 doi:10.20944/preprints202102.0147.v1

Article

QSAR models for active substances against *Pseudomonas aeruginosa* using disk-diffusion test data

Cosmin Alexandru Bugeac [1], Robert Ancuceanu [2,*] and Mihaela Dinu [2]

[1] Student, Faculty of Pharmacy, Carol Davila University of Medicine and Pharmacy, 6 Traian Vuia Street, Sector 2, 020956 Bucharest, Romania; bugeac.alexandru.cosmin@gmail.com

[2] Department of Pharmaceutical Botany and Cell Biology, Faculty of Pharmacy, Carol Davila University of Medicine and Pharmacy, 6 Traian Vuia Street, Sector 2, 020956 Bucharest, Romania; robert.ancuceanu@umfcd.ro ; mihaela.dinu@umfcd.ro

* Correspondence: Robert.ancuceanu@umfcd.ro

Abstract: *Pseudomonas aeruginosa* is a Gram-negative bacillus included among the six "ESKAPE" microbial species with an outstanding ability to "escape" currently used antibiotics and developing new antibiotics against it is of the highest priority. Whereas minimum inhibitory concentration (MIC) values against *Pseudomonas aeruginosa* have been used previously for QSAR model development, disk diffusion results (inhibition zones) have not been apparently used for this purpose in the literature, and we decided to explore their use in this sense. We developed multiple QSAR methods using several machine learning algorithms (Support vector classifier, K Nearest Neighbors, Random Forest Classifier, Decision Tree Classifier, AdaBoost Classifier, Logistic Regression, and Naive Bayes Classifier). The main descriptors used in building the models belonged to the families of adjacency matrix, constitutional descriptors, first highest eigenvalue of Burden matrix, centered Moreau-Broto autocorrelation, and averaged and centered Moreau-Broto autocorrelation descriptors. A total of 32 models were built, of which 28 were selected and stacked to create a meta-model. In terms of balanced accuracy, the best performance was provided by KNN, logistic regression, and decision tree classifier, but the ensemble method had slightly superior results in nested cross-validation.

Keywords: Pseudomonas; antimicrobial; QSAR; chemical descriptors; machine-learning; KNN; support vector classifier; AdaBoost

1. Introduction

Pseudomonas aeruginosa is a Gram-negative bacillus, widespread in various environments, from soil to water and from plants to animals [1]. Whereas in healthy people it seldom triggers disease, in patients with a weakened immune system it may quickly proliferate and trigger a range of serious acute and chronic infections, being an opportunistic pathogen [1,2]. It is the critical pathogen responsible for the morbidity and mortality associated with cystic fibrosis, as well as one of the major microbes causing nosocomial infections [3]. It is one of the six "ESKAPE" (*Enterococcus, Staphylococcus, Klebsiella, Acinetobacter, Pseudomonas, Enterobacter*) microbial species, characterized by their remarkable ability to "escape" the usual antibiotics and it belongs to the World Health Organization "critical" list of bacteria for which developing new antibiotics should be the highest priority [4]. Its adaptability and resilience, favored by an abundance of regulatory genes in a large genome, its high number of virulence factors and its ability to withstand most antibiotics places this bacteria among the dreadful microbial pathogens [5]. Although bloodstream infections by *Escherichia coli* and *Klebsiella* spp. are more frequent than those of *P. aeruginosa*, the latter is associated in a consistent manner with higher mortality (23-36%) [6]. Considering the ability of *P. aeruginosa* to escape currently

 doi:10.20944/preprints202102.0147.v1

used antibiotics, there is a strong need of developing new such antibacterial products, active against Gram-negative germs and, particularly, against *P. aeruginosa* [7].

The introduction of antibiotics in therapy has marked unprecedented progress in the control of disease and reduction of mortality in human history, conservative estimates indicating death reductions by 25%-75% for different diseases [8]. The development of new antibiotics has remained a challenge in the last decades, with low yields despite impressive progresses in certain areas of drug development [9]. During the 1990s and 2000, the number of new antibiotic drugs approved by the world's key regulatory agencies has suffered a steep decline [10]. Large pharmaceutical companies tend to abandon their antibiotic research programs and turn their back on developing new antibiotics because of financial risks, modest returns, and low probability of development success [11].

Drug design and development in the in the classic approach has been a toilsome, high-priced, time-consuming, and complex activity [12]. Progress in the computational field has allowed the drug discovery processes to become more efficient and less costly, with a variety of structure-based or ligand-based approaches used in this respect [13]. Among the ligand-guided approaches, QSAR methods are very popular, in this sense being stated that "one would say that nowadays no drug is developed without previous QSAR analyses" [14]. They are computational methods that attempt creating relationships between chemical structure features of a set of compounds and one of their biological activities expressed numerically [15]. The practical applications and uses of QSAR span a wide range, from establishing structural requirements for the prospective ligands to finding new prospective compounds via virtual screening, and to estimation of ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) features of a large number of chemical compounds [16]. Valid QSAR models allow virtual screening of large and very large databases of chemical compounds, resulting in identification with meager costs of chemical compounds with a high potential of being active and satisfying the preconditions of promising drugs [17].

It has been recognized that in assembling a proper training set, it is necessary to cover a wide chemical diversity; public databases, such as ChEMBL or PubChem, are most useful in achieving a diverse chemical space for the modeling exercise [12]. However, as a group of researchers investigating such public databases has stated, "there are no databases where we have not found errors" [18]. In a number of cases, wide variability is observed in activity values reported for the same chemical compound in different experiments, depending on the source of the included reports [19,20]. One of the methods widely used to assess antimicrobial susceptibility is the disk diffusion, because it has a number of advantages (simple to carry out, good reproducibility, inexpensive and without needing sophisticated equipment, easily understood by clinical practitioners, and quite flexible) [21]. In ChEMBL, for *Pseudomonas aeruginosa* as a target, the number of bioactivities based on the disc diffusion method (inhibition zone – IZ- as the "standard type" – 7056 data points as of January 18, 2020) is only second to those based on the minimum inhibitory concentration (MIC), and more extensive than other measurements (such as "activity", "inhibition", "MIC90" etc). Whereas MIC values against *Pseudomonas aeruginosa* have been used previously for QSAR model development, we could find no previous attempts of using IZ measurements for this purpose in the literature and decided to explore their use in this sense. Because "disk diffusion susceptibility testing ... provides only a qualitative result" [21], we used classification and not regression machine learning methods. Thus, we report on QSAR models developed for substances active against *Pseudomonas aeruginosa*, using IZ values from the ChEMBL database.

2. Results

The training data set consisted of 3226 observations, with IZ values varying between 0 and 54 mm. The data set's chemical diversity, estimated through the Tanimoto similarity coefficient computed based on the 166-bit MACCS fingerprints, is represented

 doi:10.20944/preprints202102.0147.v1

graphically in Figure 1. The median value of each compound's median similarity to the others (i.e. median of all column/row medians of the symmetric Tanimoto matrix) was 41.67%.

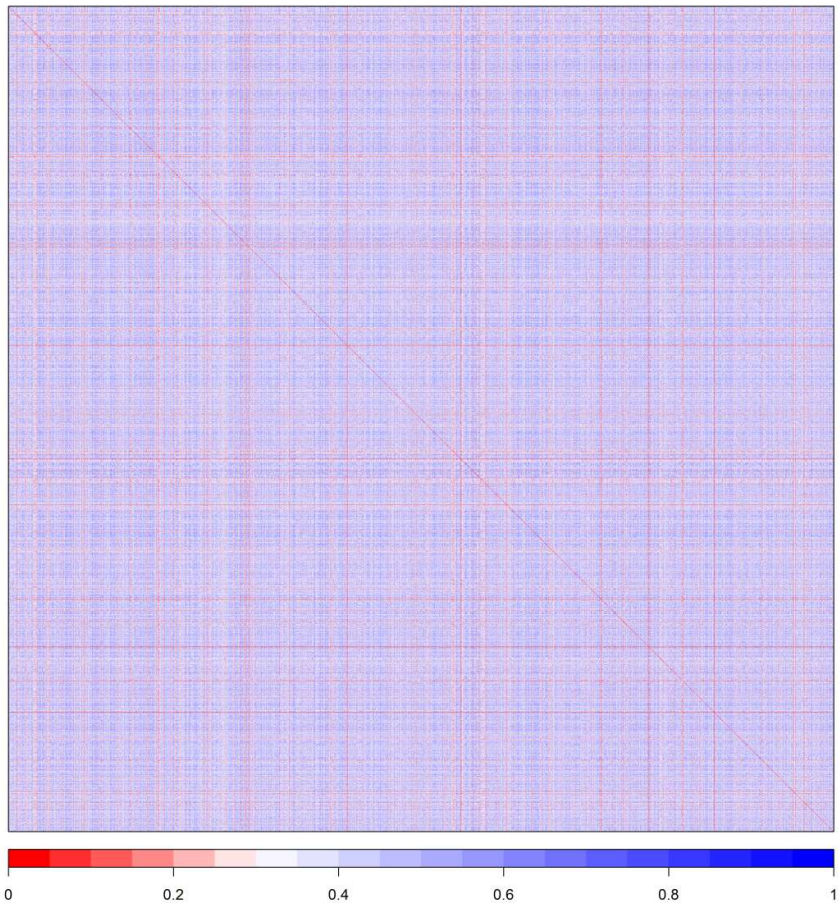

Figure 1. Dissimilarity plot based on the Tanimoto coefficient for the training data set used in this study. Lower values indicate lower dissimilarity (with zero values showing identity – the diagonal line).

Performance

28 models were created (seven classification algorithms and four feature selection methods with hyperparameter tuning in the inner loop of a nested cross-validation process) and stacked to create a meta-model. The latter was built by applying the logistic regression algorithm to the predicted probabilities of the individual models. Different values for the decision threshold were explored, and the best performance was observed for a probability threshold of 0.65. Using a higher value for the threshold led to an increase in positive predictive value with a cost in balanced accuracy and sensitivity. In nested-cross validation, individual models had a balanced accuracy (BA) varying between 48.21% and 79.11%; the stacked model had a mean BA of 72.61% (s.d. 4.61%) (Figure 2, Table S3). The models had good specificity but performed rather poorly in terms of sensitivity, a trade-off we were willing to accept, considering the imbalance between active and inactive compounds in the data set. Specificity varied between 78.09% and 99.83%; the stacked model had a mean of 92.96% (s.d. 0.12%). Sensitivity values ranged between 3.47% and 57.46 %, with best values for the stacked model (mean 56.01%, s.d. 2.06%). From our perspective, a very high specificity coupled with a lower

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2021 doi:10.20944/preprints202102.0147.v1

sensitivity was to be preferred, given the imbalanced nature of the data set. It leads to a higher positive predictive value (a smaller proportion of the active compounds is predicted correctly, but those predicted have a higher probability of being active). PPV ranged between 11.72% and 76.31%; the stacked model had a mean of 38.99% (standard deviation 1.19%). Although other models had higher PPV (e.g., SVM models over 70%), those had a very low sensitivity (a mean sensitivity of 0.4% for one SVM model). Because the dataset consisted of only about 8% compounds classified as "active", the relatively low PPV of 38.99% achieved with the stacked model implies an improvement of over 400% (compared to what is to be expected by mere random labeling). Other attempts of building meta-models were made with different algorithms (random forests, support vector classifier, k nearest neighbors, and decision trees), but the performance of those meta-models (in nested cross-validation) was inferior to the one using logistic regression.

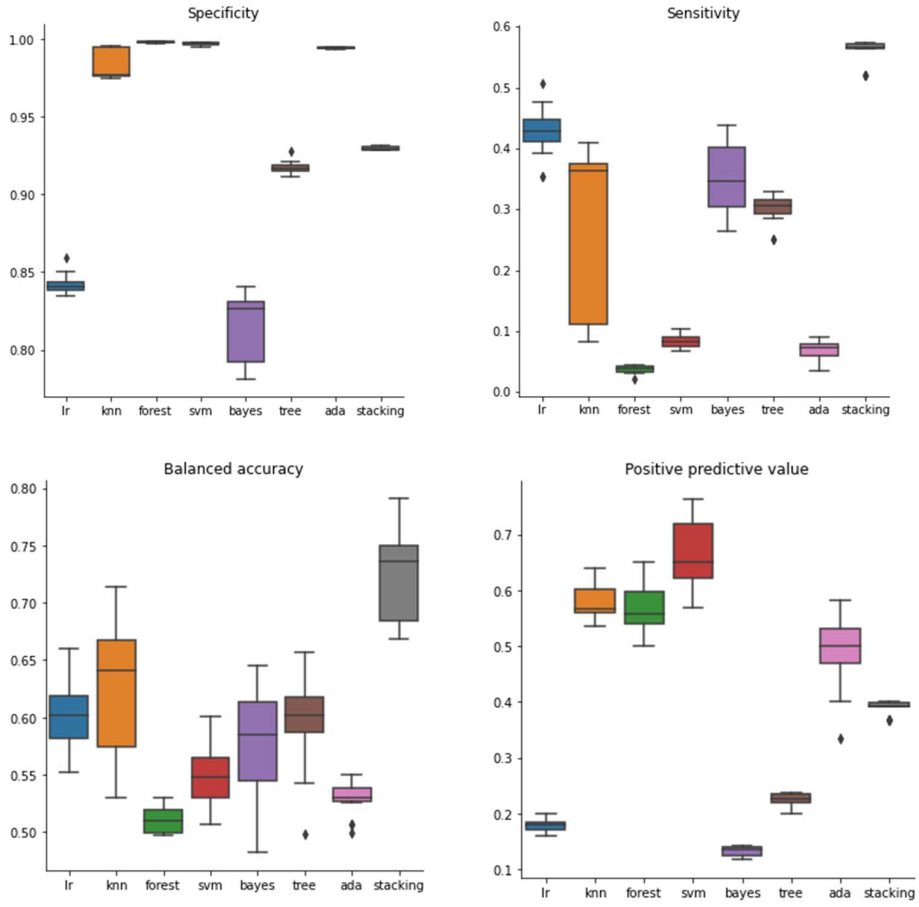

Figure 2. Performance of individual QSAR models and meta-model in the nested-cross validation

Y-randomization

The models' performance during y randomization tests was considerably worse than before randomly shuffling the response variable. Balanced accuracy was close to 50% in all cases compared to the original models, most of which had accuracy close to 60% and higher. These findings indicate that the models' results are not obtained by chance and that there is an underlying relationship between the structure of the compounds and the response variable that the models have identified. Sensitivity dropped to 2% or lower and PPV to less than 10%, a majority of the substances being classified as inactive (Fig. 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2021 doi:10.20944/preprints202102.0147.v1

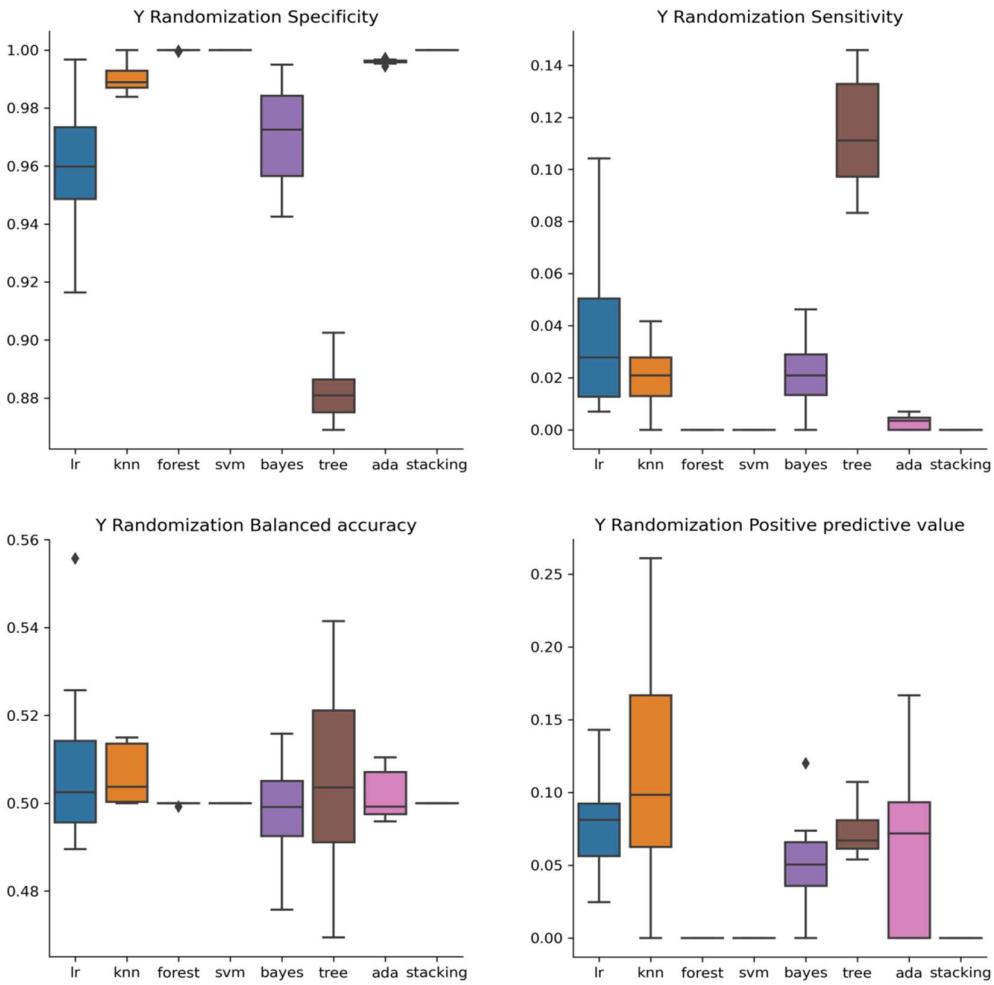

Figure 3. The performance of the QSAR models built with y-randomized data is inferior to that of the QSAR models developed with the non-permuted data set.

External Validation

We performed the external validation on the 1130 compounds from the ChEMBL database (left aside from the beginning for this purpose). The meta-model had a PPV of 40.00% on the external validation data set and balanced accuracy of 76.68% (i.e. 120 compounds were indeed active). A few individual models with higher PPV in nested cross-validation had higher PPV on the external validation set, but their sensitivity was considerably inferior to that of the meta-model (Table S4).

Outliers and applicability domain

Elimination of outliers from the dataset led to better performance of the meta-model: an increase by 2-4% at most in the different performance metrics used (BA, sensitivity, specificity, PPV).

By applying the method of F. Sahigara et al., 2013 [22], some 18-37 (1.59% - 3.27%) of the compounds from the external validation data set were found to be out of the AD of the individual models, and 76 (6.73%) were out of the AD of the meta-model.

 doi:10.20944/preprints202102.0147.v1

Descriptors

Different selection methods used selected different sets of descriptors as important for outcome prediction. VSA_EState1 (VSA EState Descriptor 1 (-inf < x < 4.78)) and BCUTi-1h (first highest eigenvalue of Burden matrix weighted by ionization potential) were the only ones that were selected by all four feature selection algorithms. The importance attributed to these two descriptors was relatively low, though (the highest rank for BCUTi-1h was 5, whereas for VSA_EState1 the highest rank was 20). Descriptors selected by three of the four selection algorithms were EState_VSA2 and ATSC2d (centered Moreau-Broto autocorrelation of lag 2 weighted by sigma electrons), only the latter being first in rank for the mutual information algorithm. The largest importance was attributed by the four selection algorithms to VR3_A (VR3 of adjacency matrix), ATSC2d, ATSC1dv (centered Moreau-Broto autocorrelation of lag 1 weighted by valence electrons), and AATS0dv (averaged Moreau-Broto autocorrelation of lag 0 weighted by valence electrons), respectively. Other features with high importance identified by the selection algorithms belonged to the classes of constitutional descriptors (number of heteroatoms, number of N atoms and number of halogen atoms), and the families of first highest eigenvalue of Burden matrix, centered Moreau-Broto autocorrelation, and averaged and centered Moreau-Broto autocorrelation descriptors (Table S5).

3. Discussion

Pseudomonas aeruginosa is one of the six ESKAPE microbial species that are justifiably worrying for the public health landscape today [4] and a worldwide need to develop new antibiotics active against this bacterial species is increasingly pressing [7]. In the available literature, QSAR models have been reported for substances potentially active against *P. aeruginosa:* local 2D-QSAR models for specific chemical classes (indolylpyrimidines [23], N-octaneamino-4-aminoquinolines [24]), local 2-D [25] and 3D-QSAR models of compounds active against specific protein targets from *P. aeruginosa*, such as the UDP-3-O-(R-3-hydroxymyristoyl)-N-acetylglucosamine deacetylase (LpxC) [26]. A multi-tasking QSAR model (oriented to both predicting anti-*Pseudomonas* activities and ADMET properties of chemical compounds) has also been developed [27]. We have developed a set of global models using individual machine learning algorithms and an ensemble model to predict antimicrobial activity of chemical compounds against *P. aeruginosa* based on the inhibition zone values measured in disk diffusion tests. Probably because of the qualitative character of the measurements in this type of testing, such ChEMBL data sets have not been very appealing to the QSAR community up to date. This reluctance mirrors the hesitation of laboratories to use the disk diffusion method, partly because of reported issues associated with disk quality or the inability of this method to provide an MIC value that can guide clinicians in their therapeutic approach [28]. Our results have shown that models with reasonable performance can be built and employed for virtual screening purposes, although their usefulness may be more restricted than that of those built with MIC, MIC90 or IC50.

One issue in using IZ measurements as an outcome variable is finding an appropriate cut-off level for the classification of compounds in active and inactive. We used a cut-off based on the literature data and clinical breakpoints established by competent organizations (CLSI, EUCAST), indicating that for an important number of currently used antibiotics the threshold between resistance and sensitivity or in some instances "areas of technical uncertainty", is 25 mm or lower [29–31]. However, this cut-off is not without problems, since clinical breakpoints used in the medical practice are substance-specific, but they are only available for a small number of authorized antibiotics, but not for the largest number of compounds tested for scientific purposes outside a clinical setting. Moreover, antimicrobial susceptibility testing breakpoints are established taking into account not only in vitro results of a large number of microbial isolates, but also pharmacokinetic, pharmacodynamics and clinical considerations specific for each

 doi:10.20944/preprints202102.0147.v1

antibiotic [32]. Furthermore, experts in the field of disk diffusion testing emphasize that IZ is dependent on the diffusion rate (through the agar gel) of the tested compound, which in its turn is dependent on certain drug features, such as its size or partition coefficient [21]. These factors make that any uniform cut-off threshold (one not individualized for each compound) should result to some extent in misclassification, and this might explain why the performance of our models, while decent (mean balanced accuracy in nested cross-validation 72.61%), was not particularly impressive.

Given the limitations of a binary cut-off, would not regression modeling be preferable? In theory, the answer would be yes, but as already discussed, the results provided by the disk diffusion method are only qualitative, and usually, "the results have large variations" [33]. Our exploration of regression machine learning models with several algorithms indicated RMSE values around 6 mm in nested cross-validation (with hyperparameter tuning in the inner loop), too large in our view to be of practical relevance.

In terms of balanced accuracy, among individual models, the best performance was provided by KNN, logistic regression, and decision tree classifier, but the ensemble method had superior results to individual models in this respect (the only model with BA generally over 70% in nested cross-validation, as well as on the external test set). KNN models have been successfully applied for other QSAR models, e.g. for different histone deacetylase inhibitors [34,35] or to predict binding affinity for different G-Protein Coupled Receptors (GPCRs) [36]. Logistic regression with regularization, although a relatively simple algorithm, has been shown to have similarly good performance as more sophisticated algorithms in QSAR models [37]. On our dataset the decision trees classifier, somewhat surprisingly, had better results than random forests, although in other cases, the latter has shown superior performance [38]. SVM and AdaBoost classifiers, which for other datasets have been very successful [39–42], in our case, did not perform as well.

The key descriptors used in building the models, selected with the help of four different selection algorithms, belonged to the families of adjacency matrix, constitutional descriptors, first highest eigenvalue of Burden matrix, centered Moreau-Broto autocorrelation, and averaged and centered Moreau-Broto autocorrelation descriptors. Adjacency matrix descriptors have also been previously used [43,44] to predict the antimicrobial effects of different chemical compounds. To a limited extent the Moreau-Broto autocorrelation descriptors were also employed [45], but we could not find in the literature references to the first highest eigenvalue of the Burden matrix in relationship with the antimicrobial activity.

4. Materials and Methods

The dataset

The data set used for model building was obtained from the ChEMBL database by searching “Pseudomonas aeruginosa” in the targets section and downloading all compounds associated with an IZ value. The initial data set consisted of 7056 compounds SMILES chemical structures (provided by ChEMBL), which were converted into 2D structures (sdf) using Bank Formatter 2017. They were filtered using the Bank Cleaner 2017 in order to remove inorganic compounds, mixtures, empty structures, salts, and duplicates; both services were provided by the FAF-Drugs4 program [46,47] and resulted in a dataset of 4520 chemical compounds. The latter were classified into 2 groups, using a cut-off of 25 mm: compounds with an IZ of more than 25mm were labeled as 1 (active) and compounds with IZ under 25mm as 0 (inactive). This dataset was divided in two subsets: one used for model development and performance assessment through nested cross-validation (3390 compounds – Table S1) and the other used for external validation purposes (1130 compounds – Table S2). Structural outliers were identified and removed from the training data set, as recommended in the QSAR literature [48] and as described

 doi:10.20944/preprints202102.0147.v1

below. Among the 3390 compounds (of which 261 were active and 3226 inactive) of the training set, 164 outliers were identified with the isolation forest algorithm and eliminated, leaving a final training data set of 3226 chemical compounds, of which 2986 inactive and 240 active. MACCS fingerprints were computed with the R "RCpi" package [49], Tanimoto coefficients with the help of the "IntClust" R package [50], whereas the dissimilarity plot was generated with the "seriation" R package [51].

Descriptors and feature selection

Using Mordred [52], a Python package, we computed a total of 1826 descriptors. The initial dataset of descriptors was filtered by eliminating those with invalid values, highly correlated (R>0.9), and those with low variance (<1%). We were thus left with a total of 199 descriptors.

Further feature selection was performed on these in order to reduce noise and eliminate redundancy by lowering the dimensions of the data [53]. We used four feature selection methods: two univariate methods (f-test and mutual info classifier), along with recursive feature elimination using a cross-validation loop to select the best number of features and finally feature selection using "SelectFfromModel" with decision trees as the estimator.

For univariate feature selection, the "select k best method" was used, which eliminates all features except those with the highest score computed with a score function, e.g., a function using ANOVA f-value. The F-test assesses the degree of association between two variables by computing the ratio of the two variances, between classes and within classes, as in the classical ANOVA test [54,55]. Mutual information is a tool used to assess the degree of statistical independence among variables, with two fundamental properties: (a) ability to capture not only linear but also non-linear relationships and (b) is invariant under any invertible transformation of the variables used as features in the modeling process [56]. The mutual information classifier used by scikit-learn for feature selection is implemented based on entropy estimation [57,58]. Recursive feature elimination is in theory superior to the previously mentioned two filter methods. It is based on an iterative procedure implying training of a classifier, ranking all the features using a specific criterion, and removing the feature with the lowest rank [59]. "SelectFfromModel" is a versatile meta-transformer implemented by scikit-learn, that removes features based on a threshold that can be predefined or found with the help of built-in heuristics [53].

By using multiple iterations, we determined that the optimal number of features for building our models was 25. Performance improved as we selected more features until this point, after which it began to flatten or even decrease, regardless of the increase in the number of features selected.

Classification algorithms.

The following algorithms were used to build classification models: Support vector classifier, K Nearest Neighbors, Random Forest Classifier, Decision Tree Classifier, Ada Boost Classifier, Logistic Regression, and Naive Bayes Classifier. All algorithms were implemented in Python (version), using the scikit-learn package.

K Nearest Neighbors classifier is based on the notion that a target can be assigned its neighbors' label, using a similarity measure such as the Euclidian distance, Hamming distance etc [60]. The shortest the distance between a number k of data points to the target, the more similar those points are to the target, and a mere majority vote decides the label. The k parameter has a strong sway over the model decision, and its tuning is needed to achieve a balance between over- and under-fitting [61,62]; the values we used ranged between five and eight.

Decision tree is a classification method that labels a data set based on a tree of dichotomic rules [63]. In the learning phase, the rules are derived (tree generation), and in

 doi:10.20944/preprints202102.0147.v1

an accuracy verification phase, random data taken from the training set is tested and rules are adjusted in order to decrease the tree size (tree pruning); in the end the unlabeled data points are classified with the rules thus developed and tested [63,64]. Simplicity, transparency, easiness to understand and to implement [65,66] are key advantages of the decision tree classifier. The key parameter influencing the tree's performance is its maximum depth, as it decides its complexity [67]; in our models, this parameter had values between two and four.

Random forest is an ensemble method that builds multiple decision trees to assign a new data point to a class by a simple majority vote [68]. From statistical and computational perspectives, random forests have multiple strengths, including powerful discriminative abilities, that make them appealing for many applications [69]. The correlation between individual trees is low due to random feature selection of each tree, which results in superior efficiency of the classifier [70]. This classifier's key parameters are the number of trees and their depth (controlled via minimum node size) [71]; our models used between 100 and 300 trees with depths between 10 and 50.

The support vector machine (SVM) is an algorithm widely used for classification purposes, based on identifying the optimal hyperplane to separate observations into the labeled classes [72]. This hyperplane is found with the help of the closest data vectors of the two classes (in the case of binary classification), which are known as "support vectors" [73]. We used the radial basis function kernel and tuned the C and gamma values between 0.5 -10 and 0.1 -1, respectively.

AdaBoost (short for adaptive boosting) is an ensemble method that integrates multiple weak classifiers (models) to build a strong one [74]. Multiple models are constructed sequentially, starting with equal weights for each observation from the training set and then gradually adjusting the weights; after multiple iterations, the results are combined [75]. The key hyperparameters are the number of estimators (we used values between 100 and 400) and the learning rate (we used 0.2 - 0.5).

Gaussian Naive Bayes (GNB) classifier is much faster than other widely used algorithms (such as SVM or even logistic regression) because it hypothesizes a diagonal covariance matrix between variables, thus avoiding the computation of the full covariance matrix [76]. The algorithm assumes Gaussian distribution of the classes, computing z-scores and converting them to p values using the Bayes' theorem, i.e. computing the probability of an observation belonging to class A or class B, given the observed data [77]. The algorithm has a "naïve" approach, not modeling the covariance between features and assuming Gaussian distribution, and the assumptions on which it is based are not necessarily valid. It may work relatively well in a surprising number of cases (but not in all, and not as powerful as more sophisticated algorithms) [77].

Logistic regression is a statistical method often employed in the machine learning applications for binary classification, having simplicity and excellent interpretability as its key advantages [78]. It computes the probability $p = 1/(1 + e^{-t})$, where $t = b0 + b_1x_1 + b_2x_2 + ... + b_nx_n$ [79]. To make predictions, a decision threshold is used (the default value being 0.5, but any other threshold may be used to optimize performance metrics) [80].

In the field of machine learning (unlike statistics), logistic regression is used with a regularization function, meant to avoid overfitting; a cost parameter C is employed for this purpose, for which we used values ranging from 0.1 to 100. We used l2 regularization and 'lbfgs' solver; we also used the 'balanced' mode for the 'class_weight' parameter, this means that the algorithm will automatically adjust class weights inversely proportional to class frequencies.

Hyperparameter tuning was done in the inner loop of nested cross-validation by selecting parameters from a specified parameter grid.

A total of 28 models (from 32 built) were stacked to build a meta-model [81], using eight algorithms and four feature selection methods. The algorithm used for stacking was logistic regression. It used the predicted probabilities from the individual models to

 doi:10.20944/preprints202102.0147.v1

make the final prediction. The stacked model used nested cross-validation as well; for stacking, we used the stacking classifier provided by sci-kit learn.

Performance evaluation

The performance was assessed using nested cross-validation with five folds in the inner loop and five folds in the outer loop. Nested (double) cross-validation outperforms simple (k-fold) cross-validation and hold-one-out validation in terms of both avoiding overfitting and underfitting [82,83].

We computed the following metrics: balanced accuracy (BA), sensitivity, specificity, and positive predictive value (PPV). Our interest was to have higher certainty about the activity of the identified substances (i.e., high specificity) while at the same time preferring not to lose too many potentially active substances (i.e., reasonable sensitivity). All metrics were computed with the classification report method offered by sci-kit learn [53]. Different seed numbers were used to assess the models' performance in the nested cross-validation setting (5 times).

A y randomization test [84] was performed to verify to what extent the results obtained are likely to have been obtained by mere chance. This test was done by randomly scrambling the activity label and rebuilding all the models using the same methods as before; this process was applied a hundred times, and the same metrics were used for performance evaluation. The results of the y randomization test should be considerably worse than the results of the models using the unshuffled data.

Outlier detection and applicability domain

An outlier is defined as a data point that differs significantly from or appears inconsistent with the rest of the data points [85]. The presence of outliers in a data set can be a problem when building a model (as it may unduly influence model parameters resulting in wrongly specified models), but outliers may also contain important information. Therefore a decision to remove them should be well-founded and not for the mere purpose of having models with an apparent better performance [85]. Detection of outliers was carried out using the isolation forest algorithm as implemented in scikit-learn [53] because in a complex benchmarking assessment, this algorithm had the best performance [86].

The applicability domain (AD) is the vector space where a mathematical model (such as a QSAR one) can be applied with reasonable confidence, in other words, the interpolation region [87]. For a new (test) compound to be inside the applicability domain, it has to be sufficiently similar to the compounds from the training set used to develop the model [88]. To determine the AD, the method proposed by F. Sahigara et al., 2013 [22] was used.

External validation

The external validation dataset (1130 compounds) consisted of 1043 inactive and 87 active compounds. The individual models and the meta-model were tested on this data set after determining which compounds were inside the models' AD. The same metrics were used to assess their performance.

Supplementary Materials: Table S1: Training data set, Table S2: External validation data set, Table S3: Mean results in nested cross-validation (5 runs), Table S4: External validation results, Table S5: Descriptors used for model building by feature selection method.

Author Contributions: Conceptualization, RA and MD; methodology, RA and CAB; software, CAB; validation, CAB, RA and MD; formal analysis, CAB and RA.; investigation, CAB and RA; data curation, CAB, MD; writing—original draft preparation, RA and CAB; writing—review and

editing, MD and CAB; visualization, CAB and RA; supervision, MD; funding acquisition, MD. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Data supporting the results in this paper have been provided as supplementary information files. More information can be obtained from the correspondence author.

Conflicts of Interest: The authors declare no conflict of interest. RA 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. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sharma, G.; Rao, S.; Bansal, A.; Dang, S.; Gupta, S.; Gabrani, R. Pseudomonas Aeruginosa Biofilm: Potential Therapeutic Targets. *Biologicals* **2014**, *42*, 1–7, doi:10.1016/j.biologicals.2013.11.001.
2. Azam, M.W.; Khan, A.U. Updates on the Pathogenicity Status of Pseudomonas Aeruginosa. *Drug Discov Today* **2019**, *24*, 350–359, doi:10.1016/j.drudis.2018.07.003.
3. Moradali, M.F.; Ghods, S.; Rehm, B.H.A. Pseudomonas Aeruginosa Lifestyle: A Paradigm for Adaptation, Survival, and Persistence. *Front Cell Infect Microbiol* **2017**, *7*, 39, doi:10.3389/fcimb.2017.00039.
4. Botelho, J.; Grosso, F.; Peixe, L. Antibiotic Resistance in Pseudomonas Aeruginosa – Mechanisms, Epidemiology and Evolution. *Drug Resistance Updates* **2019**, *44*, 100640, doi:10.1016/j.drup.2019.07.002.
5. Oliver, A.; Mulet, X.; López-Causapé, C.; Juan, C. The Increasing Threat of Pseudomonas Aeruginosa High-Risk Clones. *Drug Resist Updat* **2015**, *21–22*, 41–59, doi:10.1016/j.drup.2015.08.002.
6. Paulsson, M.; Granrot, A.; Ahl, J.; Tham, J.; Resman, F.; Riesbeck, K.; Månsson, F. Antimicrobial Combination Treatment Including Ciprofloxacin Decreased the Mortality Rate of Pseudomonas Aeruginosa Bacteraemia: A Retrospective Cohort Study. *Eur J Clin Microbiol Infect Dis* **2017**, *36*, 1187–1196, doi:10.1007/s10096-017-2907-x.
7. Tacconelli, E.; Carrara, E.; Savoldi, A.; Harbarth, S.; Mendelson, M.; Monnet, D.L.; Pulcini, C.; Kahlmeter, G.; Kluytmans, J.; Carmeli, Y.; et al. Discovery, Research, and Development of New Antibiotics: The WHO Priority List of Antibiotic-Resistant Bacteria and Tuberculosis. *The Lancet Infectious Diseases* **2018**, *18*, 318–327, doi:10.1016/S1473-3099(17)30753-3.
8. Spellberg, B. The Future of Antibiotics. *Crit Care* **2014**, *18*, 228, doi:10.1186/cc13948.
9. Pang, Z.; Raudonis, R.; Glick, B.R.; Lin, T.-J.; Cheng, Z. Antibiotic Resistance in Pseudomonas Aeruginosa: Mechanisms and Alternative Therapeutic Strategies. *Biotechnol Adv* **2019**, *37*, 177–192, doi:10.1016/j.biotechadv.2018.11.013.
10. Bettiol, E.; Harbarth, S. Development of New Antibiotics: Taking off Finally? *Swiss Med Wkly* **2015**, *145*, w14167, doi:10.4414/smw.2015.14167.
11. Gajdács, M. The Concept of an Ideal Antibiotic: Implications for Drug Design. *Molecules* **2019**, *24*, doi:10.3390/molecules24050892.
12. Wang, T.; Wu, M.-B.; Lin, J.-P.; Yang, L.-R. Quantitative Structure-Activity Relationship: Promising Advances in Drug Discovery Platforms. *Expert Opin Drug Discov* **2015**, *10*, 1283–1300, doi:10.1517/17460441.2015.1083006.
13. Macalino, S.J.Y.; Billones, J.B.; Organo, V.G.; Carrillo, M.C.O. In Silico Strategies in Tuberculosis Drug Discovery. *Molecules* **2020**, *25*, doi:10.3390/molecules25030665.
14. Andrade, C.H.; Pasqualoto, K.F.M.; Ferreira, E.I.; Hopfinger, A.J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15*, 3281–3294, doi:10.3390/molecules15053281.

15. Aleksandrov, A.; Myllykallio, H. Advances and Challenges in Drug Design against Tuberculosis: Application of in Silico Approaches. *Expert Opin Drug Discov* **2019**, *14*, 35–46, doi:10.1080/17460441.2019.1550482.
16. Halder, A.K.; Moura, A.S.; Cordeiro, MNDS. QSAR Modelling: A Therapeutic Patent Review 2010-Present. *Expert Opin Ther Pat* **2018**, *28*, 467–476, doi:10.1080/13543776.2018.1475560.
17. Dobchev, D.; Karelson, M. Have Artificial Neural Networks Met Expectations in Drug Discovery as Implemented in QSAR Framework? *Expert Opin Drug Discov* **2016**, *11*, 627–639, doi:10.1080/17460441.2016.1186876.
18. Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M.J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P.H. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discov Today* **2011**, *16*, 1019–1030, doi:10.1016/j.drudis.2011.10.005.
19. Williams, A.J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug Discov Today* **2012**, *17*, 685–701, doi:10.1016/j.drudis.2012.02.013.
20. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS ONE* **2013**, *8*, e61007, doi:10.1371/journal.pone.0061007.
21. Jorgensen, J.H.; Turnidge, J.D. Susceptibility Test Methods: Dilution and Disk Diffusion Methods. In *Manual of clinical microbiology*; Jorgensen, J.H., Pfaller, M.A., Carroll, K.C., Landry, M.L., Funke, G., Richter, S.S., Warnock, D.V., Eds.; ASM Press: Washington, DC, 2015; Vol. 1, pp. 1253–1273.
22. 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. *Journal of Cheminformatics* **2013**, *5*, 27, doi:10.1186/1758-2946-5-27.
23. Datar, PA 2D-QSAR Study of Indolylpyrimidines Derivative as Antibacterial against Pseudomonas Aeruginosa and Staphylococcus Aureus: A Comparative Approach. *Journal of Computational Medicine* **2014**, *2014*.
24. Aleksic, I.; Jeremic, J.; Milivojevic, D.; Ilic-Tomic, T.; Šegan, S.; Zlatović, M.; Opsenica, D.M.; Senerovic, L. N-Benzyl Derivatives of Long-Chained 4-Amino-7-Chloro-Quionolines as Inhibitors of Pyocyanin Production in Pseudomonas Aeruginosa. *ACS Chem Biol* **2019**, *14*, 2800–2809, doi:10.1021/acschembio.9b00682.
25. Kadam, RU; Roy, N. Cluster Analysis and Two-Dimensional Quantitative Structure-Activity Relationship (2D-QSAR) of Pseudomonas Aeruginosa Deacetylase LpxC Inhibitors. *Bioorg Med Chem Lett* **2006**, *16*, 5136–5143, doi:10.1016/j.bmcl.2006.07.041.
26. Zuo, K.; Liang, L.; Du, W.; Sun, X.; Liu, W.; Gou, X.; Wan, H.; Hu, J. 3D-QSAR, Molecular Docking and Molecular Dynamics Simulation of Pseudomonas Aeruginosa LpxC Inhibitors. *Int J Mol Sci* **2017**, *18*, doi:10.3390/ijms18050761.
27. Speck-Planche, A.; Cordeiro, M. Computer-Aided Discovery in Antimicrobial Research: In Silico Model for Virtual Screening of Potent and Safe Anti-Pseudomonas Agents. *CCHTS* **2015**, *18*, 305–314, doi:10.2174/1386207318666150305144249.
28. Humphries, R.M.; Kircher, S.; Ferrell, A.; Krause, K.M.; Malherbe, R.; Hsiung, A.; Burnham, C.-A.D. The Continued Value of Disk Diffusion for Assessing Antimicrobial Susceptibility in Clinical Laboratories: Report from the Clinical and Laboratory Standards Institute Methods Development and Standardization Working Group. *J Clin Microbiol* **2018**, *56*, e00437-18, /jcm/56/8/e00437-18.atom, doi:10.1128/JCM.00437-18.
29. Henwood, C.J.; Livermore, D.M.; James, D.; Warner, M.; Pseudomonas Study Group Antimicrobial Susceptibility of Pseudomonas Aeruginosa: Results of a UK Survey and Evaluation of the British Society for

Antimicrobial Chemotherapy Disc Susceptibility Test. *J Antimicrob Chemother* **2001**, *47*, 789–799, doi:10.1093/jac/47.6.789.

30. CLSI *Performance Standards for Antimicrobial Susceptibility Testing. 30th Ed. CLSI Supplement M100*; CLINICAL LABORATORY STAND: Wayne, PA, 2020; ISBN 978-1-68440-066-9.
31. The European Committee on Antimicrobial Susceptibility Testing Breakpoint Tables for Interpretation of MICs and Zone Diameters, Version 11.0 2021.
32. Van, T.T.; Minejima, E.; Chiu, C.A.; Butler-Wu, S.M. Don't Get Wound Up: Revised Fluoroquinolone Breakpoints for *Enterobacteriaceae* and *Pseudomonas Aeruginosa*. *J Clin Microbiol* **2019**, *57*, e02072-18, /jcm/57/7/JCM.02072-18.atom, doi:10.1128/JCM.02072-18.
33. Liu, H.; Taylor, T.H.; Pettus, K.; Trees, D. Assessment of Etest as an Alternative to Agar Dilution for Antimicrobial Susceptibility Testing of Neisseria Gonorrhoeae. *J Clin Microbiol* **2014**, *52*, 1435–1440, doi:10.1128/JCM.02131-13.
34. Cao, G.P.; Arooj, M.; Thangapandian, S.; Park, C.; Arulalapperumal, V.; Kim, Y.; Kwon, Y.J.; Kim, H.H.; Suh, J.K.; Lee, K.W. A Lazy Learning-Based QSAR Classification Study for Screening Potential Histone Deacetylase 8 (HDAC8) Inhibitors. *SAR QSAR Environ Res* **2015**, *26*, 397–420, doi:10.1080/1062936X.2015.1040453.
35. Zhao, L.; Xiang, Y.; Song, J.; Zhang, Z. A Novel Two-Step QSAR Modeling Work Flow to Predict Selectivity and Activity of HDAC Inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2013**, *23*, 929–933, doi:10.1016/j.bmcl.2012.12.067.
36. Luo, M.; Wang, X.S.; Tropsha, A. Comparative Analysis of QSAR-Based vs. Chemical Similarity Based Predictors of GPCRs Binding Affinity. *Mol Inform* **2016**, *35*, 36–41, doi:10.1002/minf.201500038.
37. Korotcov, A.; Tkachenko, V.; Russo, D.P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14*, 4462–4475, doi:10.1021/acs.molpharmaceut.7b00578.
38. Simeon, S.; Jongkon, N. Construction of Quantitative Structure Activity Relationship (QSAR) Models to Predict Potency of Structurally Diversed Janus Kinase 2 Inhibitors. *Molecules* **2019**, *24*, doi:10.3390/molecules24234393.
39. Heikamp, K.; Bajorath, J. Support Vector Machines for Drug Discovery. *Expert Opin Drug Discov* **2014**, *9*, 93–104, doi:10.1517/17460441.2014.866943.
40. Darnag, R.; Minaoui, B.; Fakir, M. QSAR Models for Prediction Study of HIV Protease Inhibitors Using Support Vector Machines, Neural Networks and Multiple Linear Regression. *Arabian Journal of Chemistry* **2017**, *10*, S600–S608, doi:10.1016/j.arabjc.2012.10.021.
41. Goya-Jorge, E.; Giner, RM; Sylla-Iyarreta Veitía, M.; Gozalbes, R.; Barigye, S.J. Predictive Modeling of Aryl Hydrocarbon Receptor (AhR) Agonism. *Chemosphere* **2020**, *256*, 127068, doi:10.1016/j.chemosphere.2020.127068.
42. Guan, D.; Fan, K.; Spence, I.; Matthews, S. Combining Machine Learning Models of in Vitro and in Vivo Bioassays Improves Rat Carcinogenicity Prediction. *Regul. Toxicol. Pharmacol.* **2018**, *94*, 8–15, doi:10.1016/j.yrtph.2018.01.008.
43. Marrero-Ponce, Y.; Marrero, R.M.; Torrens, F.; Martinez, Y.; Bernal, M.G.; Zaldivar, V.R.; Castro, E.A.; Abalo, R.G. Non-Stochastic and Stochastic Linear Indices of the Molecular Pseudograph's Atom-Adjacency Matrix: A Novel Approach for Computational in Silico Screening and "Rational" Selection of New Lead Antibacterial Agents. *J Mol Model* **2006**, *12*, 255–271, doi:10.1007/s00894-005-0024-8.
44. Fassihi, A.; Abedi, D.; Saghaie, L.; Sabet, R.; Fazeli, H.; Bostaki, G.; Deilami, O.; Sadinpour, H. Synthesis, Antimicrobial Evaluation and QSAR Study of Some 3-Hydroxypyridine-4-One and 3-Hydroxypyran-4-One Derivatives. *European Journal of Medicinal Chemistry* **2009**, *44*, 2145–2157, doi:10.1016/j.ejmech.2008.10.022.

45. Shanmugam, G.; Syed, M.; Natarajan, J. 2D-and 3D-QSAR Study of Acyl Homoserine Lactone Derivatives as Potent Inhibitors of Quorum Sensor, SdiA in Salmonella Typhimurium. *International Journal Bioautomation* **2016**, *20*, 441.
46. Lagorce, D.; Sperandio, O.; Baell, JB; Miteva, M.A.; Villoutreix, B.O. FAF-Drugs3: A Web Server for Compound Property Calculation and Chemical Library Design. *Nucleic Acids Research* **2015**, *43*, W200–W207, doi:10.1093/nar/gkv353.
47. Lagorce, D.; Oliveira, N.; Miteva, M.A.; Villoutreix, B.O. Pan-Assay Interference Compounds (PAINS) That May Not Be Too Painful for Chemical Biology Projects. *Drug Discov Today* **2017**, *22*, 1131–1133, doi:10.1016/j.drudis.2017.05.017.
48. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular informatics* **2010**, *29*, 476–488.
49. Cao, D.-S.; Xiao, N.; Xu, Q.-S.; Chen, A.F. Rcpi: R/Bioconductor Package to Generate Various Descriptors of Proteins, Compounds and Their Interactions. *Bioinformatics* **2015**, *31*, 279–281, doi:10.1093/bioinformatics/btu624.
50. Moerbeke, M.V. *IntClust: Integration of Multiple Data Sets with Clustering Techniques*; 2018;
51. Hahsler, M.; Hornik, K.; Buchta, C. Getting Things in Order: An Introduction to the R Package Seriation. *Journal of Statistical Software* **2008**, *25*, 1–34, doi:10.18637/jss.v025.i03.
52. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Journal of Cheminformatics* **2018**, *10*, doi:10.1186/s13321-018-0258-y.
53. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
54. Song, Q.; Jiang, H.; Liu, J. Feature Selection Based on FDA and F-Score for Multi-Class Classification. *Expert Systems with Applications* **2017**, *81*, 22–27, doi:10.1016/j.eswa.2017.02.049.
55. Akay, MF Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis. *Expert Systems with Applications* **2009**, *36*, 3240–3247, doi:10.1016/j.eswa.2008.01.009.
56. Vergara, J.R.; Estévez, P.A. A Review of Feature Selection Methods Based on Mutual Information. *Neural Comput & Applic* **2014**, *24*, 175–186, doi:10.1007/s00521-013-1368-0.
57. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* **2014**, *9*, e87357, doi:10.1371/journal.pone.0087357.
58. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating Mutual Information. *Phys. Rev. E* **2004**, *69*, 066138, doi:10.1103/PhysRevE.69.066138.
59. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine learning* **2002**, *46*, 389–422.
60. Kramer, O. K-Nearest Neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*; Intelligent Systems Reference Library; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; Vol. 51, pp. 13–23 ISBN 978-3-642-38651-0.
61. Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. *Ann Transl Med* **2016**, *4*, 218, doi:10.21037/atm.2016.03.37.
62. Batista, G.; Silva, D.F.; others How K-Nearest Neighbor Parameters Affect Its Performance. In Proceedings of the Argentine symposium on artificial intelligence; sn, 2009; pp. 1–12.
63. Lavanya, D. Ensemble Decision Tree Classifier For Breast Cancer Data. *International Journal of Information Technology Convergence and Services* **2012**, *2*, 17–24, doi:10.5121/ijitcs.2012.2103.

64. Priyanka, NA; Kumar, D. Decision Tree Classifier: A Detailed Survey. *International Journal of Information and Decision Sciences* **2020**, *12*, 246, doi:10.1504/IJIDS.2020.108141.
65. Feretzakis, G.; Kalles, D.; Verykios, V.S. On Using Linear Diophantine Equations for In-Parallel Hiding of Decision Tree Rules. *Entropy (Basel)* **2019**, *21*, doi:10.3390/e21010066.
66. Climent, M.T.; Pardo, J.; Muñoz-Almaraz, F.J.; Guerrero, M.D.; Moreno, L. Decision Tree for Early Detection of Cognitive Impairment by Community Pharmacists. *Front Pharmacol* **2018**, *9*, 1232, doi:10.3389/fphar.2018.01232.
67. Qian, Y.; Zhou, W.; Yan, J.; Li, W.; Han, L. Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery. *Remote Sensing* **2014**, *7*, 153–168, doi:10.3390/rs70100153.
68. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How Many Trees in a Random Forest? In *Machine Learning and Data Mining in Pattern Recognition*; Perner, P., Ed.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; Vol. 7376, pp. 154–168 ISBN 978-3-642-31536-7.
69. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning*; Zhang, C., Ma, Y., Eds.; Springer US: Boston, MA, 2012 ISBN 978-1-4419-9325-0.
70. Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I. Random Forests and Decision Trees. *International Journal of Computer Science Issues (IJCSI)* **2012**, *9*, 272.
71. Hastie, T.; Tibshirani, R.; Friedman, J. Random forests. In *The elements of statistical learning*; Springer, 2009; pp. 587–604.
72. Pisner, D.A.; Schnyer, DM Support vector machine. In *Machine Learning*; Elsevier, 2020; pp. 101–121 ISBN 978-0-12-815739-8.
73. Ali, L.; Wajahat, I.; Amiri Golilarz, N.; Keshtkar, F.; Bukhari, SAC LDA–GA–SVM: Improved Hepatocellular Carcinoma Prediction through Dimensionality Reduction and Genetically Optimized Support Vector Machine. *Neural Computing and Applications* **2020**, doi:10.1007/s00521-020-05157-2.
74. He, Y.-L.; Zhao, Y.; Hu, X.; Yan, X.-N.; Zhu, Q.-X.; Xu, Y. Fault Diagnosis Using Novel AdaBoost Based Discriminant Locality Preserving Projection with Resamples. *Engineering Applications of Artificial Intelligence* **2020**, *91*, 103631, doi:10.1016/j.engappai.2020.103631.
75. Rahman, S.; Irfan, M.; Raza, M.; Moyeezullah Ghori, K.; Yaqoob, S.; Awais, M. Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. *International Journal of Environmental Research and Public Health* **2020**, *17*, 1082, doi:10.3390/ijerph17031082.
76. Ontivero-Ortega, M.; Lage-Castellanos, A.; Valente, G.; Goebel, R.; Valdes-Sosa, M. Fast Gaussian Naïve Bayes for Searchlight Classification Analysis. *NeuroImage* **2017**, *163*, 471–479, doi:10.1016/j.neuroimage.2017.09.001.
77. Raizada, R.D.S.; Lee, Y.-S. Smoothness without Smoothing: Why Gaussian Naive Bayes Is Not Naive for Multi-Subject Searchlight Studies. *PLoS ONE* **2013**, *8*, e69566, doi:10.1371/journal.pone.0069566.
78. Musa, A.B. Comparative Study on Classification Performance between Support Vector Machine and Logistic Regression. *International Journal of Machine Learning and Cybernetics* **2013**, *4*, 13–24, doi:10.1007/s13042-012-0068-x.
79. Raevsky, O.A.; Grigorev, V.Y.; Yarkov, A.V.; Polianczyk, D.E.; Tarasov, VV; Bovina, EV; Bryzhakina, E.N.; Dearden, J.C.; Avila-Rodriguez, M.; Aliev, G. Classification (Agonist/Antagonist) and Regression "Structure-Activity" Models of Drug Interaction with 5-HT6. *Cent Nerv Syst Agents Med Chem* **2018**, *18*, 213–221, doi:10.2174/1871524918666180827100437.
80. Radovanović, S.; Delibašić, B.; Jovanović, M.; Vukićević, M.; Suknović, M. Framework for Integration of Domain Knowledge into Logistic Regression. In Proceedings of the Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics; ACM: Novi Sad Serbia, June 25 2018; pp. 1–8.

81. Ma, Z.; Wang, P.; Gao, Z.; Wang, R.; Khalighi, K. Ensemble of Machine Learning Algorithms Using the Stacked Generalization Approach to Estimate the Warfarin Dose. *PLOS ONE* **2018**, *13*, e0205872, doi:10.1371/journal.pone.0205872.
82. Cawley, G.C.; Talbot, N.L. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research* **2010**, *11*, 2079–2107.
83. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A.J. Machine Learning Algorithm Validation with a Limited Sample Size. *PLOS ONE* **2019**, *14*, e0224365, doi:10.1371/journal.pone.0224365.
84. 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). *Archives of Toxicology* **2019**, doi:10.1007/s00204-019-02580-w.
85. Ben-Gal, I. Outlier Detection. In: Maimon O., Rokach L., editors. Data Mining and Knowledge Discovery Handbook. Springer-. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer-Verlag: New York, NY, USA, 2005; pp. 131–146.
86. Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses. *Pattern Recognition* **2018**, *74*, 406–421.
87. Gajewicz, A. How to Judge Whether QSAR/Read-across Predictions Can Be Trusted: A Novel Approach for Establishing a Model's Applicability Domain. *Environmental Science: Nano* **2018**, *5*, 408–421, doi:10.1039/C7EN00774D.
88. Grenet, I.; Merlo, K.; Comet, J.-P.; Tertiaux, R.; Rouquié, D.; Dayan, F. Stacked Generalization with Applicability Domain Outperforms Simple QSAR on *in Vitro* Toxicological Data. *Journal of Chemical Information and Modeling* **2019**, *59*, 1486–1496, doi:10.1021/acs.jcim.8b00553.