SAR and QSAR
in Environmental Research

Taylor & Francis
Taylor & Francis Group

SAR and QSAR in Environmental Research

ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: http://www.tandfonline.com/loi/gsar20

Predicting the bioconcentration factor through a conformation-independent QSPR study

J. F. Aranda, D. E. Bacelo, M. S. Leguizamón Aparicio, M. A. Ocsachoque, E. A. Castro & P. R. Duchowicz

SAR AND QSAR IN ENVIRONMENTAL RESEARCH, 2017
https://doi.org/10.1080/1062936X.2017.1377765

Predicting the bioconcentration factor through a conformation-independent QSPR study

J. F. Aranda[a], D. E. Bacelo[b], M. S. Leguizamón Aparicio[c], M. A. Ocsachoque[c], E. A. Castro[a] and P. R. Duchowicz[a]

[a]Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, La Plata, Argentina; [b]Departamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Buenos Aires, Argentina; [c]Departamento de Química. Facultad de Ciencias Exactas (UNLP), Centro de Investigación y Desarrollo en Ciencias Aplicadas "Dr Jorge J. Ronco", Buenos Aires, Argentina

ABSTRACT

The ANTARES dataset is a large collection of known and verified experimental bioconcentration factor data, involving 851 highly heterogeneous compounds from which 159 are pesticides. The BCF ANTARES data were used to derive a conformation-independent QSPR model. A large set of 27,017 molecular descriptors was explored, with the main intention of capturing the most relevant structural characteristics affecting the studied property. The structural descriptors were derived with different freeware tools, such as PaDEL, Epi Suite, CORAL, Mold^2, RECON, and QuBiLs-MAS, and so it was interesting to find out the way that the different descriptor tools complemented each other in order to improve the statistical quality of the established QSPR. The best multivariable linear regression models were found with the Replacement Method variable sub-set selection technique. The proposed QSPR model improves previous reported models of the bioconcentration factor in the present dataset.

ARTICLE HISTORY
Received 17 May 2017
Accepted 6 September 2017

KEYWORDS
Bioconcentration factor (BCF); quantitative structure-property relationships; replacement method; pesticides; molecular descriptors

Introduction

The pesticides are agrochemical compounds playing a very important role in not only food production for the increasing demand of the human population, but also during the control of infectious diseases transmitted by insect-vectors and microorganisms [1]. During pesticide applications, these substances are in contact with plants or absorbed by them, although they can also be dissolved and washed away by the aqueous phase and introduced into its living organisms, thus moving to the entire food chain [2,3].

In bioconcentration processes, chemical compounds have a concentration in an organism exceeding the concentration in the surrounding environment, achieved through non-dietary routes such as respiratory and dermal surfaces [4,5]. The Bioconcentration Factor (BCF) represents the bioconcentration capability of a chemical, defined as the ratio between its

CONTACT P. R. Duchowicz pabloducho@gmail.com

The supplementary material for this article can be accessed at https://doi.org/10.1080/1062936X.2017.1377765

concentration in the organism and the concentration in water at steady-state equilibrium under laboratory conditions. The BCF parameter is an end-point of great relevance, due to its ecotoxicological impact: substances are identified as bioaccumulative when $\log \mathrm{BCF} > 3.3$, and as non-bioaccumulative under this limit. For the assessment, fish are generally used, due to their role in the food chain and the availability of standardized testing protocols [5].

Among the various methodologies available in the literature for predicting the properties of substances, i.e. BCF, and based on the knowledge of their chemical structure, the Quantitative Structure–Property Relationships (QSPR) Theory [6–8] has been widely used in past studies [9,10]. QSPR models constitute a fast and cost-effective alternative to the experimental evaluation of BCF through animal testing.

In the QSPR framework, the molecular structure is quantified by means of molecular descriptors; in other words, numerical quantities carrying specific information on the constitutional, topological, geometrical, hydrophobic, and/or electronic aspects [11–13]. Therefore, a descriptor set selected with an appropriate machine learning algorithm is statistically correlated to the experimental property under study, resulting in a mathematical model that can be used to find out useful structure–property parallelisms.

The first studies on BCF have involved its correlation to the octanol/water partition coefficient ($\log K_{ow}$) through linear, bilinear, and polynomial models [14]. During last decades, more complex models have been continuously proposed for specific regulatory purposes, such as the CAESAR model [15,16] implemented in the VEGA platform [17] (473 substances), the Meylan model [18] implemented in the Estimation Program Interface (EPI Suite) BCFBAF module [19] (527 substances), or the T.E.S.T. model [20] (598 substances) among others. However, the availability of newer and higher quality experimental BCF measures encourages the development of newer and alternative bioconcentration QSPR models with improved statistical quality [21].

A recent QSPR study performed by Gissi et al. [21] in 2014 employs the ANTARES dataset [22], a larger dataset than other ones reported previously for establishing predictive BCF models. It involves 851 highly heterogeneous compounds, from which 159 compounds are pesticides. The best results were obtained with a 9-descriptors Artificial Neural Network (ANN) QSPR model of standard architecture (one input layer, one hidden layer, and one output layer), leading to a satisfactory predictive capability. This properly validated model includes interpretable biokinetics descriptors, and is derived from a training set of 608 compounds and challenged against the validation and test sets containing 152 and 76 compounds, respectively.

This work resorts to the same large collection of known and verified experimental BCF data used by Gissi et al. [21], in order to report a new alternative QSPR model involving pesticide information. The conformation-independent QSPR approach [23–26] employed here does not consider the conformational representation of the chemical structures, by only relying on their constitutional and topological representations. It is worthy to note that this approach is not ‘geometry independent’, because also the so calculated descriptors depend on the geometry through the chemical graph. The exclusion of 3D-structural aspects avoids ambiguities due to the existence of compounds in various conformational states, which would lead to the loss of predictive capability of the QSPR model.

Materials and methods

Experimental dataset

The ANTARES dataset [21,22] includes experimental BCF values collected among various reliable and publicly available databases. This dataset involves compounds with different chemical classes, from which 159 compounds are pesticides. Compounds characterized by ambiguous data, inorganic compounds, or isomeric mixtures were discarded [21], thus leading to a set containing 851 experimental BCF values ranging in the interval −1.70–5.69. The complete list of compounds studied here is provided in Table 1S as Supplementary Material.

Structural representation and molecular descriptors calculation

The 851 molecular structures were first drawn with ACDLabs ChemSketch freeware [27] with molecules in MDL mol (V2000) format. All file format conversions were performed with Open Babel for Windows [28].

The conformation-independent molecular descriptors were computed as follows. We use the Pharmaceutical Data Exploration Laboratory (PaDEL) freeware program version 2.20 [29], because it has the advantage that it is a freely available and open source program. PaDEL allows us to calculate 1444 0D-2D descriptors and 12 fingerprint types (16,092 bits) [30]. The categorical (indicator) fingerprint descriptors involve the presence or count of specific chemical sub-structures: we treat the fingerprints like they are 'constitutional descriptors' describing the molecular composition, and, as such, these can be used for modelling any property of interest.

Five semi-empirical descriptors were calculated from the EPI Suite freeware modules [19], with molecules in SMILES format. EPI Suite uses a series of group contribution factors for calculating (in decimal logarithmic units): (i) octanol/water partition coefficient log $K_{ow}EPI$; (ii) water solubilities log $S_{w1}EPI$ and log $S_{w2}EPI$: the second parameter is based on log $K_{ow}EPI$; (iii) soil sorption partition coefficients log K_{oc1} and log K_{oc2}: the first parameter is based on the first order molecular connectivity index, while the second one is based on log $K_{ow}EPI$. We also calculate the Bioconcentration Factor log *BCFEPI* in order to compare EPI Suite′'s BCF predictions with the ones found in the present work.

Optimal molecular descriptors (*DCW*, descriptor of correlation weights) were also calculated in our QSPR study; in other words, descriptors that depend both on the molecular structure and the property under analysis (BCF), but they do not explicitly depend on the molecular conformation of compounds. We have already shown the importance of using optimal descriptors in previous QSPR studies [26,31–33]. The CORAL freeware [34] defines different kinds of optimal descriptors. The structural representation (SR) used, i.e. graph or SMILES, determines the structural attributes (SA) available for the linear model. Therefore, it is necessary to decide which SA combination is the most appropriate, and this is done in a stepwise fashion, i.e. first search for the best single SA, then search for a second SA that combines the best with the previous one, and so on. The *DCW* descriptor is a linear combination of correlation weights (*CW*), refer to Supplementary Table 2S. The *CW* was calculated for each SA in the training set through the Monte Carlo (MC) simulation method. The *DCW* depends on the threshold (*T*) and the number of epochs (n_{epochs}): the rare attributes occur in less than *T* compounds, and in this work *T* is a positive integer ($T = 0–2$). The molecules were provided to CORAL in SMILES format.

More molecular descriptors were calculated with the Molecular Descriptors from 2D structures (Mold^2) freeware [35], which generates 779 1D–2D structural variables with molecules in MDL sdf format.

Atomic charge density-based descriptors were calculated by means of the RECON 5.5 freeware [36], which encodes electronic and structural information relevant to the chemistry of intermolecular interactions. The robustness of RECON has previously been demonstrated elsewhere [37]. RECON is an algorithm for the reconstruction of molecular charge densities and charge density-based electronic properties of molecules, using atomic charge density fragments pre-computed from *ab initio* wave functions. The method is based on the Quantum Theory of Atoms in Molecules [38]. A library of atomic charge density fragments has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly. In the present case, the molecules were in SMILES format, as input for the generation of 248 Transferable Atom Equivalent (TAE) descriptors [39].

Finally, 2D molecular descriptors were calculated with the Quadratic, Bilinear, and N-Linear MapS (QuBiLs) [40] suite by using the Graph-Theoretic Electronic-Density Matrices and Atomic Weightings (MAS) module from the ToMoCoMD-CARDD free multi-platform freeware. The QuBiLs-MAS algebraic module calculates 8448 Quadratic, Bilinear, and Linear Maps, based on Pseudograph-Theoretic Electronic-Density Matrices and Atomic Weightings, when the program is used with the following options selected: 'bilinear', 'linear', and 'quadratic' algebraic forms; 'atom-based', 'non-chiral', and 'duplex' constraints; 'non-stochastic', 'simple stochastic', 'double stochastic', and 'mutual probability' matrix forms (maximum order 15); 'keep all' cut-off; 'total' groups; 'Ghose-Crippen LogP', 'Polarizability', 'Charge', 'Polar Surface Area', 'Electronegativity', 'Refractivity', 'Mass', and 'Van der Waals volume' properties; 'Euclidean distance', 'arithmetic mean', and 'standard deviation' invariants (non-standardized option).

Therefore, the total number of non-conformational molecular descriptors explored in this work was 27,017. It was our intention to capture, with such a great number of descriptors, the most relevant structural characteristics affecting the studied property.

Model development

Molecular descriptors selection in MLR

The 27,017 non-conformational molecular descriptors calculated with PaDEL, EPI Suite, CORAL, Mold^2, RECON, and QuBiLs-MAS were analysed in order to remove the 'collinear' descriptors. In this way, the linearly dependent pairs were identified, and only one variable from each pair was kept for further analysis. Therefore, non-informative descriptors that were not relevant to our QSPR analysis were excluded, such as descriptors with constant and near-constant values, and descriptors with at least one missing value. This process leads to a set containing 8122 linearly independent 0D-2D descriptors.

We employ the Replacement Method (RM) technique [41] in order to generate MLR models on the training set (train), by searching in a pool having $D = 8122$ descriptors for optimal sub-sets containing d descriptors (d is much lower than D), with smallest values for the standard deviation (S_{train}) or the root mean square error (RMS_{train}).

The main idea behind the RM is that one can approach the minimum of S_{train} by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors. In other words, we should find the global minimum of $S_{train}(d)$ in a sub-space of $D!/[d!(D-d)!]$ points d, where D represents the total number of available descriptors. The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors; although, of course, requires much less computational work. The RM is more computationally expensive than the Stepwise Regression (SR) and Genetics Algorithm (GA) approaches, although it produces similar or better results than GA and better results than SR [41,42].

Supplementary Table 2S includes a list of mathematical equations involved in the present study. All the MATLAB [43] programmed algorithms used in our calculations are available upon request.

Model validation

Several validation strategies have been proposed during the last years for the validation of a QSPR model [44,45], which consists on testing its ability to predict the property of compounds not considered during the model development. For this purpose, the complete molecular set was split into three sub-sets: training (train), validation (val), and test sets. The training set is used to calibrate the model and to obtain its parameters through the RM technique, while the validation set helps to partially validate the model. Finally, the test set includes compounds 'never seen' during the calibration step and demonstrates the true predictive capability of the QSPR.

The dataset partitioning was independent of the model building. We partitioned the dataset first, and afterwards we searched for the best linear correlations. This partitioning has to achieve similar structure–property relationships in the three sub-sets; in other words, the training set molecules should be representative of the validation and test set compounds. For this purpose, the split of the dataset was carried out by means of the Balanced Subsets Method (BSM) [46,47], a procedure proposed by our group that ensures that balanced subsets are generated. The BSM is based on the *k*-Means Cluster Analysis (*k*-MCA) method [48]: the essence of *k*-MCA is to create *k*-clusters or groups of compounds, in such a way that compounds in the same cluster are very similar in terms of distance metrics (i.e. Euclidean distance), and compounds in different clusters are very distinct. The so generated train, val and test sets are independent molecular sub-sets, and they would accomplish, with the model''s Applicability Domain (see above), that the experimental property range and chemical structures are similar in such sets, in line with the similar structure–property relationships principle of BSM.

The linear regression models are theoretically validated through the Leave-One-Out Cross-Validation (loo) procedure [49], and also through the more rigorous Leave-30%-Out Cross-Validation (l30%o), with 200,000 cases. According to Golbraikh and Tropsha [49], the Cross-Validation explained variances (r^2_{loo} and $r^2_{l30\%o}$) should be greater than 0.5, although this is a necessary but not sufficient condition for demonstrating the real predictive power.

The QSPR models are also validated with a new criteria based on the mean absolute error (*MAE*) [45]. The quality of the test set predictions is determined through the *MAE* parameter and its standard deviation σ, both computed from the test set predictions after omitting 5% high residual data points, in order to obviate the influence of rarely occurring high prediction errors that may significantly affect the quality of predictions for the whole external test set. For good test set predictions, it is considered that an error of 10% of the training set range should be acceptable, while an error value more than 20% of the training set range should be a very high error.

Finally, we scramble the experimental property values with Y-Randomization [50] and 10,000 cases, as a way of checking that the model is not a result of chance correlation when RMS_{rand} (RMS for Y-randomization) is greater than RMS_{train}.

Applicability domain

A predictive QSPR model is only able to predict molecules falling within its applicability domain (AD), so that the predicted property is not a result of substantial extrapolation (unreliable prediction) [51,52]. The AD definition is dependent on the model's descriptors and the experimental property.

In this work, we determine the AD through two alternative methodologies. The first one is based on the well-known leverage approach [53], where a test set compound *i* must have a calculated leverage h_i smaller than the warning leverage h^*. The second one is based on a simple standardization approach [52]: a given test set compound *i* having *d* standardized descriptor values s_{ik}, $k = 1, ..., d$ must have a maximum value $s_{ik}^{max} \le 3$. In the case that $s_{ik}^{max} > 3$ and its minimum value $s_{ik}^{min} < 3$, then the s_i^{new} parameter has to be calculated and must fulfil the condition: $s_i^{new} = \langle s_i \rangle + 1.28.\sigma_{s_i} \le 3$, where $\langle s_i \rangle$ is the mean of s_{ik} values for *i* and σ_{s_i} is the standard deviation for such values.

Importance of model descriptors

In order to find out the relative importance of the *j*-th descriptor in the linear QSPR model, the regression coefficients were standardized (b_j^s, see Supplementary Table 2S). The larger the absolute value of b_j^s, the greater is the importance of such a descriptor [54].

Results and discussion

The BSM technique was applied to the ANTARES dataset of 851 heterogeneous compounds, thus generating balanced sub-sets of similar size with $n_{train} = 284$, $n_{val} = 284$ and $n_{test} = 283$ compounds; Supplementary Table 1S denotes the members of each set as validation (^) and test (*). In this way, the model''s calibration compounds in train and val sets constitute 66.75% of the whole dataset.

As the next step, the most representative molecular descriptors are searched in the training set through the RM variable sub-set selection approach. The best MLR models based on 1–7 structural features are listed in Table 1, while a brief description of the descriptor's meanings is supplied in Supplementary Table 3S.

Table 1. The best multidimensional QSPR found for BCF. The selected model is in bold.

d	*Descriptors*	r^2_{train}	RMS_{train}	r^2_{val}	RMS_{val}	r^2_{test}	RMS_{test}
1	*DCW*	0.64	0.83	0.50	0.91	0.54	0.85
2	*PC406; DCW*	0.66	0.80	0.54	0.88	0.57	0.82
3	*GATS3c; Sub295; DCW*	0.69	0.77	0.56	0.86	0.62	0.77
4	*GATS3c; Sub295; AP402; DCW*	0.70	0.76	0.58	0.84	0.62	0.77
5	*AATS5e; GATS3c; Sub295; K1406; DCW*	0.71	0.74	0.58	0.83	0.65	0.75
6	***ATS8m; AATS5e; Sub295; K1406; AP391; DCW***	**0.73**	**0.71**	**0.60**	**0.81**	**0.67**	**0.71**
7	*ATS8m; AP391; DCW; D708; FPIP9; SD_B_AB_nCi_2_NS0_T_KA_psa-e_MAS; N2_B_AB_nCi_2_NS3_T_KA_a-psa_MAS*	0.75	0.69	0.60	0.82	0.67	0.71

From Table 1, it is appreciated that the RMS_{train} parameter continuously improves with the addition of molecular descriptors to the linear equation, a typical behaviour in variables sub-set selection, but RMS_{val} does not significantly improve beyond the number of six descriptors. In order to keep the model’’s size as small as possible, we select such model as the best linear regression QSPR:

$$\begin{aligned}\log \text{BCF} = \; & -3.06.10^{-5}\, ATS8m + 0.071\, AATS5e - 0.70\, Sub295 - 0.87\, K1406 \\ & +0.48\, AP391 + +0.069\, DCW + 0.51\end{aligned} \tag{1}$$

$$\begin{aligned} n_{\text{train}} &= 284, r^2_{\text{train}} = 0.73, RMS_{\text{train}} = 0.71,\ r^2_{ij\,\text{max}} = 0.13, o3 = 0, r^2_{\text{rand}} \\ &= 0.10, RMS_{\text{rand}} = 1.31,\ r^2_{\text{loo}} = 0.72, RMS_{\text{loo}} = 0.73, r^2_{\text{l30\%o}} \\ &= 0.67, RMS_{\text{l30\%o}} = 0.79,\ n_{\text{val}} = 284, r^2_{\text{val}} = 0.60, RMS_{\text{val}} \\ &= 0.81,\ n_{\text{test}} = 283, r^2_{\text{test}} = 0.67, RMS_{\text{test}} = 0.71. \end{aligned}$$

From these results $r_{ij\,\text{max}}$ is the maximum correlation coefficient between descriptor pairs, indicating the absence of serious correlation between the six selected descriptors. The $o3$ parameter indicates the number of outlier compounds in the training set having a residual (difference between experimental and predicted property) greater than 3-times RMS_{train}. Equation (1) does not involve training set compounds with very high residuals.

The plot of the predictions as a function of the experimental values is provided in Figure 1. The dispersion plot of residuals in Figure 2 tends to obey a random pattern around the zero line, suggesting that equation (1) predicts the whole dataset without systematic errors or residual bias.

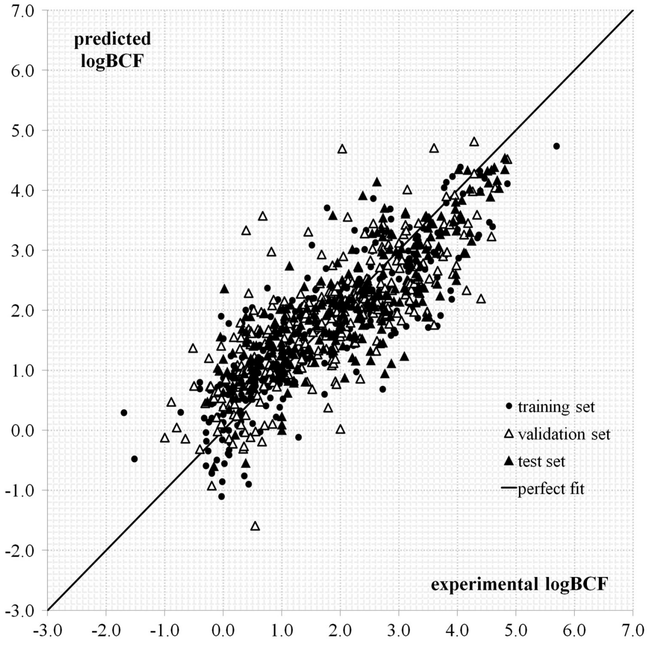

Figure 1. Predicted and experimental log BCF values according to the QSPR of equation (1).

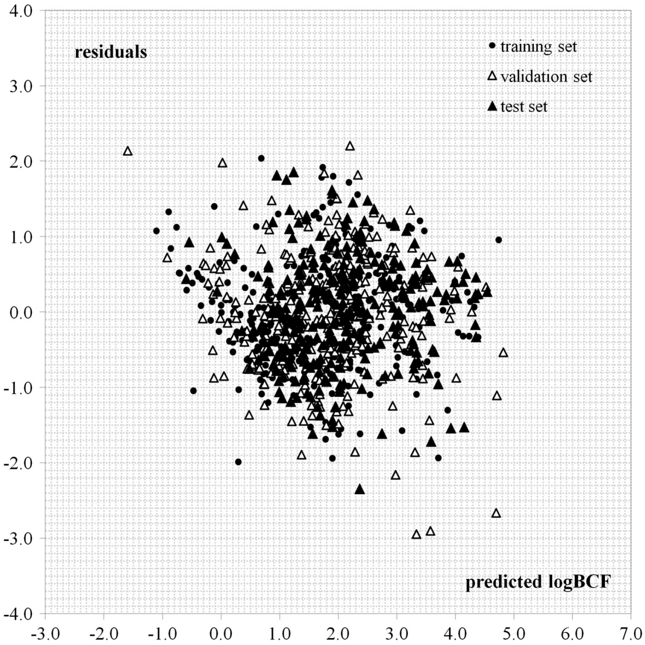

Figure 2. Dispersion plot of residuals for equation (1).

The QSPR of equation (1) has an acceptable predictive power on the external test set of 283 BCF values according to r^2_{test} and RMS_{test} parameters. Such model approves the internal validation process of Leave-One-Out and Leave-30%-Out Cross Validation, through the prediction of one or more molecules excluded at a time from the training set. The Y-randomization technique demonstrates that the model has $RMS_{\text{train}} < RMS_{\text{rand}}$ and $r^2_{\text{rand}} < r^2_{\text{train}}$, and that a valid structure-log BCF relationship is established without chance correlation. Also, the recommended external validation criteria [49] to assure predictive capability are also achieved: $1 - r^2_0/r^2_{\text{test}}(6.0410^{-4}) < 0.1$, or $1 - r'^2_0/r^2_{\text{test}}(0.21) < 0.1$; $0.85 \leq k(1.0035) \leq 1.15$, or $0.85 \leq k(0.9095) \leq 1.15$; $r^2_m(0.66) > 0.5$.

The prediction performance of our QSPR model on the 283 test set compounds is found to be 'intermediate' by the *MAE*-based criteria, which means an acceptable model [45]. For the complete test set, *MAE*(100%) = 0.57 and σ(100\) = 0.43, while omitting 5% of the high residuals compounds leads to *MAE*(95%) = 0.51 and σ(95%) = 0.34.

The six conformation-independent molecular descriptors appearing in the proposed quantitative structure–log BCF relationship are readily calculated from the molecular structure, and such variables belong to different classes [11–13]:

- two Autocorrelation of the Topological Structure descriptors: *ATS8m*, the Broto-Moreau autocorrelation – lag 8/weighted by mass, and *AATS5e*, the average Broto-Moreau autocorrelation – lag 5/weighted by Sanderson electronegativities. The structural variables introduced by Broto-Moreau are bidimensional autocorrelations between atom pairs (*i*, *j*) in a molecule, with the main purpose of capturing the degree of interaction between them. The nature of atoms is considered through a given property as atomic weight (*w*), i.e. atomic mass, polarizability, electronegativity, or volume. These indices are calculated from the graph by summing products of terms $w_i \cdot w_j$ including terminal atomic contributions in all the paths of a prescribed length (lag).

- a CORAL descriptor: *DCW*, optimal descriptor based on HSG EC2 and SMILES s attributes. In the graph approach, EC2 is the Morgan's extended connectivity index of second order. It should be noted that the index of zero-th order EC0 for vertex (atom) *j* represents the vertex degree for *j* (number of neighbour atoms to *j*), while the higher order indices ECk are obtained through a recursive formula based on EC0 [31,32]. In the SMILES approach, *s* represents a one-element attribute: i.e. if a SMILES is a sequence of elements such as 'ABCDE', then the *s* structural attribute can be represented with 'A', 'B', 'C', 'D', 'E'.

and the next descriptors have a straightforward structural interpretation:

- a 2D Atom Pairs Fingerprint descriptor: *AP391*, the presence of C-C at topological distance 6;
- a Klekota Roth Fingerprint descriptor: *K1406*, indicating the presence of the SMARTS pattern [!#1]C(=O)[OH]; and
- a Substructure Fingerprint: *Sub*295, the presence of a C_ONS bond.

All the molecular descriptors of equation (1) have positive numerical values with the exception to *DCW*, which can have either positive or negative values. The sign of the regression coefficient in the linear model indicates when the descriptor contribution increases or decreases the predicted log BCF values. Higher positive numerical values of *DCW*, *AATS*5*e* and *AP*391 and lower values for *ATS*8*m*, *Sub*295 and *K*1406 tend to predict higher log BCF values. After standardization, the most important descriptor from equation (1) is *DCW* ($b_j^s = 0.66$), thus having numerical values changing most in accordance with the numerical variations of the experimental property. The remaining descriptors *ATS8m* ($b_j^s = 0.14$), *AATS*5*e* ($b_j^s = 0.12$), *Sub*295 ($b_j^s = 0.21$) *K*1406 ($b_j^s = 0.16$), and *AP*391 ($b_j^s = 0.17$) complement each other inside the linear equation and have a comparable relevance.

The model's squared correlation matrix is provided in Supplementary Table 4S, showing the absence of high correlations between descriptors pairs, as mentioned before. We also calculate the variance inflation factor (*VIF*), a parameter that measures the multicollinearity among descriptors. A *VIF* of 1 for a specific descriptor means that there is no correlation between this descriptor and all the remaining descriptors of the model, and a *VIF* exceeding 10 indicates that multicollinearity is a problem in the dataset [55]. From Supplementary Table 4S, it is demonstrated that the *VIF* parameter for each descriptor of equation (1) is near to 1. The numerical descriptor values are given in Supplementary Table 5S.

Now we demonstrate that the proposed QSPR of equation (1) is generalizable and useful for application, that is to say, our model is not determined only by the training set composition due to the specific dataset partitioning of BSM. For this, we perform 1000 different random splitting operations and recalculate the statistics of the model proposed by us in the present work. We find that, for 1000 random external test sets, equation (1) leads to r^2_{test} ranging from 0.56–0.76 and RMS_{test} ranging from 0.66–0.85. These findings suggest that the final model of equation (1) has an acceptable stability of its predictive ability. The good predictivity of our QSPR model on the test set does not result by chance, and the molecular descriptors involved in equation (1) work satisfactorily on the different training–test sets partitions. The 1000 random training and test sets are provided in matrix form, the randtrain and randtest spreadsheets from the Random splittings xls-file of the Supplementary Material.

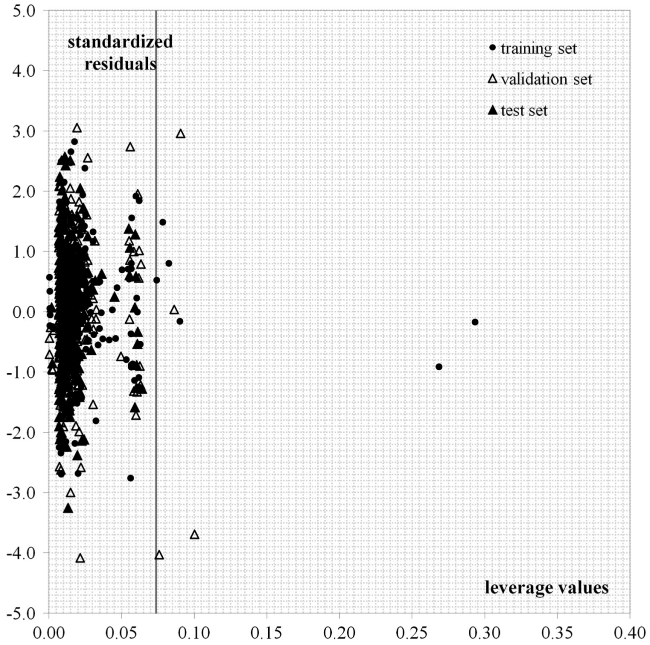

Figure 3. Williams plot for equation (1). The line indicates the warning leverage of 0.0739.

Following with the exploration of the applicability domain of the developed QSPR model, a compound with high leverage would reinforce the model if the compound is in the training set (good leverage), but such a compound in the test set could have unreliable predicted data, the result of substantial extrapolation of the model (bad leverage) [51]. In our case, it was found that the 283 test set compounds belonged to the AD, as their h_i values fall under the h^* limit (0.0739). The Williams plot for equation (1) (standardized residuals as function of the h_i values) is provided in Figure 3. Some compounds belonging to the training and validation sets have high leverages reinforcing the model, such as chemicals 41, 59, 265, 403, 427, 468, 504, 505, 522, and 659. This result obtained with the leverage approach for the test set coincides with the one obtained by using the standardization approach, as the two conditions $s_{ik}^{max} \leq 3$ or $s^{new}{}_i \leq 3$ are followed by all the 283 test set compounds. Thus, the predicted log BCF values for the test set compounds can be considered as reliable. Some compounds have standardized residuals higher than three units: this may be purely attributed to the high structurally heterogeneous dataset of 851 compounds, which cannot be expected to be modelled by using only a 6-descriptors model (equation 1).

A comparison can be done between the performance of our proposed alternative BCF QSPR model of equation (1) and the one reported by Gissi et al. [21]. By means of 836 compounds in a 608:152:76 splitting (n_{train}:n_{val}:n_{test}), the statistical quality achieved by the reported 9-descriptors ANN model appears summarized in Table 2 (model-1). We consider that our model improves such reported result due to the following four main reasons:

(i) Number of molecules treated: we contemplate all the 851 molecules in the QSPR study without excluding anyone, contrary to the reported QSPR, which employs 836 compounds and excludes 15 compounds due to limitations in the descriptor calculation software.

(ii) Model size: equation (1) involves six descriptors instead of nine.

(iii) Suitability of the dataset partitioning: we use a 284:284:283 splitting, while the reported one uses 608:152:76. Thus, more test set compounds are considered during the present QSPR study for determining the predictive capability than in model-1.

(iv) Simplicity: our linear model is simpler than the reported non-linear ANN model, and is not dependent on the molecular conformations of the heterogeneous compounds.

By means of defining the applicability domain of the reported model-1 through four independent filtering methods [21], 27 compounds are further excluded from val and test (42 compounds excluded in total from the initial dataset). Although a better statistical result is achieved by model-2 when compared to model-1 (Table 2), such a model considers only 8.53 % of the compounds in the test set, instead of the 33.25% considered by our model of equation (1). Indeed, our proposed model leads to a better result on the 283 test set compounds with $RMS_{\text{test}} = 0.71$, compared to $RMS_{\text{test}} = 0.82$ for model-2 on 69 compounds.

Table 2. Comparison of the statistical performance of different BCF QSPR models on the ANTARES dataset.

Model	*n*	*Splitting detail*	r^2_{train}	RMS_{train}	r^2_{val}	RMS_{val}	r^2_{test}	RMS_{test}
Present work (equation (1))	851	284:284:283	0.73	0.71	0.60	0.81	0.67	0.71
9-descriptors ANN model-1 [21]	836	608:152:76	0.73	0.67	0.63	0.79	0.62	0.84
9-descriptors ANN model-2 [21]	809	608:132:69	0.73	0.67	0.77	0.62	0.66	0.82
EPI Suite BCFBAF module	851	284:284:283	0.70	0.77	0.64	0.77	0.69	0.70

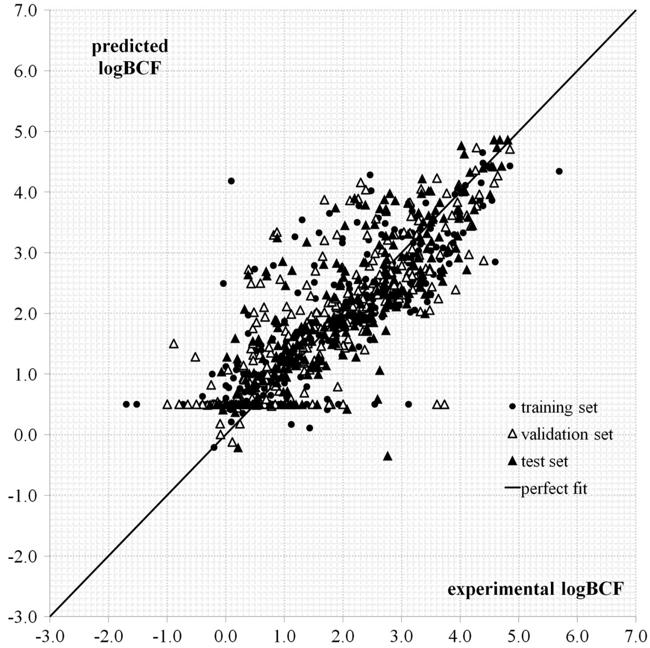

Figure 4. Predicted and experimental log BCF values according to EPI suite.

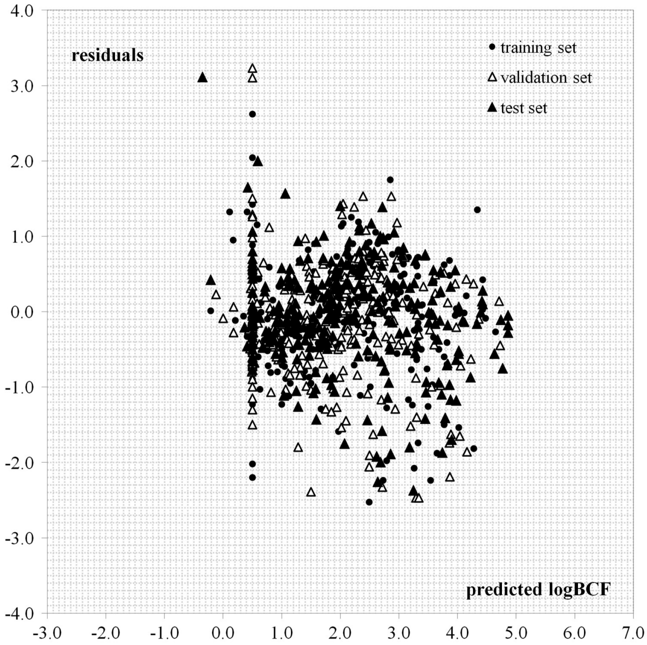

Figure 5. Dispersion plot of residuals for EPI suite log BCF.

Finally, we compare the predicted log BCF values obtained by equation (1) with the predictions calculated by using the BCFBAF module from the EPI Suite freeware and the BSM splitting. From Table 2, similar statistics were found for the training, validation, and test sets, although achieved by different methodologies in both cases. However, when plotting the predictions as a function of the experimental values for the EPI Suite results in Figure 4, together with the dispersion plot of residuals in Figure 5, it is observed that many compounds are predicted with the same value: 131 compounds have predicted log BCF = 0.50. In this sense, we consider equation (1) behaves as a better QSPR model.

Conclusions

We propose an alternative QSPR model for the bioconcentration capability of chemical compounds, for which a large number of non-conformational molecular descriptors was simultaneously analysed in order to find the best predictive capability of the relationship. The ANTARES dataset includes highly heterogeneous molecular structures together with 159 pesticides, so that the applicability domain of our best QSPR model considers in its definition different chemical classes for the BCF prediction, and, therefore, could be applied to the prediction of heterogeneous pesticides of different types.

The novelty of the present work relies on the analysis of a great pool of molecular descriptors (27,017 descriptors), in order to select the best ones in the final linear regression model. In this way, we focus our work on better describing the chemical structure, and complement different descriptor software types for improving the statistical quality of the established QSAR.

The consideration of the constitutional and topological aspects of the molecular structures in the conformation-independent QSPR approach achieves once more acceptable results, and new investigations on other physicochemical and biological properties of interest will be published soon elsewhere.

Acknowledgements

We thank the financial support provided by the National Research Council of Argentina (CONICET) PIP11220130100311 project, and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities. D.E.B., M.A.O., E.A.C. and P.R.D. are members of the scientific researcher career of CONICET.

Disclosure statement

No potential conflict of interest is reported by the authors.

Funding

This work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas [grant number PIP11220130100311 project].

References

[1] G. Matthews, *Pesticides: Health, Safety and the Environment*, 2nd ed., Wiley-Blackwell, New York, 2015.

[2] L.H. Nowell, P.D. Capel, and P.D. Dileanis, *Pesticides in Stream Sediment and Aquatic Biota: Distribution, Trends, and Governing Factors*, CRC Press, Boca Ratón, FL, 1999.

[3] G. Matthews, R. Bateman, and P. Miller, *Pesticide Application Methods*, 4th ed., Wiley-Blackwell, New York, 2014.

[4] B. Beek, *Bioaccumulation New Aspects and Developments*, The Handbook of Environmental Chemistry, Springer, Berlin, 2013.

[5] F. Grisoni, V. Consonni, S. Villa, M. Vighi, and T. Todeschini, *QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions*?, Chemosphere 127 (2015) 171–179.

[6] C. Hansch and A. Leo, *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC, 1995.

[7] E. Benfenati, *Theory, guidance and applications on QSAR and REACH*, Orchestra. Available at https://ebook.insilico.eu/insilico-ebook-orchestra-benfenati-ed1_rev-June2013.pdf 2013.

[8] K. Roy, S. Kar, and R. Narayan Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, New York, 2015.

[9] E. Benfenati, *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, Elsevier Science, Amsterdam, 2007.

[10] J.P. Knaak, C. Timchalk, and R. Tornero-Velez, *Parameters for pesticide QSAR and PBPK/PD models for human risk assessment*, Oxford University Press, Oxford, 2013.

[11] A.R. Katritzky and E.V. Goordeva, *Traditional topological indices vs. electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research*, J. Chem. Inf. Comput. Sci. 33 (1993), pp. 835–857.

[12] M.V.E. Diudea, *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, New York, 2001.

[13] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Methods and Principles in Medicinal Chemistry, Wiley-VCH, Weinheim, 2009.

[14] M. Pavan, A.P. Worth, and T.I. Netzeva, *Review of QSAR Models for Bioconcentration*, JRC report EUR EN I-21020 (2006).

[15] A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, and E. Benfenati, *Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish*, Chem. Cent. J. 4 Suppl. 1 (2010), pp. S1.

[16] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, and E. Benfenati, *A new hybrid system of QSAR models for predicting bioconcentration factors (BCF)*, Chemosphere 73 (2008), pp. 1701–1707.

[17] VEGA platform. Available at https://www.vega-qsar.eu/, 2017.

[18] W. Meylan, P. Howard, R. Boethling, D. Aronson, H. Printup, and S. Gouchie, *Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient*, Environ. Toxicol. Chem. 18 (1999) 664–672.

[19] Epi Suite, U.S.EPA. Available at https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface, 2017.

[20] T.E.S.T. model. Available at https://www.epa.gov/nrmrl/std/qsar/qsar.html, 2017.

[21] A. Gissi, D. Gadaleta, M. Floris, S. Olla, A. Carotti, E. Novelllino, E. Benfenati, and O. Nicolotti, *An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes*, ALTEX 31 (2014), pp. 23–26.

[22] A. Gissi, O. Nicolotti, A. Carotti, D. Gadaleta, A. Lombardo, and E. Benfenati, *Integration of QSAR models for bioconcentration suitable for REACH*, Sci. Total Environ. 456–457 (2013), pp. 325–332.

[23] P.R. Duchowicz, N.C. Comelli, E.V. Ortiz, and E.A. Castro, *QSAR study for carcinogenicity in a large set of organic compounds*, Curr. Drug Saf. 7 (2012), pp. 282–288.

[24] P.R. Duchowicz, D.O. Bennardi, D.E. Baselo, E.L. Bonifazi, C. Rios-Luci, J.M. Padrón, G. Burton, and R.I. Misico, *QSAR on antiproliferative naphthoquinones based on a conformation-independent approach*, Eur. J. Med. Chem. 77 (2014), pp. 176–184.

[25] P.R. Duchowicz, S.E. Fioressi, D.E. Bacelo, L.M. Saavedra, A.P. Toropova, and A.A. Toropov, *QSPR studies on refractive indices of structurally heterogeneous polymers*, Chemom. Intel. Lab. Syst. 140 (2015), pp. 86–91.

[26] J.F. Aranda, J.C. Garro Martinez, E.A. Castro, and P.R. Duchowicz, *Conformation-independent QSPR approach for the soil sorption coefficient of heterogeneous compounds*, Int. J. Mol. Sci. 17 (2016), pp 1247.

[27] ACD/ChemSketch. Available at www.acdlabs.com, 2017.

[28] Open Babel for Windows. Available at https://openbabel.org/wiki/Category:Installation, 2017.

[29] Pharmaceutical Data Exploration Laboratory (PaDEL). Available at https://www.yapcwsoft.com/, 2017.

[30] C.W. Yap, *PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints*, J. Comput. Chem. 32 (2011), pp. 1466–1474.

[31] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, T. Puzyn, D. Leszczynska, and J. Leszczynski, *Novel application of the coral software to model cytotoxicity of metal oxide nanoparticles to bacteria* Escherichia coli, Chemosphere 89 (2012), pp. 1098–1102.

[32] A.P. Toropova, A.A. Toropov, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, *Coral: QSAR modeling of toxicity of organic chemicals towards* Daphnia magna, Chemom. Intel. Lab. Syst. 110 (2012), pp. 177–181.

[33] S.E. Fioressi, D.E. Bacelo, W.P. Cui, L.M. Saavedra, and P.R. Duchowicz, *QSPR study on refractive indices of solvents commonly used in polymer chemistry using flexible molecular descriptors*, SAR QSAR Environ. Res. 26 (2015), pp. 499–506.

[34] Coral 1.5. Available at https://www.insilico.eu/coral, 2017.

[35] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, *Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics*, J. Chem. Inf. Model. 48 (2008), pp. 1337–1344.

[36] Recon. Available at https://reccr.chem.rpi.edu/Software/RECON/recondoc/ReconManual.html, 2017.

[37] B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt, *Electronic van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 1890–1905.

[38] R.F.W. Bader, *Atoms in Molecules-A Quantum Theory*, Oxford University Press, Oxford, UK, 1990.

[39] C.M. Breneman and M. Rhem, *A QSPR analysis of HPLC column capacity factors for a set of high-energy materials using electronic Van der Waals surface property descriptors computed by the transferable atom equivalent method*, J. Comput. Chem. 18 (1997), pp. 182–197.

[40] J.R. Valdes-Martini, C.R. García Jacas, Y. Marrero-Ponce, Y. Silveira Vaz 'd Almeida, and C. Morrel, *QuBiLS-MAS: Free Software for molecular descriptors calculator from Quadratic, Bilinear and Linear Maps based on Graph–Theoretic Electronic-Density Matrices and Atomic weighting*, Version 1.0. CAMD-BIR Unit, CENDA Number of register: 2373-2012, 2012.

[41] P.R. Duchowicz, E.A. Castro, and F.M. Fernández, *Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR Studies*, MATCH Commun. Math. Comput. Chem. 55 (2006), pp. 179–192.

[42] A.H. Morales, P.R. Duchowicz, M.A. Cabrera Pérez, E.A. Castro, M.N.D.S. Cordeiro, and M.P. González, *Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential*, Chemom. Intel. Lab. Syst. 81 (2006), pp. 180–187.

[43] Matlab 7.0, The MathWorks, Inc. Available at https://www.mathworks.com, 2017.

[44] P. Gramatica and A. Sangion, *A historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics and terminology*, J. Chem. Inf. Model. 56 (2016), pp. 1127–1131.

[45] K. Roy, R.N. Das, P. Ambure, and R.B. Aher, *Be aware of error measures. Further studies on validation of predictive QSAR models*, Chemom. Intel. Lab. Syst. 152 (2016), pp. 18–33.

[46] C. Rojas, P.R. Duchowicz, P. Tripaldi, and R. Pis Diez, *Quantitative structure-property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase*, J. Chromatogr. A 1422 (2015), pp. 277–288.

[47] P.R. Duchowicz, S.E. Fioressi, E.A. Castro, K. Wróbel, N.E. Ibezim, and D.E. Bacelo, *Conformation-independent QSAR study on human epidermal growth factor receptor-2 (HER2) inhibitors*, Chem. Select. 2 (2017), pp. 3725–3731.

[48] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 2005.

[49] A. Golbraikh and A. Tropsha, *Beware of q2!*, J. Mol. Graphics Modell. 20 (2002), pp. 269–276.

[50] C. Rücker, G. Rücker, and M. Meringer, *Y-Randomization and its variants in QSPR/QSAR*, J. Chem. Inf. Model. 47 (2007), pp. 2345–2357.

[51] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[52] K. Roy, S. Kar, and P. Ambure, *On a simple approach for determining applicability domain of QSAR models*, Chemom. Intel. Lab. Syst. 145 (2015), pp. 22–29.

[53] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.

[54] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, 1981.

[55] K. Roy and P.P. Roy, *Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FAMLR, PLS, GFA, G/PLS and ANN techniques*, Eur. J. Med. Chem. 44 (2009), pp. 2913–2922.