An Alternative QSAR-Based Approach for Predicting the Bioconcentration Factor for Regulatory Purposes

Andrea Gissi [1*], *Domenico Gadaleta* [1*], *Matteo Floris* [2], *Stefania Olla* [3],
Angelo Carotti [1], *Ettore Novellino* [4], *Emilio Benfenati* [5], *and Orazio Nicolotti* [1]

[1] Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari “Aldo Moro”, Bari, Italy; [2] Biomedicine Sector, CRS4, Parco Polaris, Loc. Pixinamanna, Pula, Italy; [3] IRGB-CNR, Cittadella Universitaria di Cagliari, Monserrato, Italy; [4] Dipartimento di Farmacia - Università degli Studi di Napoli “Federico II”, Napoli, Italy; [5] IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

Summary

The REACH (Registration, Evaluation, Authorization and restriction of Chemicals) and BPR (Biocidal Product Regulation) regulations strongly promote the use of non-animal testing techniques to evaluate chemical risk. This has renewed the interest towards alternative methods such as QSAR in the regulatory context. The assessment of bioconcentration factor (BCF) required by these regulations is expensive, in terms of costs, time, and laboratory animal sacrifices. Herein, we present QSAR models based on the ANTARES dataset, which is a large collection of known and verified experimental BCF data. Among the models developed, the best results were obtained from a nine-descriptor highly predictive model. This model was derived from a training set of 608 chemicals and challenged against a validation and blind set containing 152 and 76 chemicals, respectively. The model's robustness was further controlled through several validation strategies and the implementation of a multi-step approach for the applicability domain. Suitable safety margins were used to increase sensitivity. The easy interpretability of the model is ensured by the use of meaningful biokinetics descriptors. The satisfactory predictive power for external compounds suggests that the new models could represent a reliable alternative to the in vivo assay, helping the registrants to fulfill regulatory requirements in compliance with the ethical and economic necessity to reduce animal testing.

Keywords: REACH, BPR, QSAR, bioconcentration factor, biokinetics descriptors, applicability domain

1 Introduction

Quantitative Structure-Activity Relationships (QSARs) are *in silico* approaches developed to quantitatively predict a certain property/activity (e.g., pharmacological effect or toxicity, defined as endpoint) for a substance of interest. To obtain reliable models the quality of the data should be high. To this end, several preprocessing and pretreatment procedures need to be taken into account to curate preliminary data for model generation. The model is derived from the fundamental principle that the variance contained in the molecular descriptors relates to the variance in the values of the target endpoint (Nicolotti et al., 2002). Then, the model performance needs to be carefully validated to assess the quality of the predictions (Tropsha et al., 2003). There are several techniques to challenge models. In this respect, the external validation was the most appropriate approach for model control. Such an approach allows simulation of real life uses of the model. This allows prediction of the response for those chemicals included in a so-called external set, i.e., excluded from the model derivation. This procedure represents the proof of the capability of a given model to predict the properties of unknown compounds (Golbraikh and Tropsha, 2002). Despite that QSAR has acquired a more and more relevant role in numerous front-line approaches of several fields, from experimental design to ADME modeling, to drug discovery, etc. (Nicolotti et al., 2008, 2009), the actual impact on real-life applications often has been considered elusive and ineffective (Doweyko, 2004). The new European legislations, i.e., REACH (Registration, Evaluation, Authorization and restriction of Chemicals) (EC, 2006) and BPR (Biocidal Product Regulation) (EU, 2012), have refreshed and renewed the crucial role of QSAR. Indeed, REACH Article 1 encourages the use of alternative methods (*in silico* among others) for assessing the presence or absence of hazardous properties of chemical substances, which, at the same time, minimize the costs of experiments and the controversial use of vertebrate animals (EC, 2006).

Received May 22, 2013; accepted in revised form October 29, 2013; Epub November 13, 2013; http://dx.doi.org/10.14573/altex.1305221

* Authors contributed equally to this work.

To be accepted in a regulatory context, the application of QSAR has to fulfil some basic principles that ensure the reliability of the predictions. In this respect, the Organization for Economic Co-operation and Development (OECD) has stated that QSAR models have to be characterized by: 1) a well-defined endpoint; 2) an unambiguous algorithm for model derivation; 3) a clearly defined domain of applicability; 4) appropriate measures of goodness-of-fit, robustness, and predictivity; and 5) a mechanistic interpretation, if possible. Explanatory comments are provided for each point in the OECD document (OECD, 2007). In a similar manner, the European Commission (EC) has established, in Annex XI of REACH and Annex IV of BPR, four conditions for use of QSARs instead of *in vivo* testing: 1) results have to be derived from a QSAR model whose scientific validity has been well established; 2) the substances are expected to fall within the applicability domain of the QSAR model; 3) results need to be adequate for the purpose of classification and labeling and/or risk assessment; and 4) adequate and reliable documentation of the applied method has to be provided (EC, 2006). To date, a wide series of models exists that addresses almost every endpoint within REACH. In this respect, a list of software and models has been made available on the website of the European project ANTARES[1]. This open access list includes tens of models, of which some are commercial and others are freely available. Nevertheless, only some of these models have been developed for regulatory purposes and are fully compliant with the REACH requirements listed above.

Among others, the bioconcentration factor (BCF) is an endpoint of utmost relevance owing to its (eco)toxicological impact. It represents the ratio of the concentration of a substance in an aquatic organism with respect to that in water (Arnot and Gobas, 2006). There is an ongoing discussion about what is the most suitable surrogate parameter for bioaccumulation. In the future, BCF may be substituted with bioaccumulation factor (BAF), which takes also into account the exposure via the diet. To date, BCF is still the reference endpoint under REACH for Persistent, Bioacculumative and Toxic (PBT) classification (EC, 2006).

The experimental test guideline TG 305, recommended by OECD, requires the use of hundreds of fish, months of test execution (OECD, 2012), and a total cost of thousands of euros. REACH states that animal testing should be a last resort (EC, 2006). It goes without saying that the use of *in silico* approaches, like QSARs, can lead to significant savings in terms of money, time, and above all else, laboratory animals. The use of toxicological evidence from QSARs can reduce and replace the execution of further useless *in vivo* assays according to the 3Rs principle: Replace, Reduce, Refine (Russell and Burch, 1959). Continuous efforts to derive new and better predictive models have led to the development of several software programs (i.e., EPISuite BCFBAF module, T.E.S.T., and VEGA) that are, at present, widely used for the prediction of many endpoints, including BCF. However, the availability of newer and higher quality experimental BCF measures, along with the chance of using a more attractive descriptor space, prompted us to derive new QSAR models for BCF. In fact, within the ANTARES project, a dataset of 851 compounds, whose structures and experimental BCF values have been carefully checked, has been compiled to evaluate the performances of existing models. We used this newer data collection as well as biokinetics descriptors to derive and test the new models presented in this work.

2 Materials and methods

2.1 Data set

The ANTARES dataset (Gissi et al., 2013) comprises 851 chemicals. Their BCF experimental values have been collected among five reliable and publicly available databases:

- Dimitrov et al. (2005): contains 511 compounds along with unique, reliable BCF experimental data for each chemical;
- Fu et al. (2009): contains 138 ionizable chemicals. Only 10 of these are characterized by multiple BCF experimental values;
- Footprint PPDB[2] (2013): contains unique experimental values for 159 pesticides;
- Arnot and Gobas (2006): comprises only experimental data for fish species and aquatic organisms indicated by OECD 305 guidelines (OECD, 2012) (*Danio rerio, Pimephalespromelas, Cyprinus carpio, Oryziaslatipes, Poeciliareticulata, Lepomismacrochirus, Oncorhynchusmykiss, and Gasterosteusaculeatus*) with an overall reliability score of 1 (the most reliable data); contains unique or multiple experimental BCF data for 759 compounds;
- EURAS[3]: contains 511 reliable data points for fish species suggested by OECD 305.

As reported (Gissi et al., 2013), the structures were carefully checked and the values were selected if in agreement with the conditions established by OECD TG 305 (OECD, 2012). Compounds characterized by ambiguous data, inorganic compounds, or isomeric mixtures were thus discarded. The presence of duplicates was verified. Multiple values for the same chemical were mediated by geometric mean. The list of the compounds, with their SMILES, names, InChI notations and experimental values were compiled in Table S1 (supplementary data at http://www.altex-edition.org). This dataset is larger than those used for other existing BCF predictive models tailored for specific regulatory purposes, such as the CAESAR model[4] (Zhao et al., 2008; Lombardo et al., 2010) implemented in VEGA platform[5] (473 substances), the Meylan model (Meylan et al., 1999) implemented in EPISuite BCFBAF module[6] (527 substances), and the T.E.S.T. model[7] (598 substances). For the derivation of the model presented here, the ANTARES dataset was split into three subsets: about 10% (78 out of 851) of the compounds were randomly selected to form the blind set (BS) required for final validation. The remaining chemicals were split to form a training set (TS) and a vali-

[1] http://www.antares-life.eu (accessed 23.04.2013)

[2] http://sitem.herts.ac.uk/aeru/footprint/en/ (accessed 23.04.2013)

[3] http://www.cefic-lri.org (accessed 23.04.2013)

[4] http://www.caesar-project.eu (accessed 23.04.2013)

[5] http://www.vega-qsar.eu/ (accessed 23.04.2013)

[6] http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (accessed 23.04.2013)

[7] http://www.epa.gov/nrmrl/std/qsar/qsar.html (accessed 23.04.2013)

dation set (VS) containing 620 and 153 chemicals, respectively. We ensured a uniform distribution of their experimental BCF values by applying the Venetian blinds method (Consonni et al., 2009). These selection criteria were used to obtain two different and independent sets for model validation (i.e., VS) and to ensure the most realistic situation for the external compounds (i.e., BS). Given this, statistics could demonstrate the real capability of the model to predict new compounds.

2.2 Generation and selection of novel biokinetics descriptors

Many commercial and free software programs are available for the calculation of two- (2D) or three-dimensional (3D) descriptors. For instance, the widely used software Dragon version 6.0.28 (Talete srl: Milano, Italy) enables the calculation of an overwhelming number of descriptors (Todeschini, 2000) (i.e., 4885 2D and 3D molecular descriptors, divided into 29 logical boxes, such as constitutional descriptors, topological or connectivity indices, drug-like indices, matrix-based descriptors, Burden eigenvalues, autocorrelations, etc.). Conversely, the program QikProp version 3.4, included in the Maestro 9.2 suite (Schrödinger, LLC, New York), calculates a smaller number of descriptors, which are all relevant to explain ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties of organic molecules. Some of these descriptors, for instance associated with the permeation of the membranes, may be more familiar to a toxicological or pharmacological audience and suggest a mechanistic interpretation.

QikProp provides 51 2D and 3D descriptors. The calculations are fast, up to 10,000 molecules per h in the software's "normal mode" and 300,000 per h in the "fast mode." The interested reader is referred to the QikProp software guide for a detailed description. In this work, we present models based either on QikProp or Dragon descriptors, as well as "hybrid descriptor models" resulting from the combination of QikProp and Dragon descriptors.

The software Dragon failed to determine the correct Kekule structure of two heterocycle-based compounds contained within the TS (i.e., 19 and 735) that have been consequently excluded for model generation. As a consequence, Dragon models have been developed using a final TS made of 618 chemicals, while VS and BS maintained the original size of 153 and 78 chemicals, respectively.

For models based on QikProp descriptors, chemical structures were previously built using CORINA version 3.4 (Molecular Networks GmbH, Erlangen, Germany) software and then minimized by means of MacroModel 9.9 package (Schrödinger, LLC, New York) within the software Maestro 9.2. The molecular descriptors were then computed for each of those chemicals. MacroModel failed to minimize structures containing tin atoms (nine compounds). In addition, QikProp was not able to treat peroxides (two compounds) and quaternary ammoniums (four compounds). Thus, these chemicals were excluded from the relative models for further analyses, reducing the total dataset used for QikProp models to 836 molecules: 608 forming the TS, 152 the VS, and 76 the BS (the excluded compounds are listed in Table S2 in the supplementary data at http://www.altex-edition.org).

The "hybrid descriptor models" were then derived from the union of Dragon and QikProp descriptors. This led to the exclusion of those chemicals not properly computed by either Dragon or QikProp. As a result, in this case the whole dataset is made of 834 chemicals, while the corresponding TS, VS, and BS contain 606, 152, and 76 molecules, respectively.

2.3 Model derivation and validation

Several descriptor subsets have been used for the generation of models. These subsets contain an increasing number of optimal descriptors, from one to ten, which were selected using the Monte Carlo algorithm (Simulated Annealing) implemented in Canvas version 1.5 included in Maestro 9.2 Suite (Schrödinger, LLC, New York). The selection method has been applied to: 1) Dragon descriptors pool; 2) QikProp descriptors pool; and 3) the union of both. Afterwards, Multiple Linear Regression (MLR) and Neural Network algorithm (NN) implemented in Canvas 1.5 were employed to derive models from each descriptor subset. The lists of descriptors selected by Monte Carlo are shown in Table S3 (supplementary data at http://www.altex-edition.org).

All models have been derived via regression approaches, enabling the prediction of continuous BCF values, as indicated in BPR for active substances present in biocidal products and REACH for those chemicals exceeding 100 tons/year. In this regard, the determination coefficient (r^2) and the Root Mean Square of Errors (RMSE) were calculated to appreciate the goodness of regression.

The following equations were used for the calculation of r^2 and RMSE, respectively:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_{avg})^2} = 1 - \frac{RSS}{TSS}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

where y_i is the observed dependent variable (the experimental response), $\hat{y}_i$ is the calculated value, y_{avg} is the mean value of the dependent variable, RSS is the Residual Sum of Squares, and TSS is the Total Sum of Squares for n elements of the modeled data set (Nicolotti and Carotti, 2006).

In addition, the Concordance Correlation Coefficient (CCC) was measured and tabulated for fairly comparing the real external predictivity of the QSAR models discussed herein (Lin, 1989). It has been demonstrated (Chirico and Gramatica, 2011) that the CCC is one of the most reliable criteria to assess the real external predictivity of QSAR models. The following equation accounts for external data only:

$$CCC = \frac{2\sum_{i=1}^{n_{EXT}}(y_i - y_{avg})(\hat{y}_i - \hat{y}_{avg})}{\sum_{i=1}^{n_{EXT}}(y_i - y_{avg})^2 + \sum_{i=1}^{n_{EXT}}(\hat{y}_i - \hat{y}_{avg})^2 + n_{EXT}(y_{avg} - \hat{y}_{avg})^2}$$

Furthermore, the obtained regression models can be flexibly used as classifiers establishing different thresholds according to a given purpose. In this regard, the models have been first evaluated on the basis of the REACH risk thresholds established in Annex XIII to classify chemicals (EC, 2006). All substances that exceed the first threshold of log BCF = 3.3 are classified as bioaccumulative (B) while those having log BCF <3.3 are classified as non-bioaccumulative (nB) according to the PBT definition. On the other hand, all substances that exceed the second threshold of log BCF = 3.7 are classified as very bioaccumulative (vB).

Moreover, models could be easily adapted for working on different threshold values, such as that related to Classification, Labeling and Packaging (CLP) (i.e., log BCF = 2.7) (EC, 2008) and that concerned with Dangerous Substances Directive (64/548/EEC) according to risk phrase R53 (i.e., log BCF = 2.0) (EEC, 1967).

As far as classification is concerned, the Cooper statistics (Cooper et al., 1979) related to accuracy, specificity, and sensitivity have been calculated, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

Accuracy, also termed concordance, measures the correctness of the prediction. Its value is obtained by dividing the number of correct predictions by the total number of compounds. Specificity is concerned with the number of negative compounds correctly predicted (true negatives, TNs) and its value decreases with the occurrence of false positives (FPs). Sensitivity is of the utmost importance to track false negatives (FNs) whose number should be kept low to avoid the occurrence of toxic compounds wrongly predicted as not hazardous. In this respect, the regulatory context recommends a precautionary approach to minimize health and environmental risks. The sensitivity at 3.3 is used in the present work for preliminarily evaluating the results (together with the r^2) coming from different derived models.

It is worth noting that the ANTARES dataset shows an uneven distribution of nB (about 85%) and B or vB (about 15%) compounds. This reflects the common value distribution of BCF, but makes the prediction of B and vB compounds extremely challenging. This explains why lower sensitivity values occur for the threshold of log BCF equal to 3.3 and 3.7.

In this regard, the Matthews Correlation Coefficient (MCC) is useful for the evaluation of classification between two very unbalanced categories, such as in this case (Baldi et al., 2000).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The individual contribution of each descriptor was also evaluated by X-scrambling which, unlike Y-scrambling used for validation, was carried out to assess the relative weight of each descriptor within the model. Iteratively, the values of each of the model descriptors were randomly reshuffled among the TS chemicals, keeping the others descriptors unvaried. This allows the derivation of a number of so-called pseudo-models, one for each descriptor reshuffling. The performance (as r^2) of pseudo-models was compared with that of the original model to appreciate the actual relevance of each single descriptor. A substantial drop of the r^2 of a pseudo-model, as a consequence of the reshuffling of a given descriptor, flags the primary role of the descriptor for the model.

2.4 Software

The molecular geometry for chemicals in the ANTARES dataset was determined using OPLS_2005 force field and the PRGC minimization method as implemented in MacroModel version 9.9 (Schrödinger, LLC, New York). The number of iterations and the convergence threshold were equal to 50,000 and to 0.005, respectively. All the calculations were executed using water as a solvent (Banks et al., 2005).

The Monte Carlo (Simulated Annealing) implemented in Canvas 1.5 made use of the following parameters: number of Monte Carlo steps equal to 1000; initial temperature equal to 0.5 times the standard deviation in the Y variable; final temperature equal to 0.05 times the standard deviation in the Y variable.

The NN implemented in Canvas 1.5 has a standard architecture: one input layer, one hidden layer, and one output layer. The number of neurons in the input layer is equal to the number of descriptors used for the derivation of each model (from one to ten), while in the output layer there is one neuron. The number of neurons in the hidden layer is (input+output)/2. There is no fixed rule about how many hidden neurons a network should contain. A "triangular" shape (such as 9x5x1 in case of the nine-descriptor model) is a standard approach because it places a sensible limit on the number of adjustable parameters. In our case, each Canvas NN was trained using a Broyden, Fletcher, Goldfarb, and Shanno (BFGS) algorithm in the presence of a cross-validation set and training is halted when the cross-validation error starts to increase. This is an extremely effective way to prevent over-fitting. The NN made use of the following parameters: train a total of 20 networks with a number of training cycles equal to 200, keeping only the best network generated.

3 Results

3.1 Model evaluation

The results for the models based on QikProp descriptors built with NN are summarized in Table 1. More detailed statistics about regression and classification performance for these models and those based on other pools of descriptors are available in Table S3 (supplementary data at http://www.altex-edition.org).

In general we observed that NN returned better statistics than MLR. In addition, the comparison between models based on QikProp and Dragon descriptors disclosed that the former are characterized by a far better performance on TS and even higher on VS and BS. In particular, the sensitivity is approximately equal to 70% and 50% for VS and BS when using QikProp descriptors, while it drops by about 10%-20% on average when considering Dragon based models. Statistics also shows that results do not benefit from the combination of Dragon and QikProp descriptors.

Herein we discuss two QikProp based models derived via NN. The first is a straightforward three-descriptor model showing appreciable performance; the second is a nine-descriptor model returning more solid statistics.

Among others, the three-descriptor model returns the best values of r^2 (0.70) and RMSE (0.71) for the VS despite its low number of descriptors. However, the poor sensitivity results, especially those measured at the threshold equal to 3.7 (40% on TS, 41% on VS, and 33% on BS), represent a serious limit for the real use of this model, at least for the vB threshold. On the other hand, the nine-descriptor model returns very encouraging statistics except for a lower value of r^2 for VS (0.635). This event is basically due to few compounds acting as outliers and thus dropping all the statistics values relative to the dataset to which they belong. However, the exclusion of these compounds, later found to fall outside the model applicability domain (AD), allows compensation for such a drawback and results in considerable improvements (see Section 3.2).

3.2 Model applicability domain

The AD represents the space of reliability of a given QSAR model and thus predictions provided by models without a clearly defined AD should be considered useless and meaningless (Jaworska et al., 2005; Aptula and Roberts, 2006; Roberts et al., 2006). As previously described, its importance has also been cited in REACH Annex XI, BPR Annex IV, and OECD principles for the derivation of acceptable QSARs. In this respect, it goes without saying that even robust, intuitive, and validated QSAR models are unsuitable to confidently predict chemicals falling outside their AD (Weaver and Gleeson, 2008). In geometric terms this means that predictions are acceptable only if they are the result of interpolations, but not of extrapolations in the chemical space.

The AD represents a multi-faceted concept, which can be studied at different levels (Eriksson et al., 2003). We paid attention to deriving a well-defined and reproducible procedure based on simple rules for deciding whether a chemical is inside or outside the AD. In this respect, the effectiveness in the application of AD (i.e., chemicals inside/outside the AD) was further assessed by controlling statistics of new models. Our efforts have been focused on the nine-descriptor QikProp based model to which we applied a multi-step filter system to confidently designate chemicals within the AD having the matching criteria requested at any step. In particular, our approach is based on the application of four independent filtering methods, whose outputs are condensed to identify chemicals outside AD with higher confidence and transparency (Schultz et al., 2007). Thus, compounds violating even only one filter are considered outside the AD; Figure 1 shows their corresponding chemical structures.

As a first independent approach, we have explicitly considered the dataset's structural diversity. Chemicals were classified by using the organic functional group (nested) profiler available in the OECD QSAR Toolbox 3.0 software[8]. TS, VS, and BS chemicals were assigned to 102 representative chemical classes. Chemicals containing functional groups belonging to different categories were consequently assigned to multiple classes. It was observed that a good overlap exists between the structural types represented within the TS and those occurring in the VS and BS. A complete overview of the distribution of the three datasets among the defined chemical classes is reported in Table S4 (supplementary data at http://www.altex-edition.org).

To minimize the risks derivable from poor structural coverage, we assumed that the VS or BS compounds belonging to chemical classes represented by less than two TS chemicals could not be confidently predicted (Toropov and Benfenati, 2008), thus we placed them outside the AD. Such a procedure allowed us to identify 17 chemicals (14 from the VS and 3 from the BS, respectively, i.e., 146, 178, 191, 337, 443, 459, 496, 536, 543, 665,

[8] http://www.qsartoolbox.org/ (accessed 23.04.2013)

Tab. 1: Performance in regression and classification for QikProp based models derived via Neural Networks
Each model is made up using one to ten descriptors. TS, VS, and BS consist of 608, 152, and 76 chemicals, respectively.

Number of descriptors	TS		VS		BS	
	r^2	3.3 sensitivity	r^2	3.3 sensitivity	r^2	3.3 sensitivity
1	0.523	0.000	0.565	0.000	0.531	0.000
2	0.588	0.213	0.640	0.208	0.560	0.000
3	0.665	0.528	0.696	0.667	0.628	0.400
4	0.642	0.517	0.629	0.583	0.593	0.400
5	0.667	0.506	0.673	0.583	0.624	0.500
6	0.689	0.461	0.670	0.583	0.626	0.400
7	0.691	0.517	0.641	0.708	0.543	0.500
8	0.714	0.573	0.671	0.708	0.638	0.500
9	0.731	0.584	0.635	0.750	0.623	0.500
10	0.755	0.596	0.668	0.708	0.633	0.500

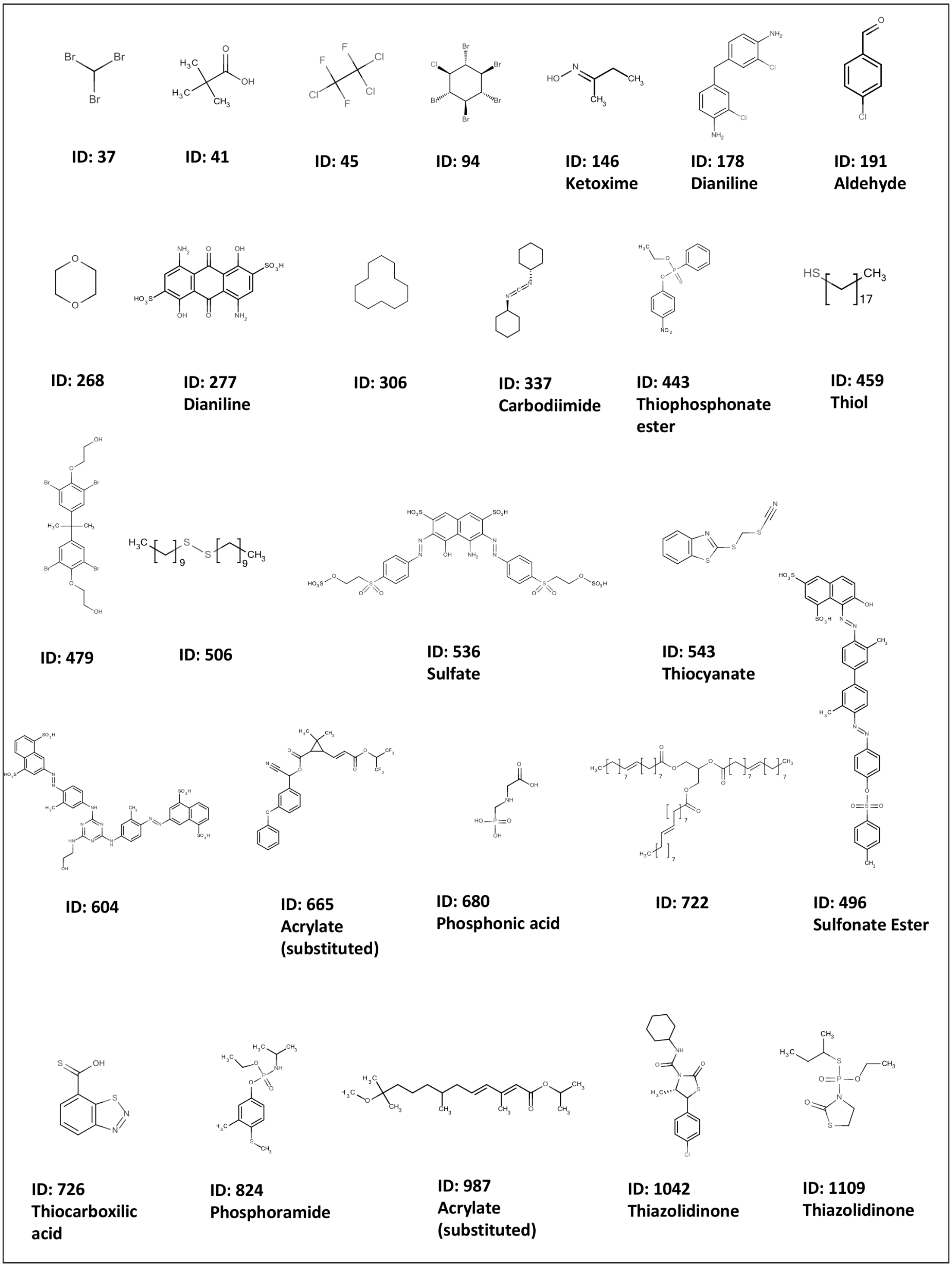


Fig. 1: Chemical structures of compounds outside the AD
Chemical classes have been explicitly reported for chemicals poorly covered in TS.

680, 824, 1042, 1109, 227, 726, 987) which were in 14 chemical classes (i.e., dianilines, sulfonate esters, thiazolidinones, substituted acrylates, phosphonic acids, phosphoramides, ketoximes, carbodiimides, aldehydes, thiophosphonate esters, sulfates, thiocynates, thiols, and thiocarboxylic acids) rarely or never represented within the TS. The exclusion of these compounds had the effect of increasing the confidence in prediction and, indirectly, improving the statistics (r^2 from 0.635 to 0.691 for the VS and from 0.623 to 0.634 for the BS). These results demonstrate the effectiveness of this step in the definition of the AD.

The second independent filter was based on the calculation of the chemical descriptors' ranges. The minimum and maximum values of the nine descriptors in the model for TS chemicals were used to define their interval of validity. Each chemical in VS or BS having one or more descriptors whose values fall outside these ranges were considered outside the AD. In this regard, four

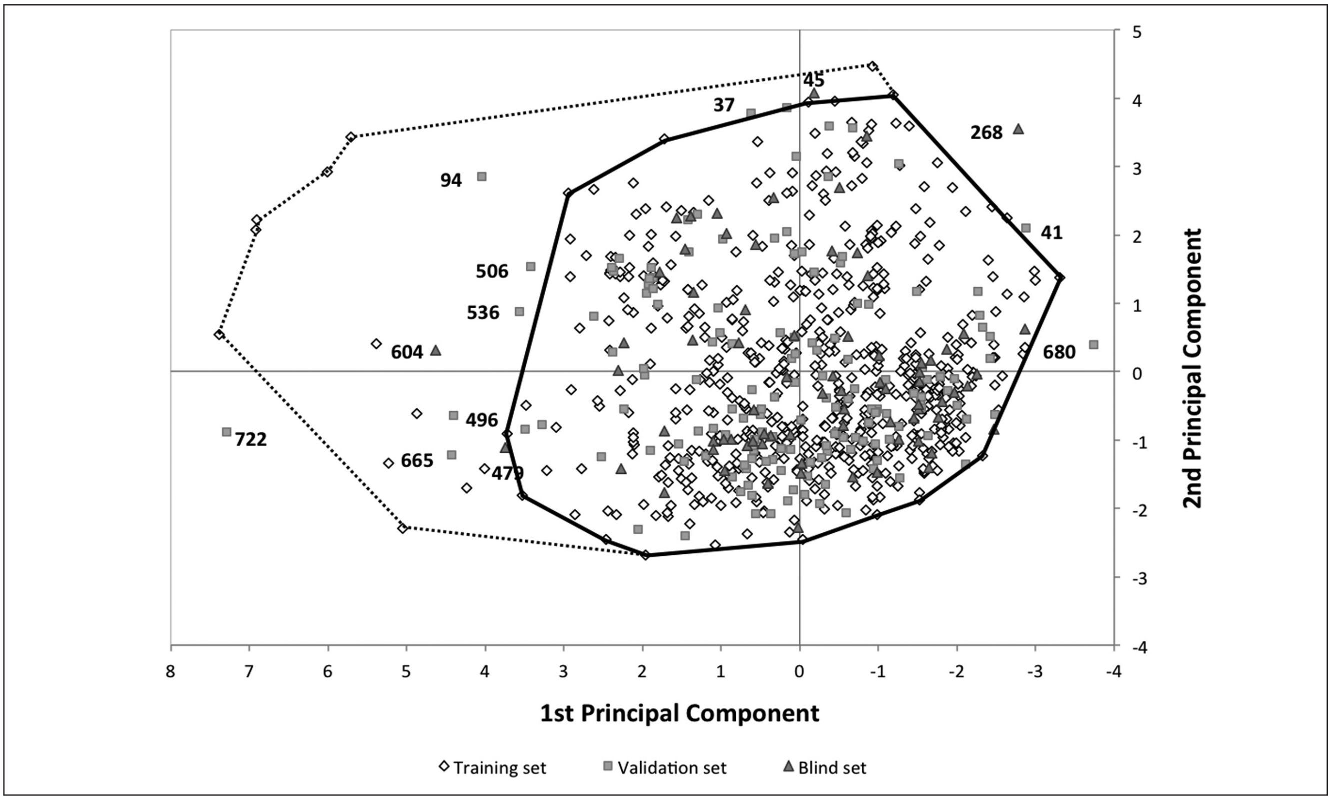


Fig. 2: Geometrical-based applicability domain based on the PCA of the nine-descriptor BCF model
The outer polygon (dashed line) takes into account all the chemicals in the TS, while the inner polygon (solid line) retains about the 98% of them on the basis of a user-dependent inclusive threshold. VS and BS chemicals outside the inner 98% polygon are flagged as outside AD.

chemicals (three from VS, one from BS, i.e., 536, 604, 680, 722) were placed outside the AD. Indirectly, the exclusion of these chemicals shifted the r^2 values from 0.635 to 0.696 for VS and from 0.623 to 0.652 for BS.

Distance-based strategies were thus applied for the definition of the AD (Minovski et al., 2013). A geometrical approach was then taken into account to identify an interpolation region space representing the smallest convex area. The borders of this area describe the perimeter of a polygon containing TS compounds, as shown in Figure 2. Being in a multivariate descriptor space, the interpolation polygon was drawn using spatial coordinates of the first two principal components, derived from the nine descriptors of the model. To avoid the inclusion of underrepresented areas likely increasing the prediction uncertainty, the polygon area was conservatively restricted to contain the top 98% TS compounds on the sole basis of their closeness to the TS centroid. Such a percentage of inclusion was well suited to ensure a uniform coverage of the TS.

In so doing, 13 VS and BS compounds falling outside the polygon (i.e., 37, 41, 94, 496, 506, 536, 665, 680, 722, 45, 268, 479, 604) were placed outside the AD. This produces an indirect effect of improving model performance by increasing the r^2 from 0.635 to 0.730 after the exclusion of nine chemicals from VS, and from 0.623 to 0.643 leaving out four chemicals from BS.

As a further filtering approach the leverage method was applied. By definition, the leverage represents the compound distance from the model's experimental space (that is the center of TS observations). Thus, it is suitable for explicitly evaluating the degree of influence that a particular TS chemical structure has on the model, or the degree of extrapolation for the prediction of VS and BS compounds. In particular, a given prediction is considered unreliable for VS and BS compounds when leverage values exceed the critical threshold of $h^* = 3p'/n$ (where p' is the number of model variables plus one and n is the number of TS compounds) (Gramatica, 2010). As a result, VS and BS compounds having leverages lower than h* are closely structurally-related to TS chemicals and thus comparable to them in terms of the probability of BCF concordance. Conversely, VS and BS chemicals with high-leverage values (>h*) are assumed to be structurally distant from TS chemicals and thus outside the model AD. The William plot shown in Figure 3 gives an immediate idea of the relationships existing between the standardized residuals and the leverage values. As clearly itemized, six compounds (five from VS and one from BS, i.e., 94, 306, 496, 536, 722, and 604) violated the h* warning threshold. The effectiveness of this procedure characterizing the AD is proved by the fact that the removal of these chemicals had the effect of improving r^2 from 0.635 to 0.735 (VS) and from 0.623 to 0.652 (BS).

The application of this harmonized approach, based on the use of the four independent filtering steps, results in an effective strategy for AD. Each step contributes to this process. The simultaneous application of the multi-filter system has the net effect of identifying outside the AD: a) 20 (13% of the initial) VS compounds, which have been shown to be mainly outliers because r^2

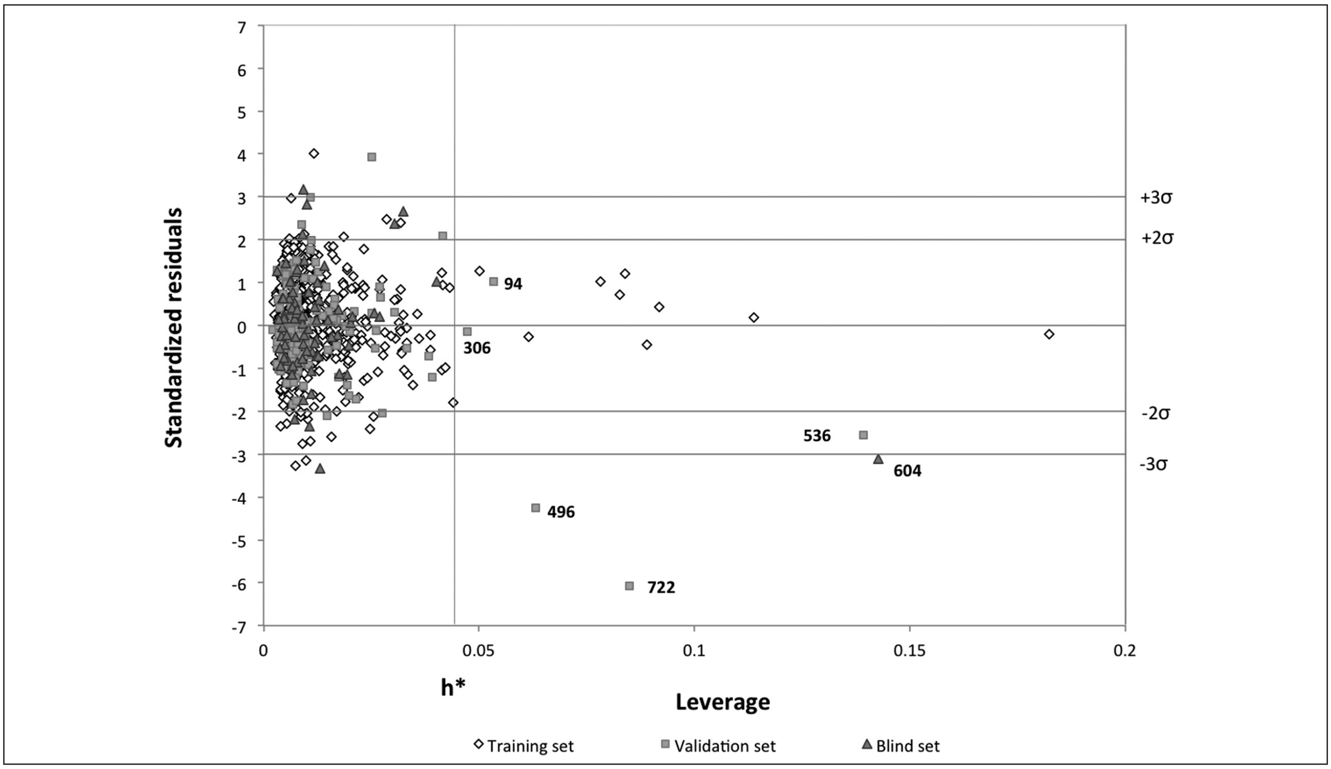


Fig. 3: William's plot reporting the leverage values for TS (diamonds), VS (squares), and BS (up-side triangles) chemicals towards the corresponding standardized residuals computed using the nine-descriptor BCF model

VS and BS chemicals exceeding the h* threshold value (see Section 3.2) are flagged as out of the model AD.

changed from 0.635 to 0.765 and RMSE from 0.794 to 0.616; b) seven (9% of the initial) BS compounds with an indirect gain of r^2 from 0.623 to 0.659 and of RMSE from 0.841 to 0.817.

3.3 Model validation

The nine-descriptor model proposed was further challenged by applying a number of additional validation tests. In this respect, the internal predictivity was controlled by measuring the Predictive Residual Sum of Squares (PRESS) and, thus, q^2:

$$PRESS = \sum_{i=1}^{n} (y_{i/i,pred} - y_i)^2$$

$$q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n} (y_i - y_{avg})^2}$$

The predicted response ($y_{i/i,pred}$) for each chemical was calculated on the basis of its experimental values (y_i), its model calculated activity ($\hat{y}_i$), and the corresponding leverage values (h_{ii}) as follows (Nicolotti and Carotti, 2006):

$$\hat{y}_i - y_{i/i,pred} = (y_i - \hat{y}_i)\frac{h_{ii}}{1 - h_{ii}}$$

A Y-scrambling procedure, based on ten random reshufflings of the response variable, was thus implemented to ascertain that the obtained nine-descriptor model is not the result of chance correlations. In doing so, we observed that the highest randomized q^2 dropped to 0.102 (being its average of 0.027 after 500 randomizations) from the unscrambled value of 0.722. Similarly, the averages of the scrambled r^2 and RMSE values were equal to 0.057 and 1.263, respectively (Table S6 in supplementary data at http://www.altex-edition.org).

Finally, we carried out a series of analyses to assess the real agreement existing between the experimental and predicted values. The perfect match would be showing that the regression line was passing through the origin with a slope equal to 1. As proposed (Golbraikh and Tropsha, 2002; Zhao et al., 2008; Chirico and Gramatica, 2011), we measured the determination coefficient calculated by forcing the regression line to pass through the origin. In doing this, the measure was carried out considering two possible regressions. The first was obtained by comparing experimental versus predicted data (i.e., r^2_0) and the second was obtained by comparing predicted versus experimental data (i.e., r'^2_0). Among the two, the one returning the highest value is retained. It would be desirable to obtain quite similar coefficients and a good match with the r^2 value. In particular, the following requisites have been proposed:

$$|r_0^2 - r'^2_0| < 0.3$$

$$\frac{(r^2 - r_0^2)}{r^2} < 0.1 \text{ or } \frac{(r^2 - r'^2_0)}{r^2} < 0.1$$

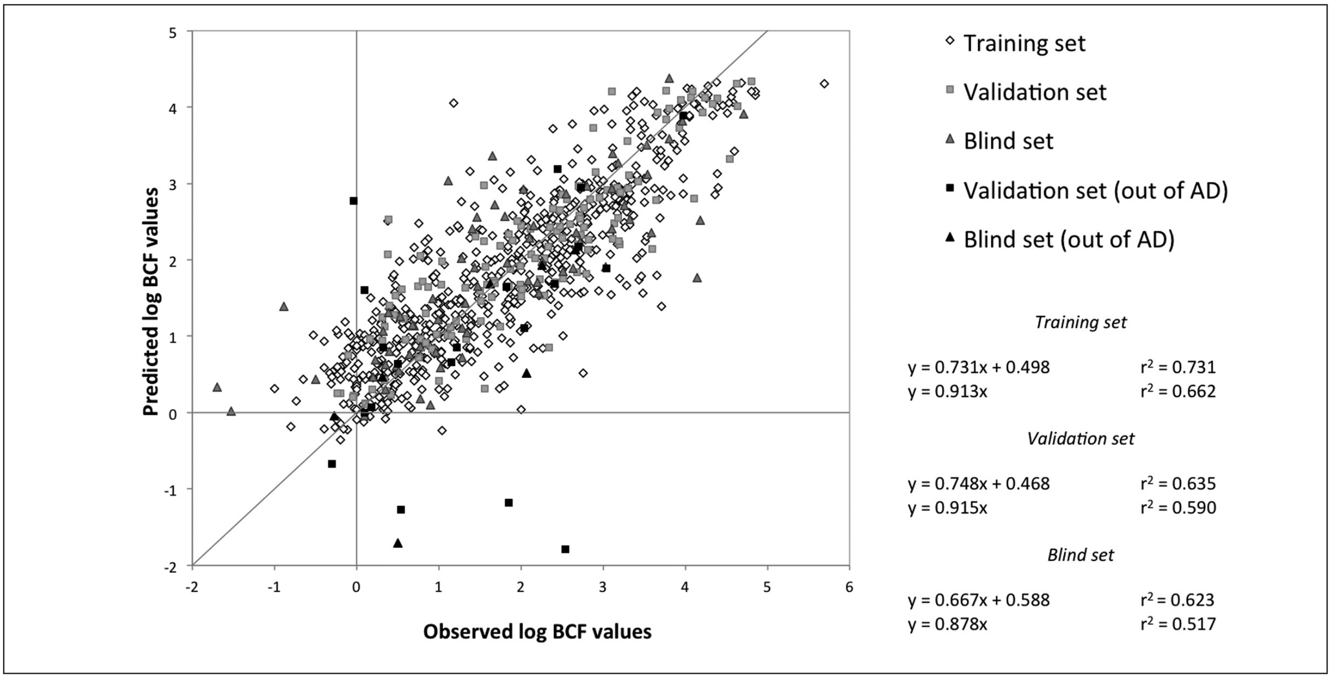


Fig. 4: Comparison of the experimental and predicted log BCF values obtained using the nine-descriptor BCF model
TS, VS, and BS chemicals are represented by white diamonds, gray squares, and up-side triangles, respectively. VS and BS chemicals outside the AD are represented by black squares and up-side triangles respectively. The continuous line represents the case of ideal correlation.

Furthermore, it must be verified that the slopes of the regression lines (i.e., k and k' related to r^2_0 and r'^2_0, respectively) are not too far from 1. It is suggested that k and k' must be in the range of 0.85 to 1.15 (Zhao et al., 2008). It must also be verified that the intercepts of the regression lines related to r^2 for the disposition of the axes (b, for experimental vs predicted data and b' for predicted vs experimental data) are both not too far from 0.

Data shown in Table 2 demonstrates that our nine-descriptor model fulfills all the prerequisites for the three datasets. Indeed, a good match exists between experimental and predicted data (Fig. 4), supporting the scientific validity of the proposed BCF model.

To confirm that the nine-descriptor QSAR model performance is not strictly dependent on the dataset composition, the ANTARES database was repartitioned five other times to obtain five diverse TS, VS, and BS (Table S7 in supplementary data at http://www.altex-edition.org). Likewise, for the initial splitting, BS was extracted at random while TS and VS were created applying the Venetian blinds method. As shown in Table S8 (supplementary data at http://www.altex-edition.org), the new derived nine-descriptor models, irrespective of the splitting, returned statistics comparable to the reference model in both regression and classification. These statistics reinforce the strength of the selected descriptors.

4 Discussion

4.1 Adequacy for REACH

The REACH and BPR legislations recommend a precautionary approach in the evaluation of QSAR predictions to reduce the number of FN compounds. In other words, a high sensitivity value for a model is necessary to consider it adequate for the regulatory context.

For this purpose, it is reasonable to omit the classification of the compounds whose predicted log BCF value falls within a safety margin just under each regulatory threshold. In light of this, it is recommended to conduct further studies to determine whether such chemicals could be definitely classified as hazardous (Gissi et al., 2013). Other QSAR models, read across predictions, considerations about stability in water, and metabolism rate shall be taken into account to reduce the uncertainty of the prediction and provide a weight of evidence conclusion. Integrated testing strategies also can be considered.

In this precautionary view, the number of compounds predicted as negative (below the threshold, non-bioaccumulative), but experimentally positive (i.e., FNs), is significantly reduced. Figure 5 displays the statistics adopting the presented precautionary classification for all the thresholds for the QikProp based three- and nine-descriptor models. In particular, the chemicals considered suspicious are counted as unclassified. The safety margin for the predicted values is gradually lowered by 0.1 log units to check the performance improvement and the number of compounds that remain unclassified. The graphs show the increase of the sensitivity for each threshold (y axis) at the increment of the precautionary margin in log units (x axis) for the nine-descriptor and three-descriptor models, respectively.

The plots clearly show that the overall best performance is obtained for the nine-descriptor model. Considering all the relevant regulatory thresholds, the sensitivity values are higher than 60%, irrespective of the safety margin. With a precautionary margin of 0.6 log units, they are greater than 80% for all the thresholds. Considering the three-descriptor model, the B and vB sensitivities are initially equal to 53% and 40%, respectively. The inclusion of an even larger safety margin does not increase the safety to 80%.

Bearing this in mind, the presented models could be employed for regulatory considerations, but for classification a safety margin must be explicitly taken into account (we would suggest a 0.6 offset). Table 3 shows the confusion matrix (Kohavi and Provost, 1998) using the nine-descriptor model as a classifier, with the 0.6 offset as safety margin. As expected, the use of a safety margin leads to a reduction of the number of compounds (e.g., a margin of 0.6 log units implies a loss of 5-18% of the predictions depending on the threshold), which remain unclassified. We want to highlight that the value of this margin of 0.6 is in the range of the experimental variability for the suggested OECD 305 BCF experimental test, which ranges from 0.4 to 0.7 log units (Dimitrov et al., 2005; Lombardo et al., 2010).

For the risk assessment of substances exceeding 100 tons/year under REACH and BPR, the prediction is not limited to an assignment of a class; instead an explicit quantification of the BCF value is mandatory. In this respect, it is worth noting that the RMSE of our best models is again in clear agreement with the above reported experimental variability.

4.2 The impact of novel biokinetics descriptors

Our nine-descriptor model is based on several properties relevant to model BCF. X-scrambling designated the CIQlogS as the descriptor likely playing the strongest influence on BCF (Table S9 in supplementary data at http://www.altex-edition.org). Such a finding is in full agreement with other published BCF solubility-based models (Piir et al., 2010).

For the sake of clarity, we report comments on each descriptor for a possible mechanistic interpretation of the model, as follows. We verified that there is no collinearity among independent variables of the nine-descriptor model.

1. CIQPlogS. The solubility is inversely related to BCF. Highly water-soluble compounds are less likely to accumulate in the lipid portion of fish tissues. However, a minimum value of solubility is necessary for the establishment of an exchange equilibrium between water media and organisms. In fact, only a few chemicals with CIQPlogS < -9 were classified as B in our model.
2. molMW. Another important aspect is the molecular weight (MW) of chemicals. Bulky chemicals are not absorbed through biological membranes. In our model, chemicals (i.e., 15 compounds) with MW values >600 Da have never been classified as B.
3. QPlogHERG. This estimates the IC_{50} values for the blockage of the hERG K^+ channel. There is evidence that such an event could be related to a number of hydrophobic interactions established by the chemicals at the hERG K^+ channel level (Aptula and Cronin, 2004). A connection exists between lipophilicity and BCF (Lombardo et al., 2010).
4. #stars. A high value for #stars flags poor druglikeness. A low value for #stars flags molecules with physiochemical properties similar to those typical of drugs. Drug-like structures are often optimized to be highly bioavailable. In this respect, chemicals with low values for #stars (higher druglikeness) are expected to have high accumulation potential, and consequently greater BCF values. As a result, the #star values are inversely related to the bioconcentration.
5. QPPCaco. The Caco2 cells permeability assay evaluates the permeability of chemicals through biological membranes. A high permeability increases the bioaccumulative potential.
6. WPSA. This describes the weakly polar component of SASA (solvent accessible surface area) and is related to the surface of a chemical. The WPSA value has a positive effect on the BCF.
7. IP(eV). The ionization potential describes the energy needed for the ionization of a chemical: low values are related to a high ionization tendency. Compounds that ionize in aqueous medium do not bioaccumulate, as they cannot cross biological membranes. Furthermore, IP(eV) may have a relationship with the polarization effect of chemicals, and consequently with their reactivity towards the numerous nucleophilic binding sites present in living organisms.
8. #noncon. It is the number of ring atoms not able to form conjugated aromatic systems. The presence of rings generally increases the lipophilicity of a chemical, positively affecting the BCF.
9. #rtFG. Such a descriptor flags the number of reactive functional groups and is related to the likelihood of metabolic transformation. The presence of reactive functional groups facilitates the conjugations between chemicals and hydrophilic endogenous substrates (i.e., phase II xenobiotic biotransformation). This favors the excretion of the molecules and the reduction of the bioconcentration rate. Reactive functional groups often have a polar nature and are thus likely to reduce the BCF.

Tab. 2: Validation parameters of the nine-descriptor model: slopes (k, k'), intercepts (b, b'), and determination coefficients (r^2, r^2_0, r'^2_0) calculated for the TS, VS, and BS for all 836 chemicals and for the 809 chemicals retained in AD

Dataset	r^2	r^2_0	r'^2_0	k	k'	b	b'	$\frac{(r^2 - r^2_0)}{r^2}$	$\frac{(r^2 - r'^2_0)}{r^2}$	$\lvert r^2_0 - r'^2_0 \rvert$
TS	0.731	0.662	0.731	0.913	0.993	0.498	-0.020	0.096	0.001	0.070
VS	0.635	0.590	0.616	0.915	0.970	0.468	0.329	0.071	0.030	0.026
BS	0.623	0.517	0.621	0.879	0.963	0.588	0.075	0.170	0.003	0.104
VS (inside AD)	0.765	0.694	0.764	0.939	0.996	0.552	0.082	0.093	0.002	0.070
BS (inside AD)	0.659	0.488	0.655	0.890	0.966	0.709	-0.154	0.259	0.006	0.167

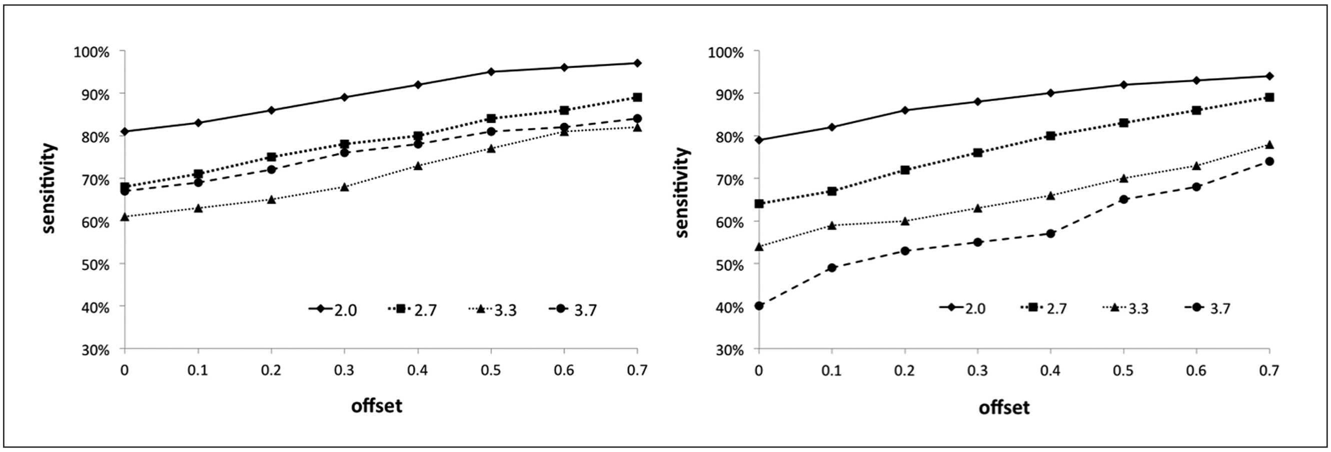


Fig. 5: Improvement in sensitivity for diverse classification thresholds reached by increasing the offset (i.e., the width of the safety margin)
The plot on the left refers to the nine-descriptor BCF model, the plot on the right to the three-descriptor BCF model.

4.3 Chemicals outside the applicability domain

The four independent approaches used for the AD derivation led to the exclusion of a total of 27 chemicals, considering both VS and BS (Fig. 1). As anticipated, 17 of the 27 chemicals were characterized by structural motifs not covered by TS chemical types (see Section 3.2).

Four chemicals were instead identified via the descriptors range approach. Interestingly, three of those compounds (i.e., 536, 604, and 772) have a high MW (i.e., MW >800 Da, which affects the molecular lipophilicity and solubility). Compound 722 is also outside the CIQlogS value range of TS chemicals. This chemical disclosed the largest error in prediction (equal to 4.33 log units). It has been recognized that testing very poorly water soluble substances may not be technically feasible. This often leads to wrong experimental values (OECD, 2012). Moreover, high MW compounds do not easily cross biological membranes. This is the case for compound 604, which is outside the AD for the extremely low value of QPPCaco, indicating a presumably poor permeability. Compound 536 shows an abnormally high value for the descriptor #stars, which would indicate bad druglikeness. That is somehow related to the capability of delivery and subsequent bioaccumulation into body tissues (see Section 4.2). Unlike the chemicals outside the AD discussed above, compound 680 has a low MW. A close structural inspection reveals the presence of a prominent hydrophilic moiety, which can reduce its permeability (i.e., low value of QPPCaco).

The remaining chemicals falling outside the AD were discarded after applying distance-based approaches, such as the polygon approach based on Principal Component Analysis (PCA) and the leverage method. Interestingly, the loadings plot resulting from PCA carried out on the pool of independent variables included in our nine-descriptor model revealed that MW and CIQlogS were the most relevant for the first two components. In this regard, 5 of 14 VS and BS chemicals designated outside AD after either leverage or PCA strategy application (i.e., 479, 496, 536, 604, 772) had MW >600 Da. Such chemicals were also characterized by large errors in prediction (on average equal to 2.343 log units).

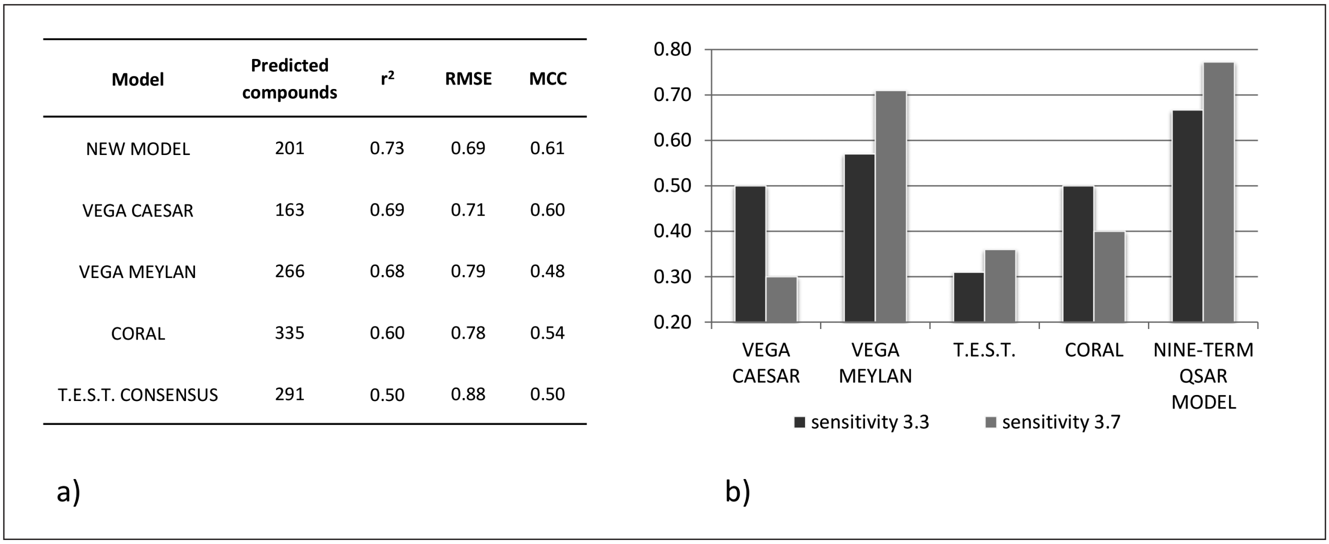
It is noteworthy that even 10 TS chemicals violated the perimeter including 98% of the polygon area or the h* threshold value. In this way we observed that all 15 (i.e., 10 from TS and 5 from VS or BS) chemicals within our dataset having MW >600 exceeded AD limits.

The relevance of MW as a cutoff is even clearer considering that 7 (i.e., 94, 479, 496, 536, 604, 665, 772) of the 10 BS and VS chemicals with MW >500 were outside the AD.

Focusing on the three compounds with MW >500 but inside AD, we noticed for instance that compound 64 (see Table S3 in

Tab. 3: Confusion matrix of the B and vB thresholds with a safety margin set to 0.6 log units
Statistics refer to the QikProp nine-descriptor model. TS, VS, and BS consist of 608, 152, and 76 chemicals, respectively.

	Predicted log BCF											
	TS	nB	B	vB	VS	nB	B	vB	BS	nB	B	vB
Observed log BCF	nB	454	4	7	nB	113	0	2	nB	56	2	0
	B	12	5	10	B	1	1	1	B	2	1	0
	vB	1	8	29	vB	0	1	15	vB	2	1	3



Model	Predicted compounds	r^2	RMSE	MCC
NEW MODEL	201	0.73	0.69	0.61
VEGA CAESAR	163	0.69	0.71	0.60
VEGA MEYLAN	266	0.68	0.79	0.48
CORAL	335	0.60	0.78	0.54
T.E.S.T. CONSENSUS	291	0.50	0.88	0.50

Fig. 6: Comparison between the herein presented nine-descriptor BCF model and others previously developed
The comparison is relative to r^2 and to sensitivity at different thresholds according to REACH (see Section 2.3). For the ease of comparison, statistics of each model refer only to chemicals not comprised in their TS and within their AD.

supplementary data at http://www.altex-edition.org) is very similar to 479. This would support the fact that MW is not the sole descriptor determining whether a compound is outside or inside the AD. Moreover, both 64 and 479 had small errors in prediction (0.335 and 0.169 log units, respectively). This suggests that these compounds are likely to be placed in a boundary zone inside and outside the AD. The exclusion of 479 is mostly due to the precautionary purposes intentionally set when defining AD.

Another important descriptor is CIQlogS. In this respect, compound 506 is among those poorly soluble belonging to VS. This is due to the presence of an extended hydrophobic moiety, similar to that of compound 772. Importantly, the 10 least soluble compounds within our dataset (including TS chemicals) exceed the 98% polygon area or the h* warning threshold.

A number of VS and BS low MW chemicals are outside the AD. Some of them (i.e., 37 and 45) are characterized by a high ratio between the number of halogens and the number of carbons, which is related to poor druglikeness.

Compound 306 is outside AD for its excessive leverage values. This is likely due to the absence of other structural homologues within the ANTARES dataset.

Compound 41 is a low molecular weight aliphatic carboxylic acid, which, despite a very small error in prediction (0.094 log unit), is outside the AD. This is mostly due to the absence of other similar low MW acids within the ANTARES dataset. The dataset instead comprised mostly bulkier aromatic acids. Equivalent considerations can be extended to compound 268, whose exclusion from AD, despite the limited error in prediction (0.223 log unit), is due to the absence of structural homologs. Similarity analyses were carried out by using VEGA software released in 2013. Both 41 and 268 violated the perimeter polygon.

4.4 Comparisons with other BCF QSAR models

At present, a number of trustworthy BCF models already exist which are based on variously sized datasets, as well as on the use of different algorithms. In 2006 and 2008 Pavan et al. made an important contribution by reviewing models for BCF (Pavan et al. 2006, 2008). A milestone study (Piir et al., 2013) based on a consensus model resulted in r^2 equal to 0.79 on TS. Note that machine-learning methods were applied to a set of 713 chemicals using eight descriptors (Strempel et al., 2013) and a value of r^2 of 0.83 on TS was found. In this case, the model was adapted to work also as a classifier, returning an accuracy of 0.99 considering the threshold at 3.3 on the whole set. Fuzzy filtering techniques were applied to a set comprising more than 500 compounds and a value of r^2 equal to 0.73 was obtained for TS (Kumar et al., 2009).

Worthy of mention are the VEGA platform, CORAL, and T.E.S.T. from US EPA. These models provided a wealth of information about the compounds used in the TS to derive the models. VEGA comprises two of the most popular BCF models (i.e., the CAESAR and the Meylan model from EPISuite BCFBAF) and returns a value, in the range 0-1, of the Applicability Domain Index (ADI) to support the reliability of a given prediction. By default, VEGA considers poorly reliable predictions as those having ADI <0.70. A value of r^2 equal to 0.86 on TS was instead obtained using models based on CORAL software (Toropova et al., 2012). T.E.S.T. software returned predictions only for compounds inside its AD.

We have compared the performance of our proposed new model with those of VEGA, CORAL, and T.E.S.T. The TS of these models are publicly available. This allows straightforward comparisons for external predictions and for assessing their reliability in a regulatory context.

As shown in Figure 6a, our proposed new nine-descriptor model discloses the best r^2 on external chemicals among the compared models. The statistics account only for those chemicals within the ANTARES dataset external to the specific model TS, but inside the AD.

As for the classification performance, Figure 6b shows that the sensitivity of our model at both the B and vB thresholds (i.e., log BCF equal to 3.3 and 3.7, respectively) was considerably higher compared to that of previous models.

5 Conclusions

The presented new nine-descriptor model mines a large volume of information, keeping in mind regulatory requirements. Its predicting power in both regression and classification assessed on external compounds makes it suitable for real-life uses. The employed biokinetics descriptors are more helpful than others, as they allow an easier mechanistic interpretation of the obtained results. The reliability of the predictions has been investigated by a well-described multi-step AD analysis. Full elucidation of the biochemical and overall processes is still a work in progress, but the practical validation provides a sound basis for the evaluation of the performance obtained by the proposed model.

References

Aptula, A. and Cronin, M. (2004). Prediction of hERG K+ blocking potency: application of structural knowledge. *SAR QSAR Environ Res 15*, 399-411.

Aptula, A. and Roberts, D. (2006). Mechanistic applicability domains for non-animal based prediction of toxicological end points: General principles and application to reactive toxicity. *Chem Res Toxicol 19*, 1097-1105.

Arnot, J. and Gobas, F. (2006). A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environ Rev 14*, 257-297.

Baldi, P., Brunak, S., Chauvin, Y., et al. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 412-424.

Banks, J., Beard, H., Cao, Y., et al. (2005). Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J Comput Chem 26*, 1752.

Chirico, N. and Gramatica, P. (2011). Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J Chem Inf Model 51*, 2320-2335.

Consonni, V., Ballabio, D., and Todeschini, R. (2009). Comments on the eefinition of the Q(2) parameter for QSAR validation. *J Chem Inf Model 49*, 1669-1678.

Cooper, J., Saracci, R., and Cole, P. (1979). Describing the validity of carcinogen screening tests. *Br J Cancer 39*, 87-89.

Dimitrov, S., Dimitrova, N., Parkerton, T., et al. (2005). Base-line model for identifying the bioaccumulation potential of chemicals. *SAR QSAR Environ Res 16*, 531-554.

Doweyko, A. (2004). 3D-QSAR illusions. *J Comput Aided Mol Des 18*, 587-596.

EC – European Commission (2006). Regulation (EC) No 1907/2006 of The European Parliament and of The Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Off J Eur Union L 396*, 1-849.

EC (2008). Regulation (EC) No 1272/2008 of The European Parliament and of Council of 16 December 2008 on Classification, Labeling and Packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006. *Off J Eur Union, L353*, 1-1979.

EEC – European Economic Community (1967). Council Directive of 27 June 1967 on the approximation of laws, regulation and administrative provisions relating to the classification, packaging and labelling of dangerous substances. *Off J Eur Union L196*, 1-98.

EU – European Union (2012). Regulation (EU) No 528/2012 of The European Parliament and of The Council of 22 May 2012 concerning the making available on the market and use of biocidal products. *Off J Eur Union L 167*, 1-123.

EPI Suite (2013). http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (accessed 23.04.2013).

Eriksson, L., Jaworska, J., Worth, A., et al. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environ Health Perspect 111*, 1361-1375.

Fu, W., Franco, A., and Trapp, S. (2009). Methods for estimating the bioconcentration factor of ionizable organic chemicals. *Environ Toxicol Chem 28*, 1372-1379.

Gissi, A., Nicolotti, O., Carotti, A., et al. (2013). Integration of QSAR models for bioconcentration suitable for REACH. *Sci Total Environ 456-457*, 325-332.

Golbraikh, A. and Tropsha, A. (2002). Beware of q(2)! *J Mol Graph Model 20*, 269-276.

Gramatica, P. (2010). Chemiometric methods and theoretical molecular descriptors in predictive QSAR modeling of the environmental behavior of organic pollutants. In T. Puzyn et al. (eds.), *Recent Advances in QSAR Studies* (327-366). Dordrecht, Heidelberg, London: Springer.

Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Altern Lab Anim 33*, 445-459.

Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning 30*, 271-274.

Kumar, S., Kumar, M., Thurow, K., et al. (2009). Fuzzy filtering for robust bioconcentration factor modeling. *Environ Modell Softw 24*, 44-53.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics 45*, 255-268.

Lombardo, A., Roncaglioni, A., Boriani, E., et al. (2010). Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem Cent J 4, Suppl 1*, S1.

Meylan, W., Howard, P., Boethling, R., et al. (1999). Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environ Toxicol Chem 18*, 664-672.

Minovski, N., Zuperl, S., Drgan, V., et al. (2013). Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study. *Analytica Chimica Acta 759*, 28-42.

Nicolotti, O., Gillet, V., Fleming, P., et al. (2002). Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. *J Med Chem 45*, 5069-5080.

Nicolotti, O. and Carotti, A. (2006). QSAR and QSPR studies of a highly structured physicochemical domain. *J Chem Inf Model 46*, 264-276.

Nicolotti, O., Miscioscia, T., Carotti, A., et al. (2008). An integrated approach to ligand- and structure-based drug design: Development and application to a series of serine protease inhibitors. *J Chem Inf Model 48*, 1211-1226.

Nicolotti, O., Giangreco, I., Miscioscia, T., et al. (2009). Improving quantitative structure-activity relationships through multi-objective optimization. *J Chem Inf Model 49*, 2290-2302.

OECD (2007). Guidance document on the validation of (quantitative) structure-activity relationships [(q)sar] models, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282007%292&doclanguage=en (accessed 29.09.2013).

OECD (2012). Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing. doi:10.1787/9789264185296-en

Pavan, M., Worth, A., and Netzeva, T. (2006). Review of QSAR models for bioconcentration. European Commission, Joint Research Centre, Ispra, Italy, EUR 22327EN.

Pavan, M., Netzeva, T., and Worth, A. (2008). Review of literature-based quantitative structure – Activity relationship models for bioconcentration. *QSAR Comb Sci 27*, 21-31

Piir, G., Sild. S., Roncaglioni, A., et al. (2010). QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects. *SAR QSAR Environ Res 21*, 711-729.

Piir, G., Sild, S., and Maran, U. (2013). Comparative analysis of local and consensus quantitative structure-activity relationship approaches for the prediction of bioconcentration factor. *SAR QSAR Environ Res 24*, 175-199.

Roberts, D., Aptula, A., and Patlewicz, G. (2006). Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the Schiff base applicability domain for skin sensitization. *Chem Res Toxicol 19*, 1228-1233.

Russell, W. M. S. and Burch, R. L. (1959). *The Principles of Humane Experimental Technique*. London, UK: Methuen.

Schultz, T., Hewitt, M., Netzeva, T., et al. (2007). Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb Sci 26*, 238-254.

Strempel, S., Nendza, M., Scheringer, M., et al. (2013). Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals. *Environ Toxicol Chem 32*, 1187-1195.

Todeschini, R. (ed.) (2000). *Handbook of Molecular Descriptors*. Weinheim, Germany: Wiley-VCH.

Toropov, A. and Benfenati, E. (2008). Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: Using rare SMILES attributes to define the applicability domain. *Bioorg Med Chem 16*, 4801-4809.

Toropova, A., Toropov, A., Benfenati, E., et al. (2012). CORAL: Quantitative models for estimating bioconcentration factor of organic compounds. *Chemometr Intell Lab Syst 118*, 70-73.

Tropsha, A., Gramatica, P., and Gombar, V. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci 22*, 69-77.

Weaver, S. and Gleeson, N. (2008). The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model 26*, 1315-1326.

Zhao, C., Boriani, E., Chana, A., et al. (2008). A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere 73*, 1701-1707.

Acknowledgements

We acknowledge the LIFE project CALEIDOS and the University of Bari "Aldo Moro" (Progetto IDEA e Giovani Ricercatori 2011). We acknowledge Ms Sophia E. Albov for her assistance in proofreading.

Correspondence to

Orazio Nicolotti, PhD
Dipartimento di Farmacia – Scienze del Farmaco
Università degli Studi di Bari "Aldo Moro"
Via E. Orabona, 4
70125 Bari
Italy
Phone: +39 080 5442551
Fax: +39 080 5442230
e-mail: orazio.nicolotti@uniba.it