Article

Not peer-reviewed version

QSAR Regression Models for Predicting HMG-CoA Reductase Inhibition Based on MACCS Molecular Fingerprints and Virtual Screening of Natural Products

Robert Ancuceanu , Patriciu Constantin Popovici , Doina Drăgănescu * , Ștefan Busnatu , Beatrice Elena Lascu , Mihaela Dinu

Posted Date: 8 October 2024

doi: 10.20944/preprints202410.0529.v1

Keywords: HMG-CoA reductase; QSAR; statins; nested cross-validation; virtual screening; Iris germanica; machine learning; feature selection; mlr3; MACCS fingerprints; molecular descriptors doi:10.20944/preprints202410.0529.v1

Article

QSAR Regression Models for Predicting HMG-CoA Reductase Inhibition Based on MACCS Molecular Fingerprints and Virtual Screening of Natural Products

Robert Ancuceanu [1], Patriciu Constantin Popovici [1], Doina Drăgănescu [2,*], Ștefan Busnatu [3], Beatrice Elena Lascu [1] and Mihaela Dinu [1]

[1] Department of Pharmaceutical Botany and Cell Biology, Carol Davila University of Medicine and Pharmacy, Faculty of Pharmacy

[2] Department of Pharmaceutical Physics, Carol Davila University of Medicine and Pharmacy, Faculty of Pharmacy

[3] Department of Cardiology, Carol Davila University of Medicine and Pharmacy; Emergency Hospital "Bagdasar-Arseni", 050474 Bucharest, Romania;

* Correspondence: doina.drăgănescu@umfcd.ro

Abstract: Background/Objectives: HMG-CoA reductase is an enzyme that regulates the initial stage of cholesterol synthesis and its inhibitors are widely used in the treatment of cardiovascular diseases. **Methods**: We have created a set of quantitative structure-activity relationship (QSAR) models for human HMG-CoA reductase inhibitors using nested cross-validation as the primary validation method. To develop the QSAR models, we employed various machine learning regression algorithms, feature selection methods, and fingerprints or descriptor datasets. **Results**: We built and evaluated a total of 300 models, selecting 21 that demonstrated good performance (coefficient of determination, $R^2 \geq 0.70$ or concordance correlation coefficient, CCC ≥ 0.85). Six of these top-performing models met both performance criteria and were used to construct five ensemble models. We identified the descriptors most important in explaining HMG-CoA inhibition for each of the six best-performing models. We used the top models to search through over 220,000 chemical compounds from a large database (ZINC 15) for potential new inhibitors. Only a small fraction (237 out of approximately 220,000 compounds) had reliable predictions with mean pIC50 values ≥8 (IC50 values ≤10 nM). Our svm-based ensemble model predicted IC50 values <10 nM for roughly 0.08% of the screened compounds. We have also illustrated the potential applications of these QSAR models in understanding the cholesterol-lowering activities of herbal extracts, such as those reported for an extract prepared from the *Iris × germanica* rhizome. **Conclusions**: Our QSAR models can accurately predict human HMG-CoA reductase inhibitors, having the potential to accelerate the discovery of novel cholesterol-lowering agents and may also be applied to understand the mechanisms underlying the reported cholesterol-lowering activities of herbal extracts.

Keywords: HMG-CoA reductase; QSAR; statins; nested cross-validation; virtual screening; *Iris germanica*; machine learning; feature selection; mlr3; MACCS fingerprints; molecular descriptors

1. Introduction

The incidence of atherosclerotic cardiovascular disease is on the rise and continues to rank as the top cause of death and disability in industrialized countries. Atherosclerosis can be slowed or even reversed with the use of lipid-lowering agents when the medicines are administered in appropriate regimens, while the plaque is still immature and has not become calcified or fibrotic [1]. Evidence from both primary and secondary prevention studies shows that HMG-CoA reductase inhibitors (also known as statins) lessen the risk of atherosclerotic cardiovascular disease, making them the first-line lipid-lowering agents recommended by various national and international clinical guidelines [2]. HMG-CoA reductase (3-hydroxy-3-methylglutaryl coenzyme A reductase) is an enzyme that catalyzes an initial stage in the biosynthesis of cholesterol. This particular step is the one controlling the overall speed of the entire sequence of reactions involved in cholesterol synthesis [3]. Besides their main effects on HMG-CoA reductase, such inhibitors appear to have a large number of pleiotropic effects, providing cardiovascular protection independent of their effect on cholesterol, by preventing the formation of intermediates in the cholesterol biosynthetic pathway. These effects result in an inhibition of post-translational modifications of intracellular proteins. These changes, in turn, have downstream effects on endothelial, inflammatory, and smooth muscle cells [4]. The pleiotropic effects of statins and their potential therapeutic uses (related to the cholesterol inhibition or their pleiotropic effects) seem to be broad, from anti-inflammatory and immunomodulatory activities [5] to neuroprotective effects [6,7], from anti-tumorigenic and anti-metastatic actions [8,9] to protection against aging [10], from preventing or reducing the risk of osteoporosis [11] to certain effects on the endocrine system [12]. This topic, however, remains controversial, and the true impact of the reduction in these intermediates has not been fully clarified because it frequently corresponds to a simultaneous fall in cholesterol [4].

The currently available statins differ widely in their solubility and pharmacokinetic properties. Some are rather lipophilic (simvastatin, fluvastatin, lovastatin, pitavastatin, and atorvastatin), can easily penetrate biological membranes and tend to be more widely distributed in the body. Others, like pravastatin and a lesser extent rosuvastatin, are more hydrophilic. They stay connected to the polar surface of the membrane and need protein transporters to get into the cell. It is thought that because they are not as widely distributed, they might have less pleiotropic effects. [13]. Whereas approved statins seem often to be similar in their efficacy and safety, there are data suggesting that different statins have different safety profiles (with respect to their muscle-related side effects [14], liver toxicity [15,16], diabetes-risk [17], Alzheimer disease risk [18], drug interactions, etc. [19]) and different efficacy [20]. Therefore, developing new HMG-CoA reductase inhibitors could result in statins with improved or modified efficacy and safety..

QSAR is a computational approach that is based on building models describing the relationship between the biological activity and certain structural properties (descriptors) of ligands that bind to a specific biologic target (or who have a specific biological effect) [21]. Over time, two primary approaches to QSAR have emerged, based on the methods used to build the models. A first, more traditional one, is based on models that are often straightforward, linear, and may be interpreted in terms of physicochemical concepts. A second approach is based on the utilization of machine learning techniques, which are more suited for predicting the relationship between structure and activity in extensive datasets with significant chemical variability [22]. Molecular descriptors can capture broad categories of molecule structure information, such as bulk characteristics, substructure frequency, or more complicated three-dimensional descriptions. To describe the level of complexity for such descriptors, different dimensionalities (levels of complexity) are used, the descriptors being labelled as 1D, 2D, 3D, and 4D [23]. While it is reasonable to assume that 3D models would provide substantially more detail regarding a compound's activity or property, in practice, such models are typically restricted to relatively small series of similar compounds, in order to eliminate conformational uncertainty. On the other hand, 2D molecule representations are commonly used for large datasets. Furthermore, the molecular graphs provided by 2D representations are also useful for interpreting QSAR models via the use of chemical structure information (molecule fragments) [24].

Rajathei et al. (2020) developed a 2D-QSAR model for HMG-CoA reductase inhibitors, but it was based on only 30 pyrrole derivatives of atorvastatin [25]. Moorthy et al. (2015) developed an interesting set of QSAR models, based on both linear regression and classification, using MOE for the calculation of the molecular descriptors (2D and 3D). However, these authors did not report on using the models for virtual screening purposes and their validation was based on the techniques of leave one out (LOO), leave many out (LMO), and bootstrapping (besides randomization and holdout testing) [26]. Nested cross-validation, which is apt to provide a more reliable estimation of model performance and a better control of overfitting was not used in this interesting paper. Moreover, it is not clear from that paper whether the HMG-CoA reductase inhibitors were evaluated on a human or rodent version of the enzyme. Samizo and Kaneko (2023) developed QSAR models using a data set of 833 compounds from the ChEMBL database, but they used a HMG-CoA reductase of rat origin, not of human origin [27]. Zang et al. (2017) built a 3D-QSAR model based on a small sample size of 19 compounds, but targeting not human, but lepidopteran HMG-CoA reductase [28]. Another QSAR model was also built on a small number (n=18) of phthalimide congeners [29]. We report in this paper on a series of QSAR models developed for human HMG-CoA reductase inhibitors, using nested cross-validation as the main validation approach, and using the best performing-models for the virtual screening of over 220,000 chemical compounds from the ZINC 15 database. As a practical application of the models, we have also used them to understand what are the natural compounds responsible for the reported LDL-cholesterol lowering effect of an *Iris* × *germanica* L. extract [30].

2. Results

2.1. Chemical Space Distribution and Diversity of the Compounds in the Training Data Set

The variation of ALogP (a measure of lipophilicity, and indirectly, of membrane permeability [31]) as a function of the molecular weight is represented graphically in Figure 1. The largest density was observed for molecular weights varying between 250 and 500 g mol^{-1} (first and third quartiles corresponded to 257.2 and 455.5, respectively), and for ALogP varying between 1 and 4. For active compounds (defined as having an IC50 < 100 nM), the minimum molecular weight in the data set was 369.4, the maximum 778.1, and the median value was 491.1 g mol^{-1}. ALogP varied between 1.4 and 8.4 for the active compounds, with a median value of 4.9. We compared these data with those for ten statins that were at least partially developed as medicinal products (lovastatin, cerivastatin, atorvastatin, fluvastatin, simvastatin, rosuvastatin, glenvastatin, pravastatin, mevastatin, and pitavastatin) and found that for the latter molecular weight varied between 390.5 and 558.6, with a median value of 422.5 g mol^{-1}. For statins, ALogP ranged between 2.1 and 5.5, with a median value of 4.2.

For the entire data set of HMG-CoA reductase inhibitors (all pairs), the average of the Tanimoto similarity coefficients was 0.59, and the first and third quartiles were 0.52 and 0.59 (Figure 2). For the compounds forming the training, the average of the Tanimoto coefficients was also 0.53, whereas for the testing set, it was slightly higher, 0.59. The Tanimoto coefficients for the whole data set and the training and test subsets indicate a reasonably large chemical diversity for the compounds used in the modeling.

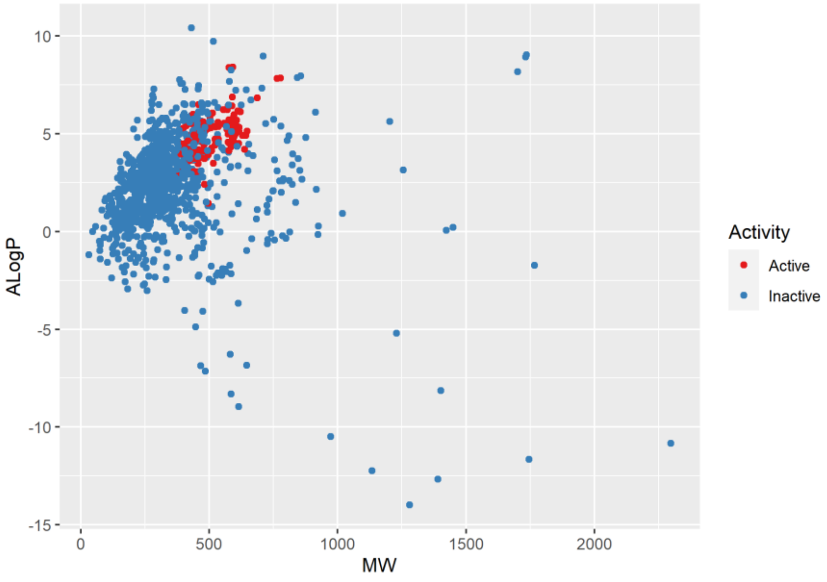


Figure 1. Chemical diversity representation of the HMGCoA-inhibitors data set (chemical space defined by the molecular weight (MW) and AK Ghose – G.M. Crippen logP (ALogP).

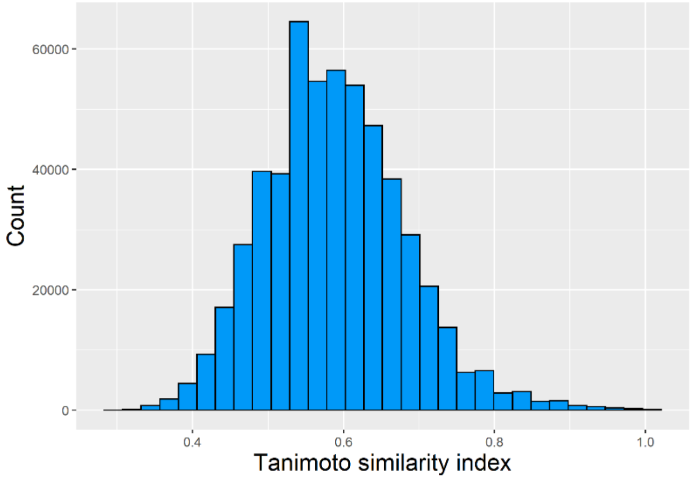


Figure 2. Chemical diversity representation of the HMGCoA-inhibitors data set - histogram of the Tanimoto similarity (based on MACCS fingerprints) for all possible unique pairs of chemical compounds from the data set.

Table 1 displays the number of failures of Lipinski's rule of five for both active and inactive compounds (defined as previously mentioned). One-third of the active compounds had no failure, whereas 19.56% had one failure, 22.46% two failures, and 24.64% three failures. No active compound had four or five failures of Lipinski’s rule. Interestingly, most inactive compounds (71.79%) had no Lipinsky failures.

Table 1. Number of failures of the Lipinski's rule of five for the compounds in the data set.

No. of failures	Active compounds	Inactive compounds
0	46	649
1	27	127
2	31	52
3	34	34
4	0	40
5	0	2

2.1. Regression Models and Their Performance

We built and evaluated a number of 300 models through nested-cross validation (using different machine learning regression algorithms, feature selection methods, and fingerprint or descriptor data sets) (Tables S1-S6). From these, we selected a number of 21 models that performed reasonably well in the nested cross-validation (either $R^2 \geqslant 0.70$ or CCC $\geqslant 0.85$, Table 2). Among the latter, only six met performance conditions ($R^2 \geqslant 0.70$ and CCC $\geqslant 0.85$), and these were selected to also build five ensemble models. To do this, we used the predicted values from the external loop of the nested cross-validation results (using as the random seed the one that gave CCC values closest to the mean value of the five seeds tested; for example, for model no. 4, we used the predicted values for the seed that gave a CCC value of 0.853, as this was the closest to the mean value of 0.851 for that model).

Table 2. Results of nested-cross validation for models with reasonably good performance. Five replicates were used with different random seeds.

No.	Regression algorithm	Descriptor set	Feature selection method	CCC (nested CV, n = 5) mean (s.d.)	R^2 (nested CV, n=5) mean (s.d.)	RMSE (n=5) mean (s.d.)

1	Random forest (“ranger”)	MACCS	“cmim”	0.837	0.707	0.872
				0.840	0.716	0.867
				0.846	0.745	0.835
				0.833	0.698	0.884
				0.825	0.720	0.877
				0.836 (0.008)	**0.717 (0.018)**	**0.867 (0.019)**
2	XGboost	MACCS	Boruta	0.848	0.726	0.848
				0.861	0.744	0.821
				0.840	0.725	0.873
				0.850	0.701	0.870
				0.848	0.741	0.835
				0.849 (0.008)	**0.727 (0.017)**	**0.849 (0.022)**
3	Random forest (“ranger”)	MACCS	Boruta	0.835	0.712	0.890
				0.827	0.696	0.903
				0.833	0.722	0.877
				0.826	0.679	0.916
				0.834	0.707	0.886
				0.831 (0.004)	**0.702 (0.016)**	**0.891 (0.018)**
4	Support vector machines	MACCS	Boruta	0.857	0.743	0.832
				0.853	0.740	0.831
				0.857	0.754	0.815
				0.843	0.708	0.872
				0.845	0.738	0.839
				0.851 (0.007)	**0.737 (0.017)**	**0.838 (0.021)**
5	Gradient boosting machine (“GBM”)	Set2	Boruta	0.858	0.752	0.815
				0.820	0.681	0.942
				0.827	0.702	0.912
				0.830	0.696	0.915
				0.829	0.667	0.926
				0.833 (0.015)	**0.700 (0.032)**	**0.902 (0.050)**
6	Support vector machines	Set2	“jmim”	0.840	0.734	0.854
				0.841	0.727	0.850
				0.850	0.756	0.824
				0.839	0.728	0.849
				0.841	0.746	0.838
				0.842 (0.004)	**0.738 (0.012)**	**0.843 (0.012)**
7	BART	Set2	Gaselect	0.846	0.730	0.858
				0.848	0.739	0.833
				0.854	0.745	0.830
				0.850	0.733	0.847
				0.845	0.689	0.864
				0.849 (0.004)	**0.727 (0.022)**	**0.846 (0.015)**
8	Random forest (“ranger”)	Set2	Gaselect	0.827	0.733	0.848
				0.830	0.730	0.849
				0.823	0.708	0.864
				0.830	0.742	0.850
				0.818	0.723	0.874
				0.826 (0.005)	**0.727 (0.013)**	**0.857 (0.011)**
9	XGboost	Set2	“jmim”	0.832	0.724	0.868
				0.843	0.724	0.852
				0.832	0.706	0.885
				0.830	0.684	0.890

				0.823	0.705	0.901
				0.832 (0.007)	**0.709 (0.017)**	**0.879 (0.019)**
10	BART	Set2	Boruta	0.867	0.764	0.797
				0.825	0.690	0.929
				0.825	0.697	0.917
				0.834	0.707	0.901
				0.829	0.674	0.919
				0.836 (0.018)	**0.704 (0.034)**	**0.893 (0.054)**
11	Rule- and instance-cased regression	Set2	Gaselect	0.837	0.724	0.860
				0.821	0.707	0.891
				0.835	0.717	0.881
				0.843	0.727	0.851
				0.823	0.660	0.893
				0.832 (0.009)	**0.707 (0.027)**	**0.875 (0.019)**
12	Support vector machines	Set2	Gaselect	0.849	0.748	0.821
				0.853	0.754	0.804
				0.840	0.715	0.846
				0.856	0.766	0.804
				0.851	0.757	0.819
				0.850 (0.006)	**0.748 (0.020)**	**0.819 (0.017)**
13	Random forest (“ranger”)	Set3	Gaselect	0.812	0.702	0.873
				0.832	0.720	0.841
				0.826	0.727	0.860
				0.826	0.731	0.862
				0.823	0.717	0.876
				0.824 (0.007)	**0.719 (0.011)**	**0.862 (0.014)**
14	BART	Set4	“jmim”	0.864	0.751	0.821
				0.845	0.710	0.888
				0.858	0.742	0.844
				0.853	0.731	0.858
				0.852	0.730	0.856
				0.854 (0.007)	**0.733 (0.015)**	**0.853 (0.024)**
15	Weighted k-Nearest Neighbor	Set4	Boruta	0.826	0.690	0.923
				0.854	0.740	0.846
				0.865	0.739	0.821
				0.848	0.709	0.874
				0.858	0.737	0.858
				0.850 (0.015)	**0.723 (0.022)**	**0.864 (0.038)**
16	BART	Set4	Gaselect	0.856	0.743	0.833
				0.847	0.726	0.865
				0.845	0.715	0.871
				0.846	0.719	0.877
				0.859	0.745	0.848
				0.851 (0.006)	**0.730 (0.014)**	**0.859 (0.018)**
17	XGboost	Set4	“jmim”	0.835	0.701	0.857
				0.832	0.702	0.896
				0.856	0.750	0.825
				0.830	0.705	0.902
				0.846	0.737	0.844
				0.840 (0.011)	**0.719 (0.022)**	**0.865 (0.033)**

18	Random forest ("ranger")	Set4	Boruta	0.831	0.702	0.861
				0.846	0.748	0.835
				0.857	0.754	0.796
				0.855	0.749	0.829
				0.847	0.738	0.848
				0.847 (0.010)	**0.738 (0.021)**	**0.834 (0.024)**
19	Rule- and instance-cased regression	Set4	Gaselect	0.851	0.726	0.859
				0.813	0.655	0.947
				0.841	0.721	0.882
				0.842	0.715	0.879
				0.837	0.688	0.896
				0.837 (0.014)	**0.701 (0.030)**	**0.893 (0.033)**
20	BART	Set4	Boruta	0.856	0.743	0.833
				0.870	0.757	0.814
				0.868	0.743	0.836
				0.872	0.760	0.805
				0.877	0.768	0.800
				0.869 (0.008)	**0.754 (0.011)**	**0.818 (0.016)**
21	XGboost	Set4	Boruta	0.850	0.737	0.848
				0.849	0.747	0.834
				0.850	0.734	0.838
				0.855	0.738	0.842
				0.843	0.711	0.893
				0.849 (0.004)	**0.733 (0.013)**	**0.851 (0.024)**

The 21 models with reasonably good performance were based on seven different algorithms: random forests (five models), BART (five models), boosting algorithms (Xgboost – four models and GBM – one model), support vector machines (three models), rule- and instance-based regression (two models) and weighted k-nearest neighbor (one model). However, the six best performing models were built with the following algorithms: support vector machines (two models), BART (three models) and weighted k-nearest neighbor (one model). With respect to feature selection algorithms, among the six best performing models three were built with the help of Boruta, two with "gaselect" and one with the "jmim" algorithm. Among the 21 selected models, nine were built with Boruta, seven with "gaselect", four with "jmim", and one with "cmim".

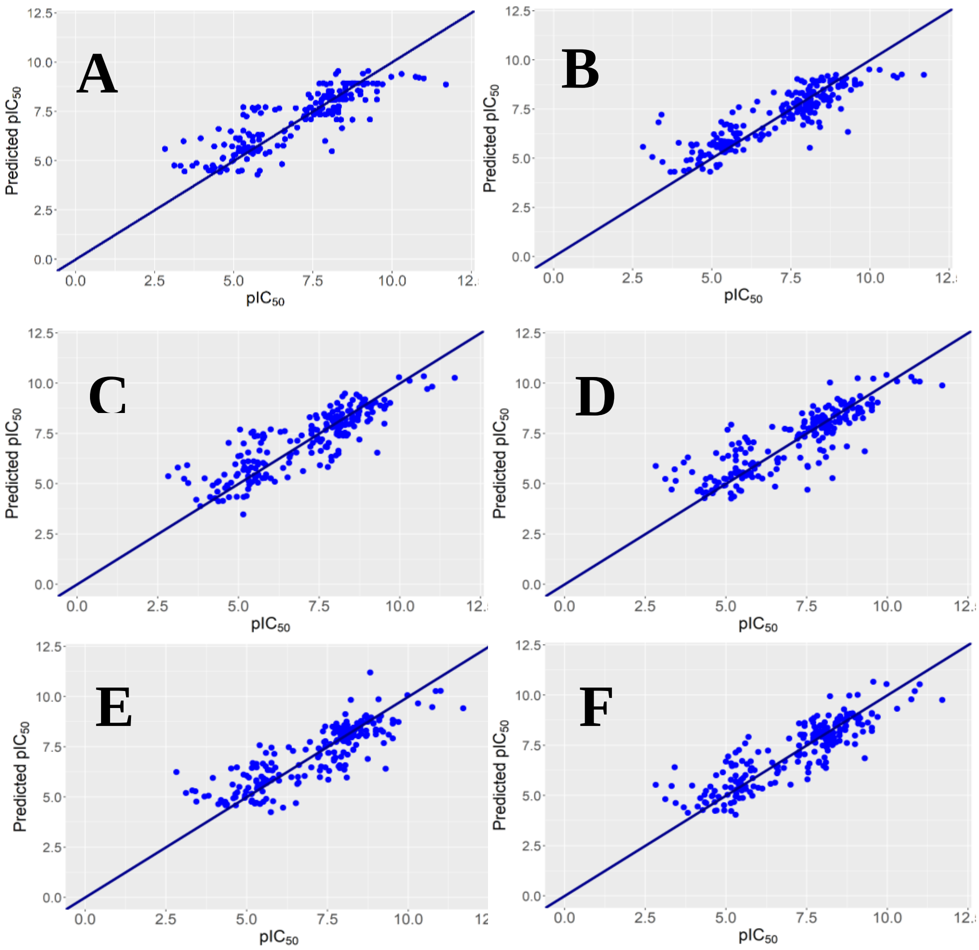


Figure 3. Experimental vs. predicted pIC50 for the six best-performing regression models (which were selected to build the ensemble models). The sloping line represents perfect agreement between actual and predicted values. Points above this line indicate overpredictions, while points below indicate underpredictions.

We built five ensemble models, each using the predicted values in the external loop of the six best performing models (models no. 4, 12, 14-16, and 20 in Table 2) and five different tree-based algorithms: support vector machines, BART, weighted k-nearest neighbor, random forests, and XGboost. The performance of these ensemble models is shown synthetically in Table 3.

Table 3. Performance of the five ensemble models.

Ensemble algorithm	CCC (nested CV)	R2 (nested cross-validation)	RMSE (nested cross-validation)
Support vector machines	0.893	0.798	0.730
BART	0.888	0.789	0.745
KKNN	0.887	0.789	0.750
Random forests	0.889	0.794	0.739
Xgboost	0.883	0.784	0.760

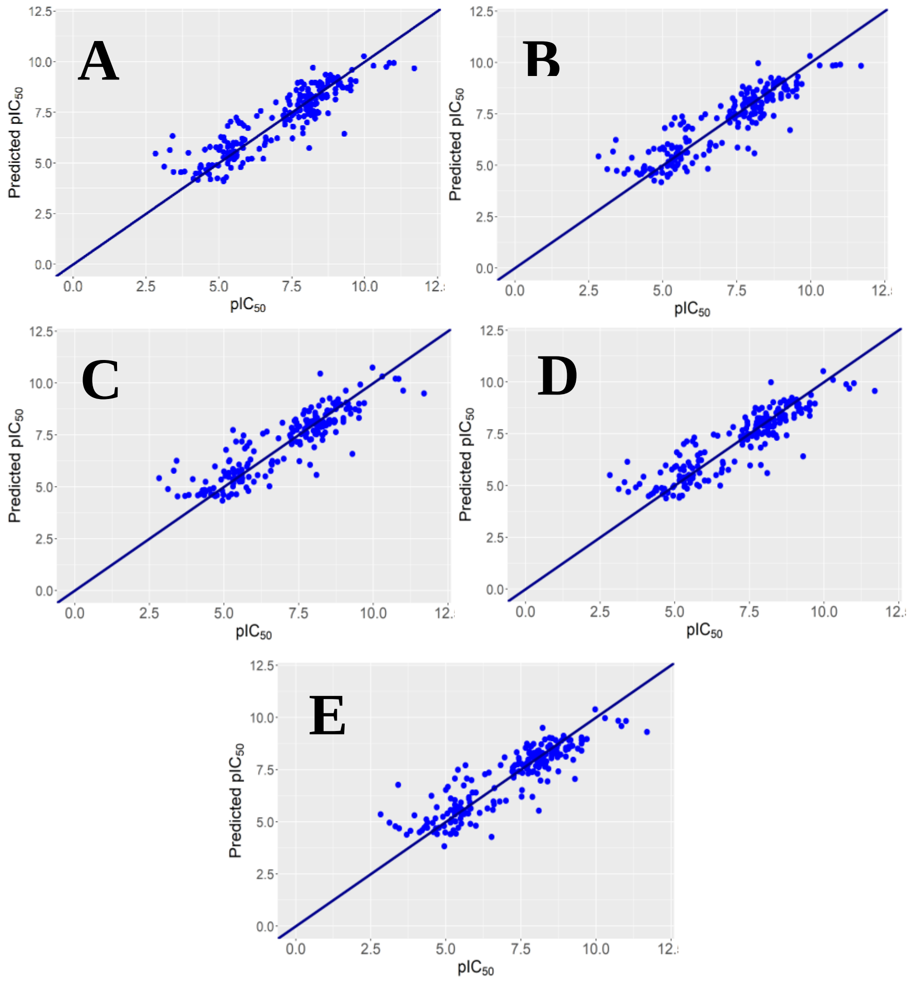


Figure 4. Experimental vs. predicted pIC_{50} for the five ensemble models. A. SVM. B. BART. C. KKNN. D. Random forests. E. Xgboost. The sloping line represents perfect agreement between actual and predicted values. Points above this line indicate overpredictions, while points below indicate underpredictions.

y-Randomization

To assess the significance of our findings, we conducted a permutation test (y-randomization test). We permuted the response variable in the initial data set and thereafter followed the same procedure of feature selection and nested cross-validation as for the authentic data set and computed the same performance metrics (Table 4). Whereas in the feature selection with simple cross-validation a few of the models had reasonable performance, in the nested cross-validation all three metrics indicated overwhelming underperformance. This provides evidence that the observed performance in the original data is not due to chance but rather reflects genuine relationships between the features and the response variable.

Table 4. Performance metrics for three representative data sets and algorithms for which the response variable was permuted 20 times.

Model whose features were randomized	**CCC (nested CV, n = 20) mean (s.d.)**	**R_r^2 (nested CV, n=20) mean (s.d.)**	**RMSE (n=20) mean (s.d.)**	**R_p^2 (for the corresponding model)**
Model 19 in Table 2	0.047 (0.055)	-0.220 (0.077)	1.804 (0.045)	0.803
Model 17 in Table 2	-0.007 (0.034)	-0.113 (0.068)	1.731 (0.027)	0.773
Model 20 in Table 2	0.078 (0.060)	-0.056 (0.040)	1.685 (0.032)	0.781

In the literature, it has been proposed that an R^2_P value should be computed as $R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$, where R^2 is the value of the non-random model and the R_r^2, the mean R2 value of the randomized models. An ideal QSAR model should have R_r^2 close to zero and and R_p^2 close to R^2 for the genuine model [32]. The y-randomization tests and the R_p^2 values have convincingly confirmed that the models selected are not the result of mere chance.

Descriptors Useful for HMGCo-A Inhibition Prediction

We used the DALEX and iml R packages to look at the most important factors that explained the HMGCoA inhibition in the six best models.

The most important MACCS keys identified in the best-performing model used with this type of descriptors (model no. 4 in Table 2), along with their structural significance and impact on activity, are shown in Table 5 and Figures S1 and S2.

Table 5. Important MACCS keys identified in model no. 4 (from Table 2), their corresponding structural patterns and their association with the HMGCoA inhibitory activity.

MACCS Key	Structural Pattern	Association
62	"A$A!A$A" (any atom – ring bond – any atom – chain bond – any atom – ring bond – any atom)	Positive
85	CN(C)C (a closed ring formed by a C-N-C chain)	Positive
105	"A$A($A)$A" (aromatic atom – substructure – aromatic atom)	Negative
22	Three-membered ring system (3M ring)	Relatively strongly negative
65	Carbon and nitrogen united by an aromatic query bond	Positive
145	6M RING > 1 (more than one six-member rings)	Positive
89	OAAAO (two oxygen atoms connected by three other atoms)	Positive
97	NAAAO (a nitrogen atom connected by a sequence of four single bonds to an oxygen atom)	Weakly negative
107	XA(A)A (where X is a halogen and A any atom)	Weakly positive
42	F (a fluorine atom)	Weakly positive

For the model based on the second set of descriptors (2D matrix-based descriptors, 2D autocorrelations, and Burden eigenvalues) (model no. 12 in Table 2), the descriptors most strongly associated with the HMGcoA reductase inhibition are summarized in Table 6 and Figures S3 and S4. For each key descriptor in the model, we have also provided the descriptors that are highly correlated with it, as a key descriptor may simply be a proxy for other highly correlated descriptor or, alternatively, may be more intuitively understood in this way, thus facilitating the interpretation of its contribution. The activity relationship is described on the basis of the partial dependence plots; although such plots are useful for understanding the way in which a feature is associated with the response variable, they may not capture the full complexity of the interactions between the contributing features [33].

Table 6. Key Descriptors Utilized in the Regression Model Constructed using the Set 2 descriptors (2D matrix-based descriptors, 2D autocorrelations, and Burden eigenvalues) (model no. 12 in Table 2). The model employed SVM as the regression algorithm with the genetic algorithm (“gaselect”) as a feature selection method.

Descriptor	Correlation coefficient	Correlated Descriptors	Activity Relationship

	(for other descriptors)		
MATS3e (Moran autocorrelation of lag 3 weighted by Sanderson electronegativity)	r = 0.846	MATS3s (Moran autocorrelation of lag 3 weighted by I-state)	Negative values → higher activity
SpMax_B(p) (Leading eigenvalue from Burden matrix weighted by polarizability)	r >0.91 r>0.80	SpDiam_B(p) (Diameter from Burden matrix weighted by polarizability) SpMax1_Bh(p) (Leading eigenvalue n. 1 of Burden matrix weighted by polarizability) piPC06 (molecular multiple path count of order 6) SpDiam_B(v) (spectral diameter from Burden matrix weighted by van der Waals volume) SpMax_B.v.	Inverted U-shape
VE1sign_B(s) (Coefficient sum of the last eigenvector from Burden matrix weighted by I-State)	N/A	None	Higher values → lower activity
SpMin1_Bh(e) (Smallest eigenvalue n. 1 of Burden matrix weighted by Sanderson electronegativity)	r = 0.99 r>0.87 -0.80	SpMin1_Bh(i) (Smallest eigenvalue n. 1 of Burden matrix weighted by ionization potential) SpMin1_Bh(v) (Smallest eigenvalue n. 1 of Burden matrix weighted by van der Waals volume) SpMin1_Bh(p) (Smallest eigenvalue n. 1 of Burden matrix weighted by polarizability) WiA_D/Dt (average Wiener-like index from distance/detour matrix)	Negative association with an asymmetric inverted U-shape
SM3_X (Spectral moment of order 3 from chi matrix)	r > 0.90 r=0.81	nR03 (Number of 3-membered rings) D/Dtr03 (Distance/detour ring index of order 3) SRW03 (Self-returning walk count of order 3) SM5_X (Spectral moment of order 5 from chi matrix) B04[N-S] (Presence/absence of N – S at topological distance 4) B06[O-S] (Presence/absence of O – S at topological distance 6) F06[O-S] (Frequency of O – S at topological distance 6)	Negative correlation with pIC50

GATS5v
(Geary autocorrelation of lag 5 weighted by van der Waals volume)
	r = -0.903
	r = 0.80
		MATS5p (Moran autocorrelation of lag 5 weighted by polarizability)
		GATS5p (Geary autocorrelation of lag 5 weighted by polarizability)
			Increasing values → higher activity

MATS1p
(Moran autocorrelation of lag 1 weighted by polarizability)
	r = 0.93
	r = 0.87
		MATS1v (Moran autocorrelation of lag 1 weighted by van der Waals volume), MATS1i (Moran autocorrelation of lag 1 weighted by ionization potential)	Inverted U-shaped relationship with activity

JGI5 (Mean topological charge index of order 5)	NA	None	Higher values → higher inhibitory activity
TI2_L (Second Mohar index from Laplace matrix)	r > 0.8 for all but none > 0.9	MSD (Mean square distance index (Balaban)) AECC (Average eccentricity) DECC (Eccentric) ICR (Radial centric information index) MaxTD (Max topological distance) S3K (3-path Kier alpha-modified shape index) IDE (Mean information content on the distance equality) HVcpx (Graph vertex complexity index) WiA_Dz(Z) (Average Wiener-like index from Barysz matrix weighted by atomic number) SpPosA_Dz(Z) (Normalized spectral positive sum from Barysz matrix weighted by atomic number) SpMaxA_Dz(Z) (Normalized leading eigenvalue from Barysz matrix weighted by atomic number) SpMAD_Dz(Z) (Spectral mean absolute deviation from Barysz matrix weighted by atomic number) WiA_Dz(m) (Average Wiener-like index from Barysz matrix weighted by mass) SpPosA_Dz(m) (Normalized spectral positive sum from Barysz matrix weighted by mass) SpMaxA_Dz(m) (Normalized leading eigenvalue from Barysz matrix weighted by mass) SpMAD_Dz(m) (Spectral mean absolute deviation from Barysz matrix weighted by mass)	Higher values → lower inhibitory activity

	WiA_Dz(v) (Average Wiener-like index from Barysz matrix weighted by van der Waals volume) SpPosA_Dz(v) (Normalized spectral positive sum from Barysz matrix weighted by van der Waals volume) SpMaxA_Dz(v) (Normalized leading eigenvalue from Barysz matrix weighted by van der Waals volume) SpMAD_Dz(v) (Spectral mean absolute deviation from Barysz matrix weighted by van der Waals volume) WiA_Dz(e) (Average Wiener-like index from Barysz matrix weighted by Sand

For model no. 14, built with descriptors from set 4 (functional group counts, atom-centred fragments, atom-type E-state indices, and pharmacophore descriptors), using BART as a regression algorithm and “jmim” as a feature selection method, the most important descriptors are summarized in Table 7 and Figures S5 and S6.

Table 7. Key Descriptors Utilized in the Regression Model Constructed using the Set 4 descriptors (functional group counts, atom-centered fragments, atom-type E-state indices, and pharmacophore descriptors) (model no. 14 in Table 2). The model employed BART as the regression algorithm with the 'jmim' as a feature selection method.

Descriptor	**Correlated Descriptors**	**Correlation coefficient(s)**	**Activity Relationship**
C-034 (R–CR..X)	nPyrroles (number of pyrrole rings), N-073 (Ar2NH / Ar3N / Ar2N-Al / R..N..R), SaasN (sum of aasN E-states), NaasN (number of atoms of type aasN)	R=0.89 – 0.90	Higher values → higher activity
SHED_AA (Shannon entropy descriptor, acceptor-acceptor)	SHED_DA (Shannon entropy descriptor, acceptor-acceptor)	r=0.91	Lower values → higher activity
C-003 (a CHR3 group)	nCt (number of total tertiary C), nCrt (number of ring tertiary C)	r=0.88 - 0.99	≤3 → lower activity, 4 or 5 → higher activity
nCrt (number of ring tertiary C)	nCt, C-003, SpMin1_Bh(s) (smallest eigenvalue n. 1 of Burden matrix weighted by I-state)	0.80 – 0.88	0 → higher activity, ≥1 → lower activity
CATS2D_04_AA (CATS2D Acceptor-Acceptor at lag 04)	F04[O-O] (Frequency of O – O at topological distance 4)	r=0.81	≥3 → Stronger activity
NsF (number of atoms of type sF, i.e. -F)	nF (number of fluorine atoms), nX (number of halogen atoms), P_VSA_e_6 (P_VSA-like on Sanderson electronegativity, bin 6),	r>0.9 or r=1.0	Fluorinated → higher activity

	F-084 (F attached to C1(sp2)), SsF (sum of sF E-states), NsF (number of atoms of type sF), F01[C-F] (frequency of C – F at topological distance 1), F02[C-F] (frequency of C – F at topological distance 2), F03[C-F] (frequency of C – F at topological distance 3), F07[C-F] (frequency of C – F at topological distance), F08[C-F] (frequency of C – F at topological distance 8)		
CATS2D_04_DA (CATS2D Donor-Acceptor at lag 04)	CATS2D_04_DD, F04[O-O]	r > 0.80	Higher values → slightly higher inhibition
SHED_AN (Shannon entropy descriptor, acceptor-negative)	SHED_DN, CATS2D_01_DN (CATS2D Donor-Negative at lag 01), CATS2D_00_NN (CATS2D Negative-Negative at lag 00, i.e. number of negative atoms)	r>0.90	Higher values → slightly lower activities
CATS2D_02_AL (CATS2D acceptor-lipophilic at lag 02)	F04[O-O]	r = 0.84	Higher values → slightly higher inhibition
CATS2D_09_DL (CATS2D Donor-Lipophilic at lag 09)	CATS2D_02_DL, CATS2D_07_DL, CATS2D_08_DL	r > 0.80	Lower values → higher inhibitory activity

For the model no. 15, built with descriptors from set 4 (functional group counts, atom-centred fragments, atom-type E-state indices, and pharmacophore descriptors), using KKNN as the regression algorithm and "Boruta" as the feature selection method, the most important descriptors are summarized in Table S7 and Figures S7 and S8.

For the model no. 16 (Table 2), built with descriptors from set 4 (functional group counts, atom-centred fragments, atom-type E-state indices, and pharmacophore descriptors), using BART as the regression algorithm and "gaselect" as the feature selection method, the most important descriptors are summarized in Table S8 and Figures S9 and S10.

For the model no. 20, built with descriptors from set 4 (functional group counts, atom-centred fragments, atom-type E-state indices, and pharmacophore descriptors), using BART as the regression algorithm and "Boruta" as the feature selection method, the most important descriptors are summarized in Table S9 and Figures S11 and S12.

Virtual Screening of a Data Set of Natural Compounds

We have used the six best-performing models (and the best ensemble model (svm)) to virtually screen set of nearly 220,000 chemical compounds (mostly natural) from the ZINC 15 database [34]. The distribution of the mean predicted pIC_{50} values is shown in Figure 5. Only 237 compounds had a mean of reliable pIC50 predictions (i.e., inside the AD) equal to or greater than 8, and 287 had a median of reliable predictions greater than 8 (i.e., had IC_{50} values equal to or lower than 10 nM). Using the svm-based ensemble model, a number of 168 compounds (about 0.08%) had predicted IC_{50} values lower than 10 nM.

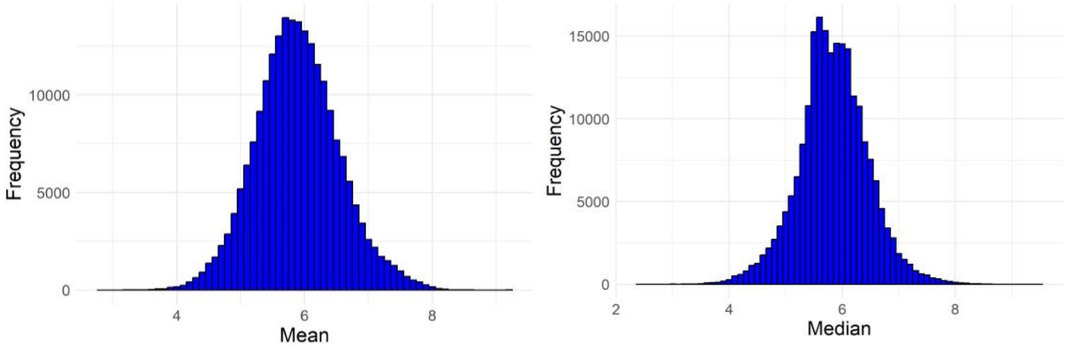


Figure 5. Histograms illustrating the distribution of the predicted mean (left) and median (right) pIC_{50} values for the 219,897 screened natural compounds from the ZINC 15 database.

The distribution of the relative standard deviation (RSD, which expresses how well the predictions for each compound agree with each other) for the virtually screened compounds is shown in Figure 6. The mean and median RSD were approximately 13% (13.27% and 13.74%), the minimum RSD was 0.04%, and the maximum RSD was 63.16%. For most compounds, the predictions were relatively close to one another; for 75% of the predictions, the RSD was less than 17.5%, i.e., there was at least moderate agreement for about three quarters of the data. However, this also means that in about a quarter of the cases, despite the selection of models with similar performance, the predictions differed to a sizeable extent.

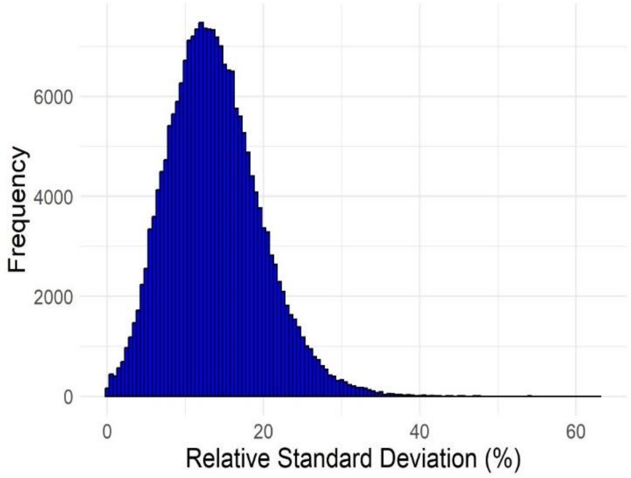


Figure 6. Distribution of the relative standard deviations (RSD) of the predictions made by the six selected models for each of the 219,897 virtually screened compounds.

A total of 81508 compounds (37.07% of all compounds screened) were inside the AD for all six models, and 88046 compounds (40.04%) were inside the AD for five of the six models. At the other extreme, 1758 compounds (0.80%) were outside the AD for all six models, 4550 (2.07%) were inside the AD for a single model, and 9143 (4.16%) were inside the AD for only two models.

Use Case Example for Herbal Extracts

A study by Iqbal Choudhary et al. (2005) found that an ethanolic extract of Iris germanica L. (rhizomes) significantly lowered all lipid components, including LDL-cholesterol [30]. The authors did not identify or discuss the chemical compounds responsible or the mechanism of action, and we were unable to identify further published research to clarify this aspect. We were therefore interested in assessing whether compounds biosynthesised by Iris × germanica have the ability to inhibit HMGCoA reductase. To this end, we downloaded from the Lotus database of natural compounds all the chemical compounds reported to date as having been identified in this species and obtained a data set of 129 compounds that were virtually screened using our selected models in a similar manner to those from ZINC. Seven compounds from this dataset were outside the AD of all six models and 12 compounds were inside the AD of a single model among the six. 60 (46.5%) compounds were inside the AD of all six models, and 38 (29.5%) were inside the AD of five out of the six models. The models predicted only two compounds, both stereoisomers of the same acetylated isoflavone basic structure, to have an IC50 less than 100 nM, while no compounds in this data set had an IC50 less than 10 nM (Figure 7). The two were [(2R,3S,4R,5R,6S)-3,4,5-triacetyloxy-6-[4-(9-acetyloxy-8-oxo-[1,3]dioxolo[4,5-g]chromen-7-yl)phenoxy]oxan-2-yl]methyl acetate and [(2S,3S,4R,5S,6R)-3,4,5-triacetyloxy-6-[4-(9-acetyloxy-8-oxo-[1,3]dioxolo[4,5-g]chromen-7-yl)phenoxy]oxan-2-yl]methyl acetate (Figure 8), with a predicted mean IC50 of 25.07 nM (RSD 11.57%; the median of the predictions for these compounds was about 36.78 nM, and the predicted IC50 made by the svm-based ensemble model was 60.26 nM).

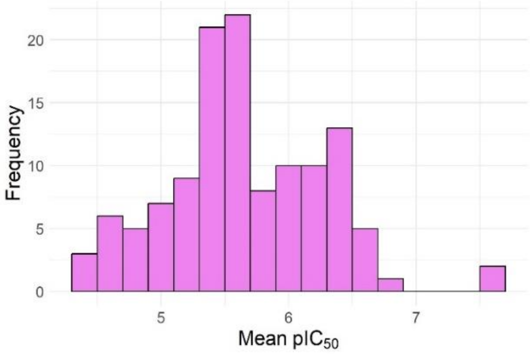
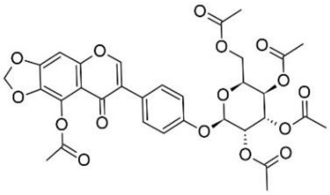
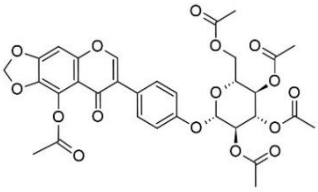


Figure 7. Histogram of the predicted pIC_{50} values for the chemical compounds reported in *Iris* × *germanica* L.



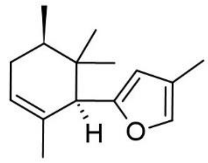
[(2R,3S,4R,5R,6S)-3,4,5-triacetyloxy-6-[4-(9-acetyloxy-8-oxo-[1,3]dioxolo[4,5-g]chromen-7-yl)phenoxy]oxan-2-yl]methyl acetate



[(2S,3S,4R,5S,6R)-3,4,5-triacetyloxy-6-[4-(9-acetyloxy-8-oxo-[1,3]dioxolo[4,5-g]chromen-7-yl)phenoxy]oxan-2-yl]methyl acetate

Figure 8. Two acetylated isoflavonoids from the rhizome of *Iris* × *germanica* predicted to be highly active against HMGCoA-reductase.

The second most active compound of *Iris* × *germanica* predicted by the six models was 4-methyl-2-[(1S,5R)-2,5,6,6-tetramethylcyclohex-2-en-1-yl]furan (Figure 9), a sesquiterpene derivative with a median predicted IC50 of 162 nM and a predicted IC_{50} of 595 nM by the svm-based ensemble model. However, the RSD for the four predictions inside AD in this case was relatively large (27.34%).



4-methyl-2-[(1S,5R)-2,5,6,6-tetramethylcyclohex-2-en-1-yl]furan

Figure 9. A sesquiterpene derivative from *Iris* × *germanica* L. with a predicted IC50 of 37.2 nM.

There were also a number of additional compounds for which the median of the predicted IC_{50} by the individual models or the predicted IC_{50} by the svm-based ensemble model were less than 1 μM and they could also contribute to the observed effect. They are listed in Table 8. Most of them belong to the isoflavonoid group; a few such additional compounds are flavonoids, terpenoids, and xanthonoids.

Table 8. Compounds predicted to have IC50 values under 1 μM (but > 100 nM).

No.	Compound	IC50* (μM)	IC50** (μM)
	Isoflavonoids		
1	irigenin (5,7,3'-trihydroxy-6,4',5'-trimethoxyisoflavone)	0.56	1.37
2	tectoridin (shekanin; 4',5-dihydro-6-methoxy-7-(o-glucoside)isoflavone)	0.84	0.72
3	irisolidone (4'-O-methyltectorigenin)	0.53	1.24
4	iristectorin A	0.89	0.82
5	iristectorigenin B	0.54	1.12
6	homotectoridin	0.87	0.70
7	germanaism A	0.52	
8	irilone 4'-O-glucoside	0.53	0.73
9	germanaism B	0.64	0.80
10	germanaism A	0.52	0.95
11	Kakkalidone (irisolidone 7-O-beta-D-glucoside and its stereoisomers)	0.59	0.75
12	homotectoridin	0.87	
13	irisflorentin	1.73	0.80
14	pratensein 7-O-glucopyranoside	2.08	0.82
15	germanaism G	2.34	0.82
16	3-(3-hydroxy-4,5-dimethoxyphenyl)-7-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxychromen-4-one	1.24	0.69
17	5-hydroxy-3-(3-hydroxy-4,5-dimethoxyphenyl)-7-[(2R,3S,4R,5R,6S)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxychromen-4-one	1.41	0.78
18	germanaism D	2.44	0.85
	flavonoids		
19	isoswertiajaponin	0.83	0.97
20	swertisin (flavocommelitin, 6-C-glucopyranosyl-7-O-methylapigenin)	1.24	0.84
21	isoswertisin (isoflavocommelitin, 7-O-methylvitexin)	1.07	0.85
22	embigenin	1.30	0.66

	terpenoids		
23	iriflorentan (2Z-2-[(2R,3S,4S)-4-hydroxy-3-(hydroxymethyl)-2-(3-hydroxypropyl)-4-methyl-3-[(3E,5E)-4-methyl-6-[(1R,3S)-2,2,3-trimethyl-6-methylidenecyclohexyl]hexa-3,5-dienyl]cyclohexylidene]propanal)	0.62	1.51
24	germanical C (2-[4-hydroxy-3-(hydroxymethyl)-2-(3-hydroxypropyl)-4-methyl-3-[4-methyl-6-(2,5,6,6-tetramethylcyclohex-2-en-1-yl)hexa-3,5-dien-1-yl]cyclohexylidene]propanal)	0.76	1.65
25	irisgermanical B (2-[4-hydroxy-3-(hydroxymethyl)-2-(3-hydroxypropyl)-4-methyl-3-[4-methyl-6-(2,2,3-trimethyl-6-methylidenecyclohexyl)hexa-3,5-dien-1-yl]cyclohexylidene]propanal)	0.62	1.51
	xanthonoids		
26	mangiferin	1.68	0.91
27	irisxanthone	1.84	0.98
28	isomangiferin	2.49	0.94

* Median IC$_{50}$ value for each compound, calculated using predictions only from models where the compound falls within the AD. ** IC$_{50}$ values predicted by the ensemble support vector machine model.

3. Discussion

The QSAR models reported here were built with 2D descriptors and not 3D. Despite the temptation to consider 2D descriptors inferior to 3D ones, previous studies have demonstrated that 2D descriptors could outperform 3D descriptors in compound discrimination across various data sets of biologically active compounds [35]. This was three decades ago, and with the progress made in compound alignment and molecular encoding, today this might not be true anymore. While 3D-QSAR techniques have multiple strengths, the 2D-QSAR approach still has a number of advantages: it is simpler, faster, and best suited for analyzing many compounds and screening large molecular databases [36]. Moreover, the performance of 2D-QSAR models can often be very similar to that of 3D-QSAR models [36]. Moreover, the performance of 2D-QSAR models can often be very similar to that of 3D-QSAR models [37]. Therefore, our focus was on developing a set of valid global 2D QSAR models for virtual screening purposes, using two sets of molecular descriptors: the MACCS keys and a variety of molecular descriptors computed by Alvadesc.

Despite their simplicity, the MACCS keys could be used to build a model that performed similarly to models built with more sophisticated descriptors. This is consistent with Brown and Martin's findings that MACCS keys achieve the highest encoding of content for a variety of properties relevant to interaction with biological targets, such as hydrophobicity, static electricity, steric interactions, dispersion interactions, and intermolecular bonding [35]. In classification models, MACCS (as well as PubChem) fingerprints have been shown to outperform other fingerprints [38], and in our regression models, MACCS yielded results only slightly inferior to the best models constructed with different sets of molecular descriptors.

Among the molecular descriptors, only two pooled sets resulted in models with a reasonably good performance, as defined in this paper: a pooled set consisting of 2D autocorrelations, 2D-matrix based descriptors, and Burden eigenvalues (one model) and another consisting of functional group counts, atom-centered fragments, atom-type E-state indices, and pharmacophore descriptors (four models).

Previously, 2D-autocorrelations have been used successfully to build QSAR models for the bioconcentration factor [39], radical scavenging activity [40], muscle relaxant activity [41], matrix metalloproteinase inhibition, and others [42,43]. This family of descriptors is relatively easy to compute and is based on summations of different autocorrelation functions (at different lags) and encodes information about the topology of the molecule or of certain parts of the molecule, as well as certain atomic properties corresponding to that topology [42]. 2D-matrix based descriptors comprise a heterogenous collection of descriptors calculated using several matrices: adjacency matrix, topological distance matrix, Laplace matrix, chi matrix, reciprocal squared distance matrix, detour matrix, Barysz matrix, and Burden matrix. More specifically, the best-performing models included several descriptors calculated from the chi matrix (SM3_X), Laplace matrix (TI2_L), and Burden matrix (SpMax_B(p), VE1sign_B(s)).

MATS3e (Moran autocorrelation of lag 3 weighted by Sanderson electronegativity) has often been identified in the literature as a "potent" descriptor capable of characterizing a variety of ligand-protein interactions [44]. It has been speculated that compounds with higher values of this descriptor have higher electronegative functionalities that favor the formation of hydrogen-bond interactions with amino acid residues of the target protein active site [45]. In the case of our models, more negative values of MATS3e tended to be associated with a more pronounced inhibitory effect. From a structural point of view, more negative values tend to indicate greater differences in electronegativity between atoms separated by 3 bonds. Its values tended to correlate well with MATS3s (Moran autocorrelation of lag 3 weighted by I-state).

MATS1p, which stands for Moran autocorrelation of lag 1 weighted by polarizability, encodes information about the distribution of polarizability in a molecule, namely between neighboring atoms (lag 1). In the literature, it was reported to correlate positively with the inhibitory activity of imidazole derivatives on glutaminyl cyclase [46] or the inhibitory activity on type I fatty acid synthase [47]. The relationship between MATS1p and the inhibitory activity on HMGCoA reductase in our model was shaped like an upside-down U.

SpMax_B(p) (leading eigenvalue from Bur-den matrix weighted by polarizability) is a less intuitive descriptor, being a leading eigenvalue derived from the Burden matrix (a mathematical instrument of representing the interactions between molecule atoms) and weighted by polarizability. It can be thought of as reflecting the contributions of all atoms in the molecule, and thus reflecting the diversity or similarity of a dataset or database [48]. Its correlation with the inhibitory effect on HMGCoA reductase has an inverted-U-shape (concave-down).

SpMin1_Bh(e) (smallest eigenvalue n. 1 of Burden matrix weighted by Sanderson electronegativity) belongs to the Burden eigenvalues and seems to have been little used in published QSAR models up to date. One study reported that it is negatively correlated with the binding affinity for the bacterial *LasR* protein [49]. We found that it has a negative association with HMGCo-A inhibitors, with an asymmetric inverted U-shape.

VE1sign_B(s) (coefficient sum of the last eigenvector from Burden matrix weighted by I-State) is a 2D-matrix based descriptor that has rarely been reported as important in QSAR studies. In a recent study, it was found to be the second most important descriptor in describing the activity of aromatase inhibitors [50]. Higher (positive) values of this descriptor were linked to more toxicity in a QSAR study that looked at how harmful chemicals were to the springtail *Folsomia candida* [51]. In our model built with Set 2 of descriptors, higher values were predictive of lower activity.

SM3_X (spectral moment of order 3 from chi matrix) was not up to date reported as an important descriptor in the QSAR literature. It provides information on the structural complexity of the molecule and could reflect certain electronic properties of the molecule. While SM3_X is less intuitive and does not lend itself to easy interpretation, at least in our dataset it was highly correlated with SM5_X, as well as with the number of 3-membered rings and the distance/detour ring index of order 3/ SRW03 (self-returning walk count of order 3), meaning that the presence of 3-membered rings (e.g., epoxides, aziridines, or cyclopropane groups) tends to be associated with lower pIC50 (i.e., less active compounds). In a recent study it was found that nR03 (a descriptor highly correlated with SM3_X) tended to decrease the toxicity of chemical compounds on *Daphnia magna* [52]. In the regression model constructed using the Set 2 descriptors (2D matrix-based descriptors, 2D autocorrelations, and Burden eigenvalues) (model no.12 in Table 1), it showed a negative correlation with pIC50.

GATS5v (Geary autocorrelation of lag 5 weighted by van der Waals volume) is a 2D autocorelation descriptor that encodes information about molecular size, shape, steric effects, and distribution of the van der Waals volume across the molecule, specifically for atoms separated by 5 bonds (lag 5). It has been shown to be an important predictor for the antagonistic activity of non-peptide compounds against the CXCR2 chemokine receptor [53] as well as for the antiproliferative activity of 3,4-dihydropyrimidin-2-(1H)-thiones [54]. In our model, higher GATS5v values were associated with higher HMGCoA reductase inhibitory properties

JGI5, or mean topological charge index of order 5, is a type of topological index whose values tend to rise as the molecular structure gets more complicated, with more branching, more ring systems, and more heteroatoms. In the literature, JGI5 has been shown to have a positive association with the antimalarial activity [55] or a stronger antioxidant activity [56]. In our model, higher values of JGI5 were associated with a higher inhibitory activity on HMGCoA reductase.

TI2_L (second Mohar index from Laplace matrix) is calculated as the inverse of the smallest non-null eigenvalue of the Laplace matrix, weighted by the amount of heavy (non-hydrogen) atoms. It ignores the presence of heteroatoms in a molecule, but it is sensitive to structural properties such as branching and the presence of rings. Its value increases with the amount of non-hydrogen atoms present. In a set of molecules of the same size, it discriminates between linear chains (higher values) and branched/cyclic structures (lower values). It has been shown to be useful in the predicting the biodegradability of molecules [57] and the permeability of the placental barrier [58]. Higher values of TI2_L are associated with a lower inhibitory activity, suggesting that some degree of branching or cyclicity is required for the HMGCoA reductase inhibition.

Among the atom-centered fragments, C-034 (R–CR..X, where X is a non-carbon heavy atom, while R is an aliphatic group) and C-003 (a CHR3 group) were shown to correlate with the inhibitory activity on HMGCoA. C-034 correlated well with several other descriptors (see Table 6), including the number of pyrrole rings, which was itself selected as a useful descriptor in other models. Higher values of C-034 were associated with increased activity. C-034 has been reported in the literature to be useful in predicting the glutaminyl cyclase inhibitory activity for imidazole derivatives [46]. C-033 (R–CH..X), has a similar effect as C-034. In a previously published model, it was found to be the most important in predicting herbicidal activity [59], but also in predicting radiosensitizing properties [60]. C-003 was found to be relevant for the binding of small molecules to the active site or the pockets of vasoactive metalloproteases [61] and in predicting the inhibitory activity of biphenylsulfonamides on aggrecanase-1 [55]. In our models, a value of 3 or less was associated with lower activity on HMG-CoA reductase, whereas values of 4 or 5 were associated with higher activity on the enzyme.

C-001 (corresponding to the number of methyl groups, which can induce a certain degree of lipophilicity [62]) has been used in published QSAR models for acetylcholinesterase inhibitors [63]. It was found that both C-001 and the number of pyrrole rings (nPyrroles) were weakly linked to the ability to stop HMG-CoA reductase. Published QSAR models do not appear to have previously selected the number of pyrrole rings among their descriptors. C-002, an atom-centered fragment describing the number of CH_2R_2 fragments, had a sawtooth-like relationship with the HMG-CoA reductase inhibitory activity, with the strongest activity being observed at the lowest value for these fragments. In published QSAR models, this descriptor was used in modeling linear retention indices for essential oil constituents [64] and the antagonistic activity of chemical compounds against the growth hormone secretagogue receptor [65]. C-006 (CH2RX, i.e., the number of carbon atoms bonded to two hydrogen atoms, a heteroatom, and another carbon atom) is a descriptor that has been used in previous research to model the MMP-13 inhibitory activity [66], the CK2 inhibitory activity [67], or the aqueous solubility of chemical compounds [68]. In our models, a higher value for this descriptor tended to be associated with lower inhibitory activity on HMG-CoA-reductase.

H-046 (defined as H attached to C0(sp3) with no X attached to the next C, i.e. a hydrogen atom joined to a carbon atom that is saturated (sp3 hybridized), with the subsequent carbon atom unattached to a heteroatom), is an atom-centred fragment descriptor that has been used to model ligand binding to the 5-HT_6 receptor [69], the inhibitory activity against CDK2 [70], or the PPARγ agonistic activity [71]. In our models, a sawtooth-like curve represented the link between this descriptor and the inhibitory activity of HMG-CoA-reductase, with the highest activity observed at the lowest values. H-053 (defined as H attached to C0(sp3) with 2X attached to the next C; in other words, H - C - C(XX), where: C is an sp3 carbon and C(XX) represents the neighboring carbon with two heteroatoms) is another atom-centred fragment that in previous research has been used in previous research in QSAR modeling of the serotonin 1A and adrenaline α1-adrenergic receptor binding activity [72], of human beta-secretase inhibitors [73], and of the antibacterial activity for pleuromutilin derivatives [74]. In our models, a flattened inverted U-shape was observed for this descriptor in relationship to HMG-CoA reductase inhibitory activity. The O-056 descriptor (number of alcohol fragments) was negatively associated with the HMG-CoA reductase inhibitory activity. It has been used in previously published research to model the odor aroma of wine components [75] or the antimicrobial activity of newly synthesized chemical compounds [76]. The number of pyrimidines (nPyrimidines) correlated positively with the HMG-CoA reductase inhibition. In the past, this descriptor has also been shown to correlate with hepatotoxicity [77] and with the CYP2C9-drug interaction [78].

nCrt (number of ring tertiary C) belong to the functional group counts and was previously reported to be a useful predictor of P-glycoprotein substrates [79]. A value of zero for nCrt was associated with higher HMG-CoA reductase inhibitory activity, whereas values of 1 or higher, were associated with lower activity. NsF (number of atoms of type sF, i.e., single bond fluoride) was also relevant for the HMG-CoA reductase inhibition, with fluorinated molecules having a higher activity. This is an aspect that has already been discussed in the literature, where a fluorine substituent in the pyrrole nucleus of atorvastatin is more effective than other ligands, and fluorine substituents in the hydrophilic side-chain of other statins have stronger inhibitory effects on the target enzyme [80].

nCconj (the number of non-aromatic conjugated carbon atoms, C(sp^2)), is a descriptor that indicates the count of carbon atoms in a molecule that are sp^2 hybridized (have a planar structure with a double bond), are involved in a conjugated system, and are not part of an aromatic ring. It has been shown to be useful in predicting the larvicidal activity of terpenoids against *Culex quinquefasciatus* [81] or the activity against *Trypanosoma cruzi*, the causative agent of the Chagas disease [82]. In our models, a higher number of non-aromatic conjugated carbon atoms was associated with greater inhibitory activity on HMG-CoA reductase.

SaaaC is an E-state descriptor, more specifically the sum of aaaC E-states, i.e., aromatic carbon atoms that have no hydrogen atoms attached and are bonded to three other aromatic atoms; the higher its value, the higher the reactivity and number of those carbon atoms. SaaaC has been shown to be negatively associated with the inhibitory activity against bacterial biofilms [83]. The same type of relationship was observed in our models (lower values of this descriptor are associated with an increase in activity). Conversely, greater values of SaaCH (the sum of aaCH E-states, i.e., all the non-substituted carbon atoms in an aromatic molecule) were correlated with slightly increased activity. This descriptor has previously been used to model algal toxicity [84] and cytotoxicity on the MCF-7 breast cancer cell line [85]. SssCH2 (sum of ssCH2 E-states, i.e., electrotopological states of a methylene group attached to the remainder of the molecule through single bonds) has been useful in modeling the histone deacetylase inhibition activity [86] and in modeling the critical micelle concentration (CMC) for anionic surfactants [85]. A slightly lower level of activity was associated with higher values of this descriptor in our models.

CATS2D_04_AA (CATS2D Acceptor-Acceptor at lag 04) belongs to the sub-block of CATS (Chemically Advanced Template Search) 2D descriptors in the pharmacophore descriptor block. A value of 3 or higher is associated with a stronger inhibitory activity on HMG-CoA reductase. In a recent paper, it was shown that CATS2D_04_AA is an important predictor of blood–brain barrier permeability [87], as well as skin permeability [88] for different substances. CATS2D_04_DA (CATS2D Donor-Acceptor at lag 04) belongs to the same descriptor block and (at least in our data set) was well correlated with CATS2D_04_AA. It has been used in previous studies to construct quantitative structure–toxicity relationship models [89] and in modeling the inhibitory activity of chemical compounds against the MAO-B enzyme [90]. CATS2D_07_DA (CATS2D Donor-Acceptor at lag 07, i.e. at a distance of seven bonds) was used to model the inhibitory activity of O6-methylguanine-DNA methyltransferase, where higher values correlated with lower activity [91]; the same type of relationship was also seen in our models. CATS2D_07_DL (CATS2D Donor-Lipophilic at lag 07) has been used in published QSAR models for *Aedes aegypti* repellents [92], models for antioxidant activity of coumarin derivatives [93], or the anticancer activity of N-(aryl/heteroaryl)-4-(1H-pyrrol-1-yl)-benzenesulfonamide derivatives [94]. In our models, the inhibitory activity against the HMG-CoA reductase was associated with higher values of this descriptor. CATS2D_06_AL (CATS2D Acceptor-Lipophilic at lag 06) is a descriptor that has been little used up to date in QSAR models; we have only identified a model where it was used in the chemometric analysis of drug groups with various pharmacological activities [95] and a model where it was used in modeling the antioxidant effects (TEAC) of chemical compounds [96]. Higher values of this descriptor tended to be associated with lower inhibitory activity on HMG-CoA reductase. CATS2D_03_DL (CATS2D Donor-Lipophilic at lag 03) has been used to model toxicity of chemical compounds against bees [97] and the binding affinity of substances with endocrine disruptor properties [98], whereas CATS2D_09_DL (CATS2D Donor-Lipophilic at lag 09) seems to have not been part of QSAR models published up to date. An increase in activity was associated in our models with lower values of these two descriptors. CATS2D_02_AL (CATS2D acceptor-lipophilic at lag 02, i.e., two bonds apart) is another pharmacophore descriptor that has been used in modeling the biological activities of SGLT2 inhibitors [99] and the multiple endpoint acute toxicity of chemical compounds (higher values, higher toxicities) [100].

First proposed in 2006 [101] Shannon entropy descriptors have not seen extensive use in QSAR models to date. SHED_AN (Shannon entropy descriptor, acceptor-negative) is a descriptor that offers information regarding the spatial arrangement of acceptor and negative atoms inside the molecule. Up to date, it has been used in models predicting the blood-brain barrier permeability [102]. Higher SHED_AN values in our models were linked to marginally lower activity. Similarly, SHED_AA (Shannon entropy descriptor, acceptor-acceptor) is an expression of the diversity or uniformity of the acceptor-acceptor interactions (acceptors being generally electronegative atoms, e.g., halogens, oxygen, and nitrogen). Lower values of SHED_AA were associated with higher HMG-Co-A inhibitory activity in our research.

As shown in the results, the virtual screening of almost 220,000 chemical compounds (mostly natural) from the ZINC 15 database predicted for only 237 compounds a mean of reliable pIC50 predictions (i.e. within the AD) equal to or higher than 8, and 287 compounds a the median of reliable predictions higher than 8 (i.e. had IC_{50} values equal to or lower than 10 nM). Using the svm-based ensemble model, a number of 168 compounds (about 0.08%) had predicted IC_{50} values lower than 10 nM. In a recent paper, Athista et al. (2023) reported on virtual screening to identify HMG-Co-A reductase inhibitors using ligand-protein docking and their predicted hit rate was of 22 natural compounds out of 558 compounds tested, i.e. 3.94% [103].

We have also shown how such QSAR models can be used to improve the understanding of non-clinical experiments performed with herbal extracts where a pharmacological mechanism of the anti-hypercholesterolemiant effect has not been explored. In our use case example, we have identified a number of natural products from *Iris germanica* L. that could explain the ability of an extract obtained from the rhizomes of this species to reduce LDL-cholesterol. Among the compounds predicted to be active by our models was mangiferin. For this compound, the median of the IC50 values predicted by the four best-performing models for which the substance was within the AD was 1.68 μM, whereas experimentally an inhibition constant of 3 ± 0.2 μM was determined [104], which seems to be in fairly good agreement. The ensemble model based on svm estimated an IC50 of 0.90 μM, which is also close to the experimental inhibition value. For irisolidone, our models predicted IC50 values of 0.53 or 1.24 μM, whereas in one experiment, an IC50 of 36 μM was estimated [105]. Such examples, where we have found experimental evidence to verify the predicted activity, tend to confirm the validity of the models and their usefulness in this setting.

4. Materials and Methods

4.1. Data Set

A set of 1170 of human HMG-CoA reductase inhibitors, whose activity was assessed on the basis of their half-maximal inhibitory concentration (IC_{50}), was downloaded from ChEMBL (target ID

CHEMBL402) [106]. The SMILE chemical formulae were carefully checked manually, and inorganic or overly simple compounds (e.g., sodium arsenite, strontium chloride hexahydrate, thioacetamide, etc.), polymers (e.g., macrogol), mixtures, or other compounds without a defined chemical structure were removed from the data set. ChemAxon Standardizer 18.8.0 (ChemAxon, Budapest, Hungary) was used to standardize the chemical structure of the compounds in the data set, using the following operations: stripping salts, neutralization, tautomerization, aromatization, clean 2D, and adding explicit hydrogens (in this order). After standardization, duplicate compounds were removed from the data set, and their IC_{50} values were replaced by the median (as this is more relevant than the mean in the presence of outliers). Compounds available in both acid and salt forms (e.g., lovastatin and lovastatin sodium, maduramicin and maduramicin ammonium) were treated as duplicates, retaining the acid form. This was done using DataWarrior (v. 6.1.0) [107], Flare™ for Academics, v.7.0 (Cresset®, Litlington, Cambridgeshire, UK), and the computing and the programming environment R, v. 4.3.1 [108]. After pre-processing operations, the final data set consisted of 1042 compounds (available with their chemical structures in SMILES notation in Table S1); their IC50 values varied between 0.002 nM and 1,500,000 nM, while their molecular weight varied between 32 g mol^{-1} and 2297 g mol^{-1}. Of the 1042 compounds, numerical IC50 values were available for only 227 compounds, while for the vast majority of the data set, IC50 values were not accessible, and therefore not suitable for use in building regression models. For modelling purposes, the IC50 values were converted to pIC_{50} values by taking the negative logarithm (log10) of the corresponding molar concentration. The 227 compounds were randomly divided into training and test data sets in a 3:1 ratio (170 and 57 compounds, respectively).

Molecular Fingerprint Calculation

The R package “Rcpi” (an open source library) [109] was used to compute MACCS keys (166 bits) under Rstudio, v. 2021.09.1, Build 372 [110]. Molecular fingerprints are a mean of representing molecular structures, encoding the presence (assigning a value of 1) or absence (assigning a value of 0) of certain fragments/substructures in a chemical molecule [111]. MACCS fingerprints were originally intended to be used for substructure searching [112], but were later widely used in QSAR modeling and are still relevant for this purpose [113]. AlvaDesc software [114] was used to compute 3874 2D molecular descriptors, grouped into 18 blocks (constitutional indices, ring descriptors, topological indices, etc.).

Chemical Space Distribution and Diversity

To investigate the diversity and distribution of the data set compounds in the chemical space, we have used two features widely used in the field: molecular weight and atomic logP (AlogP, AK Ghose-G.M. Crippen logP) [115], computed by the R package “Rcpi” [109]. We also investigated the fulfillment of Lipinski's "rule of five" as a criterion of "druggability" or "drug-likeness" for the compounds included in the modeling exercise [116], also using the “Rcpi” R package [109]. We used the average Tanimoto similarity index (computed in R with the "proxy" R package [117]) to assess the diversity of the data set.

Feature Selection, Model Building and Validation

MACCS fingerprints consist of 166 binary features/keys, whereas Alvadesc computes over 4000 of 1D or 2D descriptors. Both are large numbers that need to be reduced in order to build meaningful models, because of the so-called “curse of dimensionality” which if not properly addressed, increases the likelihood of modeling noise and obtaining useless models [118]. It is recognized that, in most cases, only a small subset of all descriptors are likely to carry the information essential for developing good mathematical models with a given data set [22]. Therefore, feature selection is an important step of the QSAR model building process, and an impressive number of methods and algorithms have been developed for this purpose. They are classified as either filter methods (faster and less computationally intensive) or wrapper methods (more robust but more time-consuming and computationally intensive) [119].

For the regression models, we have explored the use of six filter methods through the unified interface “mlr3” [120]: “carscore” (from R package “care” [121]), “correlation”, “cmim” (R package “praznik” [122]), “find_correlation”, “relief” (R package “FSelectorRcpp” [123]), and “information gain” (R package “FSelectorRcpp” [123]). We preceded feature selection by removing constant, quasi-constant (37 features removed), and highly correlated (36 additional features removed) features, using a correlation cut-off of 0.90 and the "FeatureTerminatoR" [124] R package (36 additional features removed). We coupled feature selection with a hyperparameter search and a 10-fold (and in some cases, 5-fold) cross-validation. We used this k-fold cross-validation to enhance the filtering method results, not to validate the modeling exercise (we describe and report external validation and nested cross-validation below). We divided the feature filtering methods into three groups: "carscore," "correlation," and "cmim" for the first group; "find_correlation," "relief," and "information gain" for the second group. We then used the features of the top-performing methods to construct the regression models. To achieve this, we employed the following regression algorithms:

- multiple linear regression
- elastic net regression (“glmnet” R package [125], varying the *alpha* parameter between 0.0001 and 1)
- multivariate adaptive regression splines (“earth” R package [126])
- k nearest neighbors with various kernels (“kknn” [127] and “FNN” R package [128])
- Quinlan M5 rule trees (“Cubist” R package [129] and “RWeka” R package [130]).
- random forests (“ranger” R package [131]), conditional inference trees and conditional random forests (“partykit” [132], “sandwich” [133] and “coin” [134] R packages)
- support vector machines (“e1071” R package [135]) and regularized support vector regression (“LiblineaR” R package [136])
- extreme gradient boosting (“xgboost” R package [137]) and generalized boosting models (“gbm” R package [138])
- Bayesian Additive Regression Trees (BART) (“BART” R package [139]).

For tree-based algorithms (Quinlan M5 rule trees, random forests, extreme gradient boosting, BART), numerical features were used as such (unscaled) in building and assessing the performance of the models. For the remainder of the algorithms used, features were centered and scaled (using the base R function *scale* within the mlr3 pip pipeline).

To estimate the performance of the model-building exercise, we applied a nested-cross validation procedure, using an inner loop of 10 folds, and an outer loop of 10 folds and tuning the hyperparameters for each model inside the inner loop. We have used the root mean squared error (RMSE) as a scoring function for tuning and the nested cross-validation R^2 (true q^2 [140]) as a more easily interpretable performance measure, as well as the concordance correlation coefficient (CCC, computed with the “agRee” R package [141]). We also applied the models built on the external validation data set, using the R2 and the CCC between the true values and those predicted by the models. The CCC was initially proposed by Lawrence I-Kuei Lin in 1989 as a measure of reproducibility [142], but was more recently recommended in the field of QSAR as a more conservative metric having the property of being “a true external validation measure” (using no information from the training data set) [143,144]. We rejected models for which the R^2 values for the test set were lower than 0.70; therefore, for those models we did not perform a nested-cross validation. To control for the possibility of good performance due to chance associated with a certain seed number, we have repeated the nested cross-validation five times for each model, with different random seeds.

To estimate the risk of random correlation, a y-scrambling test (described in the literature as “probably the most powerful validation procedure”) [145] was performed on three of the selected models: the model with the highest R^2 value in the nested cross-validation ($R^2 = 0.75$), one among the models with the lowest acceptable R^2 values ($R^2 = 0.70$), and one with an intermediate level for R^2 value (0.72) (all three models were built with the set 4 of Alvadesc descriptors). For each model the response variable was permuted 20 times, and the whole model building process was repeated from step zero (scaling, feature selection with the relevant methods, nested cross-validation).

To assess feature importance, identify the most important variables associated with the HMGCoA reductase inhibition in the best models, and interpret those models, the "DALEX"[146] and "iml"[147] R packages were used.

Trustworthy QSAR model applications rely on the *applicability domain* (AD), which is defined in large part by the characterization of the interpolation space [148]. We used the apd_similarity() function from the "applicable" R package [149] to estimate the AD for models built using MACCS fingerprints, which are binary variables; we considered compounds with a similarity larger than 20% versus the training set inside AD. For the molecular descriptors (computed with the Alvadesc software), the Isolation Forest algorithm was used, as implemented in the "isotree" R package [150], with a number of features randomly selected for splitting ("ntry") of 10. (The same algorithm is borrowed from the "isotree" by the "applicable" R package.)

In order to perform a virtual high-throughput screening for potential inhibitors of HMGCoA reductase, a library of approximately 220,000 chemical compounds was obtained from the ZINC database. They were downloaded in the SMILES format and the same molecular descriptors as for the training compounds were computed using Alvadesc. We then used the six best performing models to predict the pIC_{50} values for the screening chemical compounds. We assessed whether or not each compound fell within the AD of each model and calculated the median and mean of the predictions that could be trusted based on the AD assessment, as well as the relative standard deviation. The latter allows us to understand how much the predictions have varied between the models whose results were selected for pooling (the molecules being within the the AD of those models). To illustrate a practical application of the models we have also downloaded the chemical structures of all chemical compounds reported as identified in the *Iris germanica* L. species in the Lotus database [151], calculated the molecular descriptors and then virtually screened each compound in a similar way using the ZINC compound dataset.

5. Conclusions

We have developed a set of QSAR models for human HMG-CoA reductase inhibitors, employing nested cross-validation as the primary validation method, and utilizing the top-performing models for the virtual screening of approximately 220,000 chemical compounds from the ZINC 15 database. Active substances (IC50 < 100 nM) exhibited molecular weights from 369.4 to 778.1 g mol−1 and ALogP values ranging from 1.4 to 8.4. In contrast, the ten statins displayed molecular weights between 390.5 and 558.6 g mol−1 and ALogP values from 2.1 to 5.5. A number of 300 models were built using various machine learning regression algorithms, feature selection methods, and fingerprints or descriptor datasets. 21 models were selected for their good performance (R2 ≥ 0.70 or CCC ≥ 0.85), among which six met both performance criteria and were used to construct five ensemble models. Employing y-randomization, while feature selection with basic cross-validation yielded satisfactory performance for some models, nested cross-validation revealed significant underperformance across all performance measures, thus confirming the validity of the selected models. Using the DALEX and iml R packages, the descriptors that were most important in explaining HMGCoA inhibition in the six best-performing models were identified. Only 237 of about 220,000 compounds had a mean pIC50 reliable prediction (i.e., within the AD) of 8 or higher, while 287 of the compounds had a median of 8 or higher for reliable predictions (i.e., IC50 values equal to or lower than 10 nM). A total of 168 substances (or roughly 0.08%) had predicted IC50 values less than 10 nM using the svm-based ensemble model. The developed QSAR models can be successfully applied to understand the compounds involved in cholesterol-lowering activities of herbal extracts, for instance, an extract of *I. germanica* rhizome.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figures S1-S12: feature importance for models 4, 12, 14, 15, 16, and 20; Tables S1-S6: Performance of different models constructed with different fingerprints or descriptors and different data sets; Tables S7-S9: key descriptors utilized in the regression models no. 15, 16, and 20. CSV file: Preprocessed data set and maccs fingerprints; Tab-separated text document: Descriptors computed with Alvandesc for the data set compounds; Tab-separated text document: Descriptors computed with Alvadesc for compounds reported in *Iris germanica* L. The code used to perform the analyses has been made public through Figshare.

Author Contributions: "Conceptualization, R.A., M.D., and P.C.P.; methodology, R.A., P.C.P., D.D.; software, R.A.; validation, R.A., P.C.P. and M.D.; formal analysis, P.C.P, R.A., D.D.; investigation, R.A. and P.C.P.; resources, M.D., S.B.; data curation, P.C.P., R.A.; writing—original draft preparation, P.C.P., R.A., M.D., B.E.L.; writing—review and editing, S.B., D.D.; visualization, R.A.; supervision, M.D.; project administration, M.D.; funding acquisition, P.C.P., S.B. All authors have read and agreed to the published version of the manuscript."

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER, grant number XXX" and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in Figshare.

Acknowledgments: The authors would like to thank the Center for Innovation and e-Health (CiEH) of the Carol Davila University of Medicine and Pharmacy for their valuable contribution in supporting the computational analyses conducted in this study.

Conflicts of Interest: R.A. received consultancy or speakers' fees in the past from UCB, Sandoz, Abbvie, Zentiva, Teva, Laropharm, CEGEDIM, Angelini, Biessen Pharma, Hofigal, AstraZeneca, Stada, Merck. All other authors report no conflicts of interest.

References

1. Toth, P.P.; Banach, M. Statins: Then and Now. *Methodist DeBakey Cardiovascular Journal* **2019**, *15*, 23, doi:10.14797/mdcj-15-1-23.
2. Adhyaru, B.B.; Jacobson, T.A. Safety and Efficacy of Statin Therapy. *Nat Rev Cardiol* **2018**, *15*, 757–769, doi:10.1038/s41569-018-0098-5.
3. Schumacher, M.M.; DeBose-Boyd, R.A. Posttranslational Regulation of HMG CoA Reductase, the Rate-Limiting Enzyme in Synthesis of Cholesterol. *Annu. Rev. Biochem.* **2021**, *90*, 659–679, doi:10.1146/annurev-biochem-081820-101010.
4. Almeida, S.O.; Budoff, M. Effect of Statins on Atherosclerotic Plaque. *Trends in Cardiovascular Medicine* **2019**, *29*, 451–455, doi:10.1016/j.tcm.2019.01.001.
5. Arefieva, T.I.; Filatova, A.Yu.; Potekhina, A.V.; Shchinova, A.M. Immunotropic Effects and Proposed Mechanism of Action for 3-Hydroxy-3-Methylglutaryl-Coenzyme A Reductase Inhibitors (Statins). *Biochemistry Moscow* **2018**, *83*, 874–889, doi:10.1134/S0006297918080023.
6. Saeedi Saravi, S.S.; Saeedi Saravi, S.S.; Arefidoust, A.; Dehpour, A.R. The Beneficial Effects of HMG-CoA Reductase Inhibitors in the Processes of Neurodegeneration. *Metab Brain Dis* **2017**, *32*, 949–965, doi:10.1007/s11011-017-0021-5.
7. Sodero, A.O.; Barrantes, F.J. Pleiotropic Effects of Statins on Brain Cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2020**, *1862*, 183340, doi:10.1016/j.bbamem.2020.183340.
8. Stine, J.E.; Guo, H.; Sheng, X.; Han, X.; Schointuch, M.N.; Gilliam, T.P.; Gehrig, P.A.; Zhou, C.; Bae-Jump, V.L. The HMG-CoA Reductase Inhibitor, Simvastatin, Exhibits Anti-Metastatic and Anti-Tumorigenic Effects in Ovarian Cancer. *Oncotarget* **2016**, *7*, 946–960, doi:10.18632/oncotarget.5834.
9. Ahmadi, M.; Amiri, S.; Pecic, S.; Machaj, F.; Rosik, J.; Łos, M.J.; Alizadeh, J.; Mahdian, R.; Da Silva Rosa, S.C.; Schaafsma, D.; et al. Pleiotropic Effects of Statins: A Focus on Cancer. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2020**, *1866*, 165968, doi:10.1016/j.bbadis.2020.165968.
10. Bahrami, A.; Bo, S.; Jamialahmadi, T.; Sahebkar, A. Effects of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase Inhibitors on Ageing: Molecular Mechanisms. *Ageing Research Reviews* **2020**, *58*, 101024, doi:10.1016/j.arr.2020.101024.
11. Zhou, H.; Xie, Y.; Baloch, Z.; Shi, Q.; Huo, Q.; Ma, T. The Effect of Atorvastatin, 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase Inhibitor (HMG-CoA), on the Prevention of Osteoporosis in Ovariectomized Rabbits. *J Bone Miner Metab* **2017**, *35*, 245–254, doi:10.1007/s00774-016-0750-2.
12. De La Cruz, J.A.; Mihos, C.G.; Horvath, S.A.; Santana, O. The Pleiotropic Effects of Statins in Endocrine Disorders. *EMIDDT* **2019**, *19*, 787–793, doi:10.2174/1871530319666190329115003.
13. Climent, E.; Benaiges, D.; Pedro-Botet, J. Hydrophilic or Lipophilic Statins? *Front. Cardiovasc. Med.* **2021**, *8*, 687585, doi:10.3389/fcvm.2021.687585.

14. Montastruc, J. Rhabdomyolysis and Statins: A Pharmacovigilance Comparative Study between Statins. *Brit J Clinical Pharma* **2023**, *89*, 2636–2638, doi:10.1111/bcp.15757.
15. Ma, M.-M.; Xu, Y.-Y.; Sun, L.-H.; Cui, W.-J.; Fan, M.; Zhang, S.; Liu, L.; Wu, L.-Z.; Li, L.-C. Statin-Associated Liver Dysfunction and Muscle Injury: Epidemiology, Mechanisms, and Management Strategies. *International Journal of General Medicine* **2024**, 2055–2063.
16. Clarke, A.T.; Johnson, P.C.D.; Hall, G.C.; Ford, I.; Mills, P.R. High Dose Atorvastatin Associated with Increased Risk of Significant Hepatotoxicity in Comparison to Simvastatin in UK GPRD Cohort. *PLoS ONE* **2016**, *11*, e0151587, doi:10.1371/journal.pone.0151587.
17. Thakker, D.; Nair, S.; Pagada, A.; Jamdade, V.; Malik, A. Statin Use and the Risk of Developing Diabetes: A Network Meta-analysis. *Pharmacoepidemiology and Drug* **2016**, *25*, 1131–1149, doi:10.1002/pds.4020.
18. Sinyavskaya, L.; Gauthier, S.; Renoux, C.; Dell'Aniello, S.; Suissa, S.; Brassard, P. Comparative Effect of Statins on the Risk of Incident Alzheimer Disease. *Neurology* **2018**, *90*, doi:10.1212/WNL.0000000000004818.
19. Hirota, T.; Fujita, Y.; Ieiri, I. An Updated Review of Pharmacokinetic Drug Interactions and Pharmacogenetics of Statins. *Expert Opinion on Drug Metabolism & Toxicology* **2020**, *16*, 809–822, doi:10.1080/17425255.2020.1801634.
20. Zhang, X.; Xing, L.; Jia, X.; Pang, X.; Xiang, Q.; Zhao, X.; Ma, L.; Liu, Z.; Hu, K.; Wang, Z.; et al. Comparative Lipid-Lowering/Increasing Efficacy of 7 Statins in Patients with Dyslipidemia, Cardiovascular Diseases, or Diabetes Mellitus: Systematic Review and Network Meta-Analyses of 50 Randomized Controlled Trials. *Cardiovascular Therapeutics* **2020**, *2020*, 1–21, doi:10.1155/2020/3987065.
21. Leelananda, S.P.; Lindert, S. Computational Methods in Drug Discovery. *Beilstein Journal of Organic Chemistry* **2016**, *12*, 2694–2718, doi:10.3762/bjoc.12.267.
22. Khan, P.M.; Roy, K. Current Approaches for Choosing Feature Selection and Learning Algorithms in Quantitative Structure–Activity Relationships (QSAR). *Expert Opinion on Drug Discovery* **2018**, *13*, 1075–1089, doi:10.1080/17460441.2018.1542428.
23. Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure–Activity Applications: A Hands-On Approach. In *Computational Toxicology*; Nicolotti, O., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2018; Vol. 1800, pp. 3–53 ISBN 978-1-4939-7898-4.
24. Sato, A.; Miyao, T.; Jasial, S.; Funatsu, K. Comparing Predictive Ability of QSAR/QSPR Models Using 2D and 3D Molecular Representations. *J Comput Aided Mol Des* **2021**, *35*, 179–193, doi:10.1007/s10822-020-00361-7.
25. Rajathei, D.M.; Parthasarathy, S.; Selvaraj, S. Combined QSAR Model and Chemical Similarity Search for Novel HMG-CoA Reductase Inhibitors for Coronary Heart Disease. *Current Computer-Aided Drug Design* **2020**, *16*, 473–485.
26. Moorthy, N.H.N.; Cerqueira, N.M.; Ramos, M.J.; Fernandes, P.A. Ligand Based Analysis on HMG-CoA Reductase Inhibitors. *Chemometrics and Intelligent Laboratory Systems* **2015**, *140*, 102–116.
27. Samizo, S.; Kaneko, H. Predictive Modeling of HMG-CoA Reductase Inhibitory Activity and Design of New HMG-CoA Reductase Inhibitors. *ACS Omega* **2023**, *8*, 27247–27255, doi:10.1021/acsomega.3c02567.
28. Zang, Y.; Li, Y.; Yin, Y.; Chen, S.; Kai, Z. Discovery and Quantitative Structure–Activity Relationship Study of Lepidopteran HMG-COA Reductase Inhibitors as Selective Insecticides. *Pest Management Science* **2017**, *73*, 1944–1952, doi:10.1002/ps.4561.
29. Oliveira, M.A.; Araújo, R.D.C.M.U.; Lopes, C.D.C.; De Oliveira, B.G. In Silico Studies Combining QSAR Models, DFT-Based Reactivity Descriptors and Docking Simulations of Phthalimide Congeners with Hypolipidemic Activity. *Orbital: Electron. J. Chem.* **2021**, *13*, 188–199, doi:10.17807/orbital.v13i3.1493.
30. Choudhary, M.I.; Naheed, S.; Jalil, S.; Alam, J.M.; Atta-ur-Rahman Effects of Ethanolic Extract of Iris Germanica on Lipid Profile of Rats Fed on a High-Fat Diet. *Journal of Ethnopharmacology* **2005**, *98*, 217–220, doi:10.1016/j.jep.2005.01.013.
31. Naylor, M.R.; Ly, A.M.; Handford, M.J.; Ramos, D.P.; Pye, C.R.; Furukawa, A.; Klein, V.G.; Noland, R.P.; Edmondson, Q.; Turmon, A.C.; et al. Lipophilic Permeability Efficiency Reconciles the Opposing Roles of Lipophilicity in Membrane Permeability and Aqueous Solubility. *J. Med. Chem.* **2018**, *61*, 11169–11182, doi:10.1021/acs.jmedchem.8b01259.
32. De, P.; Kar, S.; Ambure, P.; Roy, K. Prediction Reliability of QSAR Models: An Overview of Various Validation Tools. *Arch Toxicol* **2022**, *96*, 1279–1295, doi:10.1007/s00204-022-03252-y.
33. Zhang, S. Partial Dependence of Breast Tumor Malignancy on Ultrasound Image Features Derived from Boosted Trees. *J. Electron. Imaging* **2010**, *19*, 023004, doi:10.1117/1.3385763.
34. Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337, doi:10.1021/acs.jcim.5b00559.
35. Brown, R.D.; Martin, Y.C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *Journal of Chemical Information and Computer Sciences* **1997**, *37*, 1–9.
36. Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in Drug Design-a Review. *Current topics in medicinal chemistry* **2010**, *10*, 95–115.

37. Hadni, H.; Elhallaoui, M. 2D and 3D-QSAR, Molecular Docking and ADMET Properties in Silico Studies of Azaaurones as Antimalarial Agents. *New Journal of Chemistry* **2020**, *44*, 6553–6565.
38. Fan, T.; Sun, G.; Zhao, L.; Cui, X.; Zhong, R. QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds. *IJMS* **2018**, *19*, 3015, doi:10.3390/ijms19103015.
39. Gramatica, P.; Papa, E. QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors. *QSAR Comb. Sci.* **2003**, *22*, 374–385, doi:10.1002/qsar.200390027.
40. Abreu, R.M.V.; Ferreira, I.C.F.R.; Queiroz, M.J.R.P. QSAR Model for Predicting Radical Scavenging Activity of Di(Hetero)Arylamines Derivatives of Benzo[b]Thiophenes. *European Journal of Medicinal Chemistry* **2009**, *44*, 1952–1958, doi:10.1016/j.ejmech.2008.11.011.
41. Sharma, S.; Prabhakar, Y.S.; Singh, P.; Sharma, B.K. QSAR Study about ATP-Sensitive Potassium Channel Activation of Cromakalim Analogues Using CP-MLR Approach. *European Journal of Medicinal Chemistry* **2008**, *43*, 2354–2360, doi:10.1016/j.ejmech.2008.01.020.
42. Fernández, M.; Caballero, J. QSAR Modeling of Matrix Metalloproteinase Inhibition by N-Hydroxy-α-Phenylsulfonylacetamide Derivatives. *Bioorganic & Medicinal Chemistry* **2007**, *15*, 6298–6310, doi:10.1016/j.bmc.2007.06.014.
43. Kadam, R.U.; 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.
44. Seraj, K.; Asadollahi-Baboli, M. In Silico Evaluation of 5-Hydroxypyrazoles as LSD1 Inhibitors Based on Molecular Docking Derived Descriptors. *Journal of Molecular Structure* **2019**, *1179*, 514–524, doi:10.1016/j.molstruc.2018.11.019.
45. Adhikari, N.; Banerjee, S.; Baidya, S.K.; Ghosh, B.; Jha, T. Ligand-Based Quantitative Structural Assessments of SARS-CoV-2 3CLpro Inhibitors: An Analysis in Light of Structure-Based Multi-Molecular Modeling Evidences. *Journal of Molecular Structure* **2022**, *1251*, 132041, doi:10.1016/j.molstruc.2021.132041.
46. Kumar, V.; Gupta, M.K.; Singh, G.; Prabhakar, Y.S. CP-MLR/PLS Directed QSAR Study on the Glutaminyl Cyclase Inhibitory Activity of Imidazoles: Rationales to Advance the Understanding of Activity Profile. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2013**, *28*, 515–522, doi:10.3109/14756366.2011.654111.
47. De Melo, E.B. Multivariate SAR/QSAR of 3-Aryl-4-Hydroxyquinolin-2(1H)-One Derivatives as Type I Fatty Acid Synthase (FAS) Inhibitors. *European Journal of Medicinal Chemistry* **2010**, *45*, 5817–5826, doi:10.1016/j.ejmech.2010.09.044.
48. Liu, Y.; Yu, X.; Chen, J. Quantitative Structure–Property Relationship of Distribution Coefficients of Organic Compounds. *SAR and QSAR in Environmental Research* **2020**, *31*, 585–596, doi:10.1080/1062936X.2020.1782468.
49. Stone, B.; Sapper, E. Machine Learning for the Design and Development of Biofilm Regulators 2018.
50. Ishfaq, M.; Aamir, M.; Ahmad, F.; M Mebed, A.; Elshahat, S. Machine Learning-Assisted Prediction of the Biological Activity of Aromatase Inhibitors and Data Mining to Explore Similar Compounds. *ACS Omega* **2022**, *7*, 48139–48149, doi:10.1021/acsomega.2c06174.
51. Lavado, G.J.; Baderna, D.; Carnesecchi, E.; Toropova, A.P.; Toropov, A.A.; Dorne, J.L.C.M.; Benfenati, E. QSAR Models for Soil Ecotoxicity: Development and Validation of Models to Predict Reproductive Toxicity of Organic Chemicals in the Collembola Folsomia Candida. *Journal of Hazardous Materials* **2022**, *423*, 127236, doi:10.1016/j.jhazmat.2021.127236.
52. Yu, X. Global Classification Models for Predicting Acute Toxicity of Chemicals towards Daphnia Magna. *Environmental Research* **2023**, *238*, 117239, doi:10.1016/j.envres.2023.117239.
53. Ghasemi, J.B.; Zohrabi, P.; Khajehsharifi, H. Quantitative Structure–Activity Relationship Study of Nonpeptide Antagonists of CXCR2 Using Stepwise Multiple Linear Regression Analysis. *Monatsh Chem* **2010**, *141*, 111–118, doi:10.1007/s00706-009-0225-4.
54. Matias, M.; Campos, G.; Santos, A.O.; Falcão, A.; Silvestre, S.; Alves, G. Synthesis, in Vitro Evaluation and QSAR Modelling of Potential Antitumoral 3,4-Dihydropyrimidin-2-(1H)-Thiones. *Arabian Journal of Chemistry* **2019**, *12*, 5086–5102, doi:10.1016/j.arabjc.2016.12.007.
55. Shekhawat, N.; Singh, P. CP-MLR/PLS Directed Structure-Activity Study in Modeling of the Aggrecanase-1 Inhibitory Activity of Biphenylsulfonamides. *Indian Journal of Chemistry* **2024**, *63*, 315–324, doi:10.56042/ijc.v63i3.6966.
56. Worachartcheewan, A.; Nantasenamat, C.; Prachayasittikul, S.; Aiemsaard, A.; Prachayasittikul, V. Towards the Design of 3-Aminopyrazole Pharmacophore of Pyrazolopyridine Derivatives as Novel Antioxidants. *Med Chem Res* **2017**, *26*, 2699–2706, doi:10.1007/s00044-017-1967-x.
57. Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867–878, doi:10.1021/ci4000213.
58. 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. *Molecules* **2015**, *20*, 8270–8286, doi:10.3390/molecules20058270.

59. Lei, B.; Li, J.; Lu, J.; Du, J.; Liu, H.; Yao, X. Rational Prediction of the Herbicidal Activities of Novel Protoporphyrinogen Oxidase Inhibitors by Quantitative Structure−Activity Relationship Model Based on Docking-Guided Active Conformation. *J. Agric. Food Chem.* **2009**, *57*, 9593–9598, doi:10.1021/jf902010g.
60. De, P.; Roy, K. QSAR and QSAAR Modeling of Nitroimidazole Sulfonamide Radiosensitizers: Application of Small Dataset Modeling. *Struct Chem* **2021**, *32*, 631–642, doi:10.1007/s11224-021-01734-w.
61. Cañizares-Carmenate, Y.; Mena-Ulecia, K.; MacLeod Carey, D.; Perera-Sardiña, Y.; Hernández-Rodríguez, E.W.; Marrero-Ponce, Y.; Torrens, F.; Castillo-Garit, J.A. Machine Learning Approach to Discovery of Small Molecules with Potential Inhibitory Action against Vasoactive Metalloproteases. *Mol Divers* **2022**, *26*, 1383–1397, doi:10.1007/s11030-021-10260-0.
62. Hasegawa, K.; Funatsu, K. Advanced PLS Techniques in Chemoinformatics Studies. *Current computer-aided drug design* **2010**, *6*, 103–127.
63. 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.
64. Noorizadeh, H. Linear and Nonlinear Quantitative Structure Linear Retention Indices Relationship Models for Essential Oils. *Eurasian Journal of Analytical Chemistry* **2013**, *8*.
65. Sharma, S.; Sharma, B.K.; Pilania, P.; Singh, P.; Prabhakar, Y.S. Modeling of the Growth Hormone Secretagogue Receptor Antagonistic Activity Using Chemometric Tools. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2009**, *24*, 1024–1033, doi:10.1080/14756360802608054.
66. Jahan, A.; Sharma, B.K.; Sharma, V.D. Quantitative Structure-Activity Relationship Study on the MMP-13 Inhibitory Activity of Fused Pyrimidine Derivatives Possessing a 1, 2, 4-Triazol-3-Yl Group as a ZBG. *GSC Biological and Pharmaceutical Sciences* **2021**, *16*, 251–265.
67. Xuan, Y.; Zhou, Y.; Yue, Y.; Zhang, N.; Sun, G.; Fan, T.; Zhao, L.; Zhong, R. Identification of Potential Natural Product Derivatives as CK2 Inhibitors Based on GA-MLR QSAR Modeling, Synthesis and Biological Evaluation. *Medicinal Chemistry Research* **2024**, 1–14.
68. Duchowicz, P.R.; Talevi, A.; Bellera, C.; Bruno-Blanch, L.E.; Castro, E.A. Application of Descriptors Based on Lipinski's Rules in the QSPR Study of Aqueous Solubilities. *Bioorganic & medicinal chemistry* **2007**, *15*, 3711–3719.
69. Choudhary, M.; Deshpande, S.; Sharma, B. CP-MLR Directed QSAR Rationales for the 1-Aryl Sulfonyl Tryptamines as 5-HT6 Receptor Ligands. *British Journal of Pharmaceutical Research* **2015**, *8*, 1–17.
70. Meena, D.K.; Sharma, B.K.; Parihar, R. Quantitative Structure-Activity Relationship Study on the CDK2 Inhibitory Activity of 6-Substituted 2-Arylaminopurines. *GSC Biological and Pharmaceutical Sciences* **2022**, *20*, 107–119.
71. Raghuraj, P.; Afsar, J.; Kishore, S.B. CP-MLR Derived QSAR Rationales for the PPARγ Agonistic Activity of the Pyridyloxybenzene-Acylsulfonamide Derivatives. *GSC Biological and Pharmaceutical Sciences* **2020**, *12*, 273–285.
72. Sharma, B.K.; Sarbhai, K.; Singh, P. A Rationale for the Activity Profile of Arylpiperazinylthioalkyls as 5-HT1A-Serotonin and A1-Adrenergic Receptor Ligands. *European Journal of Medicinal Chemistry* **2010**, *45*, 1927–1934, doi:10.1016/j.ejmech.2010.01.034.
73. Santos Cruz, D.; Santos Castilho, M. 2D QSAR Studies on Series of Human Beta-Secretase (BACE-1) Inhibitors. *Medicinal Chemistry* **2014**, *10*, 162–173.
74. Dolatabadi, M.; Nekoei, M.; Banaei, A. Prediction of Antibacterial Activity of Pleuromutilin Derivatives by Genetic Algorithm–Multiple Linear Regression (GA–MLR). *Monatsh Chem* **2010**, *141*, 577–588, doi:10.1007/s00706-010-0299-z.
75. Ojha, P.K.; Roy, K. Chemometric Modeling of Odor Threshold Property of Diverse Aroma Components of Wine. *RSC Adv.* **2018**, *8*, 4750–4760, doi:10.1039/C7RA12295K.
76. Antypenko, L.M.; Kovalenko, S.I.; Los', T.S.; Rebec', O.L. Synthesis and Characterization of Novel *N* -(Phenyl, Benzyl, Hetaryl)-2-([1,2,4]Triazolo[1,5- *c*]Quinazolin-2-ylthio)Acetamides by Spectral Data, Antimicrobial Activity, Molecular Docking and QSAR Studies. *Journal of Heterocyclic Chem* **2017**, *54*, 1267–1278, doi:10.1002/jhet.2702.
77. Abreu, R.M.V.; Ferreira, I.C.F.R.; Calhelha, R.C.; Lima, R.T.; Vasconcelos, M.H.; Adega, F.; Chaves, R.; Queiroz, M.-J.R.P. Anti-Hepatocellular Carcinoma Activity Using Human HepG2 Cells and Hepatotoxicity of 6-Substituted Methyl 3-Aminothieno[3,2-b]Pyridine-2-Carboxylate Derivatives: In Vitro Evaluation, Cell Cycle Analysis and QSAR Studies. *European Journal of Medicinal Chemistry* **2011**, *46*, 5800–5806, doi:10.1016/j.ejmech.2011.09.029.
78. Nembri, S.; Grisoni, F.; Consonni, V.; Todeschini, R. In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. *International Journal of Molecular Sciences* **2016**, *17*, 914, doi:10.3390/ijms17060914.
79. Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638–1647, doi:10.1021/ci700083n.

80. Zhao, Z.; Cui, J.; Yin, Y.; Zhang, H.; Liu, Y.; Zeng, R.; Fang, C.; Kai, Z.; Wang, Z.; Wu, F. Synthesis and Biological Evaluation of Gem-Difluoromethylenated Statin Derivatives as Highly Potent HMG-CoA Reductase Inhibitors. *Chin. J. Chem.* **2016**, *34*, 801–808, doi:10.1002/cjoc.201600180.
81. Andrade-Ochoa, S.; Correa-Basurto, J.; Rodríguez-Valdez, L.M.; Sánchez-Torres, L.E.; Nogueda-Torres, B.; Nevárez-Moorillón, G.V. In Vitro and in Silico Studies of Terpenes, Terpenoids and Related Compounds with Larvicidal and Pupaecidal Activity against Culex Quinquefasciatus Say (Diptera: Culicidae). *Chemistry Central Journal* **2018**, *12*, 53, doi:10.1186/s13065-018-0425-2.
82. Scotti, M.T.; Scotti, L.; Ishiki, H.M.; Peron, L.M.; De Rezende, L.; Do Amaral, A.T. Variable-Selection Approaches to Generate QSAR Models for a Set of Antichagasic Semicarbazones and Analogues. *Chemometrics and Intelligent Laboratory Systems* **2016**, *154*, 137–149, doi:10.1016/j.chemolab.2016.03.023.
83. Galvez-Llompart, M.; Hierrezuelo, J.; Blasco, M.; Zanni, R.; Galvez, J.; De Vicente, A.; Pérez-García, A.; Romero, D. Targeting Bacterial Growth in Biofilm Conditions: Rational Design of Novel Inhibitors to Mitigate Clinical and Food Contamination Using QSAR. *Journal of Enzyme Inhibition and Medicinal Chemistry* **2024**, *39*, 2330907, doi:10.1080/14756366.2024.2330907.
84. Seth, A.; Roy, K. QSAR Modeling of Algal Low Level Toxicity Values of Different Phenol and Aniline Derivatives Using 2D Descriptors. *Aquatic Toxicology* **2020**, *228*, 105627, doi:10.1016/j.aquatox.2020.105627.
85. Stanton, D.T.; Baker, J.R.; McCluskey, A.; Paula, S. Development and Interpretation of a QSAR Model for in Vitro Breast Cancer (MCF-7) Cytotoxicity of 2-Phenylacrylonitriles. *J Comput Aided Mol Des* **2021**, *35*, 613–628, doi:10.1007/s10822-021-00387-5.
86. Sharma, M.C.; Sharma, S. Molecular Modeling Study of Uracil-Based Hydroxamic Acids-Containing Histone Deacetylase Inhibitors. *Arabian Journal of Chemistry* **2019**, *12*, 2206–2215, doi:10.1016/j.arabjc.2014.12.030.
87. Jovanović, M.; Radan, M.; Čarapić, M.; Filipović, N.; Nikolic, K.; Crevar, M. Application of Parallel Artificial Membrane Permeability Assay Technique and Chemometric Modeling for Blood–Brain Barrier Permeability Prediction of Protein Kinase Inhibitors. *Future Medicinal Chemistry* **2024**, *16*, 873–885, doi:10.4155/fmc-2023-0390.
88. 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. *Pharmaceutical Research* **2015**, *32*, 3604–3617, doi:10.1007/s11095-015-1720-4.
89. Li, Y.; Fan, T.; Ren, T.; Zhang, N.; Zhao, L.; Zhong, R.; Sun, G. Ecotoxicological Risk Assessment of Pesticides against Different Aquatic and Terrestrial Species: Using Mechanistic QSTR and iQSTTR Modelling Approaches to Fill the Toxicity Data Gap. *Green Chem.* **2024**, *26*, 839–856, doi:10.1039/D3GC03109H.
90. Catherene Tomy, P.; Mohan, C.G. Chemical Space Navigation by Machine Learning Models for Discovering Selective MAO-B Enzyme Inhibitors for Parkinson's Disease. *Artificial Intelligence Chemistry* **2023**, *1*, 100012, doi:10.1016/j.aichem.2023.100012.
91. Sun, G.; Fan, T.; Sun, X.; Hao, Y.; Cui, X.; Zhao, L.; Ren, T.; Zhou, Y.; Zhong, R.; Peng, Y. In Silico Prediction of O6-Methylguanine-DNA Methyltransferase Inhibitory Potency of Base Analogs with QSAR and Machine Learning Methods. *Molecules* **2018**, *23*, 2892, doi:10.3390/molecules23112892.
92. Jamaludin, R.; Ibrahim, N.A.; Maarof, H. Development of Structure-Activity Modelling of Carboxamides Compounds for Aedes Aegypti Repellents. *Journal of Advanced Research Design* **2017**, *35*, 26–32.
93. Erzincan, P.; Saçan, M.T.; Yüce-Dursun, B.; Danış, Ö.; Demir, S.; Erdem, S.S.; Ogan, A. QSAR Models for Antioxidant Activity of New Coumarin Derivatives. *SAR and QSAR in Environmental Research* **2015**, *26*, 721–737, doi:10.1080/1062936X.2015.1088571.
94. Żołnowska, B.; Sławiński, J.; Brzozowski, Z.; Kawiak, A.; Belka, M.; Zielińska, J.; Bączek, T.; Chojnacki, J. Synthesis, Molecular Structure, Anticancer Activity, and QSAR Study of N-(Aryl/Heteroaryl)-4-(1H-Pyrrol-1-Yl)Benzenesulfonamide Derivatives. *IJMS* **2018**, *19*, 1482, doi:10.3390/ijms19051482.
95. Stasiak, J.; Koba, M.; Gackowski, M.; Baczek, T. Chemometric Analysis for the Classification of Some Groups of Drugs with Divergent Pharmacological Activity on the Basis of Some Chromatographic and Molecular Modeling Parameters. *CCHTS* **2018**, *21*, 125–137, doi:10.2174/1386207321666180129102149.
96. Jeličić, M.-L.; Kovačić, J.; Cvetnić, M.; Mornar, A.; Amidžić Klarić, D. Antioxidant Activity of Pharmaceuticals: Predictive QSAR Modeling for Potential Therapeutic Strategy. *Pharmaceuticals* **2022**, *15*, 791, doi:10.3390/ph15070791.
97. Mukherjee, R.K.; Kumar, V.; Roy, K. Chemometric Modeling of Plant Protection Products (PPPs) for the Prediction of Acute Contact Toxicity against Honey Bees (A. Mellifera): A 2D-QSAR Approach. *Journal of Hazardous Materials* **2022**, *423*, 127230, doi:10.1016/j.jhazmat.2021.127230.
98. He, J.; Peng, T.; Yang, X.; Liu, H. Development of QSAR Models for Predicting the Binding Affinity of Endocrine Disrupting Chemicals to Eight Fish Estrogen Receptor. *Ecotoxicology and Environmental Safety* **2018**, *148*, 211–219, doi:10.1016/j.ecoenv.2017.10.023.
99. Yuan, J.; Yu, S.; Gao, S.; Gan, Y.; Zhang, Y.; Zhang, T.; Wang, Y.; Yang, L.; Shi, J.; Yao, W. Predicting the Biological Activities of Triazole Derivatives as SGLT2 Inhibitors Using Multilayer Perceptron Neural

Network, Support Vector Machine, and Projection Pursuit Regression Models. *Chemometrics and Intelligent Laboratory Systems* **2016**, *156*, 166–173, doi:10.1016/j.chemolab.2016.06.002.

100. Daghighi, A.; Casanola-Martin, G.M.; Timmerman, T.; Milenković, D.; Lučić, B.; Rasulev, B. In Silico Prediction of the Toxicity of Nitroaromatic Compounds: Application of Ensemble Learning QSAR Approach. *Toxics* **2022**, *10*, 746, doi:10.3390/toxics10120746.
101. Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622, doi:10.1021/ci0600509.
102. Mauri, A.; Bertola, M. Alvascience: A New Software Suite for the QSAR Workflow Applied to the Blood–Brain Barrier Permeability. *IJMS* **2022**, *23*, 12882, doi:10.3390/ijms232112882.
103. Athista, M.; Hariharan, V.; Namratha, K.; Pavankumar, G.; Perciya, J.L.; Sunkar, S. Computational Identification of Natural Compounds as Potential Inhibitors for HMGCoA Reductase. *Current Trends in Biotechnology and Pharmacy* **2023**, *17*, 1457–1485.
104. Cuccioloni, M.; Bonfili, L.; Mozzicafreddo, M.; Cecarini, V.; Scuri, S.; Cocchioni, M.; Nabissi, M.; Santoni, G.; Eleuteri, A.M.; Angeletti, M. Mangiferin Blocks Proliferation and Induces Apoptosis of Breast Cancer Cells via Suppression of the Mevalonate Pathway and by Proteasome Inhibition. *Food Funct.* **2016**, *7*, 4299–4309, doi:10.1039/C6FO01037G.
105. Min, S.-W.; Kim, D.-H. Kakkalide and Irisolidone: HMG-CoA Reductase Inhibitors Isolated from the Flower of Pueraria Thunbergiana. *Biological & Pharmaceutical Bulletin* **2007**, *30*, 1965–1968, doi:10.1248/bpb.30.1965.
106. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45*, D945–D954, doi:10.1093/nar/gkw1074.
107. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473, doi:10.1021/ci500588j.
108. R Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024;
109. 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.
110. RStudio Team RStudio: Integrated Development Environment for R; RStudio, PBC.: Boston, MA, 2021;
111. Ballabio, D.; Grisoni, F.; Consonni, V.; Todeschini, R. Integrated QSAR Models to Predict Acute Oral Systemic Toxicity. *Mol. Inf.* **2019**, *38*, 1800124, doi:10.1002/minf.201800124.
112. Tomberg, A.; Boström, J. Can Easy Chemistry Produce Complex, Diverse, and Novel Molecules? *Drug Discovery Today* **2020**, *25*, 2174–2181, doi:10.1016/j.drudis.2020.09.027.
113. Gao, K.; Nguyen, D.D.; Sresht, V.; Mathiowetz, A.M.; Tu, M.; Wei, G.-W. Are 2D Fingerprints Still Valuable for Drug Discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373–8390, doi:10.1039/D0CP00305K.
114. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*; Roy, K., Ed.; Methods in Pharmacology and Toxicology; Springer US: New York, NY, 2020; pp. 801–820 ISBN 978-1-07-160149-5.
115. Shi, J.; Zhao, G.; Wei, Y. Computational QSAR Model Combined Molecular Descriptors and Fingerprints to Predict HDAC1 Inhibitors. *Med Sci (Paris)* **2018**, *34*, 52–58, doi:10.1051/medsci/201834f110.
116. Boudergua, S.; Alloui, M.; Belaidi, S.; Al Mogren, M.M.; Ellatif Ibrahim, U.A.A.; Hochlaf, M. QSAR Modeling and Drug-Likeness Screening for Antioxidant Activity of Benzofuran Derivatives. *Journal of Molecular Structure* **2019**, *1189*, 307–314, doi:10.1016/j.molstruc.2019.04.004.
117. Meyer, D.; Buchta, C. Proxy: Distance and Similarity Measures; 2021;
118. Remeseiro, B.; Bolon-Canedo, V. A Review of Feature Selection Methods in Medical Applications. *Computers in Biology and Medicine* **2019**, *112*, 103375, doi:10.1016/j.compbiomed.2019.103375.
119. Concu, R.; Cordeiro, M.N.D.S. On the Relevance of Feature Selection Algorithms While Developing Non-Linear QSARs. In *Ecotoxicological QSARs*; Roy, K., Ed.; Methods in Pharmacology and Toxicology; Springer US: New York, NY, 2020; pp. 177–194 ISBN 978-1-07-160149-5.
120. Lang, M.; Binder, M.; Richter, J.; Schratz, P.; Pfisterer, F.; Coors, S.; Au, Q.; Casalicchio, G.; Kotthoff, L.; Bischl, B. Mlr3: A Modern Object-Oriented Machine Learning Framework in R. *Journal of Open Source Software* **2019**, doi:10.21105/joss.01903.
121. Zuber, V.; Strimmer, K. Care: High-Dimensional Regression and CAR Score Variable Selection; 2021;
122. Kursa, M.B. Praznik: High Performance Information-Based Feature Selection. *SoftwareX* **2021**, *16*, 100819.
123. Zawadzki, Z.; Kosinski, M. FSelectorRcpp: “Rcpp” Implementation of “FSelector” Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support; 2021;
124. Hutson, G. FeatureTerminatoR: Feature Selection Engine to Remove Features with Minimal Predictive Power; 2021;
125. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **2010**, *33*, 1–22.

126. Milborrow, S.; Hastie, T.; Tibshirani, R. *Earth: Multivariate Adaptive Regression Splines*; 2024;
127. Schliep, K.; Hechenbichler, K. *Kknn: Weighted k-Nearest Neighbors*; 2016;
128. Beygelzimer, A.; Kakadet, S.; Langford, J.; Arya, S.; Mount, D.; Li, S. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*; 2024;
129. Kuhn, M.; Quinlan, R. Cubist: Rule- And Instance-Based Regression Modeling; 2024;
130. Hornik, K.; Buchta, C.; Zeileis, A. Open-Source Machine Learning: R Meets Weka. *Computational Statistics* **2009**, *24*, 225–232, doi:10.1007/s00180-008-0119-7.
131. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **2017**, *77*, 1–17, doi:10.18637/jss.v077.i01.
132. Hothorn, T.; Hornik, K.; Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **2006**, *15*, 651–674.
133. Zeileis, A. Object-Oriented Computation of Sandwich Estimators. *J. Stat. Soft.* **2006**, *16*, doi:10.18637/jss.v016.i09.
134. Hothorn, T.; Hornik, K.; Wiel, M.A.V.D.; Zeileis, A. Implementing a Class of Permutation Tests: The **Coin** Package. *J. Stat. Soft.* **2008**, *28*, doi:10.18637/jss.v028.i08.
135. 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; 2023;
136. Helleputte, T.; Paul, J.; Gramme, P. LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library; 2024;
137. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. *Xgboost: Extreme Gradient Boosting*; 2023;
138. Ridgeway, G.; Developers, G.B.M. Gbm: Generalized Boosted Regression Models; 2024;
139. Sparapani, R.; Spanbauer, C.; McCulloch, R. Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software* **2021**, *97*, 1–66, doi:10.18637/jss.v097.i01.
140. Majumdar, S.; Basak, S.C. Beware of External Validation! - A Comparative Study of Several Validation Techniques Used in QSAR Modelling. *CAD* **2018**, *14*, 284–291, doi:10.2174/1573409914666180426144304.
141. Feng, D. agRee: Various Methods for Measuring Agreement; 2020;
142. Lin, L.I. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268.
143. Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320–2335, doi:10.1021/ci200211n.
144. Gramatica, P. Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. *International Journal of Quantitative Structure-Property Relationships* **2020**, *5*, 61–97, doi:10.4018/IJQSPR.20200701.oa1.
145. Rücker, C.; Rücker, G.; Meringer, M. Y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357, doi:10.1021/ci700157b.
146. Biecek, P. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research* **2018**, *19*, 1–5.
147. Molnar, C.; Bischl, B.; Casalicchio, G. Iml: An R Package for Interpretable Machine Learning. *JOSS* **2018**, *3*, 786, doi:10.21105/joss.00786.
148. Hajalsiddig, T.T.H.; Osman, A.B.M.; Saeed, A.E.M. 2D-QSAR Modeling and Molecular Docking Studies on 1 *H* -Pyrazole-1-Carbothioamide Derivatives as EGFR Kinase Inhibitors. *ACS Omega* **2020**, *5*, 18662–18674, doi:10.1021/acsomega.0c01323.
149. Gotti, M.; Kuhn, M. Applicable: A Compilation of Applicability Domain Methods; 2022;
150. Cortes, D. Isotree: Isolation-Based Outlier Detection; 2023;
151. Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J.G.; Stephan, R.; Page, R.; Vondrášek, J.; et al. The LOTUS Initiative for Open Knowledge Management in Natural Products Research. *eLife* **2022**, *11*, e70780, doi:10.7554/eLife.70780.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.