Visualization and Analysis of the REACH-chemical Space with Generative Topographic Mapping

Filippo Lunghini, Gilles Marcou, Philippe Azam, Marie-hélène Enrici, Erik van Miert, Alexandre Varnek

molecular informatics

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Visualization and analysis of the REACH-chemical space with Generative Topographic Mapping

Filippo Lunghini[a,b], Gilles Marcou[a,*], Philippe Azam[b], Marie-Hélène Enrici[b], Erik Van Miert[b], Alexandre Varnek[a,*]

Abstract: In the framework of REACH (Registration Evaluation Authorization and restriction of Chemicals) regulation, industries have generated and reported a huge amount of (eco)toxicological data on substance produced or imported in Europe. The registration procedure initiated the creation of a large REACH database of well defined (eco)toxicological properties. Here, the data distribution in the REACH chemical space was analyzed with the help of the Generative Topographic Mapping (GTM) approach. GTM generates 2-dimensional maps on which each compound is represented as a data point. The 3rd dimension can be used in order to display a distribution of the given (eco)toxicological property, which can further be used for property assessment of new compounds projected on the map. We report the "Universal REACH map" which accommodates 11 endpoints, covering environmental fate and (eco)toxicological properties. This map demonstrates acceptable predictive performance: in cross-validation, balanced accuracy ranges from 0.60 to 0.78. The 11 endpoints profile has been computed for each REACH-registered substance. Some concerns related to acute aquatic toxicity have been identified, whereas for environmental fate and human health endpoints the amount of compounds predicted as of concern was much smaller. It has been demonstrated that superposition of several class landscapes allows to select the zones in the chemical space populated by compounds with a given (eco)toxicological profile.

Keywords: REACH chemical space, Generative Topographic Mapping (GTM), environmental fate, ecotoxicology, visualization

1 Introduction

The European REACH (Registration Evaluation Authorization and restriction of Chemicals) regulation[1], established in 2007, requires from industry to register all substances imported or manufactured in quantities larger than one ton/year. To this end, the registrants must submit to the European Chemicals Agency (ECHA) a technical dossier that characterizes for a given substance, its physical-chemical, environmental fate and (eco)toxicological properties, called endpoints. In order to decrease a need for expensive experimental tests, the REACH regulation allows to use some alternative methods, including predictive statistical models.

In the past years, several models predicting REACH-relevant endpoints were obtained with the help of various machine-learning methods like multiple linear regression, support vector machine or neural networks [2–5]. However, some of them suffered from absence of technical documentation complying with the REACH requirements[6], for instance concerning the model's Applicability Domain (AD) or its validation procedure[3,7]. In our recent studies[2,4], we found that for some endpoints, currently-existing tools have disappointing performances when applied to compounds coming from an industrial context, due to their restricted AD. Some existing tools used for the chemicals ranking according to their environmental and toxicological concern, such as DART (Decision Analysis by Ranking Techniques)[8], use experimental data as an input. To overcome these drawbacks, here, we propose an integrated profiling methodology based on Generative Topographic Mapping (GTM)[9]. GTM is a probability-based mapping strategy, which can be applied both for large-scale data visualization and property prediction. We chose this method because of the following reasons: (*i*) GTM allows multi-task learning, since several properties can simultaneously be accounted for; (*ii*) it produces a graphical output (i.e. a two-dimensional map of the chemical space) which can be used as support to better understand the model's output, in the light of a mechanistic interpretation. Here, a set of curated data for 11 endpoints prepared in our previous studies[2–4] ("Global dataset") has been used to train the GTM model and to delineate the REACH chemical space. The following 11 endpoints were considered: bioconcentration factor, ready biodegradability, environmental persistence in sediment, soil and water media, acute aquatic toxicity to algae, daphnia and fish, rat acute toxicity, androgen and estrogen receptor binding potential. Our goal was to demonstrate a potential of GTM to integrate available data into a multi-task modelling framework and to facilitate identification of the compounds of potential concern. For this purpose, a profile assembling selected 11 properties has been computed for each substance registered under the REACH Regulation (called "REACH-INV").

The manuscript is organized in the following way: first, we describe the GTM model built on the Global dataset, followed its application to profile the REACH-INV compounds. Second, we discuss performances of the GTM approach as multi-task learning method, highlighting its potential as a profiling tool.

2 Materials and methods

2.1 Considered endpoints

Table 1 reports the 11 considered endpoints. A brief description is provided in the following paragraphs.

[a] Laboratory of Chemoinformatics – UMR7140, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France

[b] Toxicological and Environmental Risk Assessment unit, Solvay S.A., 85, avenue des Frères Perret, 69192, St. Fons, France

** g.marcou@unistra.fr; varnek@unistra.fr; phone no.: +33-68851304*

Supporting Information for this article is available on the WWW under www.molinf.com

2.1.1 Bioconcentration Factor (BCF)

BCF estimates the tendency for a xenobiotic to concentrate inside living organisms. 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[2].

2.1.2 Ready biodegradability (RB)

Long term exposure for living organisms to many xenobiotics is dependent on the environmental fate of such chemicals which in turn is highly dependent on their biodegradation. Biodegradability is determined by a multistep procedure. This assessment usually starts with a very stringent first-tier assay, providing a binary classification whether the substance rapidly degrades in the environment, called "ready biodegradability"[4].

2.1.3 Sediment, Soil and Water Persistence (SedP, SoilP, WatP)

Unlike relatively cheap and fast ready biodegradability assay, these higher-tier simulation studies are carried out when the substance's degradation half-life (in a given environmental compartment) value actually needs to be evaluated[10].

2.1.4 Aquatic acute toxicity to Algae, Fish and Daphnia (AlgaeTox, DaphniaTox and FishTox)

Aquatic acute toxicity tests aim to estimate the short-term toxicity[11] against three species belonging to different trophic levels, considered to be representative of the aquatic ecosystem. Briefly, the test organisms are exposed to the studied substance via the water media, and the following substance-induced effects are measured: (*i*) for Algae, growth inhibition, expressed as median effective concentration (EC50) measured at 72 hours; (*ii*) for Daphnia, immobilization at 48 hours expressed as median effective concentration (EC50); (*iii*) for Fish, the median lethal concentration at 96 hours (LC50).

2.1.5 Rat acute toxicity (RatTox)

Rat acute toxicity estimates the short-term lethality (hazard) to humans following ingestion for which oral administration to rodents is used as a proxy. The REACH regulation requires its assessment even for small tonnages. Consequently, this experimental test is one of the most commonly performed animal tests which partly explains its much higher data availability compared to the other endpoints[3].

2.1.6 Androgen and Estrogen receptor binding (AR binding and ER binding)

An endocrine disrupting chemical is an exogenous substance that alters the functions of the endocrine system to the point of causing adverse effects. Possible ways for a chemical to alter the endocrine system is to bind to androgen or estrogen receptors in an agonist or antagonist way. In the framework of the "Collaborative Estrogen Receptor Activity Prediction Project" (CERAPP)[12] and "Collaborative Modelling Project for Androgen Receptor Activity" (CoMPARA)[13] international workgroups, a large number of compounds were tested for their potency to disrupt the AR/ER signaling pathway chains.

2.2 Data collection and curation

Experimental data was collected from multiple publicly available databases and scientific literature[2–4]. Among them, the main source was the database of the European Chemical Agency (ECHA)[14], which comprises the REACH-registered substances. Raw data processing and standardization were done with workflows implemented in the Konstanz Information Miner (KNIME) software[15]. The PubChem[16] online service was queried to verify SMILES correctness. Generated SMILES were then standardized with the following rules: removal of salts/solvents, removal of explicit hydrogens, aromatic representation of benzene rings, removal of stereo information and transformation of -nitro and -sulpho containing groups into canonical notation, neutralization. Duplicates were removed based on standardized SMILES matching.

As some endpoints are typically described by continuous values (e.g. acute toxicity) while others by categorical values (e.g. ready biodegradability), all the former properties were discretized into "Concern" (C) or "non-Concern" (nC) binary classes (Table 1). For this purpose, REACH-relevant threshold values were selected: for instance, a substance is defined as persistent in sediment (class C) if its degradation half-life is higher than 120 days.

Whenever it was possible, the thresholds have been selected with the goal to make them as conservative as possible. For acute aquatic toxicity, we introduced a ten-fold increase to the 1 mg/L limit (mentioned by REACH to classify the substance as hazardous for the aquatic environment), raising it to 10 mg/L. For the bioaccumulation factor, we selected the cutoff of BCF > 3 log unit, which corresponds to the bioaccumulation limit as defined by the US EPA (that is more conversative than the one reported by REACH of 3.3 log unit). The only exception was for acute rodent toxicity: in that case, a less conservative value of 300 mg/kg (corresponding to "acute toxicity category 4" under the Classification, Labelling and Packaging – CLP Regulation) has been used. To be conversative, the cutoff of >2000 mg/kg (which corresponds to "not classified" under the CLP) should have been selected to discriminate C vs. nC classes. However, this lead to an unbalanced dataset (with the majority of compounds being classified as of concern) leading to less contrasted map that we considered less useful maps in the current situation.

It is important to emphasize that these thresholds have been selected with a certain degree of arbitrariness and their relevance vary according to the field of application (for instance when considering a different regulatory framework).

For the remaining endpoints, the label C was assigned when the experimental values had an assignment of "concern" (e.g. binding to the AR/ER receptors or not readily biodegradable), as opposed to the nC label.

The Global dataset results from the merging of all the available data on the abovementioned 11 endpoints: it counts 29433 experimentally measured data points for 17762 unique compounds (as one compound can have more than one associated experimental value). This dataset has been assembled and curated in our previous studies[2–4]. Table 1 reports the sizes of the endpoint subsets. The REACH-INV set comprises the entire inventory of the substances registered under REACH, that have been extracted from the European Chemicals Agency database[17], and chemical structures were curated using the same procedure as for the Global set. As this database has been queried in our previous works[2–4], there is a certain degree of overlap between the Global dataset and the REACH-INV set, as it is shown in section 3.3. REACH-INV does not contain any experimental data, but only a list of substances concerned by the Regulation. In the end, a total of 11951 compounds (out of 22966) have been retained. The Global dataset and the REACH-INV are available through Zenodo: 10.5281/zenodo.3872735.

Table 1. Selected endpoints and data availability.

Endpoint	Acronym	C/nC threshold [a]	Unique compounds	C/nC [b]
Bioconcentration factor	BCF	3 log units	1260	299 / 961

Ready biodegradability	RB	-	3069	1168 / 1901
Persistence in Sediment	SedP	120 days	436	253 / 183
Persistence in Soil	SoilP	120 days	630	111 / 519
Persistence in Water	WatP	40 days	466	191 / 275
Algae acute toxicity	AlgaeTox	10 mg/L	1231	531 / 700
Daphnia acute toxicity	DaphniaTox	10 mg/L	2083	897 / 1186
Fish acute toxicity	FishTox	10 mg/L	2152	1046 / 1106
Rat acute toxicity	RatTox	300 mg/kg b.w.	14784	3206 / 11578
Androgen receptor binding	AR binding	-	1661	198 / 1463
Estrogen receptor binding	ER binding	-	1661	223 / 1438

[a] selected threshold to discretize continuous properties into Concern (C) or non-Concern (nC) binary labels;
[b] repartition of C and nC compounds.

2.3 Modelling workflow

The workflow is depicted by Figure 1. Its main steps are described in this section.

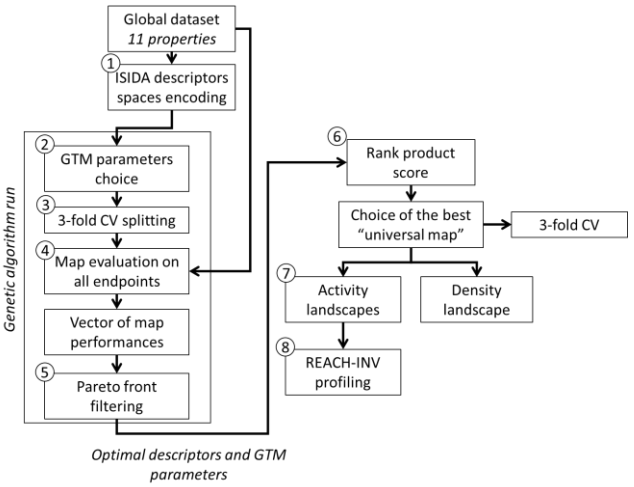

Figure 1. Modelling workflow. (1) different ISIDA descriptor spaces are generated for the Global dataset; (2) Genetic algorithm evaluates different types of descriptors and GTM parameters; (3-4) predictive performance of GTM-based models is evaluated in 3-fold cross-validation (CV) on each of the 11 properties; (5) the Pareto front filtering is applied to exclude all dominated solutions; (6) the rank product score is applied to select the best "REACH Universal Map"; (7) the chemical space is analyzed using both density and class landscapes; (8) the REACH-INV set is profiled.

2.3 Molecular descriptors

The Global dataset, is encoded (Figure 1, step 1) by several types of ISIDA Property-Label Molecular descriptors[18]. These descriptors work as substructures (fragment) counts of a molecule – for example, D1 = number of "C=O" groups, D2 =number of "C-N-C" fragments, etc. The molecule can be fragmented using two main fragmentation patterns: sequences or atom centered fragment. Moreover, in both cases the size of the fragment (length or radius, respectively) can be varied. Each unique fragmentation scheme is referred as descriptor space. Several tens of descriptor spaces were generated. This list of different fragmentation patterns was used as a starting pool for searching the most appropriate descriptor space by means of genetic algorithm selection. Table S1 reports the descriptors and parameters employed for the given GTM model.

2.4 Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a dimensionality reduction method, corresponding to a probabilistic extension of Self-Organizing Maps[19], which allows visualizing the data distribution on a 2-dimensional map. A more detailed description of GTM underlying algorithms can be found elsewhere [9]. Briefly: a squared grid of nodes is generated by inserting a flexible 2D manifold, into the initial high-dimensional space in which items occupy specific points defined by their attribute (descriptor vectors). The manifold is deformed in order to match (to approach) a maximum of these "frame" items, then it is flattened out with the above-mentioned squared grid of points, defining the latent space. The Global dataset (Figure 1, step 2) has been used to train the GTM model (i.e. to train the manifold). Genetic algorithm [20] was employed for selecting the best ISIDA descriptor space and the characteristic parameters of the GTM, as described in the next paragraph.

Once the 2D map is created, any property (here, density or class assignment) can be added as a 3rd axis forming a property landscape[9]. Here two types of landscapes are considered: (*i*) density landscape which assigns a color code depending on the number of compounds populating a given GTM node; (*ii*) class landscapes in which the color code is assigned according to the repartition of C/nC compounds.

The naming "Universal Map" refers to the GTM model showing the best overall performances for all 11 considered endpoints, see section "2.7 Ranking the performances of GTM models".

2.5 GTM's applicability domain

The "fragment control" assessment[2] is employed as method to define a model's Applicability Domain: if the test molecule has a fragment not present in the GTM's training set (i.e. the Global dataset), it is considered to be outside of applicability domain of the model. The profiling on the REACH-INV compounds has been performed only on those compounds which fulfilled the Applicability Domain requirement. Additionally, GTM models have a build-in applicability domain: any compound projected in empty regions of the chemical space are discarded. Empty regions appear as white areas on GTM landscapes (Figure 5).

2.6 Genetic algorithm optimization

A Genetic algorithm-driven optimization (Figure 1, steps 3, 4 and 5)[20] was run in order to choose the most appropriate descriptor space and GTM hyperparameters, such as the number of radial basis function centers (*m*), the radial basis functions width (*w*), the dimension of the map (*k*) and the regularization coefficient (*l*). All these parameters are encoded by a chromosome, i.e. a vector of settings needed to build a given map. The genetic algorithm therefore builds hundreds of maps based on different chromosomes. Maps' performance has been evaluated by cross-validated GTM-driven classification models for each of the considered 11 endpoints. As we are dealing with two-classes, Sensitivity (Sn), Specificity (Sp) and Balanced Accuracy (BA) parameters were computed (Table S2). The latter (BA) was chosen as scoring function for the optimization process. For each endpoint, a property-specific cross-validated BA value is returned. To obtain a more robust evaluation, this cross-validation procedure is repeated three times, and the map fitness score is based on the mean of all set-specific BAs. In the end, each map has an associated vector of 11 BA values, one per endpoint. The genetic algorithm uses the concept of the

Pareto-front optimization to select the optimal set of nondominated solutions[21] and filtering redundant configurations. The procedure resulted in 28 maps with a unique set-up of descriptor spaces and GTM hyperparameters (Table S1). Some of these maps can perform well on some tasks and poorly on others. This poses two challenges: (*i*) select the best possible map on all tasks (paragraph 2.7) and (*ii*) assembling an efficient ensemble model (paragraph 2.9). To rationalize these choices, GTM models candidates went through a Rank Product[22] scoring procedure.

2.7 Ranking the performances of GTM models

The genetic algorithm run identifies a set of different manifolds, each based on a particular type of ISIDA descriptors (28 descriptor spaces were considered). Since multiple endpoints can be predicted using the same manifold, the best "all-around" map can be selected using a score measuring overall performance of the considered GTM-based classification models. In principle, a mean value for the ensemble of balanced accuracies of individual models could be used for this purpose.

A more robust score (Figure 1, step 6) can be obtained by using the "Rank Product" scoring method[22]: (*i*) for the given property, the considered manifolds are sorted according to their BA values; (*ii*) a score *S*, equal to the rank in the sorted list, is assigned starting from the top manifold; (*iii*) this process of sorting and scoring assignment is repeated for each property; (*iv*) the overall Rank Product is calculated as the product of each property's score ($\text{Rank Product} = \prod_{i=1}^{n} S_i$); (*v*) the map having the lowest Rank Product is selected as the best, so-called "*Universal Map*" (UM), reflecting its ability to have good predictive power on all considered properties. On the other hand, the wording "*Optimal Map*" (OM) refers to the map scored by the best BA for a given property, regardless of its performances on the others. Notice that Universal Maps result from multi-task learning procedure because all 11 endpoints were used to train the GTM manifold. In comparison, Optimal Maps result from single-task learning since only one selected endpoint was used to optimize GTM parameters.

2.8 Landscapes generation

The "best" REACH Universal Map (Figure 1, steps 7 and 8) is based on the ISIDA descriptor space IIAB(2-2)[18], i.e. atom centered fragments with a radius of two. The following GTM hyperparameters were obtained in genetic algorithm optimization: k = 20*20; m = 7*7; w = 1.2; l = 0.02884. This map was used to visualize class landscapes and data distribution on Figures 3-5 and 8-9. Class landscapes built on top 5 (out of 28 considered) manifolds were used for the REACH-INV profiling.

2.9 Consensus of GTM models

Each individual map can be used to perform predictions on all 11 properties. However, the predictive performance can be improved using a consensus model combining ensemble of individual predictors[23]. To this end, the universal maps (UMs) obtained as described in section 2.7 were combined in consensus one by one following the order defined by the Rank Product. The consensus result has been calculated by majority voting of the considered UMs. When the repartition of C/nC votes was between 40 – 60 % (i.e. close to the random threshold) the result did not contribute to the performances. We observed that the performances of the consensus model are already stable after adding five maps (Figure S1).

2.10 Benchmarking of GTM with other machine learning methods

Following the strategy described in our previous works[2–4], the binary consensus classification models were generated on particular training sets (one per endpoint) extracted from the Global dataset. Each consensus model is an ensemble of several individual models, based on a different descriptor spaces and/or machine learning algorithm (chosen among Random Forest, Support Vector Machine or Naïve Bayesian). In such a way, these consensus models are optimized in terms of descriptors and methods parameters and have been trained on the same data used for GTM modelling. Therefore, 3-fold CV performances have been computed and can be directly compared with those of GTM (see section 3.6). Moreover, these models have been used to replace the missing experimental values, in order to complete the substance's environmental fate and ecotoxicological profile (see section 3.4). Table S3 report the 11 consensus models and their performances. Several other tools (such as EPI Suite or VEGA[2–4]) are relevant for property prediction under REACH. However, they have not been considered for benchmarking for the following reasons:

- Not all of them can predict the compounds on all the 11 considered properties.
- Their output is not directly comparable to the GTM models, requiring additional post-processing (e.g. a continuous value needs to be categorized).

3 Results

3.1 Overview of the curated datasets

Figure 2 depicts: (a) the repartition of available experimental data for 11 endpoints; and (b) the repartition of C/nC class for each particular endpoint. The Global dataset counts a total of 17762 unique compounds listing at least one experimental measurement for at least one property, for a total of 29433 data points.

The RatTox is the endpoint for which the amount of data was the highest, accounting for almost 50 % of the Global dataset. The datasets on environmental persistence (SedP, SoilP and WatP) were the smallest, with only few hundreds of compounds. There was a strong overlap of compounds between the acute aquatic toxicity datasets (AlgaeTox, DaphniaTox and FishTox): this is understandable, as for higher tonnage bands compounds shall be evaluated on all the three endpoints together to provide a complete acute aquatic toxicity evaluation. On the other hand, there was limited overlap between ready biodegradability (RB) and bioconcentration (BCF) and environmental persistence datasets. Normally, RB assays are conducted at the beginning of the registration process as, if the substance is demonstrated to be rapidly degraded in the environment, other endpoints do not need to be evaluated, and experimental testing can be therefore waived.

The C/nC class repartition varies as a function of endpoint: for endocrine disruption-related properties (AR/ER binding) the number of C compounds (i.e. binders) is rather small, which is a typical situation for such biological targets.

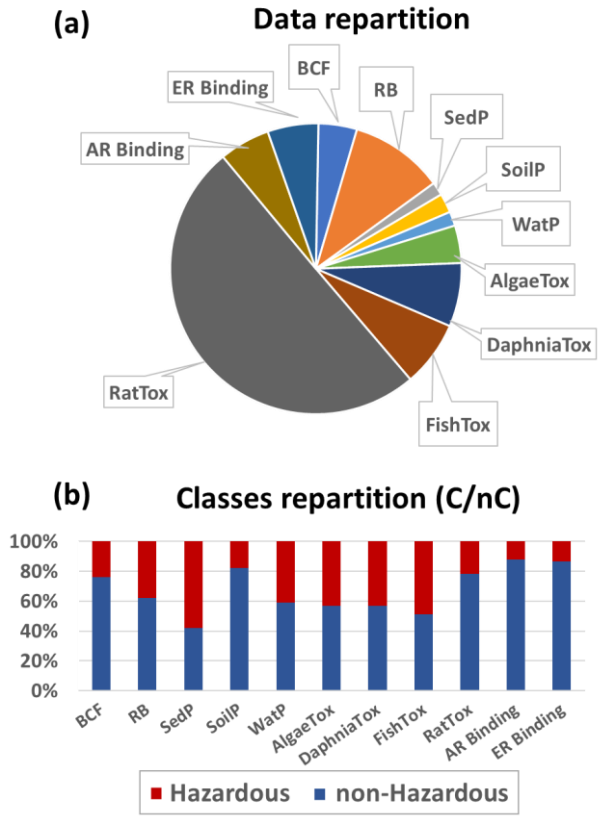

Figure 2. Datasets overview. (a) repartition of available experimental data for 11 endpoints; (b) Concern (C) and non-Concern (nC) classes repartition.

3.2 Density landscape: Chemical space qualitative analysis

- Figure 3 depicts the density landscape of the Global dataset. The colormap emphasizes how compounds are distributed in the chemical space, identifying low- and high- densely populated regions. The chemical space is characterized according to the absolute density of datapoints in a given node, which indicates the number of compounds populating that position. The density ranges from 0 (white areas) up to 40 (yellow areas) compounds. Zone 1 is populated by aromatic and aliphatic halogenated substances, for instance belonging to the chemical family of polychlorinated biphenyl.
- Zones 2a, 2b and 2c include fluorinated compounds. However, several fluorinated molecules are also found in a central area of the map delimited by the black dashed rectangles. This zone is populated by small molecules counting less than five atoms.
- Zones 3a and 3b incorporate aliphatic and aromatic compounds mainly with the ester and ether functional groups. Zone 3a contain more aromatic molecules compared to 3b. Molecules providing both functional groups are located between these two areas.
- Zones 4a and 4b agglomerate nitrogen-containing compounds. Both compounds located in 4a and 4b are characterized by the presence of nitro-containing functional groups. The structural differences between these two zones are mainly related to: (i) the presence of the diazene (N=N) moiety (absent in compounds located in 4a) and (ii) the number of hydroxyl and ketone groups (common in 4a but almost absent in 4b). Table S4 (a and b) in SI reports a list of the top-10 most similar chemicals to compounds 4a and 4b depicted in Figure 3.
- The low-density regions (e.g. two molecules identified by black dots) are populated by molecules containing “rare” chemotypes which are noticeably different from other compounds from the Global set.

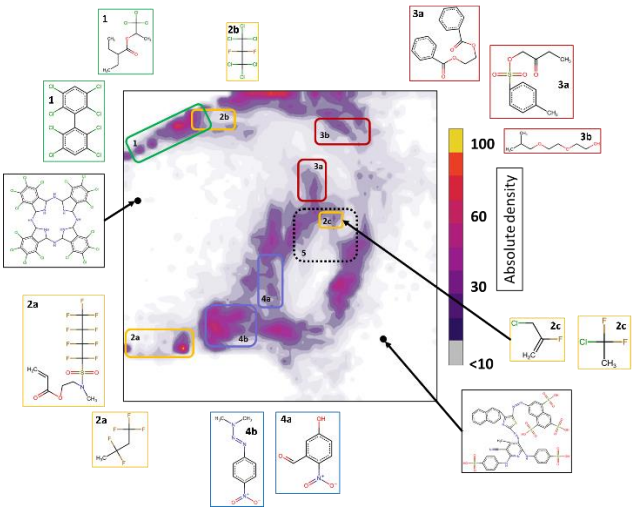

Figure 3. Density landscape or the Global dataset. The color scale refers to the number of compounds populating a given zone. Colored rectangles delimit map regions referred in the text. Representative structures populating selected zones of the map are shown.

3.3 Chemical space comparison: Global dataset vs. REACH-INV

The inventory of substances registered under the REACH-regulation (REACH-INV) has been projected on the manifold trained on the Global dataset. The likelihoods distributions of REACH-INV and the Global dataset are similar (see SI, Figure S2) which means that the manifold well describes the REACH-INV compounds. Figure 4 compares data distribution of these two databases. Blue and red colors refer to zones uniquely populated by Global dataset and REACH-INV compounds, respectively whereas intermediate colors indicate mixed regions populated by compounds from both databases. A total of 5137 out of 11951 REACH-INV compounds (43 %) are overlapping with the Global dataset. On the other hand, 12624 out of 17992 Global dataset compounds (70 %) are new to the REACH-INV. Even though the Global dataset was able to accommodate a large portion of the REACH-INV chemical space, several areas uniquely populated by REACH-INV compounds were found, indicating that the Global dataset was lacking important chemotypes: long aliphatic chain (CAS 416-630-8) and highly sulphonated compounds (in particular, belonging to the chemical class of dyes as, for instance, CAS 16470-24-9) were under-sampled in the Global dataset. The REACH-INV has also some unique chemotypes concerning perfluorinated compounds (e.g. CAS 88992-45-4).

This suggests that the applicability domain of QSARs based on public data may not include some REACH related compounds issued from the industry. This observation is consistent with our earlier studies [2,4] concerning weak performances of existing models applied to compounds of industrial context.

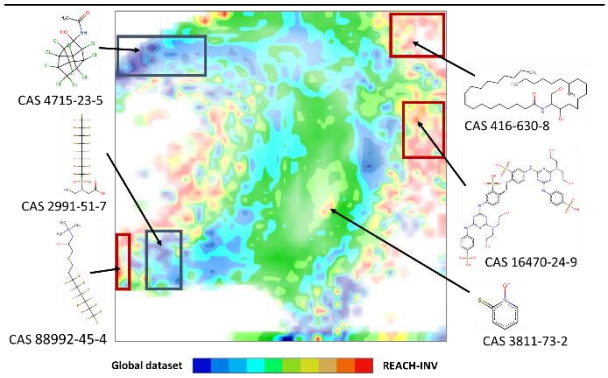

Figure 4. Global dataset and REACH-INV chemical space comparison. Blue regions are mainly populated by Global dataset compounds; red ones by the REACH-INV compounds; intermediate colors by compounds belonging to both databases. White areas display unpopulated regions.

3.4 Global dataset class landscapes

Figure 5 shows the class landscapes of the Global dataset for the 11 endpoints. Blue and red areas are populated, respectively, by nC and C compounds, yellow and green colors represent mixed populated areas in which both nC and C compounds are present. For several endpoints (e.g. BCF, RB, AR/ER binding) there is a clear separation of the classes, with very few mixed areas. On the other hand, the RatTox landscape is the one that has the worst class separation, as reflected by its lower prediction performances (Table S5).
Figure S3(a and b) in SI depicts an alternative version of the RatTox landscape, using the different cuf-off value of 2000 mg/kg to discriminate C/nC classes. White areas correspond to unpopulated regions which size is related to the absence of experimental data. Thus, for the series of RatTox, BCF, SoilP, WatP and SedP endpoints sorted according to the reduction of the dataset size (Table 1), the white areas on the related landscapes increase in the same order.

Ensemble of landscapes is a convenient tool of compounds profiling. As an example, Figure 5 considers two compounds: Chlordecone (CAS 143-50-0) and p-Phenylenediamine (CAS 106-50-3) depicted by star-shaped and circle-shaped black dots, respectively. In agreement with experimental data, Chlordecone is classified by the landscapes as C for 7 out of 11 endpoints (AR/ER binding, DaphniaTox, BCF, RatTox, SedP, WatP); while p-Phenylenediamine for 6 our of 11 endpoints (ER binding, AlgaeTox, DaphniaTox, FishTox, RatTox, RB). Chlordecone was used as an insecticide but was banned due to its deleterious effects on the environment, mainly related to persistence. Aniline derivates such as p-Phenylenediamine are of concern due to their acute toxicity effects on aquatic life organisms[24–26].

As mentioned above, the white areas indicate the absence of experimental data for a given zone of the chemical space. Generation of new data may help to fill such empty zones and, hence, to extend the area covered by the landscape. For this purpose, *in silico* predictions obtained with the help of machine-learning models described in Section 2.10, have been used instead of experimental data. Notice that the predicted values outside the applicability domain of the models were discarded. This led to the decrease of the percentage of missing data over all the 11 endpoints from 89 % to 32 %. As one may see from Figure 5 (right side), the updated landscapes cover much larger area than the initial ones. The biggest improvement is observed for the smallest datasets BCF, SedP, SoilP and WatP.

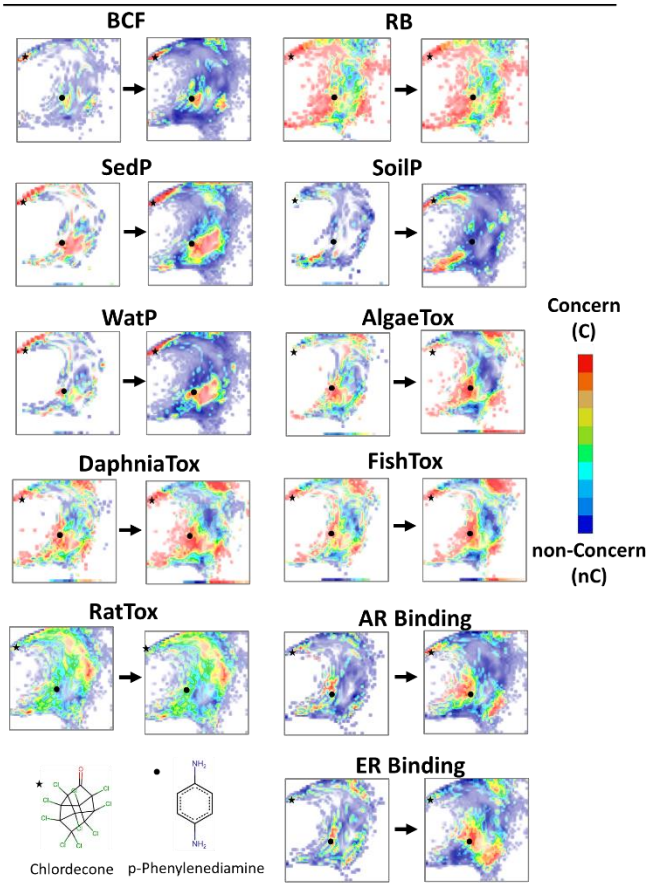

Figure 5. Class landscapes of the Global dataset for the 11 endpoints. Blue and red regions are populated by nC and C compounds, respectively. White areas correspond to unpopulated regions. Black stars and dots represent projects of two example compounds. The landscapes of the right side of the black arrows were recomputed after replacing missing experimental values with predicted ones obtained with the help of consensus classification models reported in Section 2.9. All images in a large scale are available in SI.

3.5 GTM for property prediction: single-task learning performances

Figure 6 and Table S5 reports cross-validation Balanced Accuracies for each Optimal Map (OM). Related GTM-based classification models demonstrates from moderate to satisfactory performances, with BAs ranging from 0.66 (RatTox) to 0.81 (BCF). The worst performing endpoints are ER binding and RatTox. The noticeable difference in BA between AR/ER binding is quite surprising, as these datasets show a high overlap of compounds. It seems that ER data are more noisy: these results are consistent with those reported by the CERAPP/CoMPARA workgroups [12,13] (Table S3). For all endpoints except SedP, the sensitivity (i.e. detection of truly C compounds) is always higher than specificity. The RatTox shows the largest difference between these two metrics (Sensitivity = 0.83 and Specificity = 0.52), indicating that the RatTox map frequently misclassifies nC compared to C compounds.

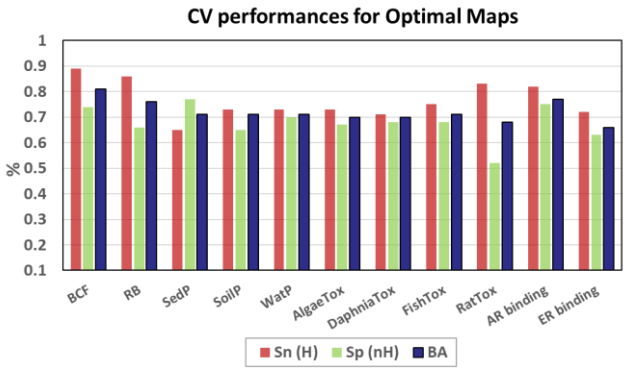

Figure 6. Optimal Maps performances.

3.6 Benchmarking studies.

Figure 7 and Table S6 reports cross-validated balanced accuracies for both the best UMs and for the Consensus of five selected UMs. Besides, performances were benchmarked against the 11 machine-learning models described in Section 2.10. As expected, for a given endpoint, the best Universal Maps perform worse than related Optimal Maps. On the other hand, Consensus of the top five Universal Maps prides with similar to OMs results. Compared to the models obtained with popular machine-learning methods (see Section 2.10), GTM displays slightly worse performances.

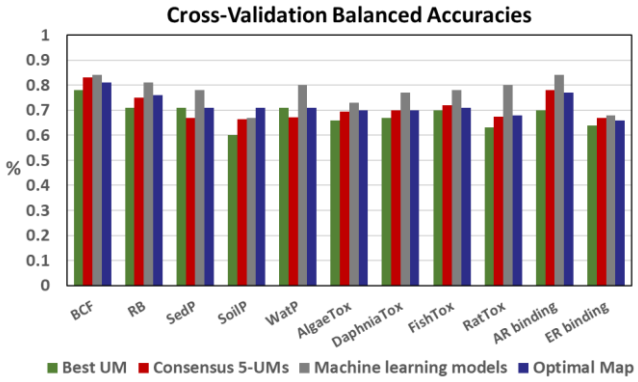

Figure 7. Results of benchmarking studies for the 11 studied endpoints. Best UM = best-performing Universal Map; Consensus 5-UMs = the top five performing Universal Maps have been ensembled in consensus; Machine-Learning models described in Section 2.10; Optimal Map is the best performing GTM model for a given endpoint.

3.7 Profiling the REACH-INV dataset

The REACH-INV dataset has been profiled by the best universal map: predictions on all properties are available on Zenodo: 10.5281/zenodo.3872735. A total of 72 % of REACH-INV compounds fell inside the applicability domain of GTM-based models according to fragment control approach. Table 2 reports the number of compounds predicted as C or nC for the given property. RB is the only property for which most of the compounds were classified as C (not readily biodegradable). This was expected as ready biodegradability assays are very stringent first-tier experiments that generally underestimate the biodegradation potential. The aquatic toxicity endpoints have a similar behavior, with roughly half of the compounds predicted as C. We also found that several chemical families, such as quaternary ammonium salts, long chain alcohols and quinones were predicted toxic for all the three trophic levels (Algae, Daphnia and Fish). Only a limited amount of compounds (3-7 %) were classified of concern (C) for bioconcentration, rat toxicity and environmental persistence.

Table 2. REACH-INV GTM profiling results for the 11 endpoints predicted with the help of Universal Maps..

Endpoint	nC	C	% (C)
BCF	11651	300	3
RB	3526	8425	70
SedP	9330	2621	22
SoilP	11581	370	3
WatP	11077	874	7
AlgaeTox	6972	4979	42
DaphniaTox	6462	5489	46
FishTox	6400	5551	47
RatTox	11410	541	5
AR binding	10591	1360	11
ER binding	9984	1967	17

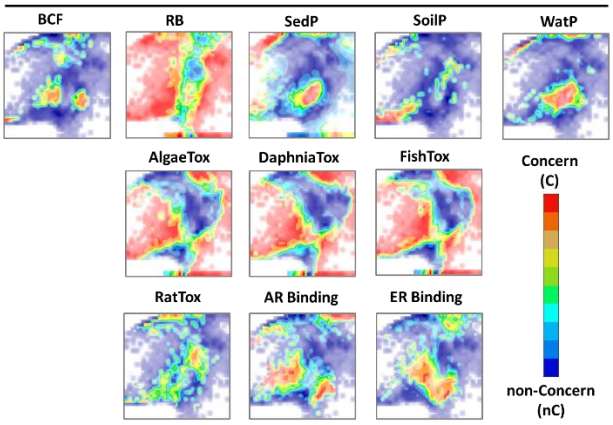

Figure 8. REACH-INV Class landscapes for the 11 endpoints predicted with the help of Universal Maps. All images in a large scale are available in SI.

To facilitate the chemical space analysis, the in-house "constrained screening" tool[27] has been used. It allows to superpose the class landscapes and, in such a way, to isolate regions of the chemical space populated by compounds possessing a given (eco)toxicological profile. Below, we provide with 3 examples of specific queries considering either all endpoints together, or only selected endpoints (e.g. aquatic toxicity or environmental fate endpoints).

Figure 9 (a,b,c) depicts the result of the superposition process, where the color code refers to the Overall Concern Score (OCS) cumulating the C labels for all considered endpoints (here, OCS varies from 0 to 11) .

In the first case (Figure 9a), all 11 landscapes shown on Figure 8 were used. No regions of the chemical space populated by the compounds labelled C with respect to endpoints were detected, the maximal OCS value was eight. Compounds located in these regions normally show acute toxicity to the aquatic environment, are not expected to rapidly degrade and in several instances exhibit acute oral toxicity. Some examples include compounds belonging the polychlorinated biphenyls (e.g. CAS 1514-82-5) which have been banned due to their deleterious effect on the environment and biota.

In the second case (Figure 9b), we focused on environmental fate landscapes (BCF and RB) aiming to extract compounds that could persist and bioconcentrate in the food chain. A consistent number of compounds belonging to the chemical family of perfluorinated compounds (e.g. CAS 118-69-4) were identified and for most of them data is scarce, especially on the bioconcentration endpoint.

In the third case (Figure 9c) the acute aquatic endpoints were considered. Chloro- and nitro- phenols (e.g., CAS 87-86-5 and 38668-48-3), some biphenyls (e.g., CAS 38668-48-3) and quaternary ammonium salts (e.g., CAS 1563-67-3) have been identified as of potential concern.

This approach is quite flexible, and it can be relevant in different regulatory contexts, for instance in order to identify Substances of Very High Concern (SVHC) under REACH. A compound is defined as SVHC if: (i) it is CMR (carcinogenic, mutagenic or toxic for reproduction); or (ii) it is PBT (persistent, bioaccumulable and toxic) or vPvB (very persistent and very bioaccumulable). In its current state, the use of reported here maps for SVHC profiling is limited, as it does not consider CMR properties nor chronic aquatic toxicity. Considered here 11 endpoints represent only a fraction of a full REACH-registration dossier, which opens a perspective to build activity landscapes for some additional endpoints. It should also be noticed that the cut-off values used to discriminate C/nC compounds should be user-controlled, in order to allow the regulator expert to adjust them depending on the context.

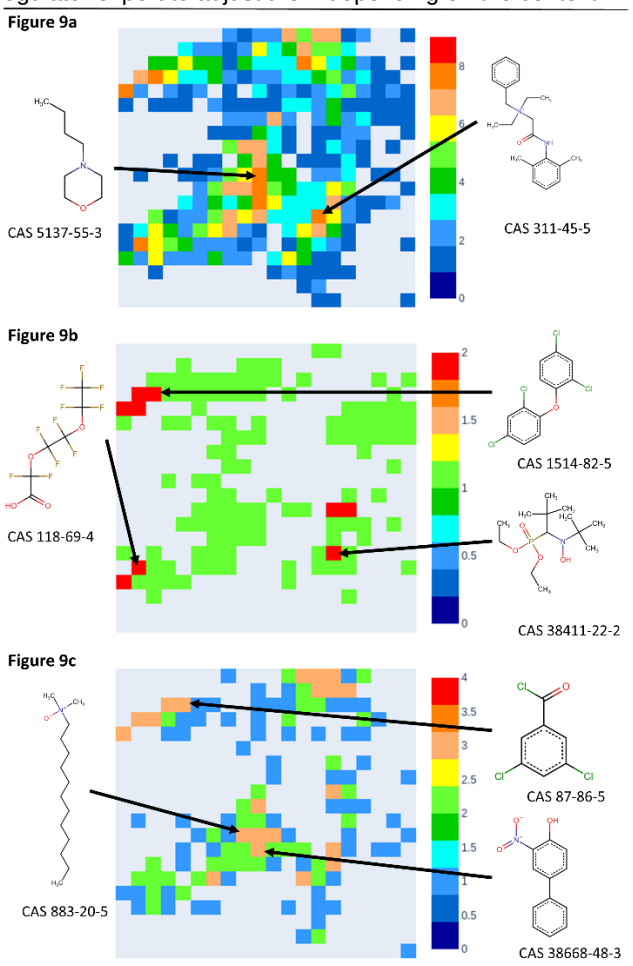

Figure 9 (a,b,c). REACH-INV profiling maps resulted from superposition of several class landscapes shown on Figure 8. Different superposition scenarios have been considered: (a) landscapes of all 11 endpoints; (b) two environmental fate endpoints; and (c) four acute aquatic toxicity endpoints. The color code refers to the Overall Concern Score (OCS) cumulating the C labels for all considered endpoints. Some compounds identified as a concern are represented along with their CAS ID.

4 Conclusions

A Global dataset on 11 toxicologically-relevant endpoints resulted from the merging of multiple public data sources has been prepared. It contains: environmental fate and pathways endpoints (bioconcentration factor, ready biodegradability, environmental persistence in sediment, soil and water compartment), ecotoxicological endpoints (acute aquatic toxicity towards algae, daphnia and fish) and human health endpoints (oral acute toxicity to rats and androgen and estrogen receptor binding). A total of 17762 unique compounds listing, at least, one experimental measurement for, at least, one endpoint have been collected. Binary Concern (C) and non-Concern (nC) labels were assigned according to relevant thresholds. The Generative Topographic Mapping approach has been employed as method for data analysis and property prediction, generating single- and multi-task learning models.

So called REACH Universal Maps trained on the Global dataset have been generated, and their ability to classify compounds of Concern from non-Concern has been tested on all the 11 endpoints. In such a way, a given compound can be profiled on multiple properties simultaneously. The best Universal Maps display acceptable predictive performance with balanced accuracies ranging from 0.60 to 0.78, as a function of the endpoint. Assembling five best Universal Maps in a consensus improves predictive performance: balanced accuracies vary from 0.67 to 0.83. The REACH-INV dataset containing 17762 substances registered under the REACH Regulation have been profiled on the considered endpoints. Superposition of several landscapes helps to identify the zones populated by compounds of a given (eco)toxicity profile.

This work proposes a novel and unique methodology for the identification and prioritization of compounds in the context of the REACH regulation. New untested compounds can be easily profiled on several endpoints using one unique model which largely facilitates the screening process. However, as we covered only a small fraction of the properties which constitute a registration dossier. This work could be further expanded by:

-Adding more endpoints to the GTM profiler (e.g. CMR properties) to make it suitable for different regulatory frameworks, such as the identification of SVHC compounds;

- Enlarging existing datasets with “waived” REACH endpoints, in particular, the outcome (C/nC) for some endpoints can be deduced from the observation of other endpoints.

- Comparing the REACH-INV chemical space versus those of other databases, such as PubChem or ChEMBL, and annotate as much as possible those databases.

- The cut-off values to discriminate C/nC compounds could be user-defined; in such a way the GTM model can be adjusted to the context (for instance, PBT or vPvB assessment).

The Global dataset and the profiled REACH-INV are available through Zenodo: 10.5281/zenodo.3872735

References

[1] European Commission, *Off. J. Eur. Union.* **2007**, *50*, 1–281.

[2] F. Lunghini, G. Marcou, P. Azam, R. Patoux, M. H. Enrici, F. Bonachera, D. Horvath, A. Varnek, *SAR QSAR Environ. Res.* **2019**, *30*, 507–524.

[3] F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert, A. Varnek, *SAR QSAR Environ. Res.* **2019**, *30*, 879–897.

[4] F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert, A. Varnek, *SAR QSAR Environ. Res.* **2020**, *31*, 171–186.

[5] ECHA, "*Practical Guide How to Use and Report (Q)SARs",* can be found under

https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf, **2016**.

[6] OECD, "*Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models"*, can be found under *https://www.oecd.org/env/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models-9789264085442-en.htm,* **2007**.

[7] A. Golbamaki, A. Cassano, A. Lombardo, Y. Moggio, M. Colafranceschi, E. Benfenati, *SAR QSAR Environ. Res.* **2014**, *25*, 673–694.

[8] Tsakovska I, Worth A, *Bioautomation* **2009**, *13*, 151–162.

[9] N. Kireeva, I. I. Baskin, H. A. Gaspar, D. Horvath, G. Marcou, A. Varnek, *Mol. Inform.* **2012**, *31*, 301–312.

[10] R. J. Larson, C. E. Cowan, *Environ. Toxicol. Chem.* **1995**, *14*, 1433–1442.

[11] ECHA, "*Guidance on Information Requirements and Chemical Safety Assessment Chapter R.7b: Endpoint Specific Guidance"*, can be found under *https://echa.europa.eu/documents/10162/13632/information_requirements_r7b_en.pdf*, **2017**.

[12] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A. M. Richard, C. M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E. B. Wedebye, F. Grisoni, G. F. Mangiatordi, G. M. Incisivo, H. Hong, H. W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N. G. Nikolov, O. Nicolotti, P. L. Andersson, Q. Zang, R. Politi, R. D. Beger, R. Todeschini, R. Huang, S. Farag, S. A. Rosenberg, S. Slavov, X. Hu, R. S. Judson, *Environ. Health Perspect.* **2016**, *124*, 1023–1033.

[13] K. Mansouri, N. Kleinstreuer, A. M. Abdelaziz, D. Alberga, V. M. Alves, P. L. Andersson, C. H. Andrade, F. Bai, I. Balabin, D. Ballabio, E. Benfenati, B. Bhhatarai, S. Boyer, J. Chen, V. Consonni, S. Farag, D. Fourches, A. T. García-Sosa, P. Gramatica, F. Grisoni, C. M. Grulke, H. Hong, D. Horvath, X. Hu, R. Huang, N. Jeliazkova, J. Li, X. Li, H. Liu, S. Manganelli, G. F. Mangiatordi, U. Maran, G. Marcou, T. Martin, E. Muratov, D. T. Nguyen, O. Nicolotti, N. G. Nikolov, U. Norinder, E. Papa, M. Petitjean, G. Piir, P. Pogodin, V. Poroikov, X. Qiao, A. M. Richard, A. Roncaglioni, P. Ruiz, C. Rupakheti, S. Sakkiah, A. Sangion, K. W. Schramm, C. Selvaraj, I. Shah, S. Sild, L. Sun, O. Taboureau, Y. Tang, I. V. Tetko, R. Todeschini, W. Tong, D. Trisciuzzi, A. Tropsha, G. Van Den Driessche, A. Varnek, Z. Wang, E. B. Wedebye, A. J. Williams, H. Xie, A. V. Zakharov, Z. Zheng, R. S. Judson, *Environ. Health Perspect.* **2020**, DOI 10.1289/EHP5580.

[14] OECD, "e*ChemPortal: Global Portal to Information on Chemical Substances*", can be found under *https://www.echemportal.org/echemportal/index.action*, **2020**.

[15] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, *SIGKDD Explor.* **2009**, *11*, 26–31.

[16] NIH, "*The PubChem Project*", can be found under https://pubchem.ncbi.nlm.nih.gov/, **2020.**

[17] ECHA, "*ECHA Website*", can be found under https://echa.europa.eu/, **2020.**

[18] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inform.* **2010**, *29*, 855–868.

[19] T. Kohonen, *Biol. Cybern.* **1982**, 43, 59-69.

[20] D. Horvath, J. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.

[21] R. Kumar, P. Rockett, *Evol. Comput.* **2002**, *10*, 283–314.

[22] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, *FEBS Lett.* **2004**, DOI 10.1016/j.febslet.2004.07.055.

[23] A. Tropsha, *Mol. Inform.* **2010**, *29*, 476–488.

[24] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, R. A. Drummond, *Environ. Toxicol. Chem.* **2005**, *16*, 948.

[25] I. Sushko, E. Salmina, V. A. Potemkin, G. Poda, I. V. Tetko, *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316.

[26] H. J. M. Verhaar, C. J. Vanleeuwen, J. L. M. Hermens, *Chemosphere* **1992**, 25, 471-491.

[27] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2015**, *55*, 84–94.

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial st