Modelling of ready biodegradability based on combined public and industrial data sources

F. Lunghini, G. Marcou, P. Gantzer, P. Azam, Dragos Horvath, E. van Miert, A. Varnek

SAR AND QSAR IN ENVIRONMENTAL RESEARCH
https://doi.org/10.1080/1062936X.2019.1697360

Modelling of ready biodegradability based on combined public and industrial data sources

F. Lunghini[a,b], G. Marcou[a], P. Gantzer[a], P. Azam[b], D. Horvath[a], E. Van Miert[b] and A. Varnek[a]

[a]Laboratory of Chemoinformatics - UMR7140, CNRS/University of Strasbourg, Strasbourg, France; [b]Toxicological and Environmental Risk Assessment Unit, Solvay S.A., St. Fons, France

ABSTRACT

The European Registration, Evaluation, Authorization and Restriction of Chemical Substances Regulation, requires marketed chemicals to be evaluated for Ready Biodegradability (RB), considering in silico prediction as valid alternative to experimental testing. However, currently available models may not be relevant to predict compounds of industrial interest, due to accuracy and applicability domain restriction issues. In this work, we present a new and extended RB dataset (2830 compounds), issued by the merging of several public data sources. It was used to train classification models, which were externally validated and benchmarked against already-existing tools on a set of 316 compounds coming from the industrial context. New models showed good performances in terms of predictive power (Balance Accuracy (BA) = 0.74–0.79) and data coverage (83–91%). The Generative Topographic Mapping approach identified several chemotypes and structural motifs unique to the industrial dataset, highlighting for which chemical classes currently available models may have less reliable predictions. Finally, public and industrial data were merged into global dataset containing 3146 compounds. This is the biggest dataset reported in the literature so far, covering some chemotypes absent in the public data. Thus, predictive model developed on the Global dataset has larger applicability domain than the existing ones.

ARTICLE HISTORY
Received 1 October 2019
Accepted 21 November 2019

KEYWORDS QSAR/QSPR; generative topographic mapping (GTM); ready biodegradability; environmental fate; reach; benchmarking

Introduction

Biodegradability is a key process which controls the environmental fate of chemicals and, as a consequence, potential exposure ways for living organisms to many xenobiotics. Indeed, chemicals which are persistent in the environment can potentially cause a long-term exposure to human beings and ecosystem on a large scale [1], for instance by reaching the marine environment and being transported to remote areas [2].

One of the most important ways for estimating biodegradation is determination of the so called 'Ready Biodegradability' (RB) binary classification parameter, corresponding to either slow (nB) or fast (B) biodegradation. There are several standardized methods for RB determination. Among them, the most widely used guideline is the Organization for

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.2019.1697360.

Economic Co-operation and Development (OECD) 301 [3], which contains several screening experimental protocols that aim to evaluate if, under aerobic conditions, the test substance can undergo easy and rapid biodegradation in the environment. Another well-known guideline is the method developed by the Japanese Ministry of International Trade and Industry (MITI) [3,4]. These protocols are considered as stringent first-tier assessments providing a binary classification, rather than measuring the actual degradation rate. Pass criteria of such tests are so strict that it can be assumed that compounds with a positive outcome will rapidly and completely biodegrade [3].

In Europe, with the implementation of the Registration, Evaluation, Authorization and Restriction of Chemical Substances (REACH, EC No 1907/2006) Regulation in 2007 [5], companies that produce or import substances for more than 1 ton/year need to provide information about their biodegradability, which would then be used for their classification as well as the evaluation of their level of exposure in the environment. The kinetic of biodegradation is also a key property in the identification of Persistent, Bioaccumulating and Toxic (PBT) or very Persistent and very Bioaccumulating (vPvB) compounds [6]. Thus, RB studies are generally performed in the very first stage of the registration process, with the aim to conclude on the absence of a possible PBT/vPvB behaviour. REACH encourages the use of alternative methods for data gap filling, including weight of evidence and read across approaches, as well as QSAR modelling [5]. However, biodegradation results are often highly dependent upon the test protocol and suffer of low reproducibility, especially when carried out by different laboratories [2,7,8]. The lack of homogeneous and high-quality datasets is a concern when generating predictive models.

Several RB models have already been built in the past years [2]. Some of them are nowadays implemented in freely-available tools, such as Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA) [9], Estimation Program Interface (EPI) Suite [10], OPEn (q)saR App (OPERA) [11] and ToxTree [12]. A brief overview of mentioned tools is reported in Table 1.

With the ending of the last REACH registration deadline (June 2018) for low-volume substances (between 1 and 100 tonnes) and the sharing of REACH study results (https://

Table 1. Already existing freely-available tools on ready biodegradability.

Model	General information	Training set size	Test set size	Sn	Sp	BA	Ref.
VEGA	Descriptors: molecular fragments Algorithm: rule-based approach	582	120	0.77	0.87	0.82	[9]
			491	0.76	0.91	0.84	[13]
			416	0.86	0.9	0.88	[14]
			757	0.89	0.93	0.91	[15]
			92	0.98	0.47	0.73	[16]
EPI Suite (Biowin 3 & 5)[a]	Descriptors: molecular fragments Algorithm: rule-based & linear model consensus	200 & 589[a]	295	0.87	0.73	0.8	[10]
			416	0.92	0.76	0.84	[13]
			199	0.6	0.83	0.72	[16]
			733	0.68	0.75	0.72	[17]
			110	0.48	0.9	0.69	[18]
OPERA	Descriptors: 2D descriptors Algorithm: k-NN	1197	411	0.81	0.77	0.79	[11]
ToxTree	Descriptors: molecular fragments Algorithm: rule-based approach	-	211	0.65	0.79	0.72	[16]

Sn = Sensitivity, Sp = Specificity, BA = Balanced Accuracy; [a]RB output is given as consensus between Biowin3 and Biowin5 models output: the two models' training sets size are reported.

iuclid6.echa.europa.eu/reach-study-results), new information is available. However, except for the recently published OPERA (2018), the training sets (from 200 to 589 compounds; Table 1) of the existing models is quite limited, and they have not been updated since several years.

In this work, we present a new and extended dataset for RB, issued from merging several public data sources. Gradual fusion of public and industrial data drove the fitting of successive models on steadily growing training sets, which were externally validated on a set of compounds coming from the industrial context ('Industrial set'). We generated three models: the first one ('ECHA model') is trained only on data coming from the ECHA's registration dossiers, which have gone careful reliability assessment; the second one ('All-Public model') comprises several sources of public data and has a much higher data coverage potential, yet at the expense of less verified data; and the last one ('Global model') is the most comprehensive model that we could build: it comes from the merging of the ECHA, the All-Public and the Industrial sets. This latter model includes important chemotypes of the industrial context; it has a much bigger training set (3146 compounds) compared to the existing tools (Table 1) and enlarged applicability domain.

Our models are available through the online In Silico Design and data Analysis (ISIDA)/Predictor platform [19], available at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi.

Methods

Modelling workflow

The modelling workflow is shown in Figure 1; the main steps will be detailed in the present section.

Data collection

Experimental data were collected from several sources: the ECHA database (accessed through the eChem portal [20]), the NITE database [4] and the training sets of already existing tools VEGA, EPI Suite and OPERA [9,10,11]. An industrial dataset (Industrial set) on biodegradation was provided by the industrial partner Solvay. Finally, additional RB data (Literature set) were collected from the work of Cheng et al. [21] and Mansouri et al. [22]. For the ECHA database, only reliable study results (i.e. with a Klimisch score [23] of 1 or 2) were retained. Curated datasets (i.e. after the data curation and standardization procedure below described) are listed in Table 2. Throughout the text, the three generated models (i.e. 'ECHA', 'All-Public' and 'Global') will be referred by the name of the dataset used for their generation. Both ECHA and All-Public models were externally validated on the Industrial set. Due to their different training set sizes, the number of truly external Industrial set compounds dropped from the initial 834 to 443 and 316, respectively. External validation for the Global model was carried out on the Literature set.

All collected public data (i.e. the All-Public set) is available on Zenodo (https://doi.org/10.5281/zenodo.3540701); the Industrial set compounds cannot be provided due to confidentiality reasons.

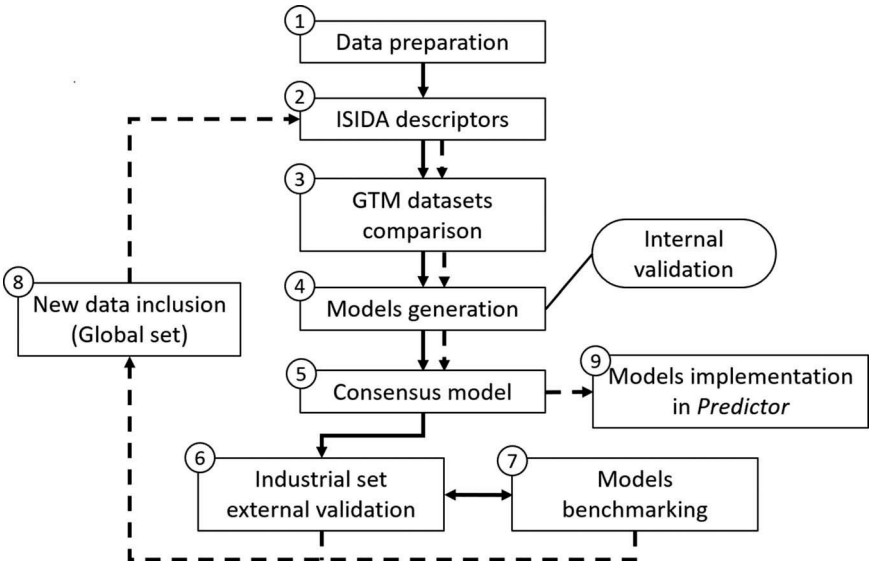

Figure 1. General workflow. (1) merging of collected data from multiple sources; (2) ISIDA descriptors are computed; (3) GTM is employed to compare the structural space of the datasets; (4), (5) individual models are trained using several machine learning algorithms and combined in consensus; (6) the Industrial set is used for external validation; (7) benchmarking against already existing tools; (8) the 'Global set' is issued by the merging of all collected data and (9) models are implemented in the Predictor platform.

Table 2. Datasets after data curation and standardization procedure.

Dataset	Size	B/nB	Ref.
NITE	861	203/658	[4]
VEGA	582	279/303	[9]
EPI SUITE	870	380/490	[10]
OPERA	1197	515/682	[11]
Industrial set	834	392/442	-
Literature set	362	36/326	[21, 22]
ECHA	1671	733/938	[20]
All-Public[a]	2830	1097/1733	-
Global[b]	3146	1197/1946	-

[a]All-Public dataset results from merging the NITE, VEGA, EPI SUITE, OPERA and ECHA datasets; [b]Global dataset, results from merging the All-Public and the Industrial datasets. The name of a particular model corresponds to the name of the dataset (e.g., ECHA model was trained on ECHA dataset).

Data curation and standardization

To check Simplified Molecular Input Line Entry System (SMILES) correctness, two online services were queried: the CADD Group Chemoinformatics Tools and User Services [24] and PubChem [25]. SMILES were generated, standardized and then cross-compared. Compounds with non-matching standardized SMILES were excluded. Chemical standardization included: removal of salts/solvents, neutralization, removal of explicit hydrogens, aromatic representation for benzene rings, removal of stereo information, transformation of -nitro and -sulpho containing groups into canonical notation. Standardization was done with workflow implemented in the Konstanz Information Miner (KNIME) software

[26]. Duplicates removal was based on standardized SMILES matching. In case of multiple values per compound, the most voted class was attributed; when the repartition of the B/nB votes was between 40 and 60%, the entry was excluded (See section 1 and Table S1 in Supplementary Material (SM)). In total, 125 compounds with discordant RB measurement were discarded by this filter. The full list, together with predicted values (by Global model), is available in Table S2 of SM. All non-relevant results (e.g. different guideline than OECD 301 or MITI-I, sampling time below the guideline threshold, etc.) as well as mixtures, polymers and 'Unknown or Variable composition, Complex reaction products or Biological materials' (UVCBs) were omitted. When the global statement of the RB behaviour was not reported but the percentage of biodegradation measured at 28 days (as requested by the OECD guideline) was available, it was manually assigned according to the relevant guideline threshold. The Literature set was processed in the same way. Out of the originally reported 1855 compounds, 362 were new to the Global model's training set. Four compounds were excluded as tested for inherent biodegradability; two compounds had wrongly reported labels which, after verifying the respective ECHA registration dossier, were corrected (Table S3 in SM). This dataset is highly unbalanced towards the nB class, as only 11% of compounds are readily biodegradable.

Molecular descriptors

ISIDA Property-Label Molecular descriptors [27] were employed. A total of 63 ISIDA descriptor spaces (DS) were generated, corresponding to molecular fragments of different sizes, topologies and 'colouration' (elements labels, physical properties mapped on the atoms explicit or implicit chemical bonds, atom pairs). Among this entire pool, the DS that led to the generation of under-performing models (see Model generation paragraph) were filtered out, retaining 19 DS (Table S4 in SM). The number of fragments depends on selected fragmentation scheme of the given DS. It varied from 203 (IA(2–6), sequences of atoms up to 6) for the ECHA model to 15872 (IIA(2–5), atom-centred fragments with radius 5) for the Global model, with an average of 6115 (SM, section 2).

Generative topographic mapping (GTM)

The chemical space of the collected datasets was compared by means of the generative topographic mapping (GTM) approach [28], a dimensionality reduction method allowing the visualization of data distribution on a two-dimensional (2D) map. A data property can be added as a third axis forming such called activity landscape. Each landscape 'spot' on the 2D map 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 [29]. Through GTM, two types of analysis were carried out: (i) a pairwise comparison between the Industrial set versus the other datasets (ECHA, NITE, training set of freely-available tools and All-Public); (ii) a characterization of how B and nB compounds are positioned in the chemical space. For the former case, the goal was to identify which chemotypes were unique to the industrial context, not represented by public data; for the latter, to visualize how the biodegradation outcome is related to the mapped structural space.

The DS (IIAB 2–3) [27] associated to the best support vector machine (SVM) radial basis function (RBF) model (in terms of Balanced Accuracy, BA) was chosen. The manifold [29] was built on the whole available chemical space (i.e. Global set). A genetic algorithm [30] was used for optimizing (with the goal to maximize the BA predicting B/nB compounds) the characteristic parameters of the GTM: the number of RBF function centres (m = 19), the RBFs width (w = 1.6) and the number of grid points, i.e. the dimension of the map (k = 19).

Model generation

SVM with linear and RBF kernels, Random Forest (RF) and Naïve Bayesian (NB) machine learning approaches were implemented. SVM models were generated with libSVM (v. 3.22) [31]; WEKA (v. 3.9.3) [32] was used for RF and for NB models. More details of the modelling process are available in Section 2 of SM, briefly:

(1) The given dataset has been randomly split (70/30%) into training and test set, and the 63 ISIDA DS were computed;
(2) SVM, RF and NB models have been fitted. SVM parameters (Cost and Gamma) were tuned by an independent genetic algorithm [29]driven optimization. For RF and NB, default WEKA settings were selected.
(3) Steps 1 and 2 were iterated 10 times. Resulting models with BA <0.70 (averaged over the iterations) were discarded.
(4) Only the best model (in terms of BA) among the three machine-learning approaches was kept for the given DS, unless it's BA <0.7. Fragmentation type and optimal method parameters corresponding to the best model were retained for the 'individual models' preparation.
(5) Finally, ensembles of 'individual models' were built on the whole dataset, each based on fragmentation and method parameters selected in previous step. Internal validation was carried out by three-fold CV by random splitting, performed for each individual model. This procedure was repeated five times. Statistics were assessed for each repetition followed by their averaging (Table 3). The influence of chance correlations was checked through Y-scrambling [33] (with 15 iterations).

This process was repeated for each dataset, i.e. ECHA, All-Public and Global, resulting into 19 individual models each. Performances were evaluated through Sensitivity (Sn), Specificity (Sp) and BA metrics, refer Section 2 in SM (Table S5).

Applicability domain

The applicability domain was evaluated through the 'fragment control' assessment (Figure 2, step 2): if a test molecule is found to have one fragment (i.e. a determined sequence of atoms and/or bonds) which was not encountered in any of the training molecules, that molecule is marked to be outside the applicability domain, since it is uncertain whether the model's predictions can be extrapolated to this not yet charted chemical space zone [27].

Table 3. Model performances.

			External validation[b]			
Model	Algorithm	BA in 3-fold CV	Sn	Sp	BA	Data coverage (%)
ECHA	SVM	0.80	0.83	0.72	0.81	81
	RF	0.81	0.82	0.77	0.8	80
	NB	0.78	0.84	0.7	0.77	78
	Consensus	0.79 (0.014)[a]	0.81	0.77	0.79	80% (353/443)
All-Public	SVM	0.79	0.78	0.71	0.74	89
	RF	0.8	0.76	0.72	0.74	92
	NB	0.77	0.81	0.62	0.72	89
	Consensus	0.79 (0.028)[a]	0.82	0.67	0.74	91% (293/316)
Global	SVM	0.8	0.62	0.84	0.73	85
	RF	0.81	0.61	0.86	0.74	83
	NB	0.77	0.61	0.84	0.72	81
	Consensus	0.81 (0.014)[a]	0.65	0.85	0.75	85% (307/362)

For each algorithm and the consensus, the Sensitivity (Sn), Specificity (Sp), Balanced Accuracy (BA) values are given in 3-fold CV and external validation (on the Industrial set). [a]In brackets, the standard deviation averaged over the CV repetitions is reported. [b]The Industrial set was used as an 'external test set' for ECHA and All-Public models; while the Literature set was used as an 'external test set' for the Global model.

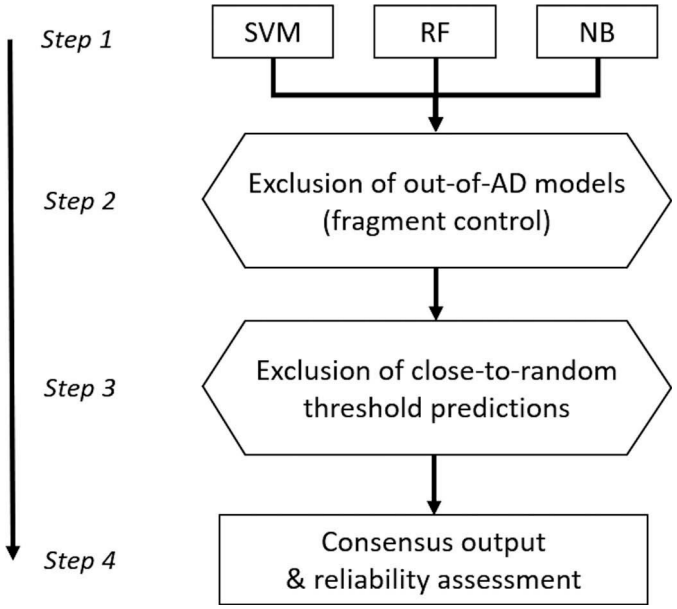

Figure 2. Consensus model workflow. Step 1: decisions of each algorithm (Support Vector Machine, Random Forest, Naïve Bayesian) are merged together; Step 2: predictions of models that failed the fragment control check are not considered; Step 3: if the percentage of votes for a given class is between 40 and 60% (i.e. close to random), the decision is rejected; Step 4: the consensus value is given with a reliability assessment.

Ensemble modelling

The graphical representation of the employed consensus strategy is shown in Figure 2. The ensemble decision is taken by a majority vote from the individual models of the employed algorithms (i.e. SVM, RF and NB) considered together (step 1). All out-of-AD decisions (based on the fragment control) are not considered for the voting (step 2). If the

percentage of the votes for a given class (B/nB) was between 40 and 60%, the decision was rejected since close to random (step 3); otherwise, the consensus prediction is given, together with its reliability (step 4) [34]. The data coverage is calculated as a ratio of the compounds accepted at steps 1 to 3 and total number in the dataset.

Benchmarking

Predictive performances on the Industrial set of the ECHA and the All-Public models were compared with those of the publicly available tools VEGA, EPI Suite, OPERA and ToxTree. To avoid potential overestimation, compounds already present in the training set of the given tool (not possible for ToxTree) were excluded. Thus, we selected a common subset of non-overlapping compounds for benchmarking. In total, seven molecules from the Industrial set were inside the training set of at least one model, reducing the number of usable compounds to 309. Moreover, the molecules outside of applicability domain of a given model were not considered (See Section 4 in SM).

Another benchmarking study concerned comparison of Global model with the publicly available tools assessed on the Literature set. At the first stage, 77 compounds from the Literature set overlapping with the training set of, at least, one of the benchmarked tools, have been excluded and, hence, the calculations were carried out on remaining 285 compounds. The Literature set together with models' predictions is reported by Table S6 in SI.

Results

GMT-driven dataset comparison

Two different types of fuzzy categorical landscapes were generated: (i) a 'dataset comparison' landscape, displaying chemical space zones occupied exclusively by members of a given dataset, zones never addressed by that dataset and zones where several datasets contribute; (ii) a two-class classification landscape of B versus nB compounds.

(i) Dataset comparison using generative topographic maps

Figure 3 shows a series of GTMs describing pairwise comparisons of the Industrial set with VEGA, EPI Suite, NITE, OPERA, ECHA and All-Public dataset. Occupied blue areas are uniquely populated by Industrial set compounds, while red ones by members of dataset x; intermediate colours are mixed populated areas. All the maps are characterized by having several constantly blue spots, indicating that the given areas contain Industrial set-unique compounds. Some of these areas (identified by rectangle 'A', map 6) even persist in the All-Public map: this provides a graphical interpretation of how the applicability domain could be extended with the addition of the new compound and clearly shows that there are some important structural differences between the Industrial set and the training set of the existing tools. For confidentiality reasons, the Industrial set cannot be disclosed. It comprises quite heterogeneous chemical structures, from high molecular weight compounds such as long-chain aliphatic esters highly halogenated compounds to much smaller ones such as simple alkenes. A large portion of them are silicium (e.g. siloxanes), fluorine (e.g. PFC) and phosphorous (e.g. organophosphonium cations) containing compounds, absent in public data sources. In addition, the All-Public and the Industrial

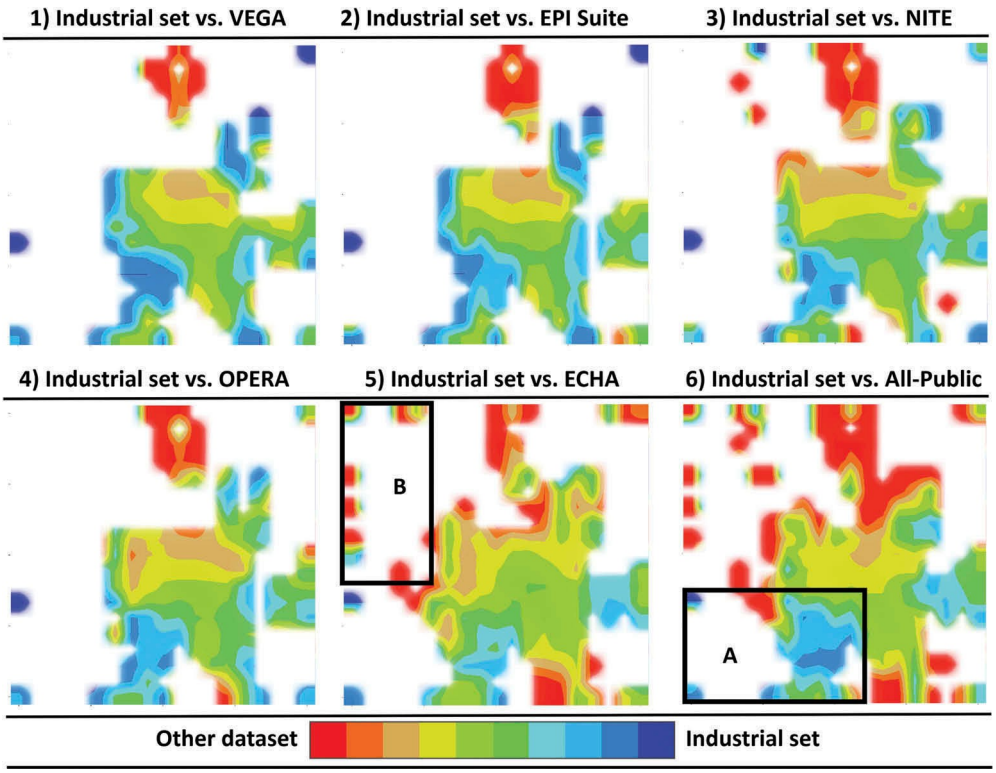

Figure 3. Pairwise datasets comparison using GTM. Each GTM map compares Industrial set versus one of publicly available datasets. Maps are sorted according to increasing size of a public dataset (from upper left to bottom right). Blue regions are mainly populated by Industrial set compounds; red ones by the public dataset compounds. White areas correspond to unpopulated regions. The manifold was prepared with the Global set.

datasets were compared by computing all pairwise Tanimoto similarities (Tc) among all their compounds (Section 3 in SM), using the DS IIAB(2–3). The average similarity value between public and industrial data resulted to Tc = 0.405, with the majority of public compounds (70%) having a Tc <0.6, indicating that the two datasets contain quite dissimilar compounds.

Finally, it is worth mentioning that there exists a strong overlap of VEGA, EPI Suite, OPERA and NITE sets: indeed, the models are mainly based on the same sources of data [4,9,10,11]. On the other hand, the ECHA set has some important structural differences, as it brings new chemotypes (rectangle 'B', map 5).

(ii) Ready biodegradability class landscape

Figure 4 depicts the B/nB class landscape. Readily biodegradable compounds are mainly clustered into one large area. Despite the fact that these compounds have quite heterogeneous structures, they share some common features, such as the absence of halogens, of heavily branched chains and of several aromatic rings. Esters and hydroxylic functional groups are known factors which increase the likelihood of rapid biodegradation [35]. It is interesting to observe that, the ECHA set is mainly adding nB entries, as

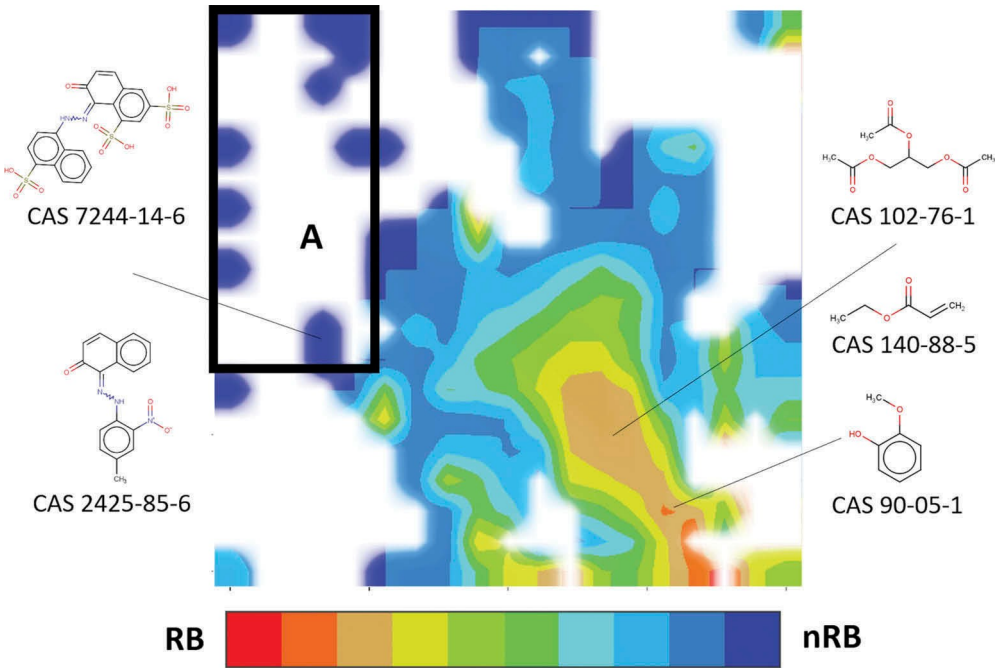

Figure 4. GTM ready biodegradability landscape. B compounds are identified by red zones, while nB by blue ones. The manifold was prepared for the Global set.

compounds in the area delimited by rectangle 'A' belong exclusively to this dataset (see Figure 3, map 5). As a consequence, since this structural information is unknown to the already existing models (Figure 3, maps 1–4), they may have missed some potentially relevant rules linked to the biodegradation property.

Model performances

Table 3 reports performances for the three generated models (ECHA, All-Public and Global model). Internal (three-fold CV and Y-scrambling) and external (the Industrial set) validation statistics are reported for each machine-learning algorithm and the consensus. Performances on the Industrial set (BA = 0.74–0.79) are not too different from those determined by CV (BA = 0.79), which supports the model robustness and absence of overfitting. In addition, the performance of 'scrambled' models is close to the random threshold (BA = 0.51–0.55; standard deviation among repetitions = 0.12–0.17), which confirms that models are unlikely to be biased by chance correlations. In external validation, the ECHA model showed a BA of 0.79 with a data coverage, here defined as the percentage of reliably predicted compounds (Figure 2) out of the total, of 80% (353 out of 443 compounds); while the All-Public model scored a BA of 0.74 and data coverage of 91% (293 out of 316). Thus, the latter model has an extended applicability domain at the expense of a lower accuracy (with a drop in BA of 5%): this supports our starting hypothesis concerning experimental data reliability. It is important to highlight that the two models were evaluated on a different set of compounds. Therefore, in order to strictly

compare performances, an evaluation on exactly the same compounds should be performed. A similar trend was noticed in the benchmarking comparison, which was based only on the smallest common subset (see 'Model benchmarking' paragraph).

Models' performances evaluated without any AD filter (Figure 2) are degraded, with BA of 0.73 (Sn = 0.81, Sp = 0.66) and 0.74 (Sn = 0.79, Sp = 0.68) for the ECHA and the All-Public model, respectively.

Even though the enlarged training set of the All-Public model, some chemotypes (e.g. siloxanes) remained unique to the Industrial set: therefore, the inclusion of new compounds from the industrial context is a necessary step in order to create RB dataset as comprehensive as possible. For this reason, the Industrial set was combined with the available public data leading to the 'Global set' of 3146 compounds, which, in turn, was externally validated on the Literature set.

Relatively small value of Sn (0.65, Table 3) resulting from the application of the Global model on the Literature set can be explained by the imbalance of the latter (the ratio of 'B' over 'nB' is only 0.11). Furthermore, we noticed that the experimental 'B' value for some of the wrongly predicted compound may be uncertain: for instance, CAS 84-65-1 is considered to be readily biodegradable even though it failed the '10-day window' condition [3]; from PubChem, [25] CAS 78-48-8 shows very high degradation half-lives in all environmental compartments; while CAS 88-06-2 (2,4,6-trichlorophenol) is reported to be biodegradable, despite all other chlorinated phenols family members in collected datasets are nB.

Model benchmarking

Table 4 reports the Industrial and Literature set performances for our models versus the considered tools. On the former set, considering accuracy and data coverage, the ECHA and the All-Public models and EPI Suite scored the best performance, with comparable BA values (0.77, 0.74 and 0.73, respectively). VEGA had one of the highest BA (0.71) as well, but its data coverage was rather limited to 44%. Furthermore, it has a very good propensity to recognize B compounds (Sn = 0.95) but tends to be 'overcautious' with the nB class (Sp = 0.48), often mispredicted as B. As a limitation, all models (except for EPI Suite) failed to predict most part of exclusive chemicals of the Industrial set (e.g. organophosphonium cations), due to applicability domain restrictions. This indicates that the availability of current public RB data was not enough to cover all the main chemotypes of the Industrial set, in agreement with the findings of GTM analysis (Figure 3).

Table 4. Benchmarking of different models on the industrial set.

	Industrial set				Literature set			
Model	Sn	Sp	BA	Data coverage	Sn	Sp	BA	Data coverage
ECHA	0.85	0.68	0.77	83%	-	-	-	-
All-Public	0.82	0.67	0.74	91%	-	-	-	-
Global	-	-	-	-	0.88	0.93	0.91	86%
VEGA	0.95	0.48	0.71	44%	0.87	0.91	0.89	58%
EPI Suite	0.65	0.74	0.69	99%	0.58	0.96	0.77	100%
OPERA	0.71	0.65	0.68	84%	0.83	0.88	0.86	80%
ToxTree	0.61	0.73	0.67	84%	0.58	0.92	0.75	96%

Statistics are computed on the common set of non-overlapping compounds of the Industrial (309) and Literature (285) sets. Compounds' out-of-ADs were not considered for performances estimation.

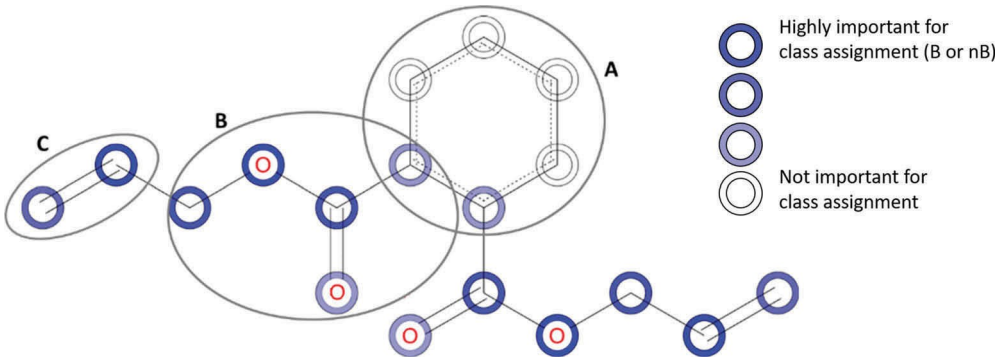

Figure 5. ColorAtom output representation. Colours refer to sensitivity of classification model to presence of a given fragment or atom: the darker the colour, the more that fragment (atom) is important for assigning the molecule to a given class.

All models with the exception of EPI Suite (BA = 0.77) and ToxTree (BA = 0.75) scored very good performances on the Literature set (BA = 0.86–0.91). Notice that performances of Global model on the Literature set given in Tables 3 and 4 differ because of different number considered test set compounds. Thus, Table 3 reports calculations performed on the entire Literature set (362 compounds), whereas in Table 4 only non-overlapping with other tools 285 compounds were used. Much higher BA = 0.91 value reported in Table 4 compared to BA = 0.75 reported in Table 3 can be explained by filtering out some noisy data.

Coloratom: structure-activity dependence analysis

The 'ColorAtom' utility assigns a colour code to each fragment or atom showing the Sn of classification model to its presence in molecular structure [36] For a given fragment, dark colour shows that its presence is crucial to assign the molecule to a given class, while completely transparent colour means that the model is insensitive to its presence. As an example, a graphical representation of Diallyl phthalate (CAS 131-17-9) is shown in Figure 5. It can be noticed that the benzene ring (ellipse A) does not affect the RB outcome, by contrast to the two carbon chains. The ester and end-chain ethene functional groups (ellipses B and C) were found to be particularly significant for RB determination. These functional groups are known to be reactive in the environment [35].

Discussion

Already-existing tools performed worse on the Industrial set (Table 4) when compared to other evaluations retrieved from the literature (Table 1). Such low performances may be attributed to the different nature of the compounds of the Industrial set: as also highlighted by GTM, there exist some noticeable structural differences between the training set of the models and the Industrial set compounds. For example, for both VEGA and OPERA, prediction accuracy reported in the literature was be significantly higher compared to our analysis (average BA of VEGA and OPERA of 0.85 and 0.79 vs. 0.71 and 0.68,

respectively). Both the ECHA and the All-Public model scored the best-BA (BA of 0.77 and 0.74) and data coverage (83 and 91%) on the Industrial set. As shown by GTM, the inclusion of the ECHA dataset brought unique structural features shared with the Industrial set which were unknown to the other tools. Despite the fact that ToxTree is a relatively simple ensemble of structural alert rule set, with an AD implicitly limited to existence of rules that apply to a given compound (otherwise, outcome is 'unknown'), it showed reasonable accuracy. In addition, its output provides the set of rules that have been used to generate the prediction. Data coverage on the Industrial set varies largely, ranging from 44 to 99% for VEGA and EPI Suite, respectively. However, for the latter, its AD is not clearly defined [37,38]. It is remarkable that some tools (e.g. OPERA and ToxTree) have an opposite behaviour in terms of Sn and Sp: ToxTree is biased in favour of B class assignment, with a higher rate of false positives, while OPERA would rather fail to recognize some B compounds and thus limits the number of false positives.

Both our models possess several strengths: the ECHA model showed a wide data coverage and the best accuracy among the other tools, while the All-Public model has a much higher data coverage potential, yet at the expense of prediction accuracy (Table 4). The Global model has a much larger training set (3146 compounds) compared to all the other already-existing tools (Table 1) and incorporates a significant subset of compounds (316) which include important chemotypes of the industrial context.

The developed models follow the OECD principles [39]: the endpoint (RB) is well defined; goodness-of-fit, robustness and predictivity were evaluated using three-fold CV, Y-scrambling, and external validation [33]; the AD of the models was defined using a fragment control assessment [27] together with a reliability scoring function.

Conclusions

In this work we reported preparation of new extended datasets for RB and related classification models (B/nB).

Gradual fusion of public source and industrial data led to successive RB models on steadily growing training sets. The first 'ECHA model' was built on 1671 compounds collected from the ECHA database. A second 'All-Public model' was generated by the merging of ECHA data with several other public databases, producing a public RB dataset as comprehensive as possible, counting 2830 compounds. Both models were externally validated on a set of 316 compounds coming from the industrial context provided by Solvay ('Industrial set'). Compared to the ECHA model, the All-Public model showed a decrease in BA (from 0.79 to 0.74), on one hand, and an improvement in data coverage which is consistent with the addition of new information (from 83 to 91%), on the other hand. The former suggests that noise has been added with the merging of all the available data.

A benchmarking against the already existing tools showed that the ECHA model scored the best predictive power (BA = 0.77), followed by the All-Public model, VEGA, EPI Suite, OPERA and ToxTree, with BA values of 0.74, 0.71, 0.69, 0.68 and 0.67, respectively. This comparison demonstrated that each model has specific strong points: for example, VEGA is able to correctly classify true-positive B compounds, whereas EPI Suite has the highest data coverage among all the tools and our models the best accuracy. Nevertheless, an important common downside to all the models was the limitation to predict several compounds classes of industrial interest (e.g. siloxanes and organophosphonium cations), because their

training sets lack such instances. These structural differences of compounds in the Industrial set and public datasets were highlighted through Generative Topographic Mapping. Finally, collected public data and the Industrial set have been merged into the 'Global' dataset containing 3146 compounds which is the biggest RB set reported so far covering important representative chemotypes of the industrial context. The 'Global' model built on this dataset was externally validated on a set of 362 new compounds taken from the literature, scoring a BA of 0.75. Our models are available for the users at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi. Collected public data are freely accessible on Zenodo (https://doi.org/10.5281/zenodo.3540701).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

F. Lunghini http://orcid.org/0000-0002-4625-6736
G. Marcou http://orcid.org/0000-0003-1676-6708
P. Gantzer http://orcid.org/0000-0001-7494-458X
P. Azam http://orcid.org/0000-0002-2974-2484
D. Horvath http://orcid.org/0000-0003-0173-5714
E. Van Miert http://orcid.org/0000-0001-6653-1371
A. Varnek http://orcid.org/0000-0003-1886-925X

References

[1] P. Gramatica and E. Papa, Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure, Environ. Sci. Technol. 41 (2007), pp. 2833–2839. doi:10.1021/es061773b.
[2] M. Pavan and A.P. Worth, Review of estimation models for biodegradation, QSAR Comb. Sci. 27 (2008), pp. 32–40. doi:10.1002/(ISSN)1611-0218.
[3] OECD, Test No. 301: Ready biodegradability, Tech. Rep. 9789264070349, Organisation for Economic Co-operation Development, Paris, FR, 1992.
[4] 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/.
[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] 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.
[7] A. Kowalczyk, T.J. Martin, O.R. Price, J.R. Snape, R.A. van Egmond, C.J. Finnegan, H. Schäfer, R. J. Davenport, and G.D. Bending, Refinement of biodegradation tests methodologies and the proposed utility of new microbial ecology techniques, Ecotoxicol. Environ. Saf. 111 (2015), pp. 9–22. doi:10.1016/j.ecoenv.2014.09.021.
[8] T.J. Martin, J.R. Snape, A. Bartram, A. Robson, K. Acharya, and R.J. Davenport, Environmentally relevant inoculum concentrations improve the reliability of persistent assessments in biodegradation screening tests, Environ. Sci. Technol. 51 (2017), pp. 3065–3073. doi:10.1021/acs.est.6b05717.

[9] 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 on CEUR Workshop Proceedings Vol 1107.
[10] US EPA, Estimation Programs Interface Suite™ for Microsoft® Windows V 4.11, US Environmental Protection Agency, Washington DC, 2012; software available at https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface.
[11] K. Mansouri, C.M. Grulke, R.S. Judson, and A.J. Williams, OPERA models for predicting physico-chemical properties and environmental fate endpoints, J. Cheminform. 10 (2018), pp. 1–19. doi:10.1186/s13321-018-0263-1.
[12] JRC, ToxTree 3.1.0 - Toxic hazard estimation by decision tree approach, Joint Research Centre (JRC), Ispra, Italy, 2018; software available at https://ec.europa.eu/jrc/en/eurl/ecvam.
[13] A. Lombardo, F. Pizzo, E. Benfenati, A. Manganaro, T. Ferrari, and G. Gini, A new in silico classification model for ready biodegradability, based on molecular fragments, Chemosphere 108 (2014), pp. 10–16. doi:10.1016/j.chemosphere.2014.02.073.
[14] D. Ballabio, F. Biganzoli, R. Todeschini, and V. Consonni, Qualitative consensus of QSAR ready biodegradability predictions, Toxicol. Environ. Chem. 99 (2017), pp. 1193–1216.
[15] A. Fernández, R. Rallo, and F. Giralt, Prioritization of in silico models and molecular descriptors for the assessment of ready biodegradability, Environ. Res. 142 (2015), pp. 161–168. doi:10.1016/j.envres.2015.06.031.
[16] R. Boethling, Comparison of ready biodegradation estimation methods for fragrance materials, Sci. Total Environ. 497–498 (2014), pp. 60–67. doi:10.1016/j.scitotenv.2014.07.090.
[17] E. Rorije, H. Loonen, M. Müller, G. Klopman, and W.J.G.M. Peijnenburg, Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test, Chemosphere 38 (1999), pp. 1409–1417. doi:10.1016/S0045-6535(98)00543-8.
[18] R. Posthumus, T.P. Traas, W.J.G.M. Peijnenburg, and E.M. Hulzebos, External validation of EPIWIN biodegradation models, SAR QSAR Environ. Res. 16 (2005), pp. 135–148. doi:10.1080/10629360412331319899.
[19] G. Marcou, D. Horvath, F. Bonachera, and A. Varnek, Laboratoire De Chemoinformatique UMR 7140 CNRS, University of Strasbourg, Strasbourg, FR, 2019; available at http://infochim.u-strasbg.fr/ [Accessed 1 May 2019].
[20] OECD, Data from: EChemPortal: Global portal to information on chemical substances, Organisation for Economic Co-operation Development; dataset available at https://www.echemportal.org/echemportal/index.action.
[21] F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu, P.W. Lee, and Y. Tang, In silico assessment of chemical biodegradability, J. Chem. Inf. Model. 52 (2012), pp. 655–669. doi:10.1021/ci200622d.
[22] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, Quantitative structure-activity relationship models for ready biodegradability of chemicals, J. Chem. Inf. Model. 53 (2013), pp. 867–878. doi:10.1021/ci4000213.
[23] H.-J. Klimisch, M. Andreae, and U. Tillmann, A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data, Regul. Toxicol. Pharmacol. 25 (1997), pp. 1–5. doi:10.1006/rtph.1996.1076.
[24] NCI, CADD Group Chemoinformatics Tools and User Services, National Cancer Institute, Chemical Biology Laboratory, Bethesda, Maryland, 2019; available at https://cactus.nci.nih.gov/.
[25] NIH, PubChem, National Library of Medicine, National Center for Biotechnology Information, Bethesda, Maryland, 2019; available at https://pubchem.ncbi.nlm.nih.gov/.
[26] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, KNIME - the Konstanz information miner: Version 2.0 and beyond, SIGKDD Explor. 11 (2009), pp. 26–31. doi:10.1145/1656274.1656280.
[27] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, ISIDA property-labelled fragment descriptors, Mol. Inform. 29 (2010), pp. 855–868. doi:10.1002/minf.201000099.
[28] 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.

[29] 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. Inform. 31 (2012), pp. 301–312. doi:10.1002/minf.201100163.
[30] D. Horvath, J. Brown, G. Marcou, and A. Varnek, An evolutionary optimizer of libsvm models, Challenges 5 (2014), pp. 450–472. doi:10.3390/challe5020450.
[31] 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.
[32] I.H. Witten, E. Frank, M.A. Hall, and C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, San Fransisco CA, 2016.
[33] 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.
[34] F. Lunghini, G. Marcou, P. Azam, R. Patoux, M.H. Enrici, F. Bonachera, D. Horvath, and A. Varnek, QSPR models for bioconcentration factor (BCF): Are they able to predict data of industrial interest? SAR QSAR Environ. Res. 30 (2019), pp. 507–524. doi:10.1080/1062936X.2019.1626278.
[35] R.S. Boethling and J. Costanza, Domain of EPI suite biotransformation models, SAR QSAR Environ. Res. 21 (2010), pp. 415–443. doi:10.1080/1062936X.2010.501816.
[36] 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. Inform. 31 (2012), pp. 639–642. doi:10.1002/minf.201100136.
[37] E. Hulzebos, D. Sijm, T. Traas, R. Posthumus, and L. Maslankiewicz, Validity and validation of expert (Q)SAR systems, SAR QSAR Environ. Res. 16 (2005), pp. 385–401. doi:10.1080/10659360500204426.
[38] R.S. Boethling, E. Sommer, and D. DiFiore, Designing small molecules for biodegradability, Chem. Rev. 107 (2007), pp. 2207–2227. doi:10.1021/cr050952t.
[39] OECD, Guidance document on the validation of (Quantitative) Structure Activity Relationship [(Q)SAR] models, Tech. Rep. ENV/JM/MONO(2007)2, Organisation for Economic Cooperation and Development, Paris, FR, 2007. doi:10.1094/PDIS-91-4-0467B.