.org/10.1021/acs.jpclett.0c02278 PMID: 32787328

[40] Xu, Z.; Yang, L.; Zhang, X.; Zhang, Q.; Yang, Z.; Liu, Y.; Wei, S.; Liu, W. Discovery of potential flavonoid inhibitors against COVID-19 3CL proteinase based on virtual screening strategy. *Front. Mol. Biosci.,* **2020**, *7*, 556481.
http://dx.doi.org/10.3389/fmolb.2020.556481 PMID: 33134310

[41] Bitencourt-Ferreira, G.; Duarte da Silva, A.; Filgueira de Azevedo, W. Jr. Application of machine learning techniques to predict binding affinity for drug targets: a study of cyclin-dependent kinase 2. *Curr. Med. Chem.,* **2021**, *28*(2), 253-265.
http://dx.doi.org/10.2174/2213275912666191102162959 PMID: 31729287

[42] Bitencourt-Ferreira, G.; Rizzotto, C.; de Azevedo Junior, W.F. Machine learning-based scoring functions. development and applications with SAnDReS. *Curr. Med. Chem.,* **2021**, *28*(9), 1746-1756.
http://dx.doi.org/10.2174/0929867327666200515101820 PMID: 32410551

[43] de Ávila, M.B.; Xavier, M.M.; Pintro, V.O.; de Azevedo, W.F. Jr. Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2. *Biochem. Biophys. Res. Commun.,* **2017**, *494*(1-2), 305-310.
http://dx.doi.org/10.1016/j.bbrc.2017.10.035 PMID: 29017921

[44] da Silva, A.D.; Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Taba: A Tool to Analyze the Binding Affinity. *J. Comput. Chem.,* **2020**, *41*(1), 69-73.
http://dx.doi.org/10.1002/jcc.26048 PMID: 31410856

[45] Levin, N.M.B.; Pintro, V.O.; Bitencourt-Ferreira, G.; de Mattos, B.B.; de Castro Silvério, A.; de Azevedo, W.F. Jr. Development of CDK-targeted scoring functions for prediction of binding affinity. *Biophys. Chem.,* **2018**, *235*, 1-8.
http://dx.doi.org/10.1016/j.bpc.2018.01.004 PMID: 29407904

[46] Shameera Ahamed, T.K.; Rajan, V.K.; Sabira, K.; Muraleedharan, K. QSAR classification-based virtual screening followed by molecular docking studies for identification of potential inhibitors of 5-lipoxygenase. *Comput. Bi-*

ol. Chem., **2018**, *77*, 154-166.
http://dx.doi.org/10.1016/j.compbiolchem.2018.10.002 PMID: 30321850

[47] de Ávila, M.B.; de Azevedo, W.F. Jr. Development of machine learning models to predict inhibition of 3-dehydroquinate dehydratase. *Chem. Biol. Drug Des.,* **2018**, *92*(2), 1468-1474.
http://dx.doi.org/10.1111/cbdd.13312 PMID: 29676519

[48] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophys. Chem.,* **2018**, *240*, 63-69.
http://dx.doi.org/10.1016/j.bpc.2018.05.010 PMID: 29906639

[49] Francoeur, P.G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R.B.; Snyder, I.; Koes, D.R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.,* **2020**, *60*(9), 4200-4215.
http://dx.doi.org/10.1021/acs.jcim.0c00411 PMID: 32865404

[50] Parks, C.; Gaieb, Z.; Amaro, R.E. An analysis of proteochemometric and conformal prediction machine learning protein-ligand binding affinity models. *Front. Mol. Biosci.,* **2020**, *7*, 93.
http://dx.doi.org/10.3389/fmolb.2020.00093 PMID: 32671093

[51] Soni, A.; Bhat, R.; Jayaram, B. Improving the binding affinity estimations of protein-ligand complexes using machine-learning facilitated force field method. *J. Comput. Aided Mol. Des.,* **2020**, *34*(8), 817-830.
http://dx.doi.org/10.1007/s10822-020-00305-1 PMID: 32185583

[52] Wang, D.D.; Zhu, M.; Yan, H. Computationally predicting binding affinity in protein-ligand complexes: free energy-based simulations and machine learning-based scoring functions. *Brief. Bioinform.,* **2021**, *22*(3), bbaa107.
http://dx.doi.org/10.1093/bib/bbaa107 PMID: 32591817

[53] Shen, C.; Hu, Y.; Wang, Z.; Zhang, X.; Zhong, H.; Wang, G.; Yao, X.; Xu, L.; Cao, D.; Hou, T. 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.
http://dx.doi.org/10.1093/bib/bbz173 PMID: 31982914

[54] Wang, D.; Cui, C.; Ding, X.; Xiong, Z.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. Improving the virtual screening ability of target-specific scoring functions using deep learning methods. *Front. Pharmacol.,* **2019**, *10*, 924.
http://dx.doi.org/10.3389/fphar.2019.00924 PMID: 31507420

[55] Zhang, H.; Liao, L.; Saravanan, K.M.; Yin, P.; Wei, Y. DeepBindRG: a deep learning based method for estimating effective protein-ligand affinity. *PeerJ,* **2019**, *7*, e7362.
http://dx.doi.org/10.7717/peerj.7362 PMID: 31380152

[56] Li, J.; Fu, A.; Zhang, L. An overview of scoring functions used for protein-ligand interactions in molecular docking. *Interdiscip. Sci.,* **2019**, *11*(2), 320-328.
http://dx.doi.org/10.1007/s12539-019-00327-w PMID: 30877639

[57] Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.,* **2009**, *30*(16), 2785-2791.
http://dx.doi.org/10.1002/jcc.21256 PMID: 19399780

[58] Bitencourt-Ferreira, G.; Pintro, V.O.; de Azevedo, W.F. Jr. Docking with AutoDock4. *Methods Mol. Biol.,* **2019**, *2053*, 125-148.
http://dx.doi.org/10.1007/978-1-4939-9752-7_9 PMID: 31452103

[59] Trott, O.; Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.,* **2010**, *31*(2), 455-461.
http://dx.doi.org/10.1002/jcc.21334 PMID: 19499576

[60] Thomsen, R.; Christensen, M.H. MolDock: a new technique for high-accuracy molecular docking. *J. Med. Chem.,* **2006**, *49*(11), 3315-3321.
http://dx.doi.org/10.1021/jm051197e PMID: 16722650

[61] Heberlé, G.; de Azevedo, W.F. Jr. Bio-inspired algorithms applied to molecular docking simulations. *Curr. Med. Chem.,* **2011**, *18*(9), 1339-1352.
http://dx.doi.org/10.2174/092986711795029573 PMID: 21366530

[62] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Molegro virtual docker for docking. *Methods Mol. Biol.,* **2019**, *2053*, 149-167.
http://dx.doi.org/10.1007/978-1-4939-9752-7_10 PMID: 31452104

[63] De Azevedo, W.F. Jr. MolDock applied to structure-based virtual screening. *Curr. Drug Targets,* **2010**, *11*(3), 327-334.
http://dx.doi.org/10.2174/138945010790711941 PMID: 20210757

[64] Azevedo, L.S.; Moraes, F.P.; Xavier, M.M.; Pantoja, E.O.; Villavicencio, B.; Finck, J.A.; Proenca, A.M.; Rocha, K.B.; de Azevedo, W.F. Recent progress of molecular docking simulations applied to development of drugs. *Curr. Bioinform.,* **2012**, *7*(4), 352-365.
http://dx.doi.org/10.2174/157489312803901063

[65] Dias, R.; de Azevedo, W.F. Jr. Molecular docking algorithms. *Curr. Drug Targets,* **2008**, *9*(12), 1040-1047.
http://dx.doi.org/10.2174/138945008786949432 PMID: 19128213

[66] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Exploring the scoring function space. *Methods Mol. Biol.,* **2019**, *2053*, 275-281.
http://dx.doi.org/10.1007/978-1-4939-9752-7_17 PMID: 31452111

[67] Smith, J.M. Natural selection and the concept of a protein space. *Nature,* **1970**, *225*(5232), 563-564.
http://dx.doi.org/10.1038/225563a0 PMID: 5411867

[68] Bohacek, R.S.; McMartin, C.; Guida, W.C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.,* **1996**, *16*(1), 3-50.
http://dx.doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6 PMID: 8788213

[69] Dobson, C.M. Chemical space and biology. *Nature,* **2004**, *432*(7019), 824-828.
http://dx.doi.org/10.1038/nature03192 PMID: 15602547

[70] Kirkpatrick, P.; Ellis, C. Chemical space. *Nature,* **2004**, *432*, 823.
http://dx.doi.org/10.1038/432823a

[71] Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature,* **2004**, *432*(7019), 855-861.
http://dx.doi.org/10.1038/nature03193 PMID: 15602551

[72] Shoichet, B.K. Virtual screening of chemical libraries. *Nature,* **2004**, *432*(7019), 862-865.
http://dx.doi.org/10.1038/nature03197 PMID: 15602552

[73] Stockwell, B.R. Exploring biology with small organic molecules. *Nature,* **2004**, *432*(7019), 846-854. http://dx.doi.org/10.1038/nature03196 PMID: 15602550

[74] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.,* **2007**, *35*(Database issue), D198-D201. http://dx.doi.org/10.1093/nar/gkl999 PMID: 17145705

[75] Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.,* **2016**, *44*(D1), D1045-D1053. http://dx.doi.org/10.1093/nar/gkv1072 PMID: 26481362

[76] Smith, R.D.; Clark, J.J.; Ahmed, A.; Orban, Z.J.; Dunbar, J.B. Jr.; Carlson, H.A. Updates to binding MOAD (mother of all databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.,* **2019**, *431*(13), 2423-2433. http://dx.doi.org/10.1016/j.jmb.2019.05.024 PMID: 31125569

[77] Benson, M.L.; Smith, R.D.; Khazanov, N.A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H.A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.,* **2008**, *36*(Database issue), D674-D678. http://dx.doi.org/10.1093/nar/gkm911 PMID: 18055497

[78] Ahmed, A.; Smith, R.D.; Clark, J.J.; Dunbar, J.B. Jr.; Carlson, H.A. Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Res.,* **2015**, *43*(Database issue), D465-D469. http://dx.doi.org/10.1093/nar/gku1088 PMID: 25378330

[79] Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics,* **2015**, *31*(3), 405-412. http://dx.doi.org/10.1093/bioinformatics/btu626 PMID: 25301850

[80] Liu, Z.; Li, J.; Liu, J.; Liu, Y.; Nie, W.; Han, L.; Li, Y.; Wang, R. Cross-mapping of protein - ligand binding data between ChEMBL and PDBbind. *Mol. Inform.,* **2015**, *34*(8), 568-576. http://dx.doi.org/10.1002/minf.201500010 PMID: 27490502

[81] Xavier, M.M.; Heck, G.S.; Avila, M.B.; Levin, N.M.B.; Pintro, V.O.; Carvalho, N.L.; Azevedo, W.F. Jr. SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions. *Comb. Chem. High Throughput Screen.,* **2016**, *19*(10), 801-812. http://dx.doi.org/10.2174/1386207319666160927111347 PMID: 27686428

[82] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. SAnDReS: a computational tool for docking. *Methods Mol. Biol.,* **2019**, *2053*, 51-65. http://dx.doi.org/10.1007/978-1-4939-9752-7_4 PMID: 31452098

[83] Stepniewska-Dziubinska, M.M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics,* **2018**, *34*(21), 3666-3674. http://dx.doi.org/10.1093/bioinformatics/bty374 PMID: 29757353

[84] Das, S.; Krein, M.P.; Breneman, C.M. Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model.,* **2010**, *50*(2), 298-308. http://dx.doi.org/10.1021/ci9004139 PMID: 20095526

[85] Durrant, J.D.; McCammon, J.A. NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. *J. Chem. Inf. Model.,* **2010**, *50*(10), 1865-1871. http://dx.doi.org/10.1021/ci100244v PMID: 20845954

[86] Durrant, J.D.; McCammon, J.A. NNScore 2.0: a neural-network receptor-ligand scoring function. *J. Chem. Inf. Model.,* **2011**, *51*(11), 2897-2903. http://dx.doi.org/10.1021/ci2003889 PMID: 22017367

[87] Durrant, J.D.; Friedman, A.J.; Rogers, K.E.; McCammon, J.A. Comparing neural-network scoring functions and the state of the art: applications to common library screening. *J. Chem. Inf. Model.,* **2013**, *53*(7), 1726-1735. http://dx.doi.org/10.1021/ci400042y PMID: 23734946

[88] Ballester, P.J.; Mitchell, J.B.O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics,* **2010**, *26*(9), 1169-1175. http://dx.doi.org/10.1093/bioinformatics/btq112 PMID: 20236947

[89] Ballester, P.J.; Schreyer, A.; Blundell, T.L. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.,* **2014**, *54*(3), 944-955. http://dx.doi.org/10.1021/ci500091r PMID: 24528282

[90] Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P.J. The impact of docking pose generation error on the prediction of binding affinity. In: *Computacional intelligence methods for bioinformatics and biostatitics*; Serio, C.D.I.; Liò, P.; Nonis, A.; Tagliaferri, R., Eds.; Springer: Cambridge, UK, **2014**; 8623, pp. 231-241.

[91] Li, H.; Leung, K.S.; Ballester, P.J.; Wong, M.H. istar: a web platform for large-scale protein-ligand docking. *PLoS One,* **2014**, *9*(1), e85678. http://dx.doi.org/10.1371/journal.pone.0085678 PMID: 24475049

[92] Wójcikowski, M.; Siedlecki, P.; Ballester, P.J. Building machine-learning scoring functions for structure-based prediction of intermolecular binding affinity. *Methods Mol. Biol.,* **2019**, *2053*, 1-12. http://dx.doi.org/10.1007/978-1-4939-9752-7_1 PMID: 31452095

[93] Zar, J.H. Significance testing of the spearman rank correlation coefficient. *J. Am. Stat. Assoc.,* **1972**, *67*(339), 578-580. http://dx.doi.org/10.1080/01621459.1972.10481251

[94] de Azevedo, W.F. Jr.; Dias, R. Computational methods for calculation of ligand-binding affinity. *Curr. Drug Targets,* **2008**, *9*(12), 1031-1039. http://dx.doi.org/10.2174/138945008786949405 PMID: 19128212

[95] Dias, R.; Timmers, L.F.; Caceres, R.A.; de Azevedo, W.F. Jr. Evaluation of molecular docking using polynomial empirical scoring functions. *Curr. Drug Targets,* **2008**, *9*(12), 1062-1070. http://dx.doi.org/10.2174/138945008786949450 PMID: 19128216

[96] De Azevedo, W.F. Jr. Structure-based virtual screening. *Curr. Drug Targets,* **2010**, *11*(3), 261-263. http://dx.doi.org/10.2174/138945010790711941 PMID: 20214598

[97] de Azevedo, W.F. Jr.; Dias, R. Evaluation of ligand-binding affinity using polynomial empirical scoring functions. *Bioorg. Med. Chem.,* **2008**, *16*(20), 9378-9382. http://dx.doi.org/10.1016/j.bmc.2008.08.014 PMID:

18829335
[98] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Molecular docking simulations with ArgusLab. *Methods Mol. Biol.,* **2019**, *2053*, 203-220. http://dx.doi.org/10.1007/978-1-4939-9752-7_13 PMID: 31452107
[99] Bitencourt-Ferreira, G.; Veit-Acosta, M.; de Azevedo, W.F. Jr. Van der Waals potential in protein complexes. *Methods Mol. Biol.,* **2019**, *2053*, 79-91. http://dx.doi.org/10.1007/978-1-4939-9752-7_6 PMID: 31452100
[100] Bitencourt-Ferreira, G.; Veit-Acosta, M.; de Azevedo, W.F. Jr. Electrostatic energy in protein-ligand complexes. *Methods Mol. Biol.,* **2019**, *2053*, 67-77. http://dx.doi.org/10.1007/978-1-4939-9752-7_5 PMID: 31452099
[101] Bitencourt-Ferreira, G.; Veit-Acosta, M.; de Azevedo, W.F. Jr. Hydrogen bonds in protein-ligand complexes. *Methods Mol. Biol.,* **2019**, *2053*, 93-107. http://dx.doi.org/10.1007/978-1-4939-9752-7_7 PMID: 31452101
[102] Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D.J.; Kellogg, G.E.; Mozzarelli, A. Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods. *Curr. Med. Chem.,* **2004**, *11*(23), 3093-3118. http://dx.doi.org/10.2174/0929867043363929 PMID: 15579003
[103] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.,* **2000**, *28*(1), 235-242. http://dx.doi.org/10.1093/nar/28.1.235 PMID: 10592235
[104] Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J.D.; Zardecki, C. The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.,* **2002**, *58*(1), 899-907. http://dx.doi.org/10.1107/S0907444902003451 PMID: 12037327
[105] Westbrook, J.; Feng, Z.; Chen, L.; Yang, H.; Berman, H.M. The protein data bank and structural genomics. *Nucleic Acids Res.,* **2003**, *31*(1), 489-491. http://dx.doi.org/10.1093/nar/gkg068 PMID: 12520059
[106] Delatorre, P.; de Azevedo, W.F. Jr. Simulation of electron density maps for two-dimensional crystal structures using mathematica. *J. Appl. Cryst.,* **2001**, *34*(5), 658-660. http://dx.doi.org/10.1107/S0021889801009724
[107] Rees, D.C.; Lipscomb, W.N. Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 A resolution. *J. Mol. Biol.,* **1982**, *160*(3), 475-498. http://dx.doi.org/10.1016/0022-2836(82)90309-6 PMID: 7154070
[108] Bolin, J.T.; Filman, D.J.; Matthews, D.A.; Hamlin, R.C.; Kraut, J. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate. *J. Biol. Chem.,* **1982**, *257*(22), 13650-13662. http://dx.doi.org/10.1016/S0021-9258(18)33497-5 PMID: 6815178
[109] de Azevedo, W.F. Jr.; Canduri, F.; Basso, L.A.; Palma, M.S.; Santos, D.S. Determining the structural basis for specificity of ligands using crystallographic screening. *Cell Biochem. Biophys.,* **2006**, *44*(3), 405-411. http://dx.doi.org/10.1385/CBB:44:3:405 PMID: 16679527
[110] De Azevedo, W.F.; Leclerc, S.; Meijer, L.; Havlicek, L.; Strnad, M.; Kim, S.H. Inhibition of cyclin-dependent kinases by purine analogues: crystal structure of human cdk2 complexed with roscovitine. *Eur. J. Biochem.,* **1997**, *243*(1-2), 518-526. http://dx.doi.org/10.1111/j.1432-1033.1997.0518a.x PMID: 9030780
[111] De Azevedo, W.F. Jr.; Mueller-Dieckmann, H.J.; Schulze-Gahmen, U.; Worland, P.J.; Sausville, E.; Kim, S.H. Structural basis for specificity and potency of a flavonoid inhibitor of human CDK2, a cell cycle kinase. *Proc. Natl. Acad. Sci. USA,* **1996**, *93*(7), 2735-2740. http://dx.doi.org/10.1073/pnas.93.7.2735 PMID: 8610110
[112] Krystof, V.; Cankar, P.; Frysová, I.; Slouka, J.; Kontopidis, G.; Dzubák, P.; Hajdúch, M.; Srovnal, J.; de Azevedo, W.F. Jr.; Orság, M.; Paprskárová, M.; Rolcík, J.; Látr, A.; Fischer, P.M.; Strnad, M. 4-arylazo-3,5-diamino-1H-pyrazole CDK inhibitors: SAR study, crystal structure in complex with CDK2, selectivity, and cellular effects. *J. Med. Chem.,* **2006**, *49*(22), 6500-6509. http://dx.doi.org/10.1021/jm0605740 PMID: 17064068
[113] Murugan, N.A.; Muvva, C.; Jeyarajpandian, C.; Jeyakanthan, J.; Subramanian, V. Performance of force-field- and machine learning-based scoring functions in ranking MAO-B protein-inhibitor complexes in relevance to developing Parkinson's therapeutics. *Int. J. Mol. Sci.,* **2020**, *21*(20), 7648. http://dx.doi.org/10.3390/ijms21207648 PMID: 33081086
[114] Mohanan, A.; Melge, A.R.; Mohan, C.G. Predicting the molecular mechanism of EGFR domain II dimer binding interface by machine learning to identify potent small molecule inhibitor for treatment of cancer. *J. Pharm. Sci.,* **2020**, *110*(2), 727-737. http://dx.doi.org/10.1016/j.xphs.2020.10.015 PMID: 33058896
[115] Decherchi, S.; Cavalli, A. Thermodynamics and kinetics of drug-target binding by molecular simulation. *Chem. Rev.,* **2020**, *120*(23), 12788-12833. http://dx.doi.org/10.1021/acs.chemrev.0c00534 PMID: 33006893
[116] Barra, C.; Ackaert, C.; Reynisson, B.; Schockaert, J.; Jessen, L.E.; Watson, M.; Jang, A.; Comtois-Marotte, S.; Goulet, J.P.; Pattijn, S.; Paramithiotis, E.; Nielsen, M. Immunopeptidomic data integration to artificial neural networks enhances protein-drug immunogenicity prediction. *Front. Immunol.,* **2020**, *11*, 1304. http://dx.doi.org/10.3389/fimmu.2020.01304 PMID: 32655572
[117] Stepniewska-Dziubinska, M.M.; Zielenkiewicz, P.; Siedlecki, P. Improving detection of protein-ligand binding sites with 3D segmentation. *Sci. Rep.,* **2020**, *10*(1), 5035. http://dx.doi.org/10.1038/s41598-020-61860-z PMID: 32193447
[118] D'Souza, S.; Prema, K.V.; Balaji, S. Machine learning models for drug-target interactions: current knowledge and future directions. *Drug Discov. Today,* **2020**, *25*(4), 748-756. http://dx.doi.org/10.1016/j.drudis.2020.03.003 PMID: 32171918
[119] Boyles, F.; Deane, C.M.; Morris, G.M. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics,* **2020**, *36*(3), 758-764. http://dx.doi.org/10.1093/bioinformatics/btz665 PMID: 31598630

[120] Aranha, M.P.; Spooner, C.; Demerdash, O.; Czejdo, B.; Smith, J.C.; Mitchell, J.C. Prediction of peptide binding to MHC using machine learning with sequence and structure-based feature sets. *Biochim. Biophys. Acta, Gen. Subj.,* **2020**, *1864*(4), 129535. http://dx.doi.org/10.1016/j.bbagen.2020.129535 PMID: 31954798

[121] Zhao, L.; Wang, J.; Pang, L.; Liu, Y.; Zhang, J. GANsDTA: predicting drug-target binding affinity using GANs. *Front. Genet.,* **2020**, *10*, 1243. http://dx.doi.org/10.3389/fgene.2019.01243 PMID: 31993067

[122] Miyazaki, Y.; Ono, N.; Huang, M.; Altaf-Ul-Amin, M.; Kanaya, S. Comprehensive exploration of target-specific ligands using a graph convolution neural network. *Mol. Inform.,* **2020**, *39*(1-2), e1900095. http://dx.doi.org/10.1002/minf.201900095 PMID: 31815371

[123] Zheng, L.; Fan, J.; Mu, Y. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. *ACS Omega,* **2019**, *4*(14), 15956-15965. http://dx.doi.org/10.1021/acsomega.9b01997 PMID: 31592466

[124] Smith, C.C.; Chai, S.; Washington, A.R.; Lee, S.J.; Landoni, E.; Field, K.; Garness, J.; Bixby, L.M.; Selitsky, S.R.; Parker, J.S.; Savoldo, B.; Serody, J.S.; Vincent, B.G. Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes. *Cancer Immunol. Res.,* **2019**, *7*(10), 1591-1604. http://dx.doi.org/10.1158/2326-6066.CIR-19-0155 PMID: 31515258

[125] Kwofie, S.K.; Broni, E.; Teye, J.; Quansah, E.; Issah, I.; Wilson, M.D.; Miller, W.A., III; Tiburu, E.K.; Bonney, J.H.K. Pharmacoinformatics-based identification of potential bioactive compounds against Ebola virus protein VP24. *Comput. Biol. Med.,* **2019**, *113*, 103414. http://dx.doi.org/10.1016/j.compbiomed.2019.103414 PMID: 31536833

[126] Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C.J.; Duca, J.S.; Hornak, V.; Koes, D.R.; Kurtzman, T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One,* **2019**, *14*(8), e0220113. http://dx.doi.org/10.1371/journal.pone.0220113 PMID: 31430292

[127] Aldeghi, M.; Gapsys, V.; de Groot, B.L. Predicting kinase inhibitor resistance: physics-based and data-driven approaches. *ACS Cent. Sci.,* **2019**, *5*(8), 1468-1474. http://dx.doi.org/10.1021/acscentsci.9b00590 PMID: 31482130

[128] Li, Q.; Kang, C. A practical perspective on the roles of solution NMR spectroscopy in drug discovery. *Molecules,* **2020**, *25*(13), 2974. http://dx.doi.org/10.3390/molecules25132974 PMID: 32605297

[129] Geraets, J.A.; Pothula, K.R.; Schröder, G.F. Integrating cryo-EM and NMR data. *Curr. Opin. Struct. Biol.,* **2020**, *61*, 173-181. http://dx.doi.org/10.1016/j.sbi.2020.01.008 PMID: 32028106

[130] Leigh, K.E.; Navarro, P.P.; Scaramuzza, S.; Chen, W.; Zhang, Y.; Castaño-Díez, D.; Kudryashev, M. Subtomogram averaging from cryo-electron tomograms. *Methods Cell Biol.,* **2019**, *152*, 217-259. http://dx.doi.org/10.1016/bs.mcb.2019.04.003 PMID: 31326022

[131] Han, R.; Li, L.; Yang, P.; Zhang, F.; Gao, X. A novel constrained reconstruction model towards high-resolution subtomogram averaging. *Bioinformatics,* **2019**, *37*(11), 1616-1626. http://dx.doi.org/10.1093/bioinformatics/btz787 PMID: 31617571

[132] Lohning, A.E.; Levonis, S.M.; Williams-Noonan, B.; Schweiker, S.S. A practical guide to molecular docking and homology modelling for medicinal chemists. *Curr. Top. Med. Chem.,* **2017**, *17*(18), 2023-2040. http://dx.doi.org/10.2174/1568026617666170130110827 PMID: 28137238

[133] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Homology modeling of protein targets with MODELLER. *Methods Mol. Biol.,* **2019**, *2053*, 231-249. http://dx.doi.org/10.1007/978-1-4939-9752-7_15 PMID: 31452109

[134] Sali, A.; Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.,* **1993**, *234*(3), 779-815. http://dx.doi.org/10.1006/jmbi.1993.1626 PMID: 8254673

[135] Uchôa, H.B.; Jorge, G.E.; Freitas Da Silveira, N.J.; Camera, J.C. Jr.; Canduri, F.; De Azevedo, W.F. Jr. Parmodel: a web server for automated comparative modeling of proteins. *Biochem. Biophys. Res. Commun.,* **2004**, *325*(4), 1481-1486. http://dx.doi.org/10.1016/j.bbrc.2004.10.192 PMID: 15555595

[136] Wu, S.; Skolnick, J.; Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.,* **2007**, *5*, 17. http://dx.doi.org/10.1186/1741-7007-5-17 PMID: 17488521

[137] Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,* **2008**, *9*, 40. http://dx.doi.org/10.1186/1471-2105-9-40 PMID: 18215316

[138] Yang, J.; Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.,* **2015**, *43*(W1), W174-W181. http://dx.doi.org/10.1093/nar/gkv342 PMID: 25883148

[139] Simons, K.T.; Bonneau, R.; Ruczinski, I.; Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins,* **1999**, (Suppl. 3), 171-176. http://dx.doi.org/10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z PMID: 10526365

[140] Simons, K.T.; Strauss, C.; Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.,* **2001**, *306*(5), 1191-1199. http://dx.doi.org/10.1006/jmbi.2000.4459 PMID: 11237627

[141] Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C.E.; Baker, D. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins,* **2001**, (Suppl. 5), 119-126. http://dx.doi.org/10.1002/prot.1170 PMID: 11835488

[142] Peng, J.; Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins,* **2011**, *79*(Suppl 10), 161-171. http://dx.doi.org/10.1002/prot.23175 PMID: 21987485

[143] Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.,* **2012**, *7*(8),

1511-1522.
http://dx.doi.org/10.1038/nprot.2012.085 PMID: 22814390

[144] Källberg, M.; Margaryan, G.; Wang, S.; Ma, J.; Xu, J. RaptorX server: a resource for template-based protein structure modeling. *Methods Mol. Biol.,* **2014**, *1137*, 17-27.
http://dx.doi.org/10.1007/978-1-4939-0366-5_2 PMID: 24573471

[145] AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics,* **2019**, *35*(22), 4862-4865.
http://dx.doi.org/10.1093/bioinformatics/btz422 PMID: 31116374

[146] Heo, L.; Feig, M. High-accuracy protein structures by combining machine-learning with physics-based refinement. *Proteins,* **2020**, *88*(5), 637-642.
http://dx.doi.org/10.1002/prot.25847 PMID: 31693199

[147] Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D.T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). *Proteins,* **2019**, *87*(12), 1141-1148.
http://dx.doi.org/10.1002/prot.25834 PMID: 31602685

[148] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. How docking programs work. *Methods Mol. Biol.,* **2019**, *2053*, 35-50.
http://dx.doi.org/10.1007/978-1-4939-9752-7_3 PMID: 31452097

[149] Santos, L.H.S.; Ferreira, R.S.; Caffarena, E.R. Integrating molecular docking and molecular dynamics simulations. *Methods Mol. Biol.,* **2019**, *2053*, 13-34.
http://dx.doi.org/10.1007/978-1-4939-9752-7_2 PMID: 31452096

[150] van der Spoel, D.; van Maaren, P.J.; Caleman, C. GROMACS molecule & liquid database. *Bioinformatics,* **2012**, *28*(5), 752-753.
http://dx.doi.org/10.1093/bioinformatics/bts020 PMID: 22238269

[151] Sanbonmatsu, K.Y.; Tung, C.S. High performance computing in biology: multimillion atom simulations of nanoscale systems. *J. Struct. Biol.,* **2007**, *157*(3), 470-480.
http://dx.doi.org/10.1016/j.jsb.2006.10.023 PMID: 17187988

[152] Bitencourt-Ferreira, G.; de Azevedo, W.F. Jr. Molecular dynamics simulations with NAMD2. *Methods Mol. Biol.,* **2019**, *2053*, 109-124.
http://dx.doi.org/10.1007/978-1-4939-9752-7_8 PMID: 31452102

[153] de Azevedo, W.F. Jr. Molecular dynamics simulations of protein targets identified in *Mycobacterium tuberculosis. Curr. Med. Chem.,* **2011**, *18*(9), 1353-1366.
http://dx.doi.org/10.2174/092986711795029519 PMID: 21366529

[154] Fresnais, L.; Ballester, P.J. The impact of compound library size on the performance of scoring functions for structure-based virtual screening. *Brief. Bioinform.,* **2021**, *22*(3), bbaa095.
http://dx.doi.org/10.1093/bib/bbaa095 PMID: 32568385

[155] Chen, J.; Wang, X.; Zhu, T.; Zhang, Q.; Zhang, J.Z. A comparative insight into amprenavir resistance of mutations V32I, G48V, I50V, I54V, and I84V in HIV-1 protease based on thermodynamic integration and MM-PBSA methods. *J. Chem. Inf. Model.,* **2015**, *55*(9), 1903-1913.
http://dx.doi.org/10.1021/acs.jcim.5b00173 PMID: 26317593

[156] Chen, J.; Wang, X.; Pang, L.; Zhang, J.Z.H.; Zhu, T. Effect of mutations on binding of ligands to guanine riboswitch probed by free energy perturbation and molecular dynamics simulations. *Nucleic Acids Res.,* **2019**, *47*(13), 6618-6631.
http://dx.doi.org/10.1093/nar/gkz499 PMID: 31173143

[157] Chen, J.; Wang, J.; Yin, B.; Pang, L.; Wang, W.; Zhu, W. Molecular mechanism of binding selectivity of inhibitors toward BACE1 and BACE2 revealed by multiple short molecular dynamics simulations and free-energy predictions. *ACS Chem. Neurosci.,* **2019**, *10*(10), 4303-4318.
http://dx.doi.org/10.1021/acschemneuro.9b00348 PMID: 31545898