Current Topics in
Toxicology
Vol. 16, 2020

Original Article

Topology-based correlation models for antileishmanial piplartine analogues

Jean Pierre Doucet[*] and Annick Doucet-Panaye
ITODYS,Université Paris, CNRS UMR 7086, 15 Rue jean de Baîf, 75013, Paris, France.

ABSTRACT

With a rapid diffusion, Leishmaniasis now appears as a severe tropical disease with millions of people affected. Currently used drugs are not devoid of detrimental side-effects and there is a crucial need for new alternative, active anti-parasitic chemicals. Recently, Nobrega *et al.* synthetized 32 analogues of piplartine, determined their activity against *Leishmania amazonensis* promastigote forms and presented a comparative molecular field analysis (CoMFA) treatment. Here we revisited these results and proposed topology-based 2D correlation models with special attention to robustness and predictive ability. From PaDEL and QSARINS softwares, a set of 3 descriptors was selected, and imported into multilinear regression and various machine learning approaches: partial least squares (PLS), projection pursuit regression (PPR), support vector machine (SVM with linear or Gaussian kernel) and three-layer perceptron (TLP, neural network with back-propagation algorithm). Although a reduced set of structural descriptors, these different models appeared attractive with satisfactory and consistent performances. The best results, obtained from linear SVM and three-layer perceptron, suggested that these models might be applied for screening new possible drugs.

KEYWORDS: Leishmaniasis, topology descriptors, QSAR, machine learning, neural network, support vector machine.

1. INTRODUCTION

Tropical disease Leishmaniasis, caused by protozoan parasites transmitted by bites of female phlebotomine sandflies, now largely impacts about 98 sub-tropical countries in South Africa, Asia and South America. About 12 million people are severely affected by various forms of infection, cutaneous, mucocutaneous, or visceral (targeting liver and spleen) [1-2]. Currently used drugs often rely on antimonial derivatives but these drugs have severe side effects and suffer parasite resistance. Faced to the need for new effective drugs, chemoinformatics methods have been actively investigated with structure-based approaches, focusing on drug-receptor interactions and energetics and ligand-based methods linking, *via* quantitative structure activity (or property) relationships (QSARs, QSPRs), activity or property values to chemical features of tentative drugs [3]. On the other hand, it was observed that several pepper family derivatives, such as piplartine showed varied interesting antiparasitic properties, particularly regarding antileishmanial activity [4].

In this field, Nobrega *et al.* [1] recently presented the synthesis of 32 analogues of Piplartine, and their activity against the *Leishmania amazonensis* promastigote forms. They proposed a 3D structure-activity CoMFA treatment of these compounds, and discussed the influence of the structural moieties present in these chemicals. However predictive capability of this QSAR model was not deeply investigated. This prompted us to revisit the antiparasitic activity of these new compounds in 2D topology-based QSAR models

*Email id: doucet@paris7.jussieu.fr

with particular attention paid to robustness and predictive ability. 2D descriptors are real structural invariants derived by swift calculations [5] and directly attainable with knowledge of the only molecular formula, avoiding problems related to the determination of the privileged (reactive) conformations and subsequent geometry optimisation by (often heavy) quantum calculations. Numerous applications established that these descriptors may reflect the most important features of molecular structure, even in groups of chemicals showing a large structural diversity [5] and lead to successful QSAR, QSPR models [6-9].

The work was here first carried out in the framework of OLS-MLR models (Ordinary Least Squares Multilinear Regression) using PaDEL [10] and QSARINS [11, 12] softwares. The free availability of these tools largely eases applications for predicting activity of new compounds. In addition to this MLR analysis, various machine learning approaches were further investigated. These methods are now largely used in QSAR/QSPR studies [13-27], and were recently extended to model properties of nanoparticles [28-32]. Such approaches usually do not propose any explicit, directly usable formula for property prediction. However they offer easy settings, rapid training and generally guarantee to find the global minimum on the error surface. In many examples, they gave definitively improved performance in property (or activity) data fitting or prediction for new compounds.

2. MATERIALS AND METHODS

Experimental values of activity, from MTT assays, expressed as 50% inhibitory concentration (IC_{50} in µM) of *L. amazonensis* promastigote forms, were retrieved from Nobrega *et al.* [1] and converted into pIC_{50} values. Structural formulae and activities of the investigated chemicals are reported in Table 1.

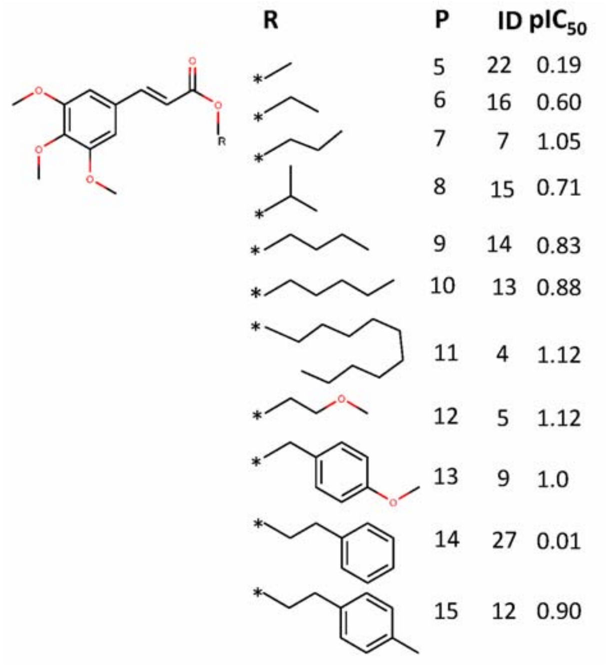
First of all, it must be noticed that these data constitute a borderline case for the development of a QSAR model: the data set is limited and the activity range is rather small: 32 compounds but, for 6 molecules, only an activity threshold is given. Activity values span only 2 log units in pIC_{50} with an error margin about 0.15 log unit on pIC_{50} for the middle of the activity range. Owing to the restricted extent of the data set, we limit the number of parameters in our models to only 3 structural variables so as to maintain a ratio (number of samples/number of parameters) >5, the currently accepted threshold (Organisation for Economic Co-operation and Development (OECD) guidance instructions) [33]. For this study, focused on a 2D, topology-based, description of molecular structure, a pool of about 1100 2D structural descriptors were initially generated by the PaDEL software [10], and further incorporated as input variables in the QSARINS software [11] to develop multi linear regression models, using ordinary least squares method (OLS-MLR). These structural variables encompassed nature of atoms, autocorrelation vectors, elements of adjacency or distance matrices, E-states, etc. From this initial set, a preliminary pre-processing step was performed in QSARINS with elimination of (nearly) constant values, and pruning pairs of highly inter-correlated descriptors (R > 0.85). This led to a restricted input set of 88 (potentially significant) structural variables.

A further selection was then carried out in multiple external validation process. For this, we defined 5 subsets (m = 0 to m = 4). Each will be alternatively considered as prediction subset, while the remaining ones constituted the corresponding training set. For this, experimental activities were first ordered by decreasing pIC_{50}, after elimination of compounds devoid of precise activity and point P14 (#ID27, see *infra*). The highest value (compound P20, #ID1) was systematically included in the considered training sets (to avoid extrapolation in prediction). This precaution seemed useful so as much the corresponding point (#ID1) was rather apart from the other investigated chemicals (as clearly evidenced on the correlation graph- *vide infra*…). The same precaution did not seem useful for the lowest activity value since several compounds exhibited similar pIC_{50} activity values, limiting the corresponding extrapolation area. The remaining compounds were then portioned in the five subsets (m = 0, 1,…4) according to their ID number (rank in the activity list) modulo 5. In other words, subset m encompasses compounds for which the rest of division by 5 of the ID numbers is m. For, example subset m = 2 corresponded to compounds ID = 2, 7, 12,…, subset 3 compounds ID = 3, 8, 13… In other words, each subset included one every five compounds, regardless of structural similarity. This led to a rather homogeneous sampling over the full reactivity range for the various subsets.

Table 1. Structure and activity of the investigated compounds.

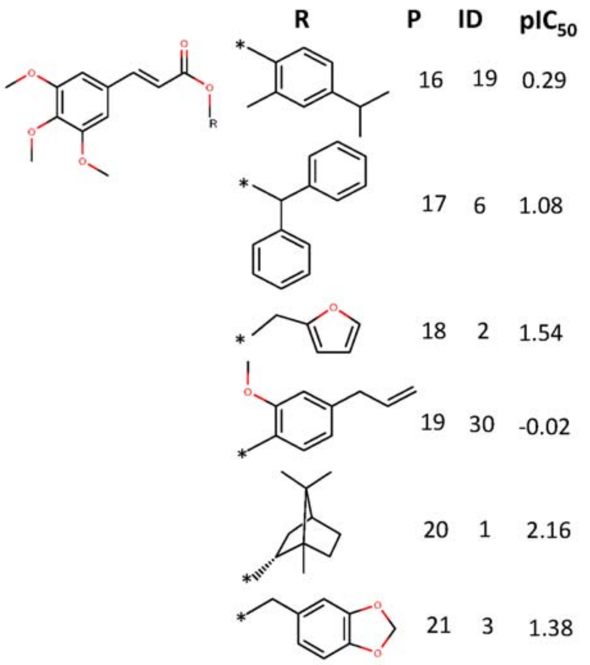
The first column indicates structural fragments common to several compounds (cinnamic esters, amides, benzoic esters), the second one, the substituent groups borne by these structures, and the following ones are the numbering in Nobrega *et al.*'s work [1]; our ID number (see text) and activity expressed as pIC_{50} (µM).

Table 1A



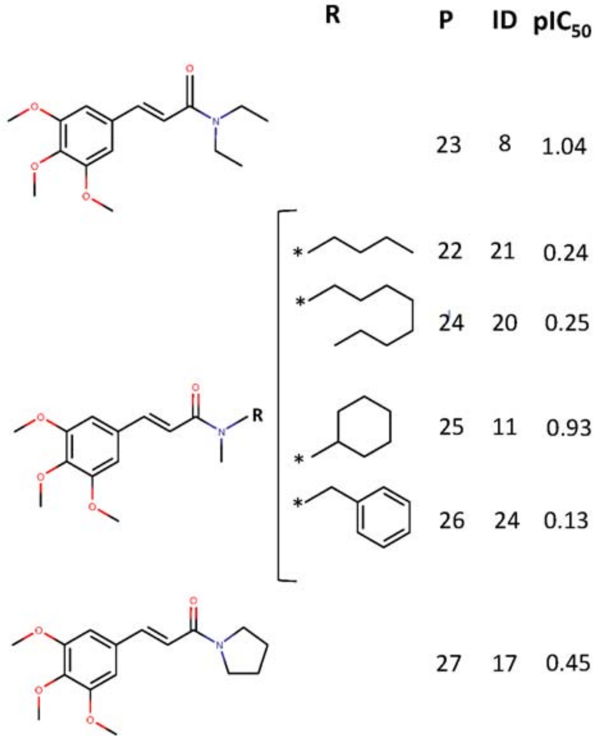
P	ID	pIC_{50}
5	22	0.19
6	16	0.60
7	7	1.05
8	15	0.71
9	14	0.83
10	13	0.88
11	4	1.12
12	5	1.12
13	9	1.0
14	27	0.01
15	12	0.90

Table 1B



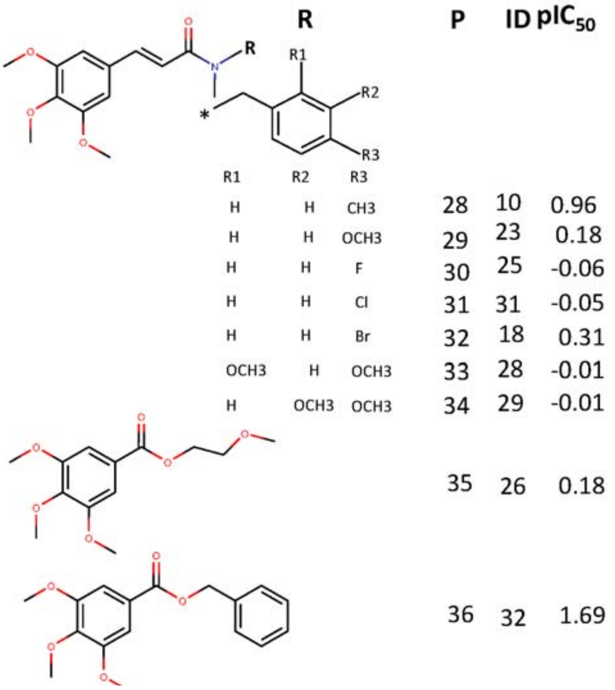
P	ID	pIC_{50}
16	19	0.29
17	6	1.08
18	2	1.54
19	30	-0.02
20	1	2.16
21	3	1.38

Table 1C



P	ID	pIC_{50}
23	8	1.04
22	21	0.24
24	20	0.25
25	11	0.93
26	24	0.13
27	17	0.45

Table 1D



R1	R2	R3	P	ID	pIC_{50}
H	H	CH3	28	10	0.96
H	H	OCH3	29	23	0.18
H	H	F	30	25	-0.06
H	H	Cl	31	31	-0.05
H	H	Br	32	18	0.31
OCH3	H	OCH3	33	28	-0.01
H	OCH3	OCH3	34	29	-0.01
			35	26	0.18
			36	32	1.69

Variable selection was then independently carried out in parallel, on the total set of compounds (that may be represented as subset m = 5) and the five subsets m = 0 to m = 4. This was performed (in each case) by exhaustive exploration of all possible triples of variables (“All subset” MLR procedure in the QSARINS package) and selection of the best MLR in loo cross validation (best Q^2 and exclusion of chance correlation by examination of the “Quick Rule” [34]).

For this analysis, compounds P19, P31, P33-P36 were discarded since only a reactivity threshold was given. Furthermore, preliminary trials showed that point P14 (ID27) largely deviates from the examined MLR recall model (in data fitting). This prompted us to also discard this point and work on a population of 25 compounds only (ID numbers #1 to #25). It may be remarked that this here-used procedure actually led to independent descriptor selections and external validation steps, since the prediction set chemicals were never involved during the development of each of the training MLR models. Although looking at first glance as some Leave-Some-Out process, things actually corresponded to a true external validation since the training sets (from one run to another) were not identical (only 66% similarity between two trials) and variable selection was independently operated on the entire reduced set of 88 descriptors rather than adjusting correlations on various sets of data with a unique choice of few, preliminary selected, descriptors.

3. RESULTS AND DISCUSSION

Descriptor selection

Examination of the selections carried out on the entire data set and the five subsets (m = 0 to m = 4) converged on the same trio of descriptors: AATS 8i, ATSC 1p, MATS 5i (acronyms detailed in Table 2). MLR built separately on these variables ranked first for the total population, and four of the five investigated subsets (m = 1 to m = 4), and second for subset m = 0. The three variables so selected will be used in all subsequent treatments.

Data Fitting (“recall”) and Predictions *via* multilinear regression

Multi linear regression by ordinary least squares (OLS-MLR) is undoubtedly the most widely used tool in QSAR modelling treatment due to its efficient and straightforward implementation. Some basic elements of this treatment are summarized in Supplementary Materials.

Data fitting

For the 25 examined compounds of the full set, a satisfactory multilinear regression model was established between observed pIC_{50} values and the 3 selected descriptors:

$$pIC_{50} = 21.092 - 0.1316\ AATS8i - 0.6729\ ATSC1p + 5.6858\ MATS5i \quad (1)$$

$R^2 = 0.8055$ RMSE = 0.22 MAE = 0.18

$Q^2loo = 0.7139$ RMSE = 0.27 MAE = 0.22

Table 2. The three selected, topology-based, 2D descriptors.

AATS8i	Averaged Broto-Moreau autocorrelation term, lag 8, weighted by the first ionisation potential. Autocorrelation term corresponds to the sum, on all pairs of atoms (i, j) separated by a given topological distance, the “lag” (here eight bonds), of the products of the property value associated to each atom of the pair (here the first ionization potential, “I”). Averaged descriptors are obtained by dividing each term by the corresponding number of contribution (avoiding dependence on the molecular size). $AATS8i = \sum_i^A \sum_j^A \delta_{ij} I_i I_j$ with $\delta_{ij} = 1$ if atoms I and j are separated by 8 bonds, zero otherwise. A is the number of atoms.
ATSC1p	Centered Broto-Moreau autocorrelation- lag1-weighted by polarisabilities.
MATS5i	Moran autocorrelation lag 5-weighted by the first ionization potential $I_k = (1/\Delta k) \sum_i^A \sum_j^A (w_i–\hat{w}) (w_j-\hat{w}) \delta_{ij} / (1/A) \sum_i^A (w_i–\hat{w})^2$ $(w_i–\hat{w})$ and $(w_j-\hat{w})$ are the centered property values (mean $\hat{w}$), and the autocorrelation values are weighted by the square of the centered property value on all atoms

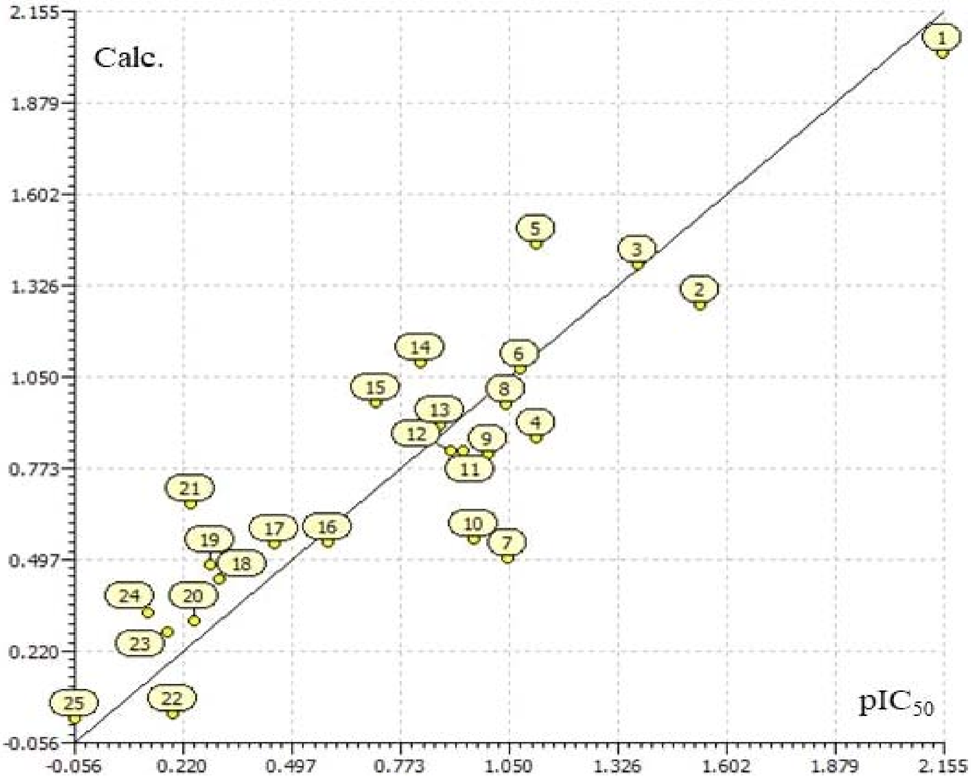


Figure 1. MLR model. Plot of calculated pIC$_{50}$ values *vs* observed ones: see equation (1) Numbers refer to compound ID.

Supplementary statistical criteria are reported in Sup. Mat. Table 1 and the correlation between calculated and experimental pIC$_{50}$ values is displayed in Figure 1. Individual activity values are collected in Sup. Mat. Table 2.

The quality of the model, in data fitting, was quantified by the determination coefficient (R^2), the root mean squared error (RMSE) and mean absolute error (MAE) [35]. Robustness is characterized by the leave one out cross-validated determination coefficient Q^2 loo, with a value that must be close to R^2. The ‘QUICK Rule’ [34] allowed for discarding chance correlation. This might be also confirmed by the low R^2value (0.13) observed in Y scrambling. Williams’ plot [11] indicated that only compound #ID =1 (P20) lies outside the applicability domain. However this compound was correctly calculated and (as previously indicated) was left in training sets to avoid extrapolated predictions (Sup. Mat. Figure 1).

The coefficient importance (according to standardized values) decreases in the sequence:

MATS5i > ATSC1p > AATS8i

Taking into account the coefficients of these three selected descriptors in MLR equation (1), it was possible to evaluate how much each descriptor intervenes in the variations of calculated pIC$_{50}$ (See Sup. Mat. Figure 2). However a detailed analysis was difficult since activity varies only on 2 log units. Among the most striking points, we could observe that the most active compound P20 (#ID 1), where the ester oxygen is linked to a bicyclic system, had a very high value of MATS5i (about one unit higher than the other chemicals) and that the corresponding values of AATS8i slightly superior to the mean, and ATSC1p (slightly inferior) mutually compensated themselves. For point P36 (ID#32); the activity (pIC$_{50}$) was also calculated rather high (in agreement with experience) with AATS8i high (the two other variable taking values near to the mean of the set). The high values of MATS5i and ATSC1p for compound P35 might be also noted, which might explain a high predicted activity (although AATS8i is rather low) not confirmed by experiment *(vide infra)*.

External validation

Predictive ability of the proposed models is quantified by external validation [12, 36]. As previously said, results obtained with the 5 subsets (m = 0 to m = 4) using the same group of descriptors (those selected) constituted five external validation processes. Quality of these models was examined on R^2tr on learning and for prediction R^2pred (which indicated how well the predicted pIC_{50} were proportional to the observed values) and Q^2 (in fact Q2-F2) which compared the errors in the prediction model and a "null model" evaluating activities as the mean of observed pIC_{50} for these compounds. Note that it had been also proposed that Q^2 loo might constitute a good quality criterion for prediction in limited datasets provided its evaluation is included in descriptor selection [37].

Indeed, in the usual train/test validation procedure, about 20% of the data set is set apart to constitute the test (aka validation or prediction) set. The model is adjusted (determination of the coefficient of the MLR equation) on the remaining compounds (those of the training set) and its validity controlled on the test set compounds. Clearly this reduction in the number of training samples leads to a loss of information available to build the model, which may be detrimental for data sets with a limited number of samples. Conversely, in leave–one-out cross validation, only one compound is, at each step, considered for evaluating the model (verifying that the calculated activity is consistent with the observed one), thus avoiding this drawback.

Results gathered in Table 3 look globally satisfactory although in some cases Q^2pred was low. This is not unexpected since, with limited prediction sets (20% of the population, that is, here, 5 or 4 compounds only), a deviation for one compound might be heavily detrimental for the global result. Validity of the model will be confirmed by carrying out a large number of random prediction tests (*vide infra*: random MLR runs).

In these treatments, each compound was included four times in learning (with quite neighbour calculated results on pIC_{50} in data fitting) and once in prediction. To get a more synthetic view of prediction capability, we gathered these predicted values (in subsets m = 0 to m = 4) in a single file and compared it to the observed pIC_{50} (last line for every method in Table 3).

Random MLR runs

To confirm the choice of the selected descriptors, we carried out 2000 runs of cross validation with 30% data left out. For a more homogeneous sampling, we separately considered about 13 compounds in the more reactive half of the population and 12 in the less reactive part in the ID-ordered list of 25 compounds. From these two parts we randomly selected for each MLR run, 4 compounds to build the validation group

Table 3. Performance of the investigated approaches.
The first column indicates for each method, the subsets (m = 0 to m = 4); m = 5 corresponds to the full set (25 compounds). The three following columns rely on data fitting (R^2tr, rmsetr, maetr), then prediction (R^2pr, rmsepr, maepr, Q^2pr) and loo (R^2loo, rmseloo, maeloo, Q^2loo). For each method, the last line corresponds to prediction on the gathered prediction files. Individuals' calculated activity values are given in Sup. Mat. Table 2.

	R^2tr	**rmsetr**	**maetr**	**R^2pr**	**rmsepr**	**maepr**	**Q^2pr**	**R^2loo**	**rmseloo**	**maeloo**	**Q^2-loo**
MLR											
0	0.853	0.20	0.15	0.682	0.37	0.30	0.301	0.749	0.26	0.20	0.747
1	0.827	0.22	0.18	0.554	0.24	0.16	0.452	0.736	0.27	0.23	0.735
2	0.859	0.19	0.15	0.766	0.36	0.29	0.410	0.769	0.25	0.20	0.768
3	0.777	0.25	0.20	0.982	0.10	0.08	0.948	0.649	0.31	0.26	0.646
4	0.824	0.22	0.17	0.686	0.24	0.24	0.632	0.726	0.28	0.22	0.745
5	0.805	0.22	0.18	NA	NA	NA	NA	0.715	0.27	0.22	0.715
				0.614	0.28	0.22	0.572				

Table 3 continued..

PLS											
0	0.853	0.20	0.15	0.682	0.37	0.30	0.305	0.658	0.30	0.21	0.647
1	0.827	0.22	0.18	0.570	0.24	0.16	0.468	0.704	0.29	0.23	0.701
2	0.859	0.19	0.15	0.763	0.36	0.30	0.407	0.698	0.29	0.22	0.685
3	0.777	0.25	0.20	0.981	0.10	0.08	0.949	0.631	0.32	0.26	0.629
4	0.824	0.22	0.17	0.683	0.24	0.24	0.627	0.664	0.31	0.23	0.661
5	0.805	0.22	0.18	NA	NA	NA	NA	0.676	0.29	0.22	0.674
				0.615	0.28	0.22	0.573				
PPR											
0	0.879	0.18	0.14	0.623	0.37	0.31	0.287	0.611	0.33	0.26	0.595
1	0.865	0.20	0.16	0.499	0.30	0.22	0.133	0.468	0.40	0.29	0.444
2	0.896	0.17	0.13	0.700	0.35	0.29	0.435	0.649	0.31	0.24	0.646
3	0.789	0.24	0.19	0.985	0.10	0.09	0.955	0.373	0.42	0.36	0.341
4	0.826	0.22	0.17	0.715	0.23	0.23	0.643	0.538	0.36	0.28	0.534
5	0.823	0.21	0.16	NA	NA	NA	NA	0.545	0.35	0.28	0.523
				0.573	0.29	0.23	0.548				
ANN											
0	0.887	0.17	0.14	0.652	0.31	0.23	0.505	0.669	0.30	0.22	0.666
1	0.875	0.19	0.15	0.698	0.25	0.17	0.432	0.621	0.33	0.26	0.621
2	0.876	0.18	0.14	0.798	0.30	0.21	0.594	0.636	0.31	0.23	0.634
3	0.844	0.21	0.16	0.957	0.12	0.08	0.933	0.485	0.38	0.30	0.478
4	0.809	0.23	0.18	0.607	0.26	0.25	0.572	0.511	0.37	0.28	0.502
5	0.856	0.19	0.15	NA	NA	NA	NA	0.616	0.31	0.23	0.615
				0.654	0.26	0.19	0.646				
SVM linear											
0	0.849	0.20	0.14	0.692	0.36	0.29	0.332	0.805	0.24	0.20	0.777
1	0.825	0.22	0.17	0.533	0.24	0.18	0.446	0.760	0.27	0.20	0.750
2	0.854	0.20	0.15	0.785	0.33	0.26	0.513	0.675	0.30	0.20	0.669
3	0.769	0.25	0.21	0.956	0.16	0.15	0.867	0.669	0.30	0.26	0.659
4	0.820	0.23	0.16	0.704	0.22	0.21	0.681	0.763	0.27	0.22	0.742
5	0.802	0.23	0.17	NaN	NaN	NaN	NaN	0.785	0.24	0.18	0.775
				0.623	0.27	0.22	0.593				
SVM Radial											
0	0.918	0.15	0.12	0.736	0.37	0.33	.0.287	0.619	0.32	0.22	0.609
1	0.914	0.16	0.12	0.996	0.17	0.15	0.727	0.497	0.40	0.27	0.447
2	0.896	0.17	0.10	0.690	0.34	0.32	0.462	0.453	0.39	0.24	0.413
3	0.882	0.18	0.16	0.946	0.15	0.13	0.889	0.554	0.35	0.27	0.554
4	0.905	0.17	0.11	0.783	0.18	0.17	0.780	0.716	0.28	0.22	0.710
5	0.901	0.17	0.13	NA	NA	NA	NA	0.527	0.35	0.24	0.521
				0.636	0.26	0.22	0.620				

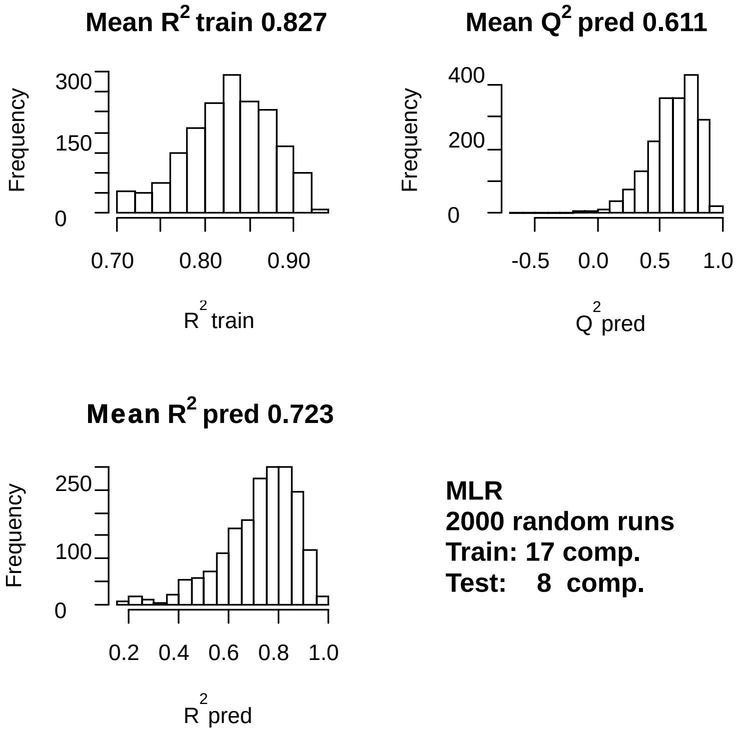


Figure 2. Histogram of statistical criteria R^2train, R^2pred, Q^2pred for 2000 random MLR runs.

(8 compounds), the remaining 17 chemicals forming the corresponding training set. The histograms of R^2 fitting, R^2 and Q^2 prediction are given in Figure 2.

The obtained values (0.827, 0.723 and 0.611 respectively) confirm the validity of the selected descriptor set. Consistency of these results prompted us to consider that the three selected structural variables led to satisfactory fitting and prediction for the various populations studied. Presumably this choice would not always be the optimal one, when looking independently at each splitting. But we considered it gives a unique set of structural variables actually applicable to the various subsets and that also could be used for the other correlation methods we proposed *via* machine learning approaches (*vide infra*).

Discarded points

In the original data set, six compounds were only assigned threshold values. And we also eliminated (after preliminary results) compound #ID 27. So, it was interesting to examine what will be the predictions of our MLR model with the 3 selected descriptors. The model correctly predicted a rather high activity for P36 (1.48 *vs* 1.70 exp. value) and low pIC_{50} (< 0.30) for P31 and P33. Strong deviations are observed for inactive compounds P19 and P34 (calculated values > 0.8). It may be noted that these compounds had two methoxy groups on adjacent positions in the “variable” part of investigated molecules, which might induce supplementary interactions. For P35 the high calculated activity, not confirmed by experiment (1.72 in place of 0.18), was consistent with a high value of MATS5i and ATSC1p and the observation that the absence of a C=C linker is more favorable [1] to activity. However these situation occurred only twice in the data set, and supplementary information would be necessary. At last, it was difficult to discuss the case of inactive P14, calculated at pIC_{50} = 0.93.

Although these results did not look quite satisfactory, it might be useful to note that points calculated as “inactive” showed low pIC_{50}values (P31 and P33) and that P36 (second point in the reactivity scale) was correctly calculated as “active”. This might be of interest to, at least, discard wrong directions in the search for new active compounds.

Machine learning results

As previously quoted, these methods are now largely used in QSAR/QSPR studies [13-32]. Several publications evidenced the efficiency of machine learning-based QSAR approaches, leading to improved results with respect to MLR analysis [14-19, 31, 32]. Using the same three-descriptor set, just selected *via* MLR correlation, we developed correlation models using partial least squares correlation (PLS), projection pursuit regression (PPR), support vector machine, with linear and radial kernels (SVM) and artificial neural network, (Three Layer Perceptron with back propagation algorithm, TLP). These methods have been largely presented in various publications [26, 27, 38-45]. So only their salient points will be reported in Supplementary Materials.

As previously noted, we might not expect drastic improvement of the results since the descriptors used have been selected by MLR. But it might be interesting to observe to what extent the modification of the structural space introduced by these machine learning methods, modified and possibly enhanced performance, in as much as projections of the initial variables in a modified descriptor space might overemphasize or underestimate some structural characteristics. Indeed, PLS operates on some (generally limited) linear combinations of the original descriptors, PPR uses selected projections of the descriptors (determined by the algorithm). SVM works on projections of the variables in a larger dimension structural space thanks to kernel functions, and TLP splits descriptor information in several parts, separately weighted by connection weights, and further recombined before transfer. No additional variable selection procedure was tempted [42]. Calculations were performed in the framework of the Cran-R-Project softwares [46] with the caret software [47, 48] and home-written routines.

The same methodology as that developed for MLR analysis was used for these machine learning approaches: Examination of performance in data fitting (full dataset, “recall”) and leave-one-out cross validation on one hand, and on the other hand, calculations on the five subsets (m = 0 to m = 4) alternatively used in train and test. Note that whereas for MLR-based descriptor selection, this process actually corresponded to an external validation step, it looked here (for machine learning approaches) like leave-some-out cv (since descriptors are already selected). The statistical elements for these different approaches are collected in Table 3. Some examples of the obtained correlation models are presented in Figure 3. Individual activity values calculated in these methods are gathered in Sup. Mat. Table 2 for data fitting, leave one out cross validation and prediction (gathering the results obtained for the five subsets).

Schematically speaking, the various approaches give highly similar and consistent results. Performance in Recall (full set data fitting) with R^2tr > 0.750 and Q^2 loo > 0.6 in most cases, looked globally satisfactory and the low differences between R^2tr and Q^2 loo are a good indication of robustness of the treatment. Results are more mitigated for subset prediction, with some mediocre performances (for instance low values of Q^2pred for subset 1 in most methods). As previously indicated, the test set encompasses here only 4 compounds, so that an important residual on one compound heavily handicaps the global result for the subset. Results on predictions, when gathered in a single file (24 compounds with ID# 1 discarded), are more homogeneous (last line for each method in the table).

Comparing the various proposed models, one can note that PLS and MLR gave nearly identical results. This is not surprising since the three selected variables are not highly correlated, and so necessary (also in PLS) for calculating activity. Projection pursuit regression (PPR), although satisfactory in data fitting, gives slightly inferior performance in loo cross validation and prediction. Support vector machine with a linear kernel, gives results very similar to MLR in recall, and in loo (except for subset m = 2, where however differences remains small) and slightly better in prediction. Radial SVM (exponential kernel) is definitely better in recall and in prediction for most subsets and the gathered predictions but clearly inferior in loo. Finally, despite of a very simple structure (3-1-1 units), the three layer perceptron (ANN) gives the best results in recall, prediction, and loo.

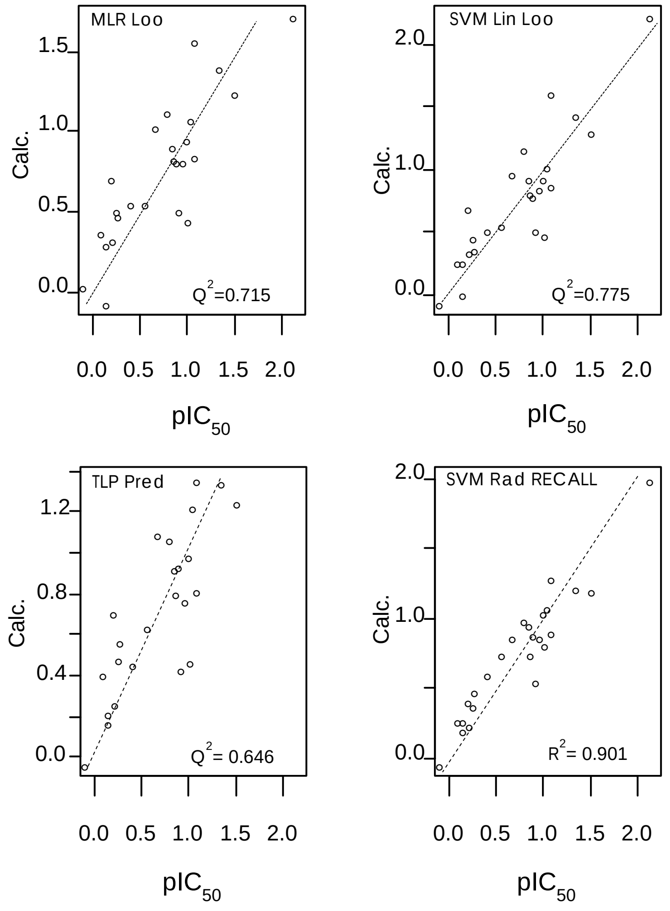


Figure 3. Examples of correlations between calculated and experimental pIC50 values.

4. CONCLUSION

In this paper, we revisited the antileishmanial activity of a population of 32 newly synthesized analogues of piplartine, a crucial question in view of the large and recent diffusion of this disease. Several topology-based 2D correlation methods were developed from a common selection of three structural descriptors extracted with the QSARINS software from a large initial pool generated by the PaDEL software. Beside MLR analysis, various machine learning approaches were developed with R routines (partial least squares, projection pursuit regression, support vector machine with linear or Gaussian kernel, and three layer perceptron with back propagation algorithm).

Special attention was paid to data fitting (“recall”), robustness (by cross validation) and predictive ability (*via* external validation). Consistent and satisfactory results were obtained by the various investigated methods. A slight advantage in performance is observed for the three layer perceptron that overwhelmed the more straightforward MLR approach in recall and prediction. The easy availability of the involved descriptors, real structural invariants, attainable through swift calculations and the simplicity of the MLR or TLP models from this limited data set made these approaches attractive for activity estimation of structural analogues and possibly for tentative proposal of new potentially active chemicals.

SUPPLEMENTARY MATERIALS

Used correlation methods

MLR

The objective is to build a model between a dependent (univariate) variable y_i (activity, property value…) and several independent variables (structural descriptors x_i for compound i).

$\mathbf{y} = \mathrm{X}\,\mathbf{b} + \mathbf{e},$

where **X** represents the matrix of the independent variables x_i, **b** and **e** being the column vectors of the coefficients and residuals respectively. The b coefficients are determined by minimizing the residuals by OLS method

$\mathbf{b} = (\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$

and the calculated response $\hat{\mathbf{y}}$ is:

$\hat{\mathbf{y}} = \mathbf{X}\,\mathbf{b}$

Performance in recall (fitting all data) is characterized by the determination coefficient R^2 of the correlation obtained between observed pIC_{50} and the corresponding structural descriptors. Another important information attainable in MLR is the applicability domain (AD) related to "influential" objects: those that in training have a heavy importance in the definition of the model, and in prediction, those falling outside of this AD, and that must be considered with caution. In the leverage approach, the influence of each object on the regression result (its "leverage") is given by the diagonal element h of the "Hat" matrix **H**

$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

For a study involving n training samples and p variables, objects with h larger than the threshold value

$h^* = 3(p+1) / n$ are considered outside the AD. Williams' plot (standardized residuals *vs* Hat diagonal values h) immediately highlights points outside the AD or outliers with residuals larger than 2.5 times the standard deviation (the common norm).

Machine learning approaches

Support vector machine (SVM): introduced by Vapnik [38, 39] and then largely used [26, 27] relies on two main ideas: the first one is to privilege robustness over an optimal recall, in view of a better predictive ability. The second one is to project (thanks to a kernel function) the initial data in a higher dimensional space where it may be hoped that a linear model might work better than in the initial data space. We used the very common linear kernel, $K(x,x') = x*x'$, x and x' being independent variables) and exponential kernel,

$K(x,x') = \exp(-\sigma(x-x')^2)$

The model depends on two tuneable parameters: the regularization constant C, trade-off between the complexity of the model and its precision (too large values tend to overfitting) and "epsilon", an estimation of the admissible error (roughly speaking the diameter of the "insensitive tube" around the regression line, where errors can be neglected when building up the model). An exponential kernel also involves a third parameter σ: "inverse width" of the Gaussian. It controls the 'weakening' of the Gaussian (and so the importance of most remote support vectors).

In this work, (operating on scaled descriptor values) we adjust them with a grid-search type procedure (ε varying from 0.05 to 0.40, C from 0.25 to 16 and σ from 0.05 to 0.25) and looking for the best loo performance.

Projection pursuit regression (PPR): operates on projections of the original variables along selected directions [27, 40-42]. The regression function, linking the property to structural variables, is approximated by a sum (empirically determined) of smooth non-linear ridge functions of these projections. An optimization routine based on cross-validation results allows for pursuing a sequence of projections revealing the most interesting data structures in the sample set. Here one projection was judged sufficient.

Partial least squares

PLS is one of the most popular approaches in QSARs [18, 27, 43, 45] in both regression and classification. PLS generates a limited set of orthogonal components spanning the descriptor space by linear combination of the original variables. These components are determined to best represent the variability in the descriptor space as well as in the property space [43].

Artificial neural network (ANN)

Three layer perceptron encompasses three layers of elementary units (the neurons). The input layer, fed with structural descriptors transmits weighted values to the hidden layer units. On each hidden unit, these weighted inputs are summed up and transmitted to the next layer units through a transfer function. Biases can be added. The sum on the output unit gives the calculated activity value [44]. To not multiply the number of connections (which may lead to overfitting) we restricted the hidden layer to a unique neuron in order to evaluate the interest of the ANN in its simplest form.

This architecture (4-1-1 units, and decay = 0.01 for 150 steps), common for all the trials, gave the best results.

Sup. Mat. Table 1. Statistical indices for the full set MLR model, equation (1), QSARINS calculation.

R^2	0.806
R^2adj	0.778
LOF	0.086
Kxx	0.320
DeltaK	0.117
RMSEtr	0.223
MAEtr	0.177
RSSt	1.247
CCCtr	0.892
s	0.244
F	29.0
Q^2loo	0.714
RMSEloo	0.27
MAEloo	0.22
PRESSloo	1.84
CCCloo	0.836
R^2Yscr	0.126
RMSEavYscr	0.47
Q^2Yscr	-0.32

Sup. Mat. Table 2. Calculated pIC_{50} in the various investigated methods.
Column PRED gathers for each method the predictions obtained in the different subsets.
Remember that compound ID1 is always in training sets.

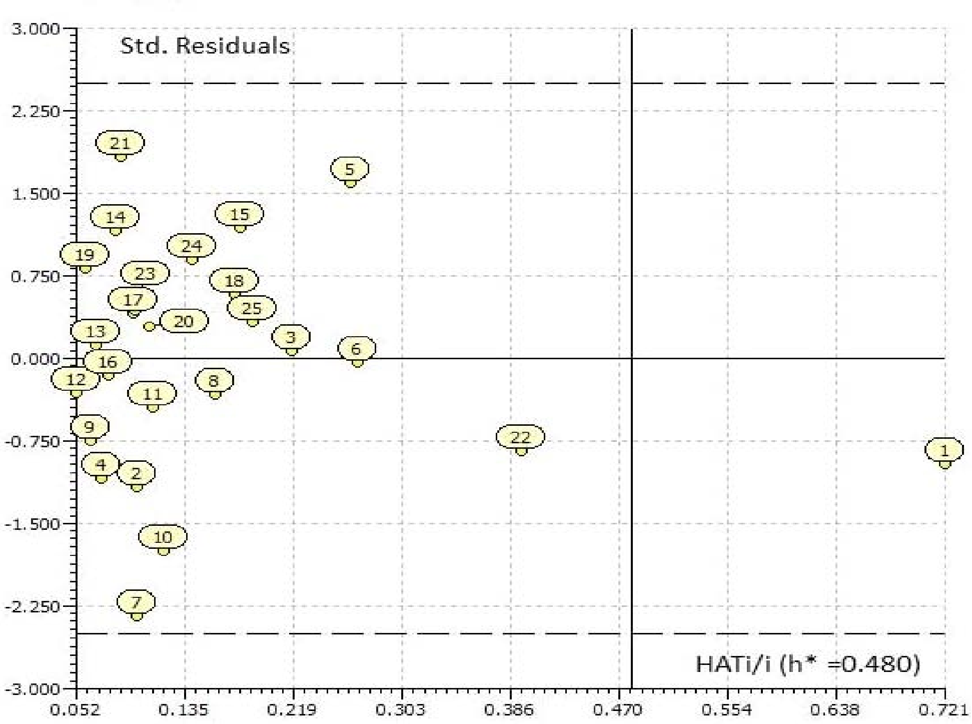
		MLR			**TLP**		
ID	ACT	RECALL	PRED	LOO	RECALL	PRED	LOO
1	2.16	2.03		1.72	2.03		1.27
2	1.54	1.27	1.21	1.24	1.30	1.24	1.24
3	1.38	1.39	1.37	1.40	1.36	1.34	1.34
4	1.12	0.87	0.86	0.85	0.90	0.81	0.87
5	1.12	1.45	1.66	1.57	1.24	1.36	1.57
6	1.08	1.07	1.03	1.07	1.19	1.22	1.32
7	1.05	0.50	0.41	0.45	0.59	0.46	0.45
8	1.04	0.97	0.95	0.95	0.99	0.98	0.98
9	1.00	0.82	0.81	0.81	0.85	0.76	0.84
10	0.96	0.56	0.46	0.51	0.54	0.43	0.45
11	0.93	0.83	0.82	0.82	0.87	0.93	0.85
12	0.90	0.83	0.79	0.83	0.82	0.80	0.81
13	0.88	0.91	0.90	0.91	0.93	0.92	0.93
14	0.83	1.10	1.09	1.12	1.12	1.07	1.17
15	0.71	0.97	1.06	1.03	0.99	1.09	1.08

Sup. Mat. Table 2 continued..

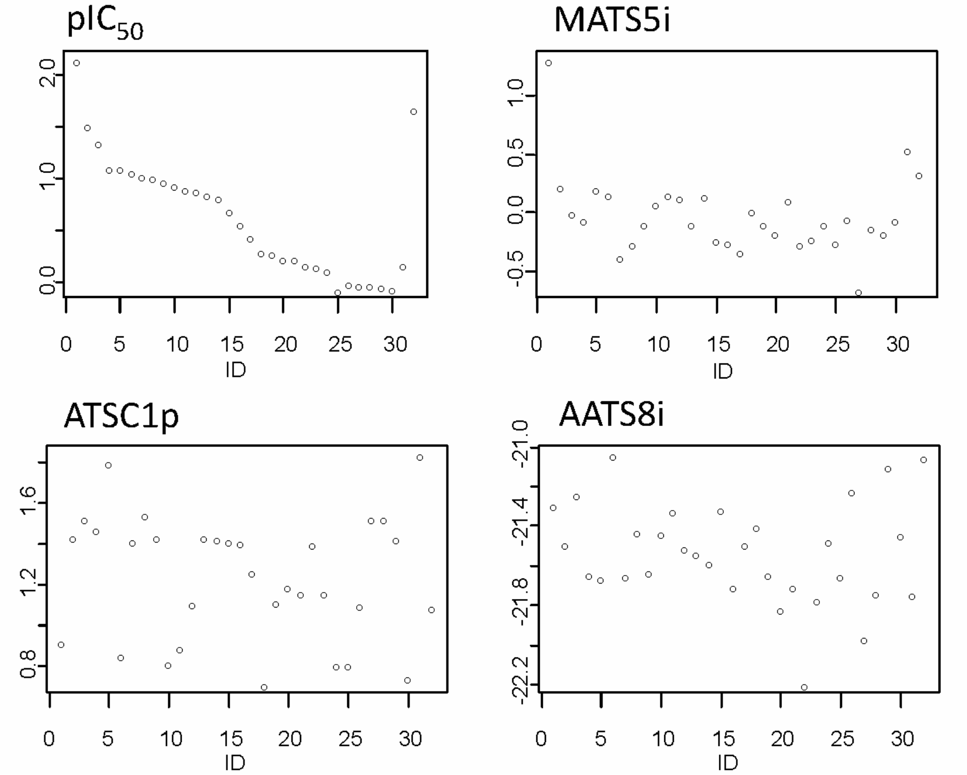
16	0.59	0.55	0.57	0.55	0.62	0.63	0.62
17	0.45	0.55	0.50	0.56	0.52	0.45	0.54
18	0.31	0.44	0.48	0.47	0.47	0.56	0.48
19	0.29	0.48	0.51	0.50	0.39	0.48	0.41
20	0.25	0.32	0.29	0.32	0.26	0.26	0.26
21	0.24	0.67	0.71	0.71	0.61	0.71	0.72
22	0.19	0.03	-0.16	-0.07	0.23	0.16	0.32
23	0.18	0.28	0.30	0.29	0.22	0.21	0.23
24	0.13	0.34	0.39	0.37	0.28	0.41	0.32
25	-0.06	0.02	-0.12	0.03	-0.02	-0.04	0.01
		SVM	**Lin.**		**SVM**	**Rad.**	
ID	ACT	RECALL	PRED	LOO	RECALL	PRED	LOO
1	2.16	2.10		2.22	2.00		0.85
2	1.54	1.32	1.27	1.30	1.20	1.10	1.17
3	1.38	1.43	1.20	1.43	1.22	1.20	1.22
4	1.12	0.90	0.88	0.87	0.90	0.87	0.88
5	1.12	1.55	1.68	1.61	1.28	1.69	1.55
6	1.08	1.03	0.98	1.02	1.08	0.92	1.02
7	1.05	0.50	0.43	0.48	0.82	0.68	0.66
8	1.04	0.99	0.81	0.92	1.05	1.06	1.05
9	1.00	0.84	0.83	0.85	0.87	0.83	0.87
10	0.96	0.51	0.47	0.51	0.55	0.47	0.51
11	0.93	0.79	0.78	0.80	0.89	0.88	0.89
12	0.90	0.82	0.81	0.82	0.75	0.62	0.73
13	0.88	0.93	0.80	0.93	0.96	0.97	0.95
14	0.83	1.14	1.13	1.17	0.99	1.08	1.14
15	0.71	0.98	1.01	0.97	0.87	0.99	1.12
16	0.59	0.56	0.57	0.56	0.75	0.68	0.83
17	0.45	0.52	0.48	0.52	0.61	0.55	0.64
18	0.31	0.37	0.49	0.37	0.47	0.57	0.48
19	0.29	0.46	0.48	0.46	0.38	0.40	0.39
20	0.25	0.30	0.31	0.35	0.23	0.13	0.24
21	0.24	0.67	0.69	0.70	0.40	0.52	0.60
22	0.19	0.05	-0.06	0.00	0.20	0.61	0.12
23	0.18	0.26	0.30	0.26	0.26	0.27	0.26
24	0.13	0.27	0.30	0.26	0.27	0.20	0.28
25	-0.06	-0.06	-0.12	-0.06	-0.05	-0.22	-0.02
		PLS			**PPR**		
ID	ACT	RECALL	PRED	LOO	RECALL	PRED	LOO
1	2.16	2.03		1.44	2.03		1.36
2	1.54	1.27	1.21	1.24	1.22	1.11	1.13
3	1.38	1.39	1.37	1.35	1.40	1.35	1.67
4	1.12	0.87	0.86	0.85	0.92	0.84	0.92
5	1.12	1.45	1.66	1.56	1.43	1.66	1.56

Sup. Mat. Table 2 continued..

6	1.08	1.07	1.04	1.07	1.05	0.99	0.56
7	1.05	0.50	0.41	0.45	0.52	0.45	0.39
8	1.04	0.97	0.95	0.94	0.97	0.97	0.85
9	1.00	0.82	0.81	0.81	0.90	0.80	0.85
10	0.96	0.56	0.46	0.51	0.57	0.46	0.52
11	0.93	0.83	0.83	0.82	0.91	0.86	0.87
12	0.90	0.83	0.79	0.83	0.89	0.85	0.88
13	0.88	0.91	0.90	0.91	0.93	0.95	0.94
14	0.83	1.10	1.09	1.12	1.04	1.03	1.07
15	0.71	0.97	1.06	1.00	0.98	1.07	1.01
16	0.59	0.55	0.57	0.55	0.57	0.74	0.90
17	0.45	0.55	0.50	0.56	0.58	0.58	0.67
18	0.31	0.44	0.48	0.47	0.42	0.47	0.47
19	0.29	0.48	0.51	0.50	0.47	0.50	0.57
20	0.25	0.32	0.29	0.33	0.26	0.27	0.25
21	0.24	0.67	0.70	0.71	0.68	0.81	0.79
22	0.19	0.03	-0.16	0.03	0.01	-0.07	-0.27
23	0.18	0.28	0.30	0.30	0.22	0.28	0.21
24	0.13	0.34	0.39	0.37	0.28	0.39	0.36
25	-0.06	0.02	-0.12	0.04	0.02	0.09	0.04



Sup. Mat. Figure 1. Williams’plot for MLR correlation (1).Numbers refer to compound ID.



Sup. Mat. Figure 2. Contributions of the three descriptors in the variations of calculated pIC_{50}.

AUTHORS' CONTRIBUTIONS

The authors equally contributed to this work.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

REFERENCES

1. Nobrega, F. R., Silva, L. V., da Silva, C., Bezerra Filho, M., Lima, T. C., Castillo, Y. P., Bezerra, D. P., Souza Lima, T. K. and de Sousa, D. P. 2019, Hindawi J. Chem., Art. ID 4785756, doi 10.1155/2019/4785756.
2. Ponte-Sucre, A., Diaz, E. and Padron-Nieves, M. 2010, Chemoinformatics: Directions Toward Combating Neglected Diseases, T. C. Ramalho (Ed.), Bentham Science.
3. Ferreira, L. L. G. and Andricopulo, A. D. 2018, Front. Pharmacol. ttps://doi.org/10.3389/fphar.2018.01278.
4. Bodiwala, H. S., Singh, G., Singh, R., Dey, C. S., Sharma, S. S., Bhutani, K. K. and Singh, I. P. 2007, J. Nat. Medicines, 61, 418.
5. Dearden, J. C. 2017, Advances in QSAR Modeling. Challenges and Advances in Computational Chemistry and Physics, 24, K. Roy (Ed), Springer International Publishing AG, 57.
6. Natarajan, R., Basak, S. C., Mills, D., Kraker, J. J. and Hawkins, D. M. 2008, Croat. Chem. Acta, 81, 333.
7. García-Domenech, R., Aguliera, J., El Moncef, A., Pocovi, S. and Gálvez, J. 2010, Mol. Diversity, 14, 321.
8. Dubois, J. E., Doucet, J. P., Panaye, A. and Fan, B. T. 1999, Topological indices and related descriptors in QSAR and QSPR, J. Devillers and A.T. Balaban (Eds), Gordon and Breach Science Publishers, The Netherlands, 613.
9. Gozalbes, R, Doucet, J. P. and Derouin, F. 2002, Current Drug Targets-Infectious Disorders, 2, 93.
10. Yap, C. W. 2011, J. Comput. Chem., 32, 1466.
11. Gramatica, P., Chirico, N., Papa, E., Cassani, S. and Kovarich, S. 2013, J. Comput. Chem., 34, 2121.

12. Gramatica, P. 2007, Comb. Sci., 26, 694.
13. Doucet, J. P. and Doucet–Panaye, A. 2018, Vector Biology Journal, 1000127.
14. Doucet, J. P., Papa, E., Doucet-Panaye, A. and Devillers, J. 2017, SAR QSAR Environ. Res., 28, 451.
15. Ren, S. 2003, J. Chem. Inf. Comput. Sci., 43, 1679.
16. Yao, X. Y., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., Hu, Z. D. and Fan, B. T. 2004, J. Chem. Inf. Comput. Sci., 44, 1257.
17. Panaye, A., Fan, B. T., Doucet, J. P., Yao, X. J., Zhang, R. S., Liu, M. C. and Hu, Z. D. 2006, SAR QSAR Environ. Res., 17, 75.
18. Thissen, U., Pepers, M., Üstün, B., Melssen, W. J. and Buydens, L. M. C. 2004, Chemom. Intell. Lab. Syst., 73, 169.
19. Tetko, I. V., Solovev, V. P., Antonov, A. V., Yao, X. J., Doucet, J. P., Fan, B. T., Hoonakker, F., Fourches, D., Jost, P., Lachiche, N. and Varnek, A. 2006, J. Chem. Inf. Model., 46, 808.
20. Xue, C. X., Zhang, R. S., Liu, H. X., Liu, M. C., Hu Z. D. and Fan, B. T. 2004, J. Chem. Inf. Comput. Sci., 44, 1267.
21. Tanabe, K., Kurita, T., Nishida, K., Lučić, B., Amić, D. and Suzuki, T. 2013, SAR QSAR Environ. Res., 24, 565.
22. Golmohammadi, H. and Dashtbozorgi, Z. 2016, SAR QSAR Environ. Res., 27, 977.
23. Ngo, T. D., Tran, T. D., Le, M. T. and Thai, K. M. 2016, SAR QSAR Environ. Res., 27, 747.
24. Drgan, V., Župerl, Š., Vračko, M., Como, F. and Novič, M. 2016, SAR QSAR Environ. Res., 27, 501.
25. Gupta, S., Basant, N., Mohan, D. and Singh, K. P. 2016, SAR QSAR Environ. Res., 27, 539.
26. Doucet, J. P., Barbault, F., Xia, H. R., Panaye, A. and Fan, B. T. 2007, Curr. Comput-Aided Drug Des., 3, 263.
27. Doucet, J. P. and Panaye, A. 2010, Three-dimensional QSAR, Applications in Pharmacology and Toxicology. CRC press, Boca Raton, FL.
28. Papa, E., Doucet, J. P. and Doucet-Panaye, A. 2015, SAR QSAR Environ. Res., 26, 647.
29. Papa, E., Doucet, J. P. and Doucet-Panaye, A. 2016, RSC Adv., 6, 68806.
30. Papa, E., Doucet, J. P., Sangion, A. and Doucet-Panaye, A. 2016, SAR QSAR Environ. Res., 27, 521.
31. Winkler, D. A., Burden, F. R., Yan, B., Weissieder, R., Tassa, C., Shaw, S. and Epa, V. C. 2014, SAR QSAR Environ. Res., 25, 161.
32. Toropova, A. P., Toropov, A. A., Benfenati, E., Puzyn, T., Leszczynska, D. and Leszczynski, J. 2014, Ecotoxicol. Environ. Safe, 108, 203.
33. OECD, 2007, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2, 1; doi:10.1787/9789264085442-en.
34. Todeschini, R., Consonni, V. and Maiocchi, A. 1999, Chemom. Intell. Lab. Sys., 46, 13.
35. Roy, K., Das, R. N., Ambure, P. and Aher, R. B. 2016, Chemom. Intell. Lab. Syst., 152, 18.
36. Consonni, V., Ballabio, D. and Todeschini, R. 2010, J. Chemom., 24, 194.
37. Hawkins, D. M., Basak, S. C. and Mills, D. 2003, J. Chem. Inf. Model., 43, 579.
38. Vapnik, V. N. 1995, The Nature of Statistical Learning Theory. Springer, New York, NY.
39. Cortes, C. and Vapnik, V. N. 1995, Mach. Learn., 20, 273.
40. Friedman, J. H. and Stuetzle, W. 1981, J. Am. Stats. Assoc., 76, 817.
41. Hu, R., Doucet, J. P., Delamar, M. and Zhang, R. 2009, Eur. J. Med. Chem., 44, 2158.
42. Doucet, J. P., Doucet-Panaye, A. and Papa, E. 2019, Mol. Inf., 38, 1900029.
43. Wold, S., Sjostrom, M. and Eriksson, L. 2001, Lab. Syst., 109-130.
44. Devillers, J. 1996, Neural Networks in QSAR and Drug Design. Academic Press, London.
45. Liew, C. Y. and Yap, C. W. 2012, Current modeling methods used in QSAR/QSPR, in Statistical Modeling of Molecular Descriptors used in QSAR/QSPR, M. Dehmer, K. Varmuza. and D. Bonchev (Eds.), Wiley-VCH Verlag GmbH, Weinheim, Germany, 1.
46. R Development Core Team, 2014, R. a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
47. Kuhn, M. 2008, J. Stat. Soft., 28, 1.
48. Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. 2004, J. Stats. Soft., 11, 1, software available at http://www.jstatsoft.org/v11/i09.