Molecular Diversity (2021) 25:1301–1314
https://doi.org/10.1007/s11030-021-10261-z

ORIGINAL ARTICLE

Molecular insights on ABL kinase activation using tree-based machine learning models and molecular docking

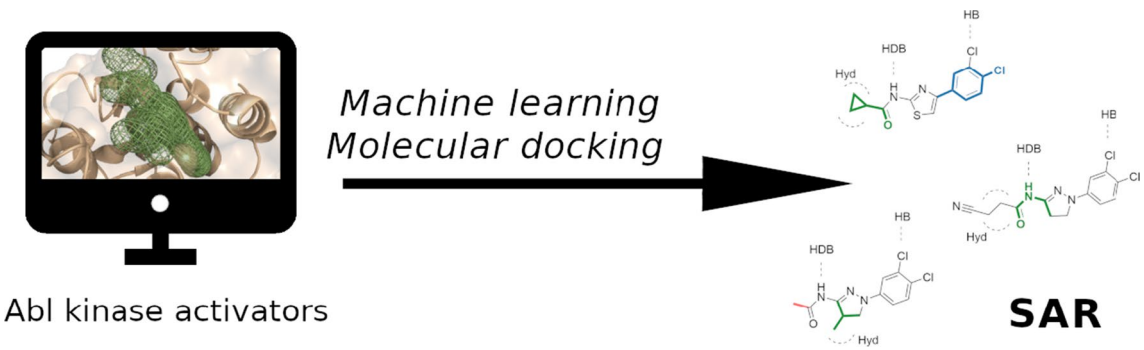
Philipe Oliveira Fernandes[1] · Diego Magno Martins[2] · Aline de Souza Bozzi[2] · João Paulo A. Martins[2] · Adolfo Henrique de Moraes[2] · Vinícius Gonçalves Maltarollo[1]

Received: 31 March 2021 / Accepted: 18 June 2021 / Published online: 30 June 2021

Abstract

Abelson kinase (c-Abl) is a non-receptor tyrosine kinase involved in several biological processes essential for cell differentiation, migration, proliferation, and survival. This enzyme's activation might be an alternative strategy for treating diseases such as neutropenia induced by chemotherapy, prostate, and breast cancer. Recently, a series of compounds that promote the activation of c-Abl has been identified, opening a promising ground for c-Abl drug development. Structure-based drug design (SBDD) and ligand-based drug design (LBDD) methodologies have significantly impacted recent drug development initiatives. Here, we combined SBDD and LBDD approaches to characterize critical chemical properties and interactions of identified c-Abl's activators. We used molecular docking simulations combined with tree-based machine learning models—decision tree, AdaBoost, and random forest to understand the c-Abl activators' structural features required for binding to myristoyl pocket, and consequently, to promote enzyme and cellular activation. We obtained predictive and robust models with Matthews correlation coefficient values higher than 0.4 for all endpoints and identified characteristics that led to constructing a structure–activity relationship model (SAR).

Graphic abstract



Keywords ABL kinase activators · Machine learning · Molecular docking · LBDD · SBDD · QSAR · SAR

Philipe Oliveira Fernandes, Diego Magno Martins and Aline de Souza Bozzi have contributed equally to this work.

✉ Vinícius Gonçalves Maltarollo
maltarollo@ufmg.br

[1] Departamento de Produtos Farmacêuticos, Faculdade de Farmácia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2] Departamento de Química, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

Abbreviations

AB	AdaBoost
CA	Cellular activation
c-Abl	Abelson kinase
CML	Chronic myeloid leukemia
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative similarity indices analysis
DT	Decision tree
EA	Enzyme activation

GA	Genetic algorithm
HQSAR	Hologram quantitative structure–activity relationship
LBDD	Ligand-based drug design
HB	Halogen bond
HBD	Hydrogen bond donor
Hyd	Hydrophobic contact
MACCS	Molecular ACCess system
MB	Myristoyl binding
MCC	Matthews correlation coefficient
$_{CV}$MCC	Cross-validation MCC
$_{EXT}$MCC	External validation MCC
PC	Principal component
PCA	Principal component analysis
PDB	Protein data bank
pEC_{50}	Logarithmic half-maximal effective concentration
pIC_{50}	Logarithmic half-maximal inhibitory concentration
PLIP	Protein–ligand interaction profiler
QSAR	Quantitative structure–activity relationship
RF	Random forest
RMSD	Root mean square-deviation
SAR	Structure–activity relationship
SBDD	Structure-based drug design
SMARTS	SMILES arbitrary target specification

Introduction

Abelson kinase (c-Abl) is a non-receptor tyrosine kinase located in many subcellular compartments, including the endoplasmic reticulum, cytoplasm, nucleus, cell cortex, and mitochondria [1]. This enzyme modulates several biological processes, including actin polymerization [2, 3], structural changes in chromatin [4], responses to DNA damage [5, 6], and other essential ones for cell proliferation, differentiation, migration, survival, and death [7].

The inhibition of c-Abl has been studied since its overactivation is associated with various diseases such as chronic myeloid leukemia (CML), cancer, immunological diseases, neurological disorders such as Parkinson's disease and others [8–11].

However, the activation of c-Abl has become the subject of investigation in the last years since the transient activation of this enzyme may have a therapeutic application such as the treatment of chemotherapy-induced neutropenia [12–14]. Moreover, c-Abl activators can block TGFβ-responsive mammary tumor growth in mice, showing a potential strategy to deal with breast cancer [15]. Abl kinases can also be oncogenes in some cases, as in CML, and tumor suppressors in others, such as prostate cancer with α3 integrin deficiency [16]. For this reason, in specific circumstances, the use of small-molecule activators of c-Abl kinases could help prostate cancer treatment [14]. Other potential applications of activators include treating ischemic injury [17] and synergistic effects when combined with BCR-Abl inhibitors that would improve treatment efficiency [18].

The c-Abl autoinhibition is mainly regulated by a series of intramolecular interactions in the SH3 and SH2 domains that stabilize the kinase domain inactive conformation [1, 19]. Another regulation mechanism is the N-terminal myristoyl binding to the myristoyl binding site in the kinase domain's C-lobe (Fig. 1a, b). This binding helps form a compact conformation in which the enzyme is in an autoinhibited state [20, 21].

Yang et al. identified in 2011 the DHP, the first cell-permeable molecule capable of activating the c-Abl [21] (Fig. 1c). Furthermore, subsequent studies carried out by the same research group [22] identified a series of compounds capable of activating c-Abl whose mechanism of action involves the binding to the myristoyl binding site, inducing conformational changes [23, 24].

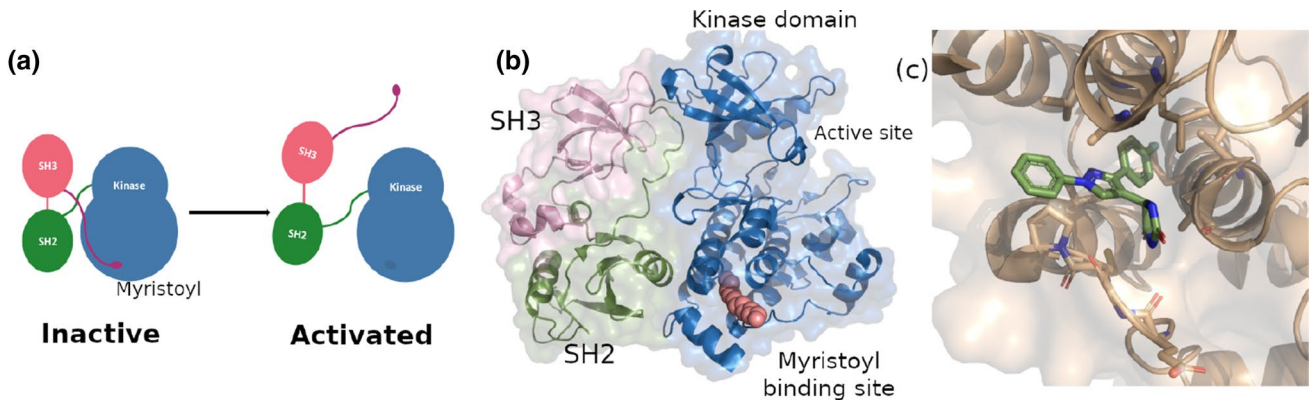


Fig. 1 **a** Representation of the conformational changes in the c-Abl kinase during their autoinhibition process; **b** cartoon representation of c-Abl kinase colored by their domains (PDB ID 2FO0) and **c** representation of DPH bounded to the myristoyl binding site (PDB ID 3PYY)

Compared to inhibitors, only a few examples of small-molecule activators of enzymes have their mechanism of action well-characterized. Moreover, beyond the possible therapeutic applications, the study of enzymatic activators can help understand how the enzyme's activation can affect a particular metabolic pathway and explain the conformational changes that govern the protein function [25]. However, even though recent studies have contributed to an advance in understanding c-Abl activation mechanisms, few investigations have explored the rational design of molecules capable of performing this function [22].

Traditional drug development had an estimated cost of 58.8 billion USD in 2015, 10% higher than in 2014 [26]. Besides being expensive, this process is also time-consuming, requiring about 10–15 years. The high cost and time associated with a low success rate in the traditional drug discovery process highlighted the necessity of using computer-aided drug discovery (CADD) in the drug development pipeline [27].

The computational strategies could be classified as structure-based drug design (SBDD) and ligand-based drug design (LBDD). The SBDD process analyzes the interactions between the molecular target and ligand to rationalize the design of novel bioactive compounds [28]. A classic example of the SBDD approach is molecular docking. This method aims to estimate the binding mode using searching algorithms and the interaction energies using a scoring function [29].

The LBDD is an indirect approach based on analysis of the physical chemistry properties or molecular features of known active compounds [30]. This approach is advantageous compared to SBDD since it does not require preliminary knowledge of the biological target. Many LBDD strategies use quantitative structure–activity relationship techniques (QSAR) such as comparative molecular field analysis (CoMFA) [31, 32], comparative similarity indices analysis (CoMSIA) [33, 34], hologram quantitative structure–activity relationship (HQSAR) [35–37], QuBiLS-MIDAS [38, 39], radial distribution function (RDF) indices [40], and GETAWAY descriptors [41]. In addition, machine learning approaches have increasingly been applied in drug design and adopted by many pharmaceutical industries [42–50].

The tree-based models proposed by Breiman et al. in 1984 [51] are good examples of machine learning algorithms for drug design purposes [52, 53]. In these models, the goal is to split the dataset into binary groups with the highest possible homogeneity. In the beginning, the chosen feature allows the highest gain on homogeneity. Then, as the tree grows, it adds other features to the splitting process for increasing the homogeneity between these groups [54]. Well-known examples of these models are the decision tree (DT) [55], random forest (RF) [56], and adaptive boosting (AdaBoost) [57]. Tree-based models are widely used in several research areas, such as metabolomics [58], disease detection [59], toxicological predictions [60–62], and stock prediction [63], due to their simplicity, efficiency, and interpretability.

In this context, this work aimed to study the structure–activity relationship (SAR) of a series of c-Abl kinase activators by combining molecular docking simulations with tree-based classification machine learning models, AdaBoost, decision tree, and random forest, for predicting myristoyl binding (MB), enzyme activation (EA), and cellular activation (CA).

Materials and methods

Dataset description and preparation

We selected compounds classified as c-Abl kinase activators identified by Simpson et al. [22]. We classified the c-Abl kinase activators according to enzyme activation capability and myristoyl binding affinity, measured as the logarithmic half-maximal inhibitory concentration (pIC_{50}), and cellular activation, measured as the logarithmic half-maximal effective concentration (pEC_{50}).

The structures of all compounds are shown in Fig. 2, and their biological activities are listed in Support Information Table S1. Among 52 reported compounds, 49 were used since the racemate and chiral compounds lacking defined stereochemistry 33, 45, and 50 were excluded. Compounds' 3D structures were generated using Avogadro 1.2.0 software [64]. Geometry optimization was performed using the Open Babel package [65] with MMFF94 force-field [66–70], Steepest Descent optimization algorithm, and Newton's method linear search.

Molecular docking

Simpson et al. [22] solved three c-Abl structures with enzyme activators using X-ray crystallography (PDB IDs: 6NPV, 6NPU, and 6NPE). The selected protein structure for docking simulation was the chain A of the one co-crystallized with compound 51 (PDB ID: 6NPV). This crystal structure was chosen because it has the best crystal resolution (1.86 Å). Then, the protein structure was prepared for docking by removing all crystallographic water and phosphate molecules, and the amino acid residues ionization states were adjusted using Discovery Studio Visualizer v19.1.0.18287 [71]. Molecular docking simulations were carried out using GOLD 5.8.1 software [72], using a grid of radius of 10 Å centered at the ALA452 CB, which contained the myristoyl binding site. All remaining parameters regarding ligand flexibility were kept as default. The genetic algorithm used in GOLD was set to maximum search efficiency

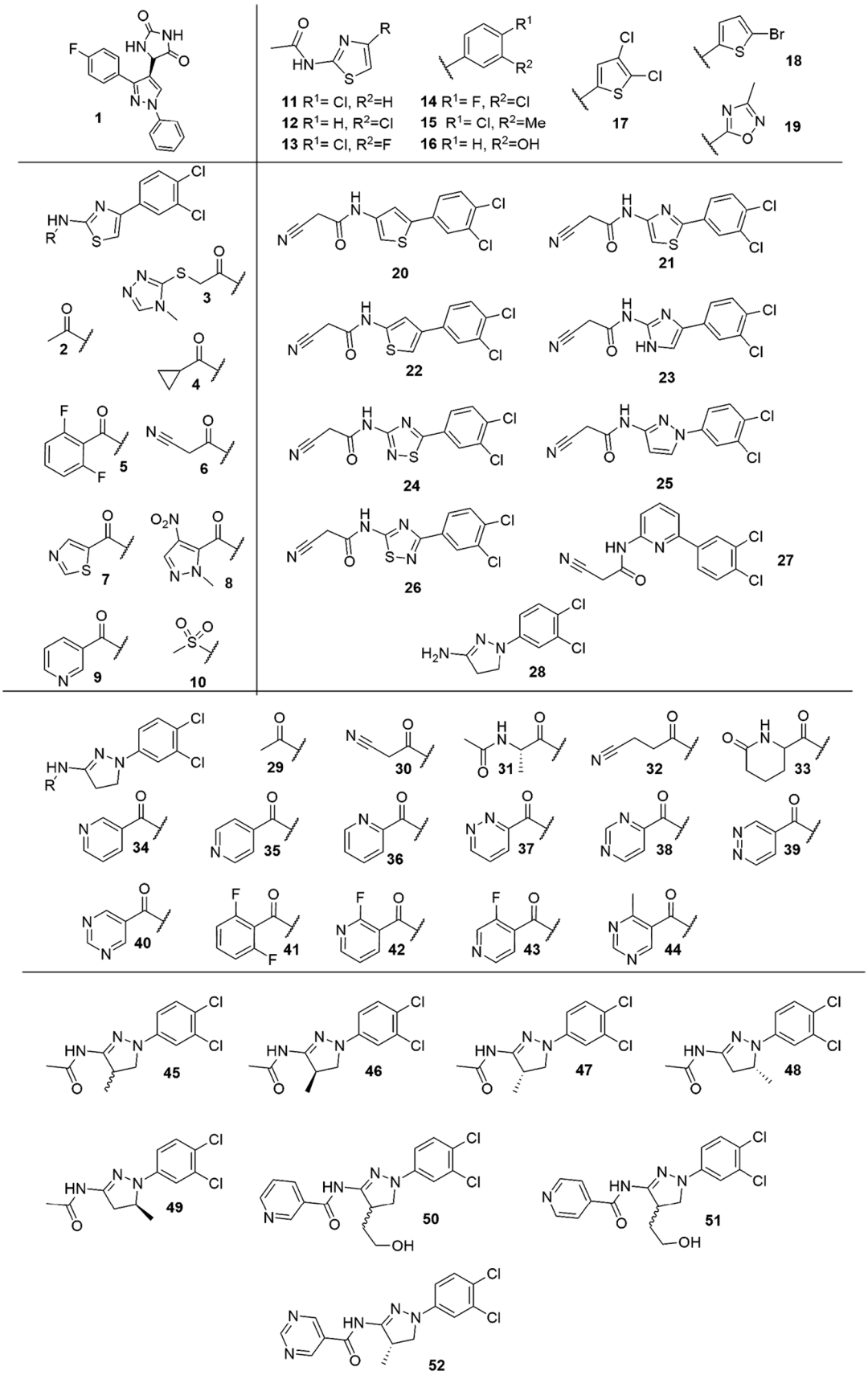


Fig. 2 Structure of the 52 c-Abl activators identified by Simpson et al. [22]

with 50 GA runs per ligand. All conformations were classified according to the ChemPLP [73] score function.

Redocking using ligand 51 and cross-dockings using ligands 6 and 29 were carried out to evaluate the experimental binding mode's predictability and validate the docking protocol. In this approach, compounds 6 and 29 were docked in the selected structure and were compared with their co-crystallized ones deposited in PDB under the ID: 6NPE (resolution of 2.15 Å), and 6NPU (resolution of 2.33 Å), respectively.

Root mean square deviation (RMSD) calculation was the criteria used to assess whether the simulation conditions were adequate when comparing the docking results with the crystallized ligands. Besides, the results of each ligand were grouped in clusters of poses that differ by a maximum of 1 Å from one another. In this process, the best score ranking poses were selected based on the most representative clusters that reproduced the experimental data and were analyzed using the software PyMOL v1.8 [74], Discovery Studio v19.1.0.18287, and PLIP algorithm [75] for poses visualization, interactions evaluation, and figures creation.

Machine learning models

The calculation of molecular fingerprints was performed with PaDEL descriptors [76]. The selected fingerprints for constructing the machine learning models were AtomPairs2D, Klekota Roth, MACCS, PubChem, and Substructure, based on their interpretability. Compounds were divided into active and inactive according to their biological activity parameters (myristoyl binding—MB, enzymatic activation—EA, and cellular activation—CA). For values of either pEC_{50} or pIC_{50} not precisely determined for the endpoint, the entry was removed from the mean activity value calculation. Thus, the MB and EA models were built using 48 compounds with mean activity values of 5.733 and 5.925, respectively, while the CA model was built from 44 compounds with a mean activity value of 5.270.

All models were constructed using similar strategies previously applied [53, 77]. Python library Scikit-learn was used for the data analysis [78] (Fig. 3). Also, the random training test splits were performed using the *train_test_split* module from Scikit-learn in an 80:20 ratio based on random selection's capability to generate predictive models [79]. The AdaBoost (AB) models were constructed varying the estimators' maximum number that ended the boosting process, from 1 to 106 (in 5 steps), 200, 500, and 600 (*n_estimators*) and the learning rate from 0.1 to 20 (in 0.1 steps) (*learning_rate*). For the Decision Tree (DT), the maximum depth of the tree varied from 10 to 100 (in 10 steps) and without limitation (*max_depth*). The minimum number of samples required to split an internal node was varied from 2 to 100 (in 2 step size) (*min_samples_split*). The random forest (RF) models were carried out by varying the same *max_depth* and *min_samples_split* from the DT models and same *n_estimators* from AB models. Accuracy, precision, recall, F1 score, and Matthews correlation coefficient (MCC) were calculated for each model using internal and external validations. Internal validation was also carried out using the fivefold cross-validation module (*cross_validate*). The model evaluation was performed using the MCC values for validation processes, and models with no positive predicted values were excluded in further analysis. The best model for each endpoint had an X-scramble validation following the SCRAMBLE'N'GAMBLE methodology [80].

The applicability domain was assessed by the bounding box approach using principal component analysis (PCA) [81, 82] for the best model for each endpoint, using the implemented method on the Scikit-learn library. Only the independent variables selected by each endpoint from the training set were employed in the model construction. The same model was used to transform the test set data. The PCA data also were applied to perform applicability domain based on the range, and the distance of each test sample from the training dataset using the Euclidean, Manhattan, Cosine, and Wasserstein (probability distribution) distances implemented in the Scipy library [83], obtaining a consensus analysis for applicability domain [84] using a threshold of 95%.

The best result for each endpoint was interpreted using the permutation importance from the Scikit-learn library (*permutation_importance*) several times to permute a feature equals 10 (*n_repeats*) using MCC as the metric. Features were interpreted using the SMARTS pattern with the SMART.plus web service [85]. Plotting was performed using Matplotlib and Seaborn libraries.

Results and discussion

Molecular docking

The RMSD values obtained in the redocking and cross-docking processes were equal to 0.878 Å (Fig. 4a), 0.402 Å (Fig. 4b), and 0.698 (Fig. 4c) for compounds 51, 29, and 6, respectively. Therefore, those results indicated that the employed docking protocol is suitable for pose prediction because RMSD values lower than 1.5 or 2 Å, depending on ligand size, were taken to indicate that the docking protocol successfully predicted the experimental binding mode [86]. Then, the same protocol was used to perform the docking simulations for all other compounds.

The myristoyl binding site is composed of several hydrophobic residues [20]. It was observed that the aromatic ring of the compounds fits in a hydrophobic pocket, a region deeper into the myristoyl binding site where a series of interactions including π-stacking with PHE512, Van der Waals

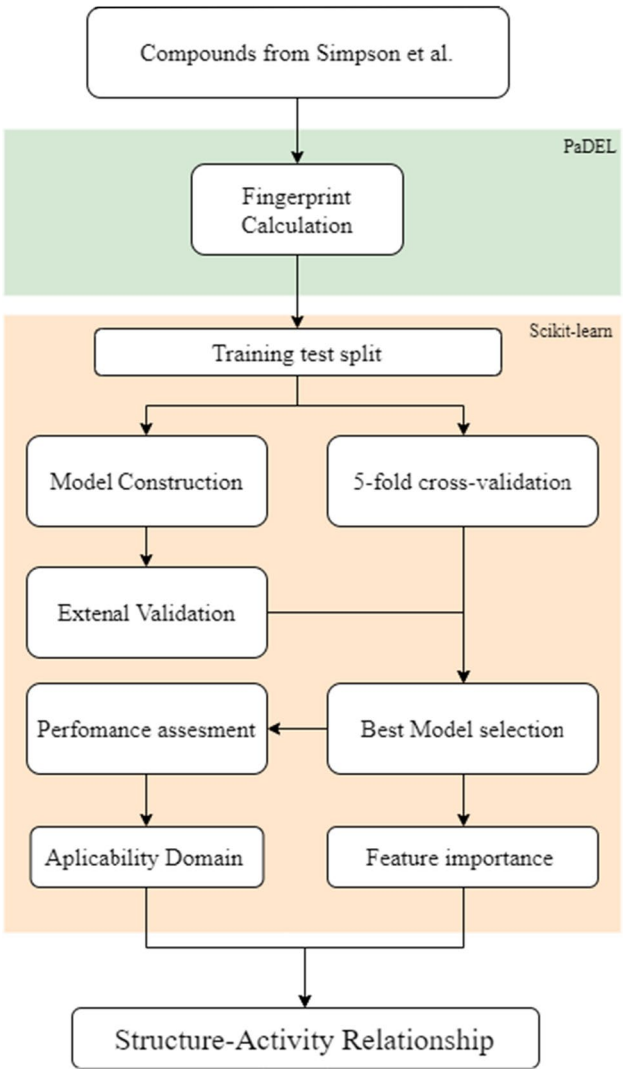


Fig. 3 Flowchart of the machine learning process applied in this work

interactions with ALA363, LEU359, LEU448, ILE451, and VAL487 [87], and halogen bonding with LEU448 occur.

In the region most exposed to the solvent, the compounds access various interaction sites such as hydrogen bond acceptors and donors and Van der Waals interactions, depending on their substituting groups.

Machine learning models

Starting from five fingerprints, 479 AdaBoost (AB) models, 549 decision tree (DT) models, and 2.903 random forest (RF) models were built and validated for each fingerprint and endpoint, summing 59,640 machine learning models (a boxplot analysis for each model and fingerprints are displayed in the Supplementary Figure S1). The cross-validation MCC value was selected to evaluate the model's performance due to their capability of classifying the performance by a single value, comprising all parameters of a confusion matrix [88, 89]. In this sense, if the model had the higher $_{CV}$MCC and $_{EXT}$MCC, then it was selected. Otherwise, the distance was considered from perfection for these two metrics to select a balanced model in both validations. Figure 5a illustrates the first situation where can be seen a model from MACCS fingerprint having the higher $_{CV}$MCC and $_{EXT}$MCC among them, while Fig. 5b, c shows the second.

For MB models, AdaBoost performed better than the others for both metrics using the PubChem fingerprint ($_{CV}$MCC of 0.445 and $_{EXT}$MCC of 0.612). Also, the most predictive random forest model used the same fingerprint. For this endpoint, the decision tree achieved the highest result using AtomPairs2D, but this model displayed the worst generalization capability compared to the others.

The selected models for the EA were considered perfect for the three methods according to the MCC value for external validation and achieved a high generalization capability in the cross-validation process. RF was the best method for

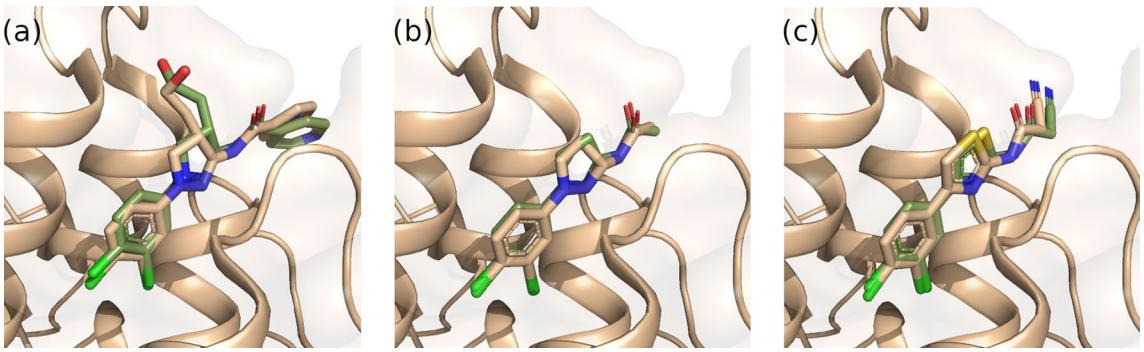


Fig. 4 Molecular docking validation results. **a** Redocking with compound 51 in the myristoyl binding site of Abl kinase crystal structure (PDB ID 6NPV); **b** Cross-docking of compound 29 in the Abl crystal structure (PDB ID 6NPU); and **c** Cross-docking of compound 6 in the Abl crystal structure (PDB ID 6NPE)

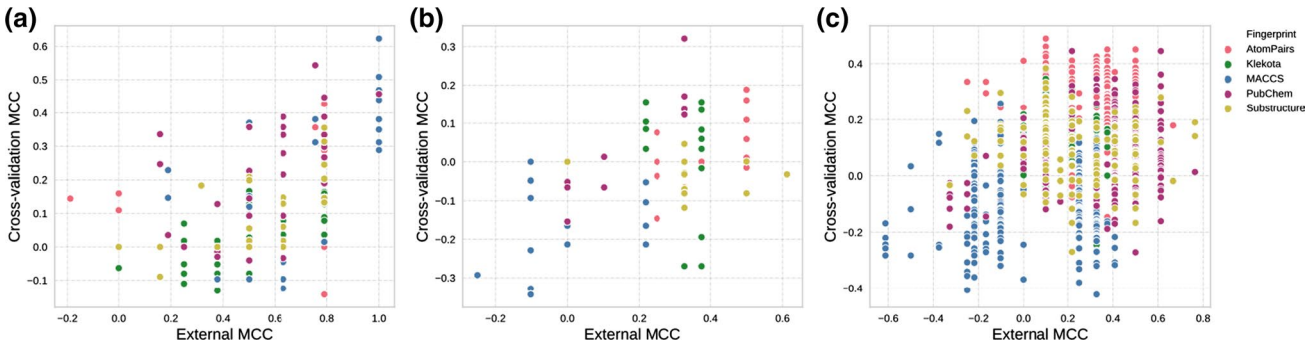


Fig. 5 Scatter plot of cross-validation and external validation MCC values from different methods and endpoints: **a** Random forest for cellular activation; **b** decision tree for myristoyl binding; and **c** Adaboost for myristoyl binding

the $_{CV}$MCC displaying a 0.616 value, followed by AB and DT with, respectively, 0.603 and 0.603 MCC values.

At least, RF performed better in the models for CA, achieving a perfect score for external validation and 0.622 in the cross-validation using MACCS fingerprint. Also, AdaBoost with AtomPairs2D achieved internal and external MCC values equal to 0.536 and 0.790, respectively. Decision tree with PubChem achieved a $_{CV}$MCC of 0.445 and $_{EXT}$MCC of 0.612. Table 1 summarizes the result of the selected models for each endpoint. Only AB with PubChem model for myristoyl binding, RF with PubChem model for enzymatic activation, and RF with MACCS model for cellular activation were considered for further analysis due their higher scores during the validation process and the remain models were disregarded.

It is important to mention that the models were generated with a small amount of chemical and biological data and, as expected, could present some bias and low variance which could led to the underfitting of the models [90]. Regarding the recall, a relevant metric for finding new active compounds, the RF models for enzymatic and cellular activation performed perfectly achieving the value equals to 1 while the model for myristoyl binding achieved a value equal to 0.625. However, despite the lower results from model for myristoyl binding model, the models were properly validated and could be used together (consensus predictions) to achieve a better success rate in finding new activators, strategies already used described in the literature [91–94]. Then, additional validations were carried out to verify the robustness of the three selected models.

The selected model for each endpoint was submitted to another validation process to ensure the model's predictability. For regression models, Y-scrambling is a widely used validation strategy [95], and by definition, it measures the prediction errors and/or validation coefficients of artificial models generated with scrambled target values (y) [96]. However, in classification models, the absolute error value is always one due the categorical nature of the target value, and for that reason, X-scrambling can provide more diversity in the scrambled input data for model generation. Thus, X-scrambling validation was performed indicating that selected models were not generated by chance (Fig. 6) because artificial models generated with X-scrambled data failed in internal and/or external validations in comparison to original models.

Table 1 Method, fingerprint, cross-validation MCC ($_{CV}$MCC), external validation MCC ($_{EXT}$MCC), cross-validation AUC ($_{CV}$AUC), external validation AUC ($_{EXT}$AUC), cross-validation F1-score ($_{CV}$F1), and external validation F1-score ($_{CV}$F1) for the selected model for each endpoint

Endpoint	Method	Fingerprint	$_{CV}$MCC	$_{EXT}$MCC	$_{CV}$AUC	$_{EXT}$AUC	$_{CV}$F1	$_{EXT}$F1
Myristoyl binding	AdaBoost	PubChem	0.445	0.612	0.692	0.875	0.659	0.857
	Decision tree	AtomPairs2D	0.188	0.500	0.579	0.813	0.605	0.769
	Random forest	PubChem	0.422	0.500	0.692	0.813	0.718	0.769
Enzymatic activation	AdaBoost	AtomPairs2D	0.603	1.000	0.777	1.000	0.876	1.000
	Decision tree	AtomPairs2D	0.496	1.000	0.707	1.000	0.830	1.000
	Random forest	PubChem	0.616	1.000	0.760	1.000	8.667	1.000
Cellular activation	AdaBoost	AtomPairs2D	0.536	0.790	0.738	0.916	0.537	0.909
	Decision tree	PubChem	0.441	0.500	0.700	0.833	0.506	0.923
	Random forest	MACCS	0.622	1.000	1.000	0.788	0.693	1.000

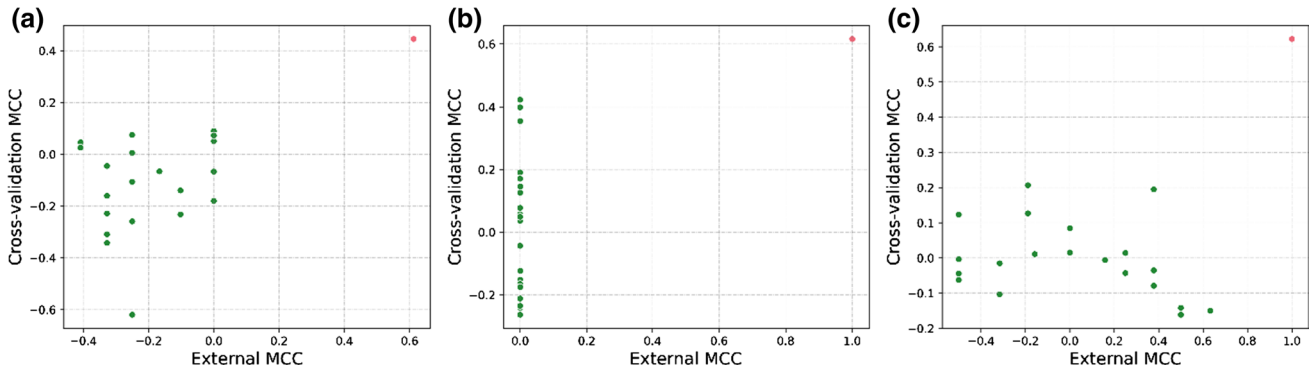


Fig. 6 Comparison between cross-validation MCC ($_{CV}$MCC) and external validation MCC ($_{EXT}$MCC) of the selected original model for each endpoint (pink) and 20 X-scrambled models (green): **a** myristoyl binding model; **b** enzymatic activation model; and **c** cellular activation model

After the validation process, the selected model for each endpoint had its applicability domain assessed by fitting a PCA into the training data and transforming the test data. The PCA bounding box approach for each endpoint is shown in Fig. 7 and it was found that the test data is within the training set applicability domain in the three analyzes. For the myristoyl binding and enzymatic activation data, the PCA constructed using PubChem (Fig. 7a, b), 3 PC's represent 60.5% of the total variance. Similarly, the PCA built using MACCS and cellular activation data, hold-out 62.1% of the total variance (Fig. 7c). Using the range and Euclidean, Manhattan, Cosine, and Wasserstein distances approaches for the applicability domain assessment, no test set compounds for the three models were considered out of the domain. Therefore, all the training and test sets splitting were suitable for the model validation.

Leonard and Roy [97] already discussed in their work methods for a rational selection of the training and test sets to obtain more predictive models. Despite this, in our work, the random selection achieved predictable models and a suitable distribution of the chemical space between the training and test sets as shown in the Hierarchical Clustering Analysis dendrograms presented in supplementary information Figure S2. These analyses may be helpful to characterize a molecular dataset, especially with multiple endpoint cases since the HCA indirectly measures the applicability domain. Also, it can be used to split data in training and test sets. Finally, the essential features of the selected models were interpreted using the permutation importance in the training data (Fig. 8). This process is defined as the decrease in a model score, in this case, MCC, in the training set when only one feature is randomly shuffled. The AdaBoost for MB model using PubChem fingerprint returned 10 features with importance different from zero, and the four most important were the positions 645, 333, 364, and 579, respectively. For the random forest for EA model, also using PubChem, the only feature important in the permutation process was the position 780. For the random forest model for CA using MACCS, 14 features were obtained differently from zero, and the four most important features were the positions 95, 160, 42, and 106, respectively. These features were selected to carry out the structure–activity relationship interpretation.

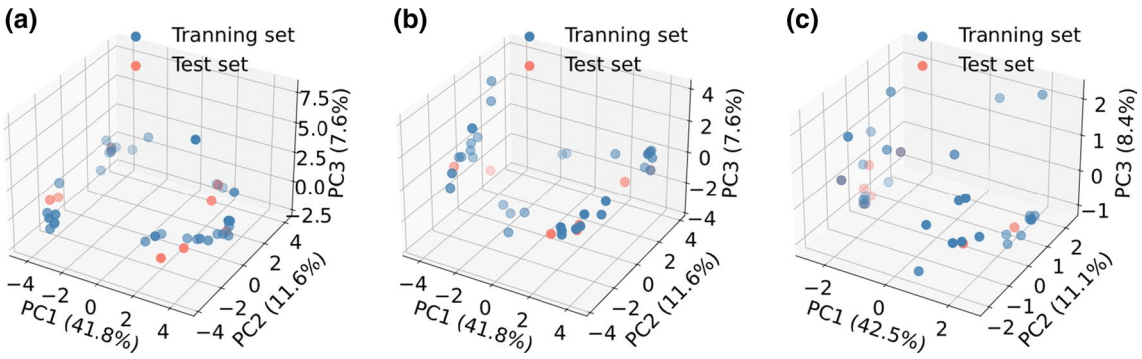


Fig. 7 The PCA bounding box approach for the applicability domain assessment for each endpoint: **a** PubChem fingerprints and myristoyl binding data; **b** PubChem fingerprints and enzymatic activation; and **c** MACCS fingerprints and cellular activation data. Each PC axis displays the contained variance

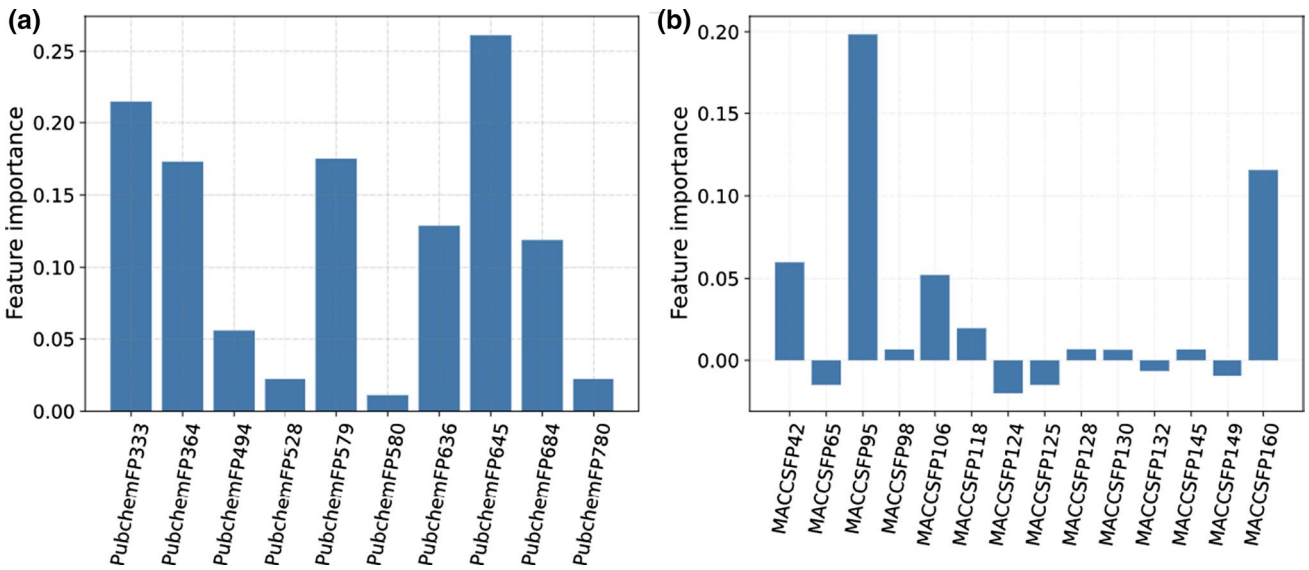


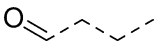
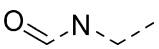
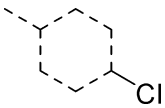
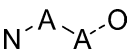
Fig. 8 Feature importance from permutation process: **a** AdaBoost for myristoyl binding model using PubChem fingerprint and **b** random forest model for cellular activation using MACCS fingerprint

Structure–activity relationship

The representation of each fingerprint key pattern and the frequency in the active/inactive compounds are shown in Table 2. With this data, it is possible to see that some fingerprint keys had substantial presence and distinct frequencies in the active and inactive compounds. Using this information, it is possible to infer structure–activity relationships for this dataset. For this analysis, features with an accumulated frequency higher than 30% were selected.

From the MB model, the fingerprint PubChemFP333 has a frequency two times higher in the active compounds. The majority position of this pattern is the methylated carbon in the pyrazoline ring. Also, it is possible to see in the crystallographic structure of compound 51 complexed with the c-Abl the Van der Waals interactions of this pattern (in this case, a carbon near to methylene) with LEU359 (Fig. 8a) and compound 47 with a methyl group from the pyrazoline ring interacting the same residue (Fig. 8b). PubChemFP579 was over six times higher in

Table 2 Visual interpretation for the fingerprints with permutation feature importance different from zero used in each endpoint

Endpoint	Fingerprint key	Presence in active compounds (%)	Presence in inactive compounds (%)	Visual interpretation
AdaBoost model for Myristoyl binding	PubChemFP333	47.916	22.916	
	PubChemFP364	8.333	6.250	
	PubChemFP579	29.166	4.166	
	PubChemFP645	37.500	16.666	
Random forest model for enzymatic activation	PubChemFP780	29.545	11.363	
Random forest model for cellular activation	MACCSFP42	10.416	4.166	F
	MACCSFP95	2.083	12.500	
	MACCSFP160	8.333	35.416	CH_3

Dashed bonds represent any bond, and A means any atom

the active compounds, showing the importance of the aliphatic bulk group for the myristoyl binding affinity model activity. This importance can be exemplified by the docking results, where compounds 4 and 32 occupy a hydrophobic site between the carbon chain in the GLU481 and TYR454 residues at the entrance of the myristoyl binding site (Fig. 8c, d). For PubChemFP645, the frequency in the active group is over two times higher than the occurrence in the inactive group. It is understandable why this pattern is important in the model because the presence of nitrogen between a carbonyl group and two carbons selects the nitrogen in the proper position to interact with ALA452 residue, acting as a hydrogen bond donor (Fig. 8e), an advance proposed by Huong et al. in 2014 [87]. This pattern also displayed the same behavior, even with different rings in both positions, as shown in compound 17 (Fig. 8f) and compound 19 (Fig. 8g), highlighting the importance of this interaction for compound recognition.

For the enzymatic activation model, the pattern in PubChemFP780 has a frequency two times higher in the active compounds when compared to the inactive compounds. The importance of this moiety is due to its optimal fitting to the hydrophobic pocket forming the Van der Waals interactions with residues LEU359, ALA452, TYR454, and PRO484 (Fig. 9h). Furthermore, the chlorine atom in *para* position may be used as a discriminant feature among all rings since its substitution to a polar group will disfavor these interactions [25]. The hydrophobic cavity involving this group can be seen in the compound 51 crystal structure and in the docking of compound 47 (Fig. 9a, b, respectively). Also, despite not being shown in the machine learning models, the chlorine atom in the meta-position can interact with LEU442 forming a halogen bond.

For the CA model, the generic methyl group recurring in the inactive compounds was the most frequent pattern. This result highlights the information from PubChemFP333, showing that only substitution in the pyrazoline ring is favorable. Also, a common position for this generic methyl group is along with the same carbonyl group of PubChemFP579, showing the importance of aliphatic bulk in this position. Finally, combining all information from ML models and docking simulations, a SAR model was reported and described in Fig. 10.

Finally, the generated models corroborated experimental binding modes and docking studies suggesting that the combination of LBDD and SBDD strategies could be employed in further drug design studies. It is well known in the literature that consensus predictions improved the predictability of QSAR models [91–94] and virtual screening protocols [98–100]. Therefore, despite the limitations of generated models (moderate predictability of MB model and the possibility of bias due the small dataset), a prediction of novel compounds followed by docking studies and following the

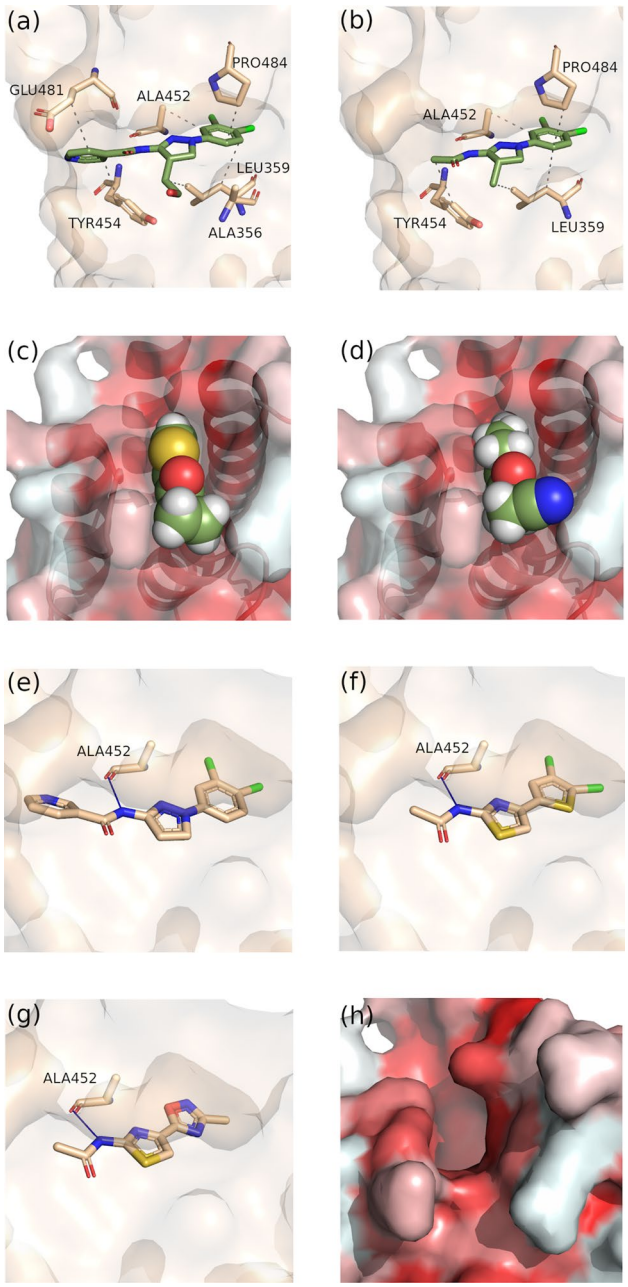


Fig. 9 Visual interpretation of the fingerprints using visual analysis: **a** crystal structure of compound 51 (PDB ID 6NPG), where Van Der Waals interactions are represented as dashed lines; **b** docking result of compound 47, where Van Der Waals interactions are represented as dashed lines; **c** docking result of compound 4, where the surface is colored by hydrophobicity; **d** docking result of compound 32, where the surface is colored by hydrophobicity; **e** docking result of compound 34, where hydrogen bond interaction is represented as a blue line; **f** docking result of compound 17, where hydrogen bond interaction is represented as a blue line; **g** docking result of compound 19, where hydrogen bond interaction is represented as a blue line; and **h** surface of myristoyl binding pocket colored by hydrophobicity

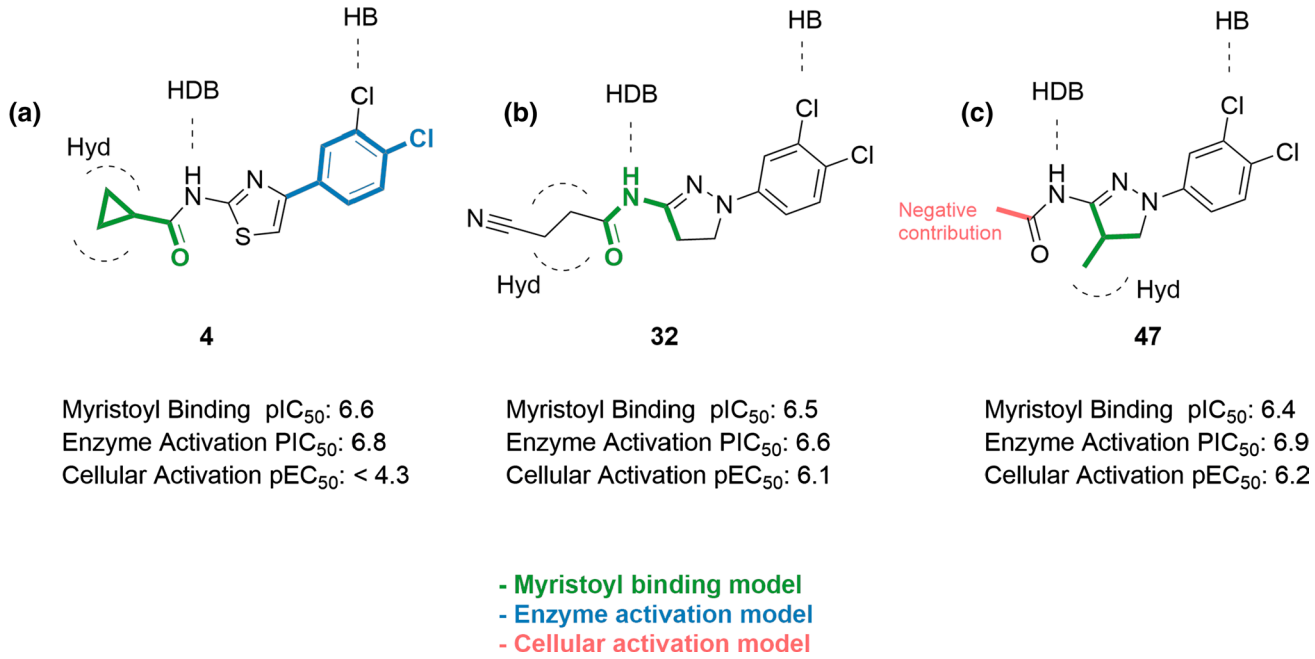


Fig. 10 SAR model for the activation of c-Abl kinase based on the molecular docking and machine learning models identified in three compounds. The patterns from de machine learning models are bold and colored. Interactions from the crystal structure and molecular docking are represented as dashed lines and display: hydrophobic contacts (Hyd), hydrogen bond donor (HBD), and halogen bond (HB)

proposed SAR by our work and the previous studies [22, 87] could be a useful to the drug design of c-Abl activators.

Conclusion

Using classification machine learning models allowed the construction of robust and predictive models for c-Abl activation, including myristoyl binding, enzyme activation, and cellular activation. For the prediction of myristoyl binding affinity, the AdaBoost algorithm using PubChem fingerprint achieved better MCC results for external and cross-validation. For enzyme activation, the random forest algorithm, and PubChem, had the best performance. Finally, for cellular activation, random forest obtained the highest MCC value using the MACCS fingerprint. It is important to mention that the using of a small dataset to train and validate the models could provide bias to the generated models, limiting the model's application and extrapolation in the SAR study. However, the combination of molecular docking with molecular fingerprints interpretation from the machine learning models corroborated SARs described by Simpson et al. [25] and provided new insights into this structure–activity relationship. This work may assist the next steps forward in identifying and designing more potent novel kinase activators.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-021-10261-z.

Acknowledgements The authors would like to thank CNPq, CAPES, and FAPEMIG for financial support.

Author contributions P.O.F. and A.S.B. carried out and interpreted the machine learning models; D.M.M. carried out and interpreted the molecular docking; POF, J.P.A.M., A.H.M., and V.G.M. designed the experiments; P.O.F., D.M.M., A.S.B., J.P.A.M., A.H.M., and V.G.M. wrote and revised the manuscript.

References

1. Hantschel O, Superti-Furga G (2004) Regulation of the c-Abl and Bcr-Abl tyrosine kinases. Nat Rev Mol Cell Biol 5:33–44. https://doi.org/10.1038/nrm1280
2. Woodring PJ (2003) Regulation of F-actin-dependent processes by the Abl family of tyrosine kinases. J Cell Sci 116:2613–2626. https://doi.org/10.1242/jcs.00622
3. Huang Y, Comiskey EO, Dupree RS et al (2008) The c-Abl tyrosine kinase regulates actin remodeling at the immune synapse. Blood 112:111–119. https://doi.org/10.1182/blood-2007-10-118232
4. Aoyama K, Fukumoto Y, Ishibashi K et al (2011) Nuclear c-Abl-mediated tyrosine phosphorylation induces chromatin structural changes through histone modifications that include H4K16 hypoacetylation. Exp Cell Res 317:2874–2903. https://doi.org/10.1016/j.yexcr.2011.09.013
5. Kharbanda S, Yuan Z-M, Weichselbaum R, Kufe D (1998) Determination of cell fate by c-Abl activation in the response to DNA damage. Oncogene 17:3309–3318. https://doi.org/10.1038/sj.onc.1202571

6. Van Etten RA (1999) Cycling, stressed-out and nervous: cellular functions of c-Abl. Trends Cell Biol 9:179–186. https://doi.org/10.1016/S0962-8924(99)01549-4
7. Wang JYJ (2014) The capable ABL: what is its biological function? Mol Cell Biol 34:1188–1197. https://doi.org/10.1128/MCB.01454-13
8. Sawyers CL (1999) Chronic myeloid leukemia. N Engl J Med 340:1330–1340. https://doi.org/10.1056/NEJM199904293401706
9. Blume-Jensen P, Hunter T (2001) Oncogenic kinase signalling. Nature 411:355–365. https://doi.org/10.1038/35077225
10. Greuber EK, Smith-Pearson P, Wang J, Pendergast AM (2013) Role of ABL family kinases in cancer: from leukaemia to solid tumours. Nat Rev Cancer 13:559–571. https://doi.org/10.1038/nrc3563
11. Brahmachari S, Karuppagounder SS, Ge P et al (2017) c-Abl and Parkinson's disease: mechanisms and therapeutic potential. JPD 7:589–601. https://doi.org/10.3233/JPD-171191
12. Caracciolo D, Valtieri M, Venturelli D et al (1989) Lineage-specific requirement of c-abl function in normal hematopoiesis. Science 245:1107–1110. https://doi.org/10.1126/science.2672339
13. Rosti V, Bergamaschi G, Lucotti C et al (1995) Oligodeoxynucleotides antisense to c-abl specifically inhibit entry into S-phase of CD34+ hematopoietic cells and their differentiation to granulocyte-macrophage progenitors. Blood 86:3387–3393. https://doi.org/10.1182/blood.V86.9.3387.bloodjournal8693387
14. Cowan-Jacob SW, Jahnke W, Knapp S (2014) Novel approaches for targeting kinases: allosteric inhibition, allosteric activation and pseudokinases. Future Med Chem 6:541–561. https://doi.org/10.4155/fmc.13.216
15. Allington TM, Galliher-Beckley AJ, Schiemann WP (2009) Activated Abl kinase inhibits oncogenic transforming growth factor-β signaling and tumorigenesis in mammary tumors. FASEB J 23:4231–4243. https://doi.org/10.1096/fj.09-138412
16. Varzavand A, Hacker W, Ma D et al (2016) α3β1 integrin suppresses prostate cancer metastasis via regulation of the hippo pathway. Cancer Res 76:6577–6587. https://doi.org/10.1158/0008-5472.CAN-16-1483
17. Cabigas EB, Liu J, Boopathy AV et al (2015) Dysregulation of catalase activity in newborn myocytes during hypoxia is mediated by c-Abl tyrosine kinase. J Cardiovasc Pharmacol Ther 20:93–103. https://doi.org/10.1177/1074248414533746
18. Dasgupta Y, Koptyra M, Hoser G et al (2016) Normal ABL1 is a tumor suppressor and therapeutic target in human and mouse leukemias expressing oncogenic ABL1 kinases. Blood 127:2131–2143. https://doi.org/10.1182/blood-2015-11-681171
19. Nagar B, Hantschel O, Seeliger M et al (2006) Organization of the SH3–SH2 unit in active and inactive forms of the c-Abl tyrosine kinase. Mol Cell 21:787–798. https://doi.org/10.1016/j.molcel.2006.01.035
20. Nagar B, Hantschel O, Young MA et al (2003) Structural basis for the Autoinhibition of c-Abl tyrosine kinase. Cell 112:859–871
21. Yang J, Campobasso N, Biju MP et al (2011) Discovery and characterization of a cell-permeable, small-molecule c-Abl kinase activator that binds to the myristoyl binding site. Chem Biol 18:177–186. https://doi.org/10.1016/j.chembiol.2010.12.013
22. Simpson GL, Bertrand SM, Borthwick JA et al (2019) Identification and optimization of novel small c-Abl kinase activators using fragment and HTS methodologies. J Med Chem 62:2154–2171. https://doi.org/10.1021/acs.jmedchem.8b01872
23. Jahnke W, Grotzfeld RM, Pellé X et al (2010) Binding or bending: distinction of allosteric Abl kinase agonists from antagonists by an NMR-based conformational assay. J Am Chem Soc 132:7043–7048. https://doi.org/10.1021/ja101837n
24. Laufkötter O, Hu H, Miljković F, Bajorath J (2021) Structure- and similarity-based survey of allosteric kinase inhibitors, activators, and closely related compounds. J Med Chem. https://doi.org/10.1021/acs.jmedchem.0c02076
25. Zorn JA, Wells JA (2010) Turning enzymes ON with small molecules. Nat Chem Biol 6:179–188. https://doi.org/10.1038/nchembio.318
26. Mullard A (2016) Biotech R&D spend jumps by more than 15%. Nat Rev Drug Discov 15:447–447. https://doi.org/10.1038/nrd.2016.135
27. Schaduangrat N, Lampa S, Simeon S et al (2020) Towards reproducible computational drug discovery. J Cheminform 12:9. https://doi.org/10.1186/s13321-020-0408-x
28. Andricopulo A, Salum L, Abraham D (2009) Structure-based drug design strategies in medicinal chemistry. CTMC 9:771–790. https://doi.org/10.2174/156802609789207127
29. Chaudhary KK, Mishra N (2016) A review on molecular docking: novel tool for drug discovery. JSM Chem 3:1029
30. Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational methods in drug discovery. Pharmacol Rev 66:334–395. https://doi.org/10.1124/pr.112.007336
31. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967. https://doi.org/10.1021/ja00226a005
32. Clark M, Cramer RD, Jones DM et al (1990) Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. Tetrahed Comput Methodol 3:47–59. https://doi.org/10.1016/0898-5529(90)90120-W
33. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146. https://doi.org/10.1021/jm00050a010
34. Lino CI, Gonçalves de Souza I, Borelli BM et al (2018) Synthesis, molecular modeling studies and evaluation of antifungal activity of a novel series of thiazole derivatives. Eur J Med Chem 151:248–260. https://doi.org/10.1016/j.ejmech.2018.03.083
35. Tong W, Lowis DR, Perkins R et al (1998) Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J Chem Inf Comput Sci 38:669–677. https://doi.org/10.1021/ci980008g
36. Heritage TW, Lowis, David R. (1999) Molecular hologram QSAR. In: Parrill AL, Reddy MR (eds) Rational drug design: novel methodology and practical applications. American Chemical Society
37. Kronenberger T, Asse LR, Wrenger C et al (2017) Studies of *Staphylococcus aureus* FabI inhibitors: fragment-based approach based on holographic structure–activity relationship analyses. Future Med Chem 9:135–151. https://doi.org/10.4155/fmc-2016-0179
38. García-Jacas CR, Marrero-Ponce Y, Acevedo-Martínez L et al (2014) QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps. J Comput Chem 35:1395–1409. https://doi.org/10.1002/jcc.23640
39. García-Jacas CR, Marrero-Ponce Y, Vivas-Reyes R et al (2020) Distributed and multicore QuBiLS-MIDAS software v2.0: computing chiral, fuzzy, weighted and truncated geometrical molecular descriptors based on tensor algebra. J Comput Chem 41:1209–1227. https://doi.org/10.1002/jcc.26167
40. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley, London, pp 366–510
41. Consonni V, Todeschini R, Pavan M (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42:682–692. https://doi.org/10.1021/ci015504a

42. Castro Gertrudes J, Maltarollo V, Silva RA et al (2012) Machine learning techniques and drug design. Curr Med Chem 19:4289–4297. https://doi.org/10.2174/092986712802884259
43. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20:318–331. https://doi.org/10.1016/j.drudis.2014.10.012
44. Lima AN, Philot EA, Trossini GHG et al (2016) Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov 11:225–239. https://doi.org/10.1517/17460441.2016.1146250
45. Schneider G (2019) Mind and machine in drug design. Nat Mach Intell 1:128–130. https://doi.org/10.1038/s42256-019-0030-7
46. Göller AH, Kuhnke L, Montanari F et al (2020) Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. Drug Discov Today 25:1702–1709. https://doi.org/10.1016/j.drudis.2020.07.001
47. Serafim MSM, dos Júnior VS, S, Gertrudes JC, et al (2021) Machine learning techniques applied to the drug design and discovery of new antivirals: a brief look over the past decade. Expert Opin Drug Discov. https://doi.org/10.1080/17460441.2021.1918098
48. Serafim MSM, Gertrudes JC, Costa DMA et al (2021) Knowing and combating the enemy: a brief review on SARS-CoV-2 and computational approaches applied to the discovery of drug candidates. Biosci Rep. https://doi.org/10.1042/BSR20202616
49. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G (2021) Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. https://doi.org/10.1080/17460441.2021.1909567
50. Huang DZ, Baber JC, Bahmanyar SS (2021) The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction. Expert Opin Drug Discov. https://doi.org/10.1080/17460441.2021.1901685
51. Breiman L, Friedman JH, Olshen RA (1984) Classification and regression tree. CRC Press
52. Neves BJ, Dantas RF, Senger MR et al (2016) Discovery of new anti-schistosomal hits by integration of QSAR-based virtual screening and high content screening. J Med Chem 59:7075–7088. https://doi.org/10.1021/acs.jmedchem.5b02038
53. Veríssimo GC, Menezes Dutra EF, Teotonio Dias AL et al (2019) HQSAR and random forest-based QSAR models for anti-*T. vaginalis* activities of nitroimidazoles derivatives. J Mol Graph Model 90:180–191. https://doi.org/10.1016/j.jmgm.2019.04.007
54. Wolff D, Neugebauer U (2019) Tree-based machine learning approaches for equity market predictions. J Asset Manag 20:273–288. https://doi.org/10.1057/s41260-019-00125-5
55. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1
56. Tin Kam Ho (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844. https://doi.org/10.1109/34.709601
57. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139. https://doi.org/10.1006/jcss.1997.1504
58. Murata T, Yanagisawa T, Kurihara T et al (2019) Salivary metabolomics with alternative decision tree-based machine learning methods for breast cancer discrimination. Breast Cancer Res Treat 177:591–601. https://doi.org/10.1007/s10549-019-05330-9
59. Suresh A, Udendhran R, Balamurgan M (2020) Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. Soft Comput 24:7947–7953. https://doi.org/10.1007/s00500-019-04066-4
60. Polishchuk PG, Muratov EN, Artemenko AG et al (2009) Application of random forest approach to QSAR prediction of aquatic toxicity. J Chem Inf Model 49:2481–2488. https://doi.org/10.1021/ci900203n
61. García-Jacas CR, Marrero-Ponce Y, Cortés-Guzmán F et al (2019) Enhancing acute oral toxicity predictions by using consensus modeling and algebraic form-based 0D-to-2D molecular encodes. Chem Res Toxicol 32:1178–1192. https://doi.org/10.1021/acs.chemrestox.9b00011
62. Mora JR, Marrero-Ponce Y, García-Jacas CR, Suarez Causado A (2020) Ensemble models based on QuBiLS-MAS features and shallow learning for the prediction of drug-induced liver toxicity: improving deep learning and traditional approaches. Chem Res Toxicol 33:1855–1873. https://doi.org/10.1021/acs.chemrestox.0c00030
63. Ampomah EK, Qin Z, Nyame G (2020) Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information 11:332. https://doi.org/10.3390/info11060332
64. Hanwell MD, Curtis DE, Lonie DC et al (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4:17. https://doi.org/10.1186/1758-2946-4-17
65. O'Boyle NM, Banck M, James CA et al (2011) Open Babel: An open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
66. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17:490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3c490::AID-JCC1%3e3.0.CO;2-P
67. Halgren TA (1996) Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 17:520–552. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3c520::AID-JCC2%3e3.0.CO;2-W
68. Halgren TA (1996) Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem 17:553–586. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3c553::AID-JCC3%3e3.0.CO;2-T
69. Halgren TA, Nachbar RB (1996) Merck molecular force field. IV. Conformational energies and geometries for MMFF94. J Comput Chem 17:587–615. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3c587::AID-JCC4%3e3.0.CO;2-Q
70. Halgren TA (1996) Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem 17:616–641. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3c616::AID-JCC5%3e3.0.CO;2-X
71. Discovery Studio Visualizer (2020) BIOVIA, Dassault Systèmes, San Diego
72. Jones G et al (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:172–748
73. Korb O, Stützle T, Exner TE (2009) Empirical scoring functions for advanced protein−ligand docking with PLANTS. J Chem Inf Model 49:84–96. https://doi.org/10.1021/ci800298z
74. Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8
75. Salentin S, Schreiber S, Haupt VJ et al (2015) PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res 43:W443–W447. https://doi.org/10.1093/nar/gkv315
76. Yap CW (2011) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/jcc.21707
77. Maltarollo VG (2019) Classification of *Staphylococcus aureus* FabI inhibitors by machine learning techniques. Int J Quant Struct Property Relationships 4:1–14. https://doi.org/10.4018/IJQSPR.2019100101

78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–30
79. Andrada MF, Vega-Hissi EG, Estrada MR, Garro Martinez JC (2017) Impact assessment of the rational selection of training and test sets on the predictive ability of QSAR models. SAR QSAR Environ Res 28:1011–1023. https://doi.org/10.1080/1062936X.2017.1397056
80. Lipiński PFJ, Szurmak P (2017) SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. Chem Pap 71:2217–2232. https://doi.org/10.1007/s11696-017-0215-7
81. Sahigara F, Mansouri K, Ballabio D et al (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17:4791–4810. https://doi.org/10.3390/molecules17054791
82. Roy K, Kar S, Das RN (2015) Validation of QSAR Models. In: Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Elsevier, pp 231–289
83. Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2
84. García-Jacas CR, Martinez-Mayorga K, Marrero-Ponce Y, Medina-Franco JL (2017) Conformation-dependent QSAR approach for the prediction of inhibitory activity of bromodomain modulators. SAR QSAR Environ Res 28:41–58. https://doi.org/10.1080/1062936X.2017.1278616
85. Ehrt C, Krause B, Schmidt R, et al (2020) SMARTS.plus—a toolbox for chemical pattern design. Mol Inform 39:2000216. https://doi.org/10.1002/minf.202000216
86. Hevener KE, Zhao W, Ball DM et al (2009) Validation of molecular docking programs for virtual screening against dihydropteroate synthase. J Chem Inf Model 49:444–460. https://doi.org/10.1021/ci800293n
87. Hong X, Cao P, Washio Y et al (2014) Structure-guided optimization of small molecule c-Abl activators. J Comput Aided Mol Des 28:75–87. https://doi.org/10.1007/s10822-014-9731-5
88. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6. https://doi.org/10.1186/s12864-019-6413-7
89. Chicco D, Tötsch N, Jurman G (2021) The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 14:13. https://doi.org/10.1186/s13040-021-00244-z
90. Mehta P, Bukov M, Wang C-H et al (2019) A high-bias, low-variance introduction to machine learning for physicists. Phys Rep 810:1–124. https://doi.org/10.1016/j.physrep.2019.03.001
91. Asikainen AH, Ruuskanen J, Tuppurainen KA (2004) Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. SAR QSAR Environ Res 15:19–32. https://doi.org/10.1080/1062936032000169642
92. Kuz'min VE, Muratov EN, Artemenko AG, et al (2009) Consensus QSAR modeling of phosphor-containing chiral AChE inhibitors. QSAR Comb Sci 28:664–677. https://doi.org/10.1002/qsar.200860117
93. Alves VM, Muratov E, Fourches D et al (2015) Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. Toxicol Appl Pharmacol 284:262–272. https://doi.org/10.1016/j.taap.2014.12.014
94. Alves VM, Muratov E, Fourches D et al (2015) Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. Toxicol Appl Pharmacol 284:273–280. https://doi.org/10.1016/j.taap.2014.12.013
95. Kiralj R, Ferreira MMC (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J Braz Chem Soc 20:770–787. https://doi.org/10.1590/S0103-50532009000400021
96. Rücker C, Rücker G, Meringer M (2007) y-Randomization and Its Variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357. https://doi.org/10.1021/ci700157b
97. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25:235–251. https://doi.org/10.1002/qsar.200510161
98. Kukol A (2011) Consensus virtual screening approaches to predict protein ligands. Eur J Med Chem 46:4661–4664. https://doi.org/10.1016/j.ejmech.2011.05.026
99. Yang J-M, Chen Y-F, Shen T-W et al (2005) Consensus scoring criteria for improving enrichment in virtual screening. J Chem Inf Model 45:1134–1146. https://doi.org/10.1021/ci050034w
100. Houston DR, Walkinshaw MD (2013) Consensus docking: improving the reliability of docking in a virtual screening context. J Chem Inf Model 53:384–390. https://doi.org/10.1021/ci300399w