Trends in Developmental Biology
Vol. 13, 2020

Original Article

Milk-Way algorithm for ligand-based virtual screening: CDK2 case study

Carmelina Figueiredo Vieira Leite[1,*], Lucianna Helene Silva Santos[2], Larissa Fernandes Leijôto[1], Diego César Batista Mariano[1], Rafael Eduardo Oliveira Rocha[2] and Marcos Augusto dos Santos[3]

[1]Laboratory of Bioinformatics and Systems; [2]Department of Biochemistry and Immunology;
[3]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte,
31270-901, Brazil.

ABSTRACT

Ligand-based screening of large molecular databases can help reduce costs with experiments by filtering and ranking promising compounds in an initial stage of the drug developing process. However, some ligand-based methods can be ineffective when presented with a high-dimensional number of attributes extracted from an extensive dataset of compounds. Herein, we propose a drug-mining algorithm that can be used to screen ligands and repurpose known drugs, from any dataset for any target. The Milk-Way algorithm combines mathematical and regression methods to select promising compounds from a high-dimensional dataset without the use of massive computational power. We carried out a prospective screening targeting cyclin-dependent kinase two (CDK2), an attractive target for therapeutics designed to arrest or recover control of the cell cycle. The combined use of the algorithm metrics and molecular docking suggested five promising drugs to be repositioned (Pramocaine, Prochlorperazine, Trifluoperazine, Methionine, and Pergolide), in which three were already mentioned as possible inhibitors of related diseases in the literature.

KEYWORDS: algorithm, drug discovery, drug repurposing, ligand-based virtual screening, logistic regression, machine learning, development.

1. INTRODUCTION

We present a new algorithm to screen novel compounds using CDK2 as the target. This is an enzyme that phosphorylates many proteins involved in cell cycle progression, DNA replication, histone synthesis, centrosome duplication, among other processes [1, 2]. Because of these functions, CDK2 represents an attractive target for therapeutics designed to arrest or recover control of the cell cycle in dividing cells [3], and since the enzyme is not essential for the cell cycle, its toxicity is not severe [4]. Despite the importance of the CDK2 protein, not many commercial drugs act against it. Thus, we investigated the use of drug repurposing as an aid to CDK2 drug development. Drug repurposing is the strategy of discovering new uses or conditions for approved drugs to not only assess the effects of the drug on a new target but also to reduce the cost of developing a new drug.

Computational approaches, such as virtual screening (VS), have emerged as alternatives to screen large libraries of small molecules in a cost-efficient manner. Although VS approaches do not substitute experimental assays, they can speed up and rationalize the process of drug discovery, enriching the number of hits since it can downsize the number of candidates to be tested [5, 6]. In structure-based virtual screening (SBVS), the three-dimensional structure of the target is known, from x-ray crystallographic, NMR, or computational

*Corresponding author: cleite@ufmg.br

modeling [7, 8]. Recently, cryo-EM (cryo-electron microscopy) was introduced as a powerful three-dimensional source and is starting to make an impact in drug discovery [9]. SBVS usually involves the molecular docking methodology, which places and ranks the compounds in the binding site according to an algorithm that predicts their possible binding affinity [10].

In the absence of three-dimensional structures of the targets, the molecular and chemical properties of known actives and tested compounds are gathered to create models of their binding using ligand-based virtual screening (LBVS) [11]. Some LBVS methodologies can assume that one or more actives share a binding mode. Thus, the screening will be done as a similarity or matching search to select potential new binders with similar chemical features to the known ones [12]. Chemical compounds to compose screening libraries are available as a free resource of bioactivity data for small molecules in various databases such as ChEBI [13], ZINC [14], PubChem [15], DrugBank [16], IUPHAR-DB [17], and KEGG [18]. Alternatively, in-house compound datasets can also be created from previously tested compounds and analogs. A powerful LBVS method is the Quantitative structure-activity relationship (QSAR). QSAR models starts by calculating chemical descriptors collected from compounds found in databases or the literature. These descriptors are correlated with biological properties using a variety of machine learning techniques [5, 19]. After created and validated, QSAR models are applied to predict novel compounds in virtual screening campaigns. Although QSAR is effective and with the use of SBVS methods also target orientated, it is still a time and computationally demanding method [5].

For many LBVS methods, in the model's generation step, known actives can be employed as training data in classification methods. These methods use this information to separate a database of compounds with unknown activity into predicted actives and inactives [20]. Classification methods are usually machine learning approaches that build models, such as decision trees [21, 22], neural networks [23, 24], and support vector machines [25], and can perform particularly well in enriching actives [20]. Established algorithms for data mining (Naive Bayes, SMO, Random Forrest, J48) are also used to classify chemical compounds [26].

However, classification methods can demand extensive knowledge over the many methodologies and need high computing power. Furthermore, these methods might not perform well when subjected to imbalanced or high-dimensional data, *i.e.*, when all features describing the chemical properties of the compounds are used. These problems can lead to an inadequate exploration of the ligands, and lack of accurate results, essential to screening [27-29]. Therefore, the proposition of an improved *in silico* approach to classified chemical compounds is a relevant issue.

In this paper, we suggest the Milk-Way algorithm (a WAY of Mathematical Interpretation of the Logistic ranK), a robust combination of well-known techniques of data-mining, logistic regression, vector space representation, and linear algebra to contribute to the ligand-based approaches to rational drug design. Our algorithm does not demand a complex computational infrastructure to select potential hits in a screening campaign. A first step in the algorithm is to collect a library of actives and inactives compound structures to be designated through descriptors (Fragment Pair/ Pharmacophore). Then, a model is created, validated, and trained to classify potential hits. The model enables us to calculate and project the probability of each ligand (P(x), where x is the ligand) outcome, which is used to distinguish possible high performing ligands. An application of repurposing commercially available drugs with the Milk-Way algorithm was conducted using the cyclin-dependent kinase 2 (CDK2) as a target to exemplify the algorithm's predictive power. CDK is a family of serine/threonine protein kinases which act as critical regulatory element in cell cycle progression and development. As the name reveals, those are enzymes that catalyze the transfer of a phosphate from ATP to a protein substrate, more precisely on a serine or threonine amino acid residue [30].

2. MATERIALS AND METHODS

All the data were processed in MATLAB R2017a, using a laptop with 4 GB RAM, 320 GB hard drive, and a processor Intel Core i5 2.53 GHz.

The Milk-Way methodology runs in the Windows operating systems.

2.1. Data collection

The starting point, in our algorithm, is the construction of a matrix, whose entities are the ligands (columns) and their attributes (lines). All the attributes were generated through the PowerMV program [31]. The attributes are binary descriptors that define both active and inactive ligands. We have chosen fragment pair (735 features) and pharmacophore fingerprints (147 features) as molecular features [32-34]. For fragment-based descriptors, PowerMV replaces atom types with groups of atoms and counts the shortest path between them. For example, two phenyl rings, which are separated by two bonds, are expressed as AR_02_AR. In total 14 groups of atoms are considered. Whereas pharmacophore fingerprints are built based on bioisosteric principles [35], *i.e.*, two atoms (or groups), predicted to have similar biological effect, are classified as the same type. For example, the disulfide (-S-) is often used to replace ester group (-O-); hence we assign these two groups to the same type. Therefore, only six classes are considered in the pharmacophore-based descriptors. However, it is possible to use any database and attributes to build the input matrix, such as molecular, physicochemical, topological, structural, pharmacological descriptors, or any property of the ligands. A binary representation is also not obligatory. The only imperative premise is a representation of ligands. To the algorithm, this is the most critical step since the projection of the ligand into the vector space depends on that, to calculate the probability.

Literature data of known inhibitors were extracted from two databases, PubChem [36] and DrugBank [16], depending on the case study. In the CDK2 study case, we used in the training set the only approved drug to CDK2 to compose the actives and 152 compounds as decoys. The test set included all 2389 commercial drugs according to the DrugBank [16], at the time the search was performed.

2.2. The screening algorithm

The algorithm consists of several consecutive steps: (i) SVD; (ii) Ad-hoc choice; (iii) Modified Logistic Regression and, (iv) Stratified feature selection through the alpha values.

2.2.1. Singular value decomposition

This step in our screening algorithm helps reducing the noise and retrieve latent patterns of the input matrix. A technique for information retrieval, using a linear algebra approach, is the singular value decomposition (SVD). This rank reduction procedure is closely related to matrix factorization, data compression, dimension reduction, and feature selection/extraction [37]. When SVD is utilized, it allows the matrix to be represented as a set of derived matrices, which can have different depictions of data without loss in their semantic meaning [37, 38]. A matrix submitted to the SVD method can be represented as:

$$A = U\Sigma V^T, \tag{1}$$

where A is a matrix of real numbers or complex numbers composed of m rows by n columns. However, now, m represents the attributes and n, the entities. The U is an orthonormal m x m matrix and the eigenvectors of AA^T; the Σ is a m x n matrix, known as the diagonal matrix, with real and non-negative numbers and contain the singular values of A. The matrix V^T is known as a conjugate transpose, a n x n unit matrix. As the diagonal values of Σ are ordered in descending order, Σ is a direct function of matrix A and distinguishes the singular values of this matrix. This sorting is from the most meaningful to the least significant values. Whereas from a subset of singular values of size $k < m$, we can obtain A_k, the approximate matrix of A, with k-dimensional:

$$A_k = U_k\Sigma_k V_k^T \tag{2}$$

The approximation will be related to how many singular eigenvalues are used [37, 39, 40]. This strategy enables an information extraction based on less data, and the data analysis execution time does not increase exponentially when the matrix size is increased. A data set represented by a smaller number of singular values than the original full-size dataset tends to cluster data that would not be clustered together if the original one was used. Therefore, the derived representation, which captures associations, is used for retrieval [38-41]. The representation in the reduced space depiction is economical, in the sense that N original index features have been replaced by the $k < M$ best-approximated surrogates. It is essential for the method that the derived k-dimensional factor space does not reconstruct the original term space correctly. The beauty of SVD, however, is that it allows a simple strategy for optimal approximate fit using smaller matrices. Everitt and Dunn [42] proposed an alternative approach where singular values whose relative variance is less than 0.7/*n*, where *n* is the number of individuals in the matrix, must be ignored. If the singular values in *Σ*, are ordered by size, the first and largest *k* may be kept and the remaining smaller ones set to zero [43].

2.2.2. Ad-hoc choice

After determining the number of singular values of our input matrix of molecular features of active and inactive compounds, we define the number of singular values as a criterion of the number of individuals to collect. Otherwise, we would have an infinitive of possibilities as well as combinations. The model was constructed for each query (ligand to predict the P(*x*)) using the closest active and inactive ligands. The main objective of this particular strategy is to create a homogeneous and specific system through the choice of closely spaced individuals. We chose Hamming distance, on account of better adjusting the matrix that is composed of zeros and ones. This distance can be defined as the number of positions in which a codeword differs between two code words [44], or, in other words, it is the minimum number of *errors* that could have transformed one string into the other.

2.2.3. Modified logistic regression

The regression method is helpful to any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The logistic regression analysis proposes the classification of individuals in different categories, with an accurate estimation for that possibility [45]. The logistic regression equation consists of assigning values to *α* = (α_0, α_1,…,α_n) to fit the logit function (3)

$$P(x) = \frac{e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}}{1 + e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}} \tag{3}$$

such that *P*(*x*) = 0/1 associated with the activity of each ligand (inactive or active, respectively). The data is represented by matrix *A*, with *m* rows and *n* columns; so the value of each position $a_{m,n}$ represents the attribute *n* of the ligand *m*. Here, the matrix is the transpose of the previous SVD-treated matrix. We will omit the indication of row *m* in the elements of vector *x,* that is *x* = (x_1, x_2,…,x_n). Associated with each row *m*, there is $P_i(x)$ = 0/1 that informs the activity of the ligand. We observed that when $e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}$ drops to zero, P_i (*x*) also goes to zero. On the other hand, if $e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}$ tends to infinity, P_i (*x*) approximates one. Viewing P_i (*x*) as the probability, the odds C_i (*x*) is given by:

$$C_i(x) = \frac{P(x)}{1-P(x)} = e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}} \tag{4}$$

To implement the method, we use $\hat{C}i(x) \approx Ci(x)$ = (0.99999 / (1-0.99999)) instead of C_i (*x*) when the odds are related to $P_i(x)$ = 1. When $P_i(x)$ = 0, we consider $\hat{C}_i(x) \approx C_i(x)$ = (0.00001 / (1-0.00001)).

Taking the logarithm on both sides of (equation 4) we have:

$$\text{In}\left[\frac{P(x)}{1-P(x)}\right] = \text{In}\left[e^{\sum_i \alpha_1 x_1 + \alpha_{n+1}}\right] = \sum_i \alpha_1 x_1 + \alpha_{n+1} \tag{5}$$

The system (equation 5) is a linear algebraic model created to determine *α*:

$$bi = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \tag{6}$$

where *i* = 1, 2, …, *m*.

Let $\bar{e} = (1, \ldots, 1)^T$ be a vector of *m* ones and $b = (b_1, b_2, \ldots, b_m)^T$. The system of linear equations (equation 6) may be represented by:

$$B\alpha = b, \text{ with } B = [\bar{e}\, A] \tag{7}$$

The solution of equation 7 is given by the solution of a least square problem but, in our case (equation 7), it has an infinite number of solutions, since *n* + 1 >> *m*. It is usual to circumvent this difficulty by suppressing the model and keeping only a small subset of the *n* attributes. This procedure resembles the feature selection in data mining, an open problem research area [35].

We approximate a solution of the combinatorial problem related to the feature selection solving the following:

$$\alpha = \textit{argument that minimize } f(\alpha) \quad (8)$$

$$\textit{where } f(\alpha) = \alpha^T \alpha + (B\alpha - b)^T * (B\alpha - b)$$

The solution of equation (8) is a convex unconstrained optimization problem [46]; after applying the optimality conditions, it can be shown that the optimal solution α* is such that it verifies the following system of linear equations

$$(I + B^T B)\,\alpha = B^T b, \quad (9)$$

where I is an identity matrix of dimension n. We point out that in Golub [47], an identity matrix to solve the rank deficiency in systems of linear equations was used. One should note that the identity matrix in equation 9 does not allow the rank to become deficient.

It was observed that the optimal solution α^* of (equation 8) is unique. So, given a query $q = (q_1, q_2, \ldots, q_n)$, the probability of q be an active ligand, is given by

$$P(q) = g(q) / (1 + g(q)), \quad (10)$$

where, $g(q) = \exp([1\ q]\,\alpha)$.

On this step, the regression gives us a probability associated with each ligand. We would like to highlight the fact that modification in the logit function allows us to use more features than ligands.

2.3. Molecular docking

A docking protocol with DOCK6 [48] was performed to establish the use of the generated docking score as the binding energy. DOCK6 provides multiple scoring functions, from force-field-based to pharmacophore-based, and the possibility of combining them. Therefore, in our case, it was useful to incorporate chemical features of known inhibitors and the binding sites to the docking of the selected compounds. Bosutinib was also subject to docking simulations since no crystallographic structure of the complex CDK2-Boustinib is available. The docking poses of the selected drugs were analyzed with the help of the Discovery Studio Visualizer [49] that showed possible interactions with CDK2 active site.

3. RESULTS

3.1. Data collection, model building, and testing of the Milk-Way algorithm

Milk-Way method is divided into five steps: (A) selection of active and inactive ligands; (B) fingerprint construction, where ligand properties are collected and stored in a matrix; (C) noise reduction using singular value decomposition (SVD); (D) model construction based on ad hoc selection; and (E) prediction using a modified logistic regression, which selects ligands based on high values of P(x) (Figure 1).

The Milk-Way algorithm requires an initial input matrix of individuals to be capable of building a classification model. In this matrix, two sets of compounds can be found — compounds with experimentally tested activity towards the desired target and compounds without any evident activity. The latter set can be formed by confirmed inactives or artificially created compounds, the so-called decoys.

Interestingly, when the weights of the attributes (α_i values in equation 3) were analyzed, we could understand the impact that each feature had on the classification into active or inactive. Since we have a significant number of features, we expected that the highest α values corresponded to active compounds. Moreover, the inverse is also true – the lowest α values correspond to inactive compounds.

3.2. Application of the Milk-Way algorithm

For the construction of the model, we selected the only commercial drug that acts against CDK2 (Bosutinib [15]) as the active entity. Decoys were generated from DUD-E [50] based on Bosutinib and two commercialized drugs for CDK4 and CDK6 [16], Ribociclib, and Palbociclib. CDK4 and CDK6 are homologous proteins to CDK2 [1] (Supplementary material Table S1). We put them with the inactives to distinguish the effect of the homology between the enzymes. The same molecular features (Fragment Pair/Pharmacophore fingerprints) previously described were also used as attributes (Supplementary material Table S2). The screening was performed using commercialized drugs retrieved from the DrugBank [15]. The drugs which obtain a probability (P(x)) of 0.98 or

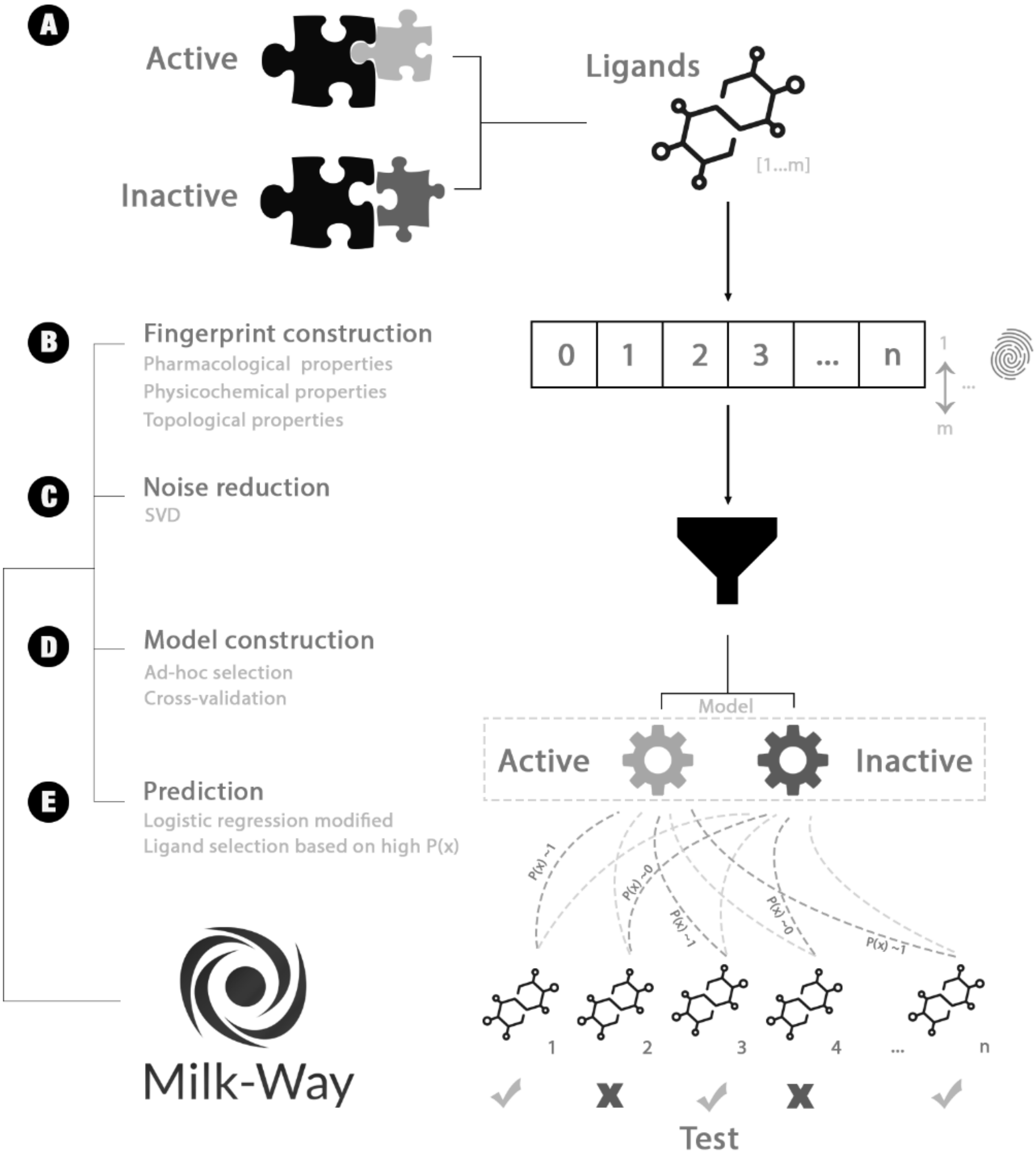


Figure 1. Workflow of the Milk-Way algorithm. It is divided into five steps: **(A)** selection of active and inactive ligands; **(B)** fingerprint construction; **(C)** noise reduction; **(D)** model construction; and **(E)** prediction. The input of the algorithm is a matrix of ligands to a specific target describe through descriptors. After the Singular Value Decomposition, the training model is ready either to calculate the probability of new binders or for repositioning marketed drugs by ad hoc selection. The output is the suggestion of new compounds to be used as ligands to the target.

higher were selected. In total, five drugs were selected: pramocaine, prochlorperazine, trifluoperazine, methionine and pergolide (Supplementary material Table S3).

We analyzed these drugs through molecular docking to probe a possible complementarity between the CDK2 structure and them. Molecular docking provided a rank through the chosen scoring function, prioritizing three compounds with predicted high affinity besides bosutinib (-26.17 kcal/mol): pramocaine (-30.83 kcal/mol), trifluoperazine (-23.67 kcal/mol), and prochlorperazine (-17.87 kcal/mol) (Supplementary material Table S4). Interestingly, pramocaine [51], prochlorperazine [52], and trifluoperazine [52, 53] are described in the literature to act on CDK2-related diseases such as Glioblastoma, breast cancer, and other tumor effects (Supplementary material Table S3).

The binding modes from docking showed that all compounds fitted very well in the CDK2 active site (Supplementary material Table S5) and interacted with key residues in the active site (Supplementary material Table S6). For instance, bosutinib binding mode was complementary to the active site (Figure 2a), and it achieved interactions with the same residues as the inhibitor present in the crystal structure, such as hydrogen bond interaction with LYS 33, hydrophobic interactions with VAL 18 and GLN 120 (Figure 2b). Similar behavior was observed with molecular docking highest-ranked compound, pramocaine (Figure 3a). Although pramocaine showed less interaction than bosutinib in the CDK2 binding site (Figure 3b), the molecular docking binding mode of this compound presented the same interacted residues as the crystal inhibitor (ILE 10, VAL 18, LYS 33, LEU 72, GLN 120, and LEU 123). The presence of these interactions from known inhibitors might indicate a possible binding of the predicted compounds.

4. DISCUSSION

Machine learning approaches are robust methodologies capable of screening drug leads from a dataset of many compounds with reasonable accuracy in a faster and cheaper manner than experimental testing. Most methods act as a classifier, separating the compounds into actives and inactives. To accomplish this classification, a model is created and validated through training sets using known actives to a specific target [12]. Milk-Way provides a novel and alternative approach to machine learning using a combination of data-mining techniques.

An interesting characteristic of Milk-Way is the usefulness of alpha values (equation 3) to perform the stratified feature selection. The positive values of the components α_i of α contribute to approximate $P(\alpha)$ to 1, and the negative values approximate $P(\alpha)$ to 0. Absolute values of α_i close to zero don't have significant impact in the

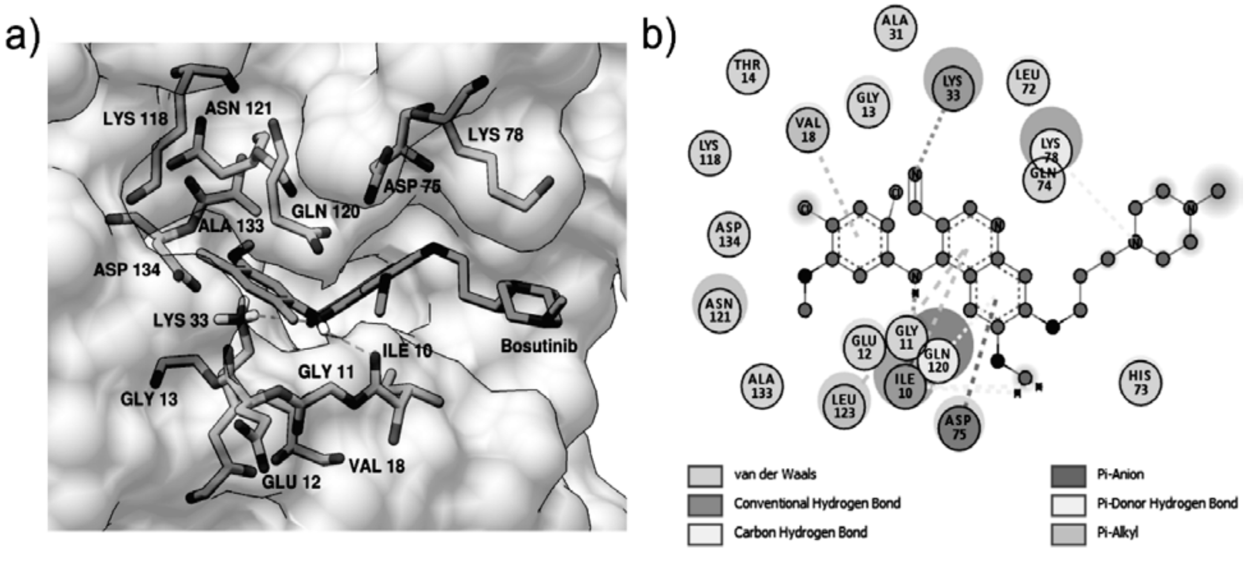


Figure 2. Bosutinib molecular docking binding mode. **a)** Bosutinib in the active site of CDK2 with the interacting residues and hydrogen bond interactions. **b)** 2D representation of the interactions between Bosutinib and the active site of CDK2.

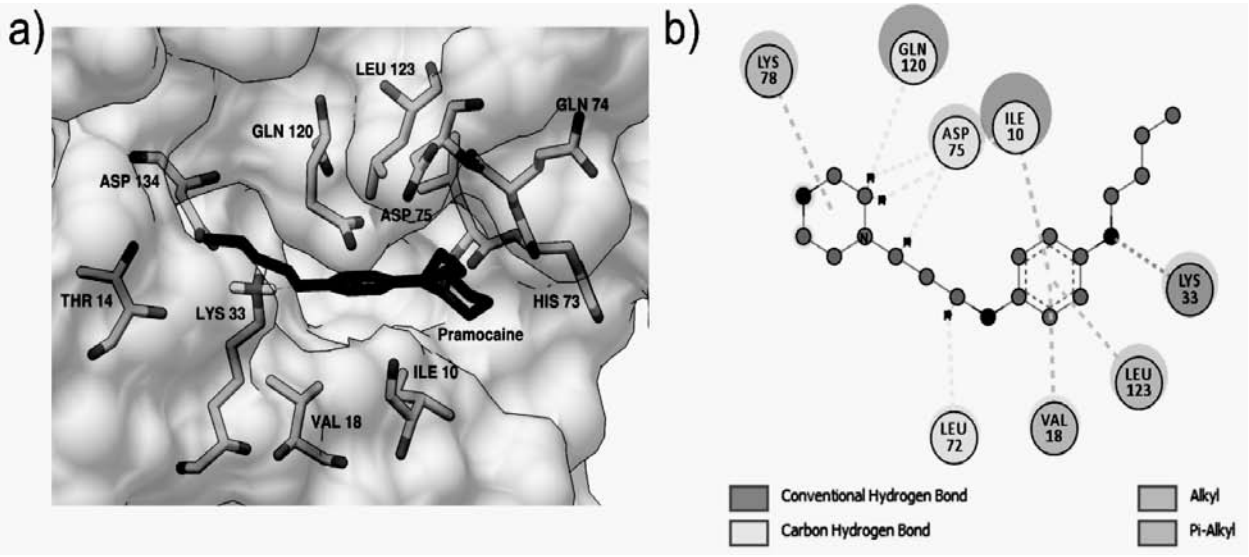


Figure 3. Pramocaine molecular docking binding mode. **a)** Pramocaine in the active site of CDK2 with the interacting residues and hydrogen bond interactions. **b)** 2D representation of the interactions between Pramocaine and the active site of CDK2.

computation of P(α). The modified logistic regression can identify the highest alpha values, which carried critical molecular features present on active compounds (Supplementary material Table S2). Low alpha values were usually observed in inactive compounds. Therefore, one advantage of the Milk-Way algorithm is that more features than compounds can be processed since the rank of the matrix is the one used in the calculations, not the entire matrix.

We performed a simulation study that showed the combination of our method to other in silico techniques. Since our algorithm is ligand-based, the ligands selected can benefit from molecular docking, a structure-based approach, when a structure of the target is known. The docking scores (Supplementary material Table S4) and interaction analysis can provide a better atomistic understanding of a possible inhibition of the CDK2 enzyme by the selected compounds (Supplementary material Tables S5 and S6). For instance, the highest scored compound pramocaine displayed all the interactions presented in another known CDK2 inhibitor (ILE 10, VAL 18, LYS 33, LEU 72, GLN 120, and LEU 123), indicating a possible strong binding of this compound. Furthermore, three of the five selected ligands (Pramocaine, Prochlorperazine, and Trifluoperazine) by the Milk-Way to repurpose, have already been described in the literature to act on CDK2-related diseases (Supplementary material Table S3) [51-53].

5. CONCLUSIONS

The development of new drugs takes about fourteen years, and the cost ranges are estimated at around 1.0 billion USD [54]. With the HTS method, it is possible to screen millions of compounds against a chosen target experimentally. However, it is an expensive process, and therefore lower-cost alternatives capable of sorting promising compounds from several other ones are sought out. These help to decrease the number of ligands to be tested in a possible experimental phase.

Our holistic approach can classify ligands with the support of the selected case studies. The use of literature data and datasets were appropriate for testing the algorithm and measuring the results. It is essential to notice, the cases investigated throughout the paper are unrelated to each other, demonstrating a practical way to prove the efficiency of the proposed algorithm for LBVS. Nevertheless, it is essential to highlight the acceptance of a higher number of attributes (ligands' features) than entities (ligands), without a problem of rank deficiency. This added factor is opposed to the classical logistic regression in which it is obligatorily to have more entities than attributes.

The proposed mathematical modulation does not require a massive infra-structure apparatus to be performed and constitutes a good strategy for the selection of promising compounds. The Milk-Way algorithm demonstrated excellent performance, and with less computational infrastructure. For a more in-depth and broader study of this algorithm, we are already applying it to other targets and other data-sets. Since there is an attempt to continually improve the efficiency of computational processes for the development of new and repositioning drugs, we proposed a robust approach that provides a general classifier to separate actives from inactives present in a dataset of ligands for any data-driven LBVS.

The Milk-Way algorithm already has an associated patent BR 10 2019 027703 3 [55].

AUTHOR CONTRIBUTIONS

Methodology, CFVL, MAS; Data curation, CFVL; Investigation, CFVL; Writing original draft, CFVL, LHS; Formal analysis, CFVL, LHS, LFL; Software, CFVL, LHS, LFL; Validation, CFVL, LHS, LFL; Project Administration, MAS; Resources, MAS; Supervision, MAS; Funding Acquisition, MAS; Writing - review & editing: CFVL, LHS, LFL, DCBM, REOR, and MAS.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

ACKNOWLEDGMENTS

The authors acknowledge Pedro Magalhães Martins for technical support.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ABBREVIATIONS

CDK2	:	cyclin-dependent kinase 2
HTS	:	High-throughput screening
LBVS	:	ligand-based virtual screening
RTI	:	reverse transcriptase inhibitors
SBVS	:	structure-based virtual screening

SUPPLEMENTARY MATERIAL

CDK2 model

Data collection

The matrix was composed of drugs described through zeros and ones, done by the Power MV [31]. The descriptors generated were: 735 fragment pair and 147 Pharmacophore fingerprints [32-34]. They are a topological representation of a chemical structure of ligands, some of which have already been used in data mining [26]. This model was based on the only drug commercialized to CDK2, bosutinib (DB06616), used for chronic myelogenous leukemia [56]. The inactives were made by generating 50 decoys, from DUD-E [50] of this drug and using two commercialized drugs for CDK4/6 and their respective decoys. Ribociclib (DB11730) and palbociclib (DB09073) inhibit tumor growth across a diversity of retinoblastoma cancers (Rb+) [57, 58]. We assign drugs for CDK4/6 in the group of inactive ligands for the sake of guaranteeing the specificity for CDK2, since the fact that they belonged to the same family could generate false positives, due to their homology. It has already been proven that CDK2 is structurally and functionally related to CDK1. Also, CDK2 has a considerably broader substrate profile than CDK4 and CDK6 [1].

As we are working with drug repositioning, the matrix of queries consisted of drugs already marketed according to the DrugBank database [16]. Each ligand was described through the same 882 descriptors of the model.

Singular value decomposition

The singular value decomposition (SVD) helps to reduce the dimension of the matrix. All the eigenvalues were chosen based on the Everitt *et al.* criteria [42].

Ad-hoc choice

It is also essential to analyze the P(x) of the drug which acts in CDK4/6, inferring about the specificity of the model, in the family of CDK, due to the fact of homology between then.

Table S1. Number of ligands and features of the training matrix CDK2 and the result of singular value decomposition.

		882*
Original matrix	Active	1
	Inactive	152
SVD	Sum Sr**	0.7005
	Eigenvalue	61
Each training model	Active	1
	Inactive	152

*number of features (fragment pair and Pharmacophore fingerprints).
**Sr = (diagonal(S)/sum(S)) [42].

Table S2. The 1% features with the highest alpha value.

Alpha features
"ARC_06_-O-"
"HY1_03_HY2"
"POS_04_POS"
"POS_04_POS"
"HY1_04_HY1"
"ARC_04_-O-"
"POS_06_-O-"
"HAL_05_HY2"

Alpha values analysis

With the system of equations represented in equation 3, the alpha value translates the impact that each feature has on the categorization of the active or inactive ligand. To get an analysis of the alpha values and how the validation model is categorizing, we compared the 1% of the highest alphas of the model. By the logic and rigor of the equation, it is expected that the higher alpha values belong to the active reference active bosutinib.

Molecular docking protocol

Molecular docking can provide a better understanding of the interactions between a target macromolecule when the target structure is known. Docking begins by sampling different orientations and conformations of the ligand within the target binding site [59, 60]. Afterward, the best positions, the so-called pose, for each ligand are determined by ranking them according to a scoring function [61]. With this strategy, we can predict a possible affinity between ligand and target. We chose DOCK6.8 [48] since it provides multiple scoring functions, from force-field based to pharmacophore-based, and the possibility of combining them.

Therefore, in our case it was useful to incorporate chemical features of known inhibitors and the binding sites to the docking of the selected compounds. We chose a protocol that calculates grid parameters to every residue interacting with the reference, the so-called Multigrid energy score (MGE). The sum of the interactions in each grid equals the interaction of a single grid representing the entire target.

Since DOCK6.8 allows the combinations of score functions, we also included the Pharmacophore Matching Similarity (PHS) score combined with the MGE. PHS is a scoring function that calculates the level of pharmacophore overlap between a reference molecule and a candidate molecule in three-dimensional space. Since MGE already uses a reference molecule, we found that including its pharmacophore overlap in the score component would be useful.

To perform the docking of the compounds chosen by the Milk-Way algorithm, we chose the 2R3Q [62] structures, since it achieved good results in pose reproduction and cross-docking experiments (data not showed). We used its native ligand as reference for the MGE and PHS scoring. OpenBabel [63] was used to convert the ligands from SMILES to the mol2 format. Geometry optimization was performed for all ligands using GAMESS [64], while AM1-BCC partial charges were assigned using AMBER’s antechamber [65].

Table S3. The references of the selected ligands (P(x) > = 0.98).

CID	Name	P(x)	Citation	Reference
DB09345	Pramocaine	0.9970	Pramocaine induced expression changes in ‘Signaling Pathways in Glioblastoma’	[51]
DB00433	Prochlorperazine	0.9919	Drugs with potential antitumor effects	[52]
DB00831	Trifluoperazine	0.9875	Trifluoperazine might be a potential available drug for treating triple-negative breast cancer with brain metastasis, which urgently needs novel treatment options	[52, 53]
DB00134	Methionine	0.9830	-	-
DB01186	Pergolide	0.9813	-	-

Table S4. Values of energy using DOCK scoring MGE + PHS.

Name:	**Bosutinib**	**Pramocaine**	**Prochlorperazine**	**Trifluoperazine**	**Methionine**	**Pergolide**
Reference PDB:	**PDB 2R3Q**					
Descriptor_Score:	-26.16698	-30.83058	-17.86901	-23.67142	-8.72689	-14.40792
MGE_Score:	-30.64930	-35.99844	-21.49072	-28.90000	-13.62098	-19.95671
PHS_Score:	4.954826	5.610463	4.244604	5.756922	5.129757	5.86296

Table S5. Binding mode of the selected ligands with P(x) > = 0.98. The system used was DOCK scoring – MGE + PHS SCORE – using PDB 2R3Q.

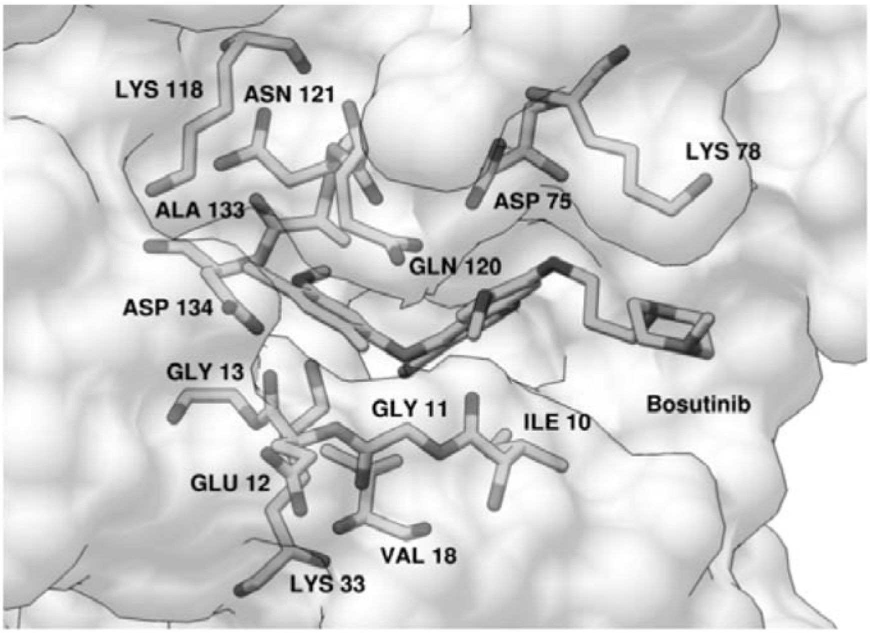
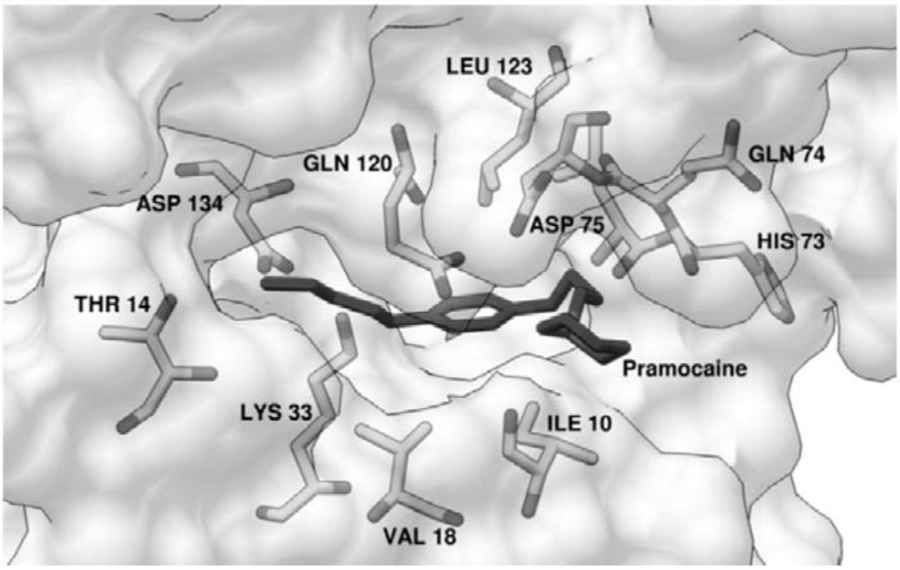
CID	Name	P(x)
DB06616	Bosutinib	-
INTERACTIONS*	 LYS 118 ASN 121 LYS 78 ALA 133 ASP 75 GLN 120 ASP 134 GLY 13 GLY 11 ILE 10 Bosutinib GLU 12 VAL 18 LYS 33	
DB09345	Pramocaine	0.9970
INTERACTIONS*	 LEU 123 GLN 120 GLN 74 ASP 134 ASP 75 HIS 73 THR 14 Pramocaine LYS 33 ILE 10 VAL 18	

Table S5 continued..

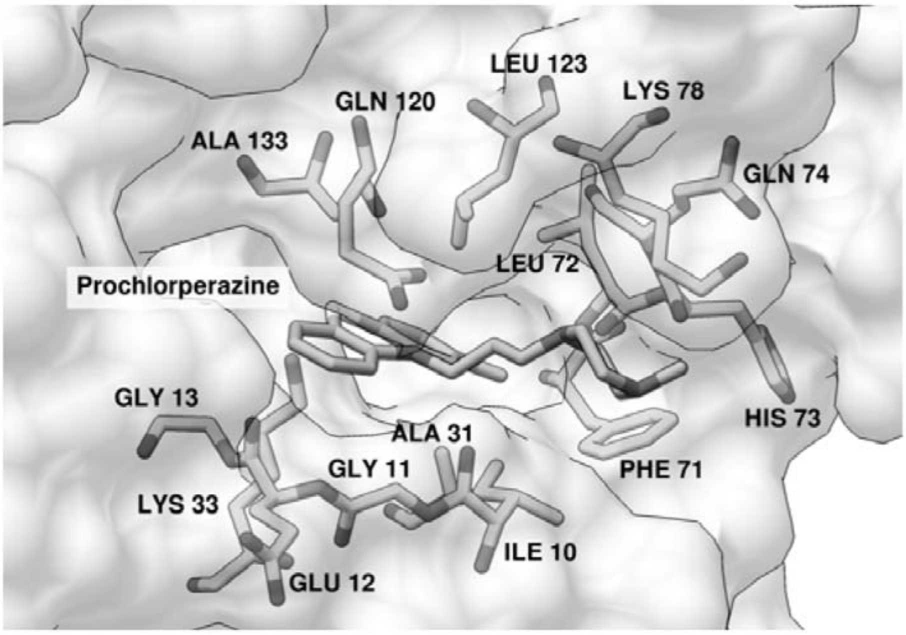
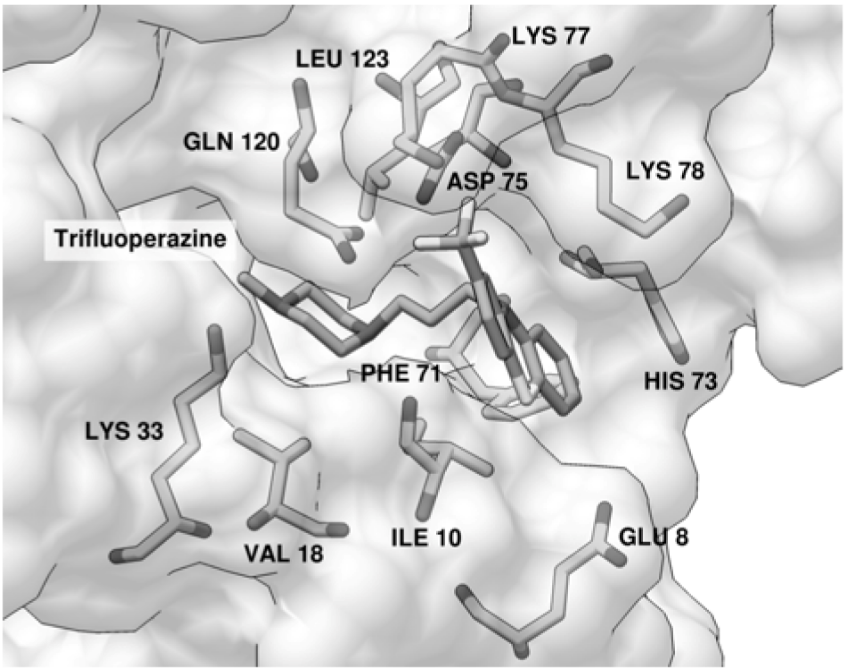
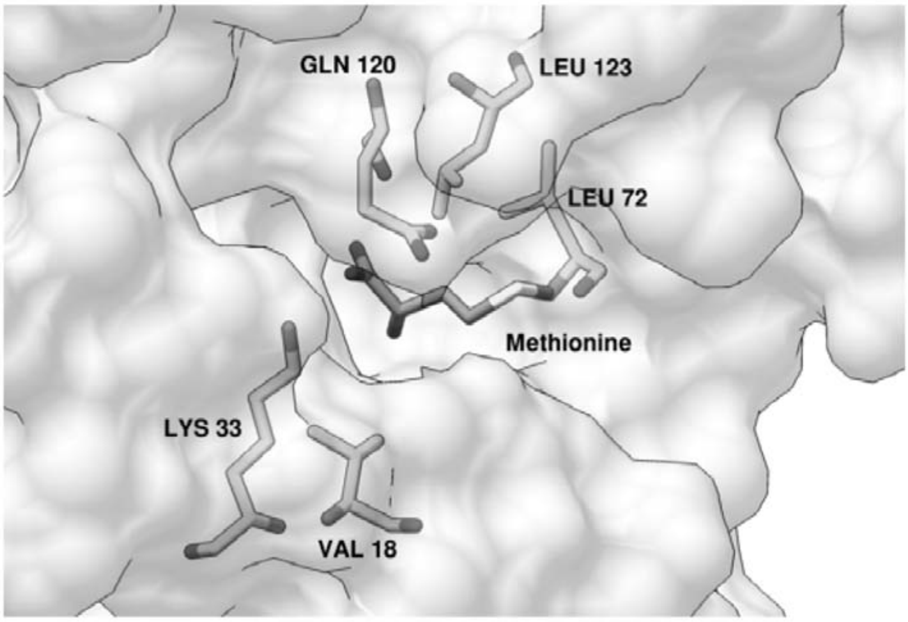
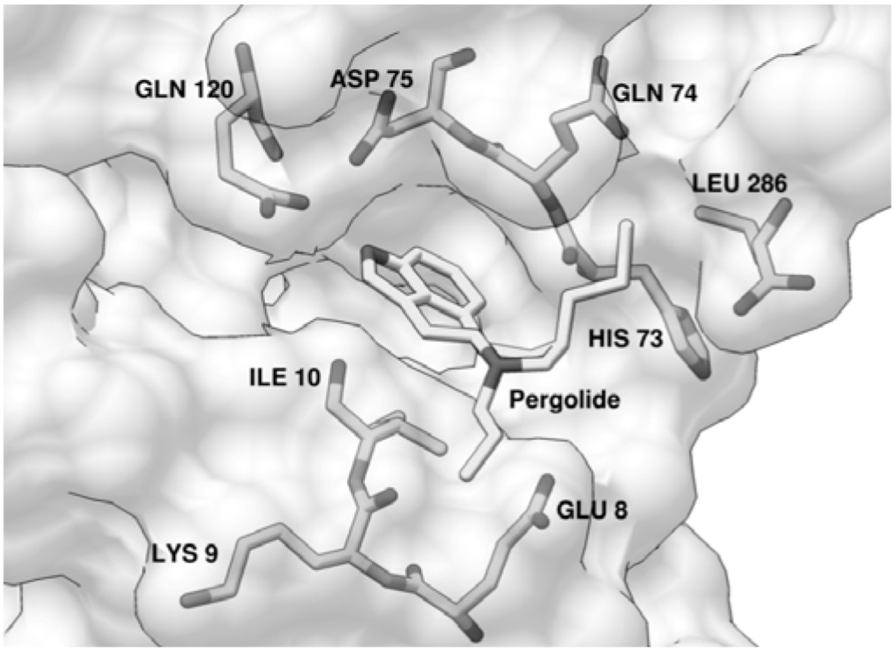
CID	Name	P(x)
DB00433	Prochlorperazine	0.9919
INTERACTIONS*	 LEU 123 LYS 78 GLN 120 ALA 133 GLN 74 LEU 72 Prochlorperazine GLY 13 HIS 73 ALA 31 GLY 11 PHE 71 LYS 33 ILE 10 GLU 12	
DB00831	Trifluoperazine	0.9875
INTERACTIONS*	 LYS 77 LEU 123 GLN 120 LYS 78 ASP 75 Trifluoperazine PHE 71 HIS 73 LYS 33 ILE 10 GLU 8 VAL 18	

Table S5 continued..

CID	Name	P(x)
DB00134	Methionine	0.9830
INTERACTIONS*	 GLN 120 LEU 123 LEU 72 Methionine LYS 33 VAL 18	
DB01186	Pergolide	0.9813
INTERACTIONS*	 GLN 120 ASP 75 GLN 74 LEU 286 HIS 73 ILE 10 Pergolide GLU 8 LYS 9	

*All images were generated with UCSF Chimera [66].

Table S6. Interactions of the selected ligands with P(x) > = 0.98. The system used was DOCK scoring – MGE + PHS SCORE – using PDB 2R3Q.

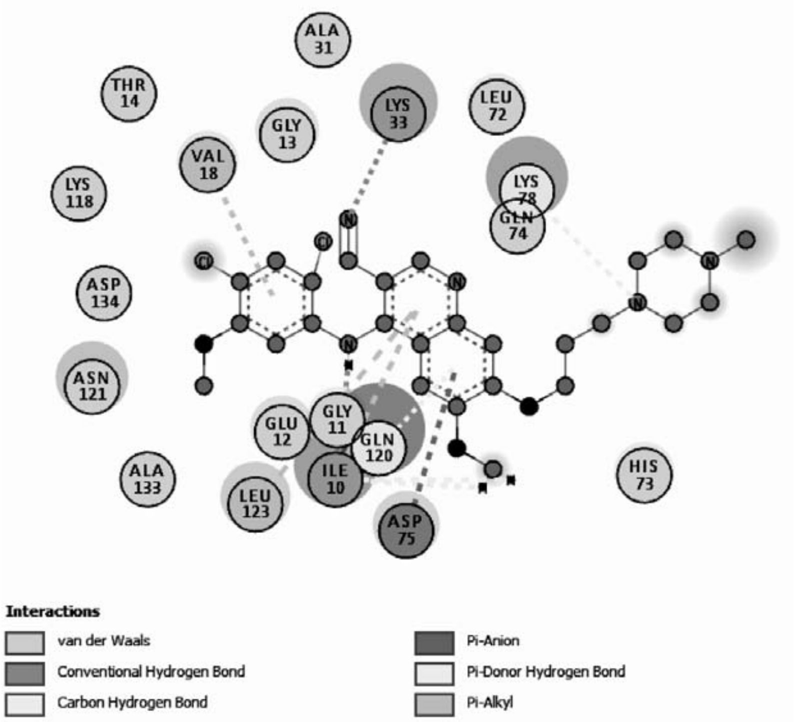
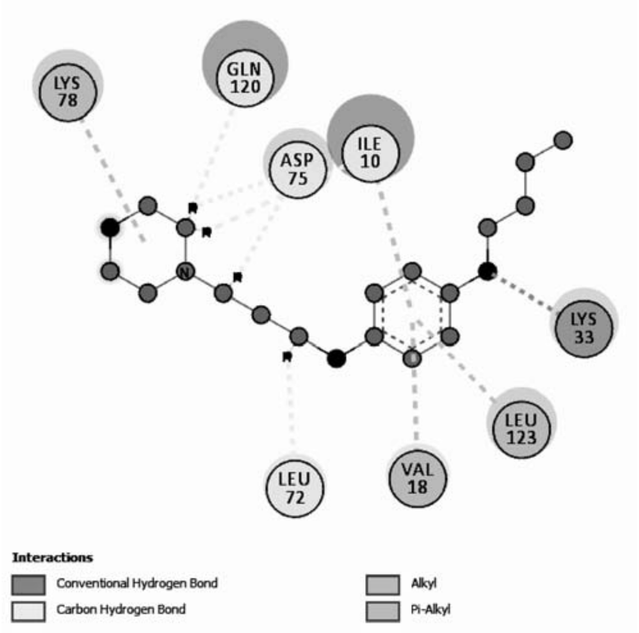
CID	Name	P(x)
DB06616	Bosutinib	-
INTERACTIONS*	 ALA 31 THR 14 GLY 13 LYS 33 LEU 72 VAL 18 LYS 118 LYS 78 GLN 74 ASP 134 ASN 121 GLU 12 GLY 11 GLN 120 ILE 10 ALA 133 LEU 123 ASP 75 HIS 73 Cl O N Interactions van der Waals Conventional Hydrogen Bond Carbon Hydrogen Bond Pi-Anion Pi-Donor Hydrogen Bond Pi-Alkyl	
DB09345	Pramocaine	0.9970
INTERACTIONS*	 LYS 78 GLN 120 ASP 75 ILE 10 LYS 33 LEU 123 LEU 72 VAL 18 N Interactions Conventional Hydrogen Bond Carbon Hydrogen Bond Alkyl Pi-Alkyl	

Table S6 continued..

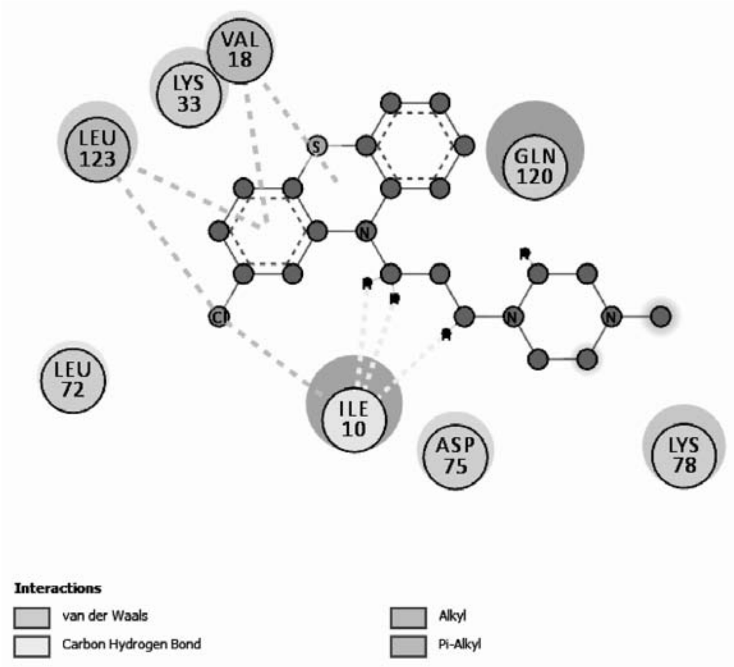
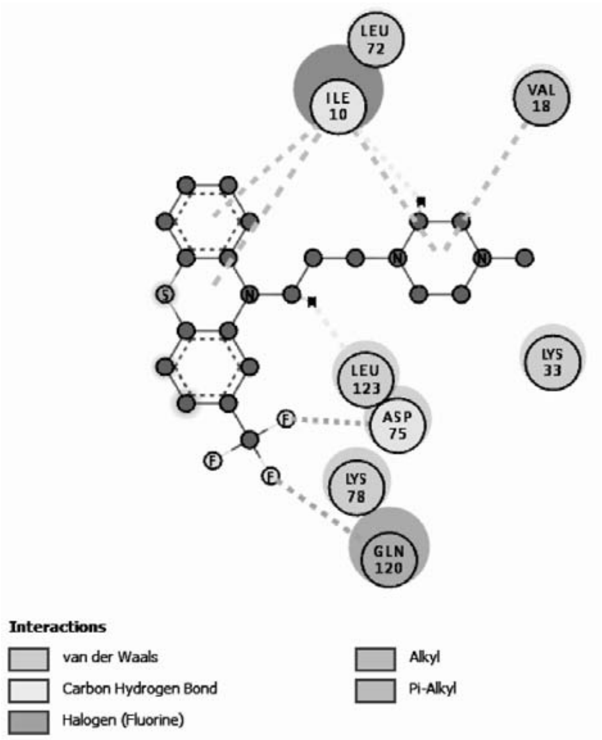
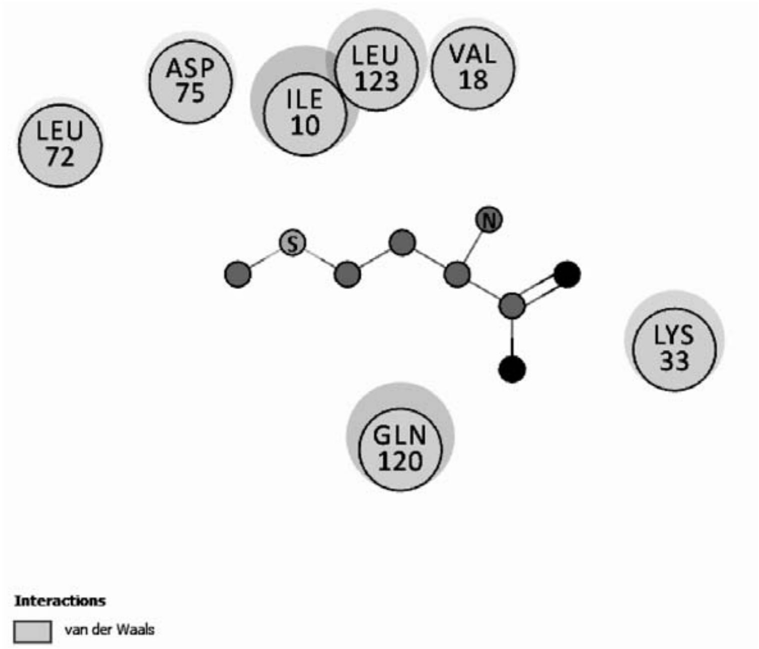
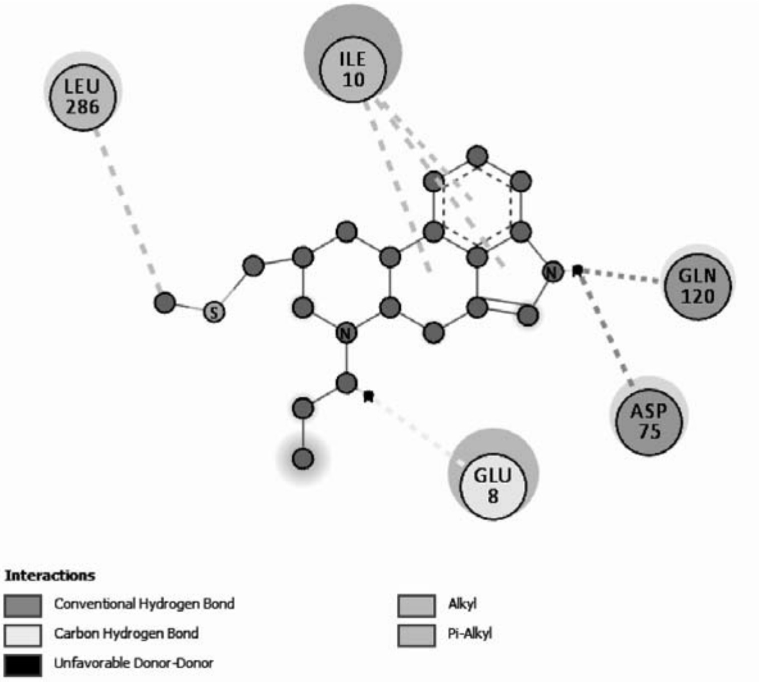
CID	Name	P(x)
DB00433	Prochlorperazine	0.9919
INTERACTIONS*	 VAL 18 LYS 33 LEU 123 S N GLN 120 Cl N N LEU 72 ILE 10 ASP 75 LYS 78 Interactions van der Waals Carbon Hydrogen Bond Alkyl Pi-Alkyl	
DB00831	Trifluoperazine	0.9875
INTERACTIONS*	 LEU 72 ILE 10 VAL 18 N N S N LEU 123 LYS 33 F ASP 75 F F LYS 78 GLN 120 Interactions van der Waals Carbon Hydrogen Bond Halogen (Fluorine) Alkyl Pi-Alkyl	

Table S6 continued..

CID	Name	P(x)
DB00134	Methionine	0.9830
INTERACTIONS*	 ASP 75 LEU 123 VAL 18 ILE 10 LEU 72 S N LYS 33 GLN 120 Interactions van der Waals	
DB01186	Pergolide	0.9813
INTERACTIONS*	 LEU 286 ILE 10 S N N GLN 120 ASP 75 GLU 8 Interactions Conventional Hydrogen Bond Carbon Hydrogen Bond Unfavorable Donor-Donor Alkyl Pi-Alkyl	

*All images were generated in Discovery Studio [49].

REFERENCES

1. Asghar, U., Witkiewicz, A. K., Turner, N. C. and Knudsen, E. S. 2015, Nature Reviews. Drug Discovery, 14(2), 130.
2. Kaldis, P. and Richardson, H. E. 2012, Development, 139(2), 225-230.
3. Betzi, S., Alam, R., Martin, M., Lubbers, D. J., Han, H., Jakkaraj, S. R., Georg, G. I. and Schönbrunn, E. 2011, ACS Chemical Biology, 6(5), 492-501.
4. Johnson, L. 2007, Biochemical Society Transactions, 35(1), 7.
5. Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N. and Andrade, C. H. 2018, Frontiers in Pharmacology, 9, 1275.
6. Seifert, M. H., Wolf, K. and Vitt, D. 2003, Biosilico, 1(4), 143-149.
7. Li, Q. and Shah, S. 2017, Methods Mol. Biol., 1558, 111-124.
8. Stockwell, B. R. 2004, Nature, 432(7019), 846-854.
9. Ceska, T., Chung, C.-W., Cooke, R., Phillips, C. and Williams, P. A. 2019, Biochemical Society Transactions, 47(1), 281-293.
10. Gimeno, A., Ojeda-Montes, M. J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G. and Garcia-Vallvé, S. 2019, International Journal of Molecular Sciences, 20(6), 1375.
11. Geppert, H., Vogt, M. and Bajorath, J. 2010, Journal of Chemical Information and Modeling, 50(2), 205-216.
12. Carpenter, K. A. and Huang, X. 2018, Curr. Pharm. Des., 24, 3347-3358.
13. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. 2008, Nucleic Acids Research, 36(Suppl. 1), D344-D350.
14. Irwin, J. J. and Shoichet, B. K. 2005, Journal of Chemical Information and Modeling, 45(1), 177-182.
15. Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J., Xiao, J., Zhang, J. and Bryant, S. H. 2010, Nucleic Acids Research, 38(Suppl. 1), D255-D266.
16. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., and Neveu, V. 2011, Nucleic Acids Research, 39(Suppl. 1), D1035-D1041.
17. Sharman, J. L., Mpamhanga, C. P., Spedding, M., Germain, P., Staels, B., Dacquet, C., Laudet, V. and Harmar, A. J. 2011, Nucleic Acids Research, 39(Suppl. 1), D534-D538.
18. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. 2010, Nucleic Acids Research, 38(Suppl. 1), D355-D360.
19. Soufan, O., Ba-alawi, W., Magana-Mora, A., Essack, M. and Bajic, V. B. 2018, Scientific Reports, 8(1), 9110.
20. Plewczynski, D., Spieser, S. A. H. and Koch, U. 2006, Journal of Chemical Information and Modeling, 46(3), 1098-1106.
21. Han, L., Wang, Y. and Bryant, S. H. 2008, BMC Bioinformatics, 9(1), 1.
22. Riniker, S., Wang, Y., Jenkins, J. L. and Landrum, G. A. 2014, Journal of Chemical Information and Modeling, 54(7), 1880-1891.
23. Chen, B., Wild, D. and Guha, R. 2009, Journal of Chemical Information and Modeling, 49(9), 2044-2055.
24. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. and Hopkins, A. L. 2006, Nature Biotechnology, 24(7), 805-815.
25. Hao, M., Wang, Y. and Bryant, S. H. 2014, Analytica Chimica Acta, 806, 117-127.
26. Schierz, A. C. 2009, Journal of Cheminformatics, 1, 21.
27. Trunk, G. V. 1979, IEEE Transactions on Pattern Analysis & Machine Intelligence, 1(3), 306-307.
28. Dai, W. and Guo, D. 2019, Molecules, 24(13), 2414.
29. Yin, L., Ge, Y., Xiao, K., Wang, X. and Quan, X. 2013, Neurocomputing, 105, 3-11.
30. Diallo, A. and Prigent, C. 2011, Bull Cancer, 98(11), 1335-1345.
31. Liu, K., Feng, J. and Young, S. S. 2005, Journal of Chemical Information and Modeling, 45(2), 515-522.
32. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S. and Pujadas, G. 2015, Methods, 71, 58-63.

33. Ahmed, H. E., Vogt, M. and Bajorath, J. R., 2010, Journal of Chemical Information and Modeling, 50(4), 487-499.

34. Awale, M. and Reymond, J. L. 2014, Journal of Chemical Information and Modeling, 54(7), 1892-1907.

35. Liu, H., Motoda, H., Setiono, R. and Zhao, Z. 2010, Feature Selection in Data Mining, 4-13.

36. Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., Shoemaker, B. A., Gindulyte, A. and Bryant, S. H. 2014, Nucleic Acids Research, 42(D1), D1075-D1082.

37. Élden, L. 2006, Acta Numerica, 15.

38. Berry, M. W., Dumais, S. T. and OBrien, G. W. 1995, SIAM Review, 37.

39. Santos, A. R., Santos, M. A., Baumbach, J., McCulloch, J. A., Oliveira, G. C., Silva, A., Miyoshi, A. and Azevedo, V. 2011, BMC Genomics, 12(4), 1-15.

40. Kumar, N., Nasser, M. and Sarker, S. C. 2011, Journal of Geography and Geology, 3(1), 227.

41. Silverio-Machado, R., Couto, B. R. and Dos Santos, M. A. 2014, Bioinformatics, 31(8), 1267-1273.

42. Everitt, B. S. D., Everitt, G. B. S. and Dunn, G. 1991, Applied Multivariate Data Analysis.

43. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990, Journal of the American Society for Information Science, 41(6), 391.

44. Morgan, S. P. 1998, Notices Amer. Math. Soc., 45(8), 972-977.

45. Cokluk, O. 2010, Educational Sciences: Theory and Practice, 10(3), 1397-1407.

46. Luenberger, D. G. and Ye, Y. 2016, Linear and Nonlinear Programming, Springer.

47. Golub, G. H. 1965, Numerical Mathematics, 7(3), 206-216.

48. Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D. and Rizzo, R. C. 2015, Journal of Computational Chemistry, 36(15), 1132-1156.

49. Biovia, D. S. 2017, Discovery Studio Modeling Environment, Dassault Systèmes: San Diego.

50. Mysinger, M. M., Carchia, M., Irwin, J. J. and Shoichet, B. K. 2012, Journal of Medicinal Chemistry, 55(14), 6582-6594.

51. Hardt, C., Beber, M. E., Rasche, A., Kamburov, A. and Herwig, R. 2019, ToxDB. Vertebrate Genomics Department at the Max Planck Institute for Molecular Genetics in Berlin, Germany.

52. Qi, L. and Ding, Y. 2013, Science China Life Sciences, 56(11), 1020-1027.

53. Feng, Z., Xia, Y., Gao, T., Xu, F., Lei, Q., Peng, C., Yang, Y., Xue, Q., Hu, X., Wang, Q., Wang, R., Ran, Z., Zeng, Z., Yang, N., Xie, Z. and Yu, L. 2018, Cell Death & Disease, 9(10), 1006.

54. Ou-Yang, S.-S., Lu, J.-Y., Kong, X.-Q., Liang, Z.-J., Luo, C. and Jiang, H. 2012, Acta Pharmacologica Sinica, 33(9), 1131-1140.

55. Figueiredo Vieira Leite, C., Dos Santos, M. A., Silva Dos Santos, L. H., Fernando Leijôto, L., Batista Mariano, D. C. and Oliveira Rocha, R. E. 2019, Método de Triagem de compostos baseados em Regressão Logística Modificada.

56. Amsberg, G. K. and Schafhausen, P. 2013, Biologics, 7, 115-122.

57. Infante, J. R., Cassier, P. A., Gerecitano, J. F., Witteveen, P. O., Chugh, R., Ribrag, V., Chakraborty, A., Matano, A., Dobson, J. R., Crystal, A. S., Parasuraman, S. and Shapiro, G. I. 2016, Clinical Cancer Research, 28(23), 5696-705.

58. DeMichele, A., Clark, A. S., Tan, K. S., Heitjan, D. F., Gramlich, K., Gallagher, M., Lal, P., Feldman, M., Zhang, P., Colameco, C., Lewis, D., Langer, M., Goodman, N., Domchek, S., Gogineni, K., Rosen, M., Fox, K. and O'Dwyer, P. 2015, Clinical Cancer Research, 21(5), 995-1001.

59. Meng, X. Y., Zhang, H. X., Mezei, M. and Cui, M. 2011, Current Computational Aided Drug Design, 7(2), 146-157.

60. Yuriev, E. and Ramsland, P. A. 2013, Journal Molecular Recognition, 26(5), 215-239.

61. Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T. and Altman, R. B. 2012, Journal of the Royal Society Interface, 9(72), 1409-1437.

62. Fischmann, T. O., Hruza, A., Duca, J. S., Ramanathan, L., Mayhood, T., Windsor, W. T., Le, H. V., Guzi, T. J., Dwyer, M. P., Paruch, K., Doll, R. J., Lees, E., Parry, D., Seghezzi, W. and Madison, V. 2008, Biopolymers, 89(5), 372-379.
63. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R. 2011, Journal of Cheminformatics, 3(1), 33.
64. Gordon, M. S. and Schmidt, M. W. 2005, Advances in electronic structure theory: GAMESS a decade later. Theory and Applications of Computational Chemistry: the first forty years, C. E. Dykstra, G. Frenking, K. S. Kim and G. E. Scuseria (Eds), Elsevier, 1167-1189.
65. Wang, J., Wang, W., Kollman, P. A. and Case, D. A. 2006, Journal of Molecular Graphics and Modelling, 25(2), 247-260.
66. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. 2004, Journal of Computational Chemistry, 25(13), 1605-1612.