Send Orders for Reprints to reprints@benthamscience.net

Current Medicinal Chemistry, XXXX, *XX,* XX-XX

REVIEW ARTICLE

Computational Prediction of Binding Affinity for CDK2-ligand Complexes. A Protein Target for Cancer Drug Discovery

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

[1]*Western Michigan University, 1903 Western, Michigan Ave, Kalamazoo, MI 49008, 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*

Abstract: ***Background*:** CDK2 participates in the control of eukaryotic cell-cycle progression. Due to the great interest in CDK2 for drug development and the relative easiness in crystallizing this enzyme, we have over 400 structural studies focused on this protein target. This structural data is the basis for the development of computational models to estimate CDK2-ligand binding affinity.

***Objective*:** This work focuses on the recent developments in the application of supervised machine learning modeling to develop scoring functions to predict the binding affinity of CDK2.

***Method*:** We employed the structures available at the protein data bank and the ligand information accessed from the BindingDB, Binding MOAD, and PDBbind to evaluate the predictive performance of machine learning techniques combined with physical modeling used to calculate binding affinity. We compared this hybrid methodology with classical scoring functions available in docking programs.

***Results*:** Our comparative analysis of previously published models indicated that a model created using a combination of a mass-spring system and cross-validated Elastic Net to predict the binding affinity of CDK2-inhibitor complexes outperformed classical scoring functions available in AutoDock4 and AutoDock Vina.

***Conclusion*:** All studies reviewed here suggest that targeted machine learning models are superior to classical scoring functions to calculate binding affinities. Specifically for CDK2, we see that the combination of physical modeling with supervised machine learning techniques exhibits improved predictive performance to calculate the protein-ligand binding affinity. These results find theoretical support in the application of the concept of scoring function space.

ARTICLE HISTORY

Received: March 25, 2021
Revised: June 15, 2021
Accepted: June 22, 2021

DOI:
10.2174/0929867328666210806105810

Keywords: chemical space, physical modeling, CDK2, scoring function space, drug design, crystal structure, machine learning

1. INTRODUCTION

Evaluation of binding affinity data based on the structures of receptor-ligand complexes is an open problem in the application of docking simulations for drug discovery [1-5]. Considering the available structural data for a specific enzyme, we may use this information to understand the basis for enzymatic inhibition. The increase in the number of structures at the protein data bank (PDB) made available the experimental data necessary to analyze protein-ligand interactions crucial for the understanding of the molecular recognition process with a focus on the binding of drugs to receptor targets [6-8].

X-ray diffraction crystallography has been the most successful technique to determine the structures of protein-ligand complexes. Taking the complexes for which binding affinity data is available, we have over 99 % of

*Address correspondence to this author at the Pontifical Catholic University of Rio Grande do Sul (PUCRS). Av. Ipiranga, 6681 Porto Alegre/RS 90619-900 Brazil; Tel/Fax: ++55- 51-3320-3545; E-mails: walter.junior@pucrs.br; walter@azevedolab.net

0929-8673/XX $65.00+.00

the structural information generated by crystallography [5, 9].

Studying the physical basis of intermolecular interactions in protein systems, we know that the key aspects defining the molecular recognition process involve van der Waals contacts [10], electrostatic interactions [2, 11], hydrogen bonding [12], and entropy [13]. The most robust theoretical approach to calculate the energetics of intermolecular interaction is the application of quantum mechanics [14-21], where the intermolecular interactions can be evaluated with precision [22]. Quantum-mechanics approaches have been successful in drug discovery applications using protein-ligand docking simulations and scoring function development [17].

Considering the potential of quantum mechanics for drug discovery, we may highlight that in the future, the application of quantum computing methodologies and supervised-machine learning software to drug discovery will generate few false-positive leads in the application of docking screens for drug discovery [23]. In the opposition to quantum mechanics methods, we may approach intermolecular interactions of a drug and a macromolecule through molecular dynamics simulations of protein-ligand complexes [24-29]. Besides quantum mechanics and molecular dynamics, we may also address protein-ligand interactions through the training of machine learning models targeted to specific protein systems.

In this scenario, the study of intermolecular interactions through the combination of protein-ligand docking simulations and machine learning methods to develop targeted scoring functions has shown the potential of generating robust computational models to predict binding affinity [30-32]. These approaches are also adequate to assess the structural features responsible for the molecular recognition process. This type of integration of structural data and machine learning techniques has been successfully applied to a wide range of protein targets, such as cyclin-dependent kinases (EC 2.7.11.22) [33, 34], proteases [35-38], and more recently to SARS-CoV-2 drug targets [39-43].

In recent years, due to the availability of machine learning methods implemented in libraries using Python and R programming languages, we have witnessed a great number of computational tools devoted to generating models to calculate affinity based on the atomic coordinates of protein-ligand complexes. Among the recently published machine learning programs to estimate binding affinity or thermodynamic parameters, we may highlight the following computational tools: Statistical Analysis of Docking Results and Scoring Functions (SAnDReS) [44, 45], Tool to Analyze the Binding Affinity (Taba) [46, 47], Pafnucy [48], property-encoded shape distributions together with standard support vector machine (PESD-SVM) [49], Neural-Network-Based Scoring function (NNScore series) [50-52], and Random Forest Score (RF-Score series) [53-57].

In this review, we describe recent applications of machine learning methods to estimate the binding affinity of ligands against targets. These computational methods use experimental protein-ligand structures for which binding data is available. The synergism of crystal data and machine learning techniques paved the way to explore the scoring function space [9, 58, 59], which establishes a theoretical framework to address the challenging studies of protein-drug interactions. The development of a theoretical basis to address the creation of targeted scoring functions is the solid basis necessary to fortify the computational models designed for specific protein targets, making them much more than fortuitous statistical models to predict a biology response.

This scenario makes it clear that the study of complex systems found in cells targeted by drugs is viable to an abstraction brought about by the concept of scoring function space [9, 58, 59]. Here, we focus on the application of machine learning methods to crystallographic structures of cyclin-dependent kinase 2 (CDK2). This protein target has experimental information for three-dimensional structures and the binding affinity, which makes this system ideal for the development of targeted-scoring functions through the application of machine learning techniques.

2. METHODS

The PDB has recently reached over 175,000 structures (search carried out on March 24, 2021). This amount of structural information adds to the experimental data about binding affinity available at BindingDB [60, 61], Binding MOAD (Mother of All Databases) [62-64], and PDBbind [65, 66]. These three databases are integrated at the PDB, which allows us to perform searches to recover structures for which binding affinity or thermodynamic parameters are known.

To highlight the recent progress in the application of machine learning techniques, we describe computational models to predict the affinity of ligands for CDK2. To focus on previously published results of this protein class, we bring a recent update in the number of structures for which experimental binding affinity data is available.

2.1. CDK2Ki Dataset

We considered only CDK2 crystal structures for which experimental inhibition constant data is available. We updated a recently published CDK dataset [46], where we have not only CDK2 but also CDK9. We eliminated CDK9 data (search carried out on March 24, 2021). We show the selected PDB access codes in Table **1**. Supplementary material 1 brings the CDK2 inhibitor structures for all entries in the CDK2Ki dataset.

Table 1. PDB access codes of the CDK2Ki dataset.

Type of Dataset	PDB Access Codes
Training set	1E1X,1H1S,1OGU,1PXN,1PXP,2CLX, 2EXM,2FVD,3LFN,4ACM,4BCK, 4BCM,4BCN,4BCO,4BCP,4BCQ,4EOP, 4FKO,4NJ3,5D1J
Test set	1E1V,1JSV,1PXM,1PXO,1PYE,1V1K, 2XMY,2XNB,3LFS

As previously highlighted, we find binding affinity data for CDK2 in the BindingDB [57, 58], Binding MOAD (Mother of All Databases) [59-61], and PDBbind [62, 63]. The data about IC_{50} relies on a wide range of techniques to evaluate the binding. On the other hand, K_i focuses on a smaller set of experimental techniques, but there is no uniform method to address the ligand binding to CDK2 [46]. One possible potential technique to generate a more reliable experimental approach to calculate the binding would be to address the energetics of the CDK2-ligand interactions using isothermal microcalorimetry (ITC). Unfortunately, the experimental data about Gibbs free energy of binding for CDK2-ligand complexes using ITC is scarce [30, 31]. Due to these challenges, we focused our analysis on previously published machine learning modeling of K_i data. We eliminated repeated ligands and for CDK2-ligand complexes with more than one source of binding affinity, we chose the most recent published results.

2.2. Classical Scoring Functions

We calculated binding affinity using the atomic coordinates of the protein-ligand complexes available in the CDK2Ki dataset employing the classical scoring functions implemented in the docking programs AutoDock4 [67, 68] and AutoDock Vina (version 1.1.2) [69]. Ligand and protein atomic partial charges were assigned using the Partial Equalization of Orbital Electronegativities (PEOE) algorithm [70] employing AutoDockTools4 [67] (version 1.5.6). No protein-ligand docking simulations were carried out, the binding affinity calculation was based exclusively on the atomic coordinates of the crystallographic structures.

2.3. Combining Physical Modeling with Machine Learning

The program Taba (version 1.0) estimates ligand binding affinity based on an approach that models protein-ligand interactions as a mass-spring system [46, 47]. In this method, we consider that the key determinants for protein-ligand binding affinity are already registered in the three-dimensional structure of the complexes and estimate the energy of the system using a polynomial equation where each term (independent variable) of this expression considers an isolate mass-spring system composed of a potential equation for a pair of atoms.

Taba scoring function relies on simple modeling of protein-ligand energetics. We consider protein-ligand interaction as a mass-spring system, as delineated in Fig. (**1**). In Fig. (**1**), we see that the energetics of the intermolecular interactions are imprinted in the three-dimensional structure and model this complex net of interactions as independent mass-spring systems. In this representation, the energetics of the protein-ligand complex is the summation of each type of pair of atoms found in the structure. We express the potential energy of the protein system (V) by the following equation,

$$V(x,y,z) = \sum_i \sum_j \omega_{i,j}\left(d_{i,j} - d_{0,i,j}\right)^2 \quad (1)$$

In Equation (**1**), $\omega_{i,j}$ is the weight of each independent variable. We determine these weights through the application of machine-learning techniques. The double summation in equation (**1**) is taken over all protein (i) and ligand atoms (j) inside a defined volume of the structure. The term $d_{0,i,j}$ is the average interatomic distance for a given pair of atoms i and j, which is calculated for all structures in the training set. The program Taba calculates the terms ($\omega_{i,j}$, and $d_{0,i,j}$), taking all structures in the training set. The term $d_{i,j}$ is the Euclidean distance for a pair of atoms for one specific structure (not averaged for all structures) [46].

Taba reads the coordinates of all structures in a dataset and calculates the average distance involving the atoms in the protein (P) and the ligand (L). In this approach, we have average distances for different types of pairs of atoms, one for carbon (P)-carbon (L), another for oxygen (P)-nitrogen (L), and so on. In each pair of atoms (PL), we take one atom from the ligand and the second from the protein. Taba considers these average interatomic distances as the equilibrium distances of

our mass-spring system and determines the relative weights of each energy term using supervised machine learning techniques [71]. In the final model developed using Taba, we keep only the most relevant energy terms. In this review, we describe the machine learning models [72] developed for the CDK2Ki dataset. Taba takes an elegant combination of physical modeling with supervised machine learning techniques to address protein-ligand interactions. Fig. (**2**) outlines a schematic flowchart with the major steps of the Taba methodology [46].

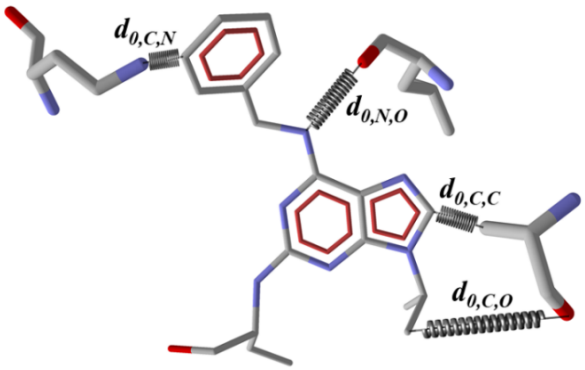

Fig. (1). Schematic representation of protein-ligand interactions as mass-spring systems. We employ $d_{0,i,j}$ to indicate the average interatomic distance for a given pair of atoms i and j. Thin lines represent covalent bonds for the ligand and the thicker lines indicate the amino acids in the protein *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

We briefly describe the Taba methodology, we start the application of Taba by defining a dataset of structures for which binding affinity data is available. Taba splits this dataset into training and test sets. For the structures in the training set, Taba calculates equilibrium distances, as previously defined. Taba employs this method to determine the weights of each independent variable defined in equation (**1**). Also, to avoid overfitting Taba employs standard k-fold cross-validation [46], where it splits the data into k subsets, called folds. In this method, Taba takes a five-fold cross-validation procedure. Taba employs training and test sets in the cross-validated supervised machine learning methods. Taba determines the predictive performance of each model based on correlation coefficients [46] and returns the best model.

2.4. SAnDReS

SAnDReS (version 1.1) is a suite of programs that aims to develop machine learning models based on the energy terms calculated by docking programs such as AutoDock4 [67, 68] and AutoDock Vina [69] in one computational tool [44, 45]. SAnDReS makes use of supervised machine learning methods available in the scikit-learn library [71] to generate polynomial empirical scoring functions to predict binding affinity [73-76]. These polynomial equations employ the energy terms calculated using the previously highlighted protein-ligand docking programs [67-69] and the crystallographic coordinates of protein and ligand. SAnDReS applies the same k-fold cross-validation approach described for the program Taba.

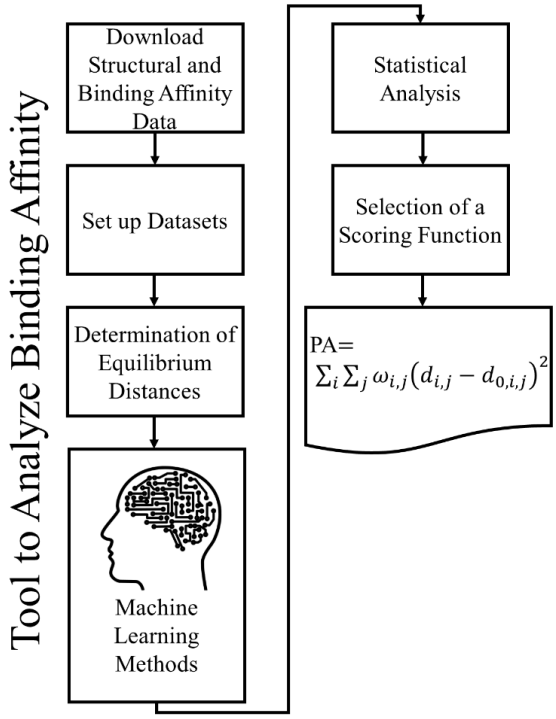

Fig. (2). Schematic flowchart for Taba methodology [46], where we combine physical modeling with machine learning methods. In the first two steps, we define the structures in the datasets. Taba automatically downloads structural and binding data from the PDB and binding affinity databases. In the next step, Taba determines the equilibrium distances for all structures in the training set. Taba considers protein-ligand interactions as a mass-spring system. Following this, Taba applies supervised machine learning to calculate the relative weights of each independent variable in Equation (**1**) to generate models to calculate the predicted affinity (PA) based on the atomic coordinates of protein-ligand complexes. Finally, Taba selects the scoring function based on the predictive performance against the structures in the test set.

2.5. Machine Learning Models

Taba and SAnDReS rely on scikit-learn [71] to generate machine learning models to predict binding affinity based on the atomic coordinates of protein-ligand complexes. In the SAnDReS approach, we have the development of machine learning models using energy

terms available in docking programs. On the other hand, Taba employs physical modeling of the intermolecular interactions. SAnDReS and Taba have the following supervised machine learning methods taken from scikit-learn [71]: Ridge, Lasso, Elastic Net, and Ordinary Linear Regression. For the first three methods, we have an additional option with cross-validation. Taken together, we have seven supervised machine-learning techniques in each program.

2.6. Statistical Analysis

We assessed the predictive power of the scoring functions calculating the correlation coefficients [77], *p*-values, and root-mean-squared error (RMSE) between the experimental data and the predicted binding affinity determined using the classical scoring functions [67-71], the empirical polynomial scoring functions developed using SAnDReS [44, 45], and the Taba mass-spring models [46, 47]. We generated the machine learning models using approximately 70 % of the structures in the CDK2Ki dataset (training set) and ~30 % of the dataset as a test set as suggested by Cichero *et al*. 2010 [78].

Taba uses four significant figures to express the interatomic distances to model protein systems. With this number of significant figures for distances, we have values with 1/1000 of Å as adopted in the PDB to express atomic coordinates of macromolecular structures [6, 7]. For interatomic distances, the X-ray diffraction crystallographic resolution is not the associated error in the atomic coordinates. These errors are not necessarily in the range of 0.001 Å for the atomic coordinates and distances, but using statistical analysis of experimental X-ray diffraction data and the final structure model such as the Luzzati plot, we have an associated error in the range 0.2 Å for a CDK2 with a resolution of 2.4 Å. For log (K_i), we adopted two significant figures, taking the experimental data for K_i from the binding databases [60-66].

3. RESULTS AND DISCUSSION

3.1. Biological System

In this review, we focus on the application of machine learning techniques to predict the binding affinity of ligands against structures of CDK2. This protein comprises an interesting biological system for the development of scoring functions for two main reasons. Firstly, the availability of crystallographic structures for which binding affinity data is known. Secondly, it is due to the importance of CDK2 for drug discovery and development [79]. CDK2 is a target for the development of anticancer drugs [80-83]. A search on clinicaltrials.gov using as keywords CDK2 and cancer returned 11 trials, including six which are either recruiting or active (search carried out on March 24, 2021). Among the CDK inhibitors identified so far, we may highlight the FDA-approved drug palbociclib, which can treat postmenopausal women with breast cancer [84-91].

Considering the filtered CDK2Ki dataset [46], where we removed the data related to CDK9 and eliminated ligands for which information about binding affinity from the PDB showed inconsistencies in the information associated with the PDBBind, BindingDB, and Binding MOAD, we ended up with 29 structures. These inconsistencies are related to different values of binding affinity for the same ligand.

In the CDK2Ki dataset, all crystallographic structures have competitive inhibitors non-covalently bound to the ATP-binding pocket of CDK2, with resolution ranging from 1.55 to 2.8 Å. The CDK2 has two domains with the N-terminal composed of a distorted beta-sheet and the C-terminal made mostly of alpha-helical structures as indicated in Fig. (**3**). The ATP-binding pocket of CDK2 lies between the two domains. All competitive inhibitors bind to this pocket. Calculation of the volume of this binding site using Molegro Virtual Docker (MVD) (version 6) [92-95] and a probe with a radius of 1.2 Å indicated a volume of 201.728 $Å^3$, which allows a wide range of different ligand structures to fit into this volume [82, 83].

In Fig. (**4**), we have the binding pocket of the structure of CDK2 in complex with roscovitine [96], where we highlight the two main residues of CDK2 participating in intermolecular interactions. Previously published intermolecular contact analyses of the residues participating in interactions involving inhibitors and the ATP-binding pocket indicated the participation of main-chain oxygen and nitrogen atoms of Leu 83 and Glu 81 of CDK2 in most complexes with high specific CDK2 inhibitors [97-112].

We have 415 structures of CDK2 deposited in the PDB (search carried out using UniProt Molecule Name as cyclin-dependent kinase 2 on March 24, 2021). Among these structures, 212 entries have validated inhibitors bound to the ATP-binding pocket of CDK2. Most of these ligands have data about IC_{50} and the minority about K_i. Analysis of the intermolecular interactions of these complex structures indicated that most of the inhibitors show intermolecular hydrogen bonds involving main-chain nitrogen and oxygen of Leu 83 and Glu 81, forming a sequence of spots for the binding of

inhibitors named the molecular fork [82]. In Fig. (**4**), we see two of these intermolecular hydrogen bonds with the participation of main-chain atoms of the residue Leu 83.

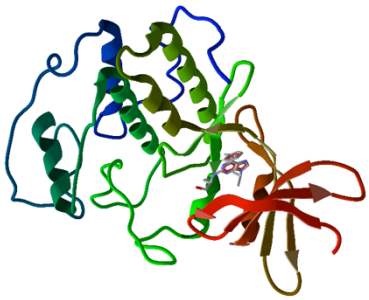

Fig. (3). Structure of human CDK2 in complex with the inhibitor roscovitine (PDB access code: 2A4L). The roscovitine is indicated with thick lines and the ribbons represent the protein structure. We used the program MVD (version 6) [92-95] to generate this figure *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

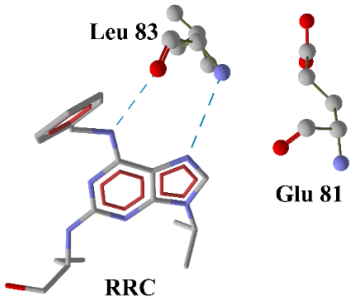

Fig. (4). Intermolecular hydrogen bonds of the inhibitor roscovitine (RRC) with the residue Leu 83 of the CDK2. Dashed lines indicate hydrogen bonds. On the right, we have the structure of the residue Glu 81 that participates in intermolecular hydrogen bonds in other CDK2-inhibitor structures. We used the program MVD (version 6) [92-95] to generate this figure *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

3.2. Binding Affinity with Classical Scoring Function

AutoDock4 and AutoDock Vina have been successfully applied to different protein targets to identify potential hits. Their scoring functions are fully described in the following references [113-115]. On the other hand, the evaluation of binding affinity using the available scoring functions in these programs was not reliable [46, 47, 116-118]. Application of AutoDock4 to predict binding affinity using the crystallographic position of the ligands in the CDK2Ki (training set) generated a Spearman rank correlation coefficient (ρ) of 0.358 with a *p*-value of 0.12. Analysis of the predictive power of each energy term used in the AutoDock4 scoring function generated ρ ranging from -0.348 to 0.359, all with *p*-values > 0.1. Analysis of the predictive performance for the structures in the test set produced ρ ranging from -0.183 to 0.367, all with *p*-values > 0.1. Supplementary materials 2 and 3 bring the predicted and experimental binding affinities for all structures in the CDK2Ki training and test sets, respectively.

Assessment of the binding affinity of the structures in the training set using a full scoring function and energy terms available in AutoDock Vina generated ρ ranging from -0.171 to 0.224, with *p*-values > 0.1. The highest correlation was obtained for the full scoring function of AutoDock Vina. Analysis of the correlation for the structures in the test set showed ρ ranging from -0.417 to 0.117, with *p*-values > 0.1. Supplementary materials 4 and 5 have the binding affinities calculated using AutoDock Vina for all structures in the CDK2Ki training and test sets, respectively.

The predictive performance of both classical scoring functions is poor. One possible reason for this failure in predicting binding affinity using classical scoring functions is the methodology applied in the creation of these computational models. We may highlight that most of the classical scoring functions use energy terms for van der Waals, electrostatic energy, hydrogen bonding, and solvation effects and then determine the relative weight of each energy term based on the regression method [1, 119-124]. Such an approach creates a model bias against the structures not employed in the training set so that these computational models to predict binding affinity are prone to work for proteins present in the training set used to determine the relative weights of each energy term in the empirical scoring function. On the other hand, protein systems not present in the original training set or poorly represented in it could be out of the scope of the classical scoring function [44, 45], which generates a low correlation with experimental data as observed for the structures in the CDK2Ki dataset.

3.3. Binding Affinity with SAnDReS

SAnDReS aims to integrate all necessary steps to create machine learning models in one suite of programs [44]. SAnDReS has been applied to a wide range of different protein systems [125-143] and has successfully generated machine learning models that outperform classical scoring functions in the prediction of binding affinity [45, 144-151].

Application of the machine learning methods of SAnDReS to the structures in the training set and the energy terms calculated using the program AutoDock4 generated polynomials equations with three independent variables (features in the machine learning terminology), which gives more than six observations for each independent variable. These polynomial scoring equations are described elsewhere [125-143]. The ratio of five observations (structures) per independent variable (or descriptor) is the minimum requested by the rule of thumb recommended for regressions models [152, 153]. Amongst the seven machine learning methods available in SAnDReS, the Elastic Net with cross-validation showed the best overall predictive performance. The highest correlation model has an $\rho = 0.319$, a *p*-value of 0.17, and an RMSE = 1.1, a correlation worse than the classical scoring functions. The correlation for the structures in the test set was also poor, with an $\rho = -0.183$, *p*-value = 0.64, and an RMSE of 1.7. Fig. (**5a**) shows the scattering plot for the test set structures. The polynomial equation for the predicted affinity (PA) generated using SAnDReS is shown in equation (**2**),

$$\text{PA (model 1)} = -6.5 - 0.0416(\text{Final Intermolecular Energy}) + 0.0416(\text{vdW} + \text{Hbond} + \text{Desolvation Energy}) + 0.192(\text{Final Total InternalEnergy}) \quad (2)$$

In equation (**2**), the variables vdW and Hbond represent van der Waals and hydrogen bond energy terms, respectively. A detailed description of the expression of each energy term is available elsewhere [67].

Using the same approach for the energy terms calculated using AutoDock Vina, we have the highest correlation model with an $\rho = 0.537$, a *p*-value of 0.015, and RMSE = 1.0. The expression of this machine learning model is in equation (**3**), as follows,

$$\text{PA (model 2)} = -3.2 - 0.00288(\text{Gauss2}) + 0.00982(\text{Repulsion}) + 0.0220(\text{Hydrophobic}) \quad (3)$$

The descriptions for the energy terms in equation (**3**) are presented elsewhere [69]. Analysis of this polynomial model against the structures in the test indicated an $\rho = 0.067$, a *p*-value of 0.86, and an RMSE = 2.0. In Fig. (**5b**), we have the scattering plot for the test set structures. Although the model generated using the energy terms of the AutoDock Vina showed some promising results for the training set, the evaluation against the test set showed poor predictive power.

3.4. Binding Affinity with Taba

Taba addresses protein-drug interactions as a mass-spring system and combines it with an integrated application of supervised machine learning techniques to generate a targeted scoring function where the independent variables are mass-spring micro-systems composed of pairs of atoms. A previously published study using this approach to generate a computational model calibrated for CDK structures [46, 47] was able to predict binding affinity with superior performance compared with classical scoring functions.

In the present study, we focused on a previously described CDK machine learning model [46] filtered for CDK2 structures to create a CDK2-targeted scoring function. In this model, we deleted the CDK9 data to have only CDK2 structures. We used a polynomial equation with three independent variables taking as energy terms the contributions of the following pairs of atoms: C-C, C-S, and O-O. Each independent variable is a mass-spring potential energy function. The equilibrium distances for each pair of atoms were calculated using the average distance taking all structures in the training set. We tested seven machine learning approaches available in Taba, and the Elastic Net with cross-validation also showed the highest correlation between experimental and predicted affinities. The machine learning model determined using Taba has the following expression (equation (**4**)),

$$\text{PA (model 3)} = \omega_0 + \omega_1(d_{i,j} - d_{0,C,C})^2 + \omega_2(d_{i,j} - d_{0,C,S})^2 + \omega_3(d_{i,j} - d_{0,O,O})^2 \quad (4)$$

In equation (**4**), the weights are the following: $\omega_0 = -6.6$, $\omega_1 = 0.132$, $\omega_2 = 0.461$, and $\omega_3 = 0.226$. The equilibrium distances have the following values: $d_{0,C,C} = 4.078$ Å, $d_{0,C,S} = 4.120$ Å, and $d_{0,O,O} = 3.663$ Å.

For the training set, the Taba model has a correlation $\rho = 0.750$ with a *p*-value = 0.0001 and RMSE = 5.1. For the test set, we have $\rho = 0.817$ with a *p*-value = 0.007 and a RMSE = 5.7. Supplementary materials 6 and 7 bring the predicted binding affinity for the training and test sets, respectively. In Fig. (**5c**) we have the scattering plot for the test set structures.

Considering the correlation, the Taba model showed the best predictive performance, compared with the classical scoring functions and the SAnDReS machine learning models. Nevertheless, the RMSE values for training and test sets are relatively high for the Taba model, over 5.0. RMSE values of the SAnDReS models were all below 2.1. This might be due to the simplicity of the mass-spring approach to protein-ligand interactions.

Taking the classical scoring functions and the three machine learning models, we may say that these models show some potential but failed in at least one key aspect of the statistical analysis of the predictive per-

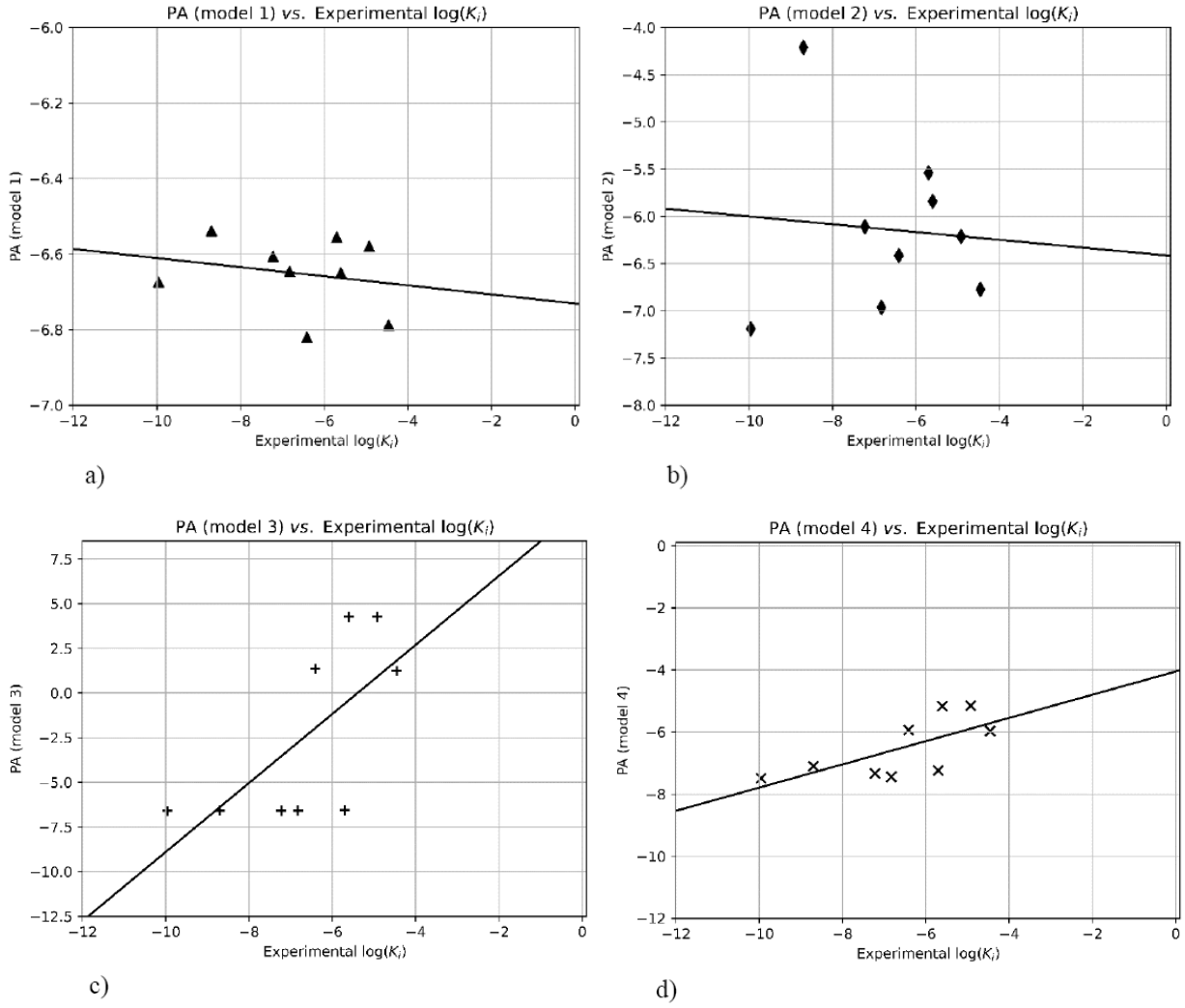

Fig. (5). Scatter plots for predicted and experimental binding affinities. **a**) PA generated using energy terms available in AutoDock4 with machine learning modeling performed with SAnDReS (model 1) (ρ = -0.183, *p*-value = 0.64, and RMSE = 1.7). **b**) PA generated using energy terms available in AutoDock Vina with regression modeling carried out with SAnDReS (model 2) (ρ = 0.067, *p*-value = 0.86, and RMSE = 2.0). **c**) Mass-spring model generated using Taba (model 3) (ρ = 0.817, *p*-value = 0.007, and RMSE = 5.7). d) Machine learning model involving the three previous models performed using SAnDReS (model 4) (ρ = 0.733, *p*-value = 0.03, and RMSE = 1.3). We used cross-validated Elastic Net to generate all machine learning models. We used the program SAnDReS [44] to generate all plots in this figure.

formance. Using the suite of programs SAnDReS, we developed a novel scoring function (model 4) considering the models generated using terms from AutoDock4 (model 1), AutoDock Vina (model 2), and Taba (model 3). Applying the cross-validated Elastic Net method taking the previously generated models, we have the following expression (equation (**5**)),

$$\text{PA (model 4)} = -0.36 + 0.770\ \text{PA(model 1)} + 0.0931\ \text{PA(model 2)} + 0.200\ \text{PA(model 3)} \quad (\mathbf{5})$$

Taking the training set, the new machine learning model (model 4) has a ρ = 0.750 with a *p*-value = 0.0001, and an RMSE = 0.7. For the test set, we have an ρ = 0.733 with a *p*-value = 0.03, and an RMSE = 1.3. In Fig. (**5d**), we have the scattering plot for the test set structures. The correlation for the training set is the same and for the test set, we have a worse result, when compared with model 3. Taking the RMSE, we observe a significant improvement of model 4 in the training and test sets. This progress in model 4 is due to the addition of terms for electrostatics, desolvation, and hydrogen bonding, not present in model 3.

These differences in the predictive performance of the machine learning models should always be considered in the context where we applied it and keeping in mind the limitations of the training sets for protein systems. We chose to focus on structures for which experimental data for atomic coordinates and inhibition con-

stants are available. This criterion limits the ratio observations per independent variable but creates machine learning models strictly based on robust experimental information. Also, although we have a poor ratio of observations per independent variable from the machine learning point of view, considering the criteria used for modeling scoring functions to assess binding affinitybased on atomic coordinates, we satisfy a well-established rule of thumb [152, 153]. In summary, considering RMSE and *p*-value, model 4 exhibits the best overall performance for the CDK2Ki dataset.

3.5. Scoring Function Space

The success of the application of targeted-scoring functions to predict binding affinity established the basis for the creation of a mathematical abstraction for the development of computational models to address protein-ligand interactions [9, 58]. Taking a systems-level approach to address this problem, we may investigate the relation involving the chemical [126, 154-161] and protein [162, 163] spaces. Defining a subset of the chemical space as formed by the inhibitors of a specific enzyme and seeing this protein as an element of the protein space, we may envisage this relation as a base to search the scoring function space. This mathematical space has all potential computational models able to predict the binding affinity taking as input the atomic coordinates of protein-ligand complexes. We apply machine learning methods to identify an adequate function to predict binding affinity for an element of the protein space considering the relation with a subset of the chemical space.

Fig. (**6**) illustrates the relations involving protein, chemical, scoring function spaces. We consider CDK2 as an element of the protein space. We highlight a subset in the chemical space composed of CDK2 inhibitors. Then, we may use machine learning approaches to identify a model to predict binding affinity for this enzyme [9, 46-54]. With this mathematical abstraction, we have a solid theoretical background to explain the superior predictive performances of machine learning models developed using SAnDReS [44] and Taba [46], when compared with classical scoring functions. Targeted scoring functions are the results of explorations of the scoring function space. So, we define their functions for a single protein.

One way to think about this abstraction is taking the experimental binding and the crystallographic data available for a given protein as a system, where through the application of machine learning methods we create a computational model tailored to this biological system. In doing so, we give up to find a general scoring function for all proteins; we address this

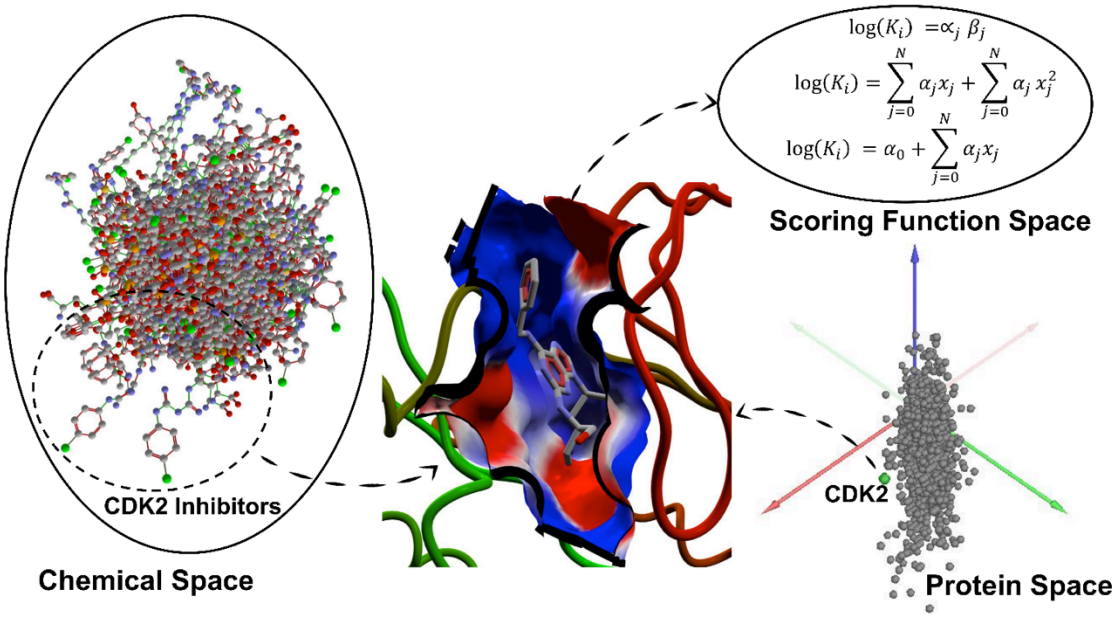

Fig. (6). Schematic diagram illustrating the relations involving protein, chemical, and scoring function spaces. On the right, we take an element of the protein space, indicated by the green sphere. This element is the CDK2. Then, we highlight a subset of the chemical space composed of CDK2 inhibitors. Finally, we apply machine learning techniques to explore the scoring function space to find an adequate model to predict the binding affinity. We used the program MVD (version 6) [92-95] to generate the chemical and protein spaces in this figure *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

problem by generating a fine-tuned computational model. This approach seems realistic considering the restricted volume of experimental data, especially seeing the complex structures for which experimental binding affinity is available. Also, we assume that as being proteins dependent on the evolution and integrated into a complex chemical environment, as found in the biological systems, the use of a targeted machine-learning model is suitable to predict binding affinity. Besides the studies reviewed here, other authors carried out machine-learning studies focused on CDK [164, 165].

CONCLUSION

The application of machine learning methods for the development of empirical scoring functions to predict protein-ligand binding affinity gave support to the use of these techniques to address the energetics of these molecular systems. The superior predictive performance of targeted scoring served as the basis for the development of the concept of scoring function space [9]. This mathematical abstraction makes it possible to integrate computational systems biology with machine learning techniques to address protein-ligand interactions. By the use of supervised machine learning techniques, we can explore this scoring function space to build a computational model targeted to a specific protein system. Here, we highlighted the superior predictive power of supervised machine learning approaches when compared to classical scoring functions using CDK2 as an example. Specifically for CDK2, machine learning models outperform classical scoring functions available in protein-ligand docking programs (AutoDock4 and AutoDock Vina). Although target-specific scoring functions show superior predictive performance compared with generalized approaches, there are two major weaknesses of this approach. Targeted machine learning models capture the essence of the binding focused on a specific pocket. Therefore allosteric ligands exhibiting a different binding mode would require another targeted machine learning model for the same enzyme, and most surely, protein targets with highly flexible binding pockets would prove to be very challenging to cope with since the training of targeted scoring function rely exclusively on crystallographic structures, at least for the methods we presented here.

LIST OF ABBREVIATIONS

CDK2	=	Cyclin-dependent Kinase 2
CDK2Ki	=	CDK2 Dataset for which Ki is Known
CDK9	=	Cyclin-dependent Kinase 9
FDA	=	Food and Drug Administration
K_i	=	Inhibition Constant
ITC	=	Isothermal Microcalorimetry
MOAD	=	Mother of All Databases
MVD	=	Molegro Virtual Docker
NNScore	=	Neural-network-based Scoring Function
PA	=	Predicted Affinity
PDB	=	Protein Data Bank
PESD-SVM	=	Property-encoded Shape dIstributions Together with Standard support Vector Machine
ρ	=	Spearman Rank Correlation Coefficient
RF-Score	=	Random Forest Score
RMSE	=	root-mean-square Error
SAnDReS	=	Statistical Analysis of Docking Results and Scoring Functions
Taba	=	Tool to Analyze the Binding Affinity

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

WFA 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.

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, this manuscript would not be possible. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001. MVA acknowledges the receipt of the Dieter H. Haenicke Scholarship (Haenicke Institute for Global Education).

SUPPLEMENTARY MATERIAL

Supplementary material can be found on the publishers website along with the published article.

REFERENCES

[1] Roviello, V.; Musumeci, D.; Mokhir, A.; Roviello, G.N. Evidence of protein binding by a nucleopeptide based on a thymine-decorated L-diaminopropanoic acid through CD and *in silico* studies. *Curr. Med. Chem.,* **2021**, *28*(24), 5004-5015. http://dx.doi.org/10.2174/0929867328666210201152326 PMID: 33593247

[2] Bitencourt-Ferreira, G.; de Azevedo, W.F.Jr. Electrostatic potential energy in protein-drug complexes. *Curr. Med. Chem.,* **2021**, *28*(24), 4954-4971. http://dx.doi.org/10.2174/0929867328666210201150842 PMID: 33593246

[3] Bondžić, A.M.; Vasić Anićijević, D.D.; Janjić, G.V.; Zeković, I.; Momić, T.; Nikezić, A.V.; Vasić, V.M. Na, K-ATPase as a biological target for gold(III) complexes: a theoretical and experimental approach. *Curr. Med. Chem.,* **2021**, *28*(23), 4742-4798. http://dx.doi.org/10.2174/0929867328999210101233801 PMID: 33397227

[4] Sulimov, V.B.; Kutov, D.C.; Sulimov, A.V. Advances in docking. *Curr. Med. Chem.,* **2019**, *26*(42), 7555-7580. http://dx.doi.org/10.2174/0929867325666180904115000 PMID: 30182836

[5] Veit-Acosta, M.; de Azevedo, W.F.Jr. The impact of crystallographicdata for the development of machine learning models to predict protein-ligand binding affinity. *Curr. Med. Chem.,* **2021**. Online ahead of print. http://dx.doi.org/10.2174/0929867328666210210121320 PMID: 33568025

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] Peters, M.B.; Raha, K.; Merz, K.M.Jr. Quantum mechanics in structure-based drug design. *Curr. Opin. Drug Discov. Devel.,* **2006**, *9*(3), 370-379. PMID: 16729734

[15] Gupta, A.; Kumar, V.; Aparoy, P. Role of topological, electronic, geometrical, constitutional and quantum chemical based descriptors in QSAR: mPGES-1 as a case study. *Curr. Top. Med. Chem.,* **2018**, *18*(13), 1075-1090. http://dx.doi.org/10.2174/1568026618666180719164149 PMID: 30027847

[16] Cavasotto, C.N.; Adler, N.S.; Aucar, M.G. Quantum chemical approaches in structure-based virtual screening and lead optimization. *Front Chem.,* **2018**, *6*, 188. http://dx.doi.org/10.3389/fchem.2018.00188 PMID: 29896472

[17] Crespo, A.; Rodriguez-Granillo, A.; Lim, V.T. Quantum-mechanics methodologies in drug discovery: applications of docking and scoring in lead optimization. *Curr. Top. Med. Chem.,* **2017**, *17*(23), 2663-2680. http://dx.doi.org/10.2174/1568026617666170707120609 PMID: 28685695

[18] Barbault, F.; Maurel, F. Simulation with quantum mechanics/molecular mechanics for drug discovery. *Expert Opin. Drug Discov.,* **2015**, *10*(10), 1047-1057. http://dx.doi.org/10.1517/17460441.2015.1076389 PMID: 26289577

[19] Habgood, M.; James, T.; Heifetz, A. Conformational searching with quantum mechanics. *Methods Mol. Biol.,* **2020**, *2114*, 207-229. http://dx.doi.org/10.1007/978-1-0716-0282-9_14 PMID: 32016896

[20] Heifetz, A.; Townsend-Nicholson, A. Characterizing rhodopsin-arrestin interactions with the fragment molecular orbital (FMO) method. *Methods Mol. Biol.,* **2020**, *2114*, 177-186. http://dx.doi.org/10.1007/978-1-0716-0282-9_12 PMID: 32016894

[21] Świderek, K.; Tuñón, I.; Moliner, V.; Bertran, J. Computational strategies for the design of new enzymatic functions. *Arch. Biochem. Biophys.,* **2015**, *582*, 68-79. http://dx.doi.org/10.1016/j.abb.2015.03.013 PMID: 25797438

[22] Morao, I.; Heifetz, A.; Fedorov, D.G. Accurate scoring in seconds with the fragment molecular orbital and density-functional tight-binding methods. *Methods Mol. Biol.,* **2020**, *2114*, 143-148. http://dx.doi.org/10.1007/978-1-0716-0282-9_9 PMID: 32016891

[23] Thomford, N.E.; Senthebane, D.A.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int. J. Mol. Sci.,* **2018**, *19*(6), 1578.

http://dx.doi.org/10.3390/ijms19061578 PMID: 29799486

[24] 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

[25] Sforça, M.L.; Oyama, S., Jr; Canduri, F.; Lorenzi, C.C.; Pertinhez, T.A.; Konno, K.; Souza, B.M.; Palma, M.S.; Ruggiero Neto, J.; Azevedo, W.F.Jr.; Spisni, A. How C-terminal carboxyamidation alters the biological activity of peptides from the venom of the eumenine solitary wasp. *Biochemistry,* **2004**, *43*(19), 5608-5617. http://dx.doi.org/10.1021/bi0360915 PMID: 15134435

[26] Hernández-Rodríguez, M.; Rosales-Hernández, M.C.; Mendieta-Wejebe, J.E.; Martínez-Archundia, M.; Basurto, J.C. Current tools and methods in molecular dynamics (MD) simulations for drug design. *Curr. Med. Chem.,* **2016**, *23*(34), 3909-3924. http://dx.doi.org/10.2174/0929867323666160530144742 PMID: 27237821

[27] de Azevedo, W.F.Jr.; Canduri, F.; Fadel, V.; Teodoro, L.G.; Hial, V.; Gomes, R.A. Molecular model for the binary complex of uropepsin and pepstatin. *Biochem. Biophys. Res. Commun.,* **2001**, *287*(1), 277-281. http://dx.doi.org/10.1006/bbrc.2001.5555 PMID: 11549287

[28] Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.,* **2005**, *26*(16), 1781-1802. http://dx.doi.org/10.1002/jcc.20289 PMID: 16222654

[29] 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

[30] Singh, A.V.; Rosenkranz, D.; Ansari, M.H.D.; Singh, R.; Kanase, A.; Singh, S.P.; Johnston, B.; Tentschert, J.; Laux, P.; Luch, A. Artificial intelligence and machine learning empower advanced biomedical material design to toxicity prediction. *Adv. Intell. Syst,* **2020**, *2*, 2000084. http://dx.doi.org/10.1002/aisy.202000084

[31] Singh, A.V.; Maharjan, R.S.; Kanase, A.; Siewert, K.; Rosenkranz, D.; Singh, R.; Laux, P.; Luch, A. Machine-learning-based approach to decode the influence of nanomaterial properties on their interaction with cells. *ACS Appl. Mater. Interfaces,* **2021**, *13*(1), 1943-1955. http://dx.doi.org/10.1021/acsami.0c18470 PMID: 33373205

[32] Singh, A.V.; Ansari, M.H.D.; Rosenkranz, D.; Maharjan, R.S.; Kriegel, F.L.; Gandhi, K.; Kanase, A.; Singh, R.; Laux, P.; Luch, A. Artificial intelligence and machine learning in computational nanotoxicology: unlocking and empowering nanomedicine. *Adv. Healthc. Mater.,* **2020**, *9*(17), e1901862. http://dx.doi.org/10.1002/adhm.201901862 PMID: 32627972

[33] 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

[34] 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

[35] 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

[36] Yang, Y.; Lu, J.; Yang, C.; Zhang, Y. Exploring fragment-based target-specific ranking protocol with machine learning on cathepsin S. *J. Comput. Aided Mol. Des.,* **2019**, *33*(12), 1095-1105. http://dx.doi.org/10.1007/s10822-019-00247-3 PMID: 31729618

[37] Li, F.; Wang, Y.; Li, C.; Marquez-Lago, T.T.; Leier, A.; Rawlings, N.D.; Haffari, G.; Revote, J.; Akutsu, T.; Chou, K.C.; Purcell, A.W.; Pike, R.N.; Webb, G.I.; Ian Smith, A.; Lithgow, T.; Daly, R.J.; Whisstock, J.C.; Song, J. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Brief. Bioinform.,* **2019**, *20*(6), 2150-2166. http://dx.doi.org/10.1093/bib/bby077 PMID: 30184176

[38] Pethe, M.A.; Rubenstein, A.B.; Khare, S.D. Large-scale structure-based prediction and identification of novel protease substrates using computational protein design. *J. Mol. Biol.,* **2017**, *429*(2), 220-236. http://dx.doi.org/10.1016/j.jmb.2016.11.031 PMID: 27932294

[39] Kabra, R.; Singh, S. Evolutionary artificial intelligence based peptide discoveries for effective Covid-19 therapeutics. *Biochim. Biophys. Acta Mol. Basis Dis.,* **2021**, *1867*(1), 165978. http://dx.doi.org/10.1016/j.bbadis.2020.165978 PMID: 32980462

[40] 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

[41] 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

[42] Gao, K.; Nguyen, D.D.; Chen, J.; Wang, R.; Wei, G.W. Repositioning of 8565 existing drugs for COVID-19. *J. Phys. Chem. Lett.,* **2020**, *11*(13), 5373-5382. http://dx.doi.org/10.1021/acs.jpclett.0c01579 PMID: 32543196

[43] Onawole, A.T.; Sulaiman, K.O.; Kolapo, T.U.; Akinde, F.O.; Adegoke, R.O. COVID-19: CADD to the rescue. *Virus Res.,* **2020**, *285*, 198022. http://dx.doi.org/10.1016/j.virusres.2020.198022 PMID: 32417181

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] Li, H.; Leung, K.-S.; Wong, M.-H. The impact of docking pose generation error on the prediction of binding affinity. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*; DI Serio, C.; Liò, P.; Nonis, A.; Tagliaferri, R., Eds.; Springer: Cham, **2015**, pp. 231-241. https://doi.org/10.1007/978-3-319-24462-4_20

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] 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. https://doi.org/10.1093/nar/gkm911 PMID: 18055497

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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.

https://doi.org/10.1002/jcc.21334 PMID: 19499576

[70] Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron,* **1980**, *36*(22), 3219-3228. http://dx.doi.org/10.1016/0040-4020(80)80168-2

[71] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Verplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikitlearn: machine learning in python. *J. Mach. Learn. Res.,* **2011**, *12*, 2825-2830.

[72] Zou, H.; Hastie, T. Regularization and variable selection *via* the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.,* **2005**, *67*(2), 301-220. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x

[73] 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

[74] 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

[75] Ducati, R.G.; Basso, L.A.; Santos, D.S.; de Azevedo, W.F.Jr. Crystallographic and docking studies of purine nucleoside phosphorylase from *Mycobacterium tuberculosis*. *Bioorg. Med. Chem.,* **2010**, *18*(13), 4769-4774. http://dx.doi.org/10.1016/j.bmc.2010.05.009 PMID: 20570524

[76] de Azevedo, W.F.Jr.; Dias, R. Experimental approaches to evaluate the thermodynamics of protein-drug interactions. *Curr. Drug Targets,* **2008**, *9*(12), 1071-1076. http://dx.doi.org/10.2174/138945008786949441 PMID: 19128217

[77] 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

[78] Cichero, E.; Cesarini, S.; Mosti, L.; Fossa, P. CoMFA and CoMSIA analyses on 1,2,3,4-tetrahydropyrrolo[3,4-b]indole and benzimidazole derivatives as selective CB2 receptor agonists. *J. Mol. Model.,* **2010**, *16*(9), 1481-1498. http://dx.doi.org/10.1007/s00894-010-0664-1 PMID: 20174844

[79] Wang, S.; Griffiths, G.; Midgley, C.A.; Barnett, A.L.; Cooper, M.; Grabarek, J.; Ingram, L.; Jackson, W.; Kontopidis, G.; McClue, S.J.; McInnes, C.; McLachlan, J.; Meades, C.; Mezna, M.; Stuart, I.; Thomas, M.P.; Zheleva, D.I.; Lane, D.P.; Jackson, R.C.; Glover, D.M.; Blake, D.G.; Fischer, P.M. Discovery and characterization of 2-anilino-4-(thiazol-5-yl)pyrimidine transcriptional CDK inhibitors as anticancer agents. *Chem. Biol.,* **2010**, *17*(10), 1111-1121. http://dx.doi.org/10.1016/j.chembiol.2010.07.016 PMID: 21035734

[80] Tadesse, S.; Anshabo, A.T.; Portman, N.; Lim, E.; Tilley, W.; Caldon, C.E.; Wang, S. Targeting CDK2 in cancer: challenges and opportunities for therapy. *Drug Discov. Today,* **2020**, *25*(2), 406-413. http://dx.doi.org/10.1016/j.drudis.2019.12.001 PMID: 31839441

[81] Volkart, P.A.; Bitencourt-Ferreira, G.; Souto, A.A.; de Azevedo, W.F. Cyclin-dependent Kinase 2 in cellular senescence and cancer. A structural and functional review. *Curr. Drug Targets,* **2019**, *20*(7), 716-726. http://dx.doi.org/10.2174/1389450120666181204165344 PMID: 30516105

[82] Levin, N.M.B.; Pintro, V.O.; de Ávila, M.B.; de Mattos, B.B.; De Azevedo, W.F.Jr. Understanding the structural basis for inhibition of cyclin-dependent kinases. New pieces in the molecular puzzle. *Curr. Drug Targets,* **2017**, *18*(9), 1104-1111. http://dx.doi.org/10.2174/1389450118666161116130155 PMID: 27848884

[83] de Azevedo, W.F.Jr. Opinion paper: targeting multiple cyclin-dependent kinases (CDKs): a new strategy for molecular docking studies. *Curr. Drug Targets,* **2016**, *17*(1), 2. http://dx.doi.org/10.2174/1389450117011512171009O7 PMID: 26687602

[84] Pondé, N.; Wildiers, H.; Awada, A.; de Azambuja, E.; Deliens, C.; Lago, L.D. Targeted therapy for breast cancer in older patients. *J. Geriatr. Oncol.,* **2020**, *11*(3), 380-388. http://dx.doi.org/10.1016/j.jgo.2019.05.012 PMID: 31171494

[85] Schoninger, S.F.; Blain, S.W. The ongoing search for biomarkers of CDK4/6 inhibitor responsiveness in breast cancer. *Mol. Cancer Ther.,* **2020**, *19*(1), 3-12. http://dx.doi.org/10.1158/1535-7163.MCT-19-0253 PMID: 31909732

[86] Yuan, L.; Alexander, P.B.; Wang, X.F. Cellular senescence: from anti-cancer weapon to anti-aging target. *Sci. China Life Sci.,* **2020**, *63*(3), 332-342. http://dx.doi.org/10.1007/s11427-019-1629-6 PMID: 32060861

[87] Frassoldati, A.; Biganzoli, L.; Bordonaro, R.; Cinieri, S.; Conte, P.; Laurentis, M.; Mastro, L.D.; Gori, S.; Lauria, R.; Marchetti, P.; Michelotti, A.; Montemurro, F.; Naso, G.; Pronzato, P.; Puglisi, F.; Tondini, C.A. Endocrine therapy for hormone receptor-positive, HER2-negative metastatic breast cancer: extending endocrine sensitivity. *Future Oncol.,* **2020**, *16*(5), 129-145. http://dx.doi.org/10.2217/fon-2018-0942 PMID: 31849236

[88] Tamura, K. Differences of cyclin-dependent kinase 4/6 inhibitor, palbociclib and abemaciclib, in breast cancer. *Jpn. J. Clin. Oncol.,* **2019**, *49*(11), 993-998. http://dx.doi.org/10.1093/jjco/hyz151 PMID: 31665472

[89] Rozeboom, B.; Dey, N.; De, P. ER+ metastatic breast cancer: past, present, and a prescription for an apoptosis-targeted future. *Am. J. Cancer Res.,* **2019**, *9*(12), 2821-2831. PMID: 31911865

[90] Bonelli, M.; La Monica, S.; Fumarola, C.; Alfieri, R. Multiple effects of CDK4/6 inhibition in cancer: from cell cycle arrest to immunomodulation. *Biochem. Pharmacol.,* **2019**, *170*, 113676. http://dx.doi.org/10.1016/j.bcp.2019.113676 PMID: 31647925

[91] Grizzi, G.; Ghidini, M.; Botticelli, A.; Tomasello, G.; Ghidini, A.; Grossi, F.; Fusco, N.; Cabiddu, M.; Savio, T.; Petrelli, F. Strategies for increasing the effectiveness of aromatase inhibitors in locally advanced breast cancer: an evidence-based review on current options. *Cancer Manag. Res.,* **2020**, *12*, 675-686. http://dx.doi.org/10.2147/CMAR.S202965 PMID: 32099464

[92] 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

[93] 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

[94] 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

[95] 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

[96] 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

[97] 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

[98] Canduri, F.; Perez, P.C.; Caceres, R.A.; de Azevedo, W.F.Jr. CDK9 a potential target for drug development. *Med. Chem.,* **2008**, *4*(3), 210-218. http://dx.doi.org/10.2174/157340608784325205 PMID: 18473913

[99] Canduri, F.; de Azevedo, W.F.Jr. Structural basis for interaction of inhibitors with cyclin-dependent kinase 2. *Curr. Comput. Aided Drug Des,* **2005**, *1*(1), 53-64. http://dx.doi.org/10.2174/1573409052952233

[100] Canduri, F.; Uchoa, H.B.; de Azevedo, W.F.Jr. Molecular models of cyclin-dependent kinase 1 complexed with inhibitors. *Biochem. Biophys. Res. Commun.,* **2004**, *324*(2), 661-666. http://dx.doi.org/10.1016/j.bbrc.2004.09.109 PMID: 15474478

[101] 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

[102] Kim, S.H.; Schulze-Gahmen, U.; Brandsen, J.; de Azevedo Júnior, W.F. Structural basis for chemical inhibition of CDK2. *Prog. Cell Cycle Res.,* **1996**, *2*, 137-145. http://dx.doi.org/10.1007/978-1-4615-5873-6_14 PMID: 9552391

[103] Schulze-Gahmen, U.; De Bondt, H.L.; Kim, S.H. High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J. Med. Chem.,* **1996**, *39*(23), 4540-4546. http://dx.doi.org/10.1021/jm960402a PMID: 8917641

[104] Schulze-Gahmen, U.; Brandsen, J.; Jones, H.D.; Morgan, D.O.; Meijer, L.; Vesely, J.; Kim, S.H. Multiple modes of ligand recognition: crystal structures of cyclin-dependent protein kinase 2 in complex with ATP and two inhibitors, olomoucine and isopentenyladenine. *Proteins,* **1995**, *22*(4), 378-391. http://dx.doi.org/10.1002/prot.340220408 PMID: 7479711

[105] Oudah, K.H.; Najm, M.A.A.; Samir, N.; Serya, R.A.T.; Abouzid, K.A.M. Design, synthesis and molecular docking of novel pyrazolo[1,5-a][1,3,5]triazine derivatives as CDK2 inhibitors. *Bioorg. Chem.,* **2019**, *92*, 103239. http://dx.doi.org/10.1016/j.bioorg.2019.103239 PMID: 31513938

[106] Ikwu, F.A.; Isyaku, Y.; Obadawo, B.S.; Lawal, H.A.; Ajibowu, S.A. *In silico* design and molecular docking study of CDK2 inhibitors with potent cytotoxic activity against HCT116 colorectal cancer cell line. *J. Genet. Eng. Biotechnol.,* **2020**, *18*(1), 51. http://dx.doi.org/10.1186/s43141-020-00066-2 PMID: 32930901

[107] Teng, M.; Jiang, J.; He, Z.; Kwiatkowski, N.P.; Donovan, K.A.; Mills, C.E.; Victor, C.; Hatcher, J.M.; Fischer, E.S.; Sorger, P.K.; Zhang, T.; Gray, N.S. Development of CDK2 and CDK5 dual degrader TMX-2172. *Angew. Chem. Int. Ed. Engl.,* **2020**, *59*(33), 13865-13870. http://dx.doi.org/10.1002/anie.202004087 PMID: 32415712

[108] Shawky, A.M.; Abourehab, M.A.S.; Abdalla, A.N.; Gouda, A.M. Optimization of pyrrolizine-based Schiff bases with 4-thiazolidinone motif: design, synthesis and investigation of cytotoxicity and anti-inflammatory potency. *Eur. J. Med. Chem.,* **2020**, *185*, 111780. http://dx.doi.org/10.1016/j.ejmech.2019.111780 PMID: 31655429

[109] Viegas, D.J.; Edwards, T.G.; Bloom, D.C.; Abreu, P.A. Virtual screening identified compounds that bind to cyclin dependent kinase 2 and prevent herpes simplex virus type 1 replication and reactivation in neurons. *Antiviral Res.,* **2019**, *172*, 104621. http://dx.doi.org/10.1016/j.antiviral.2019.104621 PMID: 31634495

[110] Zhu, J.; Wu, Y.; Xu, L.; Jin, J. Theoretical studies on the selectivity mechanisms of glycogen synthase kinase 3β (GSK3β) with pyrazine ATP-competitive inhibitors by 3DQSAR, molecular docking, molecular dynamics simulation and free energy calculations. *Curr. Computeraided Drug Des.,* **2020**, *16*(1), 17-30. http://dx.doi.org/10.2174/1573409915666190708102459 PMID: 31284868

[111] Fassio, A.V.; Santos, L.H.; Silveira, S.A.; Ferreira, R.S.; de Melo-Minardi, R.C. nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM Trans. Comput. Biol. Bioinformatics,* **2020**, *17*(4), 1317-1328. https://doi.org/10.1109/TCBB.2019.2892099 PMID: 30629512

[112] Zhang, X.; Shi, G.; Wu, X.; Zhao, Y. Gypensapogenin H from hydrolyzate of total *Gynostemma pentaphyllum* saponins induces apoptosis in human breast carcinoma cells. *Nat. Prod. Res.,* **2020**, *34*(11), 1642-1646. http://dx.doi.org/10.1080/14786419.2018.1525370 PMID: 30470142

[113] 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

[114] Cardamone, F.; Pizzi, S.; Iacovelli, F.; Falconi, M.; Desideri, A. Virtual screening for the development of dual-inhibitors targeting topoisomerase IB and tyrosyl-DNA

phosphodiesterase 1. *Curr. Drug Targets,* **2017**, *18*(5), 544-555.
http://dx.doi.org/10.2174/1389450116666150727114742 PMID: 26212266

[115] Biesiada, J.; Porollo, A.; Velayutham, P.; Kouril, M.; Meller, J. Survey of public domain software for docking simulations and virtual screening. *Hum. Genomics,* **2011**, *5*(5), 497-505.
http://dx.doi.org/10.1186/1479-7364-5-5-497 PMID: 21807604

[116] Bitencourt-Ferreira, G.; Rizzotto, C.; de Azevedo, W.F.Jr. 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

[117] 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

[118] Ballester, P.J. Machine Learning for Molecular Modelling in Drug Design. *Biomolecules,* **2019**, *9*(6), 216.
http://dx.doi.org/10.3390/biom9060216 PMID: 31167503

[119] 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

[120] Figueroa-Villar, J.D.; Petronilho, E.C.; Kuca, K.; Franca, T.C.C. Review about structure and evaluation of reactivators of acetylcholinesterase inhibited with neurotoxic organophosphorus compounds. *Curr. Med. Chem.,* **2021**, *28*(7), ~~1422-~~1442.
http://dx.doi.org/10.2174/0929867327666200425213215 PMID: 32334495

[121] Russo, S.; de Azevedo, W.F. Computational analysis of dipyrone metabolite 4-aminoantipyrine as a cannabinoid receptor 1 agonist. *Curr. Med. Chem.,* **2020**, *27*(28), 4741-4749.
http://dx.doi.org/10.2174/0929867326666190906155339 PMID: 31490743

[122] Scotti, M.T.; Monteiro, A.F.M.; de Oliveira Viana, J.; Mendonça, F.J.B.Jr.; Ishiki, H.M.; Tchouboun, E.N.; De Araújo, R.S.A.; Scotti, L. Recent theoretical studies concerning important tropical infections. *Curr. Med. Chem.,* **2020**, *27*(5), 795-834.
http://dx.doi.org/10.2174/0929867326666190711121418 PMID: 31296154

[123] Lungu, C.N.; Bratanovici, B.I.; Grigore, M.M.; Antoci, V.; Mangalagiu, I.I. Hybrid imidazole-pyridine derivatives: an approach to novel anticancer DNA intercalators. *Curr. Med. Chem.,* **2020**, *27*(1), 154-169.
http://dx.doi.org/10.2174/0929867326666181220094229 PMID: 30569842

[124] Halder, A.K.; Dias Soeiro Cordeiro, M.N. Advanced *in silico* methods for the development of anti- leishmaniasis and anti-trypanosomiasis agents. *Curr. Med. Chem.,* **2020**, *27*(5), 697-718.
http://dx.doi.org/10.2174/0929867325666181031093702 PMID: 30378482

[125] Zhu, Y.; Liang, M.; Li, H.; Ni, H.; Li, L.; Li, Q.; Jiang, Z. A mutant of *Pseudoalteromonas carrageenovora* arylsulfatase with enhanced enzyme activity and its potential application in improvement of the agar quality. *Food Chem.,* **2020**, *320*, 126652.
http://dx.doi.org/10.1016/j.foodchem.2020.126652 PMID: 32229399

[126] Taguchi, A.T.; Boyd, J.; Diehnelt, C.W.; Legutki, J.B.; Zhao, Z.G.; Woodbury, N.W. Comprehensive prediction of molecular recognition in a combinatorial chemical space using machine learning. *ACS Comb. Sci.,* **2020**, *22*(10), 500-508.
http://dx.doi.org/10.1021/acscombsci.0c00003 PMID: 32786325

[127] Jehangir, I.; Ahmad, S.F.; Jehangir, M.; Jamal, A.; Khan, M. Integration of bioinformatics and *in vitro* analysis reveal anti-leishmanial effects of azithromycin and nystatin. *Curr. Bioinform.,* **2019**, *14*(5), 450-459.
http://dx.doi.org/10.2174/1574893614666181217142344

[128] Lushington, G.H. Chemistry, Screening, and the democracy of publishing. *Comb. Chem. High Throughput Screen.,* **2019**, *22*(5), 288-289.
http://dx.doi.org/10.2174/1386207322999190715161959 PMID: 31446889

[129] Zhao, J.; Cao, Y.; Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.,* **2020**, *18*, 417-426.
http://dx.doi.org/10.1016/j.csbj.2020.02.008 PMID: 32140203

[130] Zhang, W.; Li, W.; Zhang, J.; Wang, N. Data integration of hybrid microarray and single cell expression data to enhance gene network inference. *Curr. Bioinform.,* **2019**, *14*(3), 255-268.
http://dx.doi.org/10.2174/1574893614666190104142228

[131] Wu, Y.; Guo, Y.; Xiao, Y.; Lao, S. AAE-SC: a scRNA-Seq clustering framework based on adversarial autoencoder. *IEEE Access,* **2020**, *8*, 178962-178975.
http://dx.doi.org/10.1109/ACCESS.2020.3027481

[132] Li, M.; Zhang, S.; Yang, B. Urea transporters identified as novel diuretic drug targets. *Curr. Drug Targets,* **2020**, *21*(3), 279-287.
http://dx.doi.org/10.2174/1389450120666191129101915 PMID: 31782365

[133] Safarizadeh, H.; Garkani-Nejad, Z. Investigation of MI-2 analogues as MALT1 inhibitors to treat of diffuse large B-cell lymphoma through combined molecular dynamics simulation, molecular docking and QSAR techniques and design of new inhibitors. *J. Mol. Struct.,* **2019**, *1180*, 708-722.
http://dx.doi.org/10.1016/j.molstruc.2018.12.022

[134] 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

[135] Sun, B.; Wang, W.; He, Z.; Zhang, M.; Kong, F.; Sain, M. Biopolymer substrates in buccal drug delivery: current status and future trend. *Curr. Med. Chem.,* **2020**, *27*(10), 1661-1669.
http://dx.doi.org/10.2174/0929867325666181001114750 PMID: 30277141

[136] Aleksandrov, A.; Myllykallio, H. Advances and challenges in drug design against tuberculosis: application of *in silico* approaches. *Expert Opin. Drug Discov.,* **2019**, *14*(1), 35-46.
http://dx.doi.org/10.1080/17460441.2019.1550482 PMID: 30477360

[137] Cavada, B.S.; Osterne, V.J.S.; Lossio, C.F.; Pinto-Junior, V.R.; Oliveira, M.V.; Silva, M.T.L.; Leal, R.B.; Nascimen-

to, K.S. One century of ConA and 40 years of ConBr research: a structural review. *Int. J. Biol. Macromol.,* **2019**, *134*, 901-911. http://dx.doi.org/10.1016/j.ijbiomac.2019.05.100 PMID: 31108148

[138] Jiang, M.; Li, Z.; Bian, Y.; Wei, Z. A novel protein descriptor for the prediction of drug binding sites. *BMC Bioinformatics,* **2019**, *20*(1), 478. http://dx.doi.org/10.1186/s12859-019-3058-0 PMID: 31533611

[139] Cavada, B.S.; Araripe, D.A.; Silva, I.B.; Pinto-Junior, V.R.; Osterne, V.J.S.; Neco, A.H.B.; Laranjeira, E.P.P.; Lossio, C.F.; Correia, J.L.A.; Pires, A.F.; Assreuy, A.M.S.; Nascimento, K.S. Structural studies and nociceptive activity of a native lectin from *Platypodium elegans* seeds (nPELa). *Int. J. Biol. Macromol.,* **2018**, *107*(Pt A), 236-246. https://doi.org/10.1016/j.ijbiomac.2017.08.174 PMID: 28867234

[140] Abbasi, W.A.; Asif, A.; Ben-Hur, A.; Minhas, F.U.A.A. Learning protein binding affinity using privileged information. *BMC Bioinformatics,* **2018**, *19*(1), 425. http://dx.doi.org/10.1186/s12859-018-2448-z PMID: 30442086

[141] Ribeiro, F.F.; Mendonca Junior, F.J.B.; Ghasemi, J.B.; Ishiki, H.M.; Scotti, M.T.; Scotti, L. Docking of natural products against neurodegenerative diseases: general concepts. *Comb. Chem. High Throughput Screen.,* **2018**, *21*(3), 152-160. http://dx.doi.org/10.2174/1386207321666180313130314 PMID: 29532756

[142] Lemos, A.; Melo, R.; Preto, A.J.; Almeida, J.G.; Moreira, I.S.; Dias Soeiro Cordeiro, M.N.D.S *In silico* studies targeting G-protein coupled receptors for drug research against Parkinson's disease. *Curr. Neuropharmacol.,* **2018**, *16*(6), 786-848. http://dx.doi.org/10.2174/1570159X16666180308161642 PMID: 29521236

[143] Leal, R.B.; Pinto-Junior, V.R.; Osterne, V.J.S.; Wolin, I.A.V.; Nascimento, A.P.M.; Neco, A.H.B.; Araripe, D.A.; Welter, P.G.; Neto, C.C.; Correia, J.L.A.; Rocha, C.R.C.; Nascimento, K.S.; Cavada, B.S. Crystal structure of DlyL, a mannose-specific lectin from *Dioclea lasiophylla* Mart. Ex Benth seeds that display cytotoxic effects against C6 glioma cells. *Int. J. Biol. Macromol.,* **2018**, *114*, 64-76. http://dx.doi.org/10.1016/j.ijbiomac.2018.03.080 PMID: 29559315

[144] de Ávila, M.B.; Bitencourt-Ferreira, G.; de Azevedo, W.F.Jr. Structural basis for inhibition of enoyl-[Acyl carrier protein] reductase (InhA) from *Mycobacterium tuberculosis. Curr. Med. Chem.,* **2020**, *27*(5), 745-759. http://dx.doi.org/10.2174/0929867326666181203125229 PMID: 30501592

[145] Freitas, P.G.; Elias, T.C.; Pinto, I.A.; Costa, L.T.; de Carvalho, P.V.S.D.; Omote, D.Q.; Camps, I.; Ishikawa, T.; Arcuri, H.A.; Vinga, S.; Oliveira, A.L.; Junior, W.F.A.; da Silveira, N.J.F. Computational approach to the discovery of phytochemical molecules with therapeutic potential targets to the PKCZ protein. *Lett. Drug Des. Discov.,* **2018**, *15*(5), 488-499. http://dx.doi.org/10.2174/1570180814666170810120150

[146] Russo, S.; de Azevedo, W.F. Advances in the understanding of the cannabinoid receptor 1 - focusing on the inverse agonists interactions. *Curr. Med. Chem.,* **2019**, *26*(10), 1908-1919. http://dx.doi.org/10.2174/0929867325666180417165247 PMID: 29667549

[147] Wolin, I.A.V.; Heinrich, I.A.; Nascimento, A.P.M.; Welter, P.G.; Sosa, L.D.V.; De Paul, A.L.; Zanotto-Filho, A.; Nedel, C.B.; Lima, L.D.; Osterne, V.J.S.; Pinto-Junior, V.R.; Nascimento, K.S.; Cavada, B.S.; Leal, R.B. ConBr lectin modulates MAPKs and Akt pathways and triggers autophagic glioma cell death by a mechanism dependent upon caspase-8 activation. *Biochimie,* **2021**, *180*, 186-204. http://dx.doi.org/10.1016/j.biochi.2020.11.003 PMID: 33171216

[148] 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

[149] Pinto-Junior, V.R.; Osterne, V.J.; Santiago, M.Q.; Correia, J.L.; Pereira-Junior, F.N.; Leal, R.B.; Pereira, M.G.; Chicas, L.S.; Nagano, C.S.; Rocha, B.A.; Silva-Filho, J.C.; Ferreira, W.P.; Rocha, C.R.; Nascimento, K.S.; Assreuy, A.M.; Cavada, B.S. Structural studies of a vasorelaxant lectin from *Dioclea reflexa* hook seeds: crystal structure, molecular docking and dynamics. *Int. J. Biol. Macromol.,* **2017**, *98*, 12-23. http://dx.doi.org/10.1016/j.ijbiomac.2017.01.092 PMID: 28130130

[150] 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

[151] Amaral, M.E.A.; Nery, L.R.; Leite, C.E.; de Azevedo, W.F.Jr.; Campos, M.M. Pre-clinical effects of metformin and aspirin on the cell lines of different breast cancer subtypes. *Invest. New Drugs,* **2018**, *36*(5), 782-796. http://dx.doi.org/10.1007/s10637-018-0568-y PMID: 29392539

[152] Borisa, A.; Bhatt, H. 3D-QSAR (CoMFA, CoMFA-RG, CoMSIA) and molecular docking study of thienopyrimidine and thienopyridine derivatives to explore structural requirements for aurora-B kinase inhibition. *Eur. J. Pharm. Sci.,* **2015**, *79*, 1-12. http://dx.doi.org/10.1016/j.ejps.2015.08.017 PMID: 26343315

[153] Gramatica, P. On the development and validation of QSAR models. *Methods Mol. Biol.,* **2013**, *930*, 499-526. http://dx.doi.org/10.1007/978-1-62703-059-5_21 PMID: 23086855

[154] Triggle, D.J. The chemist as astronaut: searching for biologically useful space in the chemical universe. *Biochem. Pharmacol.,* **2009**, *78*(3), 217-223. http://dx.doi.org/10.1016/j.bcp.2009.02.015 PMID: 19481639

[155] Kell, D.B.; Samanta, S.; Swainston, N. Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently. *Biochem. J.,* **2020**, *477*(23), 4559-4580. http://dx.doi.org/10.1042/BCJ20200781 PMID: 33290527

[156] Johnson, E.O.; Hung, D.T. A point of inflection and reflection on systems chemical biology. *ACS Chem. Biol.,* **2019**, *14*(12), 2497-2511. http://dx.doi.org/10.1021/acschembio.9b00714 PMID: 31613592

[157] Fotis, C.; Antoranz, A.; Hatziavramidis, D.; Sakellaropoulos, T.; Alexopoulos, L.G. Network-based technologies for

early drug discovery. *Drug Discov. Today,* **2018**, *23*(3), 626-635.
http://dx.doi.org/10.1016/j.drudis.2017.12.001 PMID: 29294361

[158] Kirkpatrick, P.; Ellis, C. Chemical space. *Nature,* **2004**, *432*(7019), 823.
http://dx.doi.org/10.1038/432823a

[159] 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

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

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

[162] 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

[163] Hou, J.; Jun, S.R.; Zhang, C.; Kim, S.H. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*(10), 3651-3656.
http://dx.doi.org/10.1073/pnas.0409772102 PMID: 15705717

[164] Singh, A.V.; Chandrasekar, V.; Janapareddy, P.; Mathews, D.E.; Laux, P.; Luch, A.; Yang, Y.; Garcia-Canibano, B.; Balakrishnan, S.; Abinahed, J.; Al Ansari, A.; Dakua, S.P. Emerging application of nanorobotics and artificial intelligence to cross the BBB: advances in design, controlled maneuvering, and targeting of the barriers. *ACS Chem. Neurosci.,* **2021**, *12*(11), 1835-1853.
http://dx.doi.org/10.1021/acschemneuro.1c00087 PMID: 34008957

[165] Singh, A.V.; Jahnke, T.; Wang, S.; Xiao, Y.; Alapan, Y.; Kharratian, S.; Onbasli, M.C.; Kozielski, K.; David, H.; Richter, G.; Bill, J.; Laux, P.; Luch, A.; Sitti, M. Anisotropic gold nanostructures: optimization *via in silico* modeling for hyperthermia. *ACS Appl. Nano Mater.,* **2018**, *1*(11), 6205-6216.
http://dx.doi.org/10.1021/acsanm.8b01406