Use of QSAR global models and molecular docking for developing new inhibitors of c-src tyrosine kinase

Robert Ancuceanu[1], Bogdan Tamba[2], Cristina Silvia Stoicescu[3], Mihaela Dinu[1]*
[1]Faculty of Pharmacy, University of Medicine and Pharmacy "Carol Davila", Bucharest, Romania
[2]Advanced Research and Development Center for Experimental Medicine (CEMEX), Grigore T. Popa University of Medicine and Pharmacy of Iasi, Romania (bogdan.tamba@umfiasi.ro);
[3]Institute of Physical Chemistry "Ilie Murgulescu", Bucharest, Romania

Abstract

Prototype of a family of at least nine members, c-src tyrosine kinase is a therapeutically interesting target, because its inhibition might be of interest not only in a number of malignancies, but also in a diverse array of conditions, from neurodegenerative pathologies to certain viral infections. Computational methods in drug discovery are considerably cheaper than conventional methods and offer opportunities of screening very large numbers of compounds in conditions that would be simply impossible within the wet lab experimental settings. We have explored the use of global QSAR models and molecular ligand docking in the discovery of new c-src tyrosine kinase inhibitors. Using a data set of 1038 compounds from ChEMBL and 19 blocks of molecular descriptors, we have developed over 200 QSAR classification models, based on six machine learning algorithms and 17 feature selection methods. We have selected 49 with reasonably good performance (positive predictive value and balanced accuracy higher than 70% in nested cross validation) and the models were assembled by stacking with a simple majority vote and used for the virtual screening of over the "named" ZINC data set (over 100,000 compounds). 744 compounds were predicted by at least 50% of the QSAR models as active, 147 compounds were within the applicability domain and predicted by at least 75% of the models to be active. The latter 147 compounds were submitted to molecular ligand docking using Vina and Ledock, and a number of 90 were predicted to be active based on the binding energy. External data from CHEMBL and PUBCHEM confirmed that at least 7.83% (in the case of QSAR) or 6.67% (in the case of integrated QSAR and molecular docking) of the compounds are active on the c-src target.

Keywords: QSAR; c-src; tyrosine kinase; docking; virtual screening

Introduction

Src (c-src, pp60-src, or p60-src) is a non-receptor, cytoplasmic tyrosine kinase, the first of its kind to be discovered (in the 1970s) in the living world, whereas the corresponding gene has been the first oncogene to be uncovered (1). It is the prototype of a larger family, comprising at least nine members, most of them with little activity in normal cells in the absence of stimulatory signals (2). Src kinases have been suggested to be involved in the exacerbation of neurodegenerative pathologies, whereas their inhibition would diminish microgliosis and mitigate inflammation, findings that are in line with experimental effects seen for non-specific src inhibitors such as bosutinib or LCB-03-0110 (3). Inhibition of src kinases has been suggested by non-clinical evidence as a potential method of therapy for the pulmonary vascular remodeling and right ventricular hypertrophy in pulmonary hypertension (4), although several reports indicate that dual Abl/src inhibitor dasatinib may actually induce pulmonary hypertension (5–7); it was more recently suggested that this dasatinib effect may in fact be independent of the src inhibition (7). This family of kinases has been recently shown to be involved in the subgenomic RNA translation and replication of alpha-viruses, their inhibition being suggested as a potentially effective way of treating infections with such viral particles (8). Thus, targeting src kinases opens wide doors for multiple therapeutic applications in a variety of pathologies and there is a constant interest for understanding the pharmacology of this class of compounds, as well as for developing new src inhibitors.

The first member of this family (c-src), has been suggested to be more important than other members of the same family in certain pathologies or clinical contexts. For instance, c-src, but not Lyn and Fyn src kinases from the same family, is up-regulated by hypoxia and plays a major role in prostate cancer metastasis of hypoxic tumours (hypoxia is a negative prognostic factor in this malignancy) (9). Besides, c-src tyrosin kinase has been shown to be abnormally activated or over-expressed in a number of different malignancies and to stimulate processes associated with tumour progression, such as proliferation, angiogenesis or metastasis (10). Src tyrosin kinase inhibitors have been explored as potential new therapies in a variety of malignancies such as melanoma (one such inhibitor showing *in vitro* that is active on a variety of melanoma cells, including some BRAF^{V600} mutant cells (11), but a report that src inhibition would increase induces melanogenesis in melanoma cells has also been published (12)), papillary thyroid carcinoma (13), clear-cell renal carcinoma (14), pancreatic (15) or ovarian cancer (16).

The space of the universe is expanding, but so is the "chemical space". Currently Pubchem includes about 96 million different chemical compounds (17), an impressive number, but minuscule when compared with the number of chemical compounds that might be synthesized in the coming years. GDB-17, probably the largest database of molecules up to date, included in 2015 no less than 166 billion compounds, and these are limited to only a few types of atoms (C, N, O, S, and halogens) and maximum 17 atoms per molecule (18). Theoretical calculations using constraints for circumscribing the drug-like chemical space have suggested that the number of molecules obeying to the Lipinsky's rules is about 10^{33} (19), an estimate intermediary between 10^{60} (as proposed earlier by R.S. Bohacek et al. (20)) and 10^{23} (as proposed later by P. Ertl. (21)). How to assess all these substances for their pharmacological, toxicological or biological effects (in all contexts, for all targets etc)? It is simply „mission: impossible" by the traditional route of wet lab experiments. Here the computing power of our age finds its place with surprisingly good results (although far from perfect).

Built on three pillars (biological data, chemical knowledge and modeling algorithms), QSAR (*quantitative structure-activity relationship*) (22) methodologies allow the development of computational tools for predicting with reasonable confidence (when validated appropriately) a wide variety of biological activities from the molecular structure of chemical compounds. Although the QSAR approaches have not gained in popularity as fast as the molecular docking modeling, the field has been far from being inert in the last decade or so, with various new approaches with respect to the mathematical algorithms used or the with respect to the biological activities explored (23). The models developed and validated may then be applied for virtual screening purposes to a large number of compounds, allowing quick identification of a sizeable number of compounds of interest (with certain activities or biological properties). Such virtual screening exercises may also be further coupled with other computational methods, such as ligand-target docking for confirmation of activity (24,25). Whereas the classical drug development process was very costly and tedious, computational methods have high efficiency and are inexpensive (26). In this context we developed a set of QSAR models with different descriptors and machine learning classification algorithms, integrated by stacking, to be used for virtual screening purposes of c-src tyrosin kinase inhibitors. A number of 49 QSAR models with reasonably good performance have been developed and their performance assessed by nested cross-validation. They were applied for the virtual screening of over 100,000 chemical compounds from the ZINC database, and 147 with the highest probability of being active were also assessed with molecular docking, for 90 of them the docking data being consistent with a hypothesis of activity. Data from CHEMBL and PUBCHEM externally validated the virtual screening results for a number of compounds.

Materials and methods

Dataset

The dataset was downloaded from CHEMBL (https://www.ebi.ac.uk/chembl) and included experimental data for c-src as a target (target code CHEMBL267). Only the records with K_i values expressed in nM were kept. Records with “=” values in the field “Relation” were kept for analysis and labeled as “active” if ki < 1000 nM and “inactive” if ki ≥ 1000 nM; records with “>” or “<” values in the field “Relation”were kept for analysis only if they allowed unequivocal classification (e.g. records with ki > 5000 nM were kept and labeled as “inactive”, whereas those with ki > 100 nM were discarded; similarly, records with ki < 5000 nM were discarded). A threshold of 1000 nm for the formal discrimination between “active” and “inactive” compounds is usual in the field and has been used in other publications (27). We have used classification rather than regression, because the data come from different laboratories and experimental settings, and although ki values have less variability than IC50, published experimental ki values still vary considerably (of the 75 compounds in our data set with multiple ki values, the relative standard deviation (RSD) of ki varied from 0% to 103%; for the first three quartiles, RSD was relatively low, under 13.85%, but for the last quartile it was quite high). Inorganic compounds were removed. For the detection and removal of duplicate compounds we proceeded in two steps: First, canonical SMILES (available in the downloaded dataset) were searched for duplicates in R (v. 3.6.0) and their ki values were replaced by the average of the duplicates. We then used ChemAxon Standardizer v. 18.8.0 (ChemAxon, Budapest, Hungary) for the standardization of the molecules, and then employed the ISIDA/Duplicates software (http://infochim.u-strasbg.fr; University of Strasbourg, France) software for the identification of potential further duplicates. We used Discovery Studio Visualizer v16.1.0.15350 (Dassault Systèmes BIOVIA, San Diego, CA, USA) to convert the standardized SMILES to 2D chemical structures (sdf). Following the removal of duplication, our dataset decreased from an initial number of 1151 compounds to 1038, of which 286 were labeled as “active” and 752 as “inactive”.

Descriptors

Molecular descriptors of the dataset molecules were computed using the Dragon 7 software (version 7.0, https://chm.kode-solutions.net; Kode SRL, Milano, Italy). 19 blocks of molecular descriptors were computed: constitutional descriptors (n=47), ring descriptors (n=32), topological indices (n=75), walk and path counts (n=46), connectivity indices (n=37), information indices (n=50), 2D matrix-based descriptors (n=607), 2D-autocorrelations (n=213), Burden eigenvalues (n=96), P-VSA-like descriptors (n=55), ETA indices (n=23), edge adjacency indices (n=324), functional groups count (153), atom-centred fragments (n=115), atom-type E-state indices (n=172), CATS 2D (n=150), 2D atom pairs (n=1596), molecular properties (n=20), and drug-like indices (n=28). All descriptors thus computed were 3839.

Feature selection

Because the number of computed descriptors is very large (almost 4000), the “dimensionality curse” precludes optimal operation of the classification or regression algorithms, which are generally designed for a relatively small number of variables, and tends to result in overfitting (28). Feature selection, which is a process of filtering a high number of variables while keeping only the most relevant of them increases the performance of machine learning algorithms, reduces the computational costs and strengthens the generalization ability of the models built (28). Multiple algorithms of feature selection have been proposed in the literature, with variable performance, often depending on the nature and particularities of the data. We have used 17 different feature selection algorithms, implemented directly in the “mlr” R package (29) or through other R packages: based on an ANOVA test, on a Kruskal test, on the Area Under the Curve (AUC), variance, and an univariate model performance score (‘mlr’), based on a permutation importance of random forest (as implemented in the R package ‘party’, (30)), based on a chi-square test, gain ratio, information gain, OneR classifier, RELIEF algorithm, and symmetrical uncertainty (methods implemented in the ‘FSelector’ R package (31)), three algorithms based on random forest importance (as implemented in the randomForest (32) and randomForestSRC (33) packages), and two algorithms based on node impurity and permutation in random forests, as implemented in the ‘ranger’ R package (34). The feature selection algorithms were applied after pre-processing consisting in removal of constant and quasi-constant features (i.e. those where less than 1% of the observations differed from the mode value) and highly correlated features (defined as those with a correlation coefficient higher than 0.9).

Machine learning algorithms and model building

For building the models we have used the following algorithms: random forests, support vector machines, ada Boosting M1, Bayesian additive regression trees, binomial regression, and C5.0 decision trees and rule-based models.

Based on an arbitrary number of decision trees used as an ensemble with a majority vote to decide on the most probable class assigned to each data point, random forests (RF) are a popular classification algorithm often used with very good performance in QSAR models (35–37). Each decision tree is constructed using bootstrap sets of the training set and subsets of descriptors that are selected in a random manner(38).

The support vector machines (SVM) algorithm is able to address data sets with high number of variables and has often been used with very good performance in a variety of classification and regression tasks, including QSAR applications (39,40). It uses a variety of kernel functions (e.g. linear, polynomial, radial etc) to project features in a vector space maximizing the partitioning boundary between classes and to identify the hyperplane that best discriminates the classes (41).

The adaboost M1 (Adaptive Boosting) algorithms were described as "widely used in QSAR studies" (42), although they are probably less used than RF or SVM. AdaBoost is an iterative algorithm that uses weights to improve the performance of "weak" classifiers (particularly decision tress), giving higher weights to the trees with better performance (smaller misclassification rates) (42).

Bayesian Additive Regression Trees (BART) is non-linear regression technique based on a Bayesian approach, whose performance in QSAR modelling has been stated to be competitive with that of other machine learning methods (43). Unlike other decision trees, where decision is taken based on a majority vote or with the help of empirical weights, BART makes use of prior knowledge and likelihood to improve the performance of the decision trees.

Binomial regression (logistic regression), despite the term „regression" is a relatively simple algorithm used for classification purposes, because it linearly models the probability that an observation belongs to one of two categorical outcomes (44). In other words, logistic regression computes the probability $P=1/(1+e^{-t})$, where $t= a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n$ (45).

C5.0 decision trees and rule-based models represent an extension of a classification algorithm proposed by R. Quinlan in 1993, under the name "C4.5" and builds models that can take either the form of a decision tree or a set of rules (in simple or boosted versions) (46). Although apparently less used in QSAR modeling than other machine learning algorithms, when employed, it gave excellent performance, comparable with that of random forests or support vector machines (47).

All models were built and their performance was assessed in the computing and programming environment R, v. 3.6.0 (48), using 'mlr' package (29) coupled with "parallelMap" (49) for parallel computing, and to a small extent, the "caret" package (50). Classification algorithms were used from the corresponding R packages implementing them: 'randomForest' (32), 'e1071' (51) (for SVM), 'RWeka' (52,53) (for adaboost M1), 'bartMachine' (54) (for BART), 'stats'(48) (for the logistic regression), and 'C50' (for the C5.0 algorithm) (46). Gower distances were computed with the "cluster" R package (55). Graphs were built in "ggplot2" (56) and (for the dissimilarity plot) "seriation" (57). All values were standardized by centering and scaling, and values larger than two standard deviations were capped to 2.

Performance evaluation

Nested cross-validation using 5 folds in the inner loop and 10 folds in the outer loop was used to evaluate the performance of the models selected, except for the Bayesian Additive Regression Trees, for which 5 folds were also used in the external loop (due to the long time taken by this classifier). The assessment of QSAR model performance should include both internal and external, and the external validation is generally deemed as "the gold standard" (58,59). However, the concept of "external validation" has received different interpretations and most often is assumed to describe a holdout data set, obtained by an initial one-time split (i.e. a set that has not been seen by the model during any adjustments or hyperparameter optimization) (60). Despite its apparent advantages of objectivity and ability to assess the generalization of the selected model(s), the use of a hold-out data set is fraught with thorny issues: the split may be simply fortunate leading to overestimation of performance (or of contrary, it may be unfortunate, leading to underestimation of performance), it requires the holdout sample to be large (which in practice may be costly or a requirement impossible to satisfy), and the sample size needed for holdout is larger than it is necessary for cross-validation to estimate the prediction error with a similar degree of precision (58). For these reasons, using nested cross-validation (also known as double cross-validation) not only does not reject the idea of external validation, but it extends it to the entire data set (61).

All models were assessed by computing (within the nested cross-validation) the balanced accuracy (BA), mean misclassification error (MMCE), sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), area under the Receiver Operating Characteristics curve (AUC) and positive predictive value (PPV), with their widely known definitions and equations (27,62). Particularly for virtual screening purposes PPV is important (because it indicates the likely proportion of positive values among the values predicted as positive). We therefore selected only models with a PPV higher than 70% and BA higher than 70%.

To make sure that the performance of the models is not consequential to chance, a Y-scrambling procedure was applied, where for multiple models the dependent variable (in our case the ki values) was shuffled through 1000 permutations (using the R package ‘gtools’ (63)), then the models were rebuilt using the same procedure from the first steps (i.e. applying the same feature selection algorithms, in the same order) and their performance evaluated. If there is a real relationship between the activity and the descriptors, following the y randomization the performance of the new models thus built should be worse.

Y-randomization test. To ascertain that the models are not the result of chance association we applied a classical y-scrambling test (64) by permuting the activity label of the compounds from the data set and re-building the models following the same steps as for the construction of the „true“ models. This process was repeated ten times and the new models were evaluated for their performance in terms of balanced accuracy, sensitivity, specificity and positive predictive value, the expectation being that the performance would be (considerably) worse than that of the QSAR models based on the initial data set.

Applicability domain

We have used two local density-based outlier methods implemented in the DDoutlier R package (65) - the Kernel Density Estimation Outlier Score (KDEOS) algorithm with gaussian kernel (66), and the INFLO algorithm (which compares the density in the neighborhood of an observed value with the density in the “reverse neighborhood”) (67) -, adding each new test observation once at a time and computing whether it is or not an outlier in comparison with the reference (i.e. training) data set. We have also applied the KNN approach proposed by Sahigara *et al* (2013) (68) and the method advanced by Roy *et al* (2015) (69) using R code written in house.

Virtual screening by QSAR

49 best-performing QSAR models were used to predict the activity of a data set consisting of 104619 Zinc database compounds (the "named" subset, i.e. compounds that have names in the Zinc 15 database (70)). The 49 models were stacked using a simple majority (plurality) voting for the decision; the performance of the stacking was assessed by applying the same majority voting to the independent predictions in the nested cross-validation loops. The compounds were ranked in decreasing order, from those predicted by 100% of the models to those predicted by only 51% of the models.

Molecular Docking Study

Crystallographic data available in the PDB database (PDB ID: 4MXO (71), PDB ID: 3QLG (72)) show that src-tyrosin kinase inhibitors engage the enzyme primarily at the hinge residues, a few amino acid residues having a particular relevance: Val281, Ala 293, Met314, Ile 336, Met341, Leu 393 (73). We intended to evaluate whether the molecules ranked in our virtual screening as active with highest confidence bind in the back pocket of the src-tyrosin kinase in a similar way with dasatinib or bosutinib. Docking was performed using VINA (74) with default parameters under Yasara (version 19.7.20), and LeDock. Human c-src protein (PDB ID: 2src (75)) was used as a target. For Vina, the protein preparation was performed in Chimera (Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco) using the Dock Prep module (deleting the ligand and water molecules, eliminating alternate locations of residues, replacing selenomethionine with methionine etc). The active site for the Vina docking was defined as a cubic cell of 5 Å around the selected residues (mentioned above). For LeDock the protein preparation was carried out using the LePrep module (with the default values) and the docking was run with the default values of the LeDock module. The SMILES structures corresponding to the ZINC codes of the compounds predicted as active in the virtual screening by at least 75% of the models were downloaded in Python with the help of the smilite package; they were then converted to sdf format in DataWarrior (adding 3D coordinates) and then to mol2 format (with hydrogens added) in Biovia Discovery Studio and batch split to individual mol2 files with Open Babel. Ligand energy minimization was performed with Marvin Sketch, v. 19.19. The mol2 files were used in the LeDock software for virtual screening.

To estimate the performance of the docking a subset of the training set comprising 175 compunds (33 with ki < 20 nM, 67 with 500 < ki < 1,000 nM, 32 with 1,500 < ki < 2,000 nM and 43 compounds with ki > 10,000 nM) was used and "cutpointr" R package was employed to define the best cut-off point of computed binding energies between actives and inactives, based on the sum of sensitivity and specificity. We have also computed various ligand efficiency metrics, which have been reported in the literature to improve the docking scoring; they are computed by dividing the free energy of binding to the molecular weight (ΔG/MW), number of heavy atoms (ΔG /nHM), number of carbon atoms (ΔG/nC), partition coefficient (-log(ΔG /P)), and Wiener index (ΔG/Wap) (76). We also explored computing ligand efficiencies by dividing the free binding energy to the squared value of the partition coefficient (ΔG/ALOGP2), to the total surface area (P_VSA-like descriptors), McGowan volume, van der Waals volume from McGowan volume, and van der Waals volume from Zhao-Abraham-Zissimos equation (metrics not reported previously). The "cutpointr" R package (77) was used to define the best cut-off point of computed binding energies between active and inactive compounds, based on the sum of sensitivity and specificity. For further validation we have also docked the co-crystallized ligand from the c-src protein (PDB ID 2csrc), namely the phosphoaminophosphonic acid-adenylate ester, and RMSD was computed for the first cluster of poses predicted by LeDock. RMSD computation was performed in R based on the well-known formula and the results were compared with those obtained with the online DockRMSD (78), the values obtained being identical.

Results

Data set analysis

In our study, the final data set included 1038 small organic molecules with a molecular weight varying from 188 to 1032 Da, a range usual in the QSAR modeling, with a median value of 390 Da and 75% of the values less than 440 Da. The number of atoms per molecule varied between 14 and 143, the median and mean value being 46 and 46.6, respectively. All molecules had at least one ring system and maximum six rings (with a median of 3). Only 46 of the 1038 molecules satisfied the Lipinsky's rule of five, of which 32 were labeled as "active" (ki < 1000 nM), and 14 as "inactive" (ki >= 1000nM). The variability of the data set illustrated by several simple constitutional descriptors or molecular properties is illustrated graphically in Fig. 1.

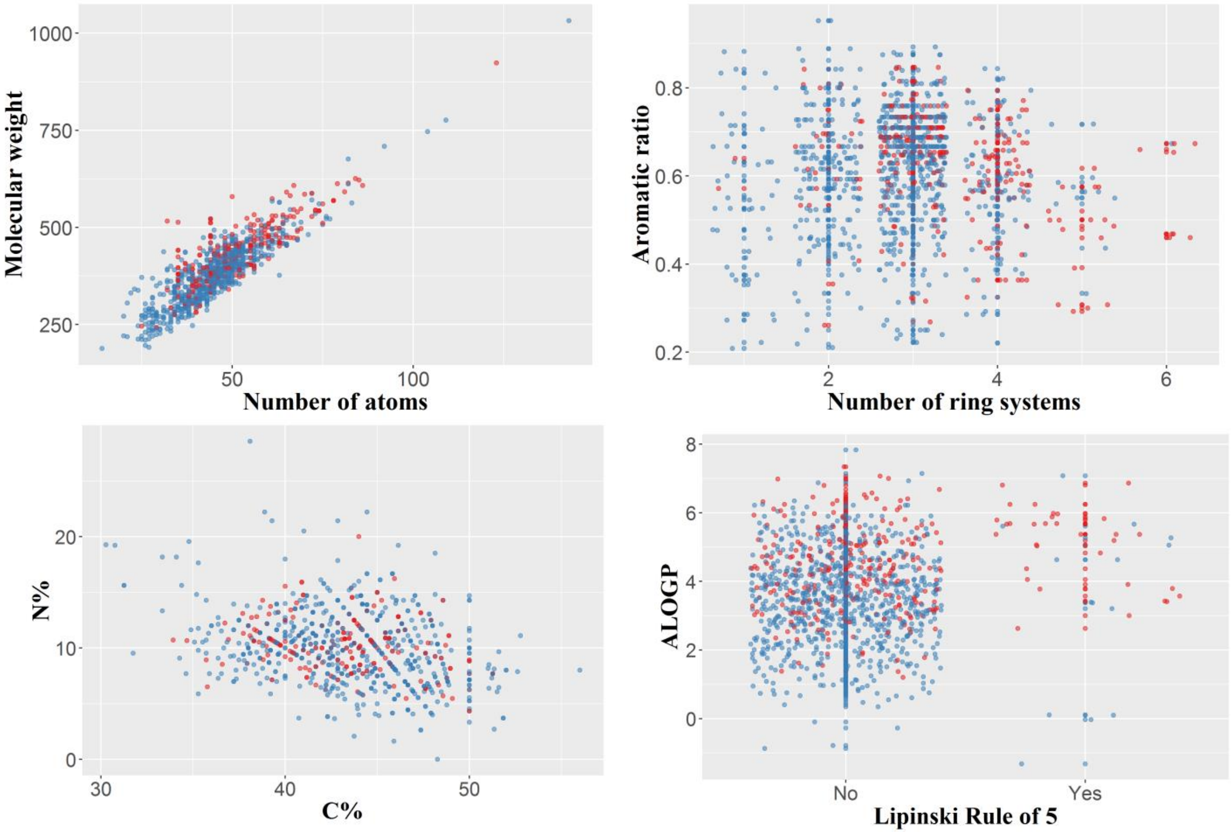

Fig. 1. Variability of the data set illustrated by several simple constitutional descriptors or molecular properties. Blue – inactive compounds; red – active compounds.

To estimate the dissimilarity of the 1038 compounds, a dissimilarity matrix based on the Gower distance was computed (the Gowever distance is appropriate for data of a heterogeneous nature), using 783 most relevant descriptors (remained after removing auto-correlated and quasi-constant features). Although Gower distance takes values between 0 and 1, because it tends to give larger weights to binary variables (because a distance to a categorical variables may only take values 0 or 1) (79), we rescaled the distance matrix and plotted it as a dissimilarity plot (Fig 2.) (before rescaling the maximum value of the Gower distance was 0.404, following rescaling it became 1). Examining the dissimilarity matrix showed that most compounds in the data set had other very similar compounds (with scaled distances under 0.1), but most compounds were quite dissimilar from other compounds (with scaled distances larger than 0.6 (Supplementary Figures S1-S3). The median (scaled) dissimilarity values were mostly around 0.2-0.3, suggesting that the chemical diversity in the data set was rather limited.

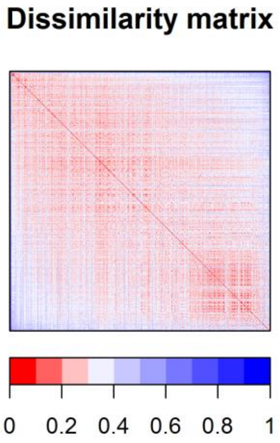

Fig. 2. Dissimilarity matrix illustrating the variability among the data set based on the Gower distances between the compounds.

Performances of models in nested cross-validation

Using a variety of classification algorithms (six), of feature selection methods (17), and numbers of features (between 3 and 40 - for instance, for binomial regression we used models with 3, 5, 10 and 20 features, and thus the number of models built for this classifier was 68), a total number of over 350 models were built and their performance was assessed by nested cross-validation. We only selected the models with an acceptable performance, defined as having both a balanced accuracy higher than 70% and a positive predictive value higher than 70% in the nested cross-validation (Table I). Where for the same classifier and selection algorithm several models (with different numbers of features) had good performance (over the threshold of 70% as explained above), we only tabulated the model we judged as best.

Model*	**BA (%)**	**PPV (%)**	**MMCE (%)**	**AUC (%)**	**TPR (%)**	**TNR (%)**
RF_anova_23	70.24	78.26	18.60	82.56	45.39	95.08
RF_auc_20	70.07	78.08	18.69	82.85	45.04	95.09
RF_cforest_13	70.07	79.39	18.60	82.96	44.80	95.34
RF_kruskal_30	70.52	77.42	18.60	82.61	46.35	94.68
RF_RFimp_30	71.54	80.04	17.73	86.03	47.69	95.39

Model*	BA (%)	PPV (%)	MMCE (%)	AUC (%)	TPR (%)	TNR (%)
RF_RF.SRCimp_20	71.01	77.44	18.31	83.76	47.18	94.83
RF_RF.SRCvarselect_10	72.93	78.72	17.34	86.01	51.29	94.56
RF_impurity_15	70.67	76.43	18.69	83.72	46.91	94.43
RF_permutation_10	71.53	80.51	17.83	83.63	47.86	95.20
RF_univariate_30	71.48	83.49	17.44	84.31	46.80	96.16
SVM_anova_30	71.83	71.26	19.07	82.08	51.60	92.05
SVM_auc_30	72.02	71.56	18.98	83.25	51.99	92.05
SVM_cforest_30	75.11	74.96	17.05	85.60	57.65	92.57
SVM_chi.sq_30	71.91	75.44	18.59	82.45	50.86	92.97
SVM_gainratio_30	72.03	72.78	18.98	82.85	51.99	92.07
SVM_information_30	72.44	73.34	18.59	83.91	52.54	92.35
SVM_kruskal_20	72.06	72.29	18.98	82.06	52.06	92.05
SVM_oneR_30	72.49	78.08	17.73	81.16	50.68	94.31
SVM_RFimp_30	74.74	74.71	17.25	86.92	57.16	92.32
SVM_RF.SRCimp_30	75.92	77.07	16.28	86.20	58.57	93.28
SVM_RF.SRCvarselect_20	76.33	76.22	16.28	86.75	60.10	92.56
SVM_impurity_30	73.96	73.86	17.82	84.27	55.61	92.30
SVM_permutation_20	72.14	73.82	18.59	84.37	51.58	92.71
SVM_relief_30	72.42	71.93	19.08	82.15	53.57	91.26
SVM_sym.uncertain_20	71.91	73.31	18.69	83.33	50.99	92.84
Adabm1_RFimp_30	71.06	73.50	19.08	83.49	49.11	93.00
Adabm1_RF.SRCvarselect_20	71.15	70.36	19.56	81.96	50.36	91.95
Adabm1_impurity_20	71.22	73.34	18.80	83.66	49.18	93.26
Adabm1_univariate_30	70.50	74.30	19.27	82.36	47.61	93.39
BartM_chi.sq_30	73.15	73.28	18.11	83.54	53.87	92.42
BartM_gainratio_20	71.61	70.19	19.37	82.45	51.57	91.64
BartM_information_20	73.56	73.52	17.92	84.08	54.68	92.44
BartM_RFimp_25	74.24	71.45	18.02	85.28	57.13	91.36
BartM_impurity_20	73.48	70.94	18.50	83.79	55.74	91.22
BartM_permutation_22	74.70	71.64	17.82	85.04	58.17	91.23
BartM_sym.uncertain_30	73.59	71.19	18.31	84.36	55.69	91.49
C50_anova_30	75.96	72.56	17.05	84.73	60.70	91.23
C50_auc_20	74.00	72.03	18.12	83.75	56.80	91.19
C50_cforest_20	75.08	71.62	17.73	85.06	59.32	90.84
C50_chi.sq_30	75.55	70.40	17.73	83.55	60.79	90.32
C50_gainratio_30	75.26	70.85	17.82	84.43	60.08	90.45
C50_kruskal_30	74.56	71.35	18.02	84.52	58.03	91.10
C50_oneR_30	73.91	72.78	18.41	83.62	57.06	90.76
C50_RFimp_30	78.56	75.39	15.32	87.24	65.23	91.89
C50_RF.SRCimp_30	76.21	72.82	17.05	85.45	61.32	91.10
C50_RF.SRCvarselect_20	77.64	72.08	16.76	87.84	65.43	89.86
C50_impurity_20	76.40	76.14	16.10	86.70	60.13	92.66
C50_permutation_30	75.93	72.28	16.96	86.29	60.51	91.36
C50_univariate_30	75.44	70.55	17.73	85.47	60.46	90.43

*Each model name is formed of three parts separated by an underscore: the first part of the name indicates the classifier, the second part the feature selection algorithm (in an abbreviated form) and the third part the number of features used to build the model. For instance, RF_anova_20 was a random forest based on features selected based on ANOVA (as implemented in "anova.test" within "mlr" R package) and the number of features used was 20.

Because in the nested cross-validation the models applied are always based on only a subset of the data, the estimation of performance should be conservative (i.e. applying the selected models on the whole data set has better performance).

Y -randomization test. As expected, despite following the same steps in building the models, scrambling the activity labels had a strong impact on the performance of the models, which was clearly inferior to those based on the initial (unscrambled) data: the average balanced accuracy of all 10 y-scrambling tests (nested cross-validation performed in the same conditions and following the same pre-processing as the true data) was 50.23%, with a standard deviation of 0.59% (minimum value 49.73% and maximum 51.45%). In a similar way, the mean value of the positive predictive (PPV) was 20.38%, and its value varied between 0.00% and 30.00%.

Descriptors associated with c-src inhibitory activity

Although for all models the number of features was relatively high (in most cases betwee 20 and 30), the largest predictive effect could be attributed to no more than 5 features. For instance, in the case of random forest, using ANOVA as a feature selection (filtering) algorithm, with 23 features the AUC was 82.56% and balanced accuracy 70.24%; however, using only the first most important five molecular descriptors, the AUC was 77.53%, and balanced accuracy 66.39%. Although there was an improvement for the larger number of features, the first five explained the largest part of the variability in the training and testing data sets. We therefore focused on the first five descriptors selected by each of the 17 selection algorithms and found that most algorithms identified the same features as being the most important. These are shown in Table II.

Table II. The most important molecular descriptors associated with the inhibition of the c-src tyrosine kinase

Name	**Interpretation**	**Descriptor block (group)**	**Frequency occurring among the first 5 most important features**
SpMax4_Bh(m)	largest eigenvalue n. 4 of Burden matrix weighted by mass	Burden eigenvalues	14
DECC	eccentric topological index	Topological indices	11
SpMax5_Bh(m)	largest eigenvalue n. 5 of Burden matrix weighted by mass	Burden eigenvalues	8
SpMax3_Bh(m)	largest eigenvalue n. 3 of Burden matrix weighted by mass	Burden eigenvalues	8

Name	Interpretation	Descriptor block (group)	Frequency occurring among the first 5 most important features
J_D	Balaban-like index from topological distance matrix (Balaban distance connectivity index)	2D matrix-based descriptors	6
F06[C-N]	frequency of C - N at topological distance 6	2D Atom Pairs	5
Chi1_EA(dm)	connectivity-like index of order 1 from edge adjacency mat. weighted by dipole moment	Edge adjacency indices	4
P_VSA_MR_6	P_VSA-like on Molar Refractivity, bin 6	P_VSA-like descriptors	3
SpMax6_Bh(m)	largest eigenvalue n. 6 of Burden matrix weighted by mass	Burden eigenvalues	3
N-073	Ar2NH / Ar3N / Ar2N-Al / R..N..R	Atom-centred fragments	2
F05[C-N]	Frequency of C - N at topological distance 5	2D Atom Pairs	2

19 other descriptors occurred only once among the 5 most important features identified by each of the 17 feature selection algorithms.

Virtual screening and external validation

We applied the models to the 104619 Zinc compounds and ranked them based on the percentage of models predicting the compounds as active. Using a threshold of 50% (i.e. compounds predicted as „active“ by more than 50% of all models applied) a number of 744 compounds were identified. Our validation data (using the predictions on the test sets from the nested cross-validation) indicated that the PPV for this threshold was 78.57%. Increasing the decision threshold to 75% the number of compounds decreased to 158, but after eliminating the compounds that had been part of the training set and the duplicates (multiple ZINC ids may correspond to the same substance), their number decreased to 115 (table SI); the validation data indicated a PPV value for this threshold of 85.43%. For a threshold of 90% the PPV in the validation was also close to 90% (90.1%), but the number of unique compounds was limited to 37.

For external validation purposes, we searched Pubchem and ChEMBL for biological data related to the activity of the predicted compounds on the src tyrosine kinase, so as to have at least partial confirmation on the accuracy of the predictions. We found that among the 115 substances predicted as being active, for 9 compounds (i.e. 7.83%) there is available evidence that they are active on the c-src tyrosine kinase; because we could not find ki values for the 9 compounds, but in most cases rather mean inhibition (as a percentage) at 1.0 µM or 0.1 µM, taking into account that IC50 values are always higher than ki values for a competitive inhibitor, and the fact that percent inhibition is dependent on both substrate and inhibitor concentration, we considered compounds with percentage inhibition values of at least 30% as active. When a compound was labeled as "active" on the src target in one of the two public databases without further information on the endpoint or bioassay used, we also considered that compound as active (that was the case for balamapimod, reported by Pubchem). Of the 9 compounds labeled by us as "active" three had a mean % inhibition higher than 50%, one had a ki less than 1000 nM (20 nM to be precise), one was stated as "active" by Pubchem with no further information and four had a mean % inhibition between 30% and 42.23% at 1 µM). 34 additional substances (29.56%) predicted by the large majority of models as being active were in fact proven to be inactive on src-tyrosine kinase, whereas 72 of the substances (62.61%) predicted to be active, seems to have never been tested for their effect on src tyrosine kinase. If the 43 compounds that were indeed tested were representative for the rest, the rate of success for the predictions would be of 20.93%).

Applicability domain

We have used a variety of algorithms to assess the applicability domain for the predictions of the QSAR virtual screening by different models. According to the method advanced by Roy et al (2015) (65), none of the compounds predicted by more than 50% of our models to be active, were outside of the applicability model. This was not very surprising, because that method uses a decision tree based on three standard deviations (values outside three standard deviations from the mean are deemed outliers), whereas we capped centered and scaled values to 2. Using the KDEOS algorithm (with minimum 3 and maximum 10 neighbours), the number of outliers among the 744 compounds predicted as active by the majority of the QSAR models was small for each model, not higher than 15% of the total number (and a median proportion toward 5%), and selecting the compounds after filtering them based on the applicability model did not change the hierarchization of the compounds predicted as active. The INFLO algorithm (with k=5 neighbours) and that of F. Sahigara *et al* (2013) (68) identified a much larger proportion of compounds as outside de applicability method: for the latter, for instance, the proportion of outliers varied (for the different models) between 1.75% and 44.35%, with a median of 32.39% of the total of 744 compounds (fig. 3). A number of 147 compounds (of which 5 had been in the training data set) were predicted by 75% of the models as being active, after limiting the votes to those compounds that were within the applicability domain estimated with the F. Sahigara et al. (2013) method. All compounds identified by the virtual screening (before checking the applicability domain) were for at least some of the models within the applicability domain, but the degree of confidence in the predictions changed after checking for the applicability domain.

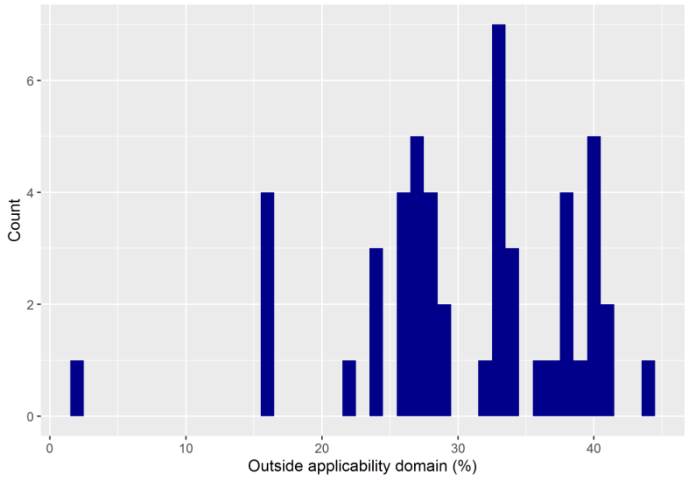

Fig. 3. Variation of the proportion of compounds estimated to be outside the applicability domain (F. Sahigara et al. method) for the 49 QSAR models used in virtual screening.

Molecular docking

In order to assess the performance of docking for the two software programs used (Vina and LeDock) we first compared the estimated binding energies for 175 compounds of the training set, with known activities on the target enzyme. With LeDock, the mean binding energy in the active compound group was -8.02, whereas in the inactive compound group it was -7.29 ($p<10^{-7}$, Welch t-test). For the very active compounds (ki<20 nM), the mean binding energy was -8.43 ($p<10^{-8}$ versus all inactive compounds, Welch t-test). Using the "cutpointr" package, an optimal cut-off was found at a binding energy of -7.17, which ensured an accuracy of 70.29%, with high sensitivity (90%), but low specificity (44%). In order to minimize the false positive, a cut-off point of -9.21 was necessary; at this level the specificity was 100% (i.e. none of the inactive compounds had such a low binding energy in the docking runs), but very low sensitivity (only 9% of the active compounds had this low estimated binding energy) (fig. 4). Because our interest was to minimize the false-positive rate, we docked the 147 compounds predicted by the QSAR models to be active and within the applicability domain and somewhat surprisingly no less than 90 of them (61.22%) had such a low binding energy, in other words they could be considered as active (Table III). Considering that in our training subset the sensitivity at this cut-off point (-9.21) was of only 9%, this high value does suggest that an important proportion of the compounds predicted by the QSAR models to be active might be indeed active, although when using docking one must be very cautious (80). The RMSD computed for the first cluster of poses of the ANP was 1.25, under the conventional threshold of 2.0, which may be considered reasonably well. The visual examination of the pose indicated that the ring pose was very well predicted, whereas the side chain prediction was less accurate (fig. 5).

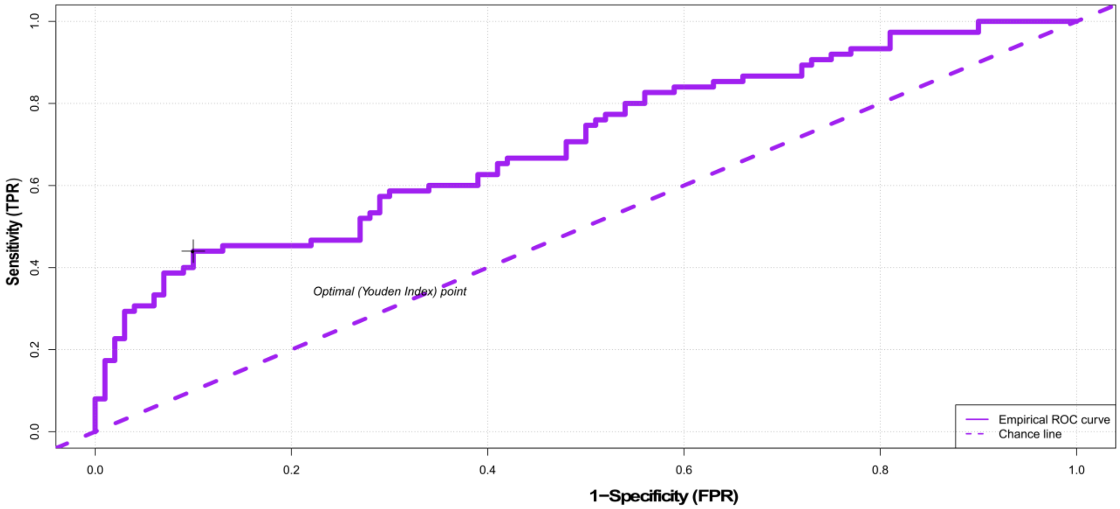

Fig. 4. Receiver operating characteristic curve for the performance of molecular docking using LeDock software on the training set (n=175 compounds, as described in the text).

Table III. Compounds predicted to be active by both the assembled QSAR models and ligand docking

Zinc Code	Substance name	Confirmation in wet lab experiments	In the training set?	Binding energy
ZINC000001550477	Lapatinib	?	Yes	-10.07
ZINC000034638188	Pf-562271	?	Yes	-9.3
ZINC000063298074	Ilorasertib	?	Yes	-10.09
ZINC000034800096	Gw583373a	No (6.34% at 1 uM, 1.76% at 0.1 uM). Active on Yes 1, EGFR and ERBB4	No	-11.02
ZINC000027184814	Vibriobactin	?	No	-9.77
ZINC000034800093	Gw580496a	No (6.5-6.7%). Active on EGFR, ERBB2, and ERBB4	No	-9.33
ZINC000150528975	Vedroprevir	?	No	-11.51
ZINC000034800112	Gw576484x	No (8.28% at 1 uM). Active on ERBB1	No	-10.36
ZINC000072190218	Avatrombopag	?	No	-9.28
ZINC000034800091	Gw576609a	No (8.56-10.17%). Reported as "active" on Yes1. Active on EGFR and ERBB4	No	-11.38
ZINC000044418656	Gw784684x	No (13.41%-18.34%). Active on Yes TK	No	-10.77
ZINC000042804069	Gsk-182497a	No (3.86%-6.56%). Active on EGFR and ERBB4	No	-9.57
ZINC000103297739	Defactinib	No. FAK inhibitor	No	-10.23
ZINC000004215255	Cefpimizole	?	No	-10.54
ZINC000042834127	Gsk1751853a	No (4.53%-11.77%). Active on EGFR and ERBB4	No	-10.34
ZINC000014945166	Gw830365a	No (4.5%-18.43%). Stated "active" on Yes1	No	-9.53
ZINC000150339466	Ciluprevir	?	No	-10.95
ZINC000043195317	Golvatinib	No (>30 uM). Active on LCK	No	-14
ZINC000042201866	Gw566221a	No (8.85%-10.94%). Stated "active" on Yes, EGFR and ERBB4	No	-10.06
ZINC000095615094	Patellamide G	?	No	-9.32
ZINC000003604326	Vaneprim	?	No	-11.01
ZINC000002007399	Gw458787a	No. Active on Yes (42% inhibition at 1uM, 463 nM potency), EGFR and ERBB4	No	-10.95
ZINC000028639340	Posaconazole	?	No	-10.92
ZINC000072122048	Gsk259178a	No (8.86% at 1 uM). Active on EGFR and ERBB4	No	-12.44

Zinc Code	Substance name	Confirmation in wet lab experiments	In the training set?	Binding energy
ZINC000068204830	Daclatasvir	?	No	-10.75
ZINC000043131420	Fostamatinib	?	No	-10.77
ZINC000169289453	Simeprevir	?	No	-11.45
ZINC000042834162	Gw869810x	No (-0.25%, 11.35%). Active on EGFR and ERBB4	No	-12.11
ZINC000049709569	Asperazine	?	No	-11.6
ZINC000096928979	Deleobuvir	?	No	-10.2
ZINC000042201868	Gw568377a	No (3.56 - 4.6%). Stated "active" on Yes, EGFR, ERBB4	No	-9.36
ZINC000014945147	Gw809897x	Yes (65.38% at 1 uM)	No	-10.44
ZINC000014945171	Gw830263a	Yes (42.23% at 1 uM). Stated "active" on Yes	No	-10.53
ZINC000014945045	Gw569530a	No (6.29-7.64%). Inconclusive on Yes1, active on EGFR	No	-9.52
ZINC000003925087	Gw806742x	Yes (86.65% at 1 uM)	No	-10.43
ZINC000095618748	Candesartan O-Glucuronide	?	No	-9.71
ZINC000098052868	Olcegepant	?	No	-9.55
ZINC000049833405	Preulicyclamide	?	No	-11.13
ZINC000034800110	Gw574782a	No (9.13-12.49%). Active on EGFR, ERBB1, ERBB2	No	-10.42
ZINC000014965596	Gw683134a	Yes (36.99% at 1 uM). Active on LYN and Yes1	No	-10.91
ZINC000034800112	Gw576484x	No (8.28% at 1 uM). Active on ERBB1	No	-9.93
ZINC000019862646	Fedratinib	Yes	No	-10.23
ZINC000150377731	Bms-247243	?	No	-10.42
ZINC000003986669	Bx-795	Yes (27-30% inhibition on human src, 77%-90% on Gallus gallus src), active on Yes1	No	-9.28
ZINC000095615898	Tyrokeradine A	?	No	-11.14
ZINC000003919988	L-766892	?	No	-9.59
ZINC000095544067	Ulithiacyclamide F	?	No	-9.76
ZINC000049889335	Edulirin A	?	No	-11.45
ZINC000003995140	Gw621823a	No (3.8% - 6.22%). Active on Yes1, EGFR, ERBB1, ERBB2, ERBB4	No	-10.63
ZINC000040379218	Gw684626b	No (2.7%) at 1 uM. Inconclusive on Yes 1. Active on EGFR	No	-10.46
ZINC000034800121	Gw567808a	No (15.81% at 1 uM). Active on EGFR, ERBB1, ERBB2	No	-10.42

doi:10.20944/preprints201910.0113.v1

Zinc Code	Substance name	Confirmation in wet lab experiments	In the training set?	Binding energy
ZINC000169306513	Hydroxyitraconazole	?	No	-9.78
ZINC000169368380	Kni-1039	?	No	-10.13
ZINC000150601177	Ombitasvir	?	No	-10.07
ZINC000040404350	Gsk-969786a	No (6.58%). Active on EGFR and ERBB4	No	-10.2
ZINC000150592451	Micromide	?	No	-12.96
ZINC000028249631	Pd-170292	?	No	-10.1
ZINC000169366333	Porphyrin	?	No	-11.05
ZINC000034800119	Gw576924a	No (8.51%-5.28%). Active on Yes 1, EGFR and ERBB4	No	-10.18
ZINC000150362888	Lissoclinamide 2	?	No	-10.23
ZINC000100057121	Tegobuvir	?	No	-10.55
ZINC000103213128	Heptamethylene 1,7-Bis-Imadacloprid	?	No	-9.58
ZINC000169291993	Sansanmycin F	?	No	-9.5
ZINC000230052516	Urobilin	?	No	-10.9
ZINC000003994828	Brecanavir	?	No	-10.41
ZINC000169363931	Ansacarbamitocin C	?	No	-10.56
ZINC000095535868	Rwj-58259	?	No	-10.09
ZINC000003921862	Tallimustine	?	No	-9.76
ZINC000150362887	Pyropheophytin B	?	No	-10.18
ZINC000063933734	Rebastinib	No (ki=5.85 uM). Active on Yes1, Lyn, LCK.	No	-9.73
ZINC000095615652	Patellamide C	?	No	-9.46
ZINC000197688172	S-[(3e,5z)-3,5-Octadienoate	?	No	-9.6
ZINC000014965588	Gw709042a	No (11.2%). Inconclusive on Yes1. Active on ABL1, PDGFRA, and KIT	No	-9.89
ZINC000085537136	Barixibat	?	No	-9.72
ZINC000169291499	Kibdelomycin	?	No	-10.99
ZINC000003946578	Mitratapide	?	No	-10.41
ZINC000001481922	Setipafant	?	No	-10.05
ZINC000072173092	Deoxyvobstusine Lactone	?	No	-9.66
ZINC000006717126	Quarfloxin	?	No	-9.85
ZINC000077301904	Losartan N2-Glucuronide	?	No	-10.86
ZINC000150609364	Pseudoceratinazole A	?	No	-11.38
ZINC000095616246	Ulithiacyclamide E	?	No	-9.35
ZINC000068151111	Narlaprevir	?	No	-9.96
ZINC000150351429	Phytosulfokine B	?	No	-9.7

Zinc Code	Substance name	Confirmation in wet lab experiments	In the training set?	Binding energy
ZINC000003989268	Ceftaroline Fosamil	?	No	-9.84
ZINC000008552132	Pristinamycin	?	No	-11.01
ZINC000095618880	Clofazimine Glucuronide	?	No	-9.65
ZINC000096006065	Xv638	?	No	-9.56
ZINC000169292535	Rifapentine	?	No	-12.81
ZINC000150341961	Mafodotin	?	No	-9.32

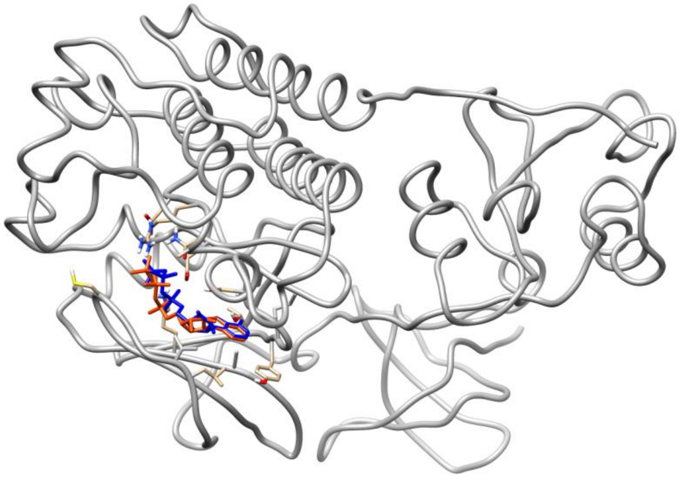

Fig. 5. Crystallographic pose of the NAP ligand within c-src tyrosine kinase (in red) and predicted pose by LeDock (in blue). It may be seen that the rings overlap very closely, whereas the free aliphatic chains do not overlap so well.

Vina performance was inferior to that of LeDock: on the same 175 compounds from the training set, the mean binding energy was -10.30 for the active compounds and -10.03 for the inactive (p=0.21, Welch t-test). An optimal cut-off for the Vina compounds was at -9.26, which ensured an accuracy of only 62.86%, with a sensitivity of 87.00 % and a specificity of only 30.67%. Because the performance of Vina was inferior to that of LeDock, we preferred to use only LeDock for virtual screening.

Computing various ligand efficiency metrics did not improve the predictions in the case of LeDock results: the accuracy rather decreased with all ligand efficiency measured attempted. In the case of Vina, using different ligand efficiency measures changed the values of accuracy, sensitivity, and specificity, with no spectacular improvement. For instance, dividing the binding energy to the molecular weight decreased sensitivity (from 87% to 43%), increased specificity (from 30.67 to 81.33%), and slightly increased the AUC (from 56.85% to 62.87%), but it also slightly decreased the accuracy (from 62.86% to 59.43%). Of the different ligand efficiency measures, for the Vina results the best was obtained by dividing the binding energy to the squared Ghose-Crippen octanol-water partition coefficient: 78% sensitivity, 49.33% specificity, 65.71% accuracy, and 65.05% AUC. Even with this ligand efficiency measure, the results were inferior to those obtained with LeDock based on the binding energies.

Discussions

Several studies of QSAR models for c-src tyrosine kinase inhibitors of have been published up to date in the scientific literature. A number of five such studies have explored the use of 3D-QSAR, and all of them used relatively small number of compounds (80, 42, 156, and 39, respectively), with the same basic chemical structure within each study (pyrrolo-pyrimidine, quinazoline, anilinoquinazoline and quinolinecarbonitrile, quinolinecarbonitrile, and 4,6-substituted-(diaphenylamino)quinazolines); they could, therefore, be considered "local" models (81–84). In the QSAR field, the term "local" is used to designate models based on a data set consisting of compounds related by their chemical structure, unlike global models, that are based on data sets consisting of structurally diverse chemical substances (85). An additional paper reported on the use of 2D-QSAR for c-src inhibitors, but these models were also local, focused on ethynyl-3-quinolinecarbonitriles (86).Therefore, our study is the first one focused on global QSAR models for inhibitors targeting the c-src tyrosine kinase. It has been argued (and it stands to reason) that local models tend to have limited predictive power, even when their apparent performance indicates that they are robust (85). Our global models are expected to have a higher predictive power, as partially confirmed in our external validation.

By far the most important descriptor in our work, identified by multiple feature selection algorithms, was SpMax4_Bh(m), the largest eigenvalue n. 4 of Burden matrix weighted by mass. This has not generally been reported in previous works as correlating with pharmacological activities. Other two Burden eigenvalues (SpMax3_Bh(m), SpMax5_Bh(m)) have also been among the most descriptors correlating with the inhibition of c-src. SpMax3_Bh(m) has been used in predicting depuration rate constants for environmental pollutants of the polychlorinated biphenyls group (87), and the less relevant (in our case) SpMax6_Bh(m) has been used to predict chronic toxicity of substances to *Pseudokirchneriella subcapitata* (88). The second most important descriptor for our data set was DECC (eccentric topologic index) has been previously reported to be important in the prediction of MAO-A activity (89,90) , placental barrier permeability (91), and gas chromatographic retention times (92). F06[C-N] was used in a model to describe the anti-proliferative effect of phenyl 4-(2-oxoimidazolidin-1-yl)-benzenesulfonates (local QSAR model) (93), anti-malaric effect (94), or skin permeability of substances (95). P_VSA_MR_6 has also been used for modeling of skin permeability (95), whereas we have identified the use of Chi1_EA(dm) only for the QSPR modeling of fluorescence properties of a number of fluorescent dyes (96). The aromatic nitrogen (N-073) has been shown to correlate positively with HIV-1 integrase activity inhibition (97) and negatively with the inhibition of the fibroblast growth factor (FGFR) (98). We found no previous reports on the use of the Balaban distance connectivity index (J_D) in other models in the biological field, neither of the F05[C-N].

Another aspect worth noting is that rarely the 49 QSAR models with similarly good performance converged in their predictions. Only 8 compounds were predicted by all models to be active, and half of them (n=4) were already in the training data set; for the large majority of compounds at least one or more of the models had contradictory results. This illustrates the need to avoid making decisions based on the results of a single or a small number of models.

As shown in the results section, for a number of 9 compounds (7.83% of the 115 substances with the best predictions) it has been confirmed from independent experiments that they are active. How good is such a measure for a virtual screening exercise? If we compare it with the PPV value in the nested cross-validation, the results are rather disappointing and indicate that one should always be cautious in interpreting results even when using double cross-validation, because the real world data are likely to be different from the data set used for training and testing. For instance, it is likely that the proportion of actives in the available data set used for the construction of the models is higher than the proportion of actives in the „real world" (i.e. the wide chemical space used for virtual screening), and this may lead to a decrease in the positive predictive value in the real world. However, if we compare the results of the virtual screening with those of the most costly high throughput screening (HTS), the results are far from being bad. It has been reported that the hit rate of HTS should be expected to be less than 1% (99) and even less than 0.1% or 0.01% (100). In one study adding a computer-aided virtual screen was able to increase the screening hit proportion to 5.8% (99).Thus, our success rate of at least 7.83% is reasonably good. If we compute the confirmation rate against the compounds that were assessed for their effect on src-tyrosine kinase (20.93%), the results are even better. Our virtual screening results showed, however, additional interesting facts.

16 additional false positives, were in fact reported to be active on other members of the src family members, particularly Yes1 tyrosine kinase. This suggests that although our virtual screening exercise failed in multiple cases, the failure was often not far from the true target. Thus, from a total of 43 molecules that were tested for their effects on the src and other tyrosine kinases, 58.14% (25 compounds) were inhibitors of one or several members of the src-tyrosine kinase family (most often Yes1, sometimes also LCK or LYN tyrosin kinase).

Other false positives of the virtual screening exercise are inhibitors of proteins that src tyrosine kinase interacts directly, either activating them or being activated by them. It is known, for instance, that EGFR (epidermal growth factor receptor) can be activated by src without the presence of the EGFR ligand and that there is a direct correlation between EGFR overexpression and Src activation (101). Rather surprisingly for us, 13 compounds wrongly predicted by our models to be src tyrosine kinase inhibitors, are in point of fact inhibitors of EGFR, and 10 additional compounds that were inactive on src or other members of src family, were reported to be inhibitors of EGFR. Most of these 10 additional compounds (as well as most of the compounds active on src or yes1 tyrosine kinase) are also active on ERBB4, and it has been reported that ErbB4-derived phosphopeptides are able to interact with the SH2 domain of src (102), that following stimulation by EGF, c-src is rapidly recruited to ErbB receptor complexes (103) and that activated src binds to ERBB4s80 (E4ICD), a cleaved fragment of ERBB4 (104). Moreover, dasatinib, described often as a src inhibitor (105), has also shown to be one of the most potent ligands of ERBB4 (106). Defactinib, apparently a false positive of our virtual screening is a potent FAK (focal adhesion kinase) inhibitor; it is known that FAK and non-receptor src tyrosin kinase are both part of a focal adhesion complex (together with other structural, enzymatic or adapter proteins), where they interact directly (107). Three false positives of the virtual screening results were KIT and PDGFR inhibitors; KIT promotes phosphorylation of src and is activated by src (108), while src and PDGFR interact and phosphorylates each other at certain Tyr positions (109).

Such findings tend to suggest that where the QSAR virtual screening fails is often not far from the target (but this is not less a failure). How could these failures been explained, considering that multiple models converge in predicting a certain molecule as active on the target of interest (src tyrosine kinase)? It seems that the models manage to predict the tyrosine kinase properties of certain compounds, without having sufficient specificity to always separate those active on src from those active on other tyrosine kinases. We hypothesize that the training set is too small and does not include (a sufficient number of) molecules with selective src inhibitory properties; we intend to evaluate whether extending the data set with additional molecules inactive on src but active on other tyrosine kinases may improve the results of the virtual screening. It is also worth exploring the combining of more diverse descriptor sets in the final assemble of models with a view of improving the performance of the virtual screening.

Among the results produced by our virtual screening there is a sizeable number of antiviral molecules (vedroprevir, daclatasvir, ciluprevir, deleobuvir, ledipasvir, faldaprevir, tegobuvir, elbasvir, ombitasvir, narlaprevir), all of them approved or developed against hepatitis C viruses. They either target the NS3/NS4A (vedroprevir, ciluprevir, faldaprevir, narlaprevir) (110) or NS5A (daclatasvir, elbasvir, ombitasvir, ledipasvir) (111) or NS5B (deleobuvir, tegobuvir) (112) non-structural proteins of the virus. It is not very surprising to see inhibitors of NS5A and NS5B here, considering that is already known that NS5A protein binds to tyrosine kinases from the src-family (113), and c-src is an essential host protein involved in the formation of the HCV replication complex, together with NS5A and NS5B (114). It was less expected to see also inhibitors of the NS3/NS4A among the results of the virtual screening, because no direct interaction was reported between the Ns3/NS4A complex and src tyrosine kinase. This list of HCV antivirals might consist only of false positives, but it is worth testing in wet lab experiments.

The docking applied to 147 compounds predicted with a high probability by the QSAR models to be active reduced their number to about 61% of the initial number. For a number (27.78%) of these 90 compounds, predicted by both QSAR and docking to be active, data available in CHEMBLE or PUBCHEM (from a single wet lab test) indicate that they are inactive, and for others (6.67%), that they are active, as discussed for the QSAR models. This suggests that computational results have to be interpreted with cautious even when different models, with different methodologies and assumptions converge in their predictions. On the other hand, the last decade has witnessed a growing realization of what has been dubbed “the reproducibility crisis”, ascribed to the inappropriate quality of antibodies used as reagents (115), insufficiently described methodologies or simply to the biology itself (116). Whereas positive findings have often not been reproduced when experiments were repeated in other laboratories, it is not impossible that negative findings could also not be replicable and some of the compounds shown by databases to be inactive might as a matter of fact be active. However, in the absence of contrary evidence, such compounds have to be considered inactive.

Conclusions

A number of 49 global QSAR models have been developed, predicting the c-src tyrosine kinase inhibition with reasonable accuracy (> 70%) and positive predictive value (> 70%). The 49 models were assembled by stacking and used for the virtual screening of over 100,000 named compounds from the ZINC database. Several hundreds of compounds were predicted to be active, depending on the decision threshold used. Those with the highest probability of being active were also subjected to molecular docking and for the majority (about 61%) of them the binding energies obtained were consistent with a hypothesis of activity. External data from CHEMBL and PUBCHEM confirmed that at least 7.83% (in the case of QSAR) or 6.67% (in the case of integrated QSAR and molecular docking) of the compounds are active on the c-src target. The proportions of active compounds are less than what was to be expected from the nested cross-validation data, but still better than what one should expect from high-throughput screening experiments.

Author Contributions

Conceptualization, R.A. and M.D.; Methodology, R.A.; Formal Analysis, B.T.; Investigation, R.A., M.D., and C.S.; Writing – Original Draft Preparation, R.A. and M.D. X.X.; Writing – Review & Editing, B.T. and C.S.; Visualization, R.A.

Conflicts of Interest

The authors declare no conflict of interest. R.A has received consultancy and speakers' fees from various pharmaceutical companies. The companies had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

1. Oneyama C, Okada M. MicroRNAs as the fine-tuners of Src oncogenic signalling. J Biochem (Tokyo). 2015 Jun;157(6):431–8.

2. Parsons JT, Parsons SJ. Src family protein tyrosine kinases: cooperating with growth factor and adhesion signaling pathways. Curr Opin Cell Biol. 1997 Apr;9(2):187–92.

3. Fowler AJ, Hebron M, Missner AA, Wang R, Gao X, Kurd-Misto BT, et al. Multikinase Abl/DDR/Src Inhibition Produces Optimal Effects for Tyrosine Kinase Inhibition in Neurodegeneration. Drugs RD. 2019 Jun;19(2):149–66.

4. Liu P, Gu Y, Luo J, Ye P, Zheng Y, Yu W, et al. Inhibition of Src activation reverses pulmonary vascular remodeling in experimental pulmonary arterial hypertension via Akt/mTOR/HIF-1<alpha> signaling pathway. Exp Cell Res. 2019 Jul 1;380(1):36–46.

5. Montani D, Seferian A, Savale L, Simonneau G, Humbert M. Drug-induced pulmonary arterial hypertension: a recent outbreak. Eur Respir Rev Off J Eur Respir Soc. 2013 Sep 1;22(129):244–50.

6. Guignabert C, Phan C, Seferian A, Huertas A, Tu L, Thuillet R, et al. Dasatinib induces lung vascular toxicity and predisposes to pulmonary hypertension. J Clin Invest. 2016 01;126(9):3207–18.

7. Özgür Yurttaş N, Eşkazan AE. Dasatinib-induced pulmonary arterial hypertension. Br J Clin Pharmacol. 2018;84(5):835–45.

8. Broeckel R, Sarkar S, May NA, Totonchy J, Kreklywich CN, Smith P, et al. Src Family Kinase Inhibitors Block Translation of Alphavirus Subgenomic mRNAs. Antimicrob Agents Chemother. 2019 Apr;63(4).

9. Dai Y, Siemann D. c-Src is required for hypoxia-induced metastasis-associated functions in prostate cancer cells. OncoTargets Ther. 2019;12:3519–29.

10. Molinari A, Fallacara AL, Di Maria S, Zamperini C, Poggialini F, Musumeci F, et al. Efficient optimization of pyrazolo[3,4-d]pyrimidines derivatives as c-Src kinase inhibitors in neuroblastoma treatment. Bioorg Med Chem Lett. 2018 15;28(21):3454–7.

11. Halaban R, Bacchiocchi A, Straub R, Cao J, Sznol M, Narayan D, et al. A novel anti-melanoma SRC-family kinase inhibitor. Oncotarget. 2019 Mar 19;10(23):2237–51.

12. Ku K-E, Choi N, Oh S-H, Kim W-S, Suh W, Sung J-H. Src inhibition induces melanogenesis in human G361 cells. Mol Med Rep. 2019 Apr;19(4):3061–70.

13. Henderson YC, Toro-Serra R, Chen Y, Ryu J, Frederick MJ, Zhou G, et al. Src inhibitors in suppression of papillary thyroid carcinoma growth. Head Neck. 2014 Mar;36(3):375–84.

14. Roelants C, Giacosa S, Pillet C, Bussat R, Champelovier P, Bastien O, et al. Combined inhibition of PI3K and Src kinases demonstrates synergistic therapeutic efficacy in clear-cell renal carcinoma. Oncotarget. 2018 Jul 10;9(53):30066–78.

15. Ahn K, O YM, Ji YG, Cho HJ, Lee DH. Synergistic Anti-Cancer Effects of AKT and SRC Inhibition in Human Pancreatic Cancer Cells. Yonsei Med J. 2018 Aug;59(6):727–35.

16. Simpkins F, Jang K, Yoon H, Hew KE, Kim M, Azzam DJ, et al. Dual Src and MEK Inhibition Decreases Ovarian Cancer Growth and Targets Tumor Initiating Stem-Like Cells. Clin Cancer Res Off J Am Assoc Cancer Res. 2018 Oct 1;24(19):4874–86.

17. PubChem Data Counts [Internet]. [cited 2019 Jun 25]. Available from: https://pubchemdocs.ncbi.nlm.nih.gov/statistics

18. Reymond J-L. The Chemical Space Project. Acc Chem Res. 2015 Mar 17;48(3):722–30.

19. Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des. 2013 Aug;27(8):675–9.

20. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996 Jan;16(1):3–50.

21. Ertl P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. J Chem Inf Comput Sci. 2003 Mar;43(2):374–80.

22. Gini G. QSAR: What Else? In: Nicolotti O, editor. Computational Toxicology [Internet]. New York, NY: Springer New York; 2018 [cited 2019 Jun 24]. p. 79–105. Available from: http://link.springer.com/10.1007/978-1-4939-7899-1_3

23. Bellera CL, Talevi A. Quantitative structure–activity relationship models for compounds with anticonvulsant activity. Expert Opin Drug Discov. 2019 Jul 3;14(7):653–65.

24. Ai S, Lin G, Bai Y, Liu X, Piao L. QSAR Classification-Based Virtual Screening Followed by Molecular Docking Identification of Potential COX-2 Inhibitors in a Natural Product Library. J Comput Biol J Comput Mol Cell Biol. 2019 Jun 24;

25. Allam L, Fatima G, Wiame L, Hamid EA, Azeddine I. Molecular screening and docking analysis of LMTK3and AKT1 combined inhibitors. Bioinformation. 2018;14(9):499–503.

26. Zhou Y, Peng J, Li P, Du H, Li Y, Li Y, et al. Discovery of novel indoleamine 2,3-dioxygenase 1 (IDO1) inhibitors by virtual screening. Comput Biol Chem. 2019 Feb;78:306–16.

27. Lagunin AA, Romanova MA, Zadorozhny AD, Kurilenko NS, Shilov BV, Pogodin PV, et al. Comparison of Quantitative and Qualitative (Q)SAR Models Created for the Prediction of Ki and IC50 Values of Antitarget Inhibitors. Front Pharmacol. 2018;9:1136.

28. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature Selection: A Data Perspective. ACM Comput Surv. 2017 Dec 6;50(6):1–45.

29. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J Mach Learn Res. 2016;17(170):1–5.

30. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional Variable Importance for Random Forests. BMC Bioinformatics [Internet]. 2008;9(307). Available from: http://www.biomedcentral.com/1471-2105/9/307

31. Romanski P, Kotthoff L. FSelector: Selecting Attributes [Internet]. 2016. Available from: https://CRAN.R-project.org/package=FSelector

32. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

33. Ishwaran H, Kogalur UB. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC) [Internet]. manual; 2019. Available from: https://cran.r-project.org/package=randomForestSRC

34. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77(1):1–17.

35. Ancuceanu R, Dinu M, Neaga I, Laszlo F, Boda D. Development of QSAR machine learning-based models to forecast the effect of substances on malignant melanoma cells. Oncol Lett [Internet]. 2019 Feb 25 [cited 2019 Jul 9]; Available from: http://www.spandidos-publications.com/10.3892/ol.2019.10068

36. Hdoufane I, Bjij I, Soliman M, Tadjer A, Villemin D, Bogdanov J, et al. In Silico SAR Studies of HIV-1 Inhibitors. Pharm Basel Switz. 2018 Jul 13;11(3).

37. Gadaleta D, Manganelli S, Roncaglioni A, Toma C, Benfenati E, Mombelli E. QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. J Chem Inf Model. 2018 Aug 27;58(8):1501–17.

38. Hodyna D, Kovalishyn V, Semenyuta I, Blagodatnyi V, Rogalsky S, Metelytsia L. Imidazolium ionic liquids as effective antiseptics and disinfectants against drug resistant S. aureus : In silico and in vitro studies. Comput Biol Chem. 2018 Apr;73:127–38.

39. Sun Y, Shi S, Li Y, Wang Q. Development of quantitative structure-activity relationship models to predict potential nephrotoxic ingredients in traditional Chinese medicines. Food Chem Toxicol Int J Publ Br Ind Biol Res Assoc. 2019 Jun;128:163–70.

40. Chen H, Chen L. Support Vector Machine Classification of Drunk Driving Behaviour. Int J Environ Res Public Health. 2017 23;14(1).

41. Idakwo G, Luttrell J, Chen M, Hong H, Zhou Z, Gong P, et al. A review on machine learning methods for in silico toxicity prediction. J Environ Sci Health Part C Environ Carcinog Ecotoxicol Rev. 2018;36(4):169–91.

42. Lei T, Sun H, Kang Y, Zhu F, Liu H, Zhou W, et al. ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning Approaches. Mol Pharm. 2017 06;14(11):3935–53.

43. Feng D, Svetnik V, Liaw A, Pratola M, Sheridan RP. Building Quantitative Structure-Activity Relationship Models Using Bayesian Additive Regression Trees. J Chem Inf Model. 2019 Jun 24;59(6):2642–55.

44. Dieguez-Santana K, Pham-The H, Rivera-Borroto OM, Puris A, Le-Thi-Thu H, Casanola-Martin GM. A Two QSAR Way for Antidiabetic Agents Targeting Using α-Amylase and α-Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. Lett Drug Des Discov [Internet]. 2017 Jul 20 [cited 2019 Jul 9];14(8). Available from: http://www.eurekaselect.com/147708/article

45. Raevsky OA, Grigorev VY, Yarkov AV, Polianczyk DE, Tarasov VV, Bovina EV, et al. Classification (Agonist/Antagonist) and Regression “Structure-Activity” Models of Drug Interaction with 5-HT6. Cent Nerv Syst Agents Med Chem. 2018 26;

46. Kuhn M, Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models [Internet]. 2018. Available from: https://CRAN.R-project.org/package=C50

47. Bharti DR, Lynn AM. QSAR based predictive modeling for anti-malarial molecules. Bioinformation. 2017;13(5):154–9.

48. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

49. Bischl B, Lang M. parallelMap: Unified Interface to Parallelization Back-Ends [Internet]. 2015. Available from: https://CRAN.R-project.org/package=parallelMap

50. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training [Internet]. 2019. Available from: https://CRAN.R-project.org/package=caret

51. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2019. Available from: https://CRAN.R-project.org/package=e1071

52. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.

53. Hornik K, Buchta C, Zeileis A. Open-Source Machine Learning: R Meets Weka. Comput Stat. 2009;24(2):225–232.

54. Kapelner A, Bleich J. bartMachine: Machine Learning with Bayesian Additive Regression Trees. J Stat Softw. 2016;70(4):1–40.

55. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. 2019.

56. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

57. Hahsler M, Hornik K, Buchta C. Getting things in order: An introduction to the R package seriation. J Stat Softw. 2008 Mar;25(3):1–34.

58. Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J Cheminformatics. 2014;6(1):47.

59. Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007 May;26(5):694–701.

60. Roy K, Ambure P. The “double cross-validation” software tool for MLR QSAR model development. Chemom Intell Lab Syst. 2016 Dec;159:108–26.

61. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, et al. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model. 2008 Sep;48(9):1733–46.

62. Capuzzi SJ, Sun W, Muratov EN, Martínez-Romero C, He S, Zhu W, et al. Computer-Aided Discovery and Characterization of Novel Ebola Virus Inhibitors. J Med Chem. 2018 26;61(8):3582–94.

63. Warnes GR, Bolker B, Lumley T. gtools: Various R Programming Tools [Internet]. 2018. Available from: https://CRAN.R-project.org/package=gtools

64. Yang H, Du Z, Lv W-J, Zhang X-Y, Zhai H-L. In silico toxicity evaluation of dioxins using structure–activity relationship (SAR) and two-dimensional quantitative structure–activity relationship (2D-QSAR). Arch Toxicol [Internet]. 2019 Sep 24 [cited 2019 Oct 3]; Available from: http://link.springer.com/10.1007/s00204-019-02580-w

65. Madsen JH. DDoutlier: Distance & Density-Based Outlier Detection [Internet]. 2018. Available from: https://CRAN.R-project.org/package=DDoutlier

66. Schubert E, Zimek A, Kriegel H-P. Generalized Outlier Detection with Flexible Kernel Density Estimates. In: Proceedings of the 2014 SIAM International Conference on Data Mining [Internet]. Society for Industrial and Applied Mathematics; 2014 [cited 2019 Sep 25]. p. 542–50. Available from: https://epubs.siam.org/doi/10.1137/1.9781611973440.63

67. Jin W, Tung AKH, Han J, Wang W. Ranking Outliers Using Symmetric Neighborhood Relationship. In: Ng W-K, Kitsuregawa M, Li J, Chang K, editors. Advances in Knowledge Discovery and Data Mining [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006 [cited 2019 Sep 25]. p. 577–93. Available from: http://link.springer.com/10.1007/11731139_68

68. Sahigara F, Ballabio D, Todeschini R, Consonni V. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J Cheminformatics. 2013;5(1):27.

69. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst. 2015 Jul;145:22–9.

70. Sterling T, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. J Chem Inf Model. 2015 Nov 23;55(11):2324–37.

71. Levinson NM, Boxer SG. human Src kinase bound to kinase inhibitor bosutinib. (:unav) [Internet]. 2013 Dec 4 [cited 2019 Aug 7]; Available from: ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/mx/pdb4mxo.ent.gz

72. Boubeva R, Pernot L, Perozzo R, Scapozza L. Crystal structure of the L317I mutant of the C-src tyrosine kinase domain complexed with dasatinib. (:unav) [Internet]. 2012 Feb 8 [cited 2019 Aug 7]; Available from: ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/ql/pdb3qlg.ent.gz

73. Roskoski R. Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. Pharmacol Res. 2015 Apr;94:9–25.

74. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010 Jan 30;31(2):455–61.

75. Xu W, Doshi A, Lei M, Eck MJ, Harrison SC. CRYSTAL STRUCTURE OF HUMAN TYROSINE-PROTEIN KINASE C-SRC, IN COMPLEX WITH AMP-PNP. (:unav) [Internet]. 1999 Jul 22 [cited 2019 Aug 7]; Available from: ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/sr/pdb2src.ent.gz

76. Garcia-Sosa AT, Hetenyi C, Maran U. Drug efficiency indices for improvement of molecular docking scoring functions. J Comput Chem. 2010 Jan 15;31(1):174–84.

77. Thiele C. cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks [Internet]. 2019. Available from: https://CRAN.R-project.org/package=cutpointr

78. Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. J Cheminformatics [Internet]. 2019 Dec [cited

2019 Oct 3];11(1). Available from: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0362-7

79. Wang S, Yabes JG, Chang C-CH. Hybrid Density- and Partition-based Clustering Algorithm for Data with Mixed-type Variables. ArXiv190502257 Cs Stat [Internet]. 2019 May 6 [cited 2019 Jul 12]; Available from: http://arxiv.org/abs/1905.02257

80. Chen Y-C. Beware of docking! Trends Pharmacol Sci. 2015 Feb;36(2):78–95.

81. Tintori C, Magnani M, Schenone S, Botta M. Docking, 3D-QSAR studies and in silico ADME prediction on c-Src tyrosine kinase inhibitors. Eur J Med Chem. 2009 Mar;44(3):990–1000.

82. Bairy SK, Suneel Kumar BVS, Bhalla JUT, Pramod AB, Ravikumar M. Three-dimensional quantitative structure-activity relationship studies on c-Src inhibitors based on different docking methods. Chem Biol Drug Des. 2009 Apr;73(4):416–27.

83. Cao R, Mi N, Zhang H. 3D-QSAR study of c-Src kinase inhibitors based on docking. J Mol Model. 2010 Feb;16(2):361–75.

84. Patil R, Das S, Stanley A, Yadav L, Sudhakar A, Varma AK. Optimized hydrophobic interactions and hydrogen bonding at the target-ligand interface leads the pathways of drug-designing. PloS One. 2010 Aug 16;5(8):e12029.

85. Chaudhry Q, Piclin N, Cotterill J, Pintore M, Price NR, Chrétien JR, et al. Global QSAR models of skin sensitisers for regulatory purposes. Chem Cent J. 2010;4(Suppl 1):S5.

86. Fang DQ, Wu WJ, Zhang R, Zeng GH, Zheng KC. Theoretical studies of QSAR and molecular design on a novel series of ethynyl-3-quinolinecarbonitriles as SRC inhibitors. Chem Biol Drug Des. 2012 Jul;80(1):134–47.

87. Yu X. Prediction of Depuration Rate Constants for Polychlorinated Biphenyl Congeners. ACS Omega. 2019 Sep 24;4(13):15615–20.

88. Ding F, Wang Z, Yang X, Shi L, Liu J, Chen G. Development of classification models for predicting chronic toxicity of chemicals to *Daphnia magna* and *Pseudokirchneriella subcapitata*. SAR QSAR Environ Res. 2019 Jan 2;30(1):39–50.

89. Vilar S, Ferino G, Quezada E, Santana L, Friedman C. Predicting monoamine oxidase inhibitory activity through ligand-based models. Curr Top Med Chem. 2012;12(20):2258–74.

90. Zanni R, Garcia-Domenech R, Galvez-Llompart M, Galvez J. Alzheimer: A Decade of Drug Design. Why Molecular Topology can be an Extra Edge? Curr Neuropharmacol. 2018;16(6):849–64.

91. Zhang Y-H, Xia Z-N, Yan L, Liu S-S. Prediction of placental barrier permeability: a model based on partial least squares variable selection procedure. Mol Basel Switz. 2015 May 7;20(5):8270–86.

92. Varmuza K, Filzmoser P, Dehmer M. Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS. Comput Struct Biotechnol J. 2013;5:e201302007.

93. Masand VH, Mahajan DT, Alafeefy AM, Bukhari SNA, Elsayed NN. Optimization of antiproliferative activity of substituted phenyl 4-(2-oxoimidazolidin-1-yl) benzenesulfonates: QSAR and CoMFA analyses. Eur J Pharm Sci. 2015 Sep;77:230–7.

94. Birck MG, Campos LJ, Melo EB de. Estudo computacional de 1h-imidazol-2-il-pirimidina-4,6-diaminas para a identificação de potenciais precursores de novos agentes antimaláricosf06[C-N]. Quím Nova [Internet]. 2016 [cited 2019 Oct 4]; Available from: http://www.gnresearch.org/doi/10.5935/0100-4042.20160065

95. Baba H, Takahara J, Yamashita F, Hashida M. Modeling and Prediction of Solvent Effect on Human Skin Permeability using Support Vector Regression and Random Forest. Pharm Res. 2015 Nov;32(11):3604–17.

96. Chen C-H, Tanaka K, Funatsu K. Random Forest Approach to QSPR Study of Fluorescence Properties Combining Quantum Chemical Descriptors and Solvent Conditions. J Fluoresc. 2018 Mar;28(2):695–706.

97. Zakariazadeh M, Barzegar A, Soltani S, Aryapour H. Developing 2D-QSAR models for naphthyridine derivatives against HIV-1 integrase activity. Med Chem Res. 2015 Jun;24(6):2485–504.

98. Durgapal J, Bisht N, Alam M, Sharma D, Salman M, Nandi S. QSAR and Structure-Based Docking Studies of Aryl Pyrido[2,3-d]pyrimidin-7(8H)-ones: An Attempt to Anticancer Drug Design. Int J Quant Struct-Prop Relatsh. 2018 Jan;3(1):43–73.

99. Evelyn CR, Biesiada J, Duan X, Tang H, Shang X, Papoian R, et al. Combined Rational Design and a High Throughput Screening Platform for Identifying Chemical Inhibitors of a Ras-activating Enzyme. J Biol Chem. 2015 May 15;290(20):12879–98.

100. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Front Pharmacol [Internet]. 2018 Nov 13 [cited 2019 Aug 21];9. Available from: https://www.frontiersin.org/article/10.3389/fphar.2018.01275/full

101. Kopetz S. Targeting SRC and epidermal growth factor receptor in colorectal cancer: rationale and progress into the clinic. Gastrointest Cancer Res GCR. 2007;1(4 Suppl 2):S37-41.

102. Kaushansky A, Gordus A, Budnik BA, Lane WS, Rush J, MacBeath G. System-wide investigation of ErbB4 reveals 19 sites of Tyr phosphorylation that are unusually selective in their recruitment properties. Chem Biol. 2008 Aug 25;15(8):808–17.

103. Olayioye MA, Beuvink I, Horsch K, Daly JM, Hynes NE. ErbB Receptor-induced Activation of Stat Transcription Factors Is Mediated by Src Tyrosine Kinases. J Biol Chem. 1999 Jun 11;274(24):17209–18.

104. Reactome. Search results for SRC. [Internet]. Available from: https://reactome.org/content/query?q=SRC&species=Homo+sapiens&types=Reaction&types=Pathway&cluster=true

105. Araujo J, Logothetis C. Dasatinib: a potent SRC inhibitor in clinical development for the treatment of solid tumors. Cancer Treat Rev. 2010 Oct;36(6):492–500.

106. IUPHAR/BPS Guide to Pharmacology. erb-b2 receptor tyrosine kinase 4 [Internet]. [cited 2019 Sep 1]. Available from: https://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=1799

107. Lo SH. Focal adhesions: What's new inside. Dev Biol. 2006 Jun;294(2):280–91.

108. Gene Cards. Human Gene Database. KIT gene [Internet]. [cited 2019 Sep 1]. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=KIT

109. Amanchy R, Zhong J, Hong R, Kim JH, Gucek M, Cole RN, et al. Identification of c-Src tyrosine kinase substrates in platelet-derived growth factor receptor signaling. Mol Oncol. 2009 Dec;3(5–6):439–50.

110. McCauley JA, Rudd MT. Hepatitis C virus NS3/4a protease inhibitors. Curr Opin Pharmacol. 2016 Oct;30:84–92.

111. Benzine T, Brandt R, Lovell WC, Yamane D, Neddermann P, De Francesco R, et al. NS5A inhibitors unmask differences in functional replicase complex half-life between different hepatitis C virus strains. Randall G, editor. PLOS Pathog. 2017 Jun 8;13(6):e1006343.

112. Watkins WJ. Evolution of HCV NS5B Non-nucleoside Inhibitors. In Berlin, Heidelberg: Springer Berlin Heidelberg; 2019 [cited 2019 Sep 6]. Available from: http://link.springer.com/10.1007/7355_2018_35

113. Macdonald A. The hepatitis C virus NS5A protein binds to members of the Src family of tyrosine kinases and regulates kinase activity. J Gen Virol. 2004 Mar 1;85(3):721–9.

114. Klinker S, Stindt S, Gremer L, Bode JG, Gertzen CGW, Gohlke H, et al. Phosphorylated tyrosine 93 of hepatitis C virus nonstructural protein 5A is essential for interaction with host c-Src and efficient viral replication. J Biol Chem. 2019 May 3;294(18):7388–402.

115. Baker M. Reproducibility crisis: Blame it on the antibodies. Nature. 2015 May 21;521(7552):274–6.

116. Hunter P. The reproducibility "crisis": Reaction to replication crisis should not stifle innovation. EMBO Rep. 2017;18(9):1493–6.