Ecotoxicology and Environmental Safety 171 (2019) 47–53

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Conformation-independent quantitative structure-property relationships study on water solubility of pesticides

Silvina E. Fioressi[a,*], Daniel E. Bacelo[a], Cristian Rojas[b], José F. Aranda[c], Pablo R. Duchowicz[c,*]

[a] Departamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Villanueva 1324, CP 1426, Buenos Aires, Argentina
[b] Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca, Ecuador
[c] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina

ARTICLE INFO

Keywords:
Quantitative structure-property relationships
Pesticides
Water solubility
Molecular descriptors
CORAL software

ABSTRACT

Water solubility is a key physicochemical parameter in pesticide control and regulation, although sometimes its experimental determination is not an easy task. In this study, we present Quantitative Structure-Property Relationships (QSPRs) for predicting the water solubility at 20 °C of 1211 approved heterogeneous pesticide compounds, collected from the online Pesticides Properties Data Base (PPDB). Validated and generally applicable Multivariable Linear Regression (MLR) models were established, including molecular descriptors carrying constitutional and topological aspects of the analyzed compounds. The most representative descriptors were selected from the exploration of a large number of about 18,000 structural variables. A hybrid approach that involves a molecular descriptor, a fingerprint, and a flexible descriptor showed the best predictive performance.

1. Introduction

There are an increasing number and amount of pesticides detected in water including drinking water sources. At the present time, important regulatory laws exist regarding the levels of pesticides that are allowed in surface water, groundwater and drinking water to avoid dangerous contamination (Agency, 2017; Hamilton et al., 2003). European regulations recognize the need for the use of pesticides for agricultural development, but their use cannot have adverse effects on human health, animals or the environment (Villaverde et al., 2017). Pesticides and their degradation products are distinguished by their strong toxicity and persistence in the environment. The predisposition of a pesticide to be removed from soil by runoff from rain or from irrigation water and to reach surface water is directly related to its water solubility (S_w). This parameter is defined as the concentration of a chemical dissolved in water when that water is both in contact and at equilibrium with the pure chemical.

The study of water solubility of pesticides is important to measure their environmental fate (e.g. biodegradation, bioaccumulation) and potential effects on humans and other living organisms. Water solubility measurement establishes a basis for other environmentally relevant parameters, such as the octanol/water partition coefficient and the organic carbon/water partition coefficient, among others. The experimental error of solubility measurements can be quite large, especially for compounds with a very low solubility value. The accurate evaluation of water solubility is complicated by a number of factors, including ionization, formation of salts and polymorphism. These effects may significantly alter the water solubility values (Cronin and Livingstone, 2004).

The application of Quantitative Structure-Property Relationships (QSPR) and computer-aided modeling techniques are valuable and frequently used tools to accurately predict physical and chemical properties of compounds (Cronin and Livingstone, 2004; Hamadache et al., 2017; Mas et al., 2010). Regulation agencies worldwide promote the use and development of non experimental tests to anticipate the possible health and environmental risks of pesticides (Tebes-Stevens et al., 2018; Villaverde et al., 2018). European regulatory agencies such as REACH (Registration, Evaluation, Authorization and restriction of Chemicals) and BPR (Biocidal Product Regulation) intensely encourage the use of non-animal testing techniques to evaluate the chemical risk of new pesticides. Therefore, QSPR techniques emerge as a logical and useful alternative to expensive and time-consuming experimental procedures for the prediction of water solubility of pesticides, ultimately avoiding many animal laboratory sacrifices. Trustworthy models can provide insights about the molecular characteristics that may influence the water solubility, and greatly improve the determination of this property. Many methods have been developed to predict water solubility, either exclusively from the molecular structure or by using variables that are easier to measure. A well known method is the Yalkowsky's "General Solubility Equation" (Ran et al., 2002), which

* Corresponding authors.
E-mail addresses: sfioressi@yahoo.com (S.E. Fioressi), pabloducho@gmail.com (P.R. Duchowicz).

https://doi.org/10.1016/j.ecoenv.2018.12.056
Received 7 July 2018; Received in revised form 14 December 2018; Accepted 19 December 2018

bases the solubility calculation on only two variables: the partition of liquid compounds and water (*logP*), and the melting point to take into account the transition from solid to liquid. The ESOL method developed by Clarke and Delaney (Clarke and Delaney, 2003; Delaney, 2004) produces solubility predictions comparable to the General Solubility Equation (GES) but with the advantage that it does not need to account for the experimental melting points. The ESOL method yields linear models based on the following four parameters: $log\,P_{oct}$, molecular weight, number of rotatable bonds and proportion of heavy atoms defined as 'aromatic'. Recently, predictive models have been developed for the aqueous solubility of a large set of drugs, drug-like compounds, and agrochemicals, with two dimensional descriptors called extended topochemical atom (ETA) indices, as well as other topological, structural, spatial and electronic non-ETA descriptors, and the lipophilicity parameter, $C\log P$ (Das and Roy, 2013). These models employed the genetic function approximation (GFA), genetic partial least squares (G/PLS), and stepwise multiple linear regression (MLR). On the other hand, conformation-independent descriptors were used by the Toropov group through the CORAL program to build up QSPRs for water-solubility (Toropov et al., 2013). Also, density functional theory (DFT) approximations and QSPR methods were applied to halogenated methylphenyl ethers to model their water solubility (Zeng et al., 2012). They found that solubility was strongly affected by three variables: energy of the lowest unoccupied molecular orbital, most positive atomic partial charge in the molecule, and the quadrupole moment. The water solubility of 209 congeners of chloro-trans-azobenzene was modeled using Genetic Algorithm-Artificial Neural Network (GA-ANN) (Wilczyńska-Piliszek et al., 2012). Non-ionic perfluorinated chemicals were studied using two-dimensional descriptors (Bhhatarai and Gramatica, 2010). In addition, Benfenati and coworkers have applied different predictive computer models to analyze water solubility in organic compounds (Cappelli et al., 2013). Analyzing all available models, they concluded that the values of highly soluble compounds can be more accurately predicted than those of poorly soluble ones. Recently, Kim et al. (2016) proposed a QSPR model based on the hyper-Wiener index (*WW*) of quantum-chemical descriptor for 75 polychlorinated dibenzo-p-dioxins (PCCDs). The single descriptor model obtained with the *WW* successfully predicted the water solubility of PCDDs and was able to distinguish among congeners with the same number of chlorine atoms. The authors concluded that for these pesticides, the structural information contained in the *WW* was fundamental to achieving good predictions.

The purpose of this study is to establish a QSPR model for the water solubility of pesticides using a large set of descriptors and experimental data from structurally heterogeneous pesticides reported in the literature. It is our aim to propose simple models based on an extensive and varied set of compounds. Only conformation-independent molecular descriptors were considered in order to obtain reliable but simple models.

It is well known that QSPR models solely based on constitutional and topological molecular characteristics, avoid ambiguities that may result from the existence of chemical compounds in various conformational states (Duchowicz et al., 2012; Talevi et al., 2012). Therefore, three different QSPR approaches were explored: i) conventional 0D, 1D and 2D descriptors and fingerprints generated by the freely available descriptor programs PaDEL-Descriptor (version 2.20) (Yap, 2011), EPI Suite (US, E.P.A. *Estimation Programs Interface Suite™ for Microsoft® Windows*, 411; Washington, DC), and Mold2 (Hong et al., 2008); ii) flexible descriptors obtained through the CORALSEA program (Toropova et al., 2012); and iii) the aforementioned sets of descriptors combined. Simple models including from 1 to 8 descriptors were chosen as the best predictive combinations of independently selected variables. The study complies with the principles required by the Organization for Economic Co-operation and Development (OECD), which includes the following: a defined endpoint with S values determined with equal experimental conditions; unambiguous algorithms, with reproducibility of the predictions covered by the generation of the descriptors using publicly available software; a defined applicability domain; and appropriate measures of goodness-of–fit, robustness and predictivity determined by external validation. (Gramatica, 2007; OECD, 2007)

2. Materials and methods

2.1. Experimental Dataset

The QSPR analysis was performed on 1211 approved pesticides (Table 1S). Their structures and water solubility measured at 20 °C were collected from the online Pesticide Properties DataBase (PPDB) (Lewis et al., 2018). The PPDB has been developed by the Agriculture & Environment Research Unit at the University of Hertfordshire. The solubility expressed as g/L was converted into logarithmic units ($\log S_w$).

2.2. Structural representation and molecular descriptors calculation

The molecular structures of the pesticides were generated in both SMILES notation and bi-dimensional structures, drawn with the free Discovery Studio software (Version 3.5, *Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment*, San Diego, USA) and saved in MDL mol (V2000) format without performing any geometrical optimization. Two different approaches were applied to calculate the descriptors:

a) The freely-available software PaDEL-Descriptor (version 2.20) (Yap, 2011), EPI Suite, and Mold2 (Hong et al., 2008) were used to compute 17,974 theoretical conformation-independent molecular descriptors and fingerprints. Of this total, 1444 1D and 2D descriptors and 12 types of fingerprints (16,092) were calculated with the PaDEL-Descriptor, 184 descriptors with the EPI Suite, and 254 descriptors with the Mold2. Descriptors found to be linearly-dependent and constant values were excluded from the pool of variables.
b) Flexible molecular descriptors were obtained from the CORAL freeware (Toropova et al., 2012) using the SMILES notation of the compounds as input along with the experimental $\log S_w$ values. The CORAL program allows different structural representation (SR) approaches: a chemical graph (hydrogen-suppressed graph (HSG), hydrogen-filled graph (HFG) or graph of atomic orbitals (GAO)), SMILES, or a hybrid of both (chemical graph and SMILES). The selected SR defines the local descriptors to be included in the QSPR analysis; therefore, it is crucial to look for the most appropriate combination of structural attributes (local descriptors, SA).

The CORAL framework searches for a QSPR model that correlates the experimental $\log S_w$ and a properly defined flexible descriptor (*DCW*) through a one-variable linear relationship. The *DCW* descriptor is a linear combination of special coefficients called correlation weights (CW) with values calculated for each SA type in the training set via a Monte Carlo (MC) simulation (Table 2S). The *DCW* depends upon the threshold value (T) and the number of epochs or iterations used (Toropova et al., 2012). T defines rare SMILES attributes that do not contribute to the predicted property. All SMILES attributes that take place in less than T SMILES notations of the training set were classified as rare instead of active. In this study, T ranges from 0 to 5 and the maximum number of iterations used is 50.

The programs used to calculate the descriptors were selected based on their calculation accuracy, ease of access, free availability and recognition by the scientific community. Following the OECD principles, for a QSPR model to be acceptable it must be easily and continuously applicable. These programs allowed calculations for the prediction of the endpoint to be reproduced by everyone and also applied to new compounds. The programs have already been used successfully by our group for other QSPR and QSAR studies (Duchowicz et al., 2015, 2017).

2.3. Model Validation

For building the QSPR models and verifying their predictive capability, the complete dataset was split into three subsets: a training set (404 compounds) for model development, a validation set (404 compounds) for checking whether the model is satisfactory for compounds that are absent from the training set, and a test set (403 compounds) for true external validation. To be certain that the training set is representative of the validation and test sets, the dataset was split using the Balanced Subsets Method (BSM) (Rojas et al., 2015). The procedure is based on *k*-Means Cluster Analysis (*k*-MCA) which guarantees similar structure-property relationships in the three subsets. The Replacement Method (RM) (Duchowicz et al., 2006) variable subset selection programmed in MATLAB software (The MathWorks, Inc., Natick, Massachusetts, USA) was applied to generate Multivariable Linear Regression (MLR) models on the training set. RM is a sequential method that optimizes the root-mean-squared deviation (*RMSD*) in MRL.

In order to measure the stability of the QSPR model upon inclusion/exclusion of molecules, the MLR models were internally validated through the Leave-One-Out Cross Validation (*loo*) method. It is a general validation criterion to accept the model if the coefficient of determination *loo* (R^2_{loo}) is greater than 0.5. However, this is a necessary but not sufficient condition for predictive power (Golbraikh and Tropsha, 2002). A more robust validation criterion is to apply the same principle ($R^2_{test} > 0.5$) to the external test set of 403 compounds. To rule out chance correlations, the experimental values were scrambled through the Y-Randomization method (Wold et al., 1995) in such a way that they did not correspond to the respective compounds.

2.4. Applicability domain

The applicability domain (AD) of a QSPR model is the range of data in which the training set model is developed and within which predictions for new molecules can be considered reliable. It is a theoretically defined space in which only the molecules that belong to this AD are not considered model extrapolations (Gadaleta et al., 2016; Gramatica, 2007). The AD for the proposed models were determined though the leverage approach (Eriksson et al., 2003), where each compound *i* has a calculated leverage value h_i and a warning leverage value h^* (Table 2S); if $h_i > h^*$ the prediction is considered as a model extrapolation.

3. Results and discussion

We performed a QSPR analysis on 1211 diverse compounds well known for their pesticide action. Three different QSPR approaches were explored to model the water solubility by resorting to different descriptor types: 1) conventional descriptors; 2) flexible descriptors; and 3) hybrid descriptors. The general methodology applied in the three approaches was first, to verify the predictive ability of the molecular descriptors, and then to evaluate the models for the experimental log S_w data in the test set. This allows to fully exploit the available structural and response information, and thus to enlarge the applicability domain of the designed model. The statistical parameters for each model are provided as Supplementary information (Table 3S to 11S). To fulfill the five validation principles suggested by the OECD the applicability domains were properly defined and the models were validated through Y-randomization and Cross-Validation. The detailed results, including the respective model equations and the main statistical parameters obtained are discussed in the following sections.

3.1. Conventional descriptors

The results for the best eight models found by using the first approach are shown in Table 1. Models involving from one to eight molecular descriptors and fingerprints were explored; the best predictive performances were observed for those containing six and seven descriptors. Both of these models presented similar $RMSD_{val}$, but the model with seven descriptors had a smaller difference between $RMSD_{val}$ and $RMSD_{train}$. Therefore, the seven descriptors model was selected as the best result for the conventional descriptors approach, and the calculated log S_w (Eq. (1)) versus the experimental values for this model are shown in Fig. 1.

$$\log S_w = 0.156 + 1.462GATS2m + 1.808GATS1p - 0.354CrippenLogP + 4.246SIC3 + 9.9 \times 10^{-11}\text{SpDiam\_D} - 1.336\ \text{VAdjMat} + 2.253\ \text{MACCSFP35} \quad (1)$$

$N_{train} = 404$, $R^2_{test} = 0.56$, $RMSDS_{train} = 1.57$, $N_{value} = 404$, $R^2_{val} = 0.54$, $RMSD_{val} = 1.49$, $_o(3S) = 7$

$N_{test} = 403$, $R^2_{test} = 0.56$, $RMSDS_{test} = 1.38$, $R^2_{loo} = 0.53$, $RMSD_{loo} = 1.62$, $RMSD_{rand} = 2.30$, $h^* = 0.059$ A compound having an absolute residual value (difference between experimental and calculated $\log_{10} S_w$) greater than 3 times $RMSD_{train}$ is considered an outlier. The $o(3\,S)$ parameter indicates the number of outlier compounds in the training set (Verma and Hansch, 2005). Ten compounds in this model are not within the applicability domain and seven compounds are outliers (compounds **53**, **213**, **233**, **637**, **853**, **1120**, **1155**). The water solubility data for compound **53** is not reliable, as there are at least two very different solubility values reported in the literature for this pesticide (Lewis et al., 2018; Wishart et al., 2017). The rest of the outliers are compounds with extremely low or very high solubility. For example, compound **853** exhibits the lowest solubility in the dataset, whereas compound **637** has the highest one. The solubilities of outlier compounds **213** and **1155** are significantly low and compounds **233** and **1120** are very soluble in water ($S_w \approx 10^3$ g/L). It is understandable that, in such a large and diverse molecular set, those compounds with extreme solubility values fall as outliers in the proposed model. Nonetheless, Eq. (1) satisfies the following external validation conditions (Roy, 2007):

$$1 - R^2_0/R^2_{test} < 0.1(0.0005) \text{ or } 1 - R'^2_0/R^2_{test} < 0.1\ (0.11) \text{ and, } 0.85 \le k \le 1.15(0.99)$$

$$\text{and } 0.85 \le k' \le 1.15(0.66) \text{and } R^2_m > 0.5\,(0.55)$$

Two *GATS* descriptors showed positive correlations with log S_w in this model. These are 2D-autocorrelation descriptors originated in the autocorrelation of the topological structure of Geary that encode both, the molecular structure and a physicochemical property as a vector, relating the topology of a structure with the selected physicochemical attribute. The number following the descriptor symbol represents the topological distance between atom pairs (lag), and the letter accounts for the physicochemical property considered in the weighting component for its computation. In the present model, the *GATS2m* descriptor represents an autocorrelation descriptor of lag 2 weighted by mass, whereas *GATS1p* describes the atomic polarizabilities at a topological distance of one.

The *MACCSFP35* fingerprint, which represents the presence of an alkali metal atom of group IA, the 2D matrix-based descriptor *SpDiam_D* (spectral diameter from the topological distance matrix) and the structural information content index *SIC3* (neighborhood symmetry of 3-order), also presented positive correlations with log S_w. In contrast, two descriptors in this model presented a negative effect on the water solubility: *CrippenLogP* and *VAdjMat*. The first one is an atom-based descriptor that measures the lipophobic character of a molecule, and the second one is vertex adjacency information (magnitude): $1 + \log_2 m$, where *m* is the number of heavy-heavy bonds. Therefore, this model predicts that the polarizability, the presence of alkali atoms, and the asymmetry of the molecular structure have positive contributions to the water solubility, whereas the lipophobic character and the presence of heavy elements have a negative contribution, as expected.

Table 1
Descriptors identified for modeling the water solubility in the training, validation, and test sets. The best model appears in bold text.

#Des#Desc.	Descriptors	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	R^2_{test}	$RMSD_{test}$
1	*CrippenLogP*	0.27	2.02	0.49	1.70	0.46	1.69
2	*CrippenLogP, piPC8*	0.31	1.96	0.50	1.60	0.43	1.62
3	*GATS1i, CrippenLogP, SpMAD_Dt*	0.38	1.86	0.50	1.57	0.50	1.51
4	*GATS1i, CrippenLogP, ZMIC0, SpDiam_D*	0.42	1.80	0.54	1.52	0.54	1.45
5	*ATSC1e, GATS1i, CrippenLogP, SpDiam_D, SubFPC297*	0.48	1.70	0.55	1.49	0.55	1.42
6	*ATSC1e, GATS2m, GATS1i, CrippenLogP, SpAD_D, SubFPC297*	0.52	1.64	0.53	1.50	0.55	1.39
7	***GATS2m, GATS1p, CrippenLogP, SIC3, SpDiam_D, VAdjMat, MACCSFP35***	**0.56**	**1.57**	**0.54**	**1.49**	**0.56**	**1.38**
8	*AATS3e, AATS2p, AATSC1e, GATS1p, CrippenLogP, SpDiam_D, VAdjMat, PubchemFP406*	0.59	1.52	0.51	1.55	0.55	1.40

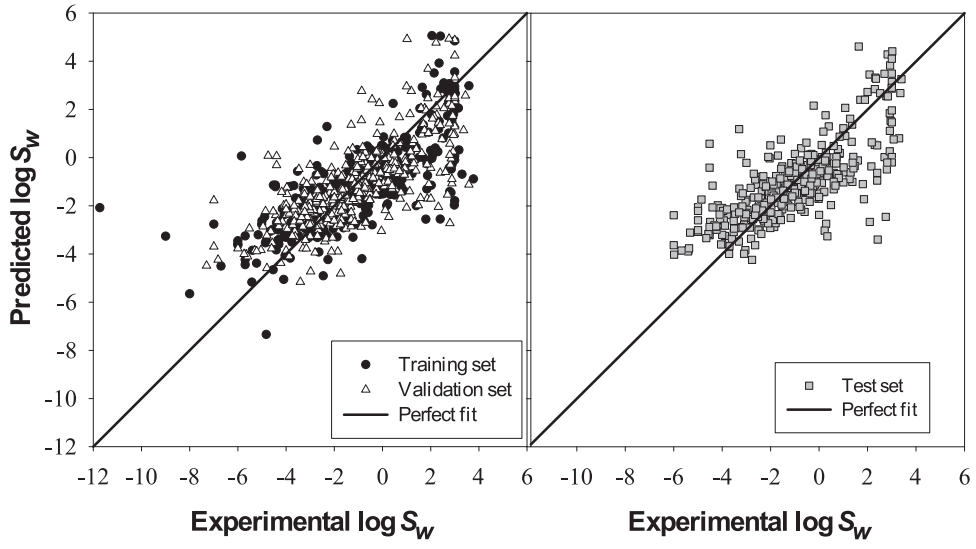

Fig. 1. Experimental and predicted values for the training, validation and test sets for the seven descriptors model (Eq. (1)) applied to 1211 pesticides.

3.2. Flexible descriptors

In order to find the most efficient structural attributes for each SR during the flexible descriptor design, the *DCW* descriptor was optimized by increasing R^2_{train}, until the model lost predictive capability in the validation set, without involving the test set. The statistical parameters for the best QSPR models found by trying different CORAL based combinations are presented in Table 2. Analysis of these results reveals that the best choice is an approach that includes HFG representations. The optimal descriptor involves three variable types, and 168 active attributes are based on them (refer to Table 10S). Fig. 2 shows that the predicted and experimental values for the training, validation, and test sets follow a straight line. The resulting equation for this model with one *DCW* descriptor is:

$$\log S_w = 0.444 + 0.103\ DCW \quad (2)$$

Table 2
Statistical parameters for the training, validation, and test sets during the search for the best QSPR model using flexible molecular descriptors. The best model appears in bold text.

Structural attributes	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	R^2_{test}	$RMSD_{test}$
^1S_k	0.45	1.76	0.44	1.65	0.46	1.53
^2S_k	0.75	1.19	0.54	1.55	0.45	1.56
^0EC_j	0.40	1.83	0.45	1.62	0.47	1.53
$Pt2_k$	0.46	1.74	0.50	1.56	0.46	1.54
NNC_j	0.52	1.64	0.50	1.56	0.49	1.49
VS2	0.59	1.52	0.45	1.7	0.48	1.51
^2S_k, NNC_j	0.69	1.31	0.54	1.52	0.50	1.49
^2S_k, $Pt2_k$	0.74	1.21	0.54	1.52	0.48	1.51
^2S_k, $Pt2_k$, NNC_j	**0.70**	**1.30**	**0.55**	**1.51**	**0.53**	**1.43**

$N_{train} = 404$, $R^2_{train} = 0.70$, $RMSD_{train} = 1.30$, $N_{val} = 404$, $R^2_{val} = 0.60$, $RMSD_{val} = 1.40$, $o(3S)) = 5$

$N_{test} = 403$, $R^2_{test} = 0.54$, $RMSD_{test} = 1.41$, $R^2_{loo} = 0.69$, $RMSD_{loo} = 1.31$, $RMSD_{rand} = 2.33$, $h^* = 0.015$

All compounds are within the applicability domain and systematic error is absent. Five compounds in the training set (**53**, **352**, **764**, **853**, **1120**) showed absolute residuals greater than 3 times $RMSD_{train}$ and were considered as outliers. We applied both Y-randomization to demonstrate that $RMSD_{train} < RMSD_{rand}$ and also the external validation criterion (Roy, 2007) to ensure that a valid structure-activity relationship was achieved:

$$1 - R^2_0/R^2_{test} < 0.1(0.000) \text{or } 1 - R'^2_0/R^2_{test} < 0.1\ (0.12)\ and,\ 0.85 \le k \le 1.15\ (0.90)$$

and $0.85 \le k' \le 1.15\ (0.75)$and $R^2_m > 0.5\ (0.59)$

Table 11S includes an example for the *DCW* calculation of compound 2. The structural attributes that contribute to such *DCW* are listed in Table 10S. The flexible molecular descriptors involved in this model (^2S_k, $Pt2_k$, NNC_j) are all local attributes (Toropov et al., 2017). The ^2S_k is a two-elements SMILES attribute, $Pt2_k$ represents a path length of two, and NNC_j is the nearest neighboring code, a local graph invariant.

3.3. Hybrid descriptors

The third approach explored combines PaDEL, EPI Suite, Mold2, and flexible CORAL descriptors and fingerprints. The combination between various flexible descriptors or between flexible descriptors and conventional molecular descriptors produced robust models with better

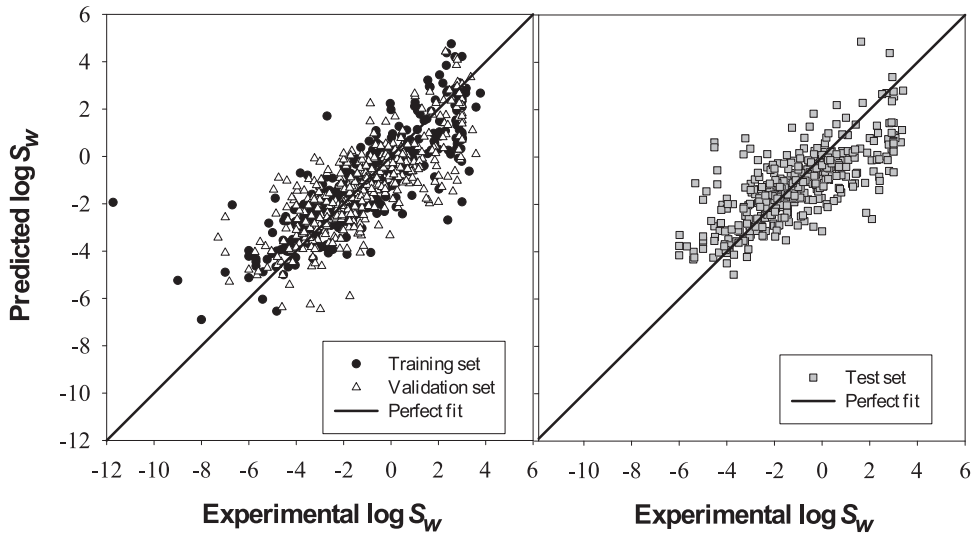

Fig. 2. Experimental and predicted values for the training, validation and test sets for the flexible-descriptor model (Eq. (2)) applied to 1211 pesticides.

predictive capability. The best hybrid model involves three descriptors (Table 3, model 3), including the descriptor called *DCW* which was the best descriptor found in the previous model (flexible molecular descriptors model, Eq. (2)). The model that contains four terms in the hybrid approach resulted in more complexity and did not yield a significantly better performance.

It can be noted from the data, that compound **853** is an outlier in all the proposed models presenting an extremely low log S_w value (–11.73) which means a very low aqueous solubility. Excluding this compound from the training set produced a better model (Table 3, model 3a) represented by Eq. (3). It can be seen from Table 3 and Fig. 3 that such a model shows the best performance among all the explored models:

$$\log S_w = -0.669 + 2.032SIC2 - 0.473MACCSFP106 + 0.108DCW \quad (3)$$

$$N_{train} = 403,\ R^2_{train} = 0.75,\ RMSD_{train} = 1.15,\ N_{val} = 404,$$
$$R^2_{val} = 0.62,\ RMSD_{val} = 1.38,\ o(3S) = 4$$

$$N_{test} = 403,\ R^2_{test} = 0.56,\ RMSD_{test} = 1.37,\ R^2_{loo} = 0.75,$$
$$RMSD_{loo} = 1.16,\ RMSD_{rand} = 2.26,\ h^* = 0.030$$

The *DCW* descriptor, and the *SIC2* descriptor, which denotes the neighborhood symmetry of 2-order both presented a positive correlation with the predicted property. The fingerprint *MACCSFP106*, which indicates the presence of non-aliphatic branching, had a negative coefficient in this model. Eq. (3) also satisfies the external validation conditions (Roy, 2007):

$$1 - R_0^2/R^2_{test} < 0.1(0.012) \text{or } 1 - R_0'^2/R^2_{test} < 0.1\ (0.16) \text{and, } 0.85 \le k \le 1.15\ (0.97)$$

$$\text{and } 0.85 \le k' \le 1.15\ (0.68) \text{and } R^2_m > 0.5(0.55)$$

Four compounds in this model are outliers (compounds **53**, **352**, **764**, **1120**). As discussed in Section 3.1, the solubility value of compound **53** is unreliable and it is an outlier in all the proposed models. The other outliers have very low (**352, 764**) or very high (**1120**) water solubility. This is understandable, since the dataset used is structurally heterogeneous and includes organic and inorganic compounds with very different water solubility values (in a more than 10 log units range). Ali et al. have shown that the sparsely populated extremities of data sets can significantly distort results for linear regression-based models (Ali et al., 2012b). These compounds can artificially inflate or depress the predictive ability of the models. In their detailed studies (Ali et al., 2012a), the ESOL and GES methods were compared with a model which combines a computationally derived molecular descriptor (topographical polar surface area, TPSA) with the experimentally determined melting point and the log *P* (TPSA model). They proposed to use a reduced data set that eliminates the compounds in sparsely populated regions and compare the statistics of the full set with those of this reduced set to assess the robustness of the model. In order to evaluate the model 3a, the compounds outside of the range $-5.00 \le \log S \le < 3.00$ were removed, generating a reduced data set of 1127 compounds. Analysis of the complete dataset including the training, validation and test sub-sets yields a R^2 of 0.65 for the full set and 0.60 for the reduced set. The *RMSD* varies from 1.30 for the full set to 1.23 when the reduced data set is used. This means that exclusion of compounds with extreme solubility values does not affect significantly the predictive ability of the model. However, the almost 8% decrease in R^2 indicates that the robustness of our model is comparable with those of ESOL and GSE (Ali et al., 2012a), but is poorer than the TPSA model that shows a reduction of less than 6% in R^2 when the reduced set is used.

The statistical parameters of this model with hybrid descriptors are

Table 3
Descriptors identified for modeling the water solubility together with the squared correlation coefficient and the standard deviation for the training, validation, and test sets.

#Des.	Descriptors	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	R^2_{test}	$RMSD_{test}$
1	*DCW*[a]	0.70	1.30	0.55	1.51	0.53	1.43
2	*DCW*[b], *MACCSFP106*	0.71	1.28	0.61	1.38	0.55	1.40
3	***DCW, SIC2, MACCSFP106***	**0.73**	**1.22**	**0.61**	**1.38**	**0.56**	**1.38**
4	*DCW, SIC2, MACCSFP106, SubFP236*	0.75	1.19	0.61	1.38	0.56	1.37
3a[b]	***DCW, SIC2, MACCSFP106***	**0.75**	**1.15**	**0.62**	**1.38**	**0.56**	**1.37**

[a] DCW refers to the descriptor obtained using CORAL in HFG representation for the attributes $Pt2_{k.}$, NNC_j, and ^2S_k.
[b] Model 3a are the results of the model 3, with the compound **853** removed from the training set.

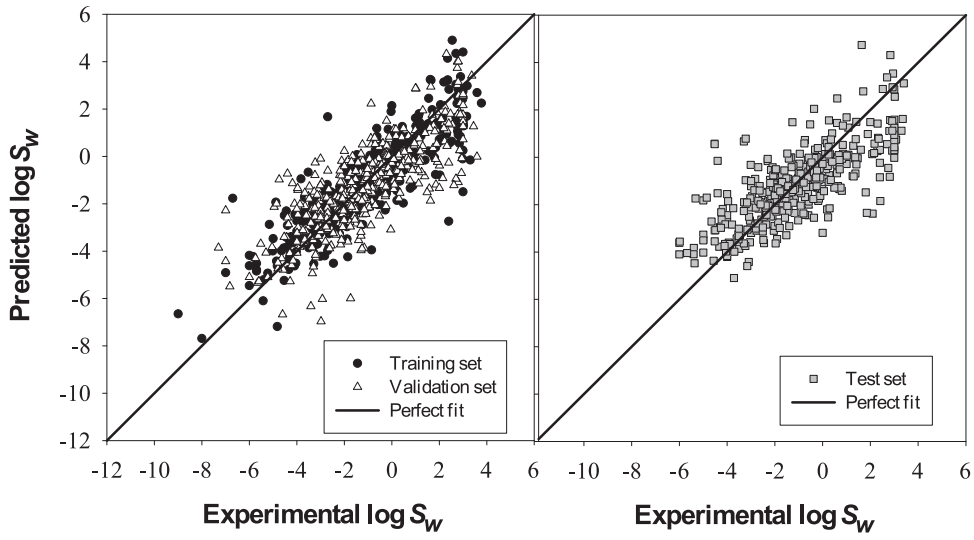

Fig. 3. Experimental and predicted values for the training, validation and test sets for the three-descriptors hybrid model (Eq. (3)) for 1210 pesticides (excluding compound **853**).

comparable with those reported for the general solubility equation of Yalkowsky (Ran et al., 2002). The GSE yields a R^2 of 0.9681 when is used to estimate the water solubility for a data set of 580 organic nonelectrolytes, but the regression coefficient falls to 0.67 when is applied to a larger (2874) and more diverse set of compounds (Delaney, 2004). The ESOL method, on the other hand, gives a R^2 of 0.72 when applied to the same set of 2874 organic compounds. The model proposed here yields a R^2 of 0.75 for the training set (Table 3) of a diverse collection of pesticides, including organic and inorganic compounds. Moreover, it has the advantage over the GSE that the water solubility prediction is based on constitutional parameters only, instead of the experimental data required to apply the GSE. The ESOL method does not use experimental data, but requires the calculation of the *C*log *P* (Delaney, 2004) and its performance is comparable to our 3a model. This means that our QSPR hybrid approach can be applied to estimate the S_w for new compounds knowing just its molecular structure and may also be helpful in the design of potentially less toxic pesticides. In addition, the applicability of our model is not limited to nonelectrolytes, while the ESOL and GSE methods have only been applied to datasets that exclude compounds that may be charged al pH 7.

4. Conclusions

We developed a simple model that successfully predicts the water solubility of a diverse and large set of pesticides through a strategy that does not require the knowledge of the molecular conformation as part of the structural representation. Analysis of the descriptors involved in the models proposed here suggests that the polarizability, the presence of alkali atoms, and the asymmetry of the molecular structure have positive contributions to the water solubility values, whereas the presence of elements heavier than carbon and the lipophobic character of the molecule have a negative correlation. The hybrid approach that involves a molecular descriptor, a fingerprint, and a flexible descriptor calculated with the CORAL software showed the best predictive performance. This model was validated through Y-randomization, Cross-Validation and included a properly defined applicability domain to fully meet the validation principles established by the OECD. Its satisfactory predictive power suggests that this new model could represent a reliable alternative to the experimental assays, helping the registrants of new pesticides to fulfill regulatory requirements in compliance with the ethical and economic necessity to reduce animal testing.

Supporting information summary

Experimental and predicted water solubility values, and *DCW* for the 1211 molecules studied along with the details of each model (correlation matrices, mathematical equations used, description of the molecular descriptors involved, correlation weight values for the structural attributes, flexible descriptor calculation example) are available online as supporting information.

Acknowledgments

We are grateful to the National Scientific and Technical Research Council of Argentina (Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET PIP0311 project); to the National University of La Plata (Argentina) for financial support; and also to the Ministry of Education, Culture, Science and Technology (Argentina) for electronic library facilities. SEF, DEB and PRD are research members of CONICET.

Conflict of interest

The authors confirm that they do not have any conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecoenv.2018.12.056.

References

Agency, USEP, 2017. Finalization of Guidance on Incorporation of Water Treatment Effects on Pesticide Removal and Transformations in Drinking Water Exposure Assessments.

Ali, J., Camilleri, P., Brown, M.B., Hutt, A.J., Kirton, S.B., 2012a. In silico prediction of aqueous solubility using simple QSPR models: the importance of phenol and phenol-like moieties. J. Chem. Inf. Model. 52, 2950–2957.

Ali, J., Camilleri, P., Brown, M.B., Hutt, A.J., Kirton, S.B., 2012b. Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. J. Chem. Inf. Model. 52, 420–428.

Bhhatarai, B., Gramatica, P., 2010. Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. Environ. Sci. Technol. 45, 8120–8128.

Cappelli, C.I., Manganelli, S., Lombardo, A., Gissi, A., Benfenati, E., 2013. Validation of quantitative structure–activity relationship models to predict water-solubility of organic compounds. Sci. Total Environ. 463, 781–789.

Clarke, E.D., Delaney, J.S., 2003. Physical and molecular properties of agrochemicals: an analysis of screen inputs, hits, leads, and products. CHIMIA Int. J. Chem. 57, 731–734.

Cronin, M.T.D., Livingstone, D.J., 2004. Predicting Chemical Toxicity and Fate. CRC press, Boca Raton.
Das, R.N., Roy, K., 2013. QSPR with extended topochemical atom (ETA) indices. 4. Modeling aqueous solubility of drug like molecules and agrochemicals following OECD guidelines. Struct. Chem. 24, 303–331.
Delaney, J.S., 2004. ESOL: estimating aqueous solubility directly from molecular structure. J. Chem. Inf. Comput. Sci. 44, 1000–1005.
Duchowicz, P.R., Castro, E.A., Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun. Math. Comput. Chem. 55, 179–192.
Duchowicz, P.R., Comelli, N.C., Ortiz, E.V., Castro, E.A., 2012. QSAR study for carcinogenicity in a large set of organic compounds. Curr. Drug Saf. 7, 282–288.
Duchowicz, P.R., Fioressi, S.E., Bacelo, D.E., Saavedra, L.M., Toropova, A.P., Toropov, A.A., 2015. QSPR studies on refractive indices of structurally heterogeneous polymers. Chemom. Intell. Lab. Syst. 140, 86–91.
Duchowicz, P.R., Fioressi, S.E., Castro, E., Wróbel, K., Ibezim, N.E., Bacelo, D.E., 2017. Conformation-Independent QSAR Study on Human Epidermal Growth Factor Receptor-2 (HER2) Inhibitors. ChemistrySelect 2, 3725–3731.
Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environ. Health Perspect. 111, 1361.
Gadaleta, D., Mangiatordi, G.F., Catto, M., Carotti, A., Nicolotti, O., 2016. Applicability domain for QSAR models: where theory meets reality. Int. J. Quant. Struct.-Prop. Relatsh. 1, 45–63.
Golbraikh, A., Tropsha, A., 2002. Beware of q2!. J. Mol. Graph Model 20, 269–276.
Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.
Hamadache, M., Amrane, A., Benkortbi, O., Hanini, S., Khaouane, L., Moussa, C.S., 2017. Environmental Toxicity of Pesticides, and its Modeling by Qsar Approaches, Advances in QSAR Modeling. Springer, Cham, pp. 471–501.
Hamilton, D., Ambrus, A., Dieterle, R., Felsot, A., Harris, C., Holland, P., Katayama, A., Kurihara, N., Linders, J., Unsworth, J., 2003. Regulatory limits for pesticide residues in water (IUPAC Technical Report). Pure Appl. Chem. 75, 1123–1155.
Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., Su, Z., Perkins, R., Tong, W., 2008. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J. Chem. Inf. Model. 48, 1337–1344.
Kim, M., Li, L.Y., Grace, J.R., 2016. Predictability of physicochemical properties of polychlorinated dibenzo-p-dioxins (PCDDs) based on single-molecular descriptor models. Environ. Pollut. 213, 99–111.
Lewis, K., Green, A., Tzilivakis, J., Warner, D., 2018. The pesticide properties database (ppdb) developed by the agriculture & environment research unit (AERU). Univ. Herts. 2006–2018.
Mas, S., de Juan, A., Tauler, R., Olivieri, A.C., Escandar, G.M., 2010. Application of chemometric methods to environmental analysis of organic pollutants: a review. Talanta 80, 1052–1067.
OECD, 2007. Guidance Document On The Validation of (Quantitative) Structure-Activity Relationship [(Q)Sar] Models, Environment Health and Safety Publications Series on Testing and Assesment No. 69.
Ran, Y., He, Y., Yang, G., Johnson, J.L., Yalkowsky, S.H., 2002. Estimation of aqueous solubility of organic compounds by using the general solubility equation. Chemosphere 48, 487–509.
Rojas, C., Duchowicz, P.R., Tripaldi, P., Diez, R.P., 2015. QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. Chemom. Intell. Lab. Syst. 140, 126–132.
Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin. Drug Discov. 2, 1567–1577.
Talevi, A., Bellera, C.L., Di Ianni, M., Duchowicz, P.R., Bruno-Blanch, L.E., Castro, E.A., 2012. An integrated drug development approach applying topological descriptors. Curr. Comput. Aided Drug Des. 8, 172–181.
Tebes-Stevens, C., Patel, J.M., Koopmans, M., Olmstead, J., Hilal, S.H., Pope, N., Weber, E.J., Wolfe, K., 2018. Demonstration of a consensus approach for the calculation of physicochemical properties required for environmental fate assessments. Chemosphere 194, 94–106.
Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2013. CORAL: QSPR model of water solubility based on local and global SMILES attributes. Chemosphere 90, 877–880.
Toropov, A.A., Toropova, A.P., Benfenati, E., Nicolotti, O., Carotti, A., Nesmerak, K., Veselinović, A.M., Veselinović, J.B., Duchowicz, P.R., Bacelo, D., 2017. QSPR/QSAR analyses by means of the CORAL software: results, challenges, perspectives, pharmaceutical sciences: breakthroughs in research and practice. IGI Glob. 929–955.
Toropova, A., Toropov, A., Martyanov, S., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2012. CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna. Chemom. Intell. Lab. Syst. 110, 177–181.
Verma, R.P., Hansch, C., 2005. An approach toward the problem of outliers in QSAR. Bioorg. Med. Chem. 13, 4597–4621.
Villaverde, J.J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J.L., Sandín-España, P., 2017. Computational methodologies for the risk assessment of pesticides in the European Union. J. Agric. Food Chem. 65, 2017–2018.
Villaverde, J.J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J.L., Sandín-España, P., 2018. Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. Sci. Total Environ. 634, 1530–1539.
Wilczyńska-Piliszek, A.J., Piliszek, S., Falandysz, J., 2012. QSAR and ANN for the estimation of water solubility of 209 polychlorinated trans-azobenzenes. J. Environ. Sci. Health, Part A 47, 155–166.
Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., 2017. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 46, D608–D617.
Wold, S., Eriksson, L., Clementi, S., 1995. Statistical validation of QSAR results. Chemom. Methods Mol. Des. 309–338.
Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474.
Zeng, X.-L., Wang, H.-J., Wang, Y., 2012. QSPR models of n-octanol/water partition coefficients and aqueous solubility of halogenated methyl-phenyl ethers by DFT method. Chemosphere 86, 619–625.