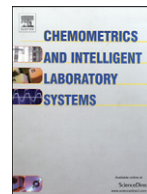




Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

QSAR study of the DPPH radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes, halophenols and caffeic acid analogues

Adam Lee^a, Andrew G. Mercader^{b,c,*}, Pablo R. Duchowicz^b, Eduardo A. Castro^b, Alicia B. Pomilio^c^a Department of Chemistry, Durham University, Durham DH1 3LE, United Kingdom^b Instituto de Investigaciones Físicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina^c PRALIB (UBA-CONICET), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956, C1113AAD Buenos Aires, Argentina

ARTICLE INFO

Article history:

Received 11 January 2012

Received in revised form 5 March 2012

Accepted 30 March 2012

Available online 11 April 2012

Keywords:

QSAR

Benzo[b]thiophenes

Di(hetero)arylamines

Halophenols

Caffeic acid analogues

Radical scavenging activity

ABSTRACT

We performed a predictive analysis based on Quantitative Structure–Activity Relationships (QSAR) of the radical scavenging activities of a set of compounds consisting of di(hetero)arylamines derivatives of benzo[b]thiophenes, halophenols, and caffeic acid analogues. Given the importance of this activity in medicinal chemistry it is of interest to develop a theoretical method for its prediction. The selection of the descriptors from a pool containing more than a thousand geometrical, topological, quantum-mechanical and electronic types of descriptors was performed using a new advanced version of the Enhanced Replacement Method (ERM). The best QSAR linear model was constructed using 52 molecular structures not previously used in this type of quantitative structure–property study, and showed good predictive attributes. The model analysis suggested that the activity depends on the atomic van der Waals volumes and on the atomic electro-negativity; and that the conformation of the molecule does not present a relevant role in the activity.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Free radicals play an important role in many physiological and pathological conditions [1]. As a general rule, excess free radicals caused by an imbalance between the creation and neutralization of radicals (known as scavenging), can lead to the development of many different diseases [2].

The disruption of living cells resulting from free radical imbalance can damage membranes, proteins, enzymes and DNA, increasing the risk of diseases such as cancer, Alzheimer's, Parkinson's, angiocardopathy, arthritis, asthma [3], diabetes, and degenerative eye disease [4,5].

By understanding the action of free radicals and their spread mechanism *in vivo*, the associated damage can be decreased and therefore reduce the expense of treating consequent diseases [6].

The diarylamine skeleton and the benzo[b]thiophene system [7] are often present in biologically active compounds, and there exist many examples of biological activities found for small molecules based on the benzo[b]thiophene moiety. Namely, they can be inhibitors of herpes simplex virus type I (HSV-1) replication, antimetotics, inhibitors of cysteine

and serine proteases, opioid receptor analgesics, and 5-HT₆ antagonists, making this a very attractive structure for medicinal chemists [8].

Bromophenols frequently isolated from various marine algae, ascidians, and sponges have attracted much research interest due to their wide spectrum of bioactivities, including antioxidative, antithrombotic, antimicrobial, anti-inflammatory, enzyme inhibition, cytotoxic and feeding deterrent activities [9–11].

Caffeic acid and its analogues are widely distributed in the plant kingdom and are found in coffee beans, olives, propolis, fruits, and vegetables. They are usually found as various simple derivatives, including amides, esters, sugar esters, and glycosides, or sometimes in rather more complex forms such as rosmarinic acid (dimer), lithospermic acid (trimer), verbascoside (heterosidic ester and glycoside of dihydroxyphenethyl-ethanol and caffeic acid), and flavonoid-linked derivatives [12]. The physiological functionality of caffeic acid, and its analogues has attracted much attention and has been studied by many research groups in recent years. The compounds are known to have antibacterial, antiviral, anti-inflammatory, antiatherosclerotic, antioxidative, antiproliferative, cytotoxic, immunostimulatory, neuroprotective and antifungal properties. These properties are associated with either their roles as antioxidants and enzyme inhibitors, or their binding activity with specific receptors [6,13,14].

A generally accepted remedy for the lack of experimental data in complex chemical phenomena is the analysis based on Quantitative Structure–Activity Relationships (QSAR) [15,16]. As a result, there exists a permanently renewed interest focused on the development of such

* Corresponding author at: Instituto de Investigaciones Físicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina. Tel.: +54 11 6091 3759; fax: +54 11 6091 2100; int: 3759.

E-mail addresses: andrewgmercader@gmail.com (A.G. Mercader), pomilio@ffyba.uba.ar (A.B. Pomilio).

kind of predictive models [17–20]. Recently, a QSAR study on the radical scavenging activity of the di(hetero)arylamine derivatives of benzo[b]thiophenes was reported [21] using a pairwise correlation analysis as the descriptor selection methodology.

The pairwise correlation analysis consists on: selecting the descriptor with the highest correlation coefficient to the experimental data, then all the remaining molecular descriptors with a correlation coefficient to that selected descriptor higher than 0.75 ($r > 0.75$) are classified as collinear and are not included in the model. Afterwards the same procedure is repeated on the molecular descriptor, still remaining on the list, with the highest correlation to experimental data. The process is continued until reaching the end of the list [21].

The main objective of the present study, is to develop a model for the prediction of the radical scavenging activity of a larger set than the previously reported one, in order to reach a more reliable, wider and more general model. In addition, a model using the original dataset and a different descriptor selection technique will be developed, aiming to make a progress on the previous findings. To carry out these tasks, a large number of structural molecular descriptors, including the definitions of all classes, were explored by the recently proposed Enhanced Replacement Method (ERM) [22], to select the best subset of variables. All data were obtained by a universal DPPH radical scavenging assay [23], to ensure continuity with the results. Results were obtained using the di(hetero)arylamine derivatives of benzo[b]thiophenes [21], halophenols [24] and caffeic acid amide and ester analogues [6].

2. Methods

2.1. Data set

In the present study we used two different size training and test sets. The first set previously reported in a QSAR study by Abreu et al. [21] was composed of 26 di(hetero)arylamine derivatives of benzo[b]thiophenes. To carry out a fair comparison, the original 18 molecule training set and 8 molecule test set from the previous report were used.

For future references, it is important to state that in the previous report there were some errors in the dataset that required many tests to check and correct. These errors were as follows: i) in the representation of the molecule **5** there was a hydroxyl at position R_2 that should have been at position R_1 ; ii) structure representation of compound **6** presented a similar mistake showing fluorine at position R_2 , whereas in the correct representation, it should have been at R_1 ; iii) the molecule **24** had a CN group at position R_1 when the correct position should have been R_2 ; and iv) molecules **17** and **18** were switched one by the other. To verify these mistakes the original papers with the experimental data were explored [25,26] and the authors were consulted.

The second structurally wider set was conformed by the addition of 18 halophenols [24], 5 caffeic acid analogues and 3 reference compounds [6] to achieve a 52 molecule total set, which was divided into 38 training and 14 test set molecules. The training and test set compounds were separated at random with the constraint that the training set compounds should have a normal distribution in the experimental values.

The experimental radical scavenging activity of the samples was expressed in terms of IC_{50} (concentration in mol/l required for a 50% decrease in absorbance of DPPH• radical at 517 nm). The halophenols added in the second test set were measured with a novel amperometric method [27], assuming that this difference will not have a significant effect on the data since there is an excellent correlation between the amperometric method and classic spectroscopic method ($R^2 = 0.9993$) [27]. Table 1 summarizes the molecular structures and experimental $-\log IC_{50}$ of the antioxidants mentioned above.

2.2. Molecular descriptors

The structures of the compounds were firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in

the Hyperchem 6.03 package [28], and the resulting geometries were further refined by means of the semiempirical method AM1 (Austin Method 1) using the Polak-Ribière algorithm and a gradient norm limit of $0.01 \text{ kcal.}\text{\AA}^{-1}$. The molecular descriptors were computed using the software e-Dragon [29,30] which calculates parameters of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups and Atom-Centered Fragments [31]. E-Dragon requires MOL2 (Sybyl) input format; in order to translate the structures to this format the software Open Babel was employed [32]. The initial descriptor matrix from e-Dragon contained 1600 descriptors. Nevertheless, some molecules did not permit the calculation of a number of descriptors; after removing these descriptors and linear dependant ones, the resulting total pool consisted of $D = 1273$ and 1327 descriptors for the datasets with $N = 26$ and 52 respectively.

2.3. Model search

It is our purpose to search the set D , containing D descriptors, for an optimal subset d , with $d \ll D$, and with minimal standard deviation S ,

$$S = \sqrt{\frac{1}{(N-d-1)} \sum_{i=1}^N res_i^2} \quad (1)$$

by means of the Multivariate Linear Regression (MLR) technique. In this equation N is the number of molecules in the training set, and res_i the residual for molecule i , is the difference between the experimental property (\mathbf{p}) and predicted property (\mathbf{p}_{pred}). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where \mathbf{d} is a point in a space of size $D!/[(d!(D-d)!)]$. A full search (FS) of optimal variables is impractical because it requires $D!/[(d!(D-d)!)]$ linear regressions. Therefore, an alternative method is necessary. We selected the optimum set of descriptors using a new advanced version of the Enhanced Replacement Method (ERM) [22] as a search algorithm that produces linear regression QSAR models with results similar to the FS, but with much less computational work. This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, \dots, X_d\}$. The ERM [33] gives models with better statistical parameters than the Forward Stepwise Regression procedure [34], and the more elaborated Genetic Algorithms [35].

ERM combines two algorithms, first a Replacement Method (RM) [36–38] is used, then a Modified RM (MRM) [33], and finally a RM is used again.

The RM consists of the following steps:

- An initial set of descriptors \mathbf{d}_k is chosen at random, one of the descriptors is replaced, denoted as X_{ki} , with all the remaining $D-d$ descriptors, one by one, and the set with the smallest value of S is kept. That is what is defined as a 'step'
- From this resulting set, the descriptor with the greatest standard deviation in its coefficient is chosen (the one changed previously is not considered) and substituted with all the remaining $D-d$ descriptors, one by one. This procedure is repeated until the set remains unmodified. In each cycle the descriptor optimized in the previous one is not modified. Thus, the candidate $\mathbf{d}_m^{(i)}$ that comes from the so-constructed path i is obtained.
- It should be noticed that if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of S , then that descriptor is not changed.

The above process is carried out for all the possible paths $i = 1, 2, \dots, d$ and the point \mathbf{d}_m with the smallest standard deviation: $\min_i S(\mathbf{d}_m^{(i)})$ is kept.

MRM follows the same strategy as RM except that in each step the descriptor with the largest error is substituted even if that substitution is not accompanied by a smaller value of S (the next smallest value of S is chosen). The main difference in MRM is that it adds some sort of noise that prevents the selected model to stay in a local minimum of S [33].

For the theoretical validation of all models, we chose the well-known Leave-One-Out (loo) and the Leave-More-Out Cross-Validation procedures ($l-n\%-o$) [39], where $n\%$ accounts for the number of molecules removed from the training set. We generated 1,000,000 cases of random data removal for 5 molecules in the case of Leave-More-Out. In our calculations we used the computational environment Matlab 5.0 (MathWorks, Natick, Massachusetts, U.S.A.).

The applicability domain (AD) for the QSAR models was explored in order to obtain a reliable prediction for external samples. The AD is a theoretical region in the chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors [40]. The AD can be characterized in various ways such as the leverage approach [41], which allows to verify whether a new chemical can be considered as interpolated and with reduced uncertainty or extrapolated outside the domain. If it is outside the model domain, a warning must be given. The leverage (h) is defined as [41]:

$$h_i = \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T \quad (i = 1, \dots, M) \quad (2)$$

where \mathbf{x}_i is the descriptor row-vector of compound i , M is the number of compounds in the dataset, and \mathbf{X} is the $N \times d$ matrix of the training set (d is the number of model descriptors, and N is the number of training set samples). The leverage is suitable for evaluating the degree of extrapolation, its limit of normal values is set as $h^* = 3(N+1)/M = 3(\Sigma h_i + 1)/M$, and a leverage greater than h^* for the training set means that the chemical is highly influential in determining the model, while for the test set, it means that the prediction is the result of substantial extrapolation of the model and may not be reliable.

The standardized residual (σ) for molecule i is defined as:

$$\sigma_i = \frac{\text{res}_i}{S_{tr}} \quad (3)$$

where res_i is the residual of molecule i and S_{tr} is the standard deviation of the training set.

In order to visualize the AD of a QSAR model a Williams plot of standardized residuals (σ) vs leverage values (h) can be used to obtain an immediate and simple graphical detection of both the response outliers (Y outliers) and the structurally influential chemicals (X outliers) of a model.

3. Results and discussion

By means of the ERM we searched the total pool of $D = 1273$ descriptors in the first set of molecules and obtained two optimal models with $d = 2$ and $d = 4$ parameters linking the molecular structure of the compounds with their radical scavenging activity. The optimal QSAR model according to ERM for two descriptors was:

$$-\log \text{IC}_{50} = 2.6052(\pm 0.1) + 0.725(\pm 0.01)\text{T}(\text{O}..S) + 2.647(\pm 0.4)\text{MATS5v} \quad (4)$$

$$N = 18, R = 0.9444, S = 0.2765, FIT = 5.6244, p < 10^{-4} \\ R_{loo} = 0.9198, S_{loo} = 0.3314, R_{l-25\%-o} = 0.8532, S_{l-25\%-o} = 0.7267 \\ RMSE_{TS} = 0.3433$$

The optimal four descriptor model was:

$$-\log \text{IC}_{50} = 3.4908(\pm 0.3) + 2.0703(\pm 0.3)\text{MATS5v} + 0.0762(\pm 0.005)\text{G}(\text{O}..S) \\ - 6.4439(\pm 1)\text{Ds} + 5.7488(\pm 1)\text{HATS1m} \quad (5)$$

$$N = 18, R = 0.9861, S = 0.1501, FIT = 13.472, p < 10^{-3} \\ R_{loo} = 0.9681, S_{loo} = 0.2280, R_{l-25\%-o} = 0.9426, S_{l-25\%-o} = 0.3425 \\ RMSE_{TS} = 0.2993$$

Using ERM the 1327 molecular descriptors of the second set of molecules were searched, arriving at the following four parameter optimal model:

$$-\log \text{IC}_{50} = 31.5327(\pm 2) - 1.1331(\pm 0.2)\text{MATS6v} + 2.4862(\pm 0.2)\text{MATS5e} \\ + 1.021(\pm 0.1)\text{EEig15x} - 7.0408(\pm 0.6)\text{BEHv1} \quad (6)$$

$$N = 38, R = 0.9374, S = 0.2039, FIT = 4.4247, p < 10^{-5} \\ R_{loo} = 0.9139, S_{loo} = 0.2378, R_{l-12\%-o} = 0.8349, S_{l-12\%-o} = 0.3350 \\ RMSE_{TS} = 0.3642$$

Here, the absolute errors of the regression coefficients are given in parentheses; p is the significance of the model, and $RMSE_{TS}$ stands for root mean squared errors of the test set.

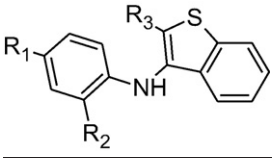
A summary of the linear models calculated by ERM is shown in Table 2. The details of the molecular descriptors of Table 2 are displayed in Table 3.

To demonstrate that Eqs. (4), (5) and (6) are not the result of happenstance, we resorted to a widely used approach to establish the model robustness: the so-called y-randomization [42]. It consists of scrambling the experimental \mathbf{p} property, so that activities do not correspond to the respective compounds. After analyzing 1,000,000 cases of y-randomization, the smallest S value obtained in this way was 0.3620, 0.2574 and 0.3240 respectively. These values were larger than those coming from the true calibration (0.2765, 0.1501 and 0.2039 respectively). For Eqs. (5) and (6), the results suggest that the models are robust, that their calibration is not a fortuitous correlations, and that we have derived reliable structure–activity relationships. In the case of Eq. (4) the difference between the y-randomization and the calibration is lower; however, as mentioned, this model is only presented as a contrast of Eq. (5) and previous findings.

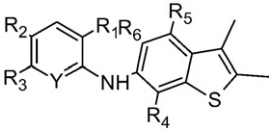
A complete contrast of the models shown in Eqs. (4) and (5) with the previously reported model [21] was not possible due to the above mentioned errors in the dataset, which prevented a complete reproduction of the previous work. Nevertheless, by looking at the calibration parameters of the previous work ($R = 0.9386$, $S = 0.2731$, $R_{loo} = 0.9187$, $R_{l-25\%-o} = 0.9039$) it can be concluded that Eq. (5) outperforms the previous model and that Eq. (4) using only two parameters shows a similar performance. The root mean square errors for the test set of the previous work ($RMSE_{TS} = 0.2216$) is lower than the one from Eq. (5), nevertheless this value is also lower than the standard deviation of models calibration implying that it could be fortuitous. The model in Eq. (5) is superior to the model in Eq. (4) in its calibration and validation parameter; Eq. (4) is only presented as a comparison to previous findings.

The plot of values predicted by Eq. (6) vs. experimental $-\log \text{IC}_{50}$ shown in Fig. 1 suggests that the 38 compounds from the training set and 14 from the test set fit a straight line. The predicted radical scavenging activities given by Eq. (6) for the training and test sets are shown in Table 1. The Williams plot of the standardized residual in terms of the leverages illustrated in Fig. 2 shows that most compounds lie within the AD of Eq. (6) and were calculated correctly. Compounds 50 and 13 are X outliers of the training set reinforcing the model. Compound 12 slightly exceeds the h limit, but is between compounds 50 and 13 of the training set and hence its prediction can be considered as reliable as those of the training chemicals. Compound

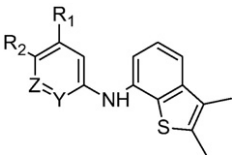
Table 1
Structure of compounds, experimental $-\log IC_{50}$, predicted $-\log IC_{50}$ by Eq. (6), predicted $-\log IC_{50}$ by Eq. (5), and residuals. (Uppercase “t” indicates test set substances for the complete dataset and uppercase “u” test set of the first smaller data set.).



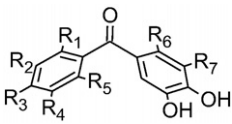
Compound	R ₁	R ₂	R ₃	Exp.	Pred. Eq. (6)	Res.	Pred. Eq. (5)	Res.
1	OCH ₃	OCH ₃	H	4.06	3.79	0.27	4.08	−0.02
2	OCH ₃	OCH ₃	COOH	3.61	3.67	−0.06	3.66	−0.05
3^u	OCH ₃	OCH ₃	COOCH ₂ CH ₃	3.77	3.75	0.02	3.82	−0.05
4	OH	OH	COOCH ₂ CH ₃	3.84	3.85	−0.01	3.81	0.03
5	OH	H	COOCH ₂ CH ₃	3.91	3.69	0.22	3.84	0.07
6^{t,u}	F	H	H	3.81	3.85	−0.04	3.35	0.46



Compound	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	Y	Exp.	Pred. Eq. (6)	Res.	Pred. Eq. (5)	Res.
7^t	OCH ₃	OCH ₃	H	H	H	CH ₃	C	4.07	3.77	0.30	4.06	0.01
8^t	H	OCH ₃	OCH ₃	H	H	CH ₃	C	4.3	4.06	0.24	4.27	0.03
9	H	OCH ₃	H	CH ₃	CH ₃	H	C	3.89	4.12	−0.23	4.01	−0.12
10^u	H	OCH ₃	H	H	H	CH ₃	C	3.38	3.21	0.17	4.06	0.08
11	H	H	OCH ₃	H	H	CH ₃	C	2.39	2.65	−0.26	3.29	0.09
12^{t,u}	H	CHO	H	H	H	CH ₃	C	2.29	2.48	−0.19	2.88	−0.49
13	H	CN	H	CH ₃	CH ₃	H	C	3.81	4.06	−0.25	1.99	0.30
14	Br	OCH ₃	OCH ₃	H	H	CH ₃	C	3.98	3.73	0.25	3.85	−0.04
15	Br	OCH ₃	H	CH ₃	CH ₃	H	C	3.69	3.76	−0.07	4.13	−0.15
16^{t,u}	Br	OCH ₃	H	H	H	CH ₃	C	3.69	3.76	−0.07	3.85	−0.16
17^t	Br	H	H	CH ₃	CH ₃	H	C	3.05	2.76	0.29	2.93	0.12
18^u	Br	H	H	H	H	CH ₃	C	2.43	2.69	−0.26	2.70	−0.27
19	I	H	H	H	H	CH ₃	C	2.91	2.82	0.09	2.93	−0.02
20^t	H	H	H	CH ₃	CH ₃	H	N	2.1	2.93	−0.83	2.30	−0.20

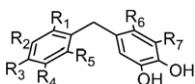


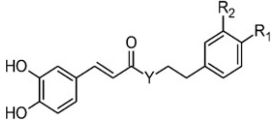
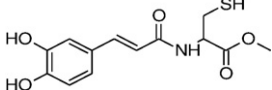
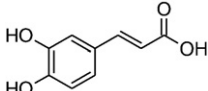
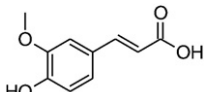
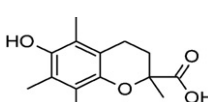
Compound	R ₁	R ₂	Y	Z	Exp.	Pred. Eq. (6)	Res.	Pred. Eq. (5)	Res.
21^u	OCH ₃	OCH ₃	C	C	4.26	3.94	0.32	3.88	0.38
22	OCH ₃	H	C	C	2.81	3.11	−0.30	2.77	0.05
23	H	OCH ₃	C	C	4.3	4.10	0.20	4.15	0.15
24^t	H	CN	C	C	1.84	2.31	−0.47	2.10	−0.26
25	H	H	C	N	2.5	2.73	−0.23	2.49	0.01
26^{t,u}	H	H	N	C	2.26	2.96	−0.70	2.12	0.14



Compound	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	Exp.	Pred. Eq. (6)	Res.
27	H	Br	OH	H	H	Br	H	4.21	3.81	0.40
28	OH	H	H	Br	H	Br	H	4.10	4.17	−0.07
29	H	H	H	H	H	Br	Br	4.06	4.13	−0.08
30	H	H	H	H	H	Br	H	4.08	4.03	0.04
31^t	Br	OH	OH	OH	Br	H	H	4.02	4.11	−0.09
32	Cl	OH	OH	OH	Cl	H	H	4.06	4.22	−0.16
39	Cl	H	H	H	H	H	H	3.95	3.80	0.15
40	H	Cl	H	H	H	H	H	3.96	4.30	−0.34
41	H	H	Cl	H	H	H	H	3.99	4.01	−0.02

Table 1 (continued)

								Exp.	Pred. Eq. (6)	Res.
Compound	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇			
33	H	Br	OH	H	H	Br	H	4.06	3.93	0.13
34	OH	H	H	Br	H	Br	H	4.05	4.13	−0.08
35	H	H	H	H	H	Br	Br	4.06	3.98	0.07
36	H	H	H	H	H	Br	H	3.79	3.91	−0.13
37	Br	OH	OH	OH	Br	H	H	4.07	3.88	0.20
38 ^t	Cl	OH	OH	OH	Cl	H	H	3.99	4.18	−0.19
42	Cl	H	H	H	H	H	H	3.88	3.56	0.32
43 ^t	H	Cl	H	H	H	H	H	3.65	3.70	−0.05
44	H	H	Cl	H	H	H	H	3.69	3.73	−0.03

					Exp.	Pred. Eq. (6)	Res.
Compound	R ₁	R ₂	Y				
45	H	H	O		4.29	4.30	−0.01
46	H	H	NH		4.37	4.47	−0.10
47	OH	H	NH		4.38	4.48	−0.10
48 ^t	OH	OH	NH		4.62	4.32	0.30
Compound number	Structure				Exp.	Pred. Eq. (6)	Res.
49					4.63	4.92	−0.29
50					4.27	4.19	0.08
51 ^t					4.24	4.38	−0.14
52					4.33	4.26	0.08

24 does not present a standardized residual higher than the limit (3σ), however has leverage much greater than the warning limit leverage h^* , which means that the compound predicted response was extrapolated from the model, and therefore, the predicted value must be used with great care [41]. Chemical **20** has a standardized residual higher than the limit and hence is considered an outlier, but belongs to the model AD; this erroneous prediction could probably be attributed to wrong experimental data rather than to the molecular structure. Compound **26** is a similar case with a standardized residual close to the limit.

Fig. 3 shows the plot of values predicted by Eq. (5) vs. experimental $-\log IC_{50}$. The Williams plot of Eq. (5) (Fig. 4) shows that most compounds lie within the AD. Compound **6** is an X outlier hence its prediction was extrapolated from the model, in addition the standardized residual is close to 3σ . Chemical **12** belongs to the model AD but is a Y outlier.

The model obtained using the larger and wider set of molecules presented in Eq. (6) showed good calibration and validation parameters. Since this model is based on a more diverse and larger group of compounds, its spectrum of application will be broader. Hence, in

general terms it is preferable to the other models and for that reason further analysis will be performed on this model.

Variance inflation factor (VIF) is defined as [43]:

$$1/(1-R_k^2) \quad (7)$$

where R_k is the coefficient of determination for regression of the i^{th} independent variable on all the other independent variables.

The correlation matrix shown in Table 4 reveals that the descriptors of the linear model are not seriously inter-correlated ($VIF \leq 2.202$, $R_{ij} \leq 0.6507$), which justifies the appearance of all parameters in the equation.

The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion and/or exclusion of compounds, measured by the statistical parameters $R_{loo} = 0.9139$ and $1 - n\% - o R_{l-12\%-o} = 0.8349$. As a general rule $R_{l-n\%-o}$ should be higher than 0.71 to have a validated model [44].

The molecular descriptors appearing in the linear Eq. (6) merge two- and three-dimensional aspects of the molecular structure, and

Table 2
Linear QSAR models for the training set of $-\log \text{IC}_{50}$.

Model	N	Descriptors used	S	R	S_{100}	R_{100}	RMSE_{TS}
M1	18	$T(\text{O..S})$, MATS5v	0.2765	0.9444	0.3314	0.9198	0.3433
M2	18	MATS5v, Ds, $G(\text{O..S})$, HATS1m	0.1501	0.9861	0.2280	0.9681	0.2993
M3	38	MATS6v, MATS5e, EEig15x, BEHv1	0.2039	0.9374	0.2378	0.9139	0.3642

can be classified as follows: (i) two 2D autocorrelation descriptors: MATS6v, Moran autocorrelation – lag 6/weighted by atomic van der Waals volumes; and MATS5e Moran autocorrelation – lag 5/weighted by atomic Sanderson electronegativities; (ii) an Edge adjacency indices descriptor: EEig15x, Eigenvalue 15 from edge adj. matrix weighted by edge degrees; and (iii) a BCUT descriptor: BEHv1, highest eigenvalue n. 1 of Burden matrix/weighted by atomic van der Waals volumes.

Different structural variables introduced by Broto, Moreau, and Geary [45] account for bi-dimensional autocorrelations between pairs of atoms in the molecule, and were defined in order to reflect the contribution of a considered atomic property to the experimental observations under investigation, in this case, $-\log \text{IC}_{50}$. The atomic properties that can be adopted to differentiate the nature of atoms are the mass (m), the polarizability (p), the electronegativity (e), or the volume (v). These indices can be readily calculated, i.e.: by summing products of atomic weights (employing atomic properties such as atomic polarizabilities, molecular volumes, etc.) of the terminal atoms of all the paths of a prescribed length. For the case of MATS6v, the path connecting a pair of atoms has length 6 and involves the van der Waals volumes as weighting scheme to distinguish their nature. MATS5e involves the atomic Sanderson electronegativities as weighting scheme and has a length of 5 for the path that connects the pair of atoms.

Edge adjacency indices are topological molecular descriptors derived from the edge adjacency matrix \mathbf{E} . Derived from the molecular graph, \mathbf{E} encodes the connectivity between graph edges. It is a square symmetric matrix of dimension $B \times B$, where B is the number of bonds, and is usually derived from an H-depleted molecular graph [46]. The entries $[\mathbf{E}]_{ij}$ of the matrix equal one if edges e_i and e_j are adjacent (the two edges thus forming a path of length two) and zero otherwise. For multigraphs, the edge adjacency matrix can be augmented by a row and a column for each multiple edge. The edge degree ε_i provides

Table 3
Symbols for molecular descriptors involved in the different models.

Molecular descriptor	Type	Description
MATS5v	2D autocorrelation	Moran autocorrelation – lag 5/weighted by atomic van der Waals volumes
$T(\text{O..S})$	Topological	Sum of topological distances between O..S
$G(\text{O..S})$	Geometrical	Sum of geometrical distances between O..S
HATS1m	GETAWAY	Leverage-weighted autocorrelation of lag 1/weighted by atomic masses
MATS5e	2D autocorrelation	Moran autocorrelation – lag 5/weighted by atomic Sanderson electronegativities
Ds	WHIM	D total accessibility index/weighted by atomic electrotopological states
nBM	Constitutional	Number of multiple bonds
$T(\text{N..O})$	Topological	Sum of topological distances between N..O
MATS6v	2D autocorrelations	Moran autocorrelation – lag 6/weighted by atomic van der Waals volumes
EEig15x	Edge adjacency indices	Eigenvalue 15 from edge adj. matrix weighted by edge degrees
BEHv1	BCUT	Highest eigenvalue n. 1 of Burden matrix/weighted by atomic van der Waals volumes

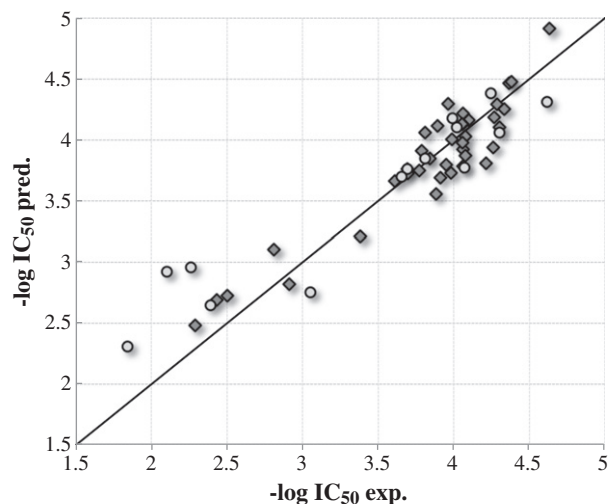


Fig. 1. Predicted (Eq. (6)) vs experimental $-\log \text{IC}_{50}$ for the training (rhombus) and test (circles) sets.

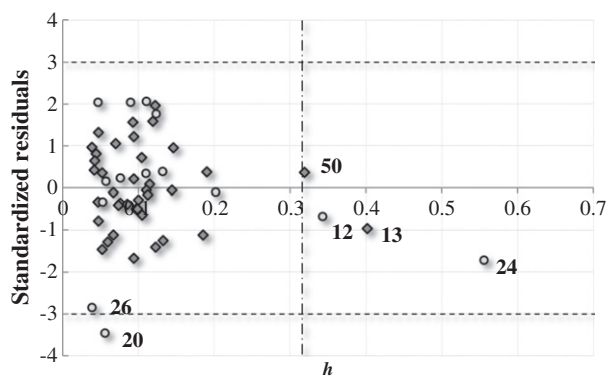


Fig. 2. Williams plot of Eq. (6) showing the Application Domain. The vertical dashed line indicates the limiting leverage h^* .

the simplest information related to the considered bond and is calculated from the edge adjacency matrix as follows:

$$\varepsilon_i = R_i(\mathbf{E}) = \sum_{j=1}^B [\mathbf{E}]_{ij} \quad (8)$$

where R_i is the row sum operator.

BCUT descriptors are the eigenvalues of a modified connectivity matrix, the Burden matrix (\mathbf{B}) [47,48]. The matrix is an H depleted molecular graph, defined as follows: diagonal elements are atomic numbers of the elements (Z_i); off diagonal elements (B_{ij}), representing bonded atoms i and j are equal to $\pi^* \cdot 10^{-1}$ where π^* is the conventional bond order (i.e. 1, 2, 3, 1.5 for single, double, triple and aromatic bonds respectively); off diagonal elements corresponding to terminal bonds are increased by 0.01 and all other matrix elements are set to 0.001. The ordered sequence of the n smallest eigenvalues of \mathbf{B} was proposed as a molecular descriptor based on the assumption that the lowest eigenvalues contain contributions from all the atoms and thus reflects topology of the molecule. The BCUT descriptors are an extension of the Burden eigenvalues and consider 3 classes of matrices, whose diagonal elements account for atomic charge related values, atomic polarizability related values and atomic H bond abilities. A variety of definitions have been used for the off diagonal terms and both 2D and 3D approaches are considered. The highest and lowest eigenvalues of these matrices have been shown to be

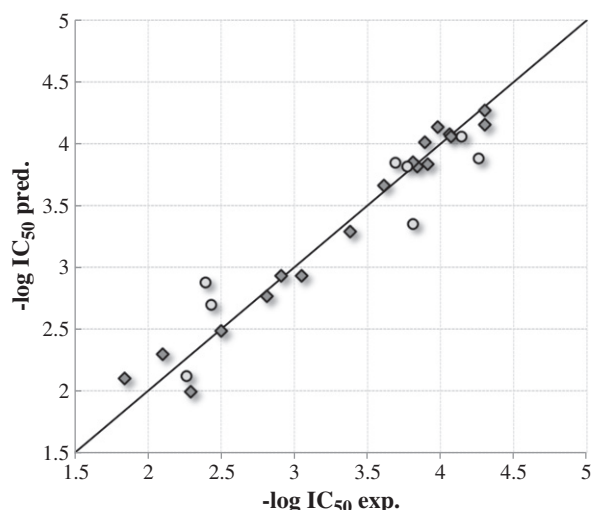


Fig. 3. Predicted (Eq. (5)) vs experimental $-\log IC_{50}$ for the training (rhombus) and test (circles) sets.

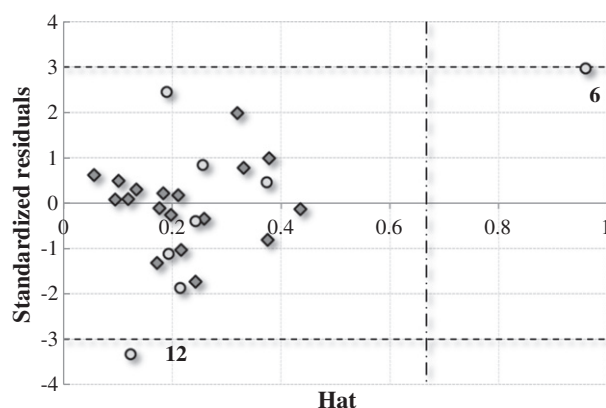


Fig. 4. Williams plot of Eq. (5) showing the Application Domain. The vertical dashed line indicates the limiting leverage h^* .

discriminating descriptors. $BEHv1$ is the highest eigenvalue of B involving atomic van der Waals volumes as weighting scheme.

The standardization of the regression coefficients of Eq. (6) allows assigning greater importance to the molecular descriptors that exhibit the largest absolute standardized coefficients [34]. In our case we have,

$$BEHv1(1.1406) > MATS5e(0.8503) > EEig15x(0.6411) > MATS6v(0.502) \quad (9)$$

where the standardized coefficients are shown in parentheses. The ranking of contributions given by Eq. (9) suggests that BCUT descriptor $BEHv1$ is the most relevant variable for the present set of molecules, thus indicating a significant dependence of the radical scavenging activity on the volume of the atoms that form the molecule. This is further

Table 4

Correlation (R_{ij}) matrix for descriptors and variance inflation factor of Eq. (6) ($N=38$). Boldface indicates highest R_{ij} and highest VIF.

	MATS6v	MATS5e	EEig15x	BEHv1	VIF
MATS6v	1	0.6485	0.2119	0.1520	1.818
MATS5e		1	0.1413	0.2370	1.857
EEig15x			1	0.6507	2.176
BEHv1				1	2.202

supported by the presence in Eq. (9) of the 2D autocorrelation $MATS6v$ that is also weighted by atomic van der Waals volumes. The second descriptor in Eq. (9) is the 2D autocorrelation $MATS5e$ implying that the activity has a significant dependence on the electronegativity of the atoms present in the molecules.

Since none of the descriptors depends on the conformation of the molecule, it is possible to argue that the radical scavenging activity of the present set of compounds has no considerable dependence on conformational changes.

4. Conclusion

In this paper we developed a predictive QSAR model for radical scavenging activity, of a group of 52 molecules that include di(hetero)arylamine derivatives of benzo[b]thiophenes, halophenols and caffeic acid analogues, using four molecular descriptors that take into account 2D- and 3D-aspects of the molecular structure. The model, showed a good predictive ability established by the theoretical and test set validations. The analysis of the model suggests that the activity depends on the atomic van der Waals volumes and electronegativity of the atoms in the molecules; the conformation of the molecule appears to have no role in the activity.

Using a previously reported smaller sized dataset containing only di(hetero)arylamine derivatives of benzo[b]thiophenes a second predictive model was obtained that presented better statistical features compared to the original work.

The proposed models are useful tools in the prediction of the antioxidant activity, in a fast and costless manner, for any future studies that may require an estimate of this activity, such as the determination of candidates for synthesis.

Acknowledgments

The authors want to thank the National Research Council of Argentina (CONICET) and Universidad de Buenos Aires (UBA) for financial support; MINCYT for the electronic library facilities. ABP is a Senior Research Member of CONICET; AGM thanks CONICET for a Post-Doctoral Research Fellowship in PRALIB (CONICET, UBA), on leave from INIFTA (CONICET, UNLP). AL thanks IAESTE International for a research fellowship in INIFTA, on leave from Durham University.

Appendix A. Supplementary data

The descriptor matrix of the dataset along with the compounds experimental activity vector and descriptor name string is available as supporting information in tabulated text format. Supplementary related to this article can be found online at doi:10.1016/j.chemolab.2012.03.016.

References

- [1] G. Vendemiale, I. Grattagliano, E. Altomare, An update on the role of free radicals and antioxidant defense in human disease, *International Journal of Clinical & Laboratory Research* 29 (1999) 49–55.
- [2] M. Valko, D. Leibfritz, J. Moncol, M.T.D. Cronin, M. Mazur, J. Telser, Free radicals and antioxidants in normal physiological functions and human disease, *The International Journal of Biochemistry & Cell Biology* 39 (2007) 44–84.
- [3] E. Keinan, A. Alt, G. Amir, L. Bentur, H. Bibi, D. Shoseyov, Natural ozone scavenger prevents asthma in sensitized rats, *Bioorganic & Medicinal Chemistry* 13 (2005) 557–562.
- [4] M.S. Cooke, M.D. Evans, M. Dizdaroglu, J. Lunec, Oxidative DNA damage: mechanisms, mutation, and disease, *The FASEB Journal* 17 (2003) 1195–1214.
- [5] T.-T. Wang, G.-C. Zeng, X.-C. Li, H.-P. Zeng, In vitro studies on the antioxidant and protective effect of 2-substituted-8-hydroxyquinoline derivatives against H_2O_2 -induced oxidative stress in BMSCs, *Chemical Biology & Drug Design* 75 (2011) 214–222.
- [6] S. Son, B.A. Lewis, Free radical scavenging and antioxidative activity of caffeic acid amide and ester analogues: structure–activity relationship, *Journal of Agricultural and Food Chemistry* 50 (2002) 468–472.
- [7] E. Campaigne, *Comprehensive Heterocyclic Chemistry* 4 (1984) 863–934.

- [8] K.R. Ronald, B.P. Jefery, Comprehensive Heterocyclic Chemistry II 2 (1996) 679–729.
- [9] K. Kurata, K. Taniguchi, K. Takashima, I. Hayashi, M. Suzuki, Feeding-deterrent bromophenols from *Odonthalia corymbifera*, Phytochemistry 45 (1997) 485–487.
- [10] W.L. Popplewell, P.T. Northcote, Colensoide A: a new nitrogenous bromophenol from the New Zealand marine red alga *Osmundaria colensoi*, Tetrahedron Letters 50 (2009) 6814–6817.
- [11] X.J. Duan, X.M. Li, B.G. Wang, Highly brominated mono- and bis-phenols from the marine red alga *Symphyocladia latiuscula* with radical-scavenging activity, Journal of Natural Products 70 (2007) 1210–1213.
- [12] J.J. Macheix, A. Fleuriot, J. Billot, Hydroxycinnamic acids, Fruit Phenolics (1990) 20–34.
- [13] S. Beyza Öztürk Sankaya, İ. Gülçin, C.T. Supuran, Carbonic anhydrase inhibitors: inhibition of human erythrocyte isozymes I and II with a series of phenolic acids, Chemical Biology & Drug Design 75 (2010) 515–520.
- [14] C.-M. Ma, T. Abe, T. Komiyama, W. Wang, M. Hattori, M. Daneshtalab, Synthesis, anti-fungal and 1,3-[beta]-d-glucan synthase inhibitory activities of caffeic and quinic acid derivatives, Bioorganic & Medicinal Chemistry 18 (2010) 7009–7014.
- [15] C. Hansch, A. Leo, Exploring QSAR: fundamentals and applications in chemistry and biology, Am. Chem. Soc., Washington, D.C., 1995.
- [16] R.A. Gupta, A.K. Gupta, L.K. Soni, S.G. Kaskhedikar, 2-(Pyrizin-2-yloxy)acetohydrazide analogs QSAR study: an insight into the structural basis of antimycobacterial activity, Chemical Biology & Drug Design 76 (2010) 441–450.
- [17] A.P. Toropova, A.A. Toropov, S.E. Martynov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*, Chemometrics and Intelligent Laboratory Systems 110 (2012) 177–181.
- [18] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on aryl-piperazine derivatives with activity on malaria, Chemometrics and Intelligent Laboratory Systems 110 (2012) 81–88.
- [19] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, S. Prachayasittikul, V. Prachayasittikul, Predicting the free radical scavenging activity of curcumin derivatives, Chemometrics and Intelligent Laboratory Systems 109 (2011) 207–216.
- [20] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometrics and Intelligent Laboratory Systems 109 (2011) 146–161.
- [21] R.M.V. Abreu, I.C.F.R. Ferreira, M.J.R.P. Queiroz, QSAR model for predicting radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes, European Journal of Medicinal Chemistry 44 (2009) 1952–1958.
- [22] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Advances in the replacement and enhanced replacement method in QSAR and QSPR theories, Journal of Chemical Information and Modeling 51 (2011) 1575–1581.
- [23] T. Yamaguchi, H. Takamura, T. Matoba, J. Terao, HPLC method for evaluation of the free radical-scavenging activity of foods by using 1,1-diphenyl-2-picrylhydrazyl, Bioscience, Biotechnology, and Biochemistry 62 (1998) 1201–1204.
- [24] W. Zhao, X. Feng, S. Ban, W. Lin, Q. Li, Synthesis and biological activity of halophenols as potent antioxidant and cytoprotective agents, Bioorganic & Medicinal Chemistry Letters 20 (2010) 4132–4134.
- [25] I.C.F.R. Ferreira, M.-J.R.P. Queiroz, M. Vilas-Boas, L.M. Estevinho, A. Begouin, G. Kirsch, Evaluation of the antioxidant properties of diarylamines in the benzo[b]thiophene series by free radical scavenging activity and reducing power, Bioorganic & Medicinal Chemistry Letters 16 (2006) 1384–1387.
- [26] M.-J.R.P. Queiroz, I.C.F.R. Ferreira, R.C. Calhelha, L.M. Estevinho, Synthesis and antioxidant activity evaluation of new 7-aryl or 7-heteroarylamino-2,3-dimethylbenzo[b]thiophenes obtained by Buchwald–Hartwig C–N cross-coupling, Bioorganic & Medicinal Chemistry 15 (2007) 1788–1794.
- [27] S. Milardovic, D. Ivekovic, B.S. Grabaric, A novel amperometric method for antioxidant activity determination using DPPH free radical, Bioelectrochemistry 68 (2006) 175–180.
- [28] HYPERCHEM, in, 6.03 (Hypercube) <http://www.hyper.com>.
- [29] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, Virtual computational chemistry laboratory – design and description, Journal of Computer-Aided Materials Design 19 (2005) 453–463.
- [30] VCCLAB, Virtual Computational Chemistry Laboratory, in, 2005.
- [31] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley VCH, Weinheim, Germany, 2000.
- [32] The Open Babel package, version 2.2.3, <http://openbabel.sourceforge.net/2006>.
- [33] A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro, Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories, Chemometrics and Intelligent Laboratory Systems 92 (2008) 138–144.
- [34] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.
- [35] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, Journal of Chemical Information and Modeling 50 (2010) 1542–1548.
- [36] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules, Chemical Physics Letters 412 (2005) 376–380.
- [37] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Alternative algorithm for the search of an optimal set of descriptors in QSAR–QSPR studies, MATCH – Communications in Mathematical and in Computer Chemistry 55 (2006) 179–192.
- [38] P.R. Duchowicz, M. Fernández, J. Caballero, E.A. Castro, F.M. Fernández, QSAR of non-nucleoside inhibitors of HIV-1 reverse transcriptase, Bioorganic & Medicinal Chemistry 14 (2006) 5876–5889.
- [39] D.M. Hawkins, S.C. Basak, D. Mills, Journal of Chemical Information and Modeling 43 (2003) 579–586.
- [40] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR and Combinatorial Science 26 (2007) 694–701.
- [41] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environmental Health Perspectives 111 (2003) 1361–1375.
- [42] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: H.V.D. Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.
- [43] J.D. Curto, J.C. Pinto, New multicollinearity indicators in linear regression models, International Statistical Review 75 (2007) 114–121.
- [44] A. Golbraikh, A. Tropsha, Beware of q²! Journal of Molecular Graphics & Modeling 20 (2002) 269–276.
- [45] G. Moreau, P. Broto, Nouveau Journal de Chimie 4 (1980) 757–764.
- [46] D. Bonchev, O. Mekenyan, Comparability graphs and electronic spectra of condensed benzenoid hydrocarbons, Chemical Physics Letters 98 (1983) 134–138.
- [47] F.R. Burden, Molecular identification number for substructure searches, Journal of Chemical Information and Computer Sciences 29 (1989) 225–227.
- [48] R.B. Frank, A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix, Quantitative Structure–Activity Relationships 16 (1997) 309–314.