



HAL
open science

QSPR models for bioconcentration factor (BCF): are they able to predict data of industrial interest?

F. Lunghini, G. Marcou, P. Azam, R. Patoux, M.H. H Enrici, F. Bonachera,
Dragos Horvath, A. Varnek

► To cite this version:

F. Lunghini, G. Marcou, P. Azam, R. Patoux, M.H. H Enrici, et al.. QSPR models for bioconcentration factor (BCF): are they able to predict data of industrial interest?. SAR and QSAR in Environmental Research, Taylor & Francis, 2019, 30 (7), pp.507-524. 10.1080/1062936X.2019.1626278 . hal-02950596

HAL Id: hal-02950596

<https://hal.archives-ouvertes.fr/hal-02950596>

Submitted on 16 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QSPR models for bioconcentration factor (BCF): are they able to predict data of industrial interest?

F. Lunghini^{a,b}, G. Marcou^a, P. Azam^b, R. Patoux^b, M.H. Enrici^b, F. Bonachera^a, D. Horvath^a and A. Varnek^a

^aLaboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France; ^bSolvay S.A., France

ABSTRACT

The bioconcentration factor (BCF), a key parameter required by the REACH regulation, estimates the tendency for a xenobiotic to concentrate inside living organisms. In silico methods can be valid alternatives to costly data measurements. However, in the industrial context, these theoretical approaches may fail to predict BCF with reasonable accuracy. We analyzed whether models built on public data only have adequate performances when challenged to predict industrial compounds. A new set of 1129 compounds has been collected by merging publicly available datasets. Generative Topographic Mapping was employed to compare this chemical space with a set of new compounds issued from the industry. Some new chemotypes absent in the training set (such as siloxanes) have been detected. A new BCF model has been built using ISIDA (In Silico design and Data Analysis) fragment descriptors, support vector regression and random forest machine-learning methods. It has been externally validated on: (i) collected data from the literature and (ii) industrial data. The latter also served as benchmark for the freely available tools VEGA, EPISuite, TEST, OPERA. New model performs (RMSE of 0.58 log BCF units) comparably to existing ones but benefits of an extended applicability, covering the industrial set chemical space (78% data coverage).

ARTICLE HISTORY

Received 9 April 2019
Accepted 29 May 2019

KEYWORDS

QSAR/QSPR; generative topographic mapping (GTM); bioconcentration factor; REACH; benchmarking

Introduction

In environmental risk assessment, the bioconcentration factor (BCF) is a key parameter to be considered. It estimates the tendency for a xenobiotic to concentrate inside living organisms and it is defined as the process of concentration of the chemical from the water phase through non-dietary routes, such as absorption from respiratory surfaces (e.g. lungs/gills) or skin. Xenobiotics' concentration inside organisms can thus reach hazardous levels, with long-term deleterious effects, such as modified behaviours, impacts on reproduction, which in the end may lead to endanger some species [1]. Organisms at the upper of the food-chain (e.g. fishes) are particularly in danger and, as a direct consequence of their consumption, man might be the ultimate impacted species. BCF is defined as the ratio of the steady state concentration of the chemical

CONTACT G. Marcou ✉ g.marcou@unistra.fr; A. Varnek varnek@unistra.fr

Supplemental data for this article can be accessed at: <https://doi.org/10.1080/1062936X.2019.1626278>.

in aquatic organisms (such as fish, mussels, algae, etc.) and the corresponding freely dissolved chemical concentration in the surrounding water media (Equation (1)) [2].

$$\text{BCF} \approx \frac{C_f}{C_w} \quad (1)$$

Where C_f and C_w are the concentrations at steady state of the chemical inside the fish and the water media, expressed in mg/Kg and mg/L, respectively. The duration of the uptake phase is usually 28 days, however it can be lengthened if necessary, or shortened if the steady-state has been reached earlier [3]. BCF is expressed in L/Kg. Typically the fish is used as test model due to its importance in the food web and the availability of standardized guidelines.

The determination of BCF is a key requirement for regulatory frameworks such as the European Union Registration, Evaluation, Authorisation and Restriction of Chemical Substances Regulation (REACH, EC No 1907/2006) for the PBT/vPvB (Persistent Bioaccumulative and Toxic/very Persistent very Bioaccumulative) substances assessment. In Europe, there are two relevant bioconcentration thresholds which will usually determine if a substance fulfills the 'bioaccumulative' criterion or the 'very bioaccumulative' criterion. The former is set at a BCF value of 2000 L/Kg (or 3.3 log unit), while the latter is set at 5000 L/Kg (or 3.7 log unit). Below 2000 L/Kg, a substance is not considered to possess a significant bioaccumulation potential [4]. Due to the expensive nature of BCF experiments and the high number of required animals, the use of *in silico* methods is encouraged [5].

During the past decades, empirical predictors have been proposed to estimate the BCF, which are mainly based on the octanol-water partition coefficient (log P) alone [6–9], as it is a key-determining factor linked to this property. More recently, other types of molecular descriptors have been employed [8,10,11], and many QSAR models are nowadays implemented in commercial or freely-available software, such as VEGA (Virtual models for property Evaluation of chemicals within a Global Architecture) [12], Toxicity Estimation Software Tool (TEST) [13], Estimation Program Interface (EPI Suite) [14], OPERA (OPEn (q)saR App) [11], Chemical Properties Estimation Software System (ChemProp) [15], CORAL [16], ACD/log D Suite [17] and OASIS-Catalogic [18]. Table 1 summarizes other authors' evaluations of the models considered in the present study. The number of publications is quite high, and performances can be very different, with RMSE (Root Mean Square Error) values reaching almost one unit of difference for the same model. This depends on the type of chemical families, but also on the user choices about the Applicability Domain (AD) thresholds (since for some of the tools the AD is not clearly defined), and the exclusion of compounds already present in the model's training set. The work of Petoumenou et al. [19], is the only one to evaluate data coming from the industrial context, i.e. extracted from the European Chemical Agency (ECHA) database [20]. These results are of particular interest because: (i) during the REACH registration, the available data was reviewed by the industries before submission and, eventually, new data was generated to comply with endpoint requirements; (ii) this database could potentially be more representative of the chemical families of industrial interest. To our knowledge, this study is unique of its kind. Yet, only a small subset of ECHA was used and there is no consideration of overlaps between the test set and the training sets of the benchmarked tools. In addition, most of the abovementioned tools queried the

Table 1. Overview of the existing tools considered for benchmarking.

Model	General information	Model performance			
		Compounds	r ²	RMSE	Reference
VEGA Caesar	Tr. set = 473	95	0.78	0.62	[12] ^{MD}
	Descriptors = 2D phys-chem descriptors	30	0.85	0.58	[21]
	Algorithm = Radial basis function neural network (RBFNN)	538	-	0.91	[22]
		45	-	1.57	[22]
		78	0.8	0.46	[19]
VEGA KNN	Tr. set = 832	162	-	1.33	[23]
	Descriptors = 2D phys-chem descriptors	152	-	0.81	[12] ^{MD}
	Algorithm = k-Nearest neighbours (kNN)	45	-	0.91	[23]
		95	0.78	0.47	[19]
		98	-	0.66	[23]
VEGA Meylan	Tr. set = 662	146	0.79	0.66	[12] ^{MD}
	Descriptors = 2D phys-chem descriptors	32	0.64	0.87	[21]
	Algorithm = Linear regression	349	-	0.99	[22]
		45	-	0.99	[22]
		76	0.78	0.43	[19]
TEST	Tr. set = 589	97	-	0.64	[23]
	Descriptors = CDK descriptors ^a	-	0.76	0.66	[13] ^{MD}
	Algorithm = consensus between algorithms	291	0.5	0.88	[21]
EPISuite	Tr. set = 527	527	0.83	0.50 ^a	[14] ^{MD}
	Descriptors = log P, functional groups	158	0.82	0.59 ^a	[14] ^{MD}
	Algorithm = Linear regression (log P-based with functional groups as correction factors)	432	0.59	0.87	[21]
		349	-	0.94	[22]
		45	-	1.33	[22]
OPERA	Tr. set = 685	145	0.45	0.89	[19]
	Descriptors = PaDEL Descriptors ^b Algorithm = k-Nearest neighbours (kNN)	157	0.83	0.64	[11]

^{MD} information has been taken from the model's documentation manual. ^aChemistry Development Kit (CDK) descriptors [24]. ^bPaDEL-Descriptors software [25].

same sources of data for training set collection [11–14]. This may limit their applicability when confronted to chemotypes of industrial interest which are new or under-sampled in the public data.

In this study, we analyzed whether models built on public data only show satisfactory performances when challenged to predict a set of compounds extracted from Solvay's portfolio ('industrial set'). We aimed at getting a more precise picture of the performances of publicly available models. We observed that the performances in this industrial context could decrease, and we hypothesized that the applicability domain of these tools did not match sufficiently our industrial set. As a consequence, we tried to collect the most comprehensive BCF training set by merging several publicly available datasets, used to generate a new BCF-model ('ISIDA Consensus'). ISIDA Consensus was then externally validated on the industrial set's compounds and benchmarked against the already existing tools (Table 1).

The Office of Economic Cooperation and Development (OECD) principles [26] for building robust QSAR models were followed. The five OECD principles are: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined applicability domain; (iv) appropriate measures for goodness-of-fit, robustness, and predictivity; (v) and a mechanistic interpretation, if possible. In this study, the endpoint (BCF) is well defined, Goodness-of-fit, robustness and predictivity were evaluated using internal 3-fold Cross-Validation (CV)

and against two external test sets. The AD of the models was defined using two complementary methodologies.

Our developed model is available as a web-application, called 'ISIDA Predictor' [27], available at the Laboratory of Chemoinformatics webpage: <http://infochim.u-strasbg.fr>.

Methods

The general workflow is shown in Figure 1. Its main steps will be detailed in the present study.

Data collection and curation

Bioconcentration experimental data was collected from multiple sources, including several public-available databases and literature research. Mined databases comprised: the Japanese National Institute of Technology and Evaluation (NITE) [28], the European Chemical Industry Council Long Range Initiative (CEFIC LRI) [29], the Canadian Domestic Substance List (DSL) [30] and the ECOTOXicology knowledgebase of the US Environmental Protection Agency (ECOTOX EPA) [31] (accessed through the OECD Toolbox [32]), and the database of ECHA (accessed through the eChem portal [33]). Additional values were retrieved from literature from the works of Arnot and Gobas [6], Dimitrov et al. [34] and Fu et al. [35]. Finally, a BCF dataset was provided by Solvay. Table 2 reports statistics for the given database. Detailed analysis of the populating chemotypes is given in the dedicated Generative Topographic Maps (GTM) paragraph in the results section. Training and test set public data are available in the SI; the industrial set compounds cannot be provided due to confidential data.

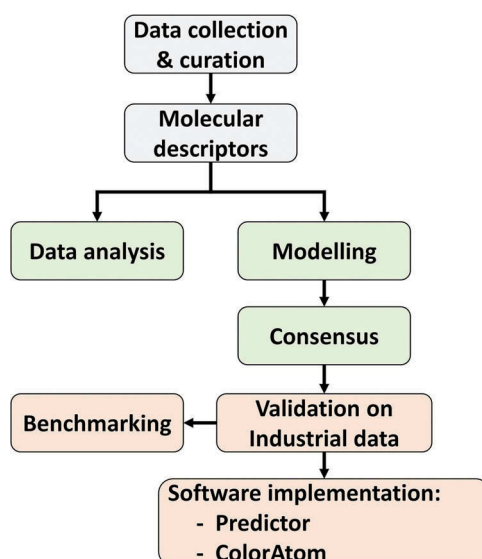


Figure 1. General workflow.

Table 2. Sources of BCF data. The upper portion of the table is referring to the curated dataset before their merging, while the bottom part reports the number of compounds that constituted to the training and the external test sets.

	Nb of compounds	log BCF range (min/max)
Database		
NITE	268	-1.0/4.4
CEFIC LRI	521	-0.8/5.3
Canadian	470	-0.7/5.8
ECOTOX	470	-2.1/5.5
ECHA chem	145	-1.0/4.3
Literature	993	-1.7/6.0
Industrial set	72 ^a	-1.1/4.9
Curated dataset		
Training set	1129	-1.0/6.0
External set	204	-1.7/5.9
Industrial set	31 ^a	-0.1/3.1

^a the number is reduced since a portion of the industrial set was already comprised in the training set.

The following entries were excluded: inorganic, polymer, Unknown or Variable composition, Complex reaction products or Biological materials (UVCBs) compounds. Furthermore, when the BCF value was not reported in L/Kg of body weight, not calculated on a whole-body measurement-basis or the test was performed on a non-recommended OECD species, the value was excluded. Since these are important study conditions that have to be explicitly stated [3], entries which were missing such details were excluded as considered of lower reliability. Chemical structures were standardized (Supplementary information, section 1.1) and duplicates were removed. When multiple data points were available, the median was taken as representative value. The median was computed according to the recommendation of the norm ISO16269-7. The median is the value at middle rank of the ordered set of observations if the set size is odd. If the set size is even, it is the arithmetic average of the two middle ranked values of the ordered set. Notice that for some substances the range of BCF values could reach two log units (Supplementary information, section 1)

Generative topographic mapping

Data visualization approaches are powerful tools that allow us to reduce a high-dimensional space to two or three dimensions which can be then more easily analyzed. For previously published BCF models, visualization techniques (e.g. Principal Component Analysis, PCA) were mainly employed as methods for defining the AD of the model [36–38], but were less often used to characterize in greater details the model's training set composition. Herein, we employed Generative Topographic Mapping, a non-linear mapping method [39]. As advantage, it introduces a probability density function for data distribution, which allows to assess the robustness of the information contained in the generated maps [39,40]. The outcome of GTM is a 2D map on which the analyzed chemical space is projected. A data property can be added as a 3rd axis forming such called activity (property) landscape. Each landscape position is coloured according to the property value (either continuous or categorical); this value is the average property of the

data subset concerned by that position on the landscape. A more detailed description of GTM underlying algorithms can be found elsewhere [39–42]. The 2D generative topographic maps were generated with ISIDA/GTM tool [27] using ISIDA descriptors selected for the best SVM model.

Encoding of chemical structures

ISIDA Property-Labelled Fragment [43] descriptors were employed. There are several types of ISIDA descriptors: (i) sequences of connected atoms and bonds, or atoms only or bonds only, (ii) ‘augmented’ atoms representing either a given atom with its close environment or selected groups of atoms and bonds, and (iii) atom triplets [44]. This led to the generation of several dozens of different descriptor spaces corresponding to different fragment sizes and topologies [45].

Model generation and validation

Support vector machine (SVM) with linear and radial basis function (RBF) kernels and random forest (RF) machine learning approaches were implemented. SVM models were generated with libSVM (v. 3.22) [46]; instead, WEKA (v. 3.9.1) [47] was used for RF models. The SVM parameters (Cost and Gamma) corresponding to minimal RMSE in 3-fold cross-validation were found in genetic algorithm driven optimization. The RMSE was estimated using a dedicated 3-fold CV, isolated from the cross-validation procedure used to evaluate the final models, mentioned below. Concerning RF, default parameters of WEKA were selected, with the number of generated trees equal to 150.

Figure 2 depicts the modelling workflow: (1) dozens of ISIDA Descriptor Spaces (DS) were generated (different fragment sizes and topologies); (2) for each DS, SVM and RF models were generated (individual models); (3) individual models were ranked according to their RMSE in 3-fold CV; (4) the best performing individual model for the given DS was retained; (5) SVM models (linear kernel) were analyzed in consensus to detect the outliers; and (6) ‘final models’ were re-built.

Each individual model was evaluated in 3-fold CV by random splitting. This procedure was repeated 5 times after reshuffling. Thus, BCF for each molecule was predicted 5 times. The r^2 and RMSE values were assessed for each repetition followed

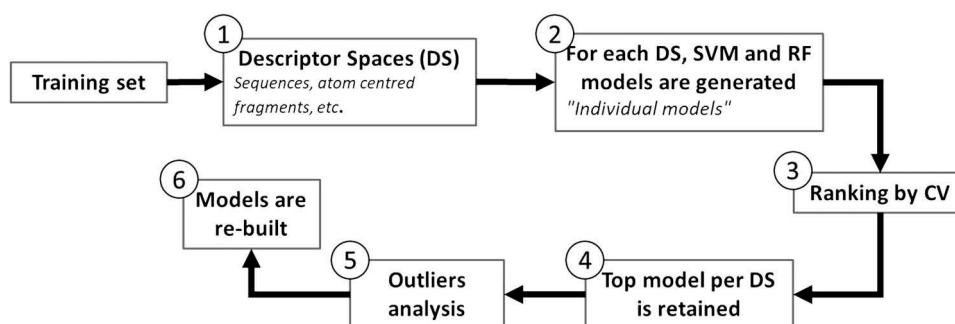


Figure 2. Modelling workflow.

by their averaging (see Table 4). During CV, no optimization of method parameters was performed. The absence of chance correlation was checked through the Y-scrambling procedure [48]. In this procedure, the log BCF values are randomly assigned to molecular structures followed by the model building. This procedure was repeated 150 times.

For the outliers, compounds consensually characterized by very high fitting errors (i.e. difference between experimental and fitted value) were ranked by the Errorscore; $\frac{1}{k} \sum_{k=1}^k \varepsilon_{i,k}$; where $\varepsilon_{i,k}$ is an absolute value of prediction error of k-th model for compound i.

Compounds with the highest scores were poorly predicted by most of the individual models. For some of poorly predicted molecules we discovered that their experimental BCF values were very different from those of their closest analogues in the training set. Unfortunately, due to missing references in databases used, we were not able to retrieve detailed information about BCF measurements for these molecules. Therefore, by precaution, we excluded the 34 compounds from the training set, which corresponded to some 3% of the initial training set (see the list of excluded compounds in SI). Thus, the final training set consisted of 1095 molecules.

The analysis of model performance relies on the r^2 determination coefficient and the RMSE parameters (Equations (2) and (3) respectively).

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

Where y_i is the experimental value of the i-th chemical; \hat{y}_i is the predicted value of the i-th chemical; y_{avg} is the mean of the experimental values of the compounds in the dataset and n is the number of compounds in the dataset.

Ensemble modelling and applicability domain

Individual models served to generate the global ISIDA consensus, and the final result is given by the calculation of the median across all the models, excluding out of AD predictions. The AD was evaluated based on a fragment control assessment: if a test molecule is found to have one fragment (i.e. a determined sequence of atoms and/or bonds) which is not present in the individual model, that molecule is marked to be outside the AD. The number of fragments involved in given individual model depends on selected fragmentation scheme. It varies from 300 (atom centred fragments with radius 1) to 5917 (sequences of atoms and bonds up to 8 atoms length), with an average of 2157. In the consensus calculation, those compounds that are predicted by less than 25% of the total generated models, are considered out of AD. Furthermore, a second assessment based on the Median Absolute Deviation (MAD) was implemented. This can be interpreted as a convergence degree: the lower the MAD, the more the models are in agreement, increasing the overall confidence of the predicted value. It was decided to

set a cut-off value for the MAD equals to 0.5: predictions above this threshold was considered of lower quality and marked as out of AD.

Predictions graphical interpretation: ColorAtom

A related utility of the ISIDA Predictor online platform [27] is the 'ColorAtom' [49]: this tool assigns a colour to each atom of the predicted molecule depending on how much, from a mathematical point of view, it contributed to the property value, either by increasing or decreasing it. The assigned colours are not meant to reflect how the given structural features are correlated to the modelled property in reality; more precisely, it is a graphical representation of how the model interpreted the molecule for calculating the predicted value. To make a comparison, this approach could be compared to the fragment constant (or group contributions) models [7], which associate numerical quantities to a specific substructure of the molecule (single atoms, functional groups, etc.) that are subsequently arithmetically added. Here, two examples of this application are reported: (i) comparison of excluded outliers to structurally analogue compounds in order to highlight the specific groups at the root of the observed differences; (ii) identification of putative chemotypes that may be associated to specific BCF value ranges.

Benchmarking on industrial data

Predictive performance of the ISIDA Consensus model on the industrial set of 72 compounds was compared with that of publicly available tools VEGA (Caesar, Knn and Meylan), TEST, EPISuite and OPERA tools [11–14]. Since industrial set and related training sets were partially overlapped, only non-overlapping compounds from the industrial set were considered for assessing the models' performance (r^2 and RMSE). Moreover, the molecules outside of applicability domain of a given model were discarded (Supplementary information, Section 4).

We also made several pairwise comparisons of ISIDA Consensus with other tools. Each pairwise benchmarking was performed on the part of the industrial set which didn't overlap with the two related training sets. Unfortunately, a common subset for all tools satisfying the above condition was too small for obtaining meaningful statistics.

Results

Overview of the curated dataset

At the end of the data cleaning procedure the number of compounds with unique BCF value was reduced to 1333. Of them, 1129 unique compounds were identified as coming from verified sources and constituted the training set; while, 204 compounds were considered as of lower reliability since there was not enough information to assess their quality (e.g. only CAS and experimental value was provided with no other stated information) and were excluded from the training set. These compounds were used in external validation (i.e. the 'External set'). The Industrial set followed the same data

curation procedure, and a total of 72 compounds were retained. Statistics of the curated datasets are reported in Table 2.

GTM: industrial set visualization and description

Figure 3 shows the GTM log BCF property landscape of the training set onto which the molecules from the Industrial dataset have been projected (represented by black dots); some examples are provided in Table 4 and Supplementary information, section 2. Here, all the 72 compounds were projected. In addition, the associated property-landscape helps characterizing the molecules' BCF profile.

Relevant areas populated by the industrial compounds are marked by numbered boxes. Examples of are showed in Table 3.

- Region 1 is very heterogeneous, including as diverse species as biphenyl derivatives, fluorinated compounds and aliphatic hydrocarbons. Some examples of herein residing unique industrial set chemotypes are: (i) long chain N-alkyl acetamides (CAS 111-57-9, 149879-98-1); (ii) aliphatic aliphatic polyphosphonic acid (CAS 2235-43-0, 29329-71-3); (iii) substituted phosphine (CAS 603-35-0); (iv) fluorinated

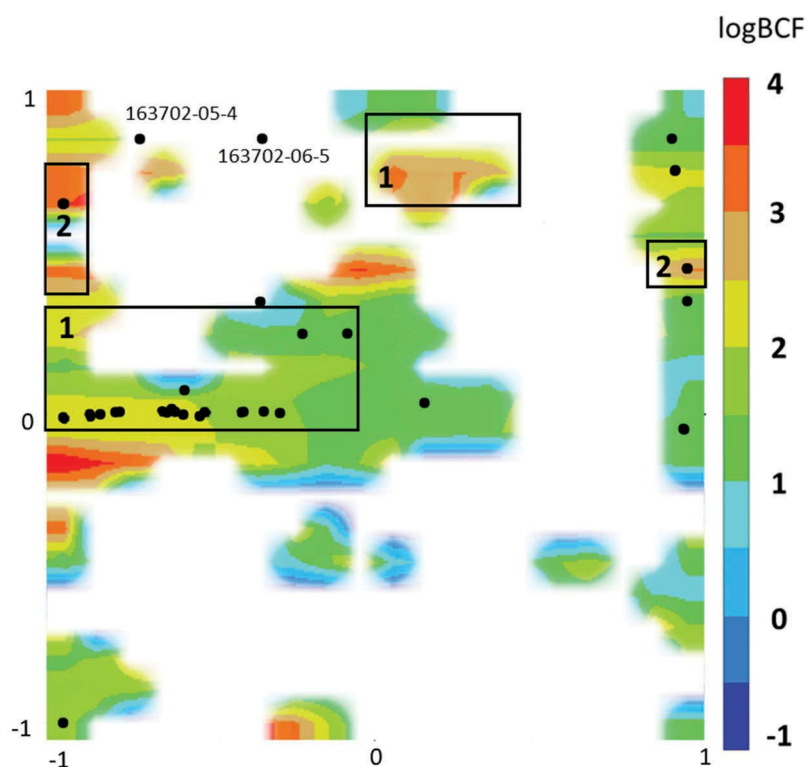


Figure 3. Log BCF property landscape of the training set. GTM is representing the density-modulated log BCF-landscape derived from training set compounds, onto which the industrial set compounds have been projected (black dots). White areas are empty regions of the map. Numbered boxes identify map regions of interest, subject to discussion.

Table 3. Example of compounds populating a given region, as represented by the GTM map (Figure 3). For each region, some molecules are given as examples. Below the molecule, its CAS no. and its experimental value are reported, respectively.

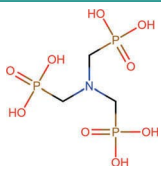
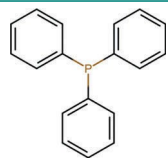
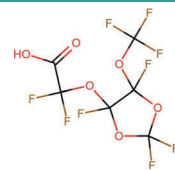
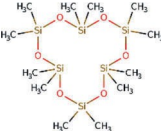
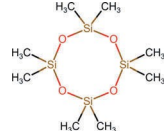
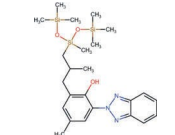
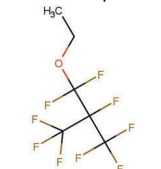
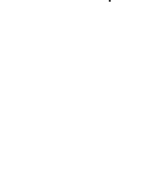
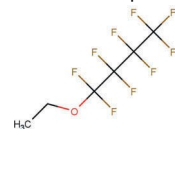


Region	Molecular structure		
1			
2	2235-43-0 1.34 	603-35-0 1.47 	1190931-27-1 0.44 
Molecules on white area	540-97-6 4.01 	556-67-2 4.09 	155633-54-8 0.93 
	163702-06-5 2.96 		163702-05-4 2.96 

Table 4. Summary of model statistics in cross-validation and for the external set. For the external set, performances were evaluated with and without out-of-AD compounds. Results are reported for each machine-learning method separately and for the consensus model. In brackets, the standard deviation computed in the 3-fold CV is reported for the r^2 and RMSE values averaged over the number of repetitions.

Model algorithm	3-fold CV			External set			
	r^2	RMSE	Y-scrb highest r^2	All compounds		Inside AD-only	
				r^2	RMSE	r^2	RMSE
SVM (Linear)	0.72 (0.068)	0.78 (0.044)	0.043	0.66	0.92	0.64	0.86
SVM (RBF)	0.75 (0.039)	0.68 (0.029)	0.042	0.77	0.76	0.75	0.71
RF	0.74 (0.038)	0.68 (0.041)	0.170	0.73	0.82	0.74	0.72
ISIDA consensus	0.75 (0.043)	0.71 (0.051)	-	0.76	0.77	0.75	0.72

sulfonamides (CAS 90076-65-6); (v) branched halogenated compounds with esters and ethers groups (CAS 642461-49-2, 1190931-27-1).

- Regions 2 include mainly silicon-containing compounds (e.g. CAS 540-97-6, 556-67-2 and 155633-54-8). Since average BCF values in these areas are high, these compounds can be potentially considered of concern.

Notice that the abovementioned compounds are absent from the training set of all studied models and contain new chemotypes which are under-sampled in the public data.

Finally, the two labelled molecules falling into the white area should be considered. These compounds (CAS 163702-06-5, 163702-05-4 respectively) have a multimodal responsibility pattern, partially residing into several disparate nodes which are

populated by analogous training set compounds. Their (X,Y) position on the map marks the barycenter of their responsibility pattern (Supplementary information, section 2).

Descriptor selection and model fitting

Table 4 summarizes the performances on training, cross-validation and on the external set for each employed algorithm and the ISIDA consensus model.

Multiple BCF values reported for some compounds were used to estimate experimental errors of BCF measurement. For each compound with at least 2 data points, a BCF range (maximum – minimum over reported values) was calculated, and the average of these range widths over concerned compounds was interpreted as experimental error. Estimated in such a way experimental variability was ± 0.61 log units, which is not too far from the value of ± 0.75 log unit reported by the work of Dimitrov et al. [34] for another BCF dataset. This experimental error is in line with the RMSE calculated in cross-validation in this work (0.71).

After the Y-scrambling procedure, shuffled models were characterized by very low determination coefficient values in cross-validation. The only exception could be random forest, since it exhibits a significantly higher r^2 compared to the other methodologies. Nevertheless, it is still much lower than the lowest r^2 of all random forest models (0.170 vs 0.697). A decrease of performances (r^2 and RMSE) in cross-validation versus the external set can be noticed, however the statistics remain comparable.

Performances on the industrial set

Table 5 reports the results on the industrial set for all the evaluated tools. Two ‘scenarios’ can be identified: (i) all the three VEGA models perform slightly better than ISIDA Consensus but, at the same time, their applicability domain is very narrow; (ii) OPERA and EPISuite have comparable or even higher coverage than ISIDA, but their accuracy is much worse. ISIDA Consensus may not be the best model in terms of precision (higher RMSE of 0.58, compared to the best VEGA model of 0.44) but, at the same time, has a much larger data coverage (ISIDA 78% vs VEGA 19%). Thus, ISIDA Consensus has an extended AD, comparable to TEST, OPERA and EPISuite, while preserving a much lower RMSE.

Table 5. Performances of the models on the industrial set.

Model	% ofAD Coverage ^a	r^2_{det}	RMSE
ISIDA Consensus	78 (25/31)	0.55	0.58
VEGA Caesar	16 (8/49)	0.70	0.58
VEGA Knn	37 (16/43)	0.74	0.50
VEGA Meylan	19 (9/47)	0.47	0.44
TEST	79 (37/47)	0.49	0.86
EPISuite	98 (45/46)	0.34	0.98
OPERA	75 (37/49)	0.40	0.91

^athe first number is the data coverage in %; the number between the parentheses is the ratio of the number of compounds inside AD and the total number of compounds.

Concerning performances on the mentioned unique chemotypes (Table 3): (i) siloxanes fell outside the AD of all the models except for VEGA Knn and TEST. However, even for the latter their prediction is error-fraught because the VEGA training set contains only one siloxane and the AD definition of TEST is very permissive; (ii) all the models failed to predict phosphonate compounds due to AD limitations; (iii) ISIDA Consensus was the only model that scored good performances on the chemotypes exemplified in Table 3.

Figure 4 shows the 'ISIDA Consensus-predicted vs experimental' scatter plot for the 31 'Industrial' compounds not used for training. Overall, predictions are well correlated to experimental values with the exception of one outlier, out of the AD (red point). Based on the % of accepted according to AD individual models, a 'traffic-light' prediction confidence score has been assigned. Three levels were defined: <25%, between 25 and 70% and >70%. They correspond to 'low (out-AD)', 'moderate' and 'high confidence', respectively.

Table 6 reports the results of the pairwise comparison between ISIDA Consensus versus all the other tools individually. With this evaluation, only predictions for compounds not in the training set, inside the AD and predicted by both tools were compared. In this case, ISIDA Consensus always shows a better accuracy except when compared with VEGA KNN (0.55 vs 0.45 of RMSE, respectively).

In the case of VEGA Caesar and VEGA Meylan, the number of compounds in common was too limited to provide a meaningful statistical evaluation and the comparison was not performed.

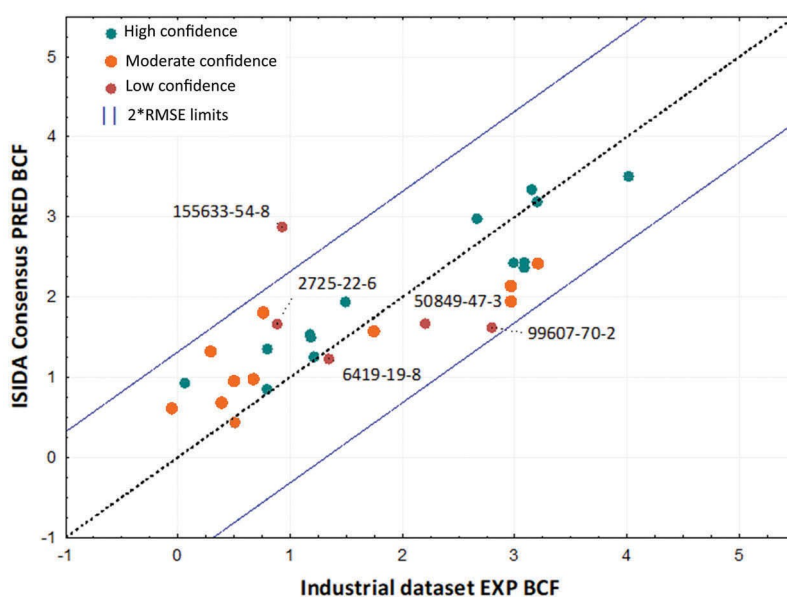


Figure 4. ISIDA consensus predicted vs industrial set experimental values. The data points labels correspond to the CAS numbers of chemicals. Red, orange and green dots = prediction confidence score based on the % in-AD models (<25%; 25–70% and >70%, respectively); Blue lines indicate $\pm 2 \times \text{RMSE}$ value given by 3-fold cross-validation.

Table 6. Pairwise comparison for overlapping compounds between ISIDA consensus vs the given tool. Comparisons against VEGA Caesar and VEGA Meylan were not considered due to the very limited number of overlapping compounds (4 and 3 respectively), which led to unmeaningful statistics.

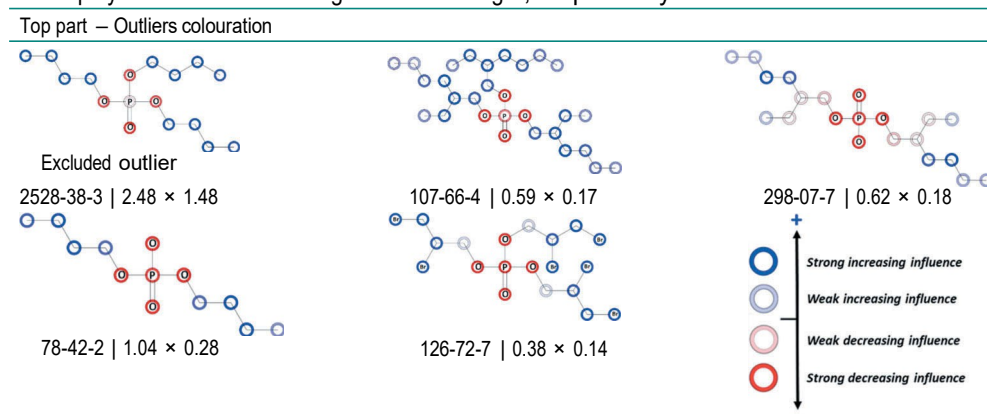
Pairwise comparison between:	Compounds in common	r^2_{det}	RMSE
ISIDA Consensus vs VEGA KNN	9	0.77 0.84	0.55 0.45
ISIDA Consensus vs TEST	19	0.73 0.57	0.63 0.80
ISIDA Consensus vs EPISuite	23	0.78 0.45	0.59 0.92
ISIDA Consensus vs OPERA	18	0.72 0.68	0.63 0.67

ColorAtom: Graphical representations

Table 7 reports one example of BCF atom contribution-coloured outlier (CAS 2528-38-3; with an absolute error of 1.0 log BCF) by contrast to similar but not mispredicted compounds. Molecules showed the same colouration pattern, with the phosphate group and the aliphatic residue being correlated to a decrease and increase of the BCF, respectively. Same colouration scheme means that the molecule was predicted using the same learned rules. However, albeit the compared species appear to be similar according to the employed ISIDA atom fragmentation scheme, the chemist will observe that the outlier, an ester, is a neutral species whilst the counterexamples have one ionizable -OH left and will be anionic species at neutral pH. Note that ISIDA fragmentation schemes using pharmacophore typing [45] are able to make this difference, but were not employed in this study. Additional examples are provided in Supplementary information, section 3.

Figure 5 shows the ColorAtom graph for phosmet (CAS 732-11-6). In this example, all the carbons of this molecule (also S and P, but to a lesser extent) positively contribute to BCF, oxygens and nitrogen are strongly correlated to a decrease of BCF values. Such a colouration pattern can also be found in other training set molecules, where these

Table 7. ColorAtom output. Example of excluded outlier with its most structurally similar compounds (based on Tanimoto score) with the respective experimental and predicted BCF. The colouration is directly referred to the modelled property (i.e. the log BCF value): blue and red atoms played a role in increasing and decreasing it, respectively.



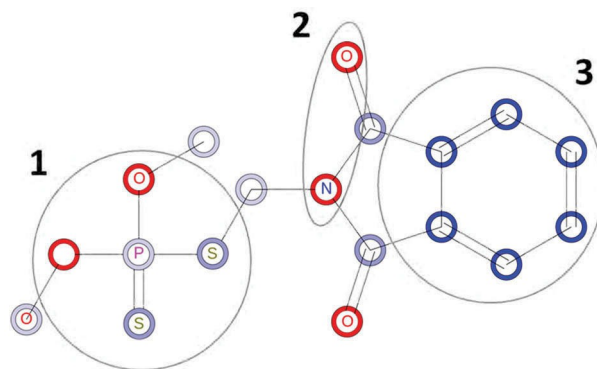


Figure 5. ColorAtom graph of phosmet. The colour scale is reported in Table 7. Numbered ellipses mark some recurring chemotypes subjected to discussion.

chemotypes (in particular, substructures no. 1, 2 and 3, as marked by black ellipses) are systematically following the same trend. Section 3 of Supplementary information reports several examples of compounds containing the mentioned structural features. Molecules in which chemotypes no. 1 and 2 are representing significant substructures are associated to lower BCF values (e.g. CAS no. 60-51-5, 2497-06-5 and 85-41-6); while the opposite happens for chemotype no. 3 (e.g. CAS no. 84-65-1, 829-26-5 and 40766-31-2). This is consistent with the more general trend between increasing hydrophobicity and bioaccumulation: the former is generally increased by aromatic rings [7].

Discussion

Applied models showed mixed performances on the industrial set. As a general trend, most accurate models have narrow data coverage (VEGA), while models with a more permissive AD had higher RMSE values (EPISuite and TEST). ISIDA Consensus was the only model that managed to obtain a good balanced between accuracy and data coverage, especially on unique chemotypes (Table 3), suggesting that its training set is more heterogeneous and diversified compared to the other tools. As a common flaw, all models failed to predict siloxanes and phosphonate compounds, either due to AD limitations or prediction accuracy. The presence in our collected training set of some silicon-containing molecules was not enough to support extension the AD to other siloxanes. Furthermore, current methods have some difficulties in measuring and interpreting the bioaccumulation property for siloxanes [50]. The compound drometrizole trisiloxane (CAS 155633-54-8; Figure 4) can be taken as example, as it showed a prediction error of almost 2 log units. This molecule is structurally similar to drometrizole (CAS 2440-22-4) and octamethyltrisiloxane (CAS 107-51-7), both of which are substructures of the former. These two compounds are present in the training set with experimental BCF values of 2.47 and 3.73 log units, respectively. Thus, the models learned to correlate these specific sequences of fragments to the respective experimental values, and drometrizole trisiloxane prediction is in the range of these two chemicals (2.86 log units). On the other hand, the experimental value reported in the REACH dossier (EC no. 422-940-4) is much lower (0.93 log units).

ColorAtom can be used as a supporting tool to interpret the model output (OECD principle #5): it was employed here to identify key structural features which were recursively correlated to the same alteration trend the property BCF.

As a novelty, (i) molecules were encoded with ISIDA Fragments, a type of descriptors never used to model this property; (ii) different machine learning algorithms were employed (i.e. support vector machine and random forest), in contrast with most of the already existing tools (Table 1). With the benchmarking, ISIDA Consensus proved to possess several strong-points, such as a bigger training set, a wider AD coverage and good accuracy (Table 6) when compared to the other models. As several structural features were identified as unique to the industrial set; model performances will benefit from the addition of such compounds, thanks to an extended AD.

Conclusions

In this work we developed a new ISIDA Consensus QSAR model for the bioconcentration factor property (BCF). The model follows the OECD principles [26] and has been internally and externally validated on two independent test sets, one of which contains relevant chemical families of the industrial context. Models showed mixed performances on the industrial compounds. Tools with the highest accuracy are associated to a very narrow AD; while models with more permissive AD had much worse RMSE. Our model scored the same accuracy (RMSE of 0.58 log BCF unit) of the most acute tool and preserved a much larger AD (78% data coverage). However, as a general limitation all models failed to predict some chemical families, such as siloxanes and highly phosphonate compounds: these are unique industrial set chemotypes which are under-sampled in the public data. In order to compensate the individual-model limitations, the use of all the available tools in consensus is encouraged to reduce uncertainty and improve the accuracy.

Comparing the performances of ISIDA Consensus with the ones from Table 1, it is possible to conclude that our findings corroborate those of other authors.

- Our results (Table 6) agree with Petoumenou et al. [19], who examined the performance of VEGA and EPISuite on data provided by the industry.
- The RMSE values of TEST and EPISuite we found are similar with those reported in Table 1.
- Finally, OPERA has never been evaluated by other authors, being a newly published model. The RMSE we obtained was higher than the one provided in the model's documentation.

In conclusion, our model can be a valid alternative tool for predicting the bioconcentration factor property within an industrial context, which is characterized by a much more heterogeneous chemical space than compounds coming from past studies, involving most of the time classical pollutants.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

G. Marcou  <http://orcid.org/0000-0003-1676-6708>

References

- [1] H.J. Geyer, G.G. Rimkus, I. Scheunert, A. Kaune, K.-W. Schramm, A. Kettrup, M. Zeeman, D. C. Muir, L.G. Hansen, and D. Mackay, Bioaccumulation and occurrence of endocrine-disrupting chemicals (EDCs), persistent organic pollutants (POPs), and other organic compounds in fish and other organisms including humans, in *Bioaccumulation—New Aspects and Developments*, B. Beek, ed., Springer Publisher, Berlin, 2000, pp. 1–166.
- [2] European Commission, Technical guidance document in support of commission directive 93/67/EEC on risk assessment for new notified substances and commission regulation (EC) No 1488/94 on risk assessment for existing substances, Tech. Rep. EUR 20418 EN/2, Institute for Health and Consumer Protection, Joint Research Centre, Ispra, IT, 2003.
- [3] OECD, Test No. 305: Bioaccumulation in fish: Aqueous and dietary exposure, Tech Rep. 9264185291, Organisation for Economic Co-operation Development, Paris, FR, 2012. doi:10.1094/PDIS-11-11-0999-PDN
- [4] ECHA, Guidance on information requirements and chemical safety assessment, r.11: PBT/vPvB assessment, Tech. Rep. ED-01-17-294, European Chemicals Agency, Helsinki, FI, 2017.
- [5] European Commission, Regulation (EC) no 1907/2006 of the european parliament and of the council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a european chemicals agency, amending directive 1999/45/ECC 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/EEC and 2000/21/EC, Off. J. Eur. Union. 50 (2007), pp. 1–281.
- [6] J.A. Arnot and F.A.P.C. Gobas, A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms, *Environ. Rev.* 14 (2006), pp. 257–297. doi:10.1139/a06-005.
- [7] M. Pavan, T.I. Netzeva, and W. Andrew, Review of literature-based quantitative structure - activity relationship models for bioconcentration, *QSAR Comb. Sci.* 27 (2008), pp. 21–31. doi:10.1002/qsar.200710102.
- [8] J.C. Dearden and M. Hewitt, QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors, *SAR QSAR Environ. Res.* 21 (2010), pp. 671–680. doi:10.1080/1062936X.2010.528235.
- [9] M. Nendza and M. Müller, Screening for low aquatic bioaccumulation (1): Lipinski's 'rule of 5' and molecular size, *SAR QSAR Environ. Res.* 21 (2010), pp. 495–512. doi:10.1080/1062936X.2010.502295.
- [10] J.F. Aranda, D.E. Babelo, M.S. Leguizamón Aparicio, M.A. Ocsachoque, E.A. Castro, and P. R. Duchowicz, Predicting the bioconcentration factor through a conformation-independent QSPR study, *SAR QSAR Environ. Res.* 28 (2017), pp. 749–763. doi:10.1080/1062936X.2017.1377765.
- [11] K. Mansouri, C.M. Grulke, R.S. Judson, and A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, *J. Cheminformatics* 10 (2018), pp. 1–19. doi:10.1186/s13321-018-0263-1.
- [12] E. Benfenati, A. Manganaro, and G. Gini, VEGA-QSAR: AI inside a platform for predictive toxicology, *Proceedings of the workshop 'Popularize Artificial Intelligence 2013'*, December 5th 2013, Turin, Italy, 2013, published in *CEUR Workshop Proceedings Vol 1107*, pp. 21–28.
- [13] T. Martin, P. Harten, and D. Young, TEST (Toxicity Estimation Software Tool) V 4.1, US Environmental Protection Agency, Washington DC, USA, 2012; software available at <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.

- [14] US EPA, Estimation Programs Interface Suite™ for Microsoft® Windows V 4.11, US Environmental Protection Agency, Washington DC, USA, 2012; software available at <https://www.epa.gov/tsca-screening-tools/epi-suite-estimation-program-interface>.
- [15] UFZ, ChemProp V 6.7, Helmholtz Centre for Environmental Research-UFZ, Leipzig, DE, 2018; software available at <http://www.ufz.de/ecochem/chemprop>.
- [16] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, Coral: Quantitative models for estimating bioconcentration factor of organic compounds, *Chemometr. Intell. Lab. 118* (2012), pp. 70–73. doi:10.1016/j.chemolab.2012.08.002.
- [17] ACD/Labs, ACD/logD V 2018.1, Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, CA, 2000; software available at <https://www.acdlabs.com>.
- [18] Catalogic: Environmental Fate and Ecotoxicity Models V 5. 11.13, OASIS Laboratory of Mathematical Chemistry, Burgas, BG, 2013; software available at <http://oasis-lmc.org/products/software/catalogic.aspx>.
- [19] M.I. Petoumenou, F. Pizzo, J. Cester, A. Fernández, and E. Benfenati, Comparison between bioconcentration factor (BCF) data provided by industry to the european chemicals agency (ECHA) and data derived from QSAR models, *Environ. Res.* 142 (2015), pp. 529–534. doi:10.1016/j.envres.2015.08.008.
- [20] ECHA Homepage, European Chemicals Agency, Helsinki, FI, 2019. Available at <https://echa.europa.eu/>.
- [21] A. Gissi, A. Lombardo, A. Roncaglioni, D. Gadaleta, G.F. Mangiatordi, O. Nicolotti, and E. Benfenati, Evaluation and comparison of benchmark QSAR models to predict a relevant reach endpoint: The bioconcentration factor (BCF), *Environ. Res.* 137 (2015), pp. 398–409. doi:10.1016/j.envres.2014.12.019.
- [22] F. Grisoni, V. Consonni, S. Villa, M. Vighi, and R. Todeschini, QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions? *Chemosphere* 127 (2015), pp. 171–179. doi:10.1016/j.chemosphere.2015.01.047.
- [23] F. Grisoni, V. Consonni, M. Vighi, S. Villa, and R. Todeschini, Expert QSAR system for predicting the bioconcentration factor under the reach regulation, *Environ. Res.* 148 (2016), pp. 507–512. doi:10.1016/j.envres.2016.04.032.
- [24] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics, *J. Chem. Inf. Comput. Sci.* 43 (2003), pp. 493–500. doi:10.1021/ci025584y.
- [25] Y.C. Wei, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *J. Comp. Chem.* 32 (2010), pp. 1466–1474.
- [26] OECD, Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models, ENV/JM/MONO(2007)2, OECD Series on Testing and Assessment No. 69. Organisation for Economic Cooperation and Development, Paris, FR, 2007.
- [27] G. Marcou, D. Horvath, F. Bonachera, and A. Varnek, Laboratoire De Chémoinformatique UMR 7140 CNRS, University of Strasbourg, Strasbourg, FR, 2019. Available at <http://infochim.u-strasbg.fr/>.
- [28] NITE, Data from: Biodegradation and bioconcentration data under CSCL National Institute of Technology and Evaluation, 2007; dataset available at <https://www.nite.go.jp/en/>.
- [29] CEFIC, Data from: BCF Bioconcentration factor database, European Chemical Industry Council Long range research initiative, 2007; dataset available at <http://cefic-lri.org/>.
- [30] Government of Canada, Data from: Canadian domestic substances list (DSL), Environment and Climate Change Canada, 1999; dataset available at <https://www.canada.ca/en/environment-climate-change/services/canadian-environmental-protection-act-registry/substances-list/domestic.html>.
- [31] US EPA, Data from: ECOTOX Knowledgebase, US Environmental Protection Agency, 2017; dataset available at <https://cfpub.epa.gov/ecotox/>.
- [32] QSAR Toolbox v 4.1, OASIS Laboratory of mathematical chemistry, Burgas, BG, 2017; software available at <http://oasis-lmc.org/products/software/toolbox.aspx>.

- [33] OECD, Data from: EChemPortal: Global portal to information on chemical substances, Organisation for Economic Co-operation Development, 2017; dataset available at <https://www.echemportal.org/echemportal/index.action>.
- [34] S. Dimitrov, N. Dimitrova, T. Parkerton, M. Comber, M. Bonnell, and O. Mekenyan, Base-line model for identifying the bioaccumulation potential of chemicals, *SAR QSAR Environ. Res.* 16 (2005), pp. 531–554. doi:10.1080/10659360500474623.
- [35] W. Fu, A. Franco, and S. Trapp, Methods for estimating the bioconcentration factor of ionizable organic chemicals, *Environ. Tox. Chem.* 28 (2009), pp. 1372–1379. doi:10.1897/08-233.1.
- [36] R.S. Boethling and J. Costanza, Domain of EPISuite biotransformation models, *SAR QSAR Environ. Res.* 21 (2010), pp. 415–443. doi:10.1080/1062936X.2010.501816.
- [37] A. Gissi, D. Gadaleta, M. Floris, S. Olla, A. Carotti, E. Novellino, E. Benfenati, and O. Nicolotti, An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes, *Altex* 31 (2014), pp. 23–36. doi:10.14573/altex.1305221.
- [38] P. Gramatica, S. Cassani, and A. Sangion, PBT assessment and prioritization by PBT Index and consensus modeling: Comparison of screening results from structural models, *Environ. Int.* 77 (2015), pp. 25–34. doi:10.1016/j.envint.2014.12.012.
- [39] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, and A. Varnek, Generative topographic mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison, *Mol. Inf.* 31 (2012), pp. 301–312. doi:10.1002/minf.v31.3/4.
- [40] C.M. Bishop, M. Svensén, C.K.I. Williams, and M. Svens, The generative topographic mapping, *Neural Comput.* 10 (1998), pp. 215–234. doi:10.1162/089976698300017953.
- [41] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, and A. Varnek, Chemical data visualization and mapping with incremental generative topographic mapping: Big data challenge, *J. Chem. Inf. Model.* 55 (2015), pp. 84–94. doi:10.1021/ci500575y.
- [42] D. Horvath, I. Baskin, G. Marcou, and A. Varnek, Generative topographic mapping of conformational space, *Mol. Inf.* 36 (2017), pp. 24–36. doi:10.1002/minf.201700036.
- [43] V.P. Solov'ev, A. Varnek, and G. Wipff, Modeling of ion complexation and extraction using substructural molecular fragments, *J. Chem. Inf. Comput. Sci.* 40 (2000), pp. 847–858.
- [44] R.E. Carhart, D.H. Smith, and R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: Definition and applications, *J. Chem. Inf. Comput. Sci.* 25 (1985), pp. 64–73. doi:10.1021/ci00046a002.
- [45] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, ISIDA property-labelled fragment descriptors, *Mol. Inf.* 29 (2010), pp. 855–868. doi:10.1002/minf.v29.12.
- [46] C. Chih-Chung and L. Chih-Jen, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.* 2 (2011), pp. 1–27. doi:10.1145/1961189.1961199.
- [47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, The WEKA machine learning workbench, in *Data Mining: Practical Machine Learning Tools and Techniques*, I. H. Witten and E. Frank, eds., Morgan Kaufmann Publishers, San Fransisco, 2005, pp. 363–449.
- [48] A. Tropsha, P. Gramatica, and V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* 22 (2003), pp. 69–77. doi:10.1002/(ISSN)1611-0218.
- [49] G. Marcou, D. Horvath, V. Solov'Ev, A. Arrault, P. Vayer, and A. Varnek, Interpretability of SAR/ QSAR models of any complexity by atomic contributions, *Mol. Inf.* 31 (2012), pp. 639–642. doi:10.1002/minf.201100136.
- [50] M.S. McLachlana, Can the stockholm convention address the spectrum of chemicals currently under regulatory scrutiny? Advocating a more prominent role for modeling in POP screening assessment, *Environ. Sci. Proces. Impacts* 20 (2018), pp. 32–37. doi:10.1039/C7EM00473G.