



Comprehensive data analysis and predictive chemoinformatics models for REACH related physicochemical and (eco)toxicity properties

Filippo Lunghini

► To cite this version:

Filippo Lunghini. Comprehensive data analysis and predictive chemoinformatics models for REACH related physicochemical and (eco)toxicity properties. Cheminformatics. Université de Strasbourg, 2020. English. NNT : 2020STRAF016 . tel-03505818

HAL Id: tel-03505818

<https://theses.hal.science/tel-03505818>

Submitted on 31 Dec 2021

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

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

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

Thèse présentée par :

Filippo LUNGHINI

Soutenue le : **29 Septembre 2020**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : **Chimie / Chémoinformatique**

**Analyse exhaustive et modèles chémoinformatiques
prédictifs des données physicochimiques et
(éco)toxicologiques concernées par REACH**

THÈSE dirigée par :

M. VARNEK Alexandre

Professeur, Université de Strasbourg

M. MARCOU Gilles

Maître des Conférences, Université de Strasbourg

RAPPORTEURS :

Mme ROTUREAU Patricia

Docteur, Institut national de l'environnement industriel et
des risques

M. BUREAU Ronan

Professeur, Université de Caen

AUTRES MEMBRES DU JURY :

M. AZAM Philippe

Docteur, Solvay S.A., France

Mme PAPA Ester

Professeur, Università degli studi dell'Insubria, Italie

Abstract

This thesis concerns the modelling of several environmental fate and (eco)toxicological properties relevant under the European Union Registration, Evaluation, Authorisation and Restriction of Chemical Substances Regulation (REACH, EC No 1907/2006).

Statistical models have been generated using state-of-the-art machine learning methods, such Support Vector Machine and Random Forest and molecular descriptors. Models have been internally and externally validated following internationally recognized guidelines, especially the OECD principles. The models are designed to be used as valid alternative to experimental testing and data-gap filling under the REACH regulation.

New models possess several advantages over already existing ones: (i) noticeable larger training sets; (ii) external validation on a significant number of compounds coming from the Industrial context (Solvay portfolio); (iii) better accuracy and extended applicability domain.

The Generative Topographic Mapping approach has been used to profile the REACH-chemical space on the modelled properties: in such a way, it is possible to identify compounds with an undesirable eco-toxicological and environmental fate profile.

All the models have been implemented in the ISIDA/Predictor platform, which is freely available for the user.



Résumé

Cette thèse concerne la modélisation de propriétés environnementales et (éco)-toxicologiques pertinentes dans le cadre du règlement de l'Union Européenne sur l'enregistrement, l'évaluation, l'autorisation et la restriction des substances chimiques (REACH, CE n ° 1907/2006).

Des modèles statistiques ont été générés à l'aide de méthodes d'apprentissage automatique, telles que les Séparateurs à Vaste Marge (SVM) ou les Forêts Aléatoires (*Random Forest*), et des descripteurs moléculaires. Le pouvoir prédictif des modèles a été estimé suivant des procédures de validation interne et externe, conformément aux directives internationalement reconnues, en particulier les principes de l'OCDE. Les modèles sont conçus pour être utilisés comme une alternative crédible aux tests expérimentaux et pour compléter les données manquantes dans le cadre du règlement REACH.

Les nouveaux modèles présentent plusieurs avantages par rapport aux modèles existants: (i) ils sont construits sur des ensembles de données sensiblement plus grands; (ii) ils sont validés sur des données externes de tailles significatives composés d'exemples issus d'un contexte industriel (l'entreprise Solvay); (iii) la précision des modèles est améliorée et leurs domaines d'applicabilité sont étendus.

La technique de Cartographie Topographique Générative a été utilisée pour profiler l'espace chimique REACH sur les propriétés modélisées: de cette manière, il est possible d'identifier rapidement les composés risquant de présenter un profil éco-toxicologique et environnemental indésirable.

Tous les modèles ont été implémentés dans la plate-forme ISIDA / Predictor, qui est disponible gratuitement pour l'utilisateur.

Table of Contents

Abstract	i
Table of Contents.....	iv
I Résumé en français	1
1.1 Introduction.....	1
1.2 Résultats et discussions.....	3
1.2.1 <i>Méthodologie de la création des modèles</i>	4
1.2.2 <i>Destin environnementale : facteur de bioconcentration</i>	5
1.2.3 <i>Destin environnementale : biodégradabilité primaire</i>	5
1.2.4 <i>Destin environnementale : persistance</i>	6
1.2.5 <i>Propriétés écotoxicologiques : toxicité aiguë pour les organismes aquatiques</i>	6
1.2.6 <i>Propriétés toxicologiques : toxicité aiguë sur le rat</i>	7
1.2.7 <i>Propriétés toxicologiques : perturbateurs endocriniens</i>	7
1.2.8 <i>GTM : analyse de l'espace chimique du REACH</i>	7
1.2.9 <i>Implémentation du logiciel ISIDA/Predictor</i>	9
1.3 Conclusion générale.....	10
II Introduction.....	11
2.1 Context.....	11
2.2 OECD requirements.....	13
2.3 State of the art	14
2.4 Organization of the thesis	14
2.5 Background of Solvay.....	16
III Tools and methods	17
3.1 QSAR: background and role in regulatory context.....	17
3.2 Modelled endpoints.....	17
3.2.1 <i>Bioconcentration Factor</i>	18
3.2.2 <i>Ready Biodegradability</i>	18
3.2.3 <i>Environmental persistence in Sediment, Soil and Water</i>	18
3.2.4 <i>Aquatic short-term toxicity on Algae, Daphnia and Fish</i>	18
3.2.5 <i>Short-term toxicity on Rodent</i>	19
3.2.6 <i>Androgen and Estrogen receptor binding</i>	19
3.3 Publicly available data sources	20
3.4 Data curation.....	21
3.5 ISIDA descriptors	21

3.6	Genetic Algorithm	22
3.7	Employed machine learning methods	23
3.7.1	<i>Support Vector Machines</i>	23
3.7.2	<i>Random Forest</i>	23
3.7.3	<i>Naïve Bayesian</i>	24
3.8	Employed tools	24
3.8.1	<i>KNIME v.4.2</i>	25
3.8.2	<i>ISIDA/Fragmentor v.2019</i>	25
3.8.3	<i>WEKA v.3.9.3</i>	26
3.8.4	<i>LibSVM v.3.21</i>	26
3.8.5	<i>ISIDA/ColorAtom v.2019</i>	26
3.8.6	<i>ISIDA/Predictor platform implementation</i>	27
3.9	Publicly available tools	27
3.9.1	<i>TEST v.4.1</i>	27
3.9.2	<i>VEGA v.1.1.5</i>	28
3.9.3	<i>EPI Suite v.4.1</i>	28
3.9.4	<i>OPERA v.2.5</i>	29
3.10	<i>ISIDA/Predictor platform implementation</i>	29
IV	Results	30
4.1	Part 1 – Modelled endpoints	30
4.1.1	<i>Bioconcentration Factor</i>	30
4.1.2	<i>Ready Biodegradability</i>	52
4.1.3	<i>Environmental persistence in Sediment, Soil and Water</i>	70
4.1.4	<i>Short-term aquatic toxicity on Algae, Daphnia and Fish</i>	92
4.1.5	<i>Short-term toxicity on Rodent</i>	116
4.1.6	<i>Androgen and Estrogen receptor binding</i>	138
4.2	Part 2 – REACH Chemical space profiling with GTM	170
4.3	Part 3 – ISIDA/Predictor software implementation	183
4.3.1	<i>ISIDA/Predictor interface</i>	183
4.3.2	<i>Input file</i>	184
4.3.3	<i>Results</i>	184
4.3.4	<i>Reliability of the prediction</i>	185
V	Conclusions and Perspectives	186
5.1	Perspectives	188
	Bibliography	190
	Appendix I. KNIME workflows	207
	Appendix 1.1 – Data extraction from QSAR Toolbox	208
	Appendix 1.2 – Data extraction from Pubchem	209
	Appendix 1.3 – Structure standardization workflow	210

I Résumé en français

1.1 Introduction

En 2007, le nouveau règlement EC N° 1907/2006 concernant l'enregistrement, l'évaluation, l'autorisation et la restriction des substances chimiques (REACH) [1], est entré en vigueur en Europe. Il s'agit d'une réponse à différents problèmes concernant l'industrie chimique, tels que le manque d'information suffisante pour évaluer la dangerosité d'une substance, l'absence d'un règlement commun aux pays de l'espace économique européen, l'insuffisance des mesures pour contrôler les risques chimiques pesant sur la population et l'environnement. La réglementation REACH se pose comme une réponse à ces problèmes, en obligeant toutes les entreprises qui veulent produire et/ou importer leurs substances sur le marché européen, pour des quantités supérieures ou égales à 1 tonne, à les enregistrer. La procédure d'enregistrement consiste à produire un dossier technique de la substance renseignant un certain nombre de caractéristiques physicochimiques (e.g. solubilité dans l'eau), environnementales (e.g. persistance) et (éco)toxicologiques (e.g. toxicité aiguë). Un dossier REACH est constitué de dizaines de ces propriétés. Le nombre de propriétés qui doivent être étudiées est fonction du tonnage annuel produit ou importé de la substance : plus il est élevé, plus le nombre de propriétés à renseigner sont nombreuses et difficiles à acquérir.

Ces propriétés sont, les plus souvent, déterminées par des tests expérimentaux, qui sont aussi éthiquement discutables en raison de l'utilisation d'un grand nombre d'animaux. Afin de limiter le recours à ces tests, les directives de REACH soutiennent l'implémentation de méthodologies alternatives, en particulier in-silico, comme les modèles Relation Quantitative Structure-Activité (QSAR). Ces outils visent à remplacer au maximum des tests expérimentaux par des valeurs estimées. Ceux-ci doivent donc être validés selon des consignes précises et des protocoles proposées par des experts [2]. L'utilisation de ces méthodologies est donc importante, et il est plausible que dans un futur relativement proche les tests sur animaux seront abandonnés et remplacés par ces méthodes

alternatives (entre autres, in silico) [3]. La réglementation REACH impacte surtout le secteur industriel qui est obligée d'enregistrer de nombreuses substances. Ces dernières années ont donc vues plusieurs modèles publiés pour l'estimation des différents endpoints (propriétés mesurées) requis par REACH [4]. Toutefois, l'application de ces outils dans le domaine industrielle est souvent restreinte, pour différentes raisons [4–6]:

- Les modèles ont été conçus pour des produits sensiblement différents de ceux dont l'industrie a besoin et souvent, leurs prédictions ne sont pas pertinentes dans ce contexte : ces produits sont souvent hors du domaine d'applicabilité (AD) des modèles ;
- Les modèles n'ont pas toujours été préparés selon les consignes éditées par les autorités ;
- Ils sont souvent insuffisamment validés sur de nouvelles données expérimentales.

Cette thèse propose de répondre à ces limitations : des nouveaux modèles ont été créés et validés sur de nouvelles données expérimentales provenant d'un contexte industriel (fournis par Solvay S.A.). Notamment, beaucoup d'efforts ont été portés à la collecte et la préparation de jeux de données les plus exhaustifs possible, afin de rendre leur AD plus étendu et améliorer leur précision. Au total, 11 endpoints REACH ont été modélisés, en utilisant différentes méthodes d'apprentissage automatique (Machine Learning, ML). Pour chaque propriété, une analyse comparative avec les modèles préexistants a été effectuée, démontrant les avantages des modèles produits pour cette thèse. Entre autres, ces modèles ont la particularité d'avoir été validés régulièrement sur des données industrielles inédites, qui permettent de donner une estimation réaliste de leur fiabilité. Tous les modèles développés sont désormais gratuitement disponibles sur une plateforme en ligne (appelée ISIDA/Predictor) [7], où les utilisateurs peuvent prédire les valeurs de ces propriétés pour leurs substances.

Les résultats de la thèse sont divisés en trois parties : la 1ere concerne la création des modèles [8]; la 2eme décrit l'utilisation de la méthode Generative Topographic Mapping (GTM) pour l'analyse et le criblage (screening) de l'espace chimique ; le 3eme rapporte l'implémentation des modèles et outils publiés dans la plateforme ISIDA/Predictor.

1.2 Résultats et discussions

Le paragraphe 2.1 décrit brièvement la méthodologie appliquée pour la création des modèles. Les paragraphes 2.2 à 2.7 résument les résultats de la modélisation des 11 propriétés, groupés selon les articles scientifiques produits qui les concernent. Le paragraphe 2.8 décrit l'application de la méthode GTM pour le criblage de l'espace chimique des molécules concernés par REACH. Le paragraphe 2.9 décrit l'implémentation de ces modèles dans la plate-forme *ISIDA/Predictor*. Les résultats sur les performances des modèles ont été groupés dans le Tableau 1.

Tableau 1. Sommaire des performances des modèles pour la propriété indiquée.

Type	Endpoint (acronyme)	Tr. set	Val. interne		Val. externe		Analyse comparative (RMSE)	DC [%]
			R ²	RMSE	R ²	RMSE		
REG	Bioconcentration (BCF)	1263	0.75	0.71	0.77	0.55	0.59	78 (25/31)
	Toxicité aiguë Algae (AlgaTox)	1231	0.61	0.69	0.48	1.07	0.99	72 (179/228)
	Toxicité aiguë Daphnia (DaphniaTox)	2083	0.67	0.78	0.58	0.93	1.03	76 (174/249)
	Toxicité aiguë Poisson (FishTox)	2152	0.67	0.73	0.54	0.97	1.09	67 (129/193)
	Toxicité aiguë rat (RodentTox)	11191	0.78	0.55	0.6	0.47	0.61	94 (186/197)
Type	Endpoint (acronyme)	Tr. set	Val. interne		Val. externe		Analyse comparative (BA)	DC [%]
			BA		BA			
CLS	Biodégradabilité (RB)	3069	0.81		0.75		0.71	85 (307/362)
	Persistance sédiment (SedP)	436	0.81		0.91		0.52	77 (101/131)
	Persistance sol (SoilP)	630	0.71		0.76		0.62	76 (693/909)
	Persistance aqueux (WatP)	466	0.80		0.77		0.57	91 (128/140)
	Récepteurs androgènes (AR binding)	1661	0.84		0.72		0.73	85 (3320/3882)
	Recepteurs estrogènes (ER binding)	1661	0.68		0.60		0.59	76 (4409/5795)

REG ou CLS = régression ou classification ; R² = coefficient de détermination ; RMSE = racine de l'erreur quadratique moyenne ; BA = précision balancée ; Benchmarking = performance du meilleur modelé existant ; DC = couverture de données, i.e. rapport entre le nombre des molécules industriel dans l'AD et le nombre total ; R² et BA peuvent être

utilisées pour résumer les performances : la première varie entre 0 (aucun modèle) et 1 (modèle idéal) ; la deuxième entre 0.5 (aucun modèle) et 1.0 (un modèle parfait).

1.2.1 Méthodologie de la création des modèles

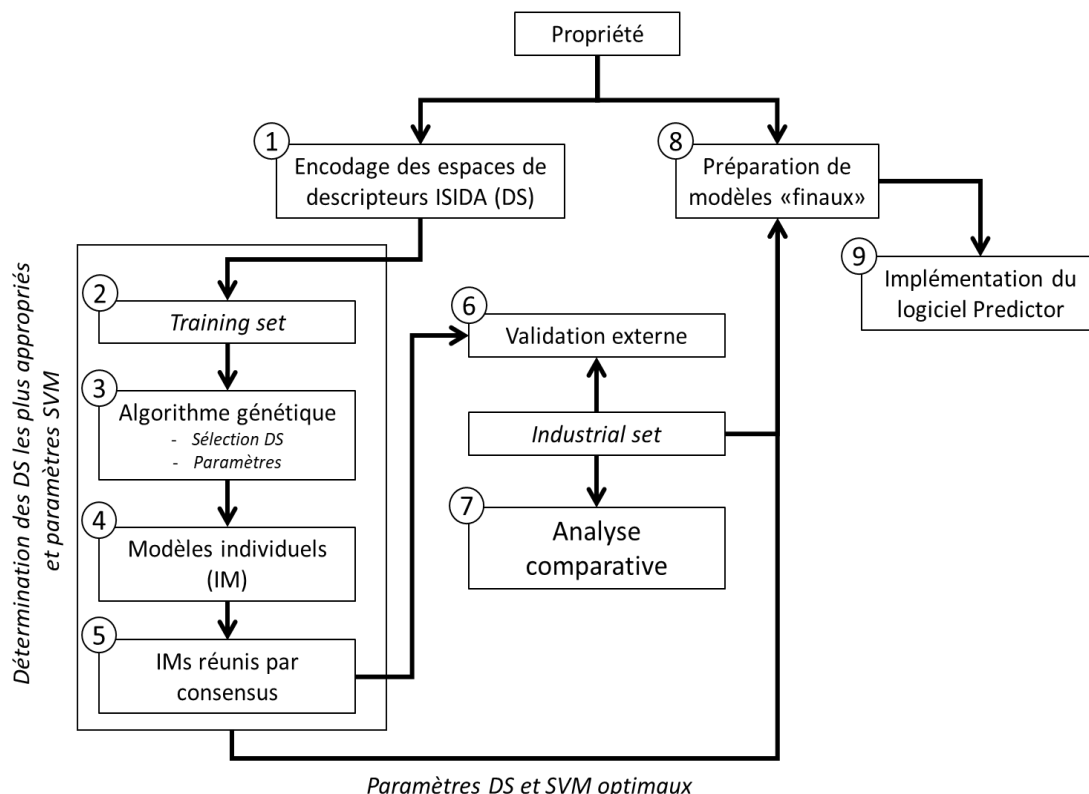
Un modèle QSAR est un processus par lequel une structure chimique est corrélée avec un effet physicochimique ou biologique bien déterminé. L'expression mathématique obtenue peut alors être utilisée comme moyen prédictif de l'effet de nouvelles molécules. La relation générale d'un QSAR est de la forme : $effet = f(\text{descripteurs moléculaires})$. Les *descripteurs moléculaires* sont des nombres qui permettent de décrire dans des termes mathématiques la structure d'une molécule. Dans cette thèse, les descripteurs ISIDA sont utilisés ; ils comptent dans une molécule, le nombre de fois qu'apparaît des motifs, i.e. des fragments moléculaires générés en « coupant » la molécule en différentes parties. La fonction f peut être déterminée à l'aide de différents algorithmes d'apprentissage automatique : les algorithmes principaux utilisés ici sont les Forêts Aléatoires (*Random Forest*, RF) et les Machines à Vecteur Support (*Support Vector Machine*, SVM).

La Figure 1 schématise la procédure suivie pour la création des modèles, appliquée pour chaque propriété. Ces étapes sont ici brièvement décrites : (1) les molécules sont encodées avec les descripteurs ISIDA ; (2) le jeu d'entraînement (*training set*), incluant les molécules utilisées pour entraîner le modèle, est assemblé ; (3) un algorithme génétique est utilisé pour optimiser le choix des meilleurs descripteurs et paramètres des méthodes utilisées ; (4) plusieurs modèles individuels (IM) sont entraînés puis (5) assemblés en un consensus qui combine les prédictions des différents IMs ; (6) les données industrielles sont utilisées pour une validation externe du modèle ; (7) une analyse comparative est effectuée pour comparer les nouveaux modèles aux préexistants (8) après leur validation, toutes les données disponibles sont assemblées pour mettre à jour les modèles ; (9) cette version finale est implémentée dans l'outil *ISIDA/Predictor*.

Les modèles créés peuvent être divisés en deux catégories : régression (la sortie est un nombre réel) et classification (la sortie est une valeur discrète, i.e. une classe). Pour simplifier, les performances des modèles de régression peuvent être résumées par le paramètre R^2 (coefficient de détermination), qui quantifie la variance expliquée par le

modèle. Ce paramètre varie entre 0 (aucun modèle) et 1 (modèle idéal). Pour les modèles de classification, la précision balancée (BA) est utilisée : elle varie entre 0.5 (aucun modèle) et 1.0 (un modèle parfait).

Figure 1. Étapes principales pour la création, validation et implémentation des modèles.



1.2.2 Destin environnementale : facteur de bioconcentration

Le *BioConcentration Factor* (BCF) estime la tendance d'une substance à se concentrer dans les tissus d'un organisme vivant. Pendant une analyse comparative avec les modèles existants, ce modèle a montré un bon équilibre entre précision et domaine d'applicabilité [4].

1.2.3 Destin environnementale : biodégradabilité primaire

Le test de *Ready Biodegradability* (RB) est extrêmement utilisé parce qu'il donne une première évaluation sur la dégradation de la substance. Si le test est positif le registrant n'est pas obligé d'effectuer certains tests supplémentaire (e.g. toxicité chronique). Aussi,

l'avantage de ce modèle est d'avoir un jeu d'entraînement bien plus large que les modèles préexistants (plus du double). Ce modèle a démontré la fiabilité de ses prédictions dans un contexte industriel [5].

1.2.4 Destin environnementale : persistance

Les propriétés de persistance dans les compartiments environnementaux, sédiments (SedP), sol (SoilP) et aquatique (WatP), peuvent être considérées comme la suite expérimentale au test RB si le résultat est négatif, car ils vont déterminer précisément le temps de dégradation dans un compartiment environnemental. Ces modèles ont le jeu d'entraînement le plus petit d'entre tous, ce qui est significatif de la carence des données dans les sources publiques. Donc leur applicabilité peut être limitée dans un contexte industriel. Pour le modèle SoilP, il est intéressant de remarquer que l'aptitude du modèle à détecter les substances les plus persistantes (Sensitivité = 0.50) est très basse : ce résultat est représentatif de l'état de l'art et reflète la variabilité des mesures expérimentales, ce qui influence négativement les modèles.

1.2.5 Propriétés écotoxicologiques : toxicité aiguë pour les organismes aquatiques

Ces propriétés sont essentielles dans un dossier REACH car elles caractérisent la toxicité d'une substance sur trois espèces considérées comme représentatives de l'écosystème aquatique. Chaque dossier REACH doit au moins inclure les résultats concernant deux espèces sur trois. Ces modèles – Algues (AlgaeTox), Crustacés (DaphniaTox) et Poissons (FishTox) – ont des performances acceptables en validation interne ($R^2 = 0.61 - 0.67$). Toutefois, les performances se dégradent ($R^2 = 0.48 - 0.58$) sur les données industrielles. Cette différence s'explique par une surestimation importante de la toxicité des molécules de bas poids moléculaire ou possédant moins de quatre atomes. Ce problème est imputé à deux facteurs principaux : (1) les résultats de toxicité sur ce genre de substances présentent une variabilité extrêmement élevée (jusqu'à deux ordres de grandeur) ; (2) leur toxicité peut être due par une réactivité spécifique une fois dans l'organisme qui se refléterait mal dans leur structure chimique. Toutefois ces résultats sont comparables à ceux de l'état de l'art.

1.2.6 Propriétés toxicologiques : toxicité aiguë sur le rat

La détermination de la toxicité aiguë sur le rat est obligatoire dans tous dossier REACH ce qui demande l'utilisation d'un grand nombre d'animaux. Aussi, une initiative américaine appelée NICEATM (*National Toxicology Programme Interagency Centre for the Evaluation of Alternative Toxicological Methods*) [9] a organisé une compétition à un niveau international pour générer des modèles prédictifs sur cette propriété. Le modèle créé dans cette thèse a été classé en troisième position (sur plus de 20 participants) en termes de précision. Par la suite, le modèle a été amélioré grâce à l'introduction de nouvelles données et testé sur les données industrielles, démontrant la qualité de ses prédictions [6].

1.2.7 Propriétés toxicologiques : perturbateurs endocriniens

Des directives sur la détermination du potentiel d'une molécule d'être un perturbateur endocrinien n'ont été publiées que récemment [10]. Ici on s'intéresse à la capacité des molécules à interagir avec des récepteurs clés du système endocrinien : les récepteurs androgènes (AR) et estrogènes (ER). Les modèles développés représentent donc une nouveauté dans la situation actuelle. Toutefois, il faut remarquer que ces modèles ont une capacité limitée à identifier certaines substances actives sur ces récepteurs ($\text{Sensitivité}_{\text{AR}} = 0.49$, $\text{Sensitivité}_{\text{ER}} = 0.34$) : ces limites ont pour origine une variabilité très élevée entre les sources des données. Ici encore, les résultats obtenus sont parfaitement comparables à ceux des compétitions international [11,12], qui ont été intégrées à nos source de données.

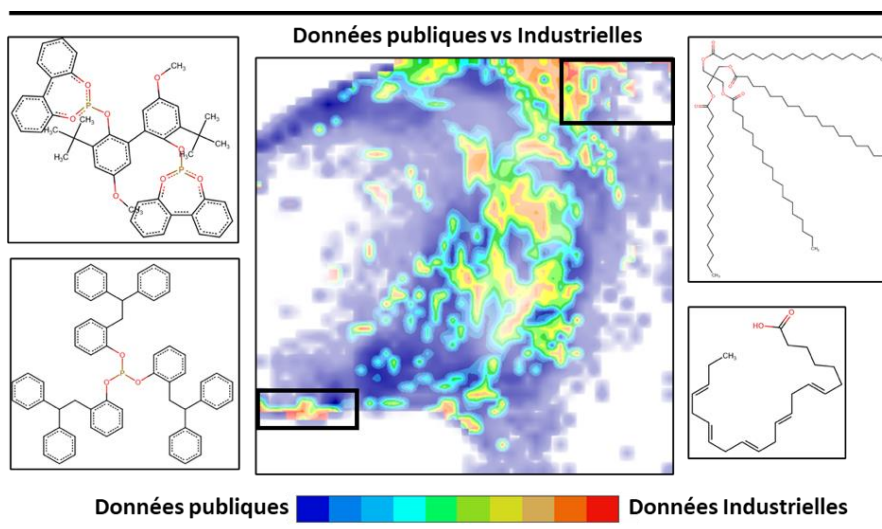
1.2.8 GTM : analyse de l'espace chimique du REACH

La multitude des propriétés à évaluer dans le contexte de REACH implique la nécessité d'avoir à disposition un modèle pour chacun d'eux (approche dite *single-task*). Pourtant, un modèle peut être préparé pour prédire, pour une substance donnée, plusieurs propriétés simultanément (approche *multi-task*). La méthodologie GTM est une stratégie de cartographie basée sur un modèle probabiliste, qui peut être utilisée pour analyser de grandes quantités de données mais aussi pour proposer des modèles prédictifs [8]. Elle produit une carte 2D représentant l'espace chimique concerné, facilitant donc une analyse de son contenu.

Plusieurs fois les limites des AD des modèles QSAR ont été évoqués et mis en relation avec les performances et l'utilité des modèles existants quand ils sont appliqués dans le domaine industriel. La GTM a été utilisée pour illustrer ces observations : l'espace chimique de toutes les données provenant de sources publiques (utilisées pour générer les modèles) a été comparé à l'espace chimique industriel (les données de validation externe). La carte GTM obtenue est illustrée en Figure 2. Plus de 18000 molécules ont contribué à caractériser cet espace chimique. Les zones rouges (soulignées par les rectangles) indiquent que les molécules qui y sont localisées appartiennent majoritairement au contexte industriel. Cela signifie que les modèles existants, qui sont développés exclusivement sur des données publiques, ne peuvent pas être appliqués sans restriction sur cet espace chimique pertinent pour l'industrie : leur AD doit être pris en compte. Toutefois, après l'inclusion des données industrielles (étape 8 en Figure 1), qui est une spécificité des nouveaux modèles produits pendant cet thèse, l'AD couvre mieux ce domaine particulier de l'espace chimique.

De plus les cartes peuvent être utilisées comme modèle prédictif : une seule et même carte peut être colorée pour plusieurs propriétés distinctes. Le modèle obtenu est alors de type *multi-task*. Ceci a permis d'utiliser la GTM pour établir le profil des risques associés aux substances enregistrées dans le REACH. Une application pratique est de faire un criblage virtuel pour identifier les substances ayant un profil (éco)toxicologique inquiétant pour l'environnement et la santé humaine. Cet avant-dernier chapitre de la thèse unifie toutes les propriétés étudiées dans un unique outil de profilage des molécules.

Figure 2. Comparaison de l'espace chimique Publique vs. Industrielle.



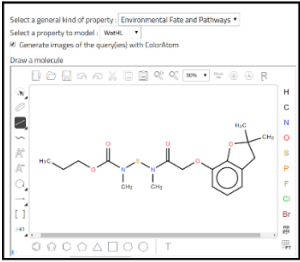
Les zones bleues sont peuplées par des molécules provenant de sources publiques ; les zones rouges, de source industrielle. Les couleurs intermédiaires, sont des zones peuplées par des composés provenant de ces deux sources. Les régions inexplorées de l'espace chimique figurent en blanc.

1.2.9 Implémentation du logiciel *ISIDA/Predictor*



Tous les modèles développés sont librement accessibles sur la plateforme *ISIDA/Predictor*: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi. L'image 3 représente un exemple d'utilisation de la plate-forme. Une fonctionnalité remarquable est le «*ColorAtom*» [4,6,13]: cet outil assigne des couleurs aux atomes de la molécule concernée en fonction de leur influence sur la valeur de la propriété prédite par un modèle. Cette information peut servir de support pour interpréter la prédiction d'un modèle, dans l'optique d'une «interprétation mécanistique», comme conseillé dans les directives internationales [2]. De plus une version desktop du *ISIDA/Predictor* a été livré à l'entreprise Solvay S.A., qui sera utilisé pour l'enregistrement des substances dans le cadre de REACH.

Figure 3. Capture d'écran du logiciel *Predictor* dans sa version SAS.

Input



Output

Predicted value	Applied models	Confidence	ColorAtom structure	Comments
nP	15/15			Optimal prediction confidence: AD satisfied for 100 % of applied models and convergence of predictions is 94 %

[Back to Main Menu](#)
[Download results](#)
[Download colored structures \(.off, .svg\)](#)

1.3 Conclusion générale

Dans le cadre de cette thèse un total de 11 propriétés importantes dans le contexte du règlement européen REACH ont été modélisées. Tous les modèles ont été validés selon les directives du règlement, afin qu'ils puissent être utilisés comme alternative aux tests expérimentaux.

Les modèles ont démontré leurs bonnes performances au cours d'analyses comparatives avec les modèles préexistants. Un point fort des modèles ici générés est l'introduction de nouvelles données provenant d'un contexte industriel. Ainsi la précision et la fiabilité des modèles sur les substances d'intérêt industrielle est améliorée.

Les modèles ont été implémentés dans la plateforme *ISIDA/Predictor*, qui est librement accessible. Une version desktop a été livrée à l'entreprise Solvay S.A. pour un usage interne.

II Introduction

2.1 Context

All the products related to a given economical market have to respond to two different kinds of regulations: the more general chemical regulation which regulates the produced substance rather than taking into accounts its final uses; and the market regulations, that provide specific obligations depending on the uses of a given products. For the former case, such regulations are generally state-dependent: in Canada there is the Canadian Environmental Protection Act [14]; in Japan the Chemical Substances Control Law [15]; in the United States the organism of reference is the Environmental Protection Agency [16]; and in Europe there is the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) [1]. Despite the differences, all these regulations are linked by the Global Harmonized System (GHS) [17], an internationally agreed-upon standard managed by the United Nations that was set up to replace the assortment of hazardous material classification and labelling schemes previously used around the world. Core elements of the GHS include standardized hazard testing criteria, universal warning pictograms, and harmonized safety data sheets which provide users of dangerous goods with a host of information. On the other hand, the market regulations are more specific to the product application. For instance, in Europe the EFSA (European Food Safety Agency) [18] is the authority responsible to regulate the markets of food and biocides. When a product is put on the market, it has to fulfil the requirements of both the chemical regulation and the market regulations (eventually more than one).

As the topics covered by this work fall into the domain REACH Regulation, it is described in more details. The REACH Regulation ((EC) N° 1907/2006) entered into force in 2007 and was born to address previous issues, including:

- Lack of information about the properties of several chemicals;
- Only few thousands substances addressed by previous legislation;
- Inadequate risk control;
- Poor information regarding risk assessment procedures between EU member Countries.

REACH was designed to pursue also other general objectives, including mapping of chemicals circulating over Europe, gain in depth knowledge about their effects on human health and the environment and replacement of hazardous substances (e.g. PBT) with safer alternatives. This system was implemented through four main actions, namely:

- *Registration* of substances imported or manufactured in quantities larger than one tonne/year;
- *Evaluation* of substances in terms of safety;
- *Authorisation* for substances of very high concern (SVHC);
- Restrictions to use.

The domain of application of REACH is very broad, and the registration of chemicals under REACH is required for: (i) all substances imported or manufactured in quantities greater than one tonne/year; (ii) monomers in polymers if present in percentages equal to or greater than 2% weight by weight and if the total quantity of monomer is greater than one tonne/year; (iii) substances in articles if the total amount is greater than one tonne/year and their release is intended under standard conditions of use.

Probably the most important modification introduced by REACH is the inversion of the burden of proof from regulators to industry, by imposing the concept of “no data, no market”. Producers are required to prove that their substances do not pose risks to human health and the environment. To this end, registrants must submit to ECHA a technical dossier, which includes information about the physico-chemical, toxicological, ecotoxicological and environmental fate properties of the substance, as specified in Annexes VII to X of REACH. A REACH dossier is composed by several tens of these properties (*endpoints*), depending on the tonnage band: the higher the amount of substance imported or produced, the higher the number of endpoints that need to be assessed. These properties are normally determined experimentally, involving the use of a significant number of animals. Therefore, in order to reduce animal testing, REACH promotes the use of alternatives methodologies, in particular *in-silico*, for data-gap filling, like Quantitative Structure-Activity Relationship (QSAR) models. REACH will accept predictions from QSARs on their own (i.e. as replacement of measured data) only under conditions that guarantee (with a certain confidence) that the results are relevant, reliable and adequate.

These conditions are enunciated in Annex XI of REACH and detailed in the most comprehensive guidance currently available for the application of QSARs within REACH, the OECD “*Guidance document on the validation of (quantitative) structure activity relationship [(Q)SAR] models*” [2].

The amount of data that needs to be generated is significant, considering the different requirements. Therefore, in the past years the interest towards alternative methodologies for data-gap filling and risk assessment greatly increased. They include, in-vitro testing, data sharing, weight of evidence and read-across approaches, and in silico modelling, among which QSARs.

2.2 OECD requirements

This document reports a list of principles that must be fulfilled in order to define the scientific validity of QSAR models. The five principles state that:

“To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- 1. a defined endpoint;*
- 2. an unambiguous algorithm;*
- 3. a defined domain of applicability;*
- 4. appropriate measures of goodness-of-fit, robustness and predictivity;*
- 5. a mechanistic interpretation, if possible.”*

The OECD guideline is an important track for the development and usage of QSAR models in the context of this work. Therefore, we paid a particular attention to answer each of the above-mentioned points.

1. Data was acquired only from verified sources and carefully curated in order to select only those experimental measurements obtained with an appropriate protocol, matching the OECD guideline requirements;
2. The processes of description calculation and machine learning algorithm employment are described precisely and include the use of peer-reviewed software;

3. Rules for defining the applicability domain have been included in order to discriminate less reliable measurements and each prediction is associated with a reliability score;
4. Appropriate validation procedures have been followed in order to estimate not only the models' ability to fit the data in the training set, but also its accuracy in predicting properties for new chemicals;
5. These are statistical models based solely on the chemical structure, with no a priori hypothesis which could influence the choice of descriptors or method parameters; an utility has been implemented in order to facilitate the interpretation of the output value.

2.3 State of the art

During the past decades, several models have already been published on REACH-relevant properties, and many QSAR models are nowadays implemented in commercial or freely-available software, such as VEGA (Virtual models for property Evaluation of chemicals within a Global Architecture) [19], Toxicity Estimation Software Tool (TEST) [20], Estimation Program Interface (EPISuite) [21] and OPERA (OPEn (q)saR App) [22]. These tools could represent a significant aid for the registration process, however their use within an industrial context is often limited, for the following reasons:

- The models have been generated for substances considerably different from those of industrial interest, and therefore industrial products are outside the applicability domain of these models;
- The models have not always been generated according to the authority's guidelines (OECD Principles);
- Insufficient external validation on new experimental data.

2.4 Organization of the thesis

This thesis aims to answer to these drawbacks, by generating new OECD-compliant models on several important properties for REACH, validated on new data coming from an

industrial context (provided by Solvay). A total of 11 endpoints have been taken into account, divided into the following REACH-sections:

1. For the environmental fate and pathway section, the bioconcentration factor, the ready biodegradability and the environmental persistence in sediment, soil and water;
2. For the ecotoxicological section, the short-term toxicity to three trophic level organisms (algae, daphnia and fish);
3. For the human toxicity section, the short-term toxicity to rats and properties related to endocrine disruption (androgen and estrogen receptor binding).

Much effort has been put in collecting the largest amount of current publicly available data, in order to constitute the most comprehensive training sets, used for models generation. These public-data models only, were then externally validated on a substantial number of industrial compounds and, at the same time, a benchmarking against already existing freely-available tools has been carried out in order to compare models performances.

All generated models showed to have strong advantages compared to state-of-the-art tools, either in terms of prediction accuracy or extended applicability domain on industrial data. Furthermore, they have been generated following OECD guidelines and all relevant documentation necessary to use them as tools for data-gap filling is readily available. All the models are now freely-available through the online platform ISIDA/Predictor [7], where users can predict these properties for their substances. Moreover, a desktop version of the platform with additional features has been provided to Solvay S.A. for internal use for REACH dossier compliance. This is only a global overview on the state-of-the-art. More detailed information can be found in each endpoint specific section, in Chapter 2.

This thesis is comprised by three main chapters: Chapter I contains the *Resumé en français*; Chapter II describes the introduction to regulatory topics in the European context; Chapter III describes the adopted methodology; Chapter IV reports the results on the main projects of the thesis (i.e. the modelled properties, the application of the Generative

Topographic Mapping method for chemical space analysis and the software implementation); Chapter V reports the conclusions and perspectives.

2.5 Background of Solvay

Solvay is a Belgian chemical company founded in 1863 by Ernest Solvay. In the early 1860s Solvay filed a patent for a method that involved the reaction of ammonium bicarbonate and salt, the product being heated to yield sodium carbonate, or soda ash. This marked the beginning of the Company's fortune. In the future years up to present, the wide-ranging activities of Solvay have been focused on following market areas:

- Agriculture: crop protection, plant nutrition, seed and grain care;
- Food industry: flavour, food ingredients, food packaging;
- Consumer goods: personal care, homecare, household goods, packaging (non-food);
- Healthcare: medical equipment and instruments, pharmaceuticals;
- Industrial applications: protective coatings, surface treatment.

Chemicals account for about one-third of the company's revenues. Solvay is among the world leaders in several commodity chemicals, including soda ash, hydrogen peroxide, persalts, barium and strontium carbonate, and caustic soda, as well as such specialty chemicals as fluorochemicals. In plastics, which account for about one-quarter of overall sales, Solvay produces fluorinated polymers and elastomers, as well as vinyls. About 19 percent of revenues come from plastic processing, including automobile fuel and air intake systems, various films, and swimming pool linings. Solvay's pharmaceutical operations, generating about a quarter of revenues, are relatively small on a global scale, ranking about 37th among the world players in the early 2000s. Drug development efforts focus on four main therapeutic fields: gastroenterology, hormone treatments, cardiology, and mental health. Solvay operates in 50 countries; more than 95 percent of its revenues are generated outside of Belgium, with 45 percent originating outside of the European Union.

As Solvay has to fulfil the REACH requirements, the use of computational approaches would be an important asset in the registration process. They can provide reliable predictions to be used for data-gap filling, ultimately reducing registration costs.

Moreover, such models can be used in the research & development area, as first tier tool to screen compounds with an undesired environmental or (eco)toxicological profile.

III Tools and methods

This chapter explains the conceptual basis of QSAR / QSPR and their link with the REACH regulation, as well as the methodologies used in this thesis, including: data acquisition, curation and preparation; encoding of molecular descriptors; data visualization; model generation, validation and applicability domain assessment; and software implementation.

3.1 QSAR: background and role in regulatory context

The study of structure-activity relationships (SARs) and their quantification (quantitative structure-activity relationships, QSARs) owes much of its development to the research carried out by Corwin Hansch, from the 1960s [23]. Hansch equation related the potency of a biological effect with lipophilic, electronic and steric properties. Since the 1960s, QSAR analysis focused more and more on the development of theoretical variables that are not derived from experiments, the so called molecular descriptors [24]. The development of a solely theoretical and computerised description of the molecular structure and its properties (i.e. theoretical molecular descriptors) was partly made possible by progresses in informatics, as well as increase in the power of computers. This development lead to the definition of a vast number of descriptors as well as more advanced machine learning methods, spanning from the simpler multi-linear regression to support vector machine, random forest and artificial neural network models [25].

3.2 Considered endpoints

Table 2 lists the 11 endpoints that have been considered; briefly described in paragraphs 3.2.1 – 3.2.6.

3.2.1 Bioconcentration Factor

The Bioconcentration Factor (BCF) estimates the tendency for a xenobiotic to concentrate inside living organisms and it is defined as the process of concentration of the chemical from the water phase through non-dietary routes, such as absorption from respiratory surfaces (e.g. lungs/gills) or skin [4].

3.2.2 Ready Biodegradability

Biodegradability is a key process that controls the environmental fate of chemicals and, as a consequence, potential exposure ways for living organisms to many xenobiotics. One of the most important ways for estimating biodegradation is the determination of the so-called “ready biodegradability” (RB) parameter, which comes from a stringent first-tier assessment, providing a binary classification whether the substance rapidly degrades in the environment [5].

3.2.3 Environmental persistence in Sediment, Soil and Water

Differently from the relatively cheap and fast RB assays, these high-tier simulation studies are carried out when the substance’s degradation half-life (in a given environmental compartment) value needs to be evaluated [26].

3.2.4 Short-term aquatic toxicity on Algae, Daphnia and Fish

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

3.2.5 Short-term toxicity on Rodent

The aim is to estimate the short-term toxicity to rodent via oral administration route. The REACH requires its assessment even for small tonnages. Consequently, this endpoint is one of the most commonly performed animal test despite its ethically debatable interest, which reflects its much higher data availability compared to the other endpoints [6].

3.2.6 Androgen and Estrogen receptor binding

An endocrine disrupting chemical is an exogenous substance that alters the functions of the endocrine system and consequently causes adverse effects. In the framework of the “Collaborative Estrogen Receptor Activity Prediction Project” (CERAPP) and “Collaborative Modelling Project for Androgen Receptor Activity” (CoMPARA) [11,12] international workgroups, a large number of compounds were tested for their potency to disrupt the AR/ER signaling pathway chains.

Table 2. Selected endpoints and data availability.

REACH section	Endpoint	Acronym	Model type	Training set	External set
Env. fate	Bioconcentration factor	BCF	REG	1263	31
	Ready biodegradability	RB	CLS	3069	362
	Persistence in sediment	SedP	CLS	436	131
	Persistence in soil	SoilP	CLS	630	909
	Persistence in water	WatP	CLS	466	140
Ecotox.	Algae acute toxicity	AlgaeTox	REG	1231	228
	Daphnia acute toxicity	DaphniaTox	REG	2083	249
	Fish acute toxicity	FishTox	REG	2152	193
	Rodent acute toxicity	RodentTox	REG	11191	197
Human tox.	Androgen receptor binding	AR binding	CLS	1661	3882
	Estrogen receptor binding	ER binding	CLS	1661	5795

3.3 Publicly available data sources

Several online freely available databases were queried for experimental data on REACH-relevant endpoints. A brief description is here provided:

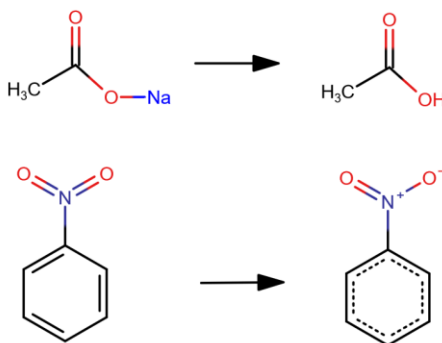
- ChEMBL: manually curated, freely available database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. The database is unique because of its focus on all aspects of drug discovery and its size, containing information on more than 1.8 million compounds and over 15 million records of their effects on biological systems. It was relevant for AR and ER binding.
- ECHA: European Chemicals Agency database storing all REACH registration dossiers. The database can be queried to download experimental data on all REACH endpoints. It counts more than 22'000 registered substances. It was queried for all endpoints.
- Already existing tools: VEGA [19], EPI Suite [21], TEST and OPERA [27], possessing several models on different properties and their training sets are, most of the times, easily accessible. They were relevant for all endpoints except for AR and ER binding.
- NITE [28]: Japanese National Institute of Technology and Evaluation (NITE), it contains useful information about environmental fate endpoints, such as BCF and RB.
- PubChem [29]: database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). It counts more than 90 million compounds. As ChemBL, it was relevant for AR and ER binding.
- QSAR Toolbox [30]: tool specifically developed for performing read-across in the context of REACH (Appendix 1.1). It was relevant for all endpoints except for AR and ER binding.

3.4 Data curation

To cross-check available chemical structure notation (i.e. SMILES) and retrieve missing notations (for instance when only CAS/EC number was available) the PubChem project was queried using automatized KNIME [31] workflows (Appendix 1.2). PubChem is currently the biggest freely accessible database containing chemical structures and identifiers and experimental measurements.

A serious source of error that can highly affect the quality of the model is related to wrongly represented structures. Certain aspect of chemical representation, such as stereochemistry, valence, charges and some functional groups notation (e.g. nitro groups; Figure 4) must be taken into account and such conventions must be homogeneously apply to all available compounds. This chemical structure “standardization” has been performed through the KNIME workflow (Appendix 1.3).

Figure 4. Example of chemical structure standardization.

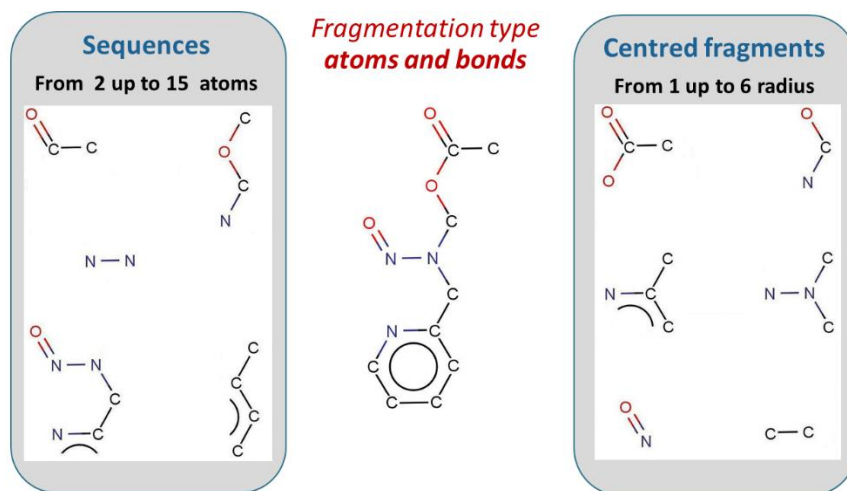


3.5 ISIDA descriptors

The Laboratory of Chemoinformatics of the University of Strasbourg has developed fragment descriptors consisting of sequences and augmented atoms of the atoms' element symbol as part of their chemoinformatics suite, named In Silico design and Data Analysis (ISIDA) [32]. ISIDA descriptors are highly configurable, as the user can decide to generate different types of descriptors (descriptor spaces, DS) based on different fragmentation patterns: sequences of increasing lengths or augmented atoms of different radiuses (Figure

5). With this approach, several tens of different DS are generated and the most appropriate one is chosen by Genetic Algorithm (GA) [33] optimization.

Figure 5. Example of ISIDA descriptors.



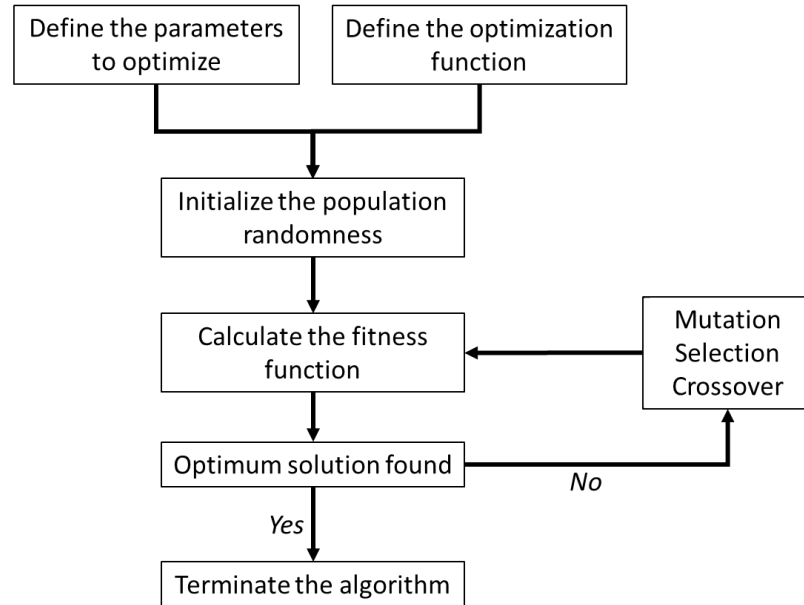
3.6 Genetic Algorithm

The GA has been employed in order to select the most appropriate set of DSs and, at the same time, tune the SVM hyperparameters [34] cost and gamma. GA algorithm is inspired by evolutionary theory by Darwin. In genetic algorithm approaches (Figure 6) to parameter space searches, chromosomes are a concatenation (vector) of the currently used operational parameter values. To each chromosome, a fitness score is associated, reporting how successful model building was when the respective choice of parameters was employed. Higher scoring chromosomes (parameter configurations) represent better fitness, and are allowed to preferentially generate “offspring”, by cross-overs and mutations, that are predisposed to result in new, potentially improved chromosome configurations.

The fitness (objective) function (SVM model quality score) to be optimized by picking the most judicious parameter combination should accurately reflect the ability of the resulting model to extrapolate/predict instances not encountered during the training stage, evaluated by 3-fold cross-validation (CV). The better the performance of this model in CV, the higher its fitting score and, as consequence, its probability to generate offspring

by cross-overs with other fit chromosomes. This translates into accumulation of well-cross validating models in the population, over time.

Figure 6. General workflow of the genetic algorithm process.



3.7 Employed machine learning methods

3.7.1 Support Vector Machines

The goal of the Support Vector Machine (SVM) [34] algorithm is to find an hyperplane in an N-dimensional space (N = the number of features) dividing the training data into two subsets corresponding to the experimental (binary) classes. As stated, this problem has many degenerated solutions. A second constraint is added: the optimal hyperplane maximizes the margin, i.e. the maximum distance between data points of both classes. Finally, if the classification problem involves more than two classes, the SVM algorithm is repeated for each class against all the others.

The number of features influences the dimension of the hyperplane: if the number of input features is 2, then the hyperplane is just a line; if it is 3, then the hyperplane becomes a two-dimensional plane; etc. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. When SVM is

employed as regression method, it is normally referred as Support Vector Regression (SVR). Similarly to the classification case, the SVR defines a solution to the regression problem based on instances of the training set. The function to optimize tolerates errors as long as they are smaller than an epsilon value and is linear with the error outside. As a consequence, the solution is based only on a subset of the training set whose members are called support vectors.

SVM models were generated with LibSVM and were employed for regression and classification models.

3.7.2 Random Forest

Random Forest is an ensemble learning method for classification and regression tasks that is constituted of a multitude of random trees [35]. The model's prediction result from a consensus of the predictions provided by each individual tree. The idea is to combine weak, relatively uncorrelated individual predictors into a strong one.

To achieve this result, Random Forest uses the technique of bootstrap aggregation (bagging). Bootstrapping is a sampling technique that choses, with replacement, k samples out of the n samples available, which are then selected as training instances for the individual learner. This strategy limits the degree of correlation between the random trees of the forest. Aggregating consists in combining the predictions of the individual learners together, which can be based on different strategies depending on the problem such as, for instance, arithmetical average, weighted average, most voted class, etc.

RF models were generated with WEKA and were employed for regression and classification models.

3.7.3 Naïve Bayesian

A Naive Bayes classifier is a probabilistic machine learning model for classification tasks. It is based on the Bayes formula [25].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

According to the theorem, it is possible to find the probability of event **A** happening, given that event **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. This formula does not help in practice because the evidence **B** is typically composed of many more simple events, B_i . For instance, **B** can be a long vector of molecular descriptors, and individual events B_i are the occurrence of a particular value for a given molecular descriptor. For this reason, $P(A|B)$ is typically under sampled and cannot be computed in practice. For the same reason, the quantity $P(B|A)$ is difficult to compute too. But it can be assumed that each event that compose **B** are independent, which is a “naïve” assumption. In this condition, this quantity factorizes in more simple and better sampled quantities $P(B_i|A)$ that can be evaluated. The naïve assumption made here is that the features (molecular descriptors) are independent: the realization of one particular feature does not affect the others.

NB models were generated with WEKA and were employed for classification models.

3.8 Employed tools

3.8.1 KNIME v.4.2

The data mining software Konstanz Information Miner (KNIME) [31] is a graphical workbench that allows to create exportable workflows through the connection of such called “nodes”. KNIME was used for data extraction and curation, file preparation (training / test set split) and results analysis (e.g. computing statistics).

3.8.2 ISIDA/Fragmentor v.2019

ISIDA/Fragmentor [32] has been used to generate ISIDA descriptors.

3.8.3 WEKA v.3.9.3

WEKA [36] is a data mining program in Java upheld by the Machine Learning group at the University of Waikato, which contains several machine learning algorithms. It has been used to generate Random Forest (RF) and Naïve Bayesian (NB) models.

3.8.4 LibSVM v.3.21

The LibSVM [34] package developed by Chih-Chung Chang and Chih-Jen Lin, has been used in this work to build SVM models. SVM hyperparameters (cost and gamma) have been tuned by GA run.

3.8.5 ISIDA/ColorAtom v.2019

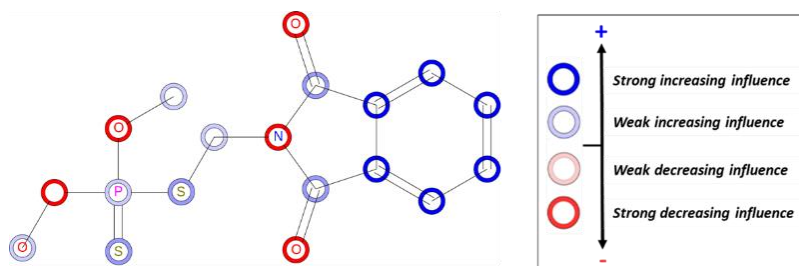
Interpretation of QSAR model is sought by chemists in order to give mechanist explanation of the observed phenomenon. However, achieving this, simultaneously with prediction efficiency is rare. The latter is often achieved using more complex algorithms (RF or SVM) or a consensus of models which cannot be readily interpreted. Another approach is to use atomic or fragment increments which sum up over the whole molecule to the predicted property. The approach developed by Marcou et al. [32] enables the interpretation of fragment descriptors into atomic increments by analysing partial derivatives of the predicted value.

An example of ColorAtom coloured graph is shown in Figure 7. Red colour means that the atom contributes to decrease the predicted property value; while blue means an increase of its value.

3.8.6 ISIDA/Predictor platform implementation

This tool is described in detail in section 3.10.

Figure 7. ColorAtom graph example.



3.9 Publicly available tools

Performances of new models generated in the thesis were compared with those of already existing tools (VEGA, EPI Suite, TEST and OPERA). These tools contain several models predicting various REACH properties and have been generated using different machine learning methods and approaches to evaluate the AD. This comparison (benchmarking) was performed on the industrial set compounds used to externally validated our models.

In such a way, it was possible to test all the models on the same set of industry-relevant compounds. Benchmarking was performed following the rules below:

- Compounds inside the training set of the models were excluded;
- Only predictions for compounds inside AD were considered. The definition of the AD varies, depending on the tool;
- Performances were compared based on the RMSE (regression) or BA (classification) parameters and the AD coverage, i.e. the percentage of compounds that fell inside the applicability domain of the model.

3.9.1 TEST v.4.1

The Toxicity Estimation Software Tool (TEST) [20] contains several models predicting physicochemical and ecotoxicological properties. Its predictions are based on a consensus of different machine learning methods including hierarchical clustering, single model (a

single multiple linear regression model), Food and Drug Administration (FDA) method, nearest neighbour. Concerning applicability domain, TEST estimates toxicity using different methods. Before any cluster model can be used to make a prediction for a test chemical, the program checks whether the chemical falls within the AD for the model. For the consensus method, the predicted toxicity is estimated by taking an average of the predicted toxicities from the above QSAR methods (provided the predictions are within the respective AD). The uncertainty in the overall prediction is calculated for each method. A benchmarking with TEST was possible for the following properties: BCF, DaphniaTox, FishTox, RatTox.

3.9.2 VEGA v.1.1.5

The VEGA [19] platform contains several QSAR models for various endpoints using different machine learning algorithms. The implementation of these models in VEGA allows an estimation of the reliability of each prediction through the Applicability Domain Index (ADI), which is a parameter that is calculated in a way independent from the QSAR model and ranges from 0 (for the lowest level of confidence, i.e. the molecule is out of the applicability domain of the model) to 1 (for the highest level of confidence). Only compounds with an ADI > 0.85 (i.e. high reliability) were included. VEGA was employed in benchmarking for the following endpoints: BCF, RB, AlgaeTox, DaphniaTox, FishTox, RatTox, SedP, SoilP, WatP.

3.9.3 EPI Suite v.4.1

The Estimation Programs Interface suite [21] contains several models for estimating environmental and ecotoxicity endpoints. The main drawback of this tool is a completely absence of an AD verification. Despite some indications are available, such as a simple molecular weight range limit, no automatic AD evaluation is carried out and therefore the users cannot easily use its predictions for dossier registration. Benchmarking with EPI Suite was carried out for: BCF, RB, AlgaeTox, DaphniaTox and FishTox.

3.9.4 OPERA v.2.5

OPERA [27] is an open-source application collecting several models. In principle, the same models proposed by EPI Suite have been update with the introduction of more data and an AD verification. The latter is performed using two methods: (i) similarity threshold between the query compound and training set compounds; and (ii) leverage approach. OPERA was used for benchmarking on: BCF and RB.

3.10 ISIDA/Predictor platform implementation

The ISIDA/Predictor platform, accessible through the Laboratory of Chemoinformatics website (http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi), is an online webservice storing all generated models. The user can either draw a query molecule or upload several compounds via SDF files. The platform automatically standardizes the query compound according to the defined rules and the prediction is then performed. A 4-grade reliability assessment (*outside AD, Average, High, Optimal*) is associated to each prediction, as a function of the combination of two independent scores: (i) the % of applied individual models (the higher, the more reliable the prediction is); (ii) the standard deviation between the predictions (the lower, the more reliable).

A desktop version of ISIDA/Predictor has been delivered to Solvay as well for internal use. This version includes the additional QPRF generation feature which is essential to speed up the registration process under REACH. The QPRF (or QSAR prediction reporting format) is a report, done according to REACH guidelines, which contains all the information needed by the authorities to evaluate the quality of a QSAR prediction, such as, for instance, the AD assessment, a mechanistic interpretation, model's description, etc. The report is automatically filled in and can be directly implemented in a REACH dossier, justifying the model's prediction.

IV Results

The results chapter is divided into three parts: part 1 lists all the 11 modelled endpoints; part 2 reports the application of GTM for data visualization and screening; part 3 provides an example of application of ISIDA/Predictor.

4.1 Part 1 – Modelled endpoints

4.1.1 Bioconcentration Factor

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

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

Models showed mixed performances on the industrial compounds. Tools with the highest accuracy are associated to a very narrow applicability domain, while models with

more permissive applicability domain performed worse as measured by the RMSE (root mean squared error). Our model scored the same accuracy (RMSE of 0.58 logBCF units) of the most accurate tool and preserved a much larger applicability domain (78 % data coverage). However, a general limitation of all models failed to predict some chemical families, such as siloxanes and highly phosphonated compounds: these are unique industrial set chemotypes which are under-sampled in the public data. In order to compensate the individual-model limitations, the use of all the available tools in consensus is encouraged to reduce uncertainty and improve the accuracy.

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



4.1.1 Facteur de Bioconcentration

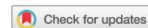
Dans l'évaluation des risques environnementaux, le Facteur de Bioconcentration (BCF) est un paramètre clé à considérer. Il estime la tendance d'un xénobiotique à se concentrer à l'intérieur des organismes vivants. Il est défini comme le processus de concentration du produit chimique depuis la phase aqueuse par des voies non alimentaires, comme l'absorption des surfaces respiratoires (par exemple les poumons / branchies) ou la peau. La concentration des xénobiotiques dans les organismes peut ainsi atteindre des niveaux dangereux, avec des effets délétères à long terme, tels que des modifications du comportement ou un impact sur la reproduction. Les organismes situés en haut de la chaîne alimentaire (par exemple, les poissons) sont particulièrement exposés et, par voie de conséquences, l'homme pourrait être l'ultime espèce touchée.

Dans cette étude, nous avons analysé si les modèles basés sur des données publiques montrent des performances satisfaisantes lorsqu'ils sont utilisés pour prédire un ensemble de composés du catalogue de l'entreprise Solvay («*Industriel set*»). Notre objectif était d'obtenir une image plus précise des performances des modèles existants et libres d'accès. Nous avons observé que les performances dans ce contexte industriel pouvaient diminuer, et nous avons émis l'hypothèse que le domaine d'applicabilité de ces

outils ne couvrait pas ces exemples industriels. Par conséquent, nous avons collecté l'ensemble de données BCF le plus complet à ce jour en fusionnant plusieurs jeux de données accessibles au public. Celui-ci a été utilisé pour générer un nouveau modèle BCF, qui a ensuite été validé selon une procédure de validation externe, sur les composés du jeu de données industriel et comparé aux outils existants.

Les modèles ont montré des performances mitigées sur les composés industriels. Les meilleures précisions sont associées à un domaine d'applicabilité étroit, tandis que les modèles avec plus permissifs sont associés à des erreurs (racine de l'erreur quadratique moyenne, RMSE) plus grandes. Notre modèle a obtenu les mêmes performances (RMSE de 0,58 unités logBCF) que le meilleur parmi les autres modèles et est prédictif sur une plus grande diversité de composés (la couverture des données de test atteint 78%). Toutefois, une limitation générale de tous les modèles concerne certaines familles chimiques, telles que les siloxanes et les composés hautement phosphonés. Ces chémotypes industriels sont spécifiques aux exemples industriels et sont sous-échantillonnés dans les bases de données publiques. Afin de compenser les limites des modèles individuels, l'utilisation de tous les outils disponibles simultanément dans un consensus est encouragée. Ceci a pour effet de réduire l'incertitude et d'améliorer la précision.

En conclusion, notre modèle est un outil alternatif pour prédire la propriété du facteur de bioconcentration dans un contexte industriel, contexte qui se caractérise par un espace chimique beaucoup plus hétérogène que les composés issus d'études antérieures, impliquant la plupart du temps des polluants classiques.



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

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

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

ABSTRACT

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

ARTICLE HISTORY

Received 9 April 2019
Accepted 29 May 2019


KEYWORDS

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

Introduction

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

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

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

© 2019 Informa UK Limited, trading as Taylor & Francis Group

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

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

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

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

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

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

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

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

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

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

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

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

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

Methods

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

Data collection and curation

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

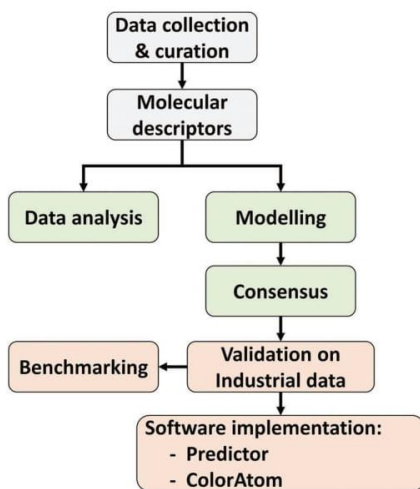


Figure 1. General workflow.

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

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

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

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

Generative topographic mapping

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

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

Encoding of chemical structures

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

Model generation and validation

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

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

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

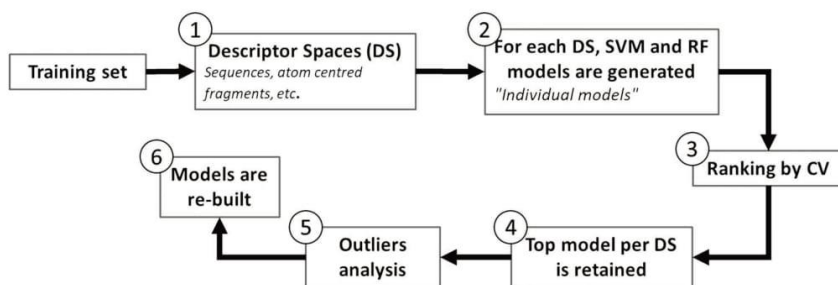


Figure 2. Modelling workflow.

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

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

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

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

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

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

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

Ensemble modelling and applicability domain

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

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

Predictions graphical interpretation: ColorAtom

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

Benchmarking on industrial data

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

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

Results

Overview of the curated dataset

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

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

GTM: industrial set visualization and description

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

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

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

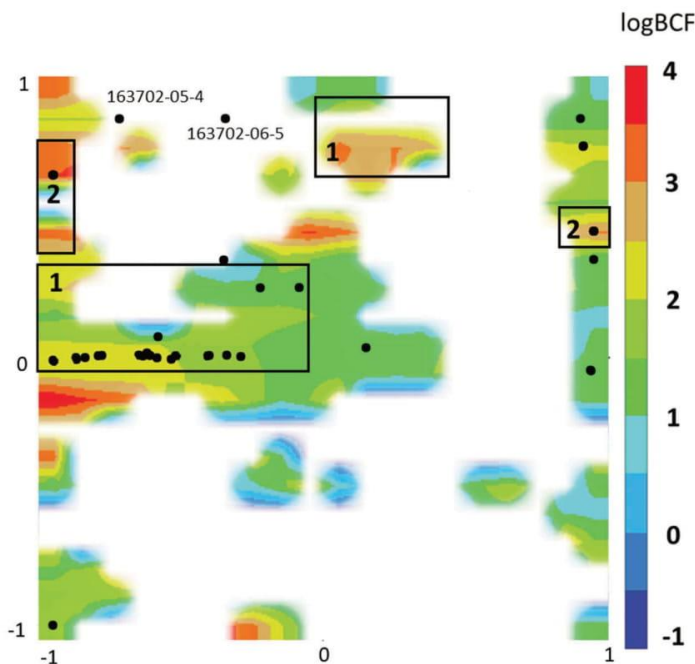


Figure 3. Log BCF property landscape of the training set. GTM is representing the density-modulated log BCF-landscape derived from training set compounds, onto which the industrial set compounds have been projected (black dots). White areas are empty regions of the map. Numbered boxes identify map regions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

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

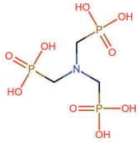
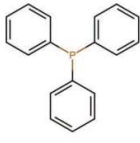
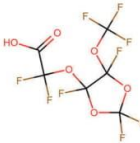
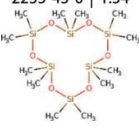
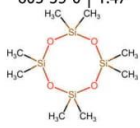
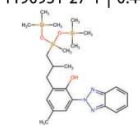
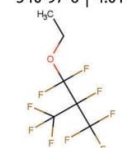
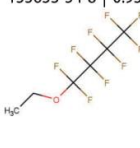
Region	Molecular structure		
1			
	2235-43-0 1.34	603-35-0 1.47	1190931-27-1 0.44
2			
	540-97-6 4.01	556-67-2 4.09	155633-54-8 0.93
Molecules on white area			
	163702-06-5 2.96		163702-05-4 2.96

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

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

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

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

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

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

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

Descriptor selection and model fitting

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

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

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

Performances on the industrial set

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

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

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

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

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

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

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

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

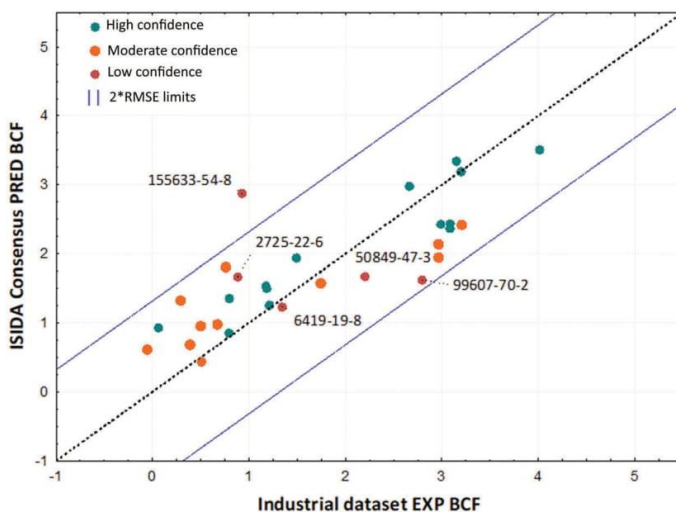


Figure 4. ISIDA consensus predicted vs industrial set experimental values. The data points labels correspond to the CAS numbers of chemicals. Red, orange and green dots = prediction confidence score based on the % in-AD models (<25%; 25–70% and >70%, respectively); Blue lines indicate $\pm 2 \times \text{RMSE}$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

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

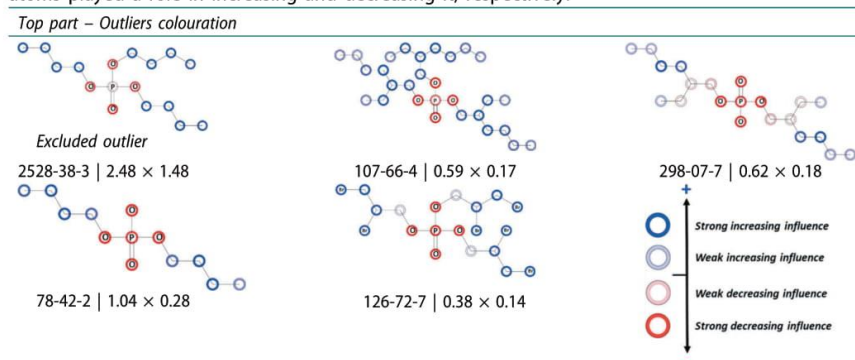
Pairwise comparison between:	Compounds in common	$r^2_{det.}$	RMSE
ISIDA Consensus vs VEGA KNN	9	0.77	0.55
ISIDA Consensus vs TEST	19	0.84	0.45
ISIDA Consensus vs EPISuite	23	0.73	0.63
ISIDA Consensus vs OPERA	18	0.57	0.80
		0.78	0.59
		0.45	0.92
		0.72	0.63
		0.68	0.67

ColorAtom: Graphical representations

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

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

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



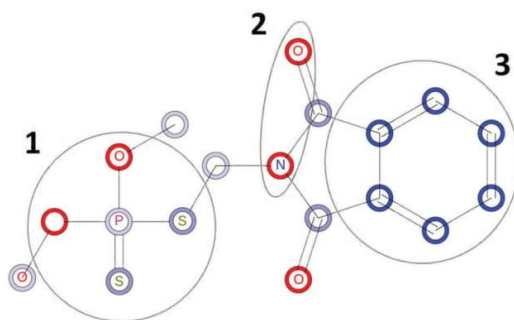


Figure 5. ColorAtom graph of phosmet. The colour scale is reported in Table 7. Numbered ellipses mark some recurring chemotypes subjected to discussion.

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

Discussion

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

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

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

Conclusions

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

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

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

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

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

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

References

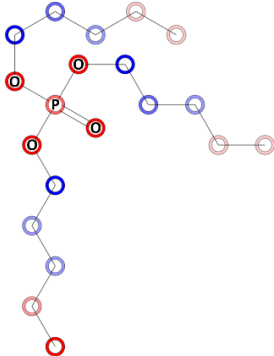
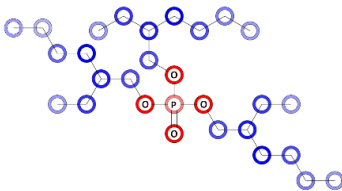
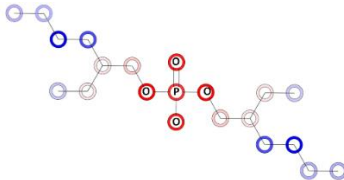
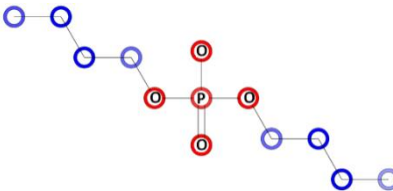
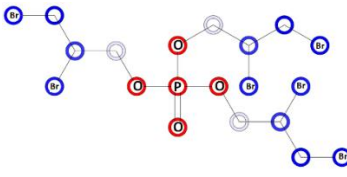
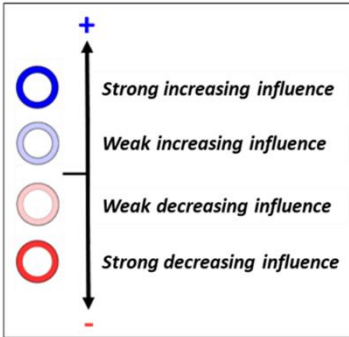
- [1] H.J. Geyer, G.G. Rimkus, I. Scheunert, A. Kaune, K.-W. Schramm, A. Kettrup, M. Zeeman, D. C. Muir, L.G. Hansen, and D. Mackay, *Bioaccumulation and occurrence of endocrine-disrupting chemicals (EDCs), persistent organic pollutants (POPs), and other organic compounds in fish and other organisms including humans*, in *Bioaccumulation–New Aspects and Developments*, B. Beek, ed., Springer Publisher, Berlin, 2000, pp. 1–166.
- [2] European Commission, *Technical guidance document in support of commission directive 93/67/EEC on risk assessment for new notified substances and commission regulation (EC) No 1488/94 on risk assessment for existing substances*, Tech. Rep. EUR 20418 EN/2, Institute for Health and Consumer Protection, Joint Research Centre, Ispra, IT, 2003.
- [3] OECD, *Test No. 305: Bioaccumulation in fish: Aqueous and dietary exposure*, Tech Rep. 9264185291, Organisation for Economic Co-operation Development, Paris, FR, 2012. doi:10.1094/PDIS-11-11-0999-PDN
- [4] ECHA, *Guidance on information requirements and chemical safety assessment, r.11: PBT/vPvB assessment*, Tech. Rep. ED-01-17-294, European Chemicals Agency, Helsinki, FI, 2017.
- [5] European Commission, *Regulation (EC) no 1907/2006 of the european parliament and of the council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a european chemicals agency, amending directive 1999/45/ECC and repealing council regulation (EEC) No 793/93 and commission regulation (EC) No 1488/94 as well as council directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/EEC and 2000/21/EC*, Off. J. Eur. Union. 50 (2007), pp. 1–281.
- [6] J.A. Arnot and F.A.P.C. Gobas, *A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms*, Environ. Rev. 14 (2006), pp. 257–297. doi:10.1139/a06-005.
- [7] M. Pavan, T.I. Netzeva, and W. Andrew, *Review of literature-based quantitative structure - activity relationship models for bioconcentration*, QSAR Comb. Sci. 27 (2008), pp. 21–31. doi:10.1002/qsar.200710102.
- [8] J.C. Dearden and M. Hewitt, *QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors*, SAR QSAR Environ. Res. 21 (2010), pp. 671–680. doi:10.1080/1062936X.2010.528235.
- [9] M. Nendza and M. Müller, *Screening for low aquatic bioaccumulation (1): Lipinski's 'rule of 5' and molecular size*, SAR QSAR Environ. Res. 21 (2010), pp. 495–512. doi:10.1080/1062936X.2010.502295.
- [10] J.F. Aranda, D.E. Babelo, M.S. Leguizamón Aparicio, M.A. Ocsachoque, E.A. Castro, and P. R. Duchowicz, *Predicting the bioconcentration factor through a conformation-independent QSPR study*, SAR QSAR Environ. Res. 28 (2017), pp. 749–763. doi:10.1080/1062936X.2017.1377765.
- [11] K. Mansouri, C.M. Grulke, R.S. Judson, and A.J. Williams, *OPERA models for predicting physicochemical properties and environmental fate endpoints*, J. Cheminformatics 10 (2018), pp. 1–19. doi:10.1186/s13321-018-0263-1.
- [12] E. Benfenati, A. Manganaro, and G. Gini, *VEGA-QSAR: AI inside a platform for predictive toxicology*, *Proceedings of the workshop 'Popularize Artificial Intelligence 2013'*, December 5th 2013, Turin, Italy, 2013, published in CEUR Workshop Proceedings Vol 1107, pp. 21–28.
- [13] T. Martin, P. Harten, and D. Young, *TEST (Toxicity Estimation Software Tool) V 4.1*, US Environmental Protection Agency, Washington DC, USA, 2012; software available at <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.

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

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

Erratum

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

Top Part - Outliers coloration				
				
Excluded outlier				
2528-38-3 2.48 1.48	78-42-2 1.04 0.28	298-07-7 0.62 0.18		
				
107-66-4 0.59 0.17	126-72-7 0.38 0.14			

4.1.2 Ready Biodegradability

Biodegradability is a key process which controls the environmental fate of chemicals and, as a consequence, potential exposure of 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, for instance by reaching the marine environment and being transported to remote areas.

One of the most important way for estimating biodegradation is determination of the so called “Ready Biodegradability” (RB) binary classification parameter, corresponding to either slow (nB) or fast (B) biodegradation.

With the ending of the last REACH registration deadline (June 2018) for low-volume substances (between 1-100 tonnes) and the sharing of REACH study results, new information is available. However, except for the recently published OPERA (2018), the training sets (from 200 to 589 compounds) 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.

A benchmarking against the already existing tools showed that the new model scored the best predictive power (BA = 0.77), followed by VEGA, EPI Suite, OPERA and ToxTree tools, 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.

Therefore, collected public data and the Industrial set have been merged into the “Global” data set containing 3146 compounds which is the largest RB set reported so far covering important representative chemotypes of the industrial context.



4.1.2 Biodégradabilité primaire

La biodégradabilité est un processus clé qui contrôle le devenir des produits chimiques dans l'environnement et, par conséquent, les voies d'exposition potentielles des organismes vivants à de nombreux xénobiotiques. En effet, les produits chimiques persistants dans l'environnement peuvent induire une exposition à long terme pour les êtres humains et l'écosystème à grande échelle (par exemple, en atteignant le milieu marin et puis en étant transportés jusque des zones reculées).

L'un des principaux moyens d'estimer la biodégradation d'un composé est la détermination de la biodégradabilité primaire (*Ready Biodegradability*, RB) un paramètre de binaire, correspondant à une biodégradation lente (nB) ou rapide (B).

Avec la fin de la période d'enregistrement de REACH (juin 2018) pour les substances de faible volume (entre 1 et 100 tonnes) et le partage des données qui en découlent, de nouvelles informations sont disponibles. Cependant, à l'exception du modèle OPERA récemment publié (2018), les jeux d'entraînement des modèles existants sont assez limités (de 200 à 589 composés) et n'ont pas été mis à jour depuis plusieurs années. Dans ce travail, nous présentons un nouveau jeu de données plus étendu concernant le RB, résultant de la fusion de plusieurs sources de données publiques

Une analyse comparative avec les outils existants a montré que le nouveau modèle avait un meilleur pouvoir prédictif (précision balancée, BA = 0,77), suivi des outils VEGA, EPI Suite, OPERA et ToxTree, (BA = 0,74, 0,71, 0,69, 0,68 et 0,67, respectivement). Cette comparaison a démontré que chaque modèle testé a des points forts spécifiques: par exemple, VEGA est capable de classer correctement les composés B positifs tandis que EPI Suite est utilisable sur la plus grande gamme de composés chimiques, et nos modèles ont la meilleure précision. Néanmoins, tous les modèles sont en défaut concernant plusieurs classes de composés d'intérêt industriel (par exemple les siloxanes et les cations organophosphonium), parce qu'ils sont essentiellement absents des jeux d'entraînement.

Par conséquent, les données publiques collectées et le jeu de données industriel ont été fusionnées dans un jeu de données «Global» contenant 3146 composés. C'est le plus grand jeu de données concernant le RB à ce jour, couvrant d'importants chémotypes représentatifs du contexte industriel.



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

^aLaboratory of Chemoinformatics - UMR7140, CNRS/University of Strasbourg, Strasbourg, France;

^bToxicological 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>.

© 2019 Informa UK Limited, trading as Taylor & Francis Group

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	-	211	0.65	0.79	0.72	[16]
	Algorithm: rule-based approach						

Sn = Sensitivity, Sp = Specificity, BA = Balanced Accuracy; ^aRB 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.

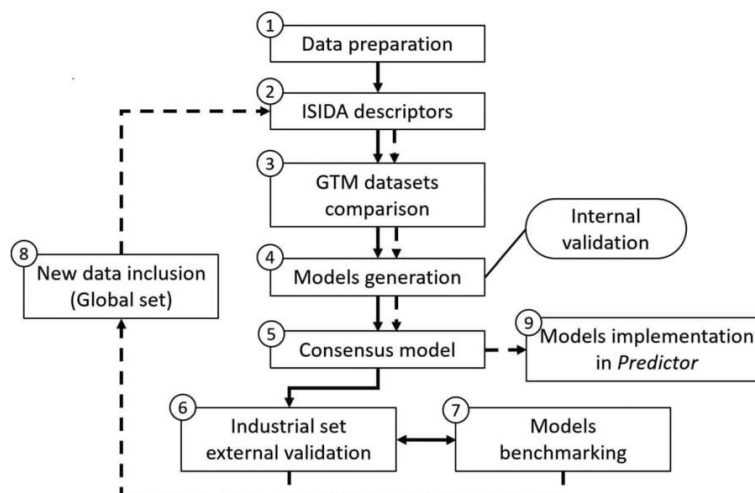


Figure 1. General workflow. (1) merging of collected data from multiple sources; (2) ISIDA descriptors are computed; (3) GTM is employed to compare the structural space of the datasets; (4), (5) individual models are trained using several machine learning algorithms and combined in consensus; (6) the 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	-

^aAll-Public dataset results from merging the NITE, VEGA, EPI SUITE, OPERA and ECHA datasets; ^bGlobal 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 its 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.

Model	Algorithm	BA in 3-fold CV	External validation ^b			
			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). ^aIn brackets, the standard deviation averaged over the CV repetitions is reported. ^bThe 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.

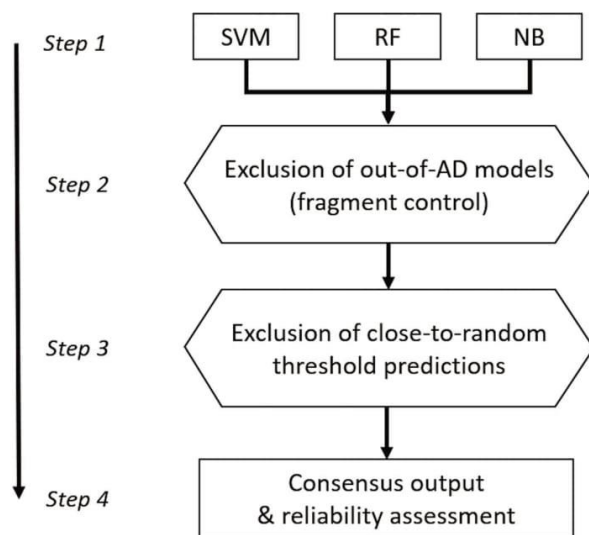


Figure 2. Consensus model workflow. Step 1: decisions of each algorithm (Support Vector Machine, Random Forest, Naïve Bayesian) are merged together; Step 2: predictions of models that failed the fragment control check are not considered; Step 3: if the percentage of votes for a given class is between 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

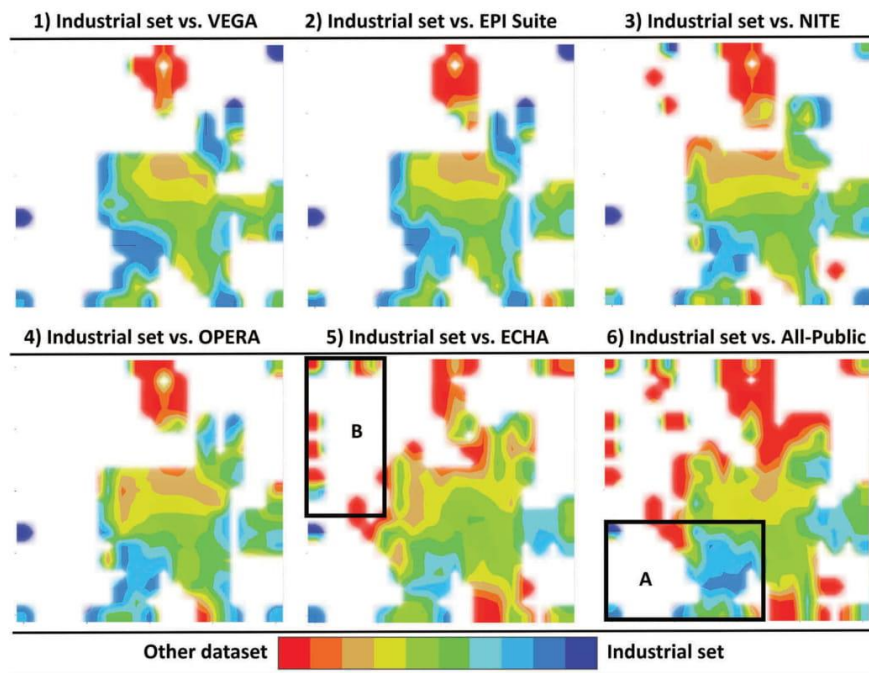


Figure 3. Pairwise datasets comparison using GTM. Each GTM map compares Industrial set versus one of publicly available datasets. Maps are sorted according to increasing size of a public dataset (from upper left to bottom right). Blue regions are mainly populated by Industrial set compounds; red ones 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 (T_c) 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 $T_c = 0.405$, with the majority of public compounds (70%) having a $T_c < 0.6$, 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

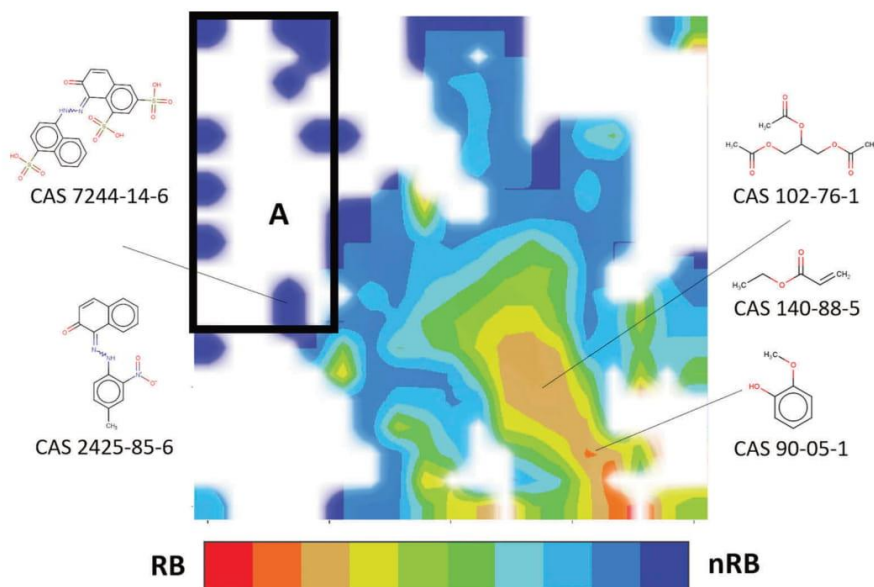


Figure 4. GTM ready biodegradability landscape. B compounds are identified by red zones, while nB by blue ones. The manifold was prepared for the Global set.

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.

Model	Industrial set				Literature set			
	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.

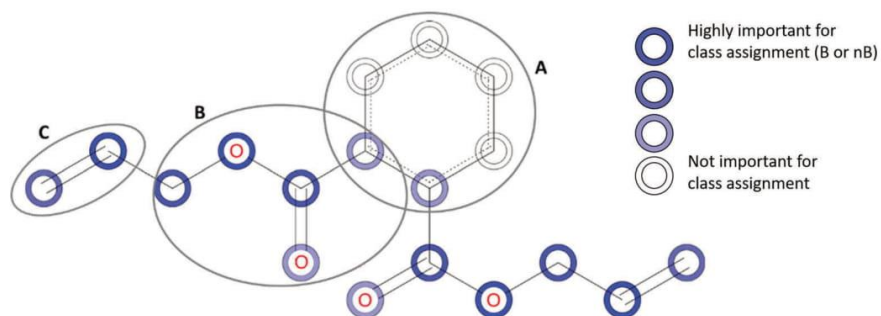


Figure 5. ColorAtom output representation. Colours refer to sensitivity of classification model to presence of a given fragment or atom: the darker the colour, the more that fragment (atom) is important for assigning the molecule to a given class.

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 S_n 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 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/EC and repealing council regulation (EEC) No 793/93 and commission regulation (EC) No 1488/94 as well as council directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/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-suite-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 MIT-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 Chimoinformatique 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/rtp.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 Francisco 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.

4.1.3 Environmental persistence in Sediment, Soil and Water

Persistence is defined as the ability of a chemical substance to stay unchanged in the environment for long time. Generally, it is expressed in terms of half-life (HL), i.e. the time it takes for half of the initial amount of substance to be removed from the environment [Larson , 1995]. According to the REACH, a substance fulfils the persistence (P) criterium when it shows a degradation half-life higher than 120, 120 or 40 days for sediments, soil or freshwater compartments, respectively.

Persistence is usually evaluated following a tiered approach, starting with relatively cheap and fast ready biodegradability assays. Since these assays have very stringent pass criteria and tend to underestimate the degree of degradation, a negative result does not necessarily mean that the substance will not be degraded under more realistic environmental conditions. In addition, those low-tier tests do not provide the required half-lives for comparison with the P and vP criteria of the Annex XIII of REACH. Therefore, more realistic high-tier simulation assays are carried out, with the aim to provide a better estimation of the substance's HL, that can be directly compared with the REACH criteria. The three environmental media of sediment, soil and water, are considered in this context, with degradation half-life thresholds as mentioned above to label a substance as persistent.

In the past years, several models estimating environmental persistence for pure compounds have been reported. However, the size of the training set is a serious limitation of existing models, as it is often restricted to specific chemical classes, such as aromatic compounds or only hydrocarbons. These latter models provide with reasonably accurate predictions, but their use for risk assessment purposes is quite limited, due to narrow Applicability Domain. Despite of the large number of existing models, only few of them have been implemented in some freely available tools. VEGA is the only one that can estimate a compound's persistence in a specific environmental compartment. In this work, we present a new and extended dataset for RB, issued from merging several public data sources. We report binary classification consensus models for persistence in sediment, soil and water environmental media. Overall, we collected a total of 1579 unique compounds, annotated by, at least, one experimental value for a given medium. This included subsets of 1533, 466 and 436 compounds whose persistence was measured in soil, water and sediment, respectively.

Existing tool VEGA showed only mediocre performances on all compartments, with balanced accuracy (BA) values ranging from 0.52 to 0.62. The main drawback of VEGA is a noticeably low data coverage, which ranges from 5% (sediment) to 44% (soil).

The sediment ($BA_{cv} = 0.81$, $BA_{ext} = 0.91$) and water ($BA_{cv} = 0.80$, $BA_{ext} = 0.77$) consensus models are noticeably more performant in both CV and external validation than the soil model ($BA_{cv} = 0.71$, $BA_{ext} = 0.76$). Lower BA values for the soil are degraded due to the noticeably lower sensitivity value, reflecting the low agreement of reported experimental values among data sources.



4.1.3 Persistance environnementale dans les sédiments, le sol et l'eau

La persistance est définie comme la capacité d'un produit chimique à rester inchangé dans l'environnement pendant long laps de temps. Généralement, il est exprimé en termes de demi-vie (*half-life*, HL), c'est-à-dire le temps qu'il faut pour que la moitié de la quantité initiale de substance disparaisse de l'environnement. Selon REACH, une substance remplit le critère de persistance (P) lorsque sa demi-vie de dégradation est supérieure à 120, 120 ou 40 jours pour les sédiments d'eau douce, le sol ou l'eau douce, respectivement.

La persistance est généralement évaluée selon une approche à plusieurs niveaux, commençant par des tests de biodégradabilité relativement bon marché et rapides. Étant donné que ces tests ont des critères de réussite très stricts et tendent à sous-estimer le degré de dégradation, un résultat négatif ne signifie pas nécessairement que la substance ne sera pas dégradée dans des conditions environnementales plus réalistes. De plus, ces tests ne fournissent pas les mesures de demi-vies correspondant aux critères de persistance P et vP de l'annexe XIII de REACH. Par conséquent, les tests de plus haut niveau, plus réalistes, sont effectués dans le but de fournir une meilleure estimation de la demi-vie de la substance, qui peut être comparée aux critères REACH. Dans ce contexte, les trois milieux environnementaux (sédiments, sol et eau) sont considérés avec des seuils de demi-vie de dégradation respectivement de 120, 120 ou 40 jours, définissant une substance comme persistante.

Au cours des dernières années, plusieurs modèles d'estimation de la persistance de l'environnement ont été publiés. Cependant, la taille et la composition des jeux

d'apprentissage sont des limitations récurrentes des modèles existants. Ils sont souvent limités à des classes chimiques spécifiques, telles que les composés aromatiques ou les hydrocarbures. Ces modèles fournissent des prévisions raisonnablement précises, mais leur utilisation à des fins d'évaluation des risques s'en trouve limitée, ceci se traduisant par un étroit domaine d'applicabilité des modèles (AD). Malgré le grand nombre de modèles existants, un petit nombre est mis en œuvre dans certains outils disponibles gratuitement. VEGA est le seul à pouvoir estimer la persistance d'un composé dans un compartiment environnemental particulier. Dans ce travail, nous avons exploré si cette nouvelle source de données de persistance peut aider à améliorer les performances des modèles QSPR. Nous avons ainsi obtenu des modèles consensus de classification binaire pour la persistance dans les milieux environnementaux sédiments, sol et eau. Au total, nous avons collecté 1579 composés uniques, annotés par au moins une valeur expérimentale pour un milieu donné. Ceux-ci se répartissent entre des sous-ensembles de 1533, 466 et 436 composés dont la persistance a été mesurée respectivement dans le sol, l'eau et les sédiments.

Les modèles VEGA ont montré des performances médiocres sur tous les compartiments, avec des mesures de performance (BA) allant de 0,52 à 0,62. Le principal inconvénient de VEGA est une couverture des données faible, variant de 5% (sédiments) à 44% (sols).

Les modèles de consensus sédiments ($BA_{cv} = 0,81$, $BA_{ext} = 0,91$) et eau ($BA_{cv} = 0,80$, $BA_{ext} = 0,77$) sont sensiblement plus performants en validation croisée et en validation externe que le modèle sol ($BA_{cv} = 0,71$, $BA_{ext} = 0,76$). Ces performances dégradées pour le compartiment sol traduisent une sensibilité sensiblement inférieure, reflétant la faible concordance entre les valeurs expérimentales référencées dans les sources de données.



Publicly available QSPR models for environmental media persistence

F. Lunghini ^{a,b}, G. Marcou ^a, P. Azam ^b, M.H. Enrici ^b, E. Van Miert ^b
and A. Varnek ^a

^aLaboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France; ^bToxicological and Environmental Risk Assessment Unit, Solvay S.A., St. Fons, France

ABSTRACT

The evaluation of persistency of chemicals in environmental media (water, soil, sediment) is included in European Regulations, in the context of the Persistence, Bioaccumulation and Toxicity (PBT) assessment. In silico predictions are valuable alternatives for compounds screening and prioritization. However, already existing prediction tools have limitations: narrow applicability domains due to their relatively small training sets, and lack of medium-specific models. A dataset of 1579 unique compounds has been collected, merging several persistence data sources annotated by, at least, one experimental dissipation half-life value for the given environmental medium. This dataset was used to train binary classification models discriminating persistent/non-persistent (P/nP) compounds based on REACH half-life thresholds on sediment, water and soil compartments. Models were built using ISIDA (In Silico design and Data Analysis) fragment descriptors and support vector regression, random forest and naïve Bayesian machine-learning methods. All models scored satisfactory performances: sediment being the most performing one ($BA_{ext} = 0.91$), followed by water ($BA_{ext} = 0.77$) and soil ($BA_{ext} = 0.76$). The latter suffer from low detection of persistent ('P') compounds ($Sn_{ext} = 0.50$), reflecting discrepancies in reported half-life measurements among the different data sources. Generated models and collected data are made publicly available.

ARTICLE HISTORY

Received 22 April 2020
Accepted 27 May 2020


KEYWORDS

QSAR/QSPR; environmental fate; persistence; generative topographic mapping (GTM); REACH

Introduction

Chemicals that are persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) are of high concern for the environment. Their low ability for degradation lead to accumulation in the environment, possible bioaccumulation and long-term effects in living organisms, and in some cases they can undergo long-range transport and contaminate remote areas [1]. Persistence is defined as the ability of a chemical to stay unchanged in the environment for a long time [2]. It is expressed in terms of half-life, i.e. the time it takes for half of the initial amount of substance to be removed from the environment [3]. More precisely, the degradation half-life takes into account only degradation processes (hydrolysis, microbial degradation, photolysis, etc.);

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

 Supplemental data for this article can be accessed at: <https://doi.org/10.1080/1062936X.2020.1776387>.

© 2020 Informa UK Limited, trading as Taylor & Francis Group

whereas the dissipation half-life (DT_{50}) is preferred when also dissipation processes (volatilization, leaching, plant uptake, etc.) played an important role in removing the substance from the compartment.

In Europe, the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) Regulation [4] compels manufactures and importers to register their substances if produced or imported for more than 1 ton/year. In addition, when the amount surpasses 10 tons/year, a Chemical Safety Assessment report must be produced by the registrant [5]. An essential point of this document is the evaluation of the substance's persistence, bioaccumulation and toxicity, which can lead to a potential PBT (persistent, bioaccumulative and toxic) or vPvB (very persistent and very bioaccumulative) status. According to the REACH Annex XIII [4], a substance fulfils the persistence (P) criterion when it shows a DT_{50} higher than 120, 120 or 40 days for freshwater sediment, soil compartment or freshwater, respectively. Similarly, the very persistent (vP) thresholds are respectively 180, 180 or 60 days. Threshold values are also available for the marine compartment.

Persistence is usually evaluated following a tiered approach, starting with the ready biodegradability assays (OECD 301 series) [5–7], which are relatively cheap and fast. Due to their very stringent pass-criteria [7], these assays tend to underestimated the degree of degradation and therefore a negative result does not necessarily mean that the substance will not be degraded under more realistic environmental conditions [5]. In addition, these tests are providing ultimate degradation results (degradation to simple molecule like CO_2 and CH_4) but they do not provide the required DT_{50} for comparison with the P and vP criteria of the Annex XIII of REACH. Therefore, more realistic high-tier simulation assays are carried out, with the aim to provide a better estimation of the substance's DT_{50} , that can be directly compared with the P and vP criteria. The three environmental media, i.e. sediment, soil and water, are considered in the context of PBT assessment [5]. Such experiments are carried out according to the OECD guidelines 307, 308 and 309 [8].

In the past years, several models estimating environmental persistence have been reported, ranging from simple regression equations to models developed with machine learning methods and expert systems [9,10]. However, the size of the training set is a serious limitation of existing models, as it is often restricted to specific chemical classes, such as aromatic compounds [11], pesticides [12], herbicides [13] or only hydrocarbons [14]. These models provide reasonably accurate predictions, but their use for risk assessment purposes is quite limited, due to narrow applicability domain (AD). Pizzo et al. [15] compiled persistence datasets on sediment, soil and water used to generate classification models. Similarly, the authors recognized that the main limitation was the relatively small number of compounds, which influenced models' performance and the relevance of extracted structural alerts. Despite a large number of existing models, only few of them have been implemented in freely available tools. VEGA (virtual models for property evaluation of chemicals within a global architecture) [16] is the only one that can estimate a compound's persistence in particular environmental compartment. Other tools like EPI Suite (estimation program interface) [17] and OPERA (OPEn (q)saR App) [14] provide quantitative estimation of persistence, but their AD is strictly limited to hydrocarbons and no distinction between media is possible.

Recently, Latino et al. [9] described the Eawag-Soil database (ESDB) assembling compounds' DT_{50} values in soil, issued from mining the EFSA's (European Food Safety Authority) pesticides registration dossiers. In this work, we explored whether this new

source of data can help to improve QSPR models performance. Concurrently, to complement the persistence evaluation in the context of the REACH PBT/vPvB assessment, we generated additional models for the sediment and water media. For this purpose, we have collected a dataset of 1579 unique compounds, annotated by, at least, one experimental value for the given medium. Our models are available through the online ISIDA/Predictor platform [18], accessible at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi.

Methods

Data sources

Experimental DT_{50} values data were collected from multiple sources (more details in Supporting Information (Table S1): (i) the RIVM report [19], (ii) the OSU Extension Pesticide Properties Database [20], (iii) the European Chemicals Agency (ECHA) database accessed through the eChem Portal [21], (iv) extraction from literature [15,22,23] and (v) the training sets of VEGA [16] (hereafter referred as '*Literature set*'), (vi) the Pesticides Properties Data Base (PPDB) [24] and (vii) the recently published ESDB [9]. Not all sources provided information on each medium. Moreover, the format of retrieved DT_{50} values varied, being either categorical (as ranges) or continuous. For instance, the *Literature set* and the PPDB are the only sources addressing all the three compartments. The former reports DT_{50} values divided into nine classes on a semi-decade logarithmic scale basis, whereas the later provides with continuous measurements. The largest collection is ESDB, with more than 10'000 quantitative $DT_{50(soil)}$ raw data values. One of its peculiarities is that detailed information concerning experimental test conditions (e.g. humidity, temperature, soil texture, etc.) are associated to the given experimental measurements, stored in the so-called '*scenarios*'. On the contrary, the RISVM, OSU, PPDB and ECHA databases report quantitative $DT_{50(soil)}$, but no information regarding experimental test conditions is given. All collected data is available on Zenodo: doi 10.5281/zenodo.3698144.

Training and test sets preparation

We decided to generate binary classification models for the following reasons: (i) only a limited amount of compounds with continuous DT_{50} values were available for sediment and water media, and soil regression models provided only mediocre results (see Discussion section); (ii) the intra- and inter-database variability residing in DT_{50} measurements was noticeably high. Therefore, the REACH-relevant 'P' thresholds of 120, 120 and 40 days were selected for sediment, soil and water, respectively. The non-persistent label ('nP') was assigned to compounds with DT_{50} values lower than the given cut-off; otherwise, the compound was label as persistent ('P').

Raw data processing and chemical structures standardization were carried out with a standardization workflow implemented in KNIME [25]. In case of duplicates, only one compound was kept, and its property was computed either as the median or the mode, for continuous and categorical DT_{50} assignments, respectively. For the latter case, when the repartition of the P/nP assignments was 50%, the entry was excluded.

Table 1. Training and test sets for the given compartment.

Model	Training set		Test set		log DT ₅₀ [days] range
	Size	nP/P	Size	nP/P	
Sediment ^a	305	128/177	131	55/76	-
Soil ^b	624	529/95	909 ^c	749/160	from -2.72 to 4
Water ^a	326	192/134	140	83/57	-

^aTraining and test set based on 70 %/30% stratified random splitting; ^bTraining set based on the entire ESDB;^cCompounds originating from RIVM, ECHA, Literature set, VEGA and PPDB.

Table 1 summarizes the composition of curated training and test sets. Since ESDB, is a more recent and manually curated database [9], we hypothesized that related soil medium data are of higher quality compared to other data sources. Therefore, we used the ESDB as training set, while the other data sources (RIVM, ECHA, Literature set, VEGA and PPDB) were merged to constitute the external test set. For the sediment and water media, models' training sets were obtained by stratified (based on nP/P binary labels) random splitting 70% and 30 %, respectively.

Molecular descriptors

ISIDA property-label molecular descriptors [26] were employed. 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 centred 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 (Figure 1, step 1). Among this entire pool, those that led to the generation of under-performing models were filtered out (see Model generation and validation section). The number of fragments varies as a function of selected fragmentation scheme. It ranged from 130 ('IAB(2-3)_AP', sequences of atoms and bonds of length up to three) for the sediment model to 5446 ('IIAB(2-5)', atom centred fragments of with radius up to 5) for the soil model, with an average of 1297. More details are given in Supporting Information (Tables S2-S4).

Generative topographic mapping

The generative topographic mapping (GTM) [27,28] is a dimensionality reduction method which can be considered as a probabilistic extension of the self-organizing maps. Briefly speaking, the algorithm injects a 2D hypersurface (manifold) [27] into an initial *D*-dimensional data space. The manifold is fitted to the data distribution and each item from the data space is projected to a 2D latent grid of *K* nodes, i.e. a 'map' showing the projections of compounds in the considered chemical space. A data property can be added as a 3rd axis forming such called class landscape [27]. Each landscape position is coloured according to the property value; this value is the average property of the data subset concerned by that position on the landscape. Therefore, to each environmental compartment, a class landscape is associated, visualizing the repartition of nP/P compounds in the chemical space. The manifold was built on the whole available chemical

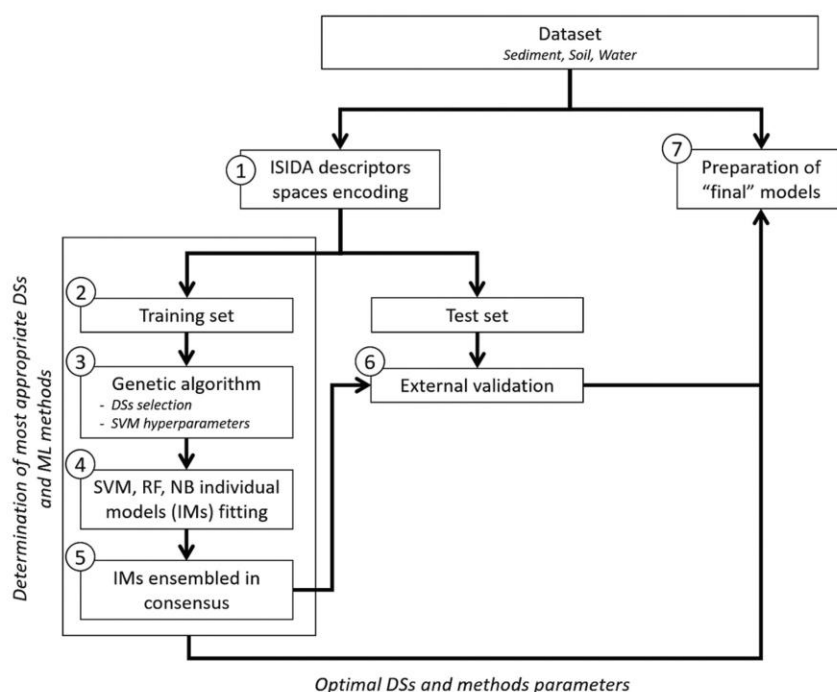


Figure 1. Model building workflow. (1) ISIDA descriptor spaces (DSs) are generated; (2) the dataset is split in training and test set; (3) best DSs and SVM hyperparameters are chosen by GA; (4) SVM, RF and NB individual models (IM) are fitted, and the best IM per DS is retained; (5) retained models are ensembled in consensus and (6) externally validated on the test set; (7) selected DSs and methods parameters are used to train the ‘final’ models, i.e. on the whole dataset.

space, i.e. by merging all the collected datasets of sediment, soil and water. GTM has four parameters (number of nodes, number of radial basis functions (RBFs), regularization coefficient, RBF’s width) to optimize according to some scoring function. Genetic algorithm (GA) [29] was employed to select ISIDA descriptor space (among the same pool used for model generation) and GTM parameters which maximise the cross-validated balanced accuracy (BA) calculated as an average of cross-validated BAs obtained for the models for sediment, soil and water. The following GTM parameters were suggested: the number of RBF centres $m = 27 \times 27$, the RBFs width $w = 1.2$ and the map resolution $k = 42 \times 42$ grid nodes. Atom centred fragments with the radius of 2 were used as descriptors.

Model generation and validation

Figure 1 depicts the modelling workflow. Support vector machine (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) [30]; WEKA (v. 3.9.3) [31] was used for RF and for NB models. For a given training set (Table 1), the ‘best’ ISIDA

descriptor spaces and the optimal SVM hyperparameters (cost and gamma) have been selected in genetic algorithm driven optimization process. In such a way, the top 15 descriptor spaces were retained (according to our experience, this threshold is a good compromise between performance and computational speed) and used to fit an equal number of optimized SVM, RF and NB 'individual models' (Figure 1, steps 2, 3, 4). For RF and NB, default WEKA settings were selected. Finally, only the best performing individual model corresponding to a given descriptor space was retained. Internal validation of each individual model was carried out by 5-fold CV repeated ten times (10*5CV) after data reshuffling. Statistics were assessed for each repetition followed by their averaging. The influence of chance correlations was checked through Y-scrambling [32] (with 50 repetitions). The 15 selected individual models were then ensembled in consensus (Figure 1, step 5), and external validation has been carried out on the given test set (Table 1). Tables S2-S4 report detailed information concerning consensus models set-up (employed descriptor space and algorithm for the given individual model) for each endpoint.

Finally, related training and test sets were merged, and models were updated using the same configuration (descriptor spaces, algorithm and method parameters) previously determined. Validation was carried out, as previously, in 10*5CV. This process was repeated for each dataset, i.e. sediment, soil, water. Performances were evaluated through the analysis of the sensitivity (Sn), specificity (Sp) and balanced accuracy (BA) metrics (Table S5 in Supporting Information).

Applicability domain and ensemble modelling

The applicability domain was evaluated based on the 'fragment control' assessment [26]: if the test molecule has a fragment not present in the training set, it is considered as 'out-of-AD'. Generated models were assembled in consensus which outcome corresponded to majority vote, without any consideration of out-of-AD predictions. In addition, we propose a 4-grade reliability scale system based on the % of models with positive AD outcome, as described in our previous works [33,34]. Briefly, depending on the % of individual models for which the compound was inside the AD a score of low ($\leq 25\%$), acceptable (25–50 %), high (50–80 %); or very high ($\geq 80\%$) was attributed. A compound is considered to be inside the AD when its reliability is higher than low.

State-of-the art models comparison

VEGA was evaluated on the whole available collected data for the given medium (merging of training and test sets; Table 1). To avoid potential sources of overestimation, molecules already present in its training set were excluded and out-of-AD predictions (i.e. 'low reliability', as stated by VEGA's output) were not considered. VEGA persistence models propose the following possible four classifications: nP (non-persistent), nP/P (close to persistent threshold), P/vP (close to very-persistent threshold), vP (very-persistent). The nP and vP classes come, respectively, from the PBT and vPvB thresholds defined under REACH; while nP/P and P/vP classes refer to a borderline classification between these two series (based on the given compartment) of thresholds. As our dataset is based only on two classes, the output of VEGA was converted accordingly: nP and nP/P predictions were treated as nP; whereas P/vP and vP were considered as P. Since the VEGA

predicted classes are not perfectly comparable to our models' output, and the models are not evaluated on exactly the same set of compounds, this comparison is not meant to be a benchmarking, but to provide an overall view on how already-existing models are performing when challenged to predict persistence of new compounds.

ColorAtom structural-activity dependence analysis

The 'ColorAtom' utility assigns a colour code to each fragment or atom showing the sensitivity of classification model to its presence in molecular structure [35]. For a given fragment, dark colour means 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. For instance, if a compound is predicted as P, dark-coloured fragments are positively associated to a persistent behaviour, and their removal is likely to change the classification of the compound to the other class (i.e. nP).

The ColorAtom brings insight to the models: through their interpretation, it is possible to check the consistency between known facts and their behaviour. We selected several compounds which contained P or nP structural alerts identified in the work of Pizzo et al. [15] which were used to generate ColorAtom graphs. Due to relatively high number of reported structural alerts, only the most significant were selected, i.e. those ones reported to have the highest accuracy. For the given medium, two structural alerts are herein reported, one for P and the other one for nP compounds.

Results

Soil dataset inter-database variability

As the soil dataset was issued by merging several sources, we analysed the degree of agreement based on assigned nP/P labels: overlapping compounds had their reported experimental label compared and the agreement is expressed in terms of accuracy. For each pair database 1 (DB1) – database 2 (DB2), Table 2 reports an overlap rate = N_i/N ,

Table 2. The overlap rate (OR) for different pairs of the soil datasets.

	EnviPath	ECHA	OSU	RIVM	PPDB	Literature
EnviPath	-	na	0.17	0.63	0.75	0.25
ECHA	na	-	na	0.80	na	na
OSU	0.98	1	-	0.23	0.36	na
RIVM	0.96	0.80	0.97	-	0.65	0.30
PPDB	0.95	1.00	0.99	0.96	-	0.22
Literature	0.84	na	0.98	1	0.98	-

For each pairwise comparison, OR for the persistent (orange background) and non-persistent class (blue background) is reported. na = not available due to absence of overlapping compounds.

where N_i is the number of overlapping compounds between DB1 and DB2 having the same label i ($i = P$ or nP) and N the total number of overlapping compounds. The matrix's upper (or lower) part reports an overlap rate for the class P (or nP). For some databases combinations, the comparison could not be performed due to the absence of common compounds. The number of overlapping compounds ranges from 11 (ECHA vs. RIVM) to 194 (PDBB vs. RIVM). Overall, the ECHA is the source showing the smallest overlapping, which is an indication that compounds that are of industrial interest are frequently missing from other datasets.

It can be noticed that for the nP class the agreement is noticeably high, with an average overlap rate of 0.96. On the contrary, the average overlap rate for the P class is only 0.44. Some examples of compounds with highly variable soil experimental values are reported in Table 3. For some of them it was possible to report the range of continuous DT_{50} measurements as well, providing that multiple ESDB scenarios were available. For example, values for the herbicide diuron span over a range of 1.3 log unit, depending on reported test conditions (e.g. pH varied from 4.6 to 7.3, humidity varied from 35 to 70 %, etc.). However, we did not find a significant correlation between test conditions and measured DT_{50} value: the highest correlation observed was 0.22, for the temperature parameter.

Chemical space analysis using GTM

Figure 2 shows the persistence activity landscape for each environmental compartment. The soil data covers a larger portion of the chemical space thanks to its bigger dataset (1555) compared to water (466) and sediment (436). Indeed, areas delimited by rectangles 'a₁', 'a₂' and 'a₃' are mainly populated by compounds for which only $DT_{50(\text{soil})}$ measurements were available. For instance, area 'a₁' delimits a well-defined chemical family of 11 siloxanes, which uniquely belong to the ECHA dataset. They show rapid degradation, with average $DT_{50(\text{soil})}$ of 4 days. Areas 'a₂' and 'a₃' are populated by several insecticides and herbicides coming from the OSU, PDBB and ESDB datasets. These compounds are mainly nP (average $DT_{50(\text{soil})}$ of 22 days) and present quite heterogeneous structures, as they belong to different pesticide chemical classes, such as pyrimidine (e.g. pyroxsulam and bispyribac-sodium), sulfonylurea (e.g. bensulfuron and amidosulfuron) or triazolone (e.g.

Table 3. Examples of compounds showing high data variability for the soil compartment.

Name (use)	CAS no.	Min/Max				nP/P count
		pH	T [°C]	Humidity [%]	log DT_{50} [days]	
Diuron (herbicide)	330-54-1	4.6-7.3	15-25	35-70	1.87/3.06	4/2
1,2-Dichloropropane (fumigant insecticide)	78-87-5	5-7.1	10-30	30-75	1.70/2.84	3/1
1,2,3-Trichloropropane (fumigant insecticide)	96-18-4	5.1-8	10-30	25-75	0.43-2.36	2/1
Imazapyr (herbicide)	81,334-34-1	4.7-7.3	19-25	55	1.04/3.33	2/2
Trifluralin (herbicide)	1582-09-8	5.5-7.9	11-21	24-55	1.73-2.34	2/3
Napropamide (herbicide)	15299-99-7	5.5-8	10-20	40-45	1.82/2.59	2/2
Mepiquat chloride (plant growth regulator)	15302-91-7	5.3-7.9	10-20	40-70	1.22/3.00	2/1

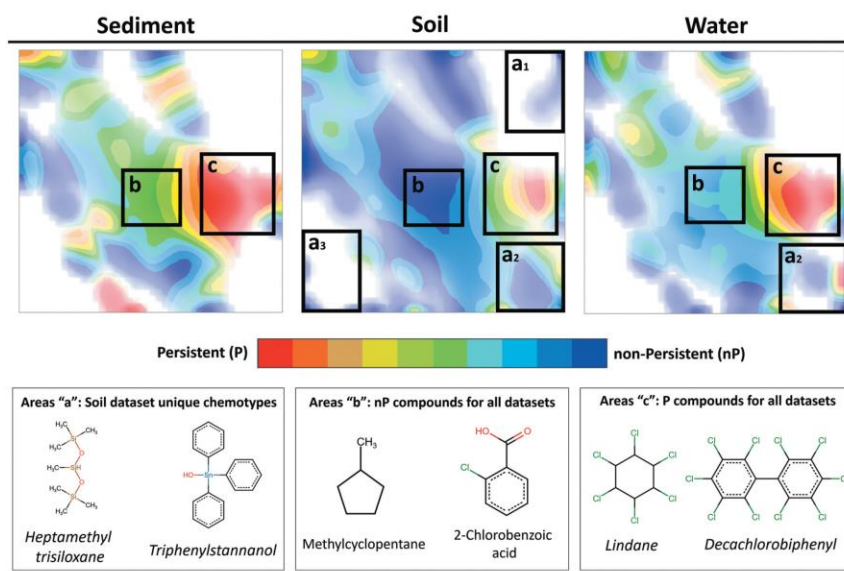


Figure 2. Class landscapes for the three environmental media. Blue regions are mainly populated by nP compounds; red ones by P compounds. White areas correspond to unpopulated regions. Black rectangles delimit map regions referred in the text.

thienicarbazone-methyl and azafedin). Chemicals in 'a₃' zone, present quite unique chemotypes, such as triphenylstannanol and some derivatives of emamectin.

Area 'b' is a densely populated area, as it concentrates roughly 40% of both water and sediment datasets. Compounds located here are generally nP, though few of them show persistence in either water or sediment medium (e.g. the herbicide fluzifop-P, with DT_{50(water)} of 45 days). Within this area, 32 chemicals show non-persistence behaviour in all media; they are characterized by: (i) relatively low molecular weight (from 32 to 166 g/mol); (ii) presence of only C, O, N elements; (iii) being mostly aliphatic with very few aromatic substructures; and (iv) the main encountered functional groups being alcohols, carboxylic acids and amines. On the contrary, area 'c' includes 58 compounds persistent in all media, with relatively high molecular weights (230 to 634 g/mol) and highly halogenated structures. These chemical features are well known to enhance persistency [36]. The GTM model was able to reproduce these known cases, which is a necessary condition to consider the map as valid. Some examples are the pesticides lindane and toxaphene, and the polychlorinated biphenyl decachlorobiphenyl.

Model performances

Table 4 reports models' performances evaluated in 5-fold CV and on the external test set (Figure 1, Step 6). The influence of chance correlation has been verified through y-scrambling: the values of the response variable (P/nP labels) are shuffled and randomly assigned to difference compounds, while the descriptor values are left intact. This procedure has

Table 4. Consensus models performances.

Compartment	5-fold CV			External test set			
	BA	Sn	Sp	BA	Sn	Sp	Data coverage [%] ^a
Sediment	0.81 (0.014)	0.81	0.80	0.91	0.92	0.90	77 (101/131)
Soil	0.71 (0.015)	0.60	0.83	0.76	0.50	0.99	76 (693/909)
Water	0.80 (0.011)	0.72	0.88	0.77	0.72	0.85	91 (128/140)

BA = Balanced Accuracy, Sn = Sensitivity, Sp = Specificity; in brackets, the standard deviation calculated on the CV repetitions is reported. ^aCalculated as the ratio of the number of compounds inside AD and the total number of compounds of the given dataset.

been repeated 15 times after reshuffling. Poor balanced accuracies values (BA = 0.49–0.51) support the significance of obtained models.

The sediment (BA_{ext} = 0.91) and water (BA_{ext} = 0.77) models are noticeably more performant in both CV and external validation than the soil model (BA_{ext} = 0.76). All the models show high specificity values (Sp > 0.85) on the external test set, reflecting their ability to discriminate true non-persistent compounds. Sensitivity values remain good for sediment and water, but drop considerably for soil, being close to random.

In an additional comparison, Table 5 reports the Soil model performances on the external set categorized according to the data provenance. The balanced accuracy and more specifically, sensitivity are degraded for the ECHA subset. Persistent compounds in soil are, in general, not correctly classified by the consensus model. This could be explained by the fact that the prior probability of a compound to be persistent is much lower in the ECHA dataset than in the other datasets (on 13 out of 140). This is maybe reflecting a bias in the constitution of the training set towards non-persistent instances compared to the situation actually observed on marketed compounds as represented by the ECHA dataset. Changing the rules of the consensus for the ECHA dataset so that a compound is considered persistent in soil if at least one individual model estimates it as persistent allows to perfectly retrieve 'P' compounds, at the expense of a lower accuracy for 'nP' instances (Sn = 1, Sp = 0.72, BA = 0.86). This conclusion is true for the other subsets as well (for instance, for the OSU: Sn = 1, Sp = 0.60, BA = 0.80). Interestingly, with this new consensus approach the average performance of the model is increased, as shown by the higher BA. However, this could be dataset dependent. Therefore, the use of the most voted class as method to ensemble individual models' predictions is a more robust and better generalizable approach.

Figure 3 depicts the performance of the selected machine learning algorithms for the given compartment. It is interesting to notice that for the sediment and water compartments, more complex higher degree algorithms (SVM and RF) scored the best performances; however, for soil the simpler Naïve Bayesian obtained reasonably good predictive power, in particular the ability to retrieve true persistent compounds, as

Table 5. Consensus models external validation performances categorized on data source.

Model	Dataset	BA	Sn	Sp	Data coverage [%]
Soil	PDBB	0.71	0.42	1	74 (572/768)
	ECHA	0.67	0.33	1	77 (107/139)
	Literature set	0.73	0.47	1	79 (244/308)
	RIVM	0.75	0.50	1	71 (83/116)
	OSU	0.72	0.45	0.99	65 (136/209)

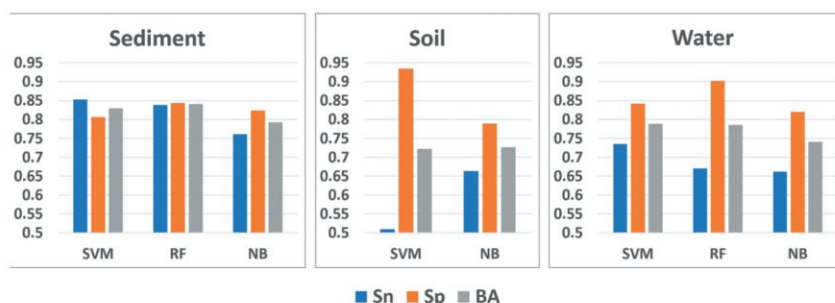


Figure 3. Selected ML algorithm performance for the given compartment.

opposed to SVM and RF. The latter algorithm was never selected (Figure 1; step 4) for the soil model, as its poor sensitivity values degraded the balanced accuracy score.

Finally, the models were rebuilt using the whole available information (merging of the training and test set) with the same descriptor spaces and methods' parameters previously determined (Tables S2-S4). Performances were evaluated in 5-fold CV, with the following results for the consensus models: $BA_{\text{sediment}} = 0.85$ ($Sn = 0.85$, $Sp = 0.84$); $BA_{\text{soil}} = 0.74$ ($Sn = 0.63$, $Sp = 0.86$); $BA_{\text{water}} = 0.81$ ($Sn = 0.71$, $Sp = 0.91$). In terms of cross-validated BA, these updated models showed a general improvement on all compartments.

Performance comparison with already-existing tools

Table 6 reports VEGA's performances on all the available data (merging of training and test set; Table 1) for the given environmental medium, both with and without considering the applicability domain filter. With the exception of the soil, data coverage on the sediment and water dataset is minimal, with only 6 and 14 compounds inside AD, respectively. All the three VEGA models do not confuse non-Persistent compounds with Persistent ones, but most Persistent compounds are erroneously classified as non-Persistent, as measured by the low sensitivity values (0.29–0.57). Without considering the AD condition, we did not notice a significant difference in BA values, with the exception of the sediment compartment, due to the very low amount (five) of compounds inside AD.

Based on these results, our models scored much better performances, with BA values ranging from 0.76–0.91 (Table 4). However, such low performances of VEGA could have been caused by the approximation that had to be done in order to convert the four output classes into a binary decision to match our collected datasets. Nevertheless,

Table 6. Performances of VEGA on the collected datasets.

Compartment	BA	Sn	Sp	Data coverage (%)
Sediment	1.00 (0.52)	1.00 (0.31)	1.00 (0.73)	5 (6/143)
Soil	0.63 (0.62)	0.25 (0.60)	1.00 (0.63)	11 (14/125)
Water	0.57 (0.57)	0.18 (0.21)	0.95 (0.92)	46 (464/996)

Values in parenthesis for BA, Sn and Sp are computed on all compounds of the given dataset, i.e. without considering the AD outcome; na = no compounds available.

concerning AD our models have a clear advantage, being able to predict most part of the external set (data coverage = 76–91 %, Table 4). The noticeable larger and diverse training sets of our models contributed to extend their ADs, and make them more suitable for risk assessment purposes.

ColorAtom analysis

Table 7 compares the selected structural alerts and the corresponding ColorAtom representation. One can see that in most of the cases, a given structural alert is highlighted on related ColorAtom graph as being important for class assignment (dark blue coloured). For instance, it is known that the dioxin and the biphenyl substructures are strongly related to a persistent behaviour. On the other hand, the coloured graphs 1, 3 and 5 show that the chlorine atoms attached to the central skeleton are the main drivers for the compound's persistency. These families are indeed well-known to persist in the environment [37]. On the contrary, functional groups such as aliphatic esters, aldehydes, carboxylic acids and hydroxyl groups (2, 4, 6) generally lead to rapid degradation [15]. Such moieties are also highlighted as important fragments in the ColorAtom graphs.

Model implementation

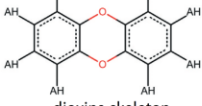

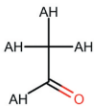

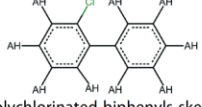

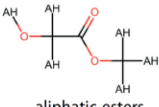
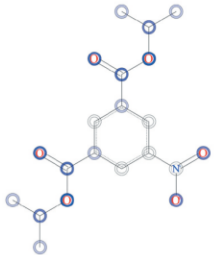
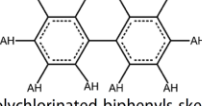
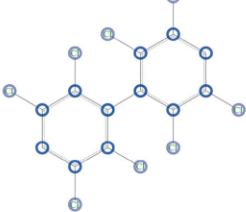
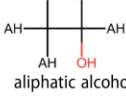
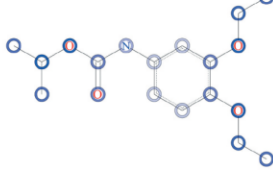
Our models are available through the online ISIDA/Predictor platform [18]: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi. The platform allows the user to either upload a batch of compounds through structure-data file format or draw a single molecule. Standardization is automatically performed. As output, the predicted value together with the number of applied models and a confidence score is provided. If selected, the ColorAtom graphs are generated as well (Figure 4).

Discussion

Data availability greatly varies depending on the environmental compartment: in abundance (soil being the largest one) and number of sources (data coming from several databases and published articles). We noticed significant differences in the experimental persistence labels for the soil compartment (nP/P, assigned by comparing the reported continuous DT_{50} values with REACH thresholds) issued from different data sources. However, these discrepancies are related almost exclusively to the P class, for which the agreement between different sources ranged from 17% to 80%. On the contrary, for the nP class, the concordance between experimental labels was higher than 90% in most of the cases (Table 2).

Low data concordance for the soil compartment stems from significant differences in measured DT_{50} data, for which the range between reported values gets up to 1.93 log unit (Table 3), meaning an uncertainty of roughly 85 days. This can completely change the classification of a given chemical under REACH. Performances of the soil consensus model (Table 4) seem to reflect this problem: the relatively low balanced accuracy ($BA_{cv} = 0.71$) is mainly caused by the difficulty to discriminate true persistent compounds in soil, as the sensitivity is only slightly better than random ($Sn_{cv} = 0.60$). On the other hand, sediment and water counterparts are noticeably more performant ($BA_{cv} = 0.80$ – 0.81). Our effort to

Table 7. ColorAtom graphs for compounds matching the given structural alert.

ID	Dataset	Structural alert	Example compound
1	Sediment	 dioxins skeleton	 CAS 30746-58-8; P
2		 aliphatic aldehyde/carboxylic acid	 CAS 103-82-2; nP
3	Soil	 polychlorinated biphenyls skeleton	 CAS 39001-02-0; P
4		 aliphatic esters	 CAS 10552-74-6; nP
5	Water	 polychlorinated biphenyls skeleton	 CAS 2136-99-4, P
6		 aliphatic alcohol	 CAS 87130-20-9; nP

For the given structural alert, the ColorAtom graph is depicted. Colour intensity 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. 'AH' stands for: 'any atom, including hydrogen'.

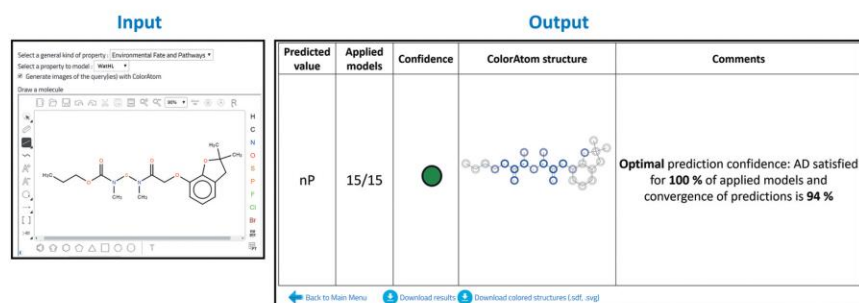


Figure 4. Predictor interface screenshot.

build the regression models using the Soil dataset, led to poor results: 5-fold CV determination coefficient $r^2 = 0.22$ and root mean squared error $RMSE = 0.70$. Latino et al. [9] built local regression models predicting soil DT_{50} including experimental assays parameters (pH, humidity, etc.) as descriptors, which led to reasonable performance ($r_{cv} = 0.63$ – 0.73). Thus, accounting for the assays conditions parameters in the modelling can be useful. We would recommend systematic recording these parameters along with DT_{50} measurements in analogy with the practice of chemical biology [38]. If the physicochemical parameters were available, then they would be interesting for the modelling. On the other hand, since the developed models would require to input additional information related to given compound, which might be difficult to acquire.

The remarkable difference of performance of the soil model, compared to sediment and water ones, is even more surprising when taking into account the size of their datasets: despite the former has a much larger training set, an improvement of its performances were not observed. We hypothesized that this could be due to following reasons:

- (i) The experimental assays in soil have more experimental conditions (such as soil texture, content of organic carbon, etc.) that need to be taken into account in order to create a subset of homogeneously determined measurements. As soil data is more abundant and was extracted from more sources than sediment and water media, this contributed to increase the uncertainty of DT_{50} measurements.
- (ii) The soil dataset is much more chemically diverse when compared to the sediment and water ones. The latter are mainly comprised by 'classic pollutants' (such as polychlorinated biphenyl, polybrominated diphenyl ethers or chlorofluorocarbons) and substances known to be easily biodegradable, such as simple alcohols or hydrocarbons. Therefore, it could be possible that persistence rules were more easily determinable for these two compartments, as opposed to soil.

The soil model has hence the disadvantage of not being able to detect truly persistent compounds but compared to the sediment and water models it has a much diverse training set which reflects its extended AD.

We identified high variability, which negatively impacted the modelling task. We suppose that there is an inherent biological variability involved but in case of persistency experiments, the differences in experimental conditions and physico-chemical properties of the tested compounds have an especially strong influence on the results. Therefore, the assignment of a P label is very dependant of the results interpretation by the experts, which induce another source of variability. So, the performance of the model in determining if a substance is persistent is comparable to experimental testing and reflects the difficulty to conclude on persistency even with reliable experimental data.

In our modelling approach we did not consider the vP criterion. We decided indeed to simplify the modelling task by considering only two classes due to: (i) the high variability of measurements; and (ii) lack a meaningful number of instances for the sediment and water media that could be used to attribute the third very-persistent class. Therefore, the lack of possibility to further discriminate between persistent and very-persistent compound denotes a limitation of our models, which could be overcome in the future with the generation of new persistence data.

Conclusions

In this work we report binary classification consensus models for persistence in sediment, soil and water environmental media. Overall, we collected a total of 1579 unique compounds, annotated by, at least, one experimental value for a given medium. This included subsets of 1533, 466 and 436 compounds which persistence was measured in soil, water and sediment, respectively.

Analysis of class landscapes on Generative Topographic Map helped us to identify some chemotypes corresponding to persistent (or non-persistent) compounds in all three media or in particular medium only.

The sediment ($BA_{cv} = 0.81$, $BA_{ext} = 0.91$) and water ($BA_{cv} = 0.80$, $BA_{ext} = 0.77$) consensus models are noticeably more performant in both CV and external validation than the soil model ($BA_{cv} = 0.71$, $BA_{ext} = 0.76$). Lower BA values for the soil are degraded due to the noticeably lower sensitivity value, reflecting the low agreement of reported experimental values among data sources. High data variability negatively influenced model performances. We believe that more predictive models could be built with the inclusion of additional parameters related to the experimental test conditions. These variables should be reported along with the test results to allow a better selection of reliable data.

Finally, the models were rebuilt on the entire sets resulted from the merging of related training and test set. In cross-validation, the new models demonstrated better performance for all three compartments ($BA_{sediment} = 0.85$; $BA_{soil} = 0.74$; $BA_{water} = 0.81$) and benefited from an enlarged applicability domain.

As only one tool (VEGA) is currently available to predict environmental media persistence, our models should be a useful addition for risk assessment and PBT classification purposes. As advantage, new models have a noticeably larger training set and were updated with the addition of recently published data which extends their applicability domain.

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. Azam  <http://orcid.org/0000-0002-2974-2484>
 M.H. Enrici  <http://orcid.org/0000-0001-5696-4376>
 E. Van Miert  <http://orcid.org/0000-0001-6653-1371>
 A. Varnek  <http://orcid.org/0000-0003-1886-925X>

References

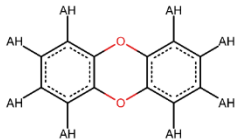
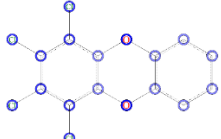
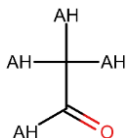
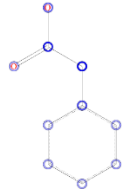


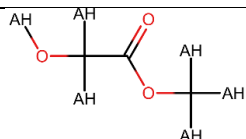
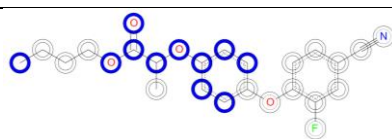
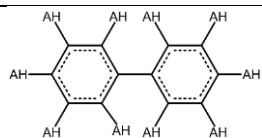
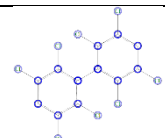
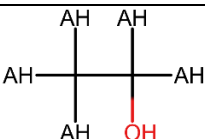

- [1] T. Junker, A. Coors, and G. Schüürmann, *Compartment-specific screening tools for persistence: Potential role and application in the regulatory context*, Integr. Environ. Assess. Manag. 15 (2019), pp. 470–481. doi:10.1002/ieam.4125.
- [2] R.J. Larson and C.E. Cowan, *Quantitative application of biodegradation data to environmental risk and exposure assessments*, Environ. Toxicol. Chem. 14 (1995), pp. 1433–1442. doi:10.1002/etc.5620140821.
- [3] ECETOC, *Persistence of chemicals in the environment*, Tech. Rep. 90, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, BE, 2003.
- [4] 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.
- [5] 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.
- [6] F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert, and A. Varnek, *Modelling of ready biodegradability based on combined public and industrial data sources*, SAR QSAR Environ. Res. 31 (2020), pp. 171–186. doi:10.1080/1062936X.2019.1697360.
- [7] OECD, *Test No. 301: Ready biodegradability*, Tech. Rep. 9789264070349, Organisation for Economic Co-operation Development, Paris, FR, 1992.
- [8] OECD, *OECD guidelines for the testing of chemicals, section 3: Environmental fate and behaviour*, Organisation for Economic Cooperation and Development, Paris, FR, 2019. Available at <http://www.oecd.org/env/ehs/testing/oecdguidelinesforhetestingofchemicals.htm>.
- [9] D.A.R.S. Latino, J. Wicker, M. Gütlein, E. Schmid, S. Kramer, and K. Fenner, *Eawag-soil in envipath: A new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data*, Environ. Sci. Process. Impacts 19 (2017), pp. 449–464. doi:10.1039/C6EM00697C.
- [10] M. Pavan and A.P. Worth, *Review of estimation models for biodegradation*, QSAR Comb. Sci. 27 (2008), pp. 32–40. doi:10.1002/qsar.200710117.
- [11] K. Acharya, D. Werner, J. Dolfing, M. Barycki, P. Meynet, W. Mrozik, O. Komolafe, T. Puzyn, and R.J. Davenport, *A quantitative structure-biodegradation relationship (QSBR) approach to predict biodegradation rates of aromatic chemicals*, Water Res. 157 (2019), pp. 181–190. doi:10.1016/j.watres.2019.03.086.
- [12] M. Salahinejad, E. Zolfonoun, and J.B. Ghasemi, *Predicting degradation half-life of organophosphorus pesticides in soil using three-dimensional molecular interaction fields*, Int. J. Quant. Struct. Relat. 2 (2017), pp. 27–35.

- [13] K. Samghani and M. HosseinFatemi, *Developing a support vector machine based QSPR model for prediction of half-life of some herbicides*, *Ecotoxicol. Environ. Saf.* 129 (2016), pp. 10–15. doi:10.1016/j.ecoenv.2016.03.002.
- [14] 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.
- [15] F. Pizzo, A. Lombardo, M. Brandt, A. Manganaro, and E. Benfenati, *A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH*, *Environ. Int.* 88 (2016), pp. 250–260. doi:10.1016/j.envint.2015.12.019.
- [16] 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.
- [17] US EPA, *Estimation Programs Interface Suite™ for Microsoft® Windows V4.11*, US Environmental Protection Agency, Washington DC, 2012; software available at <https://www.epa.gov/tsca-screening-tools/epi-suite-estimation-program-interface>.
- [18] G. Marcou, D. Horvath, F. Bonachera, and A. Varnek, *Laboratoire de Chimoinformatique UMR 7140 CNRS*, University of Strasbourg, Strasbourg, FR, 2019. available at <http://infochim.u-strasbg.fr/>.
- [19] RIVM, *Pesticides: Benefaction or Pandora's Box? A synopsis of the environmental aspects of 243 pesticides*, Tech. Rep. 679101014, National Institute of Public Health and Environmental Protection, Bilthoven, NL, 1994.
- [20] NPIC, *Data from: OSU extension pesticide properties database*. National Pesticide Information Center; dataset available at <http://npic.orst.edu/ingred/ppdmove.htm>.
- [21] 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>.
- [22] T. Gouin, I. Cousins, and D. Mackay, *Comparison of two methods for obtaining degradation half-lives*, *Chemosphere* 56 (2004), pp. 531–535. doi:10.1016/j.chemosphere.2004.04.018.
- [23] 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.
- [24] UH, *Data from: PPDB: Pesticide properties database*. Agriculture & Environment Research Unit (AERU), University of Hertfordshire.
- [25] 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.
- [26] 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.
- [27] 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.
- [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] 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.
- [30] C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, *ACM Tran. Int. Syst. Technol.* 2 (3) (2011), pp. 1–27. doi:10.1145/1961189.1961199.
- [31] 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.
- [32] 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/qsar.200390007.

- [33] 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](https://doi.org/10.1080/1062936X.2019.1626278).
- [34] F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert, and A. Varnek, *Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context*, SAR QSAR Environ. Res. 30 (2019), pp. 879–897. doi:[10.1080/1062936X.2019.1672089](https://doi.org/10.1080/1062936X.2019.1672089).
- [35] 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](https://doi.org/10.1002/minf.201100136).
- [36] R.S. Boethling, E. Sommer, and D. DiFiore, *Designing small molecules for biodegradability*, Chem. Rev. 107 (2007), pp. 2207–2227. doi:[10.1021/cr050952t](https://doi.org/10.1021/cr050952t).
- [37] E. Papa and P. Gramatica, *Screening of persistent organic pollutants by QSBR classification models: A comparative study*, J. Mol. Graph. Model. 27 (2008), pp. 59–65. doi:[10.1016/j.jmgm.2008.02.004](https://doi.org/10.1016/j.jmgm.2008.02.004).
- [38] S. Abeyruwan, U.D. Vempati, H. Küçük-McGinty, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P.A. Clemons, S. Brudz, A. Siripala, A.J. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, and S.C. Schürer, *Evolving BioAssay Ontology (BAO): Modularization, integration and applications*, J. Biomed. Semant. 5 (2014), pp. S5. doi:[10.1186/2041-1480-5-S1-S5](https://doi.org/10.1186/2041-1480-5-S1-S5).

Erratum

Table 7. ColorAtom graphs matching the given structural alert.

ID	Dataset	Structural alert	Example compound
1	Sediment	 <p>dioxins skeleton</p>	 <p>CAS 30746-58-8; P</p>
2		 <p>aliphatic aldehyde/carboxylic acid</p>	 <p>CAS 103-82-2; nP</p>
3	Soil	 <p>polychlorinated biphenyls skeleton</p>	 <p>CAS 39001-02-0; P</p>
4		 <p>aliphatic esters</p>	 <p>CAS 10552-74-6; nP</p>
5	Water	 <p>polychlorinated biphenyls skeleton</p>	 <p>CAS 2136-99-4, P</p>
6		 <p>aliphatic alcohol</p>	 <p>CAS 87130-20-9; nP</p>

4.1.4 Short-term aquatic toxicity on Algae, Daphnia and Fish

In the frame of REACH, the evaluation of acute aquatic toxicity on aquatic plants (Algae) and invertebrates (Daphnia) is mandatory for all substances that are manufactured or imported above 1 ton per year; while acute toxicity on vertebrates (Fish) is required when the substance surpasses the 10 tons/year cut-off. Algae, Daphnia and Fish organisms belong to different trophic levels, and have been considered as representative for the aquatic ecosystem. Briefly, the test organisms are exposed to the studied substance via contaminated water media, and the following effects are measured: (i) for Algae, the substance's growth inhibition effect, expressed as median effective concentration (EC50) measured at 72 hours; (ii) for Daphnia, immobilization at 48 hours and expressed as median effective concentration (EC50); (iii) for Fish, the median lethal concentration measured at 96 hours (LC50).

These endpoints have been extensively studied in the past years. However, many of these models can be defined as so-called “local models”, whose training set is restricted to few tens of compounds belonging to specific chemical families. They generally have better prediction accuracy, but their narrow applicability domain limits their use for risk assessment purposes. On the other hand, a “global model” has a much larger and more chemically diverse training set.

In this work, we aimed at building new acute aquatic toxicity models on each species based on the most comprehensive collection of available data, issued from merging several public data sources. Models were externally validated on two test sets: the former was created by splitting available public data, while the latter comprised proprietary industrial data. Performances on the former datasets were acceptable (RMSE = 0.56 – 0.78), similar to those determined by cross-validation. On the other hand, prediction accuracy on the Industrial sets were noticeably worse (RMSE = 0.92 – 1.12). The main issue was the overestimation of the toxicity of several small molecular weight molecules (absolute errors higher than 1.5 log units). It is hypothesised that these errors are due to uncertainties in experimental data and to specificities of the electronic structures that are insufficiently represented by the molecular graph of the molecules.

In addition, a benchmarking on the Industrial sets have been carried out considering the ECOSAR, VEGA and TEST freely available tools: our models scored one of the best prediction accuracies coupled with a good data coverage.

Finally, public and industrial data were merged, and models were updated: the final models' training sets are considerably larger (1806, 2529, 2591 for Algae, Daphnia and Fish, respectively) than those of already existing tools, thus extending their applicability domain. In cross-validation, these models showed R2 values of 0.60, 0.72, 0.71 and RMSE values of 0.71, 0.71, 0.69 for Algae, Daphnia and Fish, respectively.



4.1.4 Toxicité aquatique aigue pour algues, daphnies et poissons

Dans le cadre de REACH, l'évaluation de la toxicité aquatique aiguë sur les plantes aquatiques (algues) et les invertébrés (daphnies) est obligatoire pour toutes les substances produites ou importées à plus de 1 tonne par an; tandis qu'une toxicité aiguë sur les vertébrés (poisson) est requise lorsque la substance dépasse le seuil de 10 tonnes / an. Les algues, les daphnies et les poissons appartiennent à différents niveaux trophiques et ont été considérés comme représentatifs de l'écosystème aquatique. Pour résumer, les organismes choisis dans un essai sont exposés à la substance étudiée via un milieu contaminé et les effets suivants sont mesurés: (i) pour les algues, l'objectif est de déterminer la capacité de la substance à inhiber la croissance, exprimée en concentration efficace médiane (CE50) mesurée après 72 heures; (ii) pour la Daphnie, c'est l'immobilisation qui est observée après 48 heures et exprimée en concentration efficace médiane (CE50); (iii) pour le poisson, la concentration létale médiane est mesurée après 96 heures (CL50).

Ces paramètres ont été largement étudiés au cours des dernières années. Cependant, bon nombre de ces modèles peuvent être définis comme des «modèles locaux», dont le jeu d'entraînement est limité à quelques dizaines de composés appartenant à une même famille chimique. Bien qu'ils aient généralement une meilleure précision, leur domaine d'applicabilité étroit limite leur intérêt pour l'évaluation des risques. D'un autre côté, un «modèle global» est construit sur un jeu d'apprentissage beaucoup plus vaste et plus diversifié chimiquement.

Dans ce travail, nous avons cherché à construire de nouveaux modèles de toxicité aquatique aiguë sur chaque espèce à partir de la collection de données la plus exhaustive possible, issue de la fusion de plusieurs sources de données publiques. Les modèles ont été validés par une procédure de validation externe sur deux ensembles de tests: le premier est un sous-ensemble des données publiques disponibles, tandis que le second inclus des données industrielles propriétaires. Les performances sur le premier jeu de données étaient acceptables ($RMSE = 0,56 - 0,78$) et similaires à celles déterminées par validation croisée. En revanche, la précision des prédictions sur le second jeu de données, incluant les données industrielles, étaient sensiblement moins bonnes ($RMSE = 0,92 - 1,12$). Ces résultats traduisent une surestimation de la toxicité pour plusieurs molécules de petit poids moléculaire (erreurs absolues supérieures à 1,5 unités logarithmiques). Nous supposons que ces erreurs sont dues à des incertitudes dans les données expérimentales et aux spécificités des structures électroniques insuffisamment représentées par le graphe moléculaire des molécules.

De plus, une analyse comparative des modèles ECOSAR, VEGA et TEST a été réalisée sur les données industrielles. Nos modèles se sont révélés parmi les plus performant tout en bénéficiant d'une bonne couverture des données.

Enfin, les données publiques et industrielles ont été fusionnées et les modèles ont été mis à jour: les jeux de données d'entraînement des modèles finaux sont considérablement plus grands (1806, 2529, 2591 pour Algae, Daphnia et Fish, respectivement) que ceux des outils déjà existants, étendant ainsi leur domaine d'applicabilité. En validation croisée, les performances de ces modèles ont été mesurées par des valeurs de R^2 de 0,60, 0,72, 0,71 et des valeurs de RMSE de 0,71, 0,71, 0,69 pour les algues, la daphnie et le poisson, respectivement.



Consensus QSAR models estimating acute toxicity to aquatic organisms from different trophic levels: algae, *Daphnia* and fish

F. Lunghini^{a,b}, G. Marcou^a, P. Azam^b, M.H. Enrici^b, E. Van Miert^b and A. Varnek^a

^aLaboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France; ^bToxicological and Environmental Risk Assessment Unit, Solvay S.A., St. Fons, France

ABSTRACT

We report new consensus models estimating acute toxicity for algae, *Daphnia* and fish endpoints. We assembled a large collection of 3680 public unique compounds annotated by, at least, one experimental value for the given endpoint. Support Vector Machine models were internally and externally validated following the OECD principles. Reasonable predictive performances were achieved ($RMSE_{ext} = 0.56\text{--}0.78$) which are in line with those of state-of-the-art models. The known structural alerts are compared with analysis of the atomic contributions to these models obtained using the ISIDA/ColorAtom utility. A benchmarking against existing tools has been carried out on a set of compounds considered more representative and relevant for the chemical space of the current chemical industry. Our model scored one of the best accuracy and data coverage.

Nevertheless, industrial data performances were noticeably lower than those on public data, indicating that existing models fail to meet the industrial needs. Thus, final models were updated with the inclusion of new industrial compounds, extending the applicability domain and relevance for application in an industrial context. Generated models and collected public data are made freely available.

ARTICLE HISTORY

Received 28 May 2020

Accepted 15 July 2020


KEYWORDS

QSAR/QSPR; acute aquatic toxicity; generative topographic mapping (GTM); REACH

Introduction

The determination of acute aquatic toxicity is a key parameter under the European Union Regulation for the Registration, Evaluation, Authorisation and Restriction of Chemical Substances (REACH, EC No 1907/2006) [1]. In this context, the evaluation of the acute toxicity towards aquatic algae and invertebrates (*Daphnia*) is mandatory for all substances that are manufactured or imported above 1 ton per year; while acute toxicity on vertebrates (fish) is required when the substance surpasses the 10 tons/year cut-off. Algae, *Daphnia* and fish organisms belong to different trophic levels and have been considered as representative for the aquatic ecosystem [2]. These ecotoxicity tests are performed

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.2020.1797872>.

© 2020 Informa UK Limited, trading as Taylor & Francis Group

according to the OECD guidelines no. 201 (Algae), 202 (*Daphnia*) and 203 (Fish) [3]. Briefly, the test organisms are exposed to the study substance via contaminated water media, and the following effects are measured: (i) for Algae, the purpose is to determine the substance's growth inhibition effect, expressed as the median effective concentration (EC_{50}) measured at 72 hours; (ii) for *Daphnia*, mortality, which is evaluated by the immobilization of the invertebrate is recorded at 48 hours and expressed as the median effective concentration (EC_{50}); (iii) for Fish, the median lethal concentration measured at 96 hours is considered (LC_{50}).

These endpoints have been of strong interest for QSAR development in the past years. However, many of these models can be defined as 'local models', whose training set is restricted to few tens of compounds belonging to a particular chemical family [4–8]. Although this approach generally induces a better prediction accuracy for the specific chemical family, it also narrows the applicability domain and limits the use for risk assessment purposes. On the other hand, a 'global model' has a much larger and more chemically diverse training set. Table 1 reports several already-published (global) models on these three endpoints. Only models with a training set of at least 200 compounds, and for which an external validation was carried out, have been considered. Some of them have been implemented in three freely available tools: (i) Ecological Structure Activity Relationships (ECOSAR) available through the Estimation Program Interface (EPI Suite) program [9]; (ii) Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA) [10]; (iii) and Toxicity Estimation Software Tool (T.E.S.T.) [11].

Fish acute toxicity is the endpoint for which models are the most numerous. Similarly, data availability follows the same trend, with fish models having the largest training sets compared to Algae and *Daphnia* ones. In this respect, the results reported by Sheffield et al. [12] are quite interesting. The authors aimed to build a model estimating acute fish toxicity based on the biggest possible amount of available data,

Table 1. Existing models and tools for acute aquatic toxicity prediction.

Endpoint	Descriptors	Algorithm	Training set	Model performance			Ref.
				Test set	r^2	RMSE	
A*	CDK	Consensus	330	-	0.75–0.79	0.56–0.64	[2]
D	CDK	Consensus	426	-	0.66	0.67	[2]
F	Dragon	MLR	771	192	0.64	-	[2]
A (VEGA) ^a	Dragon	SVM	252	109	0.64	-	[10]
D (VEGA) ^a	2D	Various	220–269	43–68	0.49–0.68	1.02–1.49	[10]
F (VEGA) ^a	2D	Various	564–652	164–382	0.54–0.64	0.89–0.90	[10]
D (TEST) ^a	2D	Various	353	-	0.74	0.91	[11]
F (TEST) ^a	2D	Various	823	-	0.73	0.77	[11]
A	Dragon	PLS	251	83	-	0.67	[13]
D	2D and 3D	PLS	222	-	0.56–0.72	-	[14]
D	Dragon	Knn	436	110	0.43–0.72	-	[15]
D	Fragments	PNN	621	-	0.59–0.71	-	[16]
D (ECOSAR) ^a	logP	MLR	-	480	0.44	-	[9,17]
F	2D	SVM	457	114	0.80	0.51	[18]
F	2D	Knn	726	182	-	0.68–0.89	[19]
F	Dragon	MLR	841	280	0.63	0.80–0.83	[20]
F	Padel	Consensus	2124	-	0.58–0.66	0.81–0.89	[12]
F (ECOSAR) ^a	logP	MLR	-	532	0.23	1.07	[9,12]

* A, D, F = algae, *Daphnia*, fish; PLS = partial least squares; CDK = chemistry development kit; Knn = k-nearest neighbours; MLR = multiple linear regression; PNN = probabilistic neural network; SVM = support vector machine; r^2 = coefficient of determination; RMSE = root mean squared error; ^afreely-available tool implementing several models.

therefore without creating a subset of data experiments performed under homogeneous study conditions (e.g. all fish species considered, merging of different LC_{50} values taken at different times, etc.). The performance of this global model was comparable to that of more specific models. In this work, following a similar approach, we aimed to build new acute aquatic toxicity models based on the most comprehensive collection of available data from merging several public data sources. However, we filtered the data in order to obtain subsets of toxicity values determined under homogeneous test conditions, hypothesizing that a more careful data selection could help to improve model performances.

Support Vector Machine models were generated, internally and externally validated following the OECD principles [21] and ensembled in consensus. Additional external validation has been carried out on a set of compounds comprising data provided by Solvay ('Industrial set'). Finally, public and industrial data sources have been merged in order to build the most comprehensive models we could obtain, with training sets of 1806 (Algae), 2526 (*Daphnia*) and 2591 (Fish) compounds. These latter models have extended applicability domains compared to previously published models thanks to the noticeably bigger training sets and they include a significant subset of compounds containing industry-relevant chemotypes representative of the diversity of chemicals found in Chemical world.

Our models are available through the online In Silico Design and data Analysis (ISIDA)/Predictor platform [22], available at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi. The Chrome navigator is preferred for using this service.

Methods

Data sources

Experimental ecotoxicological data have been collected from multiple sources: (i) the European Chemicals Agency (ECHA) database accessed through the eChem Portal [23], (ii) the Environmental Protection Agency Fathead Minnow Acute Toxicity dataset (EPAFHM; <https://pubchem.ncbi.nlm.nih.gov/bioassay/1188>), (iii) the Japanese National Institute of Technology and Evaluation (NITE; <https://www.nite.go.jp/en/>), the (iv) ECOTOXicology knowledgebase database (ECOTOX) [24], (v) the Aquatic OASIS database (<http://oasis-lmc.org/>), the (vi) the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC; <http://www.ecetoc.org/>), (vii) the European Food Safety Authority database (EFSA; <https://www.efsa.europa.eu>), (viii) extraction from the work of Cassani et al. [25], Toropov et al. [26], Singh et al. [2], Wu et al. [27], Furuhashi et al. [28], Khan et al. [7,13] and the training sets of VEGA and T.E.S.T. tools (hereafter referred as '*Literature sets*') and finally (ix) data provided by Solvay (referred as '*Industrial sets*'), which are partly proprietary. The latter naming is referred to the ensemble of the three sets of industrial compounds available for Algae, *Daphnia* and Fish. Data coming from sources iii, iv, v, vi and vii were extracted from the QSAR Toolbox software (v.4.3) [29]. Data from the QSAR Toolbox were extracted with 'database search' function, by querying all the databases under the section 'Ecotoxicological information' (ECETOC, Japan MoE, OASIS, EFSA).

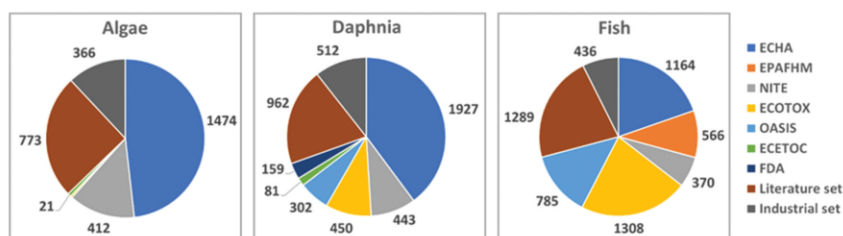


Figure 1. Data distribution for the given endpoint. The numbers refer to the amount of unique compound for the given data source, but there are overlaps between the data sources.

Table S1 report more information concerning database cardinality; **Figure 1** graphically resumes these findings. As illustrated, ECHA data represents a significant part of the available data.

Data cleaning

Raw data processing and standardization were done with workflow implemented in the Konstanz Information Miner (KNIME) software [30]. Retrieved data were cleaned in order to retain only measurements taken under similar experimental conditions as required by the OECD guideline of the given endpoint. The following criteria were taken into account: the test organism must be in the list with those recommended by the guideline; the measured effect must be growth inhibition (growth rate) measured at 72 hours for algae; and mortality measured at 48/96 hours for *Daphnia*/fish; the value must be precisely determined (i.e. values reported as ranges were excluded); the organism life stage must be in line with the recommended stage in the OECD guideline (e.g. toxicity studies performed on organisms at the larval stage were discarded). The PubChem [31] online service was queried to verify SMILES correctness. Generated SMILES were 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. In case of multiple values per compound, the median was taken as a representative value. When the ratio between the minimum and the maximum reported value per compound was >10, the entry was discarded. Finally, compounds with toxicity values higher than the experimentally determined water solubility and those which are unstable in water were excluded (these experimental values were retrieved from the ECHA database). We could perform this comparison for 781, 1045 and 743 compounds for Algae, *Daphnia* and Fish, respectively. LC_{50} and EC_{50} values originally expressed in mg/l were transformed to the inverse log of the molar dose (pLC_{50} and pEC_{50} in mMol/L).

Overall, we collected a total of 3680 unique public compounds, annotated by, at least, one experimental value for a given trophic level. This included subsets of 1440, 2120 and 2110 compounds whose acute toxicity was measured for Algae, *Daphnia* and Fish, respectively. Datasets are freely accessible on Zenodo: 10.5281/zenodo.3708082.

Training and test set preparation

In this work, two types of models were generated: (i) 'ECHA models' built on data coming exclusively from the ECHA database; and (ii) 'All-Public models' built on all collected public data. We took this decision under the hypothesis that the former database comprised data of higher quality, as a reliability evaluation (i.e. the Klimisch score [32]) is performed by the registrants. In both cases, the model's training and test sets were obtained by stratified (based on toxicity values) random splitting 70% and 30%, respectively. Table 2 summarizes the composition of the training and test sets for the given model, as well as the Industrial sets, which were used as additional external validation (these results are depicted by Figure S1). Industrial set compounds already inside the models' training sets have been excluded. For each endpoint, the external test set is composed of industrial set compounds that are neither in the ECHA set nor in the All-Public set.

Molecular descriptors

ISIDA Property-Label Molecular descriptors [33] were employed. Several tens of ISIDA descriptor spaces (DS) corresponding to molecular fragment of different sizes and topologies were generated. Among this entire pool, DSs that led to the generation of underperforming models were filtered out (see Model generation and validation section). The number of fragments varies as a function of the selected DS. It ranged from 470 ('IAB(2-4)', Type I sequences of atoms and bonds of length up to four) for the ECHA Fish model to 26405 ('IIAB(2-7)', Type II atom-centred fragments considering atoms and bonds with radius up to 7) for the All-Public Fish model. More details are given in Supporting Information (Tables S2–S4).

Generative topographic mapping

Following an analogous methodology as described in our previous works [34–36] we use Generative Topographic Mapping (GTM) [37,38] for data visualization approaches. Two types of landscapes have been prepared: (i) the *density landscape* which, as the name suggests, assigns a colour scale to the map depending on the amount of compounds populating the given position; (ii) the *property landscapes* which colour the maps according to the envisaged property, i.e. the toxicity values for three datasets of Algae, *Daphnia* and Fish.

The manifold [37] was built on the whole available chemical space, i.e. by merging all the collected datasets of Algae, *Daphnia* and Fish. Genetic algorithm (GA) [39] was

Table 2. Dataset compositions.

Endpoint	ECHA			All-Public			Industrial set	
	Training	Test	PR*	Training	Test	PR	Size	PR
Algae	625	268	−0.88–4.84	1007	433	−2.00–5.99	249	−2.52–5.55
<i>Daphnia</i>	844	363	−0.97–4.52	1484	636	−2.51–8.75	228	−2.94–4.76
Fish	588	253	−2.00–3.98	1450	660	−2.4–6.51	193	−3.1–4.02

*PR = pEC₅₀ for algae and *Daphnia* and pLC₅₀ for fish property range.

employed to select the most suitable ISIDA descriptor space (among the same pool used for model generation) and GTM parameters. The following GTM parameters were suggested: the number of radial basis function centres $m = 4 \times 4$, the RBFs width $w = 0.3$ and the map resolution $k = 17 \times 17$ grid nodes. Atom centred fragments with the radius of 3, IIAB(2–3), were used as descriptors.

Model generation and validation

The modelling workflow is depicted in Figure 2. Among other tested machine learning algorithms, including random forest and k-nearest neighbours, support vector machine (SVM) with radial basis function (RBF) kernel was employed as it scored the best performances. SVM models were generated with libSVM (v. 3.22) [40]. For a given training set (Table 2), the ‘best’ ISIDA descriptor spaces and the optimal SVM hyperparameters (cost and gamma) have been selected in the genetic algorithm-driven optimization process. Cross-validated determination coefficient was selected as fitness function. In such a way, the top 15 DS spaces were retained (according to our experience, this threshold is a good compromise between performance and computational speed) and used to fit an equal

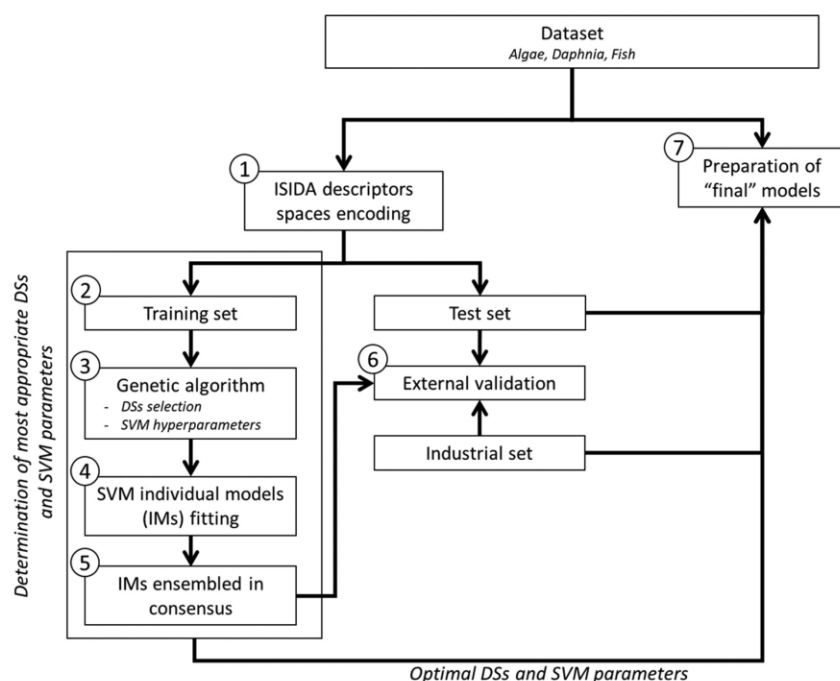


Figure 2. Model building workflow. (1) ISIDA descriptor spaces (DS) are generated; (2) the dataset is split in training and test sets; (3) optimal DSs and SVM hyperparameters are chosen by genetic algorithm; (4) SVM individual models (IM) are fitted; (5) retained models are ensembled in consensus and (6) externally validated on the Test and Industrial sets; (7) selected DSs and SVM parameters are used to train the ‘final’ models, i.e. on the whole available data.

number of optimized SVM Individual Models (IM). Internal validation of each IM was carried out by fivefold CV repeated 10 times (10*5CV) after data reshuffling. Statistics were assessed for each repetition followed by their averaging. The influence of chance correlations was checked through Y-scrambling [41] (with 50 repetitions). The 15 selected IMs were then ensembled in consensus, and external validation has been carried out on the given test set and, in addition, on the Industrial set (Table 2). Tables S2–S4 report detailed information concerning consensus model set-up (employed DS and SVM parameters for the given IM) for each endpoint. Finally, related training, test and Industrial sets were merged, and models were updated using the same DSs and SVM parameters. Validation was carried out, as previously, in 10*5CV. To evaluate the models' performance the r^2 determination coefficient and the RMSE parameters are reported (formulas are reported in Table S5).

Applicability domain and ensemble modelling

The 'fragment control' assessment [34] is employed as a method to verify a model's applicability domain (AD): if the test molecule has a fragment not present in the training set, it is considered as 'out-of-AD'. Generated models were assembled in consensus whose outcome corresponded to the mean among the IM predicted values, without any consideration of out-of-AD predictions. Given the different types of descriptors employed, each individual model has its own applicability domain, which is checked independently. In addition, we propose a 4-grade reliability scale system [35] based on the percentage of models with positive AD outcome. Briefly, depending on the percentage of individual models for which the compound was inside the AD a score of Low ($\leq 25\%$); Average (25%–50%); Good (50%–80%); or Optimal ($\geq 80\%$) was assigned. A compound is considered to be inside the AD when its reliability is higher than Low.

State-of-the-art model comparison

The ECOSAR, VEGA and T.E.S.T. freely available tools were challenged to predict the Industrial set compounds and benchmarked against the ECHA and All-Public models. The following conditions were taken in account: (i) molecules already present in the model's training set were excluded; (ii) out-of-AD or low-reliability predictions were not considered; (iii) as VEGA includes several models for each endpoint, performances are expressed as ranges. Conditions (i) and (ii) can hardly be respected for ECOSAR as the training sets are not readily available and AD evaluation is not automatically performed by the software. As ECOSAR can report more than one predicted value depending on the chemical class of the query compound, the lowest toxicity value (more conservative approach) was selected in such instances.

ColorAtom structural-activity dependence analysis

The ColorAtom [42] can be used as support to better interpret the model output [34,35]. This utility assigns a colour code to each fragment or atom depending on whether it was correlated to an increase (blue) or decrease (red) of the predicted value.

As an example of application, we selected some compounds which contained a structural alert (SA) reported to be associated to high aquatic toxicity, as identified in the work of Gini et al. [43], which were used to generate ColorAtom graphs.

Results

Experimental data inter-database variability

As data were collected from multiple sources, it was possible to compare experimental values of overlapping compounds between a given database pair. Figure 3 reports some (only the first three sources that had the highest number of common elements were considered) of these pairwise comparisons. This analysis has been performed before duplicate removal: therefore, the same compounds with more than one experimental value could be present in the given source. In such a way, it was possible to identify compounds for which determined toxicity values spanned over a large range of log units (e.g. the groups of points disposed horizontally or vertically which can be well seen in graphs 4, 7 and 9). Fish datasets had the highest number of compounds common to several sources when compared to algae and *Daphnia* ones. Moreover, the ECHA database always displays the highest number of compounds common to other sources, suggesting that it is one of the most exhaustive sources. Despite that a relatively large RMSE has been found when comparing experimental values among the different databases (RMSE up to 0.80 log unit), most data sources showed a good correlation ($r^2 = 0.50$ – 0.95). This is consistent with the error of already published and of our new

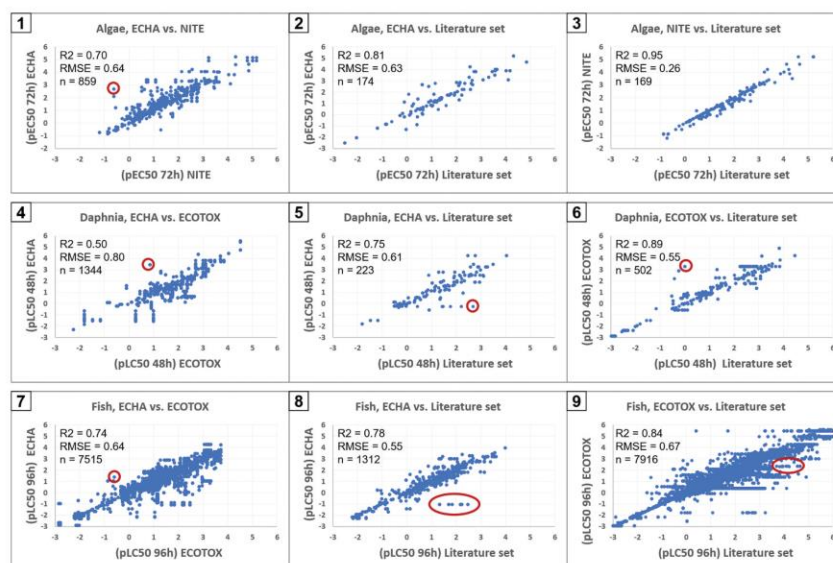


Figure 3. Experimental values comparison of inter-databases overlapping compounds. n = number of overlapping compounds for the given pair. Red circles mark some compounds reported as examples in Table 3.

models, with cross-validated RMSE values ranging from 0.56 to 0.78 (see results section). Indeed, the precision of aquatic toxicity tests is influenced by several conditions, such as the employed species, the media pH and temperature, use of solvent, whether analytical monitoring was performed, etc. Several studies reported that the inter- and intra-laboratory variability could be noticeably high, up to 3 log units [17,44–47].

Compounds marked by red ellipses (Figure 3) have been taken as examples and are reported in Table 3. We noticed that small molecules (e.g. acrolein or 2-mercaptoethanol) tend to have highly discordant toxicity values. This may be due to the fact that such compounds are generally volatile, and therefore it is technically more difficult to precisely determine their concentration in the water media, especially without an analytical verification of concentrations, which was not always performed or reported.

Chemical space analysis using GTM

Figure 4 shows the density landscape for the whole chemical space, i.e. the merging of the algae, *Daphnia* and fish datasets. The colour scale refers to the number of compounds populating a given region, ranging from 0 (white areas) up to 40 (yellow areas). The most densely populated area (zone 1) is dominated by the presence of methylated benzene structures with varying substituents, such as -nitro or -sulpho (e.g. CAS 25241-16-1 and 99-51-4). Similarly, zone 2 exhibits high density, being mainly populated by aliphatic and aromatic compounds with ester or ether functional group (e.g. CAS 105-53-3 and 103-60-6). Across this zone, there is the tendency to increase the length of the aliphatic chain and increase the aromaticity when moving horizontally. Another densely populated area is zone 6 (red rectangle), where aliphatic and aromatic alcohols can be found (e.g. CAS 111-27-3 and 4130-42-1). Similarly, it is possible to see a trend of increasing branching through this area. Zones 3, 4 and 5 delimit the well-defined chemical families of highly fluorinated aliphatic compounds (e.g. CAS 686-83-3), chlorinated phenols (e.g. CAS 88-06-2) and amines (e.g. CAS 280-57-9), respectively. Finally, in the low-density regions, it is possible to find compounds which present 'rare' chemotypes: as their structure is noticeably different from the rest of the chemical space, they are projected into a relatively isolated location (e.g. 116-95-0; identified by the black dot).

Figure 5 shows the property landscape for each endpoint. It can be noticed that the algae landscape possesses larger portions of white areas, as opposed to the *Daphnia* and fish ones. This indicates that the former dataset is lacking some chemotypes, reflecting its smaller size (Table 2). Areas delimited by rectangles '1' and '2' are populated by compounds that exhibit high toxicity (pLC₅₀ or pEC₅₀ values higher 2) for the three trophic

Table 3. Compounds showing high inter- or intra-database differences.

Name	CAS no.	Endpoint	Database	Min/Max ^a pLC or pEC ₅₀
2-Mercaptoethanol	60-24-2	Algae	ECHA vs. NITE	0.61–2.66, 2.66
Trinonylamine	68814-95-5	<i>Daphnia</i>	ECHA vs. ECOTOX	0.85–3.44, 3.44
Dodecanol	112-53-8	<i>Daphnia</i>	ECHA vs. Literature set	1.66–2.68, –0.24–2.91
Acrolein	107-02-8	<i>Daphnia</i>	ECOTOX vs. Literature set	–0.25–2.90, –0.25
2,4-Diaminotoluene	95-80-7	Fish	ECHA vs. ECOTOX	–1.07, –1.07–1.42
Diglycol chloroformate	106-75-2	Fish	ECHA vs. Literature set	–2.51–1.28, –2.51–1.67
Cetylpyridinium	123-03-5	Fish	ECOTOX vs. Literature set	0.85–3.31, 3.31

^aMinimum and maximum pLC or pEC₅₀ values for the first and the second databases (comma separated), respectively.

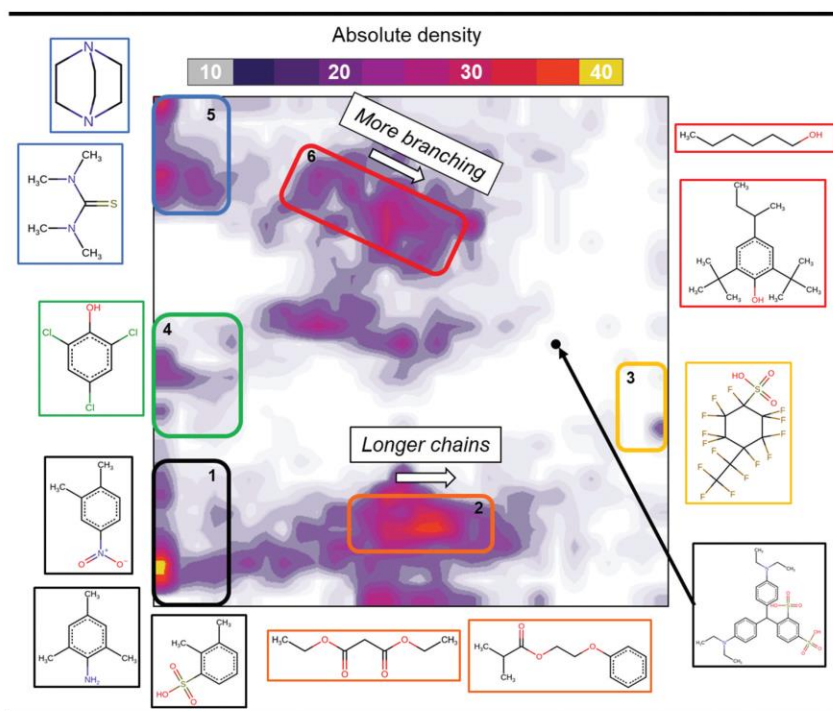


Figure 4. Density landscape of the whole chemical space. The colour scale refers to the number of compounds populating a determined zone. Coloured rectangles delimit map regions referred in the text.

levels simultaneously. Among them, it is possible to identify some chemical families which are known to be toxic for the aquatic environment [48–50], such as organochlorine compounds and polychlorinated biphenyl derivatives (e.g. chlordane and 2,2',5-trichlorobiphenyl) and long-chain aliphatic amines and quaternary ammonium salts (e.g. *N,N*-dimethylhexadecan-1-amine and trimethyl(octadecyl)azanium). On the other hand, areas '3' delimits non-toxic compounds such as aliphatic compounds with hydroxyl and carboxylic acid functional groups (2-ketoglutaric acid and citric acid). Finally, there are few compounds that exhibited acute toxicity for one species and were harmless for the others. This is the case, for instance, of dioxane (highly toxic to *Daphnia* only) or ethyl L-lactate (moderately toxic to fish only).

Models' performance

Two types of models were generated for each endpoint: (i) the ECHA models, generated using only data coming from the ECHA database; and (ii) the All-Public models, generated using all available public data.

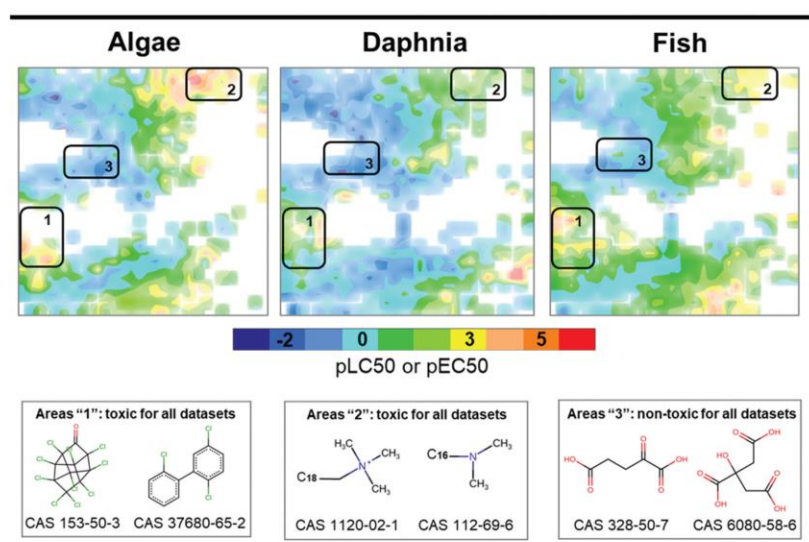


Figure 5. Property landscapes for the three datasets. Blue regions are mainly populated by non-toxic compounds; red ones by toxic compounds. White areas correspond to unpopulated regions. Black rectangles delimit map regions referred in the text.

Table 4 reports consensus models' performance evaluated in a fivefold CV and on the external test sets. The y-scrambling experiments measured the performance of 'randomized' models: r^2 values ($r^2 = -0.22$ – -0.15 ; std. dev. among repetitions = 0.06 – 0.11) were very low and noticeably far from those obtained by cross-validation. Internal validation metrics (Q^2_{loo} and Q^2_{boot} reported in Table 3) are acceptable and close to external validation values. In terms of prediction accuracy, ECHA models (RMSE = 0.56 – 0.61) are more performant than their All-Public counterparts (RMSE = 0.69 – 0.78). The variance of the All-Public dataset is also much larger than the ECHA dataset as can be deduced from Table 2. Therefore, the determination coefficient of the All-Public ($r^2 = 0.61$ – 0.67) models are better than for the ECHA models ($r^2 = 0.54$ – 0.64). On the external test sets, the data coverage (calculated as the ratio of the number of compounds inside AD and the total number of compounds of the given dataset) is around 70% (65%–74%). A compound is considered to be out-of-AD when its reliability is equal to 'Low'. External validation performances are comparable to those obtained by cross-validation, which supports the absence of overfitting. Without taking into account the applicability domain, external statistics are degraded. Respectively, for algae, *Daphnia* and fish dataset, the ECHA models scored r^2 of 0.60 , 0.67 , 0.60 and RMSE of 0.59 , 0.50 , 0.63 , while the All-Public models scored r^2 of 0.56 , 0.70 , 0.72 and RMSE of 0.74 , 0.75 , 0.69 .

Figure 6 depicts experimental values vs predicted values (EXP/PRED) scatter plots for the ECHA and All-Public models for all the three endpoints. Graphs show cross-validation predictions (grey) and external test set predictions (red). It can be noticed that the algae models have the largest scattering of data points, reflecting their lower r^2 values. An important aspect to mention is the limitation of the models to correctly predict very low

Table 4. Consensus model performance in internal and external validation.

Endpoint	Model	Q^2_{loo}		Q^2_{boot}		5-fold CV		External test set	
		Min-Max, single model		Min-Max, single model		r^2	RMSE	r^2	RMSE
Algae	ECHA	0.55–0.61		0.51–0.56		0.54 (0.007)	0.61 (0.005)	0.63	0.58
	All-Public	0.58–0.63		0.55–0.61		0.61 (0.004)	0.69 (0.003)	0.63	0.65
Daphnia	ECHA	0.59–0.65		0.57–0.62		0.60 (0.009)	0.56 (0.008)	0.75	0.43
	All-Public	0.64–0.70		0.60–0.65		0.67 (0.007)	0.78 (0.006)	0.77	0.60
Fish	ECHA	0.63–0.67		0.60–0.63		0.64 (0.005)	0.60 (0.006)	0.65	0.56
	All-Public	0.67–0.71		0.60–0.66		0.67 (0.005)	0.73 (0.006)	0.73	0.61

In brackets, the standard deviation computed on the CV repetitions is reported. ^acalculated as the ratio of the number of compounds inside AD and the total number of compounds of the given dataset. Q^2_{loo} and Q^2_{boot} = individual model's minimum and maximum Q^2 values determined in leave-one-out and bootstrapping.

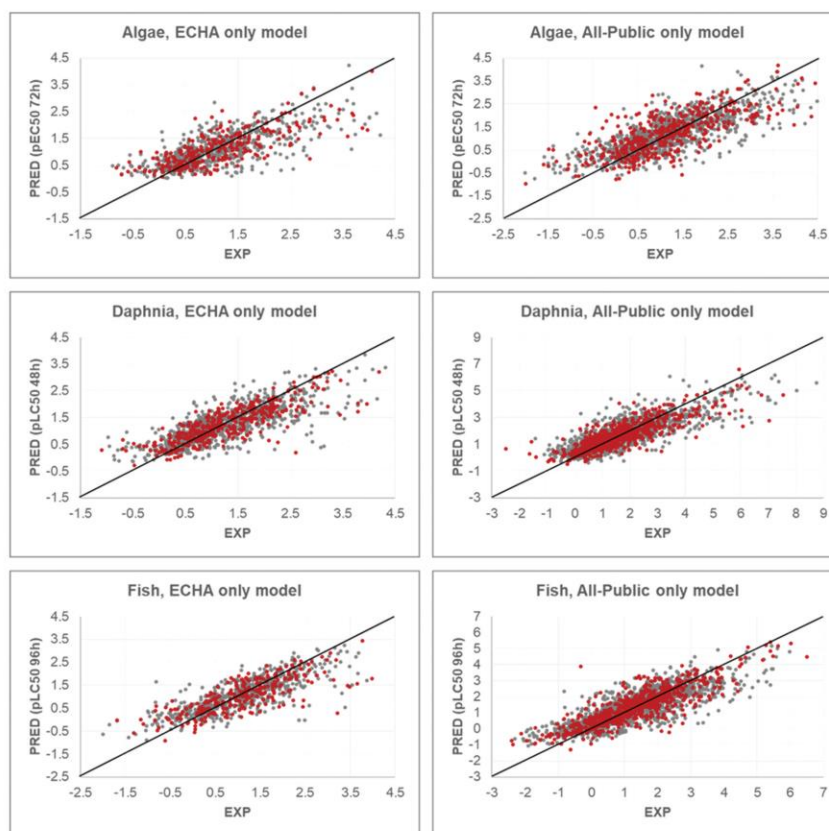


Figure 6. Internal and external validation scatter plots for Consensus models. Grey points represent the training set evaluated in CV; red points indicate external set compounds.

toxic compounds (i.e. with pLC_{50} or $ECC_{50} < -1$ log unit), which are consistently overestimated. Some of the benchmarked models (e.g. T.E.S.T.) also had this issue. There are two possible explanations to this drawback:

- Many low-molecular-weight molecules in these datasets (e.g. hydrogen cyanide, iodomethane or chloroacetonitrile) have some of the lowest toxicities. We hypothesize that these toxicities are specific to these chemical species and are explained by their reactivity. This could be taken into account by alternative approaches such as quantum chemistry. In our models, these structures are encoded, so the predictions for these will be accurate, but the models might not generalize correctly for their analogues.
- Uncertainty of measured toxicity values: when a compound is far from regulatory classification thresholds a looser estimation of its toxicity could be enough to fulfil data requirements. In other cases, the limit of solubility can be taken as toxicity value

as a worst-case approach. For instance, for 2-ethoxyethanol (110-80-5), available fish toxicity values range from 500 up to 16000 mg/l.

This latter consideration can be applied to highly-toxic compounds as well: in such cases, the amount of the test compound is very low and therefore it is technically more difficult to precisely determined its concentration (for instance, it could be close to the limit of detection of the analytical method). These factors increase the uncertainty in the data and negatively affect model's performance.

In the end, all the available public and industrial data were merged, and 'final' models were prepared. Their performances were evaluated by fivefold CV repeated 10 times. For alga, *Daphnia* and fish r^2 values were 0.60, 0.72, 0.71 and RMSE values were 0.71, 0.71, 0.69, respectively. With the exception of algae, final models have shown a slight overall improvement compared to All-Public models, but their error (RMSE) is still higher than the ECHA models.

Model performance on the industrial sets

For confidentiality reasons, this dataset cannot be disclosed, and only some general information can be provided. It comprises quite heterogeneous chemical structures, from high molecular weight compounds such as long-chain aliphatic surfactants and halogenated biphenyls to much smaller ones such as phenol derivatives and simple amides. The molecular weight ranges from 30 to 1134. A significant number of compounds belong to the chemical class of surfactants, especially ethoxylated alkylphenols. Table 5 reports ECHA and All-Public model performance on the Industrial sets. Compared to internal and external validation performances (Table 4), the models are performing considerably worse in predicting the Industrial sets, with RMSE values ranging from 0.92 (Fish) to 1.12 (Algae). Data coverage is acceptable, ranging from 64% to 76%. Chemical space projections of the Industrial set compounds are depicted in Figure 7: as expected, the majority of out-of-AD compounds are located in lower-density regions (Figure 4).

As previously observed on public data, the risks of small and very-low toxic compounds appear overestimated: as can be seen in Figure S2, several compounds with experimental pLC_{50} or $pEC_{50} < -1$ are predicted to be much more toxic.

Table 5. Consensus model performance on the Industrial set.

Endpoint	Model	Industrial set		
		r^2	RMSE	Data coverage ^a [%]
Algae	ECHA	0.44	1.12	69 (172/249)
	All-Public	0.48	1.07	72 (179/249)
<i>Daphnia</i>	ECHA	0.54	0.94	73 (166/228)
	All-Public	0.58	0.93	76 (174/228)
Fish	ECHA	0.58	0.92	64 (124/193)
	All-Public	0.54	0.97	67 (129/193)

^acalculated as the ratio of the number of compounds inside AD and the total number of compounds of the given dataset.

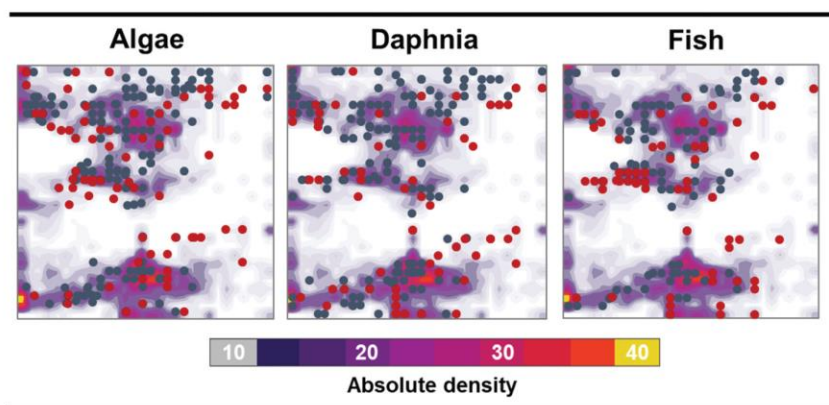


Figure 7. Industrial set compounds chemical space projections. The maps show the projects on the Industrial set compounds using the density landscape (Figure 4). Blue/red dots refer to inside/outside AD compounds.

Table 6. Performances of considered tools on the Industrial sets.

Endpoint	Tool	r^2	RMSE	Data coverage ^a (%)
Algae	ECOSAR	0.26	1.94	97
	VEGA	0.29–0.33	1.17–0.99	30–32
Daphnia	ECOSAR	0.38	1.57	99
	VEGA	0.28–0.44	1.19–1.03	20–52
	T.E.S.T.	0.42	0.98	66
Fish	ECOSAR	0.41	1.65	96
	VEGA	0.20–0.51	1.09–0.85	33–45
	T.E.S.T.	0.39	1.10	58

^acalculated as the ratio of the number of compounds inside AD and the total number of compounds of the given dataset.

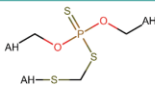

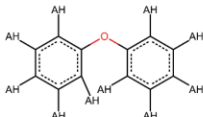
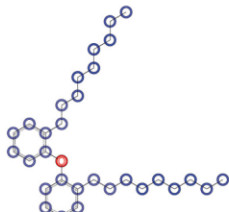
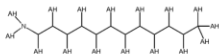

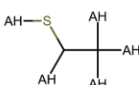

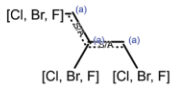
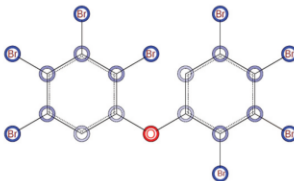
Already-existing tools performance comparison

Table 6 reports performances of the benchmarking tools on the Industrial sets. ECOSAR is performing significantly worse than any model ($RMSE = 1.57$ – 1.94), which could be caused by the absence of any AD filter. In terms of accuracy, our models are performing similarly to VEGA and T.E.S.T. tools, with comparable RMSE values. Generally, VEGA shows slightly more accurate predictions, but suffer from a narrower AD compared to our models, with data coverage of 30%–45% vs. 64%–76%, respectively.

ColorAtom analysis

Table 7 compares the selected SAs and the corresponding ColorAtom representation. For instance, molecule CAS 9014–90-8 (SA no. 2) exhibits moderate toxicity to invertebrates ($EC_{50-Daphnia} = 21$ mg/L). The aliphatic chains have been identified as the main drivers of the molecule's toxicity (dark-blue coloured): this could be caused by the fact that longer chains tend to make the molecule more lipophilic, which is generally related to an increase of toxicity towards aquatic organisms [19,27]. Similarly, molecule CAS 4337-75-1 (SA no 3)

Table 7. ColorAtom graphs for compounds matching the given SA (For the given SA, the ColorAtom graph is depicted. The colouration is directly referred to the modelled property (i.e. the pLC_{50} or pEC_{50} value): blue and red circled atoms played a role in increasing and decreasing it, respectively. AH stands for: 'any atom, including hydrogen').

ID	Structural alert	Example compound
1		 CAS 563-12-2; $EC_{50-Daphnia} = 6.8E-07$ mg/L
2		 CAS 9014-90-8; $EC_{50-Daphnia} = 21$ mg/L
3		 CAS 4337-75-1; $LC_{50-Fish} = 5.04$ mg/L
4		 CAS 111-71-1; $EC_{50-Alga} = 49.4$ mg/L
5		 CAS 32536-52-0; $EC_{50-Alga} = 0.012$ mg/L

shows the same trend, with the polar head being less related to an increase of toxicity. For molecule CAS 32536-52-0 (SA no. 5), bromine atoms attached to benzene rings have been seen by the model as positively related to a toxicity increase. A total of 57 halogenated aromatic structures matched this SA, which comprise the chemical families of polybrominated diphenyl ethers or polychlorinated biphenyl. All benzene rings and their halogen substituents show the same colouration pattern, as opposed to oxygen-containing functional groups, such as ethers or carboxylic acids.

Discussion

Ecotoxicological data quality and variability are a serious issue negatively impacting model performance, as already reported by previous authors [17,19,51,52]. Here we

noticed that same problem, with almost 20% of all compounds having measured acute toxicity values differing by more than 1 log unit. Often it is very time-consuming (if not impossible), to verify available information, for instance, due to missing experimental information. Compounds that are more difficult to test (e.g. compounds which are poorly soluble or hydrophobic, volatiles, reactive in water) [53] were also affected by the highest data variability. The fact that poorly predicted molecules are the same that are difficult to test confirms the well-known concept that the model quality is highly dependent on the dataset quality and shows that experimental measurements for these difficult substances may actually not be more reliable than model prediction.

In an effort to extract a subset of (theoretically) more reliable compounds, we tried to select only data coming from the ECHA database, as it includes a reliability evaluation performed by the registrants. Indeed, despite worse determination coefficient values as opposed to All-Public models, which can be attributed to the much narrower property range, the average errors of the ECHA models were considerably better, with an improvement of $RMSE_{cv}$ values of 0.08–0.22 (Table 4).

Compared to already-published models (Table 1), our new models not only show an improvement of accuracy, but also a significant increase of training set sizes which, in turn, means extended applicability domains.

All the employed tools showed mediocre performances on the Industrial sets, with average an RMSE of roughly 1 log unit. This denotes a limitation of currently existing models when applied to an industrial context and encourages the use of consensus to reduce uncertainty and improve the accuracy.

Models failed to predict several compounds belonging to the chemical class of surfactants. These compounds are generally more difficult to predict by QSAR models and also to be handled from an experimental point of view [17,54]. Their behaviour in water is highly dependent on the nature of their polar/non-polar portions, level of branching and molecular weight and chain length. Surfactants tend either to concentrate at the water-air phase, or to aggregate in micelles, which could affect the precision in determining their water concentration. Moreover, for small organisms such as daphnids, their toxicity could have been enhanced by a physical action of the surfactant, as the organisms could be trapped by the micelles. In addition, surfactants tend to be mixtures rather than pure compounds, and normally the most frequent component is taken as a representative member to approximate their true composition [55].

The final models (merging of public and industrial data) demonstrated to have better performances (in cross-validation) than All-Public models. Despite their error ($RMSE = 0.69$ – 0.71) is still worse than the ECHA models (0.56 – 0.61), they have a much larger training set including a substantial number of compounds bringing industrially relevant chemotypes, which noticeably extends their applicability domains.

Conclusions

In this work, we report regression consensus models of acute aquatic toxicity towards three trophic levels: algae, *Daphnia* and fish. A total of 3680 publicly available unique compounds, annotated by at least one experimental value per endpoint, were collected.

Models were externally validated on two test set: the former was created by splitting available public data, while the latter comprised proprietary industrial data. Performances

on the former datasets were acceptable (RMSE = 0.56–0.78) and similar to those determined by cross-validation. On the other hand, prediction accuracy on the Industrial sets was noticeably worse (RMSE = 0.92–1.12). The main cause was the overestimation of the toxicity of several small molecular weight molecules (absolute errors higher than 1.5 log units). It is hypothesised that these errors are due to uncertainties in experimental data and to specificities of the electronic structures that are insufficiently represented by the molecular graph of the molecules.

In addition, a benchmarking on the Industrial sets has been carried out considering the ECOSAR, VEGA and T.E.S.T. freely available tools: our models scored one of the best prediction accuracies coupled with a good data coverage.

Finally, public and industrial data were merged and models were updated: final models' training sets are considerable bigger (1806, 2529, 2591 for algae, *Daphnia* and fish, respectively) than those of already existing tools, extending therefore their applicability domain. In cross-validation, these models showed r^2 values of 0.60, 0.72, 0.71 and RMSE values of 0.71, 0.71, 0.69 for Algae, *Daphnia* and Fish, respectively.

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: 10.5281/zenodo.3708082.

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. Azam  <http://orcid.org/0000-0002-2974-2484>

M.H. Enrici  <http://orcid.org/0000-0001-5696-4376>

E. Van Miert  <http://orcid.org/0000-0001-6653-1371>

A. Varnek  <http://orcid.org/0000-0003-1886-925X>

References

- [1] 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/EEC 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.
- [2] K.P. Singh, S. Gupta, A. Kumar, and D. Mohan, *Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology*, Chem. Res. Toxicol. 27 (2014), pp. 741–753. doi:10.1021/tx400371w.
- [3] OECD, OECD Guidelines for the Testing of Chemicals, Section 3: Effects on Biotic Systems, Organisation for Economic Cooperation and Development, Paris, FR, 2019. Available at <http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm>.

- [4] V. Aruoja, M. Moosus, A. Kahru, M. Sihtmäe, and U. Maran, *Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga Pseudokirchneriella subcapitata*, Chemosphere 96 (2014), pp. 23–32. doi:10.1016/j.chemosphere.2013.06.088.
- [5] P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap, and E. Papa, QSAR modeling is not “Push a button and find a correlation”: A case study of toxicity of (Benzo-)triazoles on algae, Mol. Inform. 31 (2012), pp. 817–835. doi:10.1002/minf.201200075.
- [6] A. Levet, C. Bordes, Y. Clément, P. Mignon, C. Morell, H. Chermette, P. Marote, and P. Lantéri, *Acute aquatic toxicity of organic solvents modeled by QSARs*, J. Mol. Model. 22 (2016), pp. 288–298. doi:10.1007/s00894-016-3156-0.
- [7] K. Khan, P.M. Khan, G. Lavado, C. Valsecchi, J. Pasqualini, D. Baderna, M. Marzo, A. Lombardo, K. Roy, and E. Benfenati, *QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors*, Chemosphere 229 (2019), pp. 8–17. doi:10.1016/j.chemosphere.2019.04.204.
- [8] C. Bertinetto, C. Duce, R. Solaro, M.R. Tiné, A. Micheli, K. Héberger, A. Miličević, and S. Nikolić, *Modeling of the acute toxicity of benzene derivatives by complementary QSAR methods*, Match 70 (2013), pp. 1005–1021.
- [9] US EPA, Estimation Programs Interface Suite™ for Microsoft® Windows V4.11, US Environmental Protection Agency, Washington DC, 2012; software available at <https://www.epa.gov/tscascreening-tools/epi-suite-estimation-program-interface>.
- [10] 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.
- [11] T. Martin, P. Harten, and D. Young, (T.E.S.T.) Toxicity Estimation Software Tool V 4.1, US Environmental Protection Agency, 2012; software available at <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.
- [12] T. Sheffield and R.S. Judson, *Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure*, Environ. Sci. Technol. 53 (2019), pp. 2793–2802. doi:10.1021/acs.est.9b03957.
- [13] K. Khan and K. Roy, *Ecotoxicological QSAR modelling of organic chemicals against Pseudokirchneriella subcapitata using consensus predictions approach*, SAR QSAR Environ. Res. 30 (2019), pp. 665–681. doi:10.1080/1062936X.2019.1648315.
- [14] S. Kar and K. Roy, *QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors*, J. Hazard. Mater. 177 (2010), pp. 344–351. doi:10.1016/j.jhazmat.2009.12.038.
- [15] M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I.V. Tetko, and R. Todeschini, *Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method*, ATLA Altern. To Lab. Anim. 42 (2014), pp. 31–41. doi:10.1177/026119291404200106.
- [16] K.L.E. Kaiser and S.P. Niculescu, *Modeling acute toxicity of chemicals to Daphnia magna: Aprobabilistic neural network approach*, Env. Tox. Chem. 20 (2001), pp. 420–431. doi:10.1002/etc.5620200225.
- [17] A. Golbamaki, A. Cassano, A. Lombardo, Y. Moggio, M. Colafranceschi, and E. Benfenati, *Comparison of in silico models for prediction of Daphnia magna acute toxicity*, SAR QSAR Environ. Res. 25 (2014), pp. 673–694. doi:10.1080/1062936X.2014.923041.
- [18] Y. Wang, M. Zheng, J. Xiao, Y. Lu, F. Wang, J. Lu, X. Luo, W. Zhu, H. Jiang, and K. Chen, *Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow*, SAR QSAR Environ. Res. 21 (2010), pp. 559–570. doi:10.1080/1062936X.2010.502300.
- [19] M. Cassotti, D. Ballabio, R. Todeschini, and V. Consonni, *A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas)*, SAR QSAR Environ. Res. 26 (2015), pp. 217–243. doi:10.1080/1062936X.2015.1018938.
- [20] K. Khan, D. Baderna, C. Cappelli, C. Toma, A. Lombardo, K. Roy, and E. Benfenati, *Ecotoxicological QSAR modeling of organic compounds against fish: Application of fragment based descriptors in feature analysis*, Aquat. Toxicol. 212 (2019), pp. 162–174. doi:10.1016/j.aquatox.2019.05.011.

- [21] 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.
- [22] G. Marcou, D. Horvath, F. Bonachera, and A. Varnek, *Laboratoire De Chimoinformatique UMR 7140 CNRS*, University of Strasbourg, Strasbourg, FR, 2019. Available at <http://infochim.520u-strasbg.fr>.
- [23] 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>.
- [24] US EPA, *Data From: ECOTOX Knowledgebase*. US Environmental Protection Agency, 2017; 525 dataset available at <https://cfpub.epa.gov/ecotox/>.
- [25] S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal, and P. Gramatica, *Daphnia and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity-activity modelling*, *J. Hazard. Mater.* 258–259 (2013), pp. 50–60. doi:10.1016/j.jhazmat.2013.04.025.
- [26] A.A. Toropov, A.P. Toropova, M. Marzo, J. Lou Dorne, N. Georgiadis, and E. Benfenati, *QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database*, *Environ. Toxicol. Pharmacol.* 53 (2017), pp. 158–163. doi:10.1016/j.etap.2017.05.011.
- [27] X. Wu, Q. Zhang, and J. Hu, *QSAR study of the acute toxicity to fathead minnow based on a large dataset*, *SAR QSAR Environ. Res.* 27 (2016), pp. 147–164. doi:10.1080/1062936X.2015.1137353.
- [28] A. Furuhashi, K. Hasunuma, T.I. Hayashi, and N. Tatarazako, *Predicting algal growth inhibition toxicity: Three-step strategy using structural and physicochemical properties*, *SAR QSAR Environ. Res.* 27 (2016), pp. 343–362. doi:10.1080/1062936X.2016.1174151.
- [29] OASIS, *QSAR Toolbox V 4.3*, OASIS Laboratory of mathematical chemistry, Burgas, BG, 2017; software available at <http://oasis-lmc.org/products/software/toolbox.aspx>.
- [30] 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), 545 pp. 26–31. doi:10.1145/1656274.1656280.
- [31] NIH, *PubChem*, National Library of Medicine, National Center for Biotechnology Information, Bethesda, Maryland, 2019. Available at <https://pubchem.ncbi.nlm.nih.gov/>.
- [32] 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/rtp.1996.1076.
- [33] 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.
- [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] F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert, and A. Varnek, *Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context*, *SAR QSAR Environ. Res.* 30 (2019), pp. 879–897. doi:10.1080/1062936X.2019.1672089.
- [36] F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert, and A. Varnek, *Modelling of ready biodegradability based on combined public and industrial data sources*, *SAR QSAR Environ. Res.* 31 (2020), pp. 171–186. doi:10.1080/1062936X.2019.1697360.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.

- [41] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701. doi:10.1002/qsar.200610151.
- [42] 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.
- [43] G. Gini, G. Giuseppina, T. Ferrari, A. Lombardo, A. Cassano, and E. Benfenati, *A new QSAR model for acute fish toxicity based on mined structural alerts*, J. Toxicol. Risk Assess. 5 (2019), pp. 2572–2580.
- [44] C.I. Cappelli, A. Cassano, A. Golbamaki, Y. Moggio, A. Lombardo, M. Colafranceschi, and E. Benfenati, *Assessment of in silico models for acute aquatic toxicity towards fish under REACH regulation*, SARQSAR Environ. Res. 26 (2015), pp. 977–999. doi:10.1080/1062936X.2015.1104519.
- [45] M. Pavan, A.P. Worth, and T.I. Netzeva, *Comparative Assessment of QSAR Models for Aquatic Toxicity*, Tech. Rep. 21750EN, Joint Research Centre, Ispra, IT, 2005.
- [46] S. Lozano, E. Lescot, M.P. Halm, A. Lepaillieur, R. Bureau, and S. Rault, *Prediction of acute toxicity in fish by using QSAR methods and chemical modes of action*, J. Enzyme Inhib. Med. Chem. 25 (2010), pp. 195–203. doi:10.3109/14756360903169857.
- [47] M. Hrovat, H. Segner, and S. Jeram, *Variability of in vivo fish acute toxicity data*, Regul. Toxicol. Pharmacol. 54 (2009), pp. 294–300. doi:10.1016/j.yrtph.2009.05.013.
- [48] H.J.M. Verhaar, C.J. van Leeuwen, and J.L.M. Hermens, *Classifying environmental pollutants*, Chemosphere 25 (1992), pp. 471–491. doi:10.1016/0045-6535(92)90280-5.
- [49] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond, *Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)*, Environ. Toxicol. Chem. 16 (2005), pp. 948. doi:10.1002/etc.5620160514.
- [50] S.J. Enoch, M. Hewitt, M.T.D. Cronin, S. Azam, and J.C. Madden, *Classification of chemicals according to mechanism of aquatic toxicity: An evaluation of the implementation of the Verhaar scheme in Toxtree*, Chemosphere 73 (2008), pp. 243–248. doi:10.1016/j.chemosphere.2008.06.052.
- [51] P.C. Thomas, P. Bicherel, and F.J. Bauer, *How in silico and QSAR approaches can increase confidence in environmental hazard and risk assessment*, Integr. Environ. Assess. Manag. 15 (2019), pp. 40–50. doi:10.1002/ieam.4108.
- [52] N. Burden, S.K. Maynard, L. Weltje, and J.R. Wheeler, *The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: A retrospective validation approach*, Regul. Toxicol. Pharmacol. 80 (2016), pp. 241–246. doi:10.1016/j.yrtph.2016.05.032.
- [53] OECD, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures*, Tech. Rep. JT03442844, Organisation for Economic Co-operation Development, Paris, FR, 1992.
- [54] G.E. Bragin, C.W. Davis, M.H. Kung, B.A. Kelley, C.A. Sutherland, and M.A. Lampi, *Biodegradation and ecotoxicity of branched alcohol ethoxylates: Application of the target lipid model and implications for environmental classification*, J. Surfactants Deterg. 23 (2020), pp. 383–403. doi:10.1002/jsde.12359.
- [55] D.W. Roberts, *QSAR issues in aquatic toxicity of surfactants*, Sci. Total Environ. 109 (1991), pp.557–568. doi:10.1016/0048-9697(91)90209-W.

4.1.5 Short-term toxicity on Rodent

The estimation of the acute oral toxicity is a mandatory requirement in the frame of REACH for substances manufactured or imported in quantities of 1 ton or more per year. By the beginning of 2018, the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), as part of the effort to support the use of alternative methods, organized a worldwide workgroup to develop in-silico models of acute oral toxicity. Specifically, five relevant endpoints needed by regulatory agencies were targeted. These endpoints included (i) identification of “very toxic” chemicals (LD50 less than 50 mg/kg) and (ii) “nontoxic” chemicals (LD50 greater than or equal to 2000 mg/kg), (iii) point estimates for LD50s, (iv) categorization of toxicity hazard using the U.S. Environmental Protection Agency (EPA) and (v) the GHS classification schemes.

The NICEATM collected rat oral LD50 data on over 15,000 substances from different publicly available databases and resources. The curated dataset was split into training and validation set. In the first stage, only the former was provided to the participants. The validation set was later used to externally validate the submitted models. The committee evaluated each model qualitatively with respect to the OECD principles and quantitatively based on the predictive performance against the test set. Models were then used to screen a large prediction set of ~40 k chemicals of interest to different agencies and finally were also included into a consensus model, which leverages the strengths and compensate for the weaknesses of each individual approach.

As participants, we submitted a regression model for LD50 estimation. In this manuscript we present our modelling approach and a continuation of our work, including:

- Generation of a new multi-classification model based on GHS categories;
- Collection of additional acute oral toxicity data from several sources to extend the model’s training set (“Global models”);
- External validation against a dataset relevant for the context of the chemical industry (hereafter named “Industrial set”), provided by Solvay.

It has been demonstrated that both regression and classification Global models obtained in this work (RMSE = 0.47 and BA = 0.72) perform better than the previously

reported NICEATM models (RMSE = 0.56 and BA = 0.69) when challenged on industrial data. Moreover, the Global models have much larger applicability domain: the data coverage on the Industrial set is 85 % and 82 % (classification) and 94 % and 58 % (regression) for Global and NICEATM models, respectively.



4.1.5 Toxicité aiguë par voie orale chez le rat

L'estimation de la toxicité orale aiguë est exigée dans le cadre de REACH pour les substances fabriquées ou importées en quantités supérieure ou égale à 1 tonne par an. Au début de 2018, le National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), dans le cadre des efforts visant à soutenir l'utilisation de méthodes alternatives, a organisé un groupe de travail mondial pour développer des modèles *in silico* de toxicité aiguë par voie orale. En particulier, cinq critères d'effet pertinents requis par les organismes de réglementation ont été ciblés. Ces critères d'évaluation comprenaient (i) l'identification de produits chimiques «très toxiques» (dose létale médiane, DL50 inférieure à 50 mg / kg) et (ii) des produits chimiques «non toxiques» (DL50 supérieure ou égale à 2000 mg / kg), (iii) des estimations ponctuelles pour les DL50, (iv) la catégorie de risque de toxicité selon les standards de l'Environmental Protection Agency (EPA) aux États-Unis et (v) les schémas de classification du GHS (*Globally Harmonized System*).

Le NICEATM a collecté des données de DL50 orale sur le rat concernant plus de 15 000 substances à partir de différentes bases de données et ressources accessibles au public. L'ensemble des données sélectionnées a été divisé en un ensemble d'apprentissage et un ensemble de validation. Dans la première étape, seul le premier a été fourni aux participants. L'ensemble de validation a ensuite été utilisé pour valider, selon une procédure de validation externe, les modèles soumis. Le comité a évalué chaque modèle qualitativement par rapport aux principes de l'OCDE et quantitativement sur la base de la performance prédictive par rapport à l'ensemble de test. Les modèles ont ensuite été utilisés pour cribler un ensemble prospectif de ~40k produits chimiques d'intérêt pour différentes agences. Les prédictions de chaque modèle individuel ont été rassemblées dans un consensus qui exploite les forces et compense les faiblesses de chaque approche individuelle.

En tant que participants, nous avons soumis un modèle de régression pour l'estimation de la DL50. Dans ce manuscrit, nous présentons notre approche de modélisation et la poursuite de nos travaux, notamment:

- La génération d'un nouveau modèle de classification basé sur les catégories du GHS;
- La collection de données supplémentaires sur la toxicité orale aiguë chez le rat à partir de plusieurs sources pour étendre l'ensemble d'entraînement des modèles («modèles globaux»);
- La validation externe par rapport à un ensemble des données pertinent pour le contexte de l'industrie chimique (ci-après désigné «*industrial set*»), fourni par l'entreprise Solvay.

Il a été démontré que les modèles globaux de régression et de classification obtenus dans ce travail (RMSE = 0,47 et BA = 0,72) fonctionnent mieux que les modèles NICEATM précédemment rapportés (RMSE = 0,56 et BA = 0,69) lorsqu'ils sont utilisés sur des données industrielles. De plus, les « modèles globaux » ont un domaine d'applicabilité beaucoup plus large: la couverture des données sur l'ensemble industriel est de 85% et 82% (classification) et 94% et 58% (régression) pour les « modèles globaux » et NICEATM, respectivement.



Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context

F. Lunghini^{a,b}, G. Marcou^a, P. Azam^b, D. Horvath^a, R. Patoux^b, E. Van Miert^b and A. Varnek^a

^aLaboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France; ^bToxicological and Environmental Risk Assessment unit, Solvay S.A., St. Fons, France

ABSTRACT

We report predictive models of acute oral systemic toxicity representing a follow-up of our previous work in the framework of the NICEATM project. It includes the update of original models through the addition of new data and an external validation of the models using a dataset relevant for the chemical industry context. A regression model for LD₅₀ and multi-class classification model for toxicity classes according to the Global Harmonized System categories were prepared. ISIDA descriptors were used to encode molecular structures. Machine learning algorithms included support vector machine (SVM), random forest (RF) and naïve Bayesian. Selected individual models were combined in consensus. The different datasets were compared using the generative topographic mapping approach. It appeared that the NICEATM datasets were lacking some relevant chemotypes for chemical industry. The new models trained on enlarged data sets have applicability domains (AD) sufficiently large to accommodate industrial compounds. The fraction of compounds inside the models' AD increased from 58% (NICEATM model) to 94% (new model). The increase of training sets improved models' prediction performance: RMSE values decreased from 0.56 to 0.47 and balanced accuracies increased from 0.69 to 0.71 for NICEATM and new models, respectively.

ARTICLE HISTORY

Received 12 July 2019
Accepted 21 September 2019


KEYWORDS

QSAR/QSPR; generative topographic mapping (GTM); oral rat acute toxicity; OECD principles; REACH

Introduction

The estimation of the acute oral toxicity is a mandatory requirement under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH, EC No. 1907/2006) legislation for substances manufactured or imported in quantities of 1 ton or more per year [1]. In most cases, this information is generated by performing an animal test according to the Organisation for Economic Co-operation and Development (OECD) guidelines. Until 2002, the reference guideline was OECD 401, however it was abolished for animal welfare reasons. Nowadays, more advanced guidelines are available which demand much less testing on animals and are likely to produce more reliable results [2]. Currently used guidelines are: OECD 420 (fixed dose procedure), OECD 423 (acute toxic class method), OECD 425 (up and down procedure) [3]. These guidelines are designed to classify the substances according to the

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

 Supplementary material for this article can be accessed at: <https://doi.org/10.1080/1062936X.2019.1672089>.

© 2019 Informa UK Limited, trading as Taylor & Francis Group

Global Harmonized System (GHS) categories and LD₅₀ values are only roughly estimated, at best.

To reduce animal testing, REACH encourages the use of non-testing methodologies, such as weight of evidence approaches, read across and QSAR modelling. In the past years, several QSAR models have already been developed to predict Acute Oral toxicity [4–7]. Some models are nowadays implemented in both commercial and free software (Table 1).

By the beginning of 2018, the National Toxicology Programme Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM) [8], as part of the effort to support the use of alternative methods, organized a worldwide workgroup to develop *in silico* models of acute oral toxicity. In particular, five relevant endpoints needed by regulatory agencies were targeted. These endpoints included (i) identification of ‘very toxic’ chemicals (LD₅₀ less than 50 mg/kg) and (ii) ‘non-toxic’ chemicals (LD₅₀ greater than or equal to 2000 mg/kg), (iii) point estimates for LD₅₀s, (iv) categorization of toxicity hazard using the U.S. Environmental Protection Agency (EPA) [9] and (v) the GHS [10] classification schemes. The NICEATM collected rat oral LD₅₀ data on over 15,000 substances from different publicly available databases and resources. The curated dataset was split into training and validation set. In the first stage, only the former was provided to the participants. The validation set was later used to externally validate the submitted models. The committee evaluated each model qualitatively with respect to the OECD principles [11] and quantitatively based on the predictive performance against the test set. Models were then employed to screen a large prediction set of ≈40 k chemicals of interest to different agencies and finally were also included into a consensus model, which leverages the strengths and compensate for the weaknesses of each individual approach [12]. More information about data preparation can be found on the workgroup website [8] and described by Ballabio et al. [13].

As participants, we submitted a regression model for LD₅₀ estimation. In this manuscript we present our modelling approach and a continuation of our work, including:

- (1) Generation of a new multi-classification model based on GHS categories;
- (2) Collection of additional acute oral toxicity data from several sources to extend the model’s training set;
- (3) External validation against a dataset relevant for the context of the chemical industry (hereafter named ‘Industrial set’), provided by Solvay.

Finally, all public data was merged to constitute a ‘Global set’ (counting 11981 compounds) and models were updated. To the best of our knowledge, this is the largest reported dataset used for the development of QSARs predicting acute toxicity (Table 1).

Table 1. Tools for acute oral LD₅₀ estimation.

Model	Tr. size	Employed descriptors	Algorithm	Ref.
TEST ^F	7420	Chemistry Development Kit (CDK) [15]	Consensus on five methods	[16]
ADMET ^C	7150	2D, 3D molecular descriptors	Artificial neural network	[17]
ACD/Labs ^C	8631	Expert knowledge and structural descriptors	Expert knowledge and classification-SAR	[18]
TerraBase ^C	≈ 10000	Molecular structure descriptors	Probabilistic Neural Network	[19]
Accelrys ^C	≈ 4000	Molecular structure descriptors	Consensus on several models	[20]

^F = freely available; ^C = commercial

Our models are available through the online ISIDA/Predictor platform [14], available at the Laboratory of Chemoinformatics webpage: <http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi>.

Methods

Modelling workflow

A graphical representation of the general workflow is shown in Figure 1; its main steps will be detailed in the present chapter.

Data collection

Curated experimental data was distributed by the NICEATM workgroup. The original continuous LD₅₀ training and validation set counted respectively 6734 and 2174 compounds; analogously, for GHS classes 8960 and 2885 compounds were available. Additional oral rat LD₅₀ data was collected from the database of the European Chemicals Agency (ECHA) through the eChem portal [21], the relevant databases from the QSAR Toolbox software (SI, Section 1) [22] and the Toxicity Estimation Software Tool (TEST) training set [16]. Furthermore, a dataset on LD₅₀ (Industrial set) was provided by the industrial partner Solvay. This naming has been chosen in order to underline the existing structural differences between the compounds coming from an industrial context, which may represent new trends in large-scale production, from those available in public databases. To support

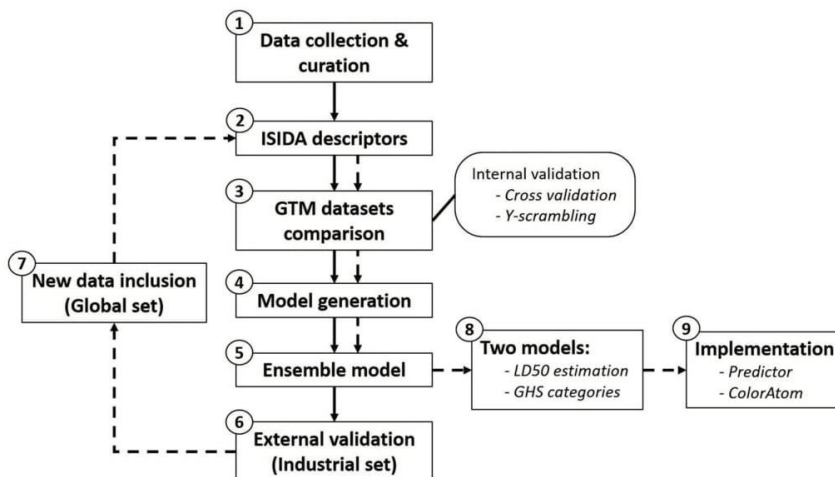


Figure 1. General workflow. (1) data is collected from different sources; (2) ISIDA descriptors encoding; (3) GTM is employed to compare the structural space of the datasets; (4), (5) individual models are trained and combined in consensus; (6) the Industrial set is used for external validation (7) the 'Global set' is issued by the merging of all public data and (8) models are updated; (9) models are published on the online platform.

this statement, collected databases were analysed through GTM (results section) and pairwise comparison of the Tanimoto similarities (SI, Section 2). Both approaches highlighted structural differences between their chemical spaces and the presence of unique chemotypes. Finally, an additional dataset of 462 compounds, not overlapping with the collected data, was provided by Solvay afterwards. This dataset (Blind set) was thus used to externally validate the last model version built on all collected data (public + industrial).

All collected public data (i.e. a total of 13682 unique compounds after the curation procedure) is available on Zenodo (DOI 10.5281/zenodo.3300664) with the respective LD₅₀ and/or GHS property; the industrial compounds cannot be provided due to confidentiality reasons.

Data curation and standardization

To avoid additional sources of variability, data was limited to rat-only assays. Mixtures, polymers and UVCBs (Unknown or Variable composition, Complex reaction products or Biological materials) were discarded. Chemical standardization included: removal of salts/solvents, neutralization, removal of explicit hydrogens, aromatic representation for benzene rings, removal of stereo information, standardization of -nitro and -sulpho containing groups. This step was performed with a standardization workflow implemented in the Konstanz Information Miner (KNIME) [23]. In case of duplicates only one structure was kept and their LD₅₀ median value was selected (computed according to norm ISO16269-7). Multiple LD₅₀ values available the same compound were used to estimate the experimental error of the measurements. For each compound with at least 2 data points, a LD₅₀ range (maximum – minimum over reported values) was calculated, and the average of these range widths over concerned compounds was interpreted as the experimental error. GHS classes [10] were assigned based on the continuous LD₅₀ value, using the following thresholds (in mg/kg): ≤5, class 1; >5 and ≤50, class 2; >50 and ≤300, class 3; >300 and ≤2000, class 4; >2000, class 5. In order to maintain the same NICEATM classification system, the GHS ‘not classified’ category (i.e. > 5000 mg/kg) and GHS Category 5 (i.e. > 2000 mg/kg) were merged together in one unique class. For the regression model, LD₅₀ values originally expressed in mg/kg body weight were transformed to the inverse log of the molar dose (pLD₅₀ in mmol/kg body weight).

Encoding of chemical structures

ISIDA property-label molecular descriptors [24] were employed. This led to the generation of dozens of different descriptor spaces which corresponds to different fragment sizes, topologies and encoded chemical information, called ‘colouration’ (elements labels, physical properties mapped on the atoms explicit or implicit chemical bonds, atom pairs). The number of fragments of the given descriptor space depends on selected fragmentation scheme. It varied from 387 (IIAB(2–2), atom centred fragments with radius 1) to 31623 (IIAB(2–5), atom centred fragments with radius 5), with an average of 7974 (SI, section 1).

Generative topographic mapping

The chemical space of the collected databases was compared by means of the GTM approach [25], a dimensionality reduction method allowing the visualization of data distribution on a 2-dimensional map. A data property can be added as a 3rd 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 [26–28]. Two types of analysis were carried out: (i) the NICEATM dataset set was pairwise compared with the other databases (i.e. QSAR Toolbox, TEST, etc.); (ii) a map was generated on the Global set and the LD₅₀ value was used as property landscape. For the former, the goal was to identify which chemotypes were unique to the industrial context and under-represented in public available data. For the latter, the goal was to visualize how toxic and non-toxic compounds are distributed in the chemical space. The ISIDA descriptor space IIB(2–2) [24] associated to the best support vector machine (SVM) model (in terms of balanced accuracy) was chosen. These descriptors are based on molecular fragments consisting in an atom and information on the corresponding chemical bonds. The manifold [21] was built on the whole available chemical space (i.e. the Global set).

Model generation

Employed machine learning approaches included: SVM with linear and radial basis function kernels, random forest (RF) and multinomial naïve Bayesian (NB). SVM models were generated with libSVM (v. 3.22) [29]; WEKA (v. 3.9.3) [30] was used for RF and for NB models. The SVM parameters (Cost and Gamma) corresponding to minimal RMSE in 3-fold CV were found by genetic algorithm driven optimization. The RMSE was estimated using a dedicated 3-fold CV, isolated from the cross-validation procedure used to evaluate the final models, mentioned below. Concerning RF, default parameters of WEKA were selected, with the number of generated trees equal to 100. No strategy was used to compensate the class imbalance in the dataset.

The modelling workflow is depicted in Figure 2: (1) dozens of ISIDA descriptor spaces (DSs) were generated (different fragment sizes and topologies); (2) for each DS, SVM and RF models were trained (individual models); (3) individual models were ranked according to their root mean squared error (RMSE) in 3-fold CV; (4) the best performing individual model

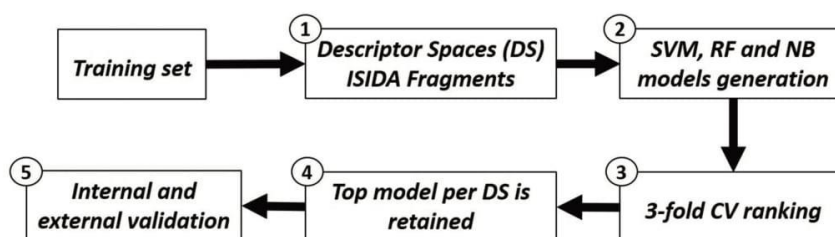


Figure 2. Model generation workflow.

for the given DS was retained; (5) models are internally and externally validated. Internal validation was carried out by random splitting 3-fold CV. This procedure was repeated 5 times after reshuffling (i.e. the property for each molecule is predicted 5 times). The Model quality criteria (see Figure 2) were assessed for each repetition followed by their averaging. During CV no further optimization of SVM parameters was performed. The absence of chance correlation was checked through the Y-scrambling procedure (repeated 100 times).

The Industrial set was used in external validation. In addition, it was predicted by the model TEST (Table 1). To evaluate the performance of regression models, the r^2 determination coefficient and the RMSE parameters are reported. For multi-class classification models, the sensitivity (Sn), specificity (Sp) and balanced accuracy (BA) are instead used. Dealing with multi-classes, the overall values for Sn, Sp and BA were computed as the weighted average among the classes based on the number of instances of the given class, following the same approach implemented in WEKA (v. 3.9.3) [30].

The following terminology is adopted:

- 'NICEATM original': the regression LD₅₀ model generated for the workgroup. Its training set is based solely on the NICEATM training set.
- 'NICEATM full': regression and multi-class classification models generated on all NICEATM data (i.e. training plus validation set).
- 'Global': regression and classification models generated on all collected data, externally validated on the Industrial set.

Applicability domain

The applicability domain was evaluated through the so-called 'fragment control' assessment (Figure 3, step 2): if a test molecule is found to have one fragment (i.e. a determined sequence of atoms and/or bonds) which is not present in the individual model, that molecule is marked to be outside the applicability domain since it is uncertain whether the model's predictions can be extrapolated to this not yet chartered chemical space zone [24].

Consensus modelling

To derive the consensus decision, the following strategy was implemented (Figure 3). The ensemble decision is taken either by computing the median (regression model) or by a majority vote (classification model) from the individual models of the different algorithms considered together (step 1). All out-of-AD predictions (based on the fragment control) are excluded (step 2) and the consensus is computed (step 3). Finally, a 4-grade reliability scale is associated to the output (step 4), based on a combined score of (i) the concordance of the predictions and (ii) the % of individual models, out of the total, for which the compound was inside the AD. The former was estimated by the median absolute deviation for regression models or the entropy value for classification models (SI, section 2).

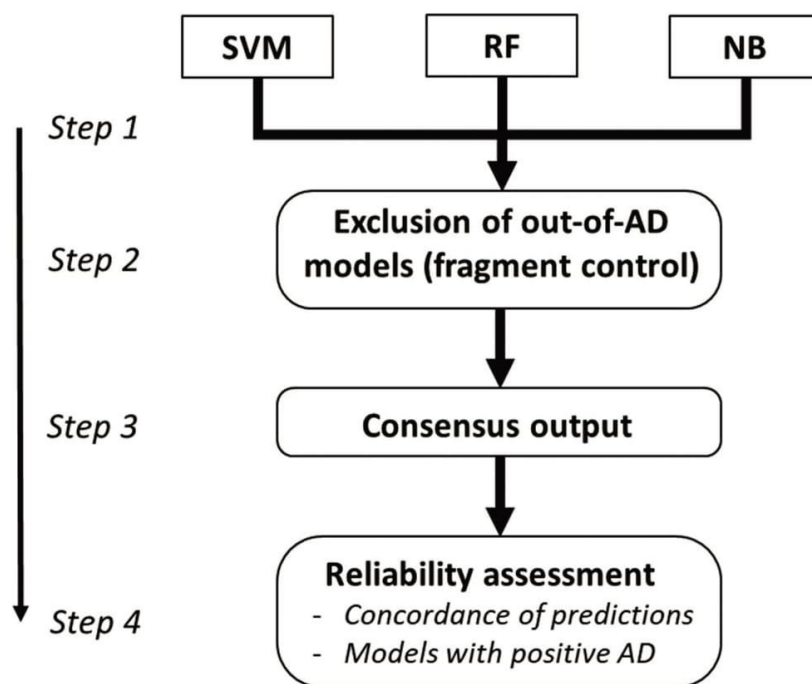


Figure 3. Consensus model workflow. Step 1: predictions for each algorithm (SVM, RF and NB) are merged together; Step 2 & 3: the consensus is the average of the predictions, excluding those models identifying the compound as out of applicability domain; Step 4: reliability assessment is associated to the output.

Graphical interpretation of predictions: coloratom

ISIDA ColorAtom [14] analyses local gradients of descriptors as reflecting their contributions to the variation of the modelled property [31]. A colour is assigned to each atom of the predicted molecule reflecting its positive or negative increment to the modelled property. This is a graphical representation of how the model interpreted the molecule for calculating the predicted value, not a mechanistic statement of the role played by each atom.

Results

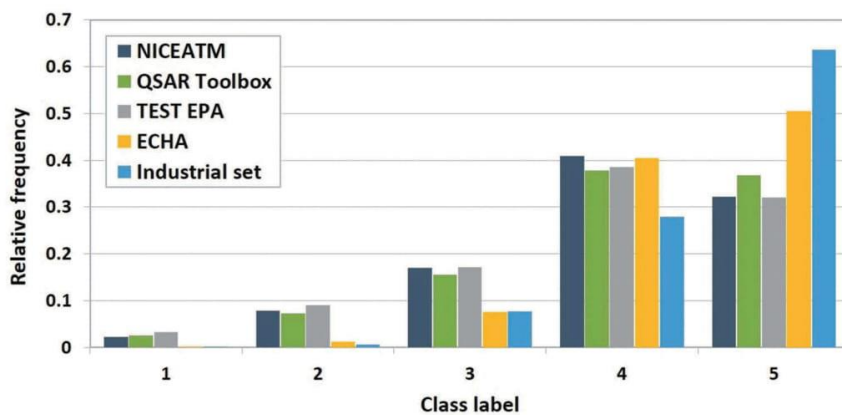
Curated datasets

Table 2 reports summary statistics of the collected datasets; Figure 4 shows the distribution in the five GHS classes (SI, Section 2). The distribution pattern is the same for NICEATM, QSAR Toolbox and TEST datasets, for which the most populated class is the GHS class 4; on the other hand, for ECHA and the Industrial set the GHS class 5 is the

Table 2. Statistics of the curated datasets.

Curated datasets	Total no.	Numerical pLD_{50} statistics			GHS class repartition				
		Min	Max	Mean	1	2	3	4	5
NICEATM ^a	10863	-2.71	4.6	-0.48	180	650	1395	3359	2643
QSAR Toolbox	10531	-3.34	4.21	-0.53	276	760	1628	3987	3880
TEST	7315	-2.71	4.21	-0.45	237	661	1250	2819	2348
ECHA	1717	-2.79	2.42	-0.97	4	20	131	694	868
Industrial set	1563	-4.57	1.31	-0.95	1	9	121	437	995
Blind set	462	-2.31	0.89	0.58	0	0	16	96	211
Global set ^b	11981	-4.57	4.21	-0.54	317	851	1773	4350	4690

^adataset used to build the 'NICEATM full' model; ^bdataset used to build the 'Global model'. The Global set^b was issued by merging of the whole public data.

**Figure 4.** Class frequency distribution for the classification model.

most abundant. The experimental variability, when multiple values for the same compound were available, was calculated to be 0.40 log unit.

Database comparison by GTM

Once the molecules are projected, landscapes are generated according to the envisaged property, and colours are assigned to the nodes of the map. In this context, two different landscapes were used: (i) the compound's database affiliation (i.e. NICEATM, QSAR Toolbox, etc.) and (ii) the LD_{50} value.

Database affiliation maps

With this analysis, the NICEATM was pairwise compared against all the remaining datasets. The goal was to verify if its set of compounds was sufficiently diverse to cover most of the chemical space, especially when confronted to the industrial context (i.e. the REACH registration dossiers on the ECHA database and the data provided by Solvay). Figure 5 shows all the pairwise comparison. Red areas are uniquely populated by the NICEATM dataset and blue by the others; intermediate colours are mixed populated areas. As visible from the first and the second landscape, NICEATM is almost

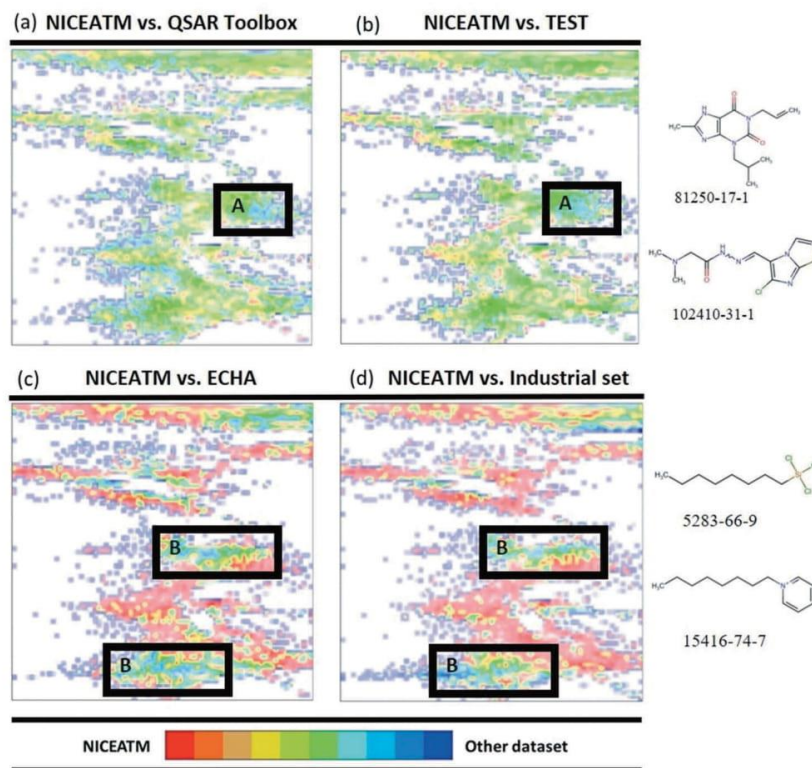


Figure 5. GTM database comparison. Each map compares the NICEATM vs. the other dataset: (a) QSAR Toolbox, (b) TEST, (c) ECHA and (d) Industrial set. Red regions are mainly populated by the NICEATM compounds and blue ones by the dataset it is compared to. White areas are empty regions of the map.

completely overlapping with QSAR Toolbox and TEST datasets. Some exceptions are two areas marked by the black rectangles 'A', indicative of some chemotypes under-sampled in the NICEATM dataset. For example, molecules with methylxanthine (CAS 81250-17-1; 66172-75-6) or imidazothiazole (CAS 102410-20-8; 102410-31-1) as substructures are almost unique to the QSAR Toolbox and TEST datasets.

For the third and fourth landscape, the situation is quite different: even though the chemical space is mainly dominated by NICEATM compounds (since its size is almost four times ECHA and the Industrial dataset), there are several spots dominated by ECHA or Industrial compounds (black rectangles 'B'). Interestingly, these areas are localized on a similar X, Y position of the map, suggesting that the NICEATM dataset is missing some chemotypes which are, however, shared between the Industrial set and ECHA. To provide few examples, the chemotype containing a sequence of Halogen-Silicium-Halogen atoms (e.g. CAS 5283-66-9) and long aliphatic chains terminating with

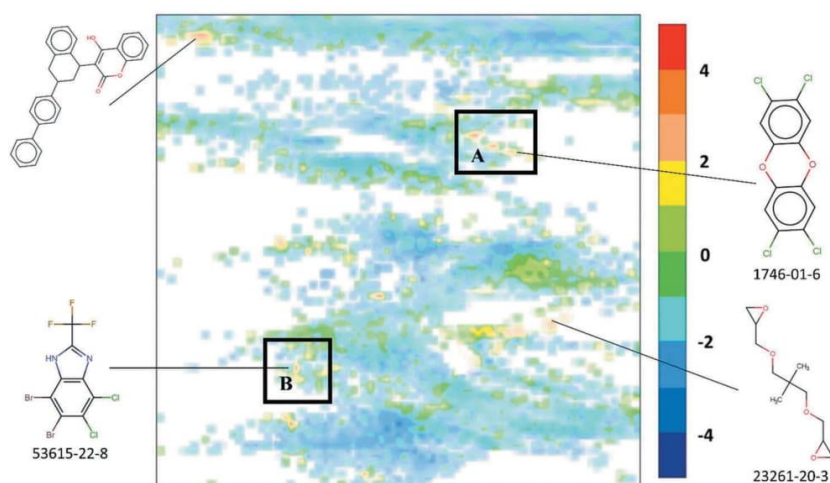


Figure 6. 'Global map' for LD₅₀. The map is built by merging all the available sources of data. Very toxic compounds are identified by red zones while less toxic compounds by blue ones.

a positively charged nitrogen-containing functional group (e.g. CAS 15416-74-7) are unknown or under-sampled to the other databases.

LD₅₀ property map

Figure 5 reports the Global map coloured according to the LD₅₀ value. There are several spots of very highly toxic chemicals (indicated by black rectangles). For example, the area delimited by rectangle 'A' is populated by members of the dioxine and furane family (such as TCDD and TCDF); while in the area of the rectangle 'B' there is a collection of chemicals with the benzimidazole as substructure (e.g. CAS 89427-34-9) (Figure 6).

Table 3. Model performances.

Regression	Model	Internal validation (3-fold CV) ^a			External validation	
		r ²	RMSE	r ² Y-scrb	RMSE	Data coverage (%) ^b
	NICEATM original	0.79 (0.050)	0.55 (0.051)	0.13	0.56	58 (287/479)
	NICEATM full	0.77 (0.045)	0.56 (0.053)	0.15	0.51	87 (205/235)
	Global model	0.78 (0.047)	0.55 (0.055)	0.12	0.47	94 (186/197)
	TEST ^c	—	—	—	0.61	90 (293/322)

Classification	Model	Internal validation (3-fold CV) ^a		External validation			
		BA	BA Y-scrb	BA	Sn	Sp	Data coverage (%) ^b
	NICEATM full	0.70 (0.031)	0.30	0.69	0.74	0.63	82 (669/811)
	Global model	0.70 (0.029)	0.32	0.72	0.76	0.69	85 (635/744)

Regression LD₅₀ model (upper part) and classification model (bottom part). ^aIn brackets, the standard deviation computed in the 3-fold CV is reported for the *r*² and RMSE values averaged over the number of repetitions. External validation is based on the Industrial set. BA = balanced accuracy, Sn = sensitivity, Sp = specificity. ^bThe first number is the data coverage in %; the number between the parentheses is a ratio of the number of compounds inside AD and the total number of compounds. ^cresults from the TEST model.

Table 4. Performance of selected machine learning methods.

Regression	Method	External validation	
		RMSE	Data coverage (%)
	Random forest	0.47	94
	SVM linear kernel	0.51	82
	SVM RBF kernel	0.50	97
	Global model	0.47	94

Classification	Method	External validation			
		BA	Sn	Sp	Data coverage (%)
	Random forest	0.74	0.82	0.66	81
	SVM linear kernel	0.69	0.81	0.56	87
	SVM RBF kernel	0.73	0.81	0.66	85
	Naïve Bayesian	0.64	0.60	0.68	80
	Global model	0.72	0.76	0.69	85

Regression LD₅₀ model (upper part) and classification model (bottom part). External validation is based on the Industrial set. BA = balanced accuracy, Sn = sensitivity, Sp = specificity.

Model performances

Table 3 reports performances of the generated models: regression LD₅₀ model (top) and classification model (bottom). In addition, the performances of the TEST tool are reported for LD₅₀. Individual machine learning algorithms performances are reported in Table 4. Overall, all the models scored a good prediction accuracy on the Industrial set, with RMSE values ranging from 0.47 to 0.56 and BA values from 0.69 to 0.72. TEST showed a good data coverage, being able to predict the 90% of the Industrial set. However, its prediction accuracy is worse (0.61 RMSE). The addition of new data is directly correlated to both an increase of prediction accuracy and data coverage. The latter increased from 58% for the NICEATM original model to 94% for the Global model (regression) and from 82 to 85% (classification models). This reflects that the NICEATM data are more comprehensive regarding GHS data. The contamination of models by chance correlations is limited as monitored by Y-scrambling: the maximum observed r^2 and BA metrics had very low values ($r^2 < 0.2$ and BA < 0.5). Overall, all the models are robust and well generalizable: performances in external validation are comparable to those in cross-validation and the data coverage reaches very high levels.

Performances on blind set

Finally, the last version of the model (built on all collected data, i.e. public + industrial) was challenged to predict a new list of 462 unique compounds made available afterwards. Of them, 224 had a precise estimation of LD₅₀; while 347 had only the categorical statement. Thus, both the regression and classification models were used.

For confidentiality reasons, this dataset cannot be disclosed, and only some general information can be provided. It comprises quite heterogeneous chemical structures,

Table 5. Performances of public and industrial data ensemble models on the blind set.

Blind set	Regression			Classification			
	r^2	RMSE	Data coverage (%)	Sn	Sp	BA	Data coverage (%)
	0.3	0.48	92 (207/224)	0.77	0.97	0.87	93 (303/323)

BA = balanced accuracy, Sn = sensitivity, Sp = specificity.

from high molecular weight compounds such as long chain aliphatic surfactants and halogenated biphenyls to much smaller ones such as phenol derivates and simple amides. A good number of compounds are organofluorine derivatives. The molecular weight ranges from 41 to 1094 with an experimental pLD_{50} from -2.31 to 0.89 log unit. This dataset is mainly 'non-toxic', as almost 60% of the compounds are not classified under the GHS system (i.e. $\text{LD}_{50} > 2000$ mg/kg).

Performances for the regression model are similar to the previous external validation ($\text{RMSE}_{\text{blind}} = 0.48$ vs. $\text{RMSE}_{\text{ext}} = 0.47$; Tables 3 and 5). In both instances, the prediction accuracy is better than the one estimated through cross-validation ($\text{RMSE}_{\text{cv}} = 0.55$). On the other hand, the classification model performed better ($\text{BA}_{\text{blind}} = 0.87$ vs. $\text{BA}_{\text{ext}} = 0.72$, Tables 3 and 5). This is probably due to the unbalanced nature of the Blind set, as the majority of the compounds belong to GHS class 5.

The Blind set r^2 value may appear disappointing at first sight. However, it must be noticed that its pLD_{50} property range is considerably smaller than the Global model's one (-4.57 – 4.21). Figure 7 depicts experimental/predicted scatterplot of the Global model's training set (evaluated in 3-fold CV) overlapped with the Blind set. As expected, the Blind set covers only a fraction of the entire property range; this explains the low determination coefficient value.

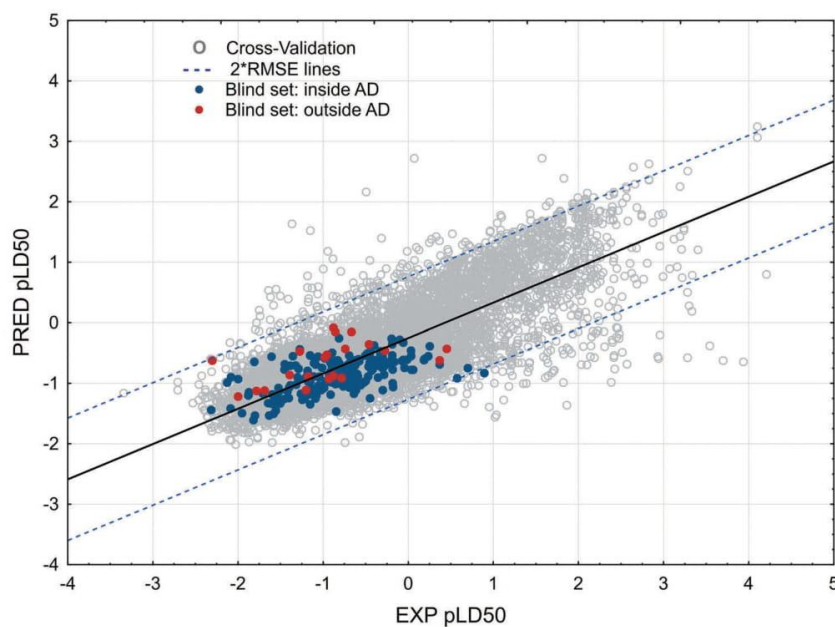
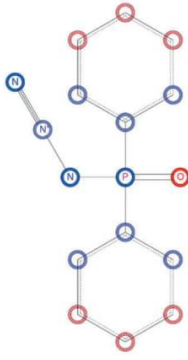
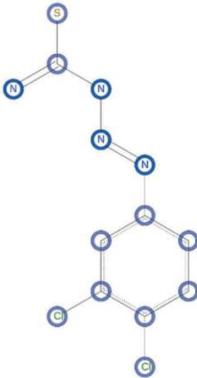
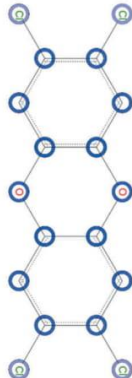


Figure 7. Blind set scatterplot. Grey points represent the training set evaluated in 3-fold CV; red and blue points indicate Blind set molecules outside and inside the AD, respectively. Blue dashed lines mark the $\pm 2 \text{ RMSE}_{\text{cv}}$ limits.

Table 6. ColorAtom output.

<i>Diphenylphosphinyl azide</i>	<i>Chloropromurite</i>	<i>TCDD</i>
CAS 4129-17-3 LD ₅₀ = 240 mg/Kg	CAS 5836-73-7 LD ₅₀ = 1.0 mg/Kg	CAS 1746-01-6 LD ₅₀ = 0.02 mg/kg
		

Colours refer to atomic contribution to the predicted value of the property (i.e. pLD₅₀ values). Red colour means that the atom contributes to decrease its value (lowering the toxicity); while blue means an increase of its value (i.e. increasing the toxicity).

Model interpretation with coloratom

For in-depth structure-activity dependence analysis, Table 6 reports three molecules chosen as examples for the ColorAtom: diphenylphosphinyl azide, chloropromurite and TCDD. As expected, as the compounds become more toxic, 'blue-coloured' atoms become dominant. For the first compound, the 'triazo-' substructure is the main driver for its correct prediction as an acute toxic. Similarly, chloropromurite presents two functional groups which are associated with enhanced toxicity: the 'diazo' (CNN) and the 'thiocyanate' (SCN). Finally, all the atoms of TCDD are represented as promoters of toxicity. In these cases, the colouration patterns are actually in agreement with the mechanistical interpretation of the analysed functional groups [32,33]. SI, Section 3 reports additional examples of compounds with the same functional groups that showed the same colouration scheme.

Discussion

Among the QSAR tools for the estimation of the oral rat acute toxicity reported in Table 1, only one is freely available (TEST). The collaborative NICEATM workshop aimed at filling this gap, by proposing a set of new models which will be freely available [12], implemented in the open source platform OPERA [34]. On the Industrial set, the predictive power of the models (regression and classification) was found to be reasonably high, with RMSE values of 0.47–0.56 and BA values of 0.69–0.71 (5 five classes) for the NICEATM and the Global models, respectively. Data coverage was quite unsatisfactory with the original NICEATM model (58% on the Industrial set), but after the addition of new data from several databases (QSAR

Toolbox, TEST, ECHA) it significantly improved: reaching 85 and 94% for the classification and the regression model, respectively. New data improved models' predictive power as well, with the biggest improvement for the regression model, where the RMSE decreased from 0.56 to 0.47. Finally, new models were built on the ensemble of public data and Industrial data. Cross-validation performances for the regression model were $r^2 = 0.78$ and RMSE = 0.53; while for the classification model BA = 0.69. These models were also externally validated on the Blind set (Table 5), showing good prediction accuracy and data coverage: RMSE = 0.47 with 92% inside AD (regression); and BA = 0.87 with 93% inside AD (classification).

GTM was employed to show positions of 109 'out-of-AD' compounds (Table 3, bottom part) in the public data chemical space, which constitutes the training sets of the models (Figure 8). As expected, the majority of them are located in the regions mainly populated by external set compounds (blue areas), indicating that their chemotypes are quite unique and non-overlapping with those in the models' training sets. For example, compound CAS 34762-90-8 presents the unique chemotype – N^+BCl_3 . Some compounds are singletons far away from the occupied chemical space, such as CAS 24108-89, a pigment characterized by a very complex and diverse chemical structure. On the other hand, there are some out-of-AD compounds projected in areas of the public data chemical space. This happens when the given molecule both shares several functional groups with the training set compound and contains new chemotype. For example, drometrizole trisiloxane (CAS 155633-54-8), contains trisiloxane

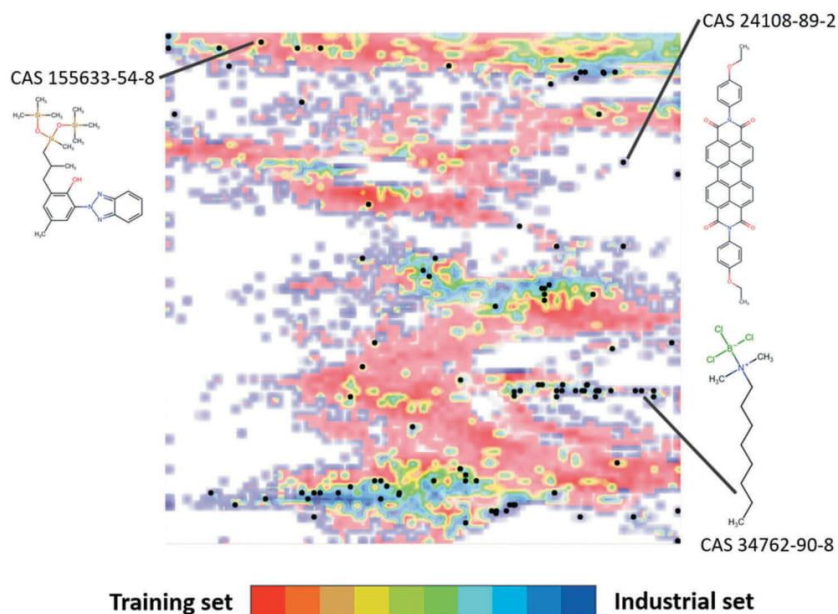


Figure 8. Tracking the out-of-AD compounds in the chemical space. Zones populated by the training set and Industrial set compounds are highlighted in colour. Black points represent the projections of 109 out of AD compounds.

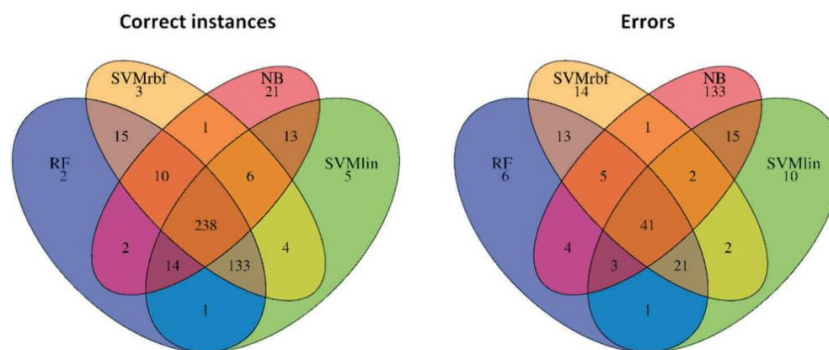


Figure 9. Venn diagrams comparing individual multi-class classification models performances in external validation on Industrial set. Left and right diagrams correspond to correct and erroneous predictions, respectively.

motif absent in the training set, and drometrizole motif present in several training set compounds.

Performances of individual models for the multi-class classification in external validation on Industrial set are represented by means of Venn diagrams (Figure 9). Comparison is performed for both the correct (left) and for erroneous (right) predictions. These results support the conclusion about the robustness of consensus model, since great majority of instances (238) were simultaneously correctly predicted by all four machine-learning algorithms.

Our developed models follow the OECD principles [11]. The endpoint (LD_{50}) is well defined. Goodness-of-fit, robustness and predictivity were evaluated using internal and external 3-fold Cross-Validation (CV), Y-scrambling, and external validation [35–37]. The AD of the models was defined using a fragment control assessment [24] together with a reliability scoring function.

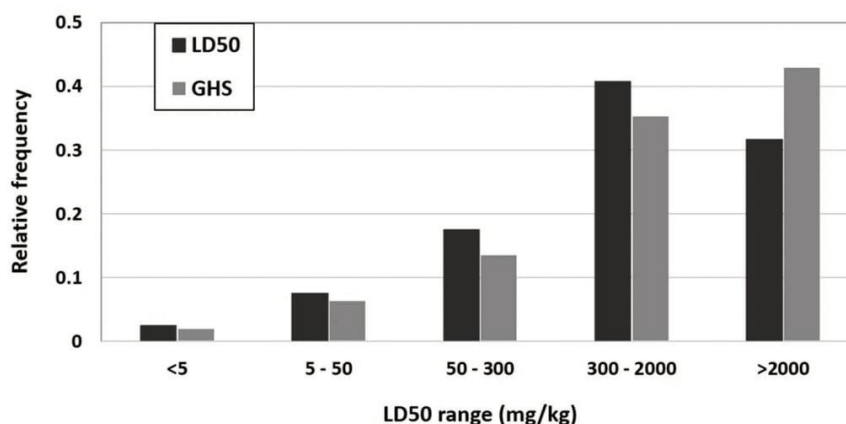


Figure 10. Continuous- LD_{50} and the GHS-classes distribution comparison.

Figure 10 depicts the relative frequency distribution for the continuous- LD_{50} and the GHS-classes for the full NICEATM dataset (training and evaluation set). It is interesting to notice that LD_{50} data is always more frequent than categorical assays for more toxic compounds, with the biggest difference (+15%) for the medium toxicity class (GHS class 4, i.e. 300–2000 mg/kg). On the other hand, for low toxicity values (GHS category 5, i.e. >2000 mg/kg,) categorical data becomes much more frequent. This is related to the current regulatory requirements: in case the substance shows high toxicity, it could be more advantageous for the registrant to have the precise LD_{50} , in order to avoid a potential overestimation of the compound's toxicity, leading to a less desirable GHS classification. On the other hand, when the substance is far from GHS thresholds, a looser toxicity estimation could be enough. This bias of the data is also reflected in the model's learned rules: we noticed that the regression LD_{50} model tends to overestimate the toxicity of some very low toxic compounds. Furthermore, as mentioned in the introduction, current guidelines do not foresee anymore the precise estimation of LD_{50} . Instead, the goal is to perform limit tests (OECD 420, 423, 425) for estimating the GHS categories, which allows the use of fewer animals. For this reason, new LD_{50} data is unlikely to be generated, and future in-silico models will have to be updated based on the new categorical data.

Conclusions

In this work we report predictive models of acute oral toxicity obtained in the context of the National Toxicology Programme Interagency Centre for the Evaluation of Alternative Toxicological Methods (NICEATM) workgroup [14,18].

The datasets including 11211 and 13680 compounds for 'Global' regression and classification models respectively, were collected from the publicly available sources. To our knowledge, these are the biggest datasets ever used for the modelling of oral acute toxicity in rodent.

The models were obtained using ISIDA fragment descriptors [24] and support vector machine, random forest and naïve Bayes machine learning methods. Compared to our contribution to the NICEATM project in this paper (i) a new classification model based on GHS toxicity categories was generated (ii) Global models were generated by collecting new data.

The predictive performance of the models was assessed on independent Industrial set provided by Solvay. It has been demonstrated that both regression and classification Global models obtained in this work (RMSE = 0.47 and BA = 0.72) perform better than the previously reported NICEATM models (RMSE = 0.56 and BA = 0.69). Moreover, the Global models have much larger applicability domain: the data coverage on the Industrial set is 85% and 82% (classification) and 94% and 58% (regression) for Global and NICEATM models, respectively. Finally, new models built on the ensemble of public data and Industrial dataset were validated on a set of 462 new structures provided by Solvay. This blind test proved reasonably high predictive power of the models: RMSE = 0.48 and BA = 0.87 for regression and classification, respectively.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

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

D. Horvath  <http://orcid.org/0000-0003-0173-5714>

References

- [1] 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/EEC 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.
- [2] A. Gissi, K. Louekari, L. Hoffstadt, N. Bornatowicz, and A.M. Aparicio, *Alternative acute oral toxicity assessment under REACH based on sub-acute toxicity values*, ALTEX 34 (2017), pp. 353–361. doi:10.14573/altex.1609121.
- [3] *OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects*, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2019. Available at <http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm>.
- [4] I. Tsakovska, I. Lessigiarska, T. Netzeva, and A.P. Worth, *A mini review of mammalian toxicity (Q)SAR models*, QSAR Comb. Sci. 27 (2008), pp. 41–48. doi:10.1002/qsar.200710107.
- [5] J.X. Guo, J.J.-Q. Wu, J.B. Wright, and G.H. Lushington, *Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: A molecular modeling study*, Chem. Res. Toxicol. 19 (2006), pp. 209–216. doi:10.1021/tx050090r.
- [6] A.P. Freidig, S. Dekkers, M. Verwei, E. Zvinavashe, J.G.M. Bessems, and J.J.M. van de Sandt, *Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds*, Toxicol. Lett. 170 (2007), pp. 214–222. doi:10.1016/j.toxlet.2007.03.008.
- [7] A.A. Toropov, B.F. Rasulev, and J. Leszczynski, *QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: Comparative analysis by MLRA and optimal descriptors*, QSAR Comb. Sci. 26 (2007), pp. 686–693. doi:10.1002/qsar.200610135.
- [8] Legal Information Institute, *Predictive Models for Acute Oral Systemic Toxicity*, National Toxicology Program, US Department of Health and Human Services, Bethesda, Maryland, US, 2019. Available at <https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/acute-systemic-tox/models/index.html>.
- [9] *Electronic Code of Federal Regulations (40 Cfr Part 156)*, United States Government Publishing Office, Washington DC, US, 2019. Available at <https://www.law.cornell.edu/cfr/text/40/part-156.html>.
- [10] The European Parliament and the Council of the European Union, *Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006*, Off. J. Eur. Union 353 (2008), pp. 1–1389.
- [11] 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.
- [12] N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrickc, and G. Patlewicz, *Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation*, Comput. Tox. 201 (2018), pp. 489–492.
- [13] D. Ballabio, F. Grisoni, V. Consonni, and R. Todeschini, *Integrated QSAR models to predict acute oral systemic toxicity*, preprint (2019), submitted for publication. Available at <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800124>.

- [14] 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/cgi-bin/predictor.cgi>.
- [15] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, *The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 493–500. doi:10.1021/ci025584y.
- [16] T. Martin, P. Harten, and D. Young, *(TEST) Toxicity Estimation Software Tool V 4.1*, US Environmental Protection Agency, 2012; software available at <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.
- [17] Simulations Plus Inc, *ADMET Predictor*, Simulations Plus Inc., Lancaster, US, 2019; software available at <http://www.simulations-plus.com/>.
- [18] ACD/Labs, *ACD/Percepta Platform V 2018.1*, Advanced Chemistry Development, Inc. (ACD/Labs), 2019; software available at <http://www.acdlabs.com/>.
- [19] TerraBase, *TerraQSAR Biological Effect Programs*, TerraBase Inc., 2006; software available at <http://www.terrabase-inc.com/>.
- [20] Accelrys, *(TOPKAT) TOXicity Prediction by Komputer Assisted Technology V 3.1*, Accelrys software Inc., San Diego, CA, US, 2019; software available at <http://www.3dsbiovia.com/>.
- [21] 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>.
- [22] OASIS, *The OECD QSAR toolbox v 4.1*, OASIS Laboratory of Mathematical Chemistry, 2017; software available at <http://www.oecd.org/chemicalsafety/risk-assessment>.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath, and A. Varnek, *Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge*, J. Chem. Inf. Model. 55 (2015), pp. 84–94. doi:10.1021/ci500575y.
- [28] D. Horvath, I. Baskin, G. Marcou, and A. Varnek, *Generative topographic mapping of conformational space*, Mol. Inform. 36 (2017), pp. 24–36. doi:10.1002/minf.201700036.
- [29] 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.
- [30] I.H. Witten and E. Frank, *The Weka Workbench*, in *Data Mining: Practical Machine Learning Tools and Techniques*, I.H. Witte, eds., Morgan Kaufman Publishers, San Fransisco, 2005, pp. 363–449.
- [31] G. Marcou, D. Hor Vath, 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.
- [32] G.M. Cramer, R.A. Ford, and R.L. Hall, *Estimation of toxic hazard-a decision tree approach*, Food Cosmet. Toxicol. 16 (1976), pp. 255–276. doi:10.1016/S0015-6264(76)80522-6.
- [33] S. Bhatia, T. Schultz, D. Roberts, J. Shen, L. Kromidas, and A. Marie Api, *Comparison of Cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgment*, Regul. Toxicol. Pharmacol. 71 (2015), pp. 52–62. doi:10.1016/j.yrtph.2014.11.005.
- [34] K. Mansouri, C.M. Grulke, R.S. Judson, and A.J. Williams, *OPERA models for predicting physicochemical properties and environmental fate endpoints*, J. Cheminform. 10 (2018), pp. 1–19. doi:10.1186/s13321-018-0263-1.
- [35] A. Tropsha, *Best practices for QSAR model development, validation, and exploitation*, Mol. Inform. 29 (2010), pp. 476–488. doi:10.1002/minf.201000061.

- [36] D. Fourches, E. Muratov, and A. Tropsha, *Trust but verify: On the importance of chemical structure curation in chemoinformatics and qsar modeling research*, J. Chem. Inf. Model. 50 (2010), pp. 1189–1204. doi:10.1021/ci100176x.
- [37] 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/qsar.200390007.

4.1.6 Androgen and Estrogen receptor binding

Endocrine hormones regulate various functions, including metabolism, sleep, growth and development. An Endocrine Disrupting Chemical (EDC) is an exogenous substance that alters the functions of the endocrine system and consequently causes adverse effects. The endocrine disruption event is the result of a series of mechanisms, which can be in part explained by the interaction with specific Nuclear hormone Receptors (NRs). The most studied NRs include the Estrogen Receptor (ER) and the Androgen Receptor (AR).

Few models have been published in the past years. However, with the exception of two more recent projects, the “Collaborative Estrogen Receptor Activity Prediction Project” (CERAPP) and the “Collaborative Modelling Project for Androgen Receptor Activity” (CoMPARA), the relatively small training set sizes (ranging from 66 to 645 compounds) are symptomatic of the limitations of existing models.

Thus several models and datasets are currently available. However, this is the first time that a meta-analysis is proposed comparing the data sources. A surprising observation is that the dataset size seems to be unrelated with the models’ performances: the addition of more information tends to dilute the relation between the chemical structure and the modeled properties. In this work we search for some roots explaining this observation and we propose an in-depth analysis about the available ER and AR data, illustrated by the generation of QSAR models.

Binary classification models were generated using data prepared within the CERAPP and CoMPARA frameworks. Additional external validation was carried out based on data collected from two sources: the “Tox21 Data Challenge 2014” (hereafter named Tox-DC) [NIH, 2014] and an extraction from PubChem database. We encountered some interesting points worth mentioning:

- Significant discrepancies of the experimental labels when comparing the different sources, indicating that the CERAPP and CoMPARA datasets shall not be fused with other kinds of assays;
- Biological events such as receptor binding, agonist or antagonist effects may refer to biological phenomena that do not necessarily have an obvious causal link;

- Some caution has to be taken considering the external validation set for the CERAPP and CoMPARA.

Generated models illustrated how these discrepancies have negatively impacted performances. Binding activity models reproduced the CERAPP/CoMPARA frameworks, showing comparable results (ER $BA_{\text{ext}} = 0.60$ and AR $BA_{\text{ext}} = 0.73$). Additional external validation was carried out on additional Tox-DC and the PubChem datasets ($BA_{\text{Tox-DC}} = 0.65 - 0.72$ and $BA_{\text{PubChem}} = 0.66 - 0.71$). The ability of the models to detect compounds truly recognized by a nuclear receptor are limited by the poor agreement between the data sources. Data sources agreement is illustrated by a sensibility measure ranging from 0.34 to 0.49. It seems that different data sources are merging results of different biological meaning, that could explain the degraded performances of the models. The CERAPP/CoMPARA datasets are standing out particularly, suggesting that they shall not be fused with other data sources.

This article has been submitted to the peer-reviewed journal “SAR and QSAR in Environmental Research”; the manuscript reported here corresponds to the submitted version.



4.1.6 Liaison aux récepteurs nucléaires androgènes et œstrogènes

Les hormones endocrines régulent diverses fonctions, dont le métabolisme, le sommeil, la croissance et le développement. Un perturbateur endocrinien chimique (EDC) est une substance exogène qui modifie les fonctions du système endocrinien et entraîne par conséquent des effets indésirables. La perturbation endocrinienne est le résultat d'une série d'évènements, qui peuvent s'expliquer en partie par l'interaction du composé chimique avec des récepteurs nucléaires hormonaux (NR) spécifiques. Les NR les plus étudiés sont le récepteur aux œstrogènes (ER) et le récepteur aux androgènes (AR).

Peu de modèles ont été publiés ces dernières années. Cependant, à l'exception de deux récents projets, le «Collaborative Estrogen Receptor Activity Prediction Project» (CERAPP) et le «Collaborative Modeling Project for Androgen Receptor Activity» (CoMPARA), la taille relativement petite des jeux d'entraînement utilisés (allant de 66 à 645 composés) est symptomatique des limites des modèles existants.

Il existe donc plusieurs modèles et jeux de données de tailles variables actuellement disponibles. Cependant, c'est la première fois qu'une méta-analyse est proposée pour comparer les sources de données. En effet, une observation surprenante est que la taille des jeux de données d'entraînement semble sans rapport avec les performances des modèles: l'ajout d'information tend même à diluer la relation entre la structure chimique et les propriétés modélisées. Dans ce travail, nous recherchons des causes expliquant cette observation ce qui nous conduit à proposer une analyse approfondie des données ER et AR disponibles, illustrée par la génération de modèles QSAR.

Des modèles de classification binaire ont été générés à partir de données préparées dans les cadres à l'occasion des collaborations CERAPP et CoMPARA. Une validation externe supplémentaire a été réalisée sur la base de données collectées à partir de deux sources: le «Tox21 Data Challenge 2014» (ci-après dénommé Tox-DC) [41] et une extraction de la base de données PubChem. Nous avons réalisé quelques observations intéressantes qui méritent d'être mentionnées:

- Des différences importantes entre les étiquettes expérimentales d'un même composé entre différentes sources, indiquant en particulier que les jeux de données CERAPP et CoMPARA ne doivent pas être fusionnés avec d'autres types de données;
- Les événements biologiques tels que la liaison à récepteur, les effets agonistes ou antagonistes peuvent se référer à des phénomènes biologiques n'ayant pas nécessairement un lien de causalité évident;
- Une certaine prudence doit être mise en œuvre quand il s'agit de traiter les jeux de données de validation externe issus des collaborations CERAPP et CoMPARA.

Les modèles générés ont illustré comment ces différences entre sources de données ont eu un impact négatif sur les performances. Les modèles d'activité sur les récepteurs nucléaires reproduisaient les conditions de CERAPP / CoMPARA, montrant des résultats comparables (ER BAext = 0,60 et AR BAext = 0,73). Une validation externe supplémentaire a été effectuée sur des jeux de données Tox-DC et PubChem supplémentaires (BATox-DC = 0,65 - 0,72 et BAPubChem = 0,66 - 0,71). La capacité des modèles à détecter les composés vraiment reconnus par un récepteur nucléaire sont limités

par le faible accord entre les bases de données. Cet accord est illustré par une mesure de sensibilité allant de 0,34 à 0,49. Il semble que différentes sources fusionnent des résultats biologiques de signification différente, ce qui expliquerait ces performances médiocres. Dans cet ensemble, les données CERAPP / CoMPARA ressortent particulièrement, indiquant qu'elles ne doivent pas être fusionnées avec d'autres sources de données.

Cet article a été soumis au journal à comité de lecture « SAR and QSAR in Environmental Research » ; le manuscrit ci-dessous correspond à la version soumise.

The limitations of the available data adversely impact the performance of models predicting the complex event of endocrine disruption

Filippo Lunghini^{1,2}, Gilles Marcou^{1,*}, Philippe Azam², Marie-Hélène Enrici², Erik Van Miert², Alexandre Varnek^{1*}

¹Laboratory of Chemoinformatics, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France; ²Toxicological and Environmental Risk Assessment unit, Solvay S.A., 85, avenue des Frères Perret, 69192, St. Fons, France

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

Abstract

In this work we present a meta-analysis of available data concerning two nuclear receptors involved in endocrine disruption: the Estrogen (ER) and the Androgen (AR) receptor. We noticed that the dataset size of already existing ER/AR models seems to be unrelated with the models' performances: the addition of more information did not appear to strengthen modeled relationship between the chemical structure and the ED properties. To analyze this issue, we collected ED-relevant data from multiple sources, including the CERAPP/CoMPARA collaborations, the Tox21 data challenge and from ChEMBL and PubChem. The Generative Topographic Mapping approach has been employed to compare the chemical space of considered data sources: datasets suffered from a low agreement between experimental values, as the average concordance for binding class labels was only 42 %.

Collected data was used to train classification binding activity and quantitative Relative Binding Affinity (RBA) and median Inhibition Concentration (IC₅₀) models. Models showed mixed performances: classification models' abilities to detect truly binding compounds are limited (sensitivities: ER = 0.34, AR = 0.49) and RBA and IC₅₀ models showed mediocre determination coefficient values in external validation ($R^2 = 0.44 - 0.76$). Such low R^2 values were caused by the presence of several outliers due to misinterpreted experimental endpoints or wrongly reported values. For analogous reasons, the merging of assays having different biological meaning affected the performances of binding activity models.

Developed models and employed training and test sets, counting a total of 6215 (ER) and 3789 (AR) unique compounds, are made freely available.

Keywords: QSAR/QSPR; Generative topographic mapping (GTM); Estrogen/Androgen Receptor; Endocrine Disruptors, REACH

Introduction

The endocrine system regulates a large numbers of biological functions, including metabolism, sleep, growth and development [1] by the release of hormones and their binding to cellular receptor. An Endocrine Disrupting Chemical (EDC) is an exogenous substance that alters the functions of the endocrine system to the point of causing adverse effects. EDCs can exert their effects by several ways: (i) mimicking normal hormones such as estrogens and androgens; (ii) antagonizing hormones receptors; (iii) altering the pattern of synthesis and metabolism of hormones or (iv) modifying hormone receptor levels [2]. Endocrine disruption (ED) can result from the interaction of an exogenous chemical with specific Nuclear hormone Receptors (NRs). The most studied NRs include: the Estrogen Receptor (ER), the Androgen Receptor (AR), the Aryl hydrocarbon Receptor (AhR), the Thyroid Hormone (TA), Retinoic Acid (RAR) and Peroxisome Proliferator-Activated Receptor (PPAR) [3].

Chemicals with an ED potential are of particular concern for human health and the environment: for instance, under the European REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) Regulation, EDs are considered to be substances of very high concern [4]. Recently, the European Chemicals Agency (ECHA) has developed a guidance document for ED screening [3,5] and an OECD guidance (OECD TG 150) has been published [6]. These documents focus mainly on the estrogenic, androgenic, thyroidal and steroidogenic (EATS) modalities, which are pathways (but not only) of which the disruption can potentially lead to endocrine disruption. This is because EATS modalities are currently the ones for which there is a relatively good mechanistic understanding of how substance-induced perturbations may lead to adverse effects. As these guidelines were published recently (2018), rigorously standardized and homogeneous data on different ED modalities has yet to be generated. Therefore, compared with other toxicological endpoints, the number of QSARs related to endocrine disruption is significantly lower. This is due to the diversity and biological complexity of this family of endpoints and the fact that endocrine disruption is not a toxicological endpoint *per se*, but one out of the many modes-of-action which may result in adverse effects [4] as well as the limited amount of data on EDCs testing [7]. Three OECD guidelines are relevant to test ER/AR binding: OECD 455, 458 and 493 [8].

In the past years several QSAR models have been generated, mainly targeting the androgen and the estrogen receptors, due to the higher amount of available information. Table 1 reports some of already published models on ER and AR. More details about available ED models can be find in the work of E. Lo Piparo and A. Worth [4]. Most of them are binary classification models

(discriminating binders “B” or non-binders “nB”); in case of continuous properties, the relative binding affinity (RBA) is commonly used; followed by the median inhibitory concentration (IC50).

Table 1. Available models for ED screening (ER and AR).

Receptor	Property	Descriptors	Algorithm ^b	Size training set	Test set ^d	Ref.
Estrogen receptor	IC50	2D	MLR	86	R ² = 0.55	[9]
	IC50	Steric field	PLS	81	R ² = 0.74	[10]
	IC50	2D	PLS	127	R ² = 0.72	[11]
	RBA	Dragon	MLR	150	R ² = 0.88, RMSE = 0.62	[12]
	RBA	Dragon	MLR	232	R ² = 0.79, RMSE = 0.99	[13]
	RBA	Dragon	k-NN	546	R ² = 0.73	[14]
	RBA	Dragon	k-NN, RF	645	BA = 0.82	[15]
	AC50	Various	CERAPP consensus	1677	BA = 0.55 – 0.66 ^a	[16]
Androgen receptor	RBA	2D & 3D	COREPA ^c	202	BA = 0.81	[17]
	IC25	Molecular fragments	PLS	523	BA = 0.77	[18]
	Binding	Molecular fragments	-	595	BA = 0.74	[19]
	AC50	Various	CoMPARA consensus	1688	BA = 0.60 – 0.85 ^a	[20]

^aminimum and maximum value among reported models; ^bMLR = Multiple Linear Regression, k-NN = k-Nearest Neighbors, RF = Random Forest, PLS = Partial Least Squares; ^cprobabilistic classification scheme as described in [17]; ^dPerformances (external validation) have been collected from the work in the “Ref.” column; R² = determination coefficient, RMSE = Root Mean Squared Error, BA = Balanced Accuracy.

With the exception of two more recent projects, i.e. the “Collaborative Estrogen Receptor Activity Prediction Project” (CERAPP) [16] and the “Collaborative Modelling Project for Androgen Receptor Activity” (CoMPARA) [20], the relatively small training set size (ranging from 66 to 645 compounds) denotes a limitation of published models, especially when compared to the huge number of structurally different man-made and natural compounds to which we could be exposed, e.g. under REACH, more than 22100 unique substances are registered (May, 2019) [5].

In this work, we present a meta-analysis comparing the data sources in the context of QSAR model development targeting AR and ER. Indeed, a surprising point shown by Table 1 is that the dataset size seems to be unrelated with the models' performances: the addition of more information did not appear to strengthen modeled relationship between the chemical structure and the ED properties. In this work we investigate some root causes explaining this observation and we present an in-depth analysis of the available ER and AR data. Available data was checked

Binary classification models were generated using data prepared within the CERAPP and CoMPARA frameworks. Additional external validation was carried out based on data collected from two sources: the "Tox21 Data Challenge 2014" (hereafter named Tox-DC) [21] and an extraction from PubChem database [22]. We encountered some interesting points worth mentioning:

- Significant discrepancies of the experimental labels when comparing the different sources, indicating that the CERAPP and CoMPARA datasets shall not be fused with other kinds of assays;
- The test sets of the CERAPP and CoMPARA collaborations must be handled with care to avoid misinterpretations;

Moreover, to get a more comprehensive picture of ED modelling, we discuss the generation of quantitative QSAR models for estimating the ED potential: the IC50 and the RBA.

All collected datasets are freely available on [Zenodo \(10.5281/zenodo.3935808\)](https://zenodo.org/record/3935808). Our models are available through the online ISIDA/Predictor platform [23], at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi.

Methods

Description of the modelled properties

We modelled three separate properties for both ER and AR: the receptor's binding activity ("B", binder / "nB", non-binder), the relative binding affinity (RBA) and the median inhibitory concentration (IC50).

Binding activity

The CERAPP and CoMPARA projects share the same data source, based on a collection of *in vitro* High-Throughput Screening (HTS) assays included in the Environmental Protection Agency (EPA) “ToxCast” program and the inter-agency “Toxicology Testing in the 21st Century”, involving the EPA, the Food and Drug Administration (FDA), the National Institutes of Health (NIH) and the National Toxicology Program (NTP) [24–28]. We generated a binary classification model using CERAPP and CoMPARA data. In this instance, it must be highlighted that the modelled property is not, strictly speaking, binding to the receptor, but an integration of several HTS assays exploring multiple sites in the ER / AR signaling pathway chain [16,20]. Respectively, 18 and 11 assays were selected for ER and AR. Subsequently, in order to assign a unique label, a mathematical model was developed to integrate *in vitro* data in a final score (ranging from 0 to 1) which is eventually converted into a label “active” / “inactive” meaning that a compound is considered as a “disruptor” / “non-disruptor” of a given pathway. The cutoff on the score to assign the labels is optimized on a Receiver Operating Characteristic (ROC) curve and is equal to 0.01 [16,25].

The CERAP and CoMPARA challenges include also agonist/antagonist labels. Yet, generation of models for agonist and antagonist events led to unacceptable results (see SI, Section 1). Therefore, we decided to focus our investigations only on the general binding event to given NRs, i.e. if the compound does interact with the receptor (“binder”) or not (“non-binder”).

Relative binding affinity

The RBA, as defined by equation (1), compares the IC₅₀ of a test chemical with that of the reference compound (i.e. 17β-estradiol for ER or R1881 for AR). The IC_{50_{ref}} and IC_{50_{test}} are the concentrations of the reference compound and of the test compound at 50% inhibition of radiolabeled [³H]-17β-estradiol or [³H]-R1881 binding to the estrogen or androgen receptor, respectively. Therefore, the RBA of estradiol and R1881 is equal to 100%. In this study we employed the logarithm of the RBA (logRBA). As reported in literature [29,30], compounds with log RBA values >0 and <-3 can be considered strong binders and non-binders, respectively.

$$RBA = \frac{IC_{50_{ref}}}{IC_{50_{test}}} * 100 \quad (1)$$

Median inhibitory concentration

The IC₅₀ is derived from a dose response curve obtained from competitive binding experiment with a radioligand (radio tagged 17 β -estradiol or R1881). It is the ligand concentration displacing 50% of the binding of the radioligand. Although this definition cannot be easily related to a dissociation constant, it is considered that in a given assay the values are comparable across the tested compounds. However, the IC₅₀ value is assay specific, thus the RBA is considered to be more relevant when combining the output of various assays. A conservative cutoff value of IC₅₀>10 nM has been reported in the literature to define non-binders [31]. IC₅₀ values originally expressed in nM were transformed to inverse log units expressed in molar concentration (i.e. -logIC₅₀ or pIC₅₀ in M).

Data collection

ER and AR data were collected from multiple sources: (i) already-curated datasets from the CERAPP (Collaborative Estrogen Receptor Activity Prediction Project; ER data) [16] and CoMPARA (Collaborative Modeling Project for Androgen Receptor Activity; AR data) [20] international workgroups; (ii) data downloaded from the “Tox21 Data Challenge 2014” website (“Tox-DC dataset”; both ER and AR data); (iii) the Endocrine Disruptor Knowledge Base (EDKB; both ER and AR data) database [32]; (iv) PubChem (gene ID searching 2099 for ER α and 367 for AR); (v) ChEMBL (IDs 240 for ER α and 1871 for AR); (vi) BindingDB (both ER and AR data) [33]; and (vii) the “Receptor Mediated Effects” database, extracted from the QSAR Toolbox (both ER and AR data) [34].

Data prepared during CERAPP and CoMPARA programs followed an analogous procedure. The training sets were constituted from several assays from the ToxCast and Tox21 programs (18 and 11 for ER and AR, respectively), run on a library of 1855 ToxCast chemicals [16,20], and assigning the binding activity (as described in the “Description of the modelled properties” paragraph). This same library used to constitute the training set was also screened in the context of the multi-agency Tox21 program. CERAPP / CoMPARA test sets were assembled, mostly, from published literature and databases (e.g. ChEMBL) sources. Both training and test sets were subject to a data curation procedure, as described in the respective publications.

The Tox-DC ER/AR datasets originate from a data challenge in 2014 sought to “crowdsource” predictive models from various researchers across the globe. As these datasets come from the

Tox21 program, there is an overlap between CERAPP / CoMPARA and the Tox-DC. A qualitative representation about the relationship of considered data sources is represented by Figure 1.

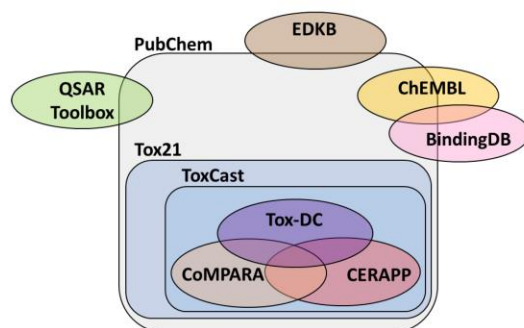


Figure 1. Qualitative representation showing overlap of compounds among the different data sources.

The databases do not report all properties: for instance, BindingDB reports only the IC₅₀ values, ChEMBL the IC₅₀ and RBA, while PubChem reports the binding activity and IC₅₀. Table 2 reports summary statistics of the cleaned datasets (i.e. after the data curation procedure described below).

Raw data processing and chemical structures standardization were carried out with a standardization workflow implemented in KNIME [35]. In case of duplicates, only one compound was kept, and its property was computed either as the median (IC₅₀ and RBA) or the mode (binding activity) of the values. For the latter, when the repartition of B / nB was close to the random threshold (i.e. between 40 - 60 %), the entry was discarded.

Consistency of experimental labels

The following analysis have been performed to better describe the available data. For the binding activity (for both ER and AR), multiple pairwise comparisons among the different sources of data (i.e. CERAPP / CoMPARA, Tox-DC and PubChem) were carried out to verify the consistency of reported experimental labels. Overlapping compounds were paired either by their unique ID number (e.g. the “CERAPP_ID”) when available, or by their chemical structure after the data

curation procedure. Finally, a given compound can be recorded simultaneously in 4 sources for ER and 3 sources for AR, with different labels or no label.

The degree of agreement is the average number of sources that agree on the label of a compound. For four ER, the levels are 1 (all sources agree), 0.75 (3 sources agree) and 0.5 (2 sources agree). The number of compounds in each level value are noted N_{100} , N_{75} and N_{50} . For AR, the levels are 1 and 0.66 and the corresponding populations are N_{100} and N_{66} . N_c is the total number of overlapping compounds. Using these notations, the degree of agreement for ER and AR are summarized in the Equations 2 and 3.

$$Agreement_{ER} = \frac{N_{100} * 1 + N_{75} * 0.75 + N_{50} * 0.5}{N_c} \quad (2)$$

$$Agreement_{AR} = \frac{N_{100} * 1 + N_{66} * 0.66}{N_c} \quad (3)$$

Multiple IC50 and RBA values reported for the same compounds were used to estimate the overall experimental measurement error for the given property, expressed through the Median Absolute Deviation (MAD) value (Equation 4). Where X_j^i is the j^{th} value in the set of N_i measures available for compound i , N_c is the total number of compounds with the given property and $\langle X^i \rangle$ is the average of values for the property of compound i .

$$MAD = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{1}{N_i} \sum_{j=1}^{N_i} |X_j^i - \langle X^i \rangle| \quad (4)$$

Molecular descriptors

ISIDA Property-Label Molecular descriptors [36] were employed. A total of 71 ISIDA descriptor spaces were generated, corresponding to molecular fragment of different sizes, topologies and “coloration” (elements labels, physical properties mapped on the atoms explicit or implicit chemical bonds, atom pairs). The number of fragments varies accordingly to the selected fragmentation scheme of the given descriptor space. It ranged from 242 (“IIAB(2-2)”, sequences of atom of length 2) for the AR RBA model (the smallest training set) to 9013 (“IIAB(2-4)”, atom centred fragments with radius up to 4) for the ER IC50 model (the largest training set), with an average of 4383 (SI, section 2).

Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a dimensionality reduction method which allows to visualize the data distribution on 2-dimensional map. A data property can be added as a 3rd axis forming such called activity landscape [37,38].

GTM was employed to graphically visualize the quantitative results obtained during the experimental labels “agreement-analysis”. We used the binding activity (B / nB labels) as function characterizing the landscapes. In such a way it was possible to identify clusters of compounds with discordant experimental values. Through multiple pairwise evaluations, the chemical space of CERAPP / CoMPARA was compared against Tox-DC and PubChem, following the following steps: (i) for the given pair of datasets, overlapping compounds were extracted; (ii) two landscapes were generated, the former colored according to the first dataset’s molecules binding activity, the latter according to the second one; (iii) graphical differences between the generated landscapes were qualitatively highlighted. These latter areas allow the identification of groups of compounds with discordant labels between the two datasets.

The ISIDA descriptor space “IAB(2-4)” [36], associated to the best Support Vector Machine (SVM) model (in terms of BA), was chosen. These descriptors are based on molecular fragments consisting in sequences of atoms and bonds up to a length four. The so-called “manifold” (which could be seen as a two-dimensional “rubber sheet” injected into the D-dimensional descriptor space) [37] was built on the CERAPP dataset. The GTM model was optimized for discriminating binders from non-binders according the CERAPP labels. Most of the CoMPARA training set chemical structures are actually the same as the CERAPP ones.

Genetic algorithm [39] was used for optimizing the characteristic parameters of the GTM: the number of Radial Basis Function centers ($m = 12$), the Radial Basis Functions width ($w = 1.8$) and the number of grid points, i.e. the dimension of the map ($k = 30$). The GA fitness function was set to the balance accuracy on the binding activity reported in CERAPP dataset.

Model generation and validation

Figure 2 depicts the modelling workflow. Support Vector Machine (SVM) with linear and Radial Basis Function kernels, Random Forest (RF) and Naïve Bayesian (NB) machine learning approaches were implemented. SVM models were generated with libSVM (v. 3.22) [40]; WEKA (v. 3.9.3) [41] was used for RF and for NB models. For a given training set (Table 2), the “best” ISIDA descriptor spaces and the optimal SVM hyperparameters (cost and gamma) have been selected in genetic algorithm driven optimization process. In such a way, the top 15 descriptor

spaces were retained (as a compromise between performance and computational speed) and used to fit an equal number of optimized SVM, RF and NB “individual models”. For RF and NB, default WEKA settings were selected. Finally, only the best performing individual model corresponding to a given descriptor space was retained. Internal validation of each individual model was carried out by 5-fold CV repeated ten times (10*5CV) after data reshuffling. Statistics were assessed for each repetition followed by their averaging. The 15 selected individual models were then ensembled in consensus. Tables S1(a-f) report detailed information concerning consensus models set-up (employed descriptor space and algorithm for the given individual models) for each endpoint.

The abovementioned procedure was followed for all the generated models: ER/AR binding affinity, RBA and IC50. Classification and regression models’ performance was evaluated through the analysis of the Sensitivity (Sn), Specificity (Sp), Balanced Accuracy (BA) and determination coefficient (R^2) and root mean squared error (RMSE), respectively (Table S2).

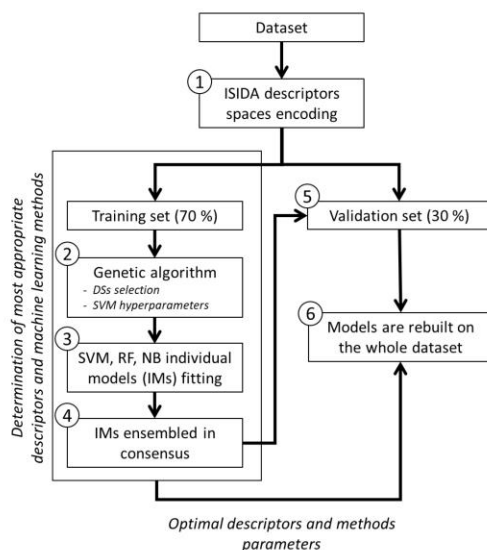


Figure 2. Model generation workflow. (1) ISIDA descriptor spaces are generated; (2) best descriptor spaces and SVM hyperparameters are chosen by genetic algorithm; (3) SVM, RF and NB individual models (IM) are fitted, and the best IM per descriptor spaces is retained; (4) retained models are ensembled in consensus and (5) externally validated on the test set; (6)

selected descriptor spaces and methods parameters are used to train the “final” models, i.e. on the whole dataset.

Training and test set splitting

Training / test set definition modality varied according to the modelled property.

- For the ER / AR binding activity, we kept the same CERAPP / CoMPARA splitting, as the goal was to reproduce models generated within the workgroups. In addition, these two models were externally validated on Tox-DC and PubChem data.
- For the IC50, only data coming from BindingDB was chosen to constitute the training set; instead, compounds coming from the remaining databases (i.e. ChEMBL and PubChem) were merged to constitute the test set. We opted for this choice under the hypothesis that the former database comprised higher-quality data, as data is manually curated before being imported [33].
- For the RBA we employed a stratified (based on the property range) random splitting, 70 % training and 30 % test set. As we had no other information to evaluate the quality of the data sources, we implemented a random splitting.

Applicability domain and ensemble modelling

The applicability domain (AD) was evaluated based on the so-called “fragment control” assessment [36]: if the test molecule has one fragment (i.e. a determined sequence of atoms and/or bonds) not present in the model’s training set, that molecule is considered out-of-AD. Generated models were ensemble in consensus to increase the overall quality of the prediction. The consensus outcome is provided either by computing the median (continuous model) or by a majority voting (classification models). In these calculations, all out-of-AD predictions are excluded. In addition, we propose a 4-grade reliability scale system based on the % of models with positive AD outcome, as described in our previous works [42,43].

Results

Overview of the curated datasets

Tables 2 and 3 report general statistics of the curated (i.e. at the end of the data curation procedure) datasets.

The percentage of binders greatly varies depending on the dataset: it ranges from 4 % for PubChem ER and Tox-DC AR, to 25 – 27 % for CERAPP test set and PubChem AR. The CoMPARA test set has an amount of binders much more similar to its training set (11 vs. 12 %), as opposed to CERAPP (25 vs. 14 %). The average agreement among the different sources (Equation 2 and 3) is 78 and 95 % for ER and AR, respectively.

On average, RBA and IC50 span over a range of 6 – 7 log unit. The noticeably higher MAD value for ER RBA is caused by the presence of several compounds with significantly different reported values. For instance, the range of values for 4'-Hydroxypropiofenone (CAS 70-70-2) is 1.66 log unit.

Table 2. Curated datasets for model generation and validation.

Receptor	Property	Data agreement ^a	Training set			Test set		
			#	Property values ^b	% B ^c	#	Property values ^c	% B
ER	Binding	78 %	1677	237 / 1440	14	5795 ^d	1458 / 4337	25
	IC50	0.78	1982	3.2 – 9.9	7	656	4.13 – 9.52	6
	RBA	1.12	1398	-4.56 – 2.98	10	600	-5.12 – 2.63	9
AR	Binding	95 %	1662	198 / 1464	12	3882 ^d	428 / 3323	11
	IC50	0.68	1622	3.34 – 9.70	13	355	3.38 – 9.15	9
	RBA	0.66	233	-3.54 – 2.27	11	101	-3.3 – 2.05	12

^afor the binding activity, value refers the average labels agreement (%; Equation 2 and 3); whereas for continuous properties, to the MAD (Equation 4, log units). ^brepartition between B / nB (binding property) or min-max experimental value (pIC50 for IC50 and logRBA for RBA); ^cpercentage of B labels, assigned according to the threshold defined in the methods section for IC50 and RBA; ^dCERAPP / CoMPARA original test set, excluding training set's overlapping compounds.

Table 3. Additional datasets for the binding activity property.

Receptor	Database	#	B / nB	% B
ER	Tox-DC	5408	527 / 4881	10
	PubChem	95496	3750 / 91746	4
AR	Tox-DC	5549	200 / 5349	4
	PubChem	8577	2304 / 6273	27

Binding activity experimental label consistency

Figure 3 represents ER / AR comparison matrices for the different datasets. The matrix's upper part (orange background) reports the recall values for the binders (i.e. sensitivity), while the bottom part (blue background) the recall for the non-binders (i.e. specificity). We found that the CERAPP train and test sets had a certain degree of overlapping compounds, as they were left on purpose during in the context of the workgroup [16]; however, this was not the case for the CoMPARA [20].

It can be noticed that already at the level of training vs. test set (CERAPP), the degree of inconsistencies is high, especially for the binders, with an agreement of only 68 %. This agreement decreases significantly when considering the comparison against Tox-DC (49 % and 16 % for ER and AR) and PubChem (27 % and 37 % for ER and AR). On the other hand, the agreement of the inactive class is almost perfect (>96 %) in all instances; the only exception is the CERAPP training vs. test set, with only 85 %.



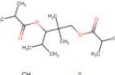


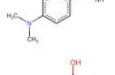
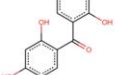

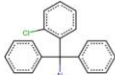
Estrogen receptor	CERAPP train				Androgen receptor	CoMPARA train			
	CERAPP train	CERAPP test	Tox-DC	PubChem		CoMPARA train	CoMPARA test	Tox-DC	PubChem
CERAPP train	-	0.68	0.49	0.27	CoMPARA train	-	na	0.16	0.37
CERAPP test	0.85	-	0.34	0.20	CoMPARA test	na	-	0.40	0.61
Tox-DC	0.96	0.98	-	0.31	Tox-DC	0.99	1.00	-	0.66
PubChem	0.97	1.00	0.99	-	PubChem	0.98	0.98	0.96	-

Figure 3. Experimental labels comparison matrices. For each pairwise comparison, the agreement (recall) of the binder (orange background) and non-binder class (blue background) for the given dataset is reported. na = not available due to absence of overlapping compounds.

Table 4 reports some compounds with discordant binding activities. One reason of such discrepancies is the combination of assay types which measure different endpoints: for instance, molecule CAS 133-06-2 in the CERAPP test set has an assigned B label based on five measurements (two logRBA, one AC50, one relative affinity and one unspecified data values) while PubChem reports a total of 36 measurements (five B, 17 nB and 15 “inconclusive”); another example is CAS 92-68-2, which has a B label in CERAPP test set based on seven measurements

(two logRBA, two relative affinities, two cell proliferation and one unspecified data values) while PubChem reports 13 measurements (one B, 21 nB, and 11 “inconclusive”).

Table 4. Examples compounds with discordant binding activities.

Receptor	Molecule	CAS	Molecule ID number ^a	Assigned experimental label ^b			
			<i>Ctr/tst^c</i> <i>Tox-DC</i> <i>PubChem</i>	<i>Ctr^c</i>	<i>Cts^c</i>	<i>Tox-DC</i>	<i>PubChem</i>
ER		133-06-2	1291 1385 8607	nB	B	nB	nB
		2425-06-1	16612 242 17038	nB	B	B	nB
		6846-50-0	18438 7635 23284	nB	B	B	nB
		137-26-8	13006 1332 5455	nB	B	nB	nB
		92-68-2	11297 20721 41690	B	nB	nB	nB
AR		2465-27-2	1 20114 1717	B	-	nB	nB
		131-55-5	108 21306 8571	B	-	nB	B
		20830-75-5	1493 2934 2724385	nB	-	B	B
		23593-75-1	1566 9871 2812	nB	-	nB	B

^athe molecule’s ID of the given dataset is provided for the given dataset; ^bthe assigned binding activity experimental label. ^c*Ctr/tst* = CERAPP/CoMPARA training/test set set.

GTM was employed to graphically visualize the results illustrated by Figure 3. Figure 4 and 5 report binding activity landscapes comparison between the considered datasets, for ER and AR respectively. For each row, the CERAPP / CoMPARA training set is compared against: (1) CERAPP test set (not possible for CoMPARA); (2) Tox-DC and (3) PubChem, for both ER and AR. Red areas are mainly populated by binders, while blue ones by the non-binders. Intermediate colors are mixed populated areas, i.e. chemical space zone occupied by the same compounds with different experimental labels. The “difference” map of the third column highlights in black discordant areas.

The output of GTM reflect the differences which were quantitatively reported in the previous paragraph. In particular, inter-datasets agreement for ER is significantly higher than for AR (27 – 68 % vs. 16 – 37 %; Figure 3), as highlighted by the fact that “differences” AR maps have much more black spots (Figure 5).

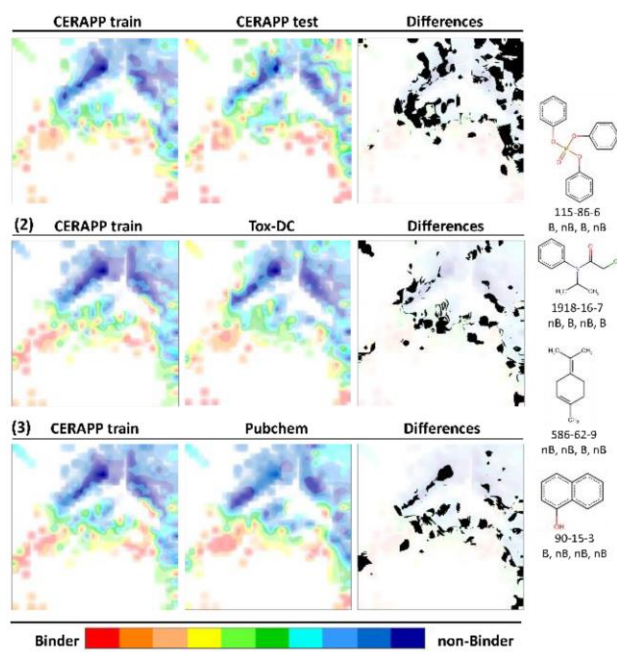


Figure 4. GTM comparison of the different sources of data for ER. The CERAPP training set is compared against the CERAPP test set, Tox-DC and PubChem. The “Differences” maps were generated in order to highlight the areas of the map that present the highest disagreement.

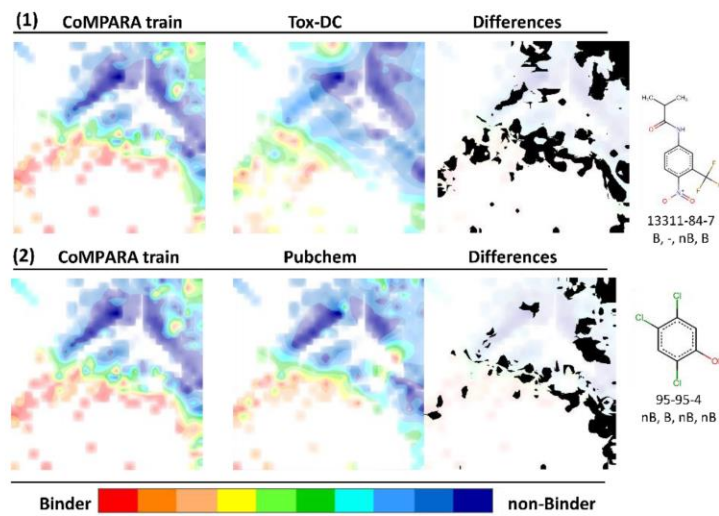


Figure 5. GTM comparison of the different sources of data for AR. The CoMPARA training set is compared against the Tox-DC and PubChem.

Model performances

Table 5 reports models performances for internal and external validation. Classification models performed similarly to those reported within the two workgroups (see Tables S3(a-b) and Discussion section), with BA_{CV} 0.68 for ER and 0.84 for AR. Modelling of AR produced more promising results than ER, especially for the detection of true binders (whereas the ER binding model scored the lowest sensitivity of 0.34). By following the same CERAPP / CoMPARA approach – i.e. removing test set compounds with less than 2 experimental assays – BA in external validation improved, increasing to 0.63 ($Sn = 0.38$, $Sp = 0.89$) for ER and 0.82 ($Sn = 0.70$, $Sp = 0.95$) for AR. Performances on Tox-DC and PubChem (Table 6; $BA = 0.65 - 0.72$) are similar to those on CERAPP / CoMPARA test sets (Table 5). Again, the model's ability to correctly predict binders is rather low, with sensitivity values ranging from 0.40 to 0.55.

Concerning regression_models IC50 models ($R^2_{AR} = 0.64 - R^2_{ER} = 0.77$) were slightly better than RBA ones ($R^2_{AR} = 0.57 - R^2_{ER} = 0.71$). Low performance ($R^2 = 0.43$) of the ER IC50 model has been caused by the presence of some outliers in the external set (orange rectangles A and B in Figure 6), analyzed in the discussion section. With the exclusion of these suspicious data points

(see Discussion section), ER IC50 performances raised to similar values obtained by cross-validation ($R^2 = 0.60$, RMSE = 0.66).

Table 5. Model performances.

Classification	Receptor	Property	5-fold CV		External validation			
			BA		BA	Sn	Sp	Data coverage ^b
	ER	Binding	0.68		0.60	0.34	0.87	76 % (4409 / 5795)
	AR	Binding	0.84		0.72	0.49	0.95	85 % (3320 / 3882)

Regression	Receptor	Property	5-fold CV		External validation		
			R^2	RMSE	R^2	RMSE	Data coverage ^b
	ER	RBA	0.71	0.84	0.76	0.76	93 % (563 / 600)
		IC50	0.77	0.67	0.43	0.93	78 % (517 / 656)
	AR	RBA	0.57	0.76	0.60	0.67	70 % (71 / 101)
		IC50	0.64	0.66	0.44	0.68	85 % (266 / 310)

Classification (upper part) and regression models (bottom part). BA = balanced accuracy, Sn = sensitivity, Sp = specificity. ^aIn brackets, the standard deviation computed in the 5-fold CV is reported for the given metric values averaged over the number of repetitions; ^bratio of the number of compounds inside AD and the total number of compounds.

Table 6. Additional validation for classification models.

Receptor	Tox-DC				PubChem			
	BA	Sn	Sp	Data coverage ^b	BA	Sn	Sp	Data coverage ^b
ER	0.65	0.44	0.85	80 % (4374 / 5408)	0.71	0.47	0.98	76 % (72829 / 95496)
AR	0.72	0.55	0.92	81 % (4485 / 5549)	0.66	0.40	0.92	82 % (7100 / 8577)

Figure 6 depicts the scatterplots (experimental vs. predicted values) for the external validation set of the four regression models. Molecules highlighted by orange rectangles were marked as suspicious outliers and are analyzed with more details in the Discussion section.

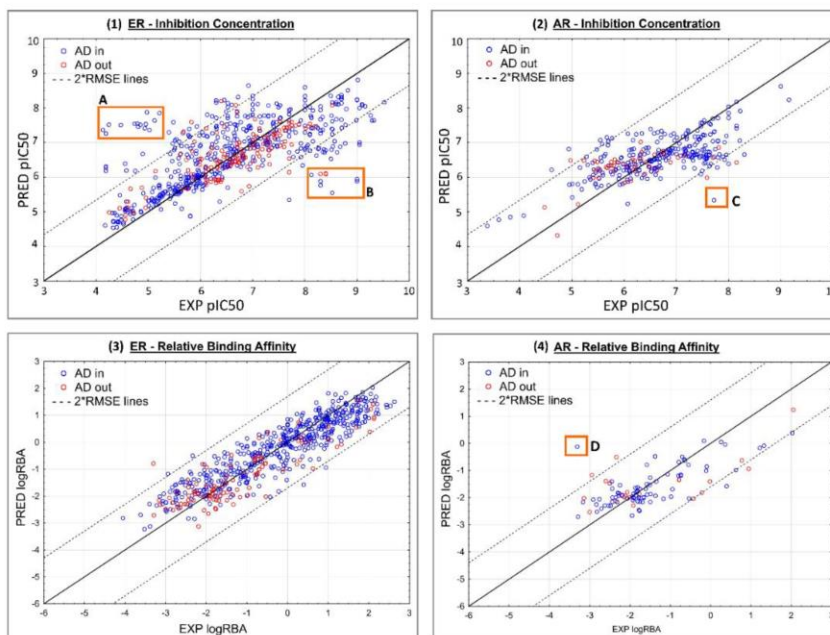


Figure 6. Experimental vs. Predicted values scatterplots. In order from upper left corner to bottom right corner: (1) ER IC₅₀; (2) AR IC₅₀; (3) ER RBA; (4) ER IC₅₀. Potential outliers were marked by the orange rectangles “A”, “B”, “C” and “D” (see Discussion section).

Discussion

Binding activity classification models

Our modelling results on the binding affinity property (Table 5) are similar to those published within the CERAPP and CoMPARA frameworks. Table S3 in SI reports the performance for each individual group: the average BAs of the generated models were 0.59 for ER ($S_n = 0.35$; $S_p = 0.82$) for ER and 0.73 ($S_n = 0.56$; $S_p = 0.88$) for AR. The relatively low performances in external validation for the CERAPP and CoMPARA models can be attributed to the different nature of the validation data sets. Contrary to the training sets that are homogeneous collections of compounds coming from the same source and processed in the same way, the validation data sets are a collection of data from multiple literature sources and databases, whose labels are adapted to be close (but not identical) to the training set data processing [16,20]. For example, experimental

inhibitory values of the validation data sets were considered to be approximatively equivalent to the endocrine disruption score of the training set, i.e. the AC50 derived from the computed AUC [16,25]. Ideally, these models should be validated with the generation of new data in an analogue approach followed by the US EPA's "ToxCast" and "Toxicology Testing in 21st Century" [24–28] programs, from where the data was originally generated. Compounds should be tested on all the individual assays and the AC50 computed to obtain really comparable labels.

IC50 and RBA regression models

As showed by Figure 6, several suspicious outliers were identified. We do not exclude the possibility that some additional points may need revision. After carefully reviewing the sources of some compounds, we noticed that the cause was a wrongly reported experimental value. Indeed, values were not always referring to the displacement of labelled radiolabeled estradiol or R1881, instead, they resulted from other kind of testing protocols. To provide an example, the clusters of points marked by the orange rectangles A and B (Figure 6, ER IC50 graph) belong, respectively, to the chemical families of benzoxepine and raloxifene. Both these groups of chemicals come from two unique studies [44,45], where the authors synthesized differently substituted version of the base scaffold. However, the tested property was not the inhibition of ER by displacement of labelled estradiol, but the inhibition of MCF7 tumor breasts cells proliferation. Furthermore, both families have very close structural analogues in the training set, but the average difference between their experimental pIC50 values was about 3 and 2 log units for benzoxepine and raloxifene compounds, respectively. For instance, the outlier-benzoxepine ChemBL188193 (pIC50 = 4.25) is closely similar to ChemBL513397 (pIC50 = 8.06) and ChemBL470993 (pIC50 = 8.72) as illustrated on Figure 7. Therefore, all external set compounds for which the experimental IC50 was determined based on a cell proliferation assay were filtered out. With this filter, 138 compounds were removed from the initial 656, and the performances improved significantly: R² increase from 0.43 to 0.60 and RMSE decreased from 0.93 to 0.66 (Table 5).

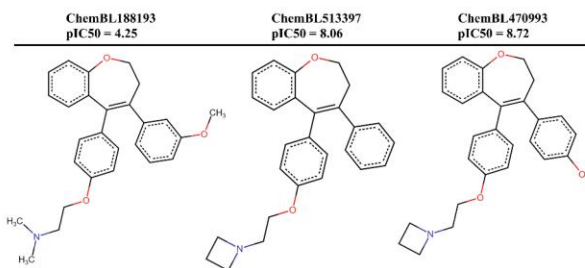


Figure 7. Closely similar structures with different ER activity value.

For AR IC50, the only extreme point is 17 α -estradiol (CAS 5864-38-0; orange rectangle in Figure 6; point C). The compound experimental pIC50 is reported at 7.8 and predicted at 5.4, an error of 2.4 log unit (SI table S4(a), PubChem CID 450). Estradiol is absent of the training set which contains close analogs which potency ranges from 3.36 (estriol) to 7.05 (3-deoxyestradiol). Actually, estradiol has been identified previously as an activity cliff generator and our result confirm this analysis [46].

For AR RBA, the compound in the frame D is androstane (CAS [24887-75-0](#)) an hydrocarbon steroid scaffold. It has an experimental logRBA of -3.3 and is predicted at -0.09, resulting in a significant error of 3.2 log unit (SI table S4(b), PubChem CID 123412). This experimental value is originating from Fang et al. [29], reporting a very large IC50 of 6.35E-4 Mol/L which could be above the aqueous solubility limit of the compound (4.8E-8 Mol/L, estimated value by PubChem). This prediction should not have been accepted, and this is highlighting a limitation of the applicability domain of our models: since other compound in the training set did share this hydrocarbon scaffold, the scaffold itself was considered as legitimate.

RBA measurements have low repeatability. This observation has been already addressed, for instance, by Fang et al. [29] (figure 1 of their work) in comparison to Waller et al [47]. When we applied our models to steroidal androgens, one of the largest errors was on testosterone. Actually, testosterone is not part of our dataset because of the large variability of reported RBA values. For instance, the RBA of testosterone is reported 12 times in ChEMBL, with 9 different values. Investigating the associated articles [29,47–49], the RBA (calculated using R1881 as reference compound) is about 6 %. Values are varying because of different reference ligand (for instance 37 % is reported using DHT as reference [48]); or because the reported value is not in the correct

column (for instance the value 417 % reported from [49] refer to sex steroid binding protein and not AR). These are some reasons explaining why RBA values are difficult to validate.

Overall, there is a lot of uncertainty associated to ED data: for instance, a search through PubChem for ER α binding activity reveals the existence of more than 300.000 compounds tested on roughly 1200 different types of assays (gene ID 2099). Furthermore, the way of reporting ED activity is multiple and not always straightforward, such as: inhibition concentration, potency, relative binding affinity, K_i , efficacy, EC50 and the binary label binder/non-binder (which may be automatically assigned by the databases based on different thresholds). All these differences around ED-related data were also promoted by the absence of a standardized ED screening guideline [6], which was published only recently (2018) by regulatory agencies.

Our developed models follow the OECD principles [50]: the endpoints (binding activity, IC50 and RBA) are well defined. Goodness-of-fit, robustness and predictivity were evaluated in cross-validation and external validation [51]. The AD of the models was defined using a fragment control assessment [35] together with a reliability scoring function [42].

Our models are available through the online ISIDA/Predictor platform [23], available at the Laboratory of Chemoinformatics webpage: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi.

Conclusions

In this work we presented a throughout analysis concerning endocrine disruption, focusing on the estrogen and androgen receptors. We considered the following three properties: binding activity, median Inhibitory Concentration (IC50) and Relative Binding Affinity (RBA).

Classification binding activity (B/nB) and quantitative IC50 and RBA datasets we collected from multiple sources, including the CERPAP/CoMPARA collaborations, the Tox21 data challenge (Tox-DC) and from PubChem. We noticed that datasets suffered from a low concordance between experimental values. For the B class, the average concordance was only 42 %, ranging from 16 % (CoMPARA training set vs. Tox-DC data) to 68 % (CERAPP training set vs. CERAPP test set). For IC50 and RBA measurements, the average data variability ranged from 0.73 up to 1.88 log units for AR_{RBA} and ER_{IC50}, respectively.

To detect clusters of compounds with such discordant experimental values, the Generative Topographic Mapping approach was employed to compare the chemical space of the collected datasets.

Generated models illustrated how these discrepancies have negatively impacted performances. Binding activity models reproduced the CERAPP/CoMPARA frameworks, showing comparable results (ER $BA_{ext} = 0.60$ and AR $BA_{ext} = 0.73$). Additional external validation was carried out on additional Tox-DC and the PubChem datasets ($BA_{Tox-DC} = 0.65 - 0.72$ and $BA_{PubChem} = 0.66 - 0.71$). The models' limited ability to detect reliably binding compounds reflect the low inter-databases concordance, with sensitivity values ranging from 0.34 to 0.49. In particular, the merging of assays with different biological meanings was responsible of such mediocre performances, indicating that CERAPP/CoMPARA datasets should not be fused with other data sources when building predictive models.

Regression RBA models were more performant than their IC50 counterparts: for the former, R^2_{ext} values ranged from 0.60 – 0.76, while for the latter from 0.43 – 0.44. Low R^2 values for IC50 were caused by the presence of several identified outliers, either due to wrongly reported information or misinterpreted experimental endpoints. This reflects the same issues encountered during the binding activity modelling, due to the merging of experimental data measurements of different meaning. The generation of standardized data sets from validated assays will allow the generation of more performant models, capable of predicting specific properties to complex and multifactorial mode of action like endocrine disruption. On the other hand, it is possible to exploit larger datasets based on a binary classification “binding/non-binding”. Of course, many details are lost in the process and these datasets maybe biased when defining the classes from the interpretation of experimental results originating from different assays.

Generated models have good predicting power when detecting non-binders. In-depth analysis of the results demonstrate that a probable cause of low accuracy for detecting binders is the merging of different experimental results from different assays.

All collected data is freely available on Zenodo ([10.5281/zenodo.3935808](https://doi.org/10.5281/zenodo.3935808)), counting a total of 6215 (ER) and 3789 (AR) unique compounds listing at least one experimental. Moreover, the Tox-DC and PubChem datasets, counting more than 100.000 compounds with binding activity label, are provided as well. This is, to the best of our knowledge, the biggest collection of ED-relevant data published so far.

Bibliography

- [1] H. Liu, X. Yang and R. Lu, *Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin*, Chemosphere 156 (2016), pp. 1–7.
- [2] E. Furusjö, A. Allard, S. Nilsson, M. Rahmberg and A. Svenson, *State of the art assessment of endocrine disruptors*, Tech. Rep. 070307/2009/550687/SER/D3, European Commission, DG Environment, Brussels, BE, 2012
- [3] N. Andersson, M. Arena, D. Auteri, S. Barmaz, E. Grignard, A. Kienzler, P. Lepper, A.M. Lostia, S. Munn, J.M. Parra Morte, F. Pellizzato, J. Tarazona, A. Terron and S. Van der Linden, *Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009*, EFSA J. 16 (2018), pp. 1–135.
- [4] E. Lo Piparo and A. Worth, *Review of QSAR Models and Software Tools for Predicting Developmental and Reproductive Toxicity*, Tech. Rep. EUR 24522 EN, European Commission JRC, Ispra, IT, 2010.
- [5] *ECHA website*. European Chemicals Agency, Helsinki, FI, 2020. Available at: <https://echa.europa.eu/>.
- [6] OECD, *Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption*, Tech. Rep. 978-92-64-30474-1, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2018.
- [7] S. Adler, D. Basketter, S. Creton, O. Pelkonen, J. Van Benthem, V. Zuang et al., *Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010*, Arch. Toxicol. 85 (2011), pp. 367–485.
- [8] *OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects*, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2019. Available at <http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm>.
- [9] M.O. Taha, M. Tarairah, H. Zalloum and G. Abu-Sheikha, *Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and in silico search for new hits*, J. Mol. Graph. Model. 28 (2010), pp. 383–400.

- [10] L.B. Salum, I. Polikarpov and A.D. Andricopulo, *Structure-based approach for the study of estrogen receptor binding affinity and subtype selectivity*, J. Chem. Inf. Model. 48 (2008), pp. 2243–2253.
- [11] T.I. Netzeva, A.G. Saliner and A.P. Worth, *Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory*, Environ. Toxicol. Chem. 25 (2006), pp. 1223–1230.
- [12] J. He, T. Peng, X. Yang and H. Liu, *Development of QSAR models for predicting the binding affinity of endocrine disrupting chemicals to eight fish estrogen receptor*, Ecotoxicol. Environ. Saf. 148 (2018), pp. 211–219.
- [13] H. Liu, E. Papa and P. Gramatica, *QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of oecd principles*, Chem. Res. Toxicol. 19 (2006), pp. 1540–1548.
- [14] L. Zhang, A. Sedykh, A. Tripathi, H. Zhu, A. Afantitis, V.D. Mouchlis, G. Melagraki, I. Rusyn and A. Tropsha, *Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure based virtual screening approaches*, Toxicol. Appl. Pharmacol. 272 (2013), pp. 67–76.
- [15] J. Li and P. Gramatica, *QSAR classification of estrogen receptor binders and pre-screening of potential pleiotropic edcs*, SAR QSAR Environ. Res. 21 (2010), pp. 657–669.
- [16] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A. Richard, C. 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, and R. Judson, *CERAPP: collaborative estrogen receptor activity prediction project*, Environ. Health Perspect. 124 (2016), pp. 1023–1033.
- [17] M. Todorov, E. Mombelli, S. Aït-Aïssa and O. Mekenyan, *Androgen receptor binding affinity: a QSAR evaluation*, SAR QSAR Environ. Res. 22 (2011), pp. 265–291.
- [18] A.M. Vinggaard, J. Niemelä, E.B. Wedebye and G.E. Jensen, *Screening of 397 chemicals*

and development of a quantitative structure-activity relationship model for androgen receptor antagonism, *Chem. Res. Toxicol.* 21 (2008), pp. 813–823.

- [19] G.E. Jensen, J.R. Niemelä, E.B. Wedebye and N.G. Nikolov, *QSAR models for reproductive toxicity and endocrine disruption in regulatory use - a preliminary investigation*, *SAR QSAR Environ. Res.* 19 (2008), pp. 631–641.
- [20] K. Mansouri, N. Kleinstreuer, E. Watt, J. Harris and R. Judson, *Compara: collaborative modeling project for androgen receptor activity*, *Environ. Health Perspect.* 124 (2020), pp. 250 - 261.
- [21] NIH, Data from: *Tox21 data challenge 2014*, National Institutes of Health, dataset available at: <https://tripod.nih.gov/tox21>.
- [22] NIH, *PubChem*, National Library of Medicine, National Center for Biotechnology Information, Bethesda, Maryland, 2019; available at <https://pubchem.ncbi.nlm.nih.gov/>.
- [23] G. Marcou, D. Horvath, F. Bonachera, and A. Varnek, *Laboratoire De Chimoinformatique UMR 7140 CNRS*, University of Strasbourg, Strasbourg, FR, 2019; available at <http://infochim.u-strasbg.fr/>
- [24] EPA, Data from: *Endocrine disruptor screening program (EDSP) in the 21st century.*, Environmental Protection Agency, dataset available at: <https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-edsp-21st-century>.
- [25] R.S. Judson, F.M. Magpantay, V. Chickarmane, C. Haskell, N. Tania, J. Taylor et al., *Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor*, *Toxicol. Sci.* 148 (2015), pp. 137–154.
- [26] D.J. Dix, R.J. Kavlock, R.W. Setzer, K.A. Houck, A.M. Richard and M.T. Martin, *The toxcast program for prioritizing toxicity testing of environmental chemicals*, *Toxicol. Sci.* 95 (2006), pp. 5–12.
- [27] R.R. Tice, C.P. Austin, R.J. Kavlock and J.R. Bucher, *Improving the human hazard characterization of chemicals: A Tox21 update*, *Environ Health Perspect* 7 (2013), pp. 756–765.
- [28] R. Kavlock, K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla, D. Reif, A. Richard, D. Rotroff, N. Sipes and D. Dix *Update on*

- EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management*, Chem Res Toxicol 7 (2012), pp. 1287-1302.
- [29] H. Fang, W. Tong, W.S. Branham, C.L. Moland, S.L. Dial, H. Hong, Q. Xie, R. Perkins, W. Owens and D.M. Sheehan, *Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor*, Chem. Res. Toxicol. 16 (2003), pp. 1338–1358.
- [30] C.M. Olsen, E.T.M. Meussen-Elholm, J.K. Hongslo, J. Stenersen and K.E. Tollefsen, *Estrogenic effects of environmental chemicals: an interspecies comparison*, Comp. Biochem. Physiol. - C Toxicol. Pharmacol. 141 (2005), pp. 267–274.
- [31] A. Rybacka, C. Rudén, I. V. Tetko and P.L. Andersson, *Identifying potential endocrine disruptors among industrial chemicals and their metabolites - development and evaluation of in silico tools*, Chemosphere 139 (2015), pp. 372–378.
- [32] D. Ding, L. Xu, H. Fang, H. Hong, R. Perkins, S. Harris, E.D. Bearden, L. Shi and W. Tong, *The EDKB: An Established Knowledge Base for Endocrine Disrupting Chemicals*, BMC Bioinformatics 11 (2010), pp. 1471–1478.
- [33] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology*, Nucleic Acids Res. 44 (2016), pp. 1045–1053.
- [34] *QSAR Toolbox v 4.1*, OASIS Laboratory of mathematical chemistry, Burgas, BG, 2017; software available at <http://oasis-lmc.org/products/software/toolbox.aspx>.
- [35] M. Berthold, N. Cebon, 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.
- [36] F. Ruggiu, G. Marcou, A. Varnek and D. Horvath, *ISIDA property-labelled fragment descriptors*, Mol. Inform. 29 (2010), pp. 855–868.
- [37] C.M. Bishop, M. Svensén, C.K.I. Williams and M. Svens, *The generative topographic mapping*, Neural Comput. 10 (1998), pp. 215–234.
- [38] 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.

- [39] D. Horvath, J. Brown, G. Marcou and A. Varnek, *An evolutionary optimizer of libsvm models*, Challenges 5 (2014), pp. 450–472.
- [40] C. Chih-Chung and L. Chih-Jen, *LIBSVM: a library for support vector machines*, ACM Trans. Intell. Syst. Technol. 2 (2011), pp. 1–27.
- [41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, *The Weka Workbench*, in *Data Mining: Practical Machine Learning Tools And Techniques*, I.H. Witte, eds., Morgan Kaufman Publishers, San Fransisco, 2005, 363-449.
- [42] 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.
- [43] F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert and A. Varnek, *Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context*, SAR QSAR Environ. Res. (2019), .
- [44] D.G. Lloyd, R.B. Hughes, D.M. Zisterer, D.C. Williams, C. Fattorusso, B. Catalanotti, G. Campiani and M.J. Meegan, *Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor*, J. Med. Chem. 47 (2004), pp. 5612–5615.
- [45] T.A. Grese, S. Cho, D.R. Finley, A.G. Godfrey, C.D. Jones, C.W.L. Iii et al., *Modifications to the 2-arylbenzothiophene core of raloxifene*, 2623 (1997), pp. 146–167.
- [46] J.J. Naveja, U. Norinder, D. Mucs, E. López-López and J.L. Medina-Franco, *Chemical space, diversity and activity landscape analysis of estrogen receptor binders*, RSC Adv. 8 (2018), pp. 38229-38237.
- [47] C.L. Waller, B.W. Juma, L. Earl Gray and W.R. Kelce, *Three-dimensional quantitative structure-activity relationships for androgen receptor ligands*, Toxicol. Appl. Pharmacol. 137 (1996), pp. 219-227.
- [48] S. Kamata, T. Matsui, N. Haga, M. Nakamura, K. Odaguchi, T. Itoh, T. Shimizu, T. Suzuki and M. Ishibashi, *Aldosterone antagonists. 2. synthesis and biological activities of 11,12-dehydropregnane derivatives*, J. Med. Chem. 30 (1987), pp. 1647–1658.
- [49] Y.S. Choe, P.J. Lidstroem, D.Y. Chi, T.A. Bonasera, M.J. Welch and J.A. Katzenellenbogen, *Synthesis of 11.beta.-[18f]fluoro-5.alpha.-dihydrotestosterone and 11.beta.-[18f]fluoro-19-nor-5.alpha.-dihydrotestosterone: preparation via*

halofluorination-reduction, receptor binding, and tissue distribution, J. Med. Chem. 38 (1995), pp. 816–825.

- [50] 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.
- [51] 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.

4.2 Part 2 – REACH Chemical space profiling with GTM

Several models have been published in the past years on REACH-relevant endpoints, spanning from simple multiple linear regression models to more complex machine learning methods, such as support vector machine or neural networks. However, we can highlight few drawbacks: (i) being single-task models, the user needs to run multiple models on the same compound, increasing the complexity of the assessment and the required time; (ii) the use of multiple models is not always straightforward due to pre- and post-treatment of data, for instance, for the different output generated, the differences in how the applicability domain (AD) is evaluated; (iii) we found that for some endpoints, the AD of currently-existing tools is rather limited when applied to an industrial context, due to the presence of specific chemotypes poorly represented in typical training sets.

Despite there already exist few tools that can be used for hazardous prioritization, such as Toxmatch or DART (Decision Analysis by Ranking Techniques), they are not able to perform predictions if the experimental value of some key properties is not known: their use is limited to endpoints for which the experimental value is known.

Hence, to overcome these drawbacks, in this work we propose an integrated screening methodology based on Generative Topographic Mapping (GTM). GTM is a probability-based mapping strategy, which can be applied both for large-scale data visualization and property prediction. We chose this method as it has two important advantages : (i) allows multi-task learning, which is quite effective in this context as several properties can be taken into account simultaneously; (ii) produces a graphical output (i.e. a 2D-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; (iii) a base model can be trained and adapted to predict new endpoints that might even not be defined at that time.

To show the potential of GTM, the entire chemical space of the REACH-registered substances was profiled. A total of 11 endpoints have been considered: bioconcentration factor (BCF), ready biodegradability (RB), environmental persistence in sediment (SedP), soil (SoilP) and water (WatP) media, acute aquatic toxicity to algae (AlgaeTox), daphnia (DaphniaTox) and fish (FishTox), rat acute toxicity (RatTox),

androgen and estrogen receptor binding (AR/ER Binding). The determination of this list of properties constitutes what is a “substance's profile”.

A total of 17762 unique compounds listing at least one experimental measurement per property have been collected into a Global dataset. Binary Hazardous (H) and non-Hazardous (nH) labels were assigned according to relevant regulatory REACH-thresholds. Generative Topographic Mapping has been employed as method for data analysis and property prediction, generating single- and multi-task learning models. The chemical spaces of the Global dataset and the REACH-registered substances (ECHA-DB) have been compared. Despite the Global dataset was able to accommodate a large portion of the ECHA-DB chemical space, several areas uniquely populated by ECHA-DB compounds were found, indicating that the Global dataset was lacking important chemotypes. This suggests the applicability domain of currently existing QSARs, which are based on public data, may have some restrictions when applied to the REACH-chemical domain.

Concerning GTM for property prediction, so called Universal Maps trained on the Global dataset have been generated, and their ability to classify H from nH compounds has been tested on all the 11 endpoints. In such a way, a compound can be screened on multiple properties simultaneously. The best Universal map showed acceptable predictive power, ranging from 0.60 to 0.78 balanced accuracy, depending on the endpoint. However, when more universal maps are ensembled in consensus, performances show a general improvement, with BA ranging from 0.67 to 0.83.

This article has been submitted to the peer-reviewed journal “Molecular Informatics”; the manuscript reported here corresponds to the submitted version.



4.2 Partie 2 – Profilage de l'espace chimique REACH avec la GTM

Au cours des dernières années, plusieurs modèles ont été publiés concernant des propriétés pertinentes pour REACH, allant de simples modèles de régression linéaire multiple à des méthodes d'apprentissage automatique plus complexes, telles que des machines à vecteur support ou des réseaux de neurones. Cependant, nous pouvons souligner quelques inconvénients: (i) ces modèles sont mono-tâches, l'utilisateur doit exécuter plusieurs modèles sur le même composé, ce qui augmente la complexité de l'évaluation et le temps

requis; (ii) l'utilisation de plusieurs modèles n'est pas toujours aisée en raison, par exemple, des différents pré- et post-traitements des données, par exemple les différences dans la façon dont le domaine d'applicabilité (AD) est évalué par chaque modèle; (iii) nous avons constaté que pour certaines propriétés, le AD des outils existants est plutôt limité lorsqu'ils sont utilisés dans un contexte industriel car il y a parfois des chemotypes spécifiques mal échantillonnés dans les jeux de données d'entraînement.

Bien qu'il existe déjà quelques outils pouvant être utilisés pour la prioriser les risques liés à des substances chimiques, tels que Toxmatch ou DART (Decision Analysis by Ranking Techniques), ils ne peuvent être utilisés que si certaines valeurs expérimentales sont connues pour certaines propriétés stratégiques : leur utilisation est donc limitée aux composés pour lesquelles ces valeurs sont déjà connues.

Pour surmonter ces limitations, nous proposons dans ce travail une méthodologie de criblage basée sur la cartographie topographique générative (GTM). La GTM est une stratégie de cartographie basée sur un modèle probabiliste des données, qui peut être appliquée à la fois pour la visualisation et la prédiction de propriétés. Nous avons choisi cette méthode car elle présente deux avantages importants: (i) elle permet un apprentissage multi-tâches prenant en charge plusieurs propriétés simultanément; (ii) elle produit une sortie graphique (c'est-à-dire une carte 2D de l'espace chimique) qui peut être utilisée comme support pour mieux comprendre les prédictions du modèle, à la lumière d'une interprétation mécanistique; (iii) un modèle de base peut être entraîné puis adapté pour prédire de nouvelles propriétés qui pourraient n'avoir pas même encore été envisagées initialement.

Pour montrer le potentiel de la GTM, tout l'espace chimique des substances enregistrées pour REACH a été profilé. Au total, 11 paramètres ont été pris en compte: le facteur de bioconcentration, la biodégradabilité primaire, la persistance environnementale dans les sédiments, les sols et les milieux aquatiques, la toxicité aquatique aiguë pour les algues, la daphnie et les poissons, la toxicité aiguë pour le rat, la liaison aux récepteurs nucléaires androgènes et œstrogènes. La liste des valeurs de ces propriétés pour un composé donné définit ce qu'est son «profil de substance».

Un total de 17762 composés uniques répertorient au moins une mesure expérimentale parmi ces propriétés ont été collectés dans un jeu de données « Global ». Les étiquettes binaires, dangereuses (H) et non dangereuses (nH), ont été attribuées en fonction des seuils réglementaires fixés par REACH. La cartographie topographique générative a été utilisée comme méthode d'analyse des données et de prédiction des propriétés, générant des modèles d'apprentissage mono-tâche et multi-tâches. Les espaces chimiques du jeu de données Global et des substances enregistrées pour REACH (ECHA-DB) ont été comparés. Bien que le jeu de données Global partage une grande partie de l'espace chimique avec ECHA-DB, plusieurs zones seulement peuplées de composés ECHA-DB ont été localisées, indiquant qu'il manque des chémotypes importants dans le jeu de données Global. Ceci suggère que le domaine d'applicabilité des modèles QSAR actuels, qui sont basés sur des données publiques, peut souffrir de certaines restrictions lorsqu'ils sont utilisés dans le contexte de REACH.

Concernant les capacités prédictives de la GTM pour les propriétés, des cartes universelles entraînées sur l'ensemble de données Global ont été générées et leur capacité à discriminer les composés H des composés nH a été testée sur les 11 propriétés. De cette manière, un composé peut être criblé sur plusieurs propriétés simultanément. La meilleure carte universelle a montré une capacité prédictive acceptable, allant de 0,60 à 0,78 de précision balancée, selon la propriété. Mais, lorsque plusieurs cartes sont rassemblées dans un consensus, les performances prédictives s'améliorent en général, la précision balancée allant de 0,67 à 0,83.

Cet article a été soumis au journal à comité de lecture « Molecular Informatics » ; le manuscrit ci-dessous correspond à la version soumise.

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. Thanks to this registration procedure initiated in 2007, a large REACH database of well defined (eco)toxicological properties has been created. Considering a high number of chemicals and experimental data registered under the REACH Regulation, their detailed analyses is an important and challenging task. Here, the data distribution in the REACH chemical space was analysed with the help of the Generative Topographic Mapping (GTM) approach. Similarly to geography, GTM allows to generate 2-dimensional maps on which each object (compound) is represented as a data point. The 3d dimension can be used in order to display a distribution of the given (eco)toxicological property (such-called "class landscape"), 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, ecotoxicological and toxicological properties. This map is able to provide with predictions for the above endpoints demonstrating acceptable predictive performance: in cross-validation, balanced accuracy ranges from 0.60 to 0.78. As case study, 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, visualisation

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 tonne/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: firstly, we describe the GTM model built on the Global dataset,

[a] Laboratory of Chemoinformatics, 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

followed its application to profile the REACH-INV compounds. Finally, 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.

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 multistep procedure. This assessment usually starts with a very stringent first-tier assessment, 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

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 explains (partly) its much higher data availability compared to the other endpoints^[3].

2.1.6 Androgen and Estrogen receptor binding (AR/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 oestrogen 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 signalling 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. However, different (more conservative) thresholds were chosen in the following instances: aquatic acute toxicity (10 instead of 1 mg/L); acute rat toxicity (300 instead of 2000 mg/kg b.w.); bioconcentration factor (3 instead of 3.3 log units). This choice was taken in the light of a precautionary approach. 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 datapoints 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	SedHL	120 days	436	253 / 183
Persistence in Soil	SoiHL	120 days	630	111 / 519
Persistence in Water	WatHL	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	ARbinding	-	1661	198 / 1463
Estrogen receptor binding	ERbinding	-	1661	223 / 1438

^aselected threshold to discretize continuous properties into Concern (C) or nonConcern (nC) binary labels;

^brepartition of C and nC compounds.

2.3 Modelling workflow

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

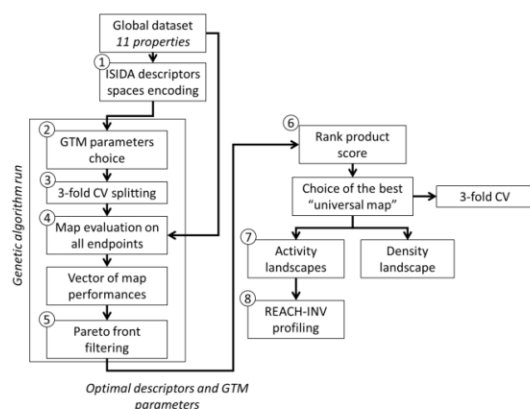


Figure 1. Modelling workflow. (1) different ISIDA descriptor spaces are generated for the Global dataset; (2) Genetic algorithm evaluates different types of descriptors and GTM parameters; (3-4) predictive performance of GTM-based models is evaluated in 3-fold cross-validation (CV) on each of the 11 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 centred 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 hyperplane, called 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 colour code depending on the number of compounds populating a given GTM node; (ii) class landscapes in which the colour 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 domain of the model. The profiling on the REACH-INV compounds has been performed only on those compounds which fulfilled the Applicability Domain requirement. 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 centres 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 (S_n),

Running title

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. However, this score can be biased toward the best performing model.

A more representative 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 is assigned starting from the top manifold; (iii) this process of sorting and score 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 co, Optimal Maps result from single 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 IAB(2-2)^[18], i.e. atom centred fragments with a radius of two. The following GTM hyperparameters were obtained in genetic algorithm optimization: $k = 20 \times 20$; $m = 7 \times 7$; $w = 1.2$; $l = 0.02884$. 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 obtained maps were

included to the consensus one by one in the order defined by the Rank Product. 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.

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.

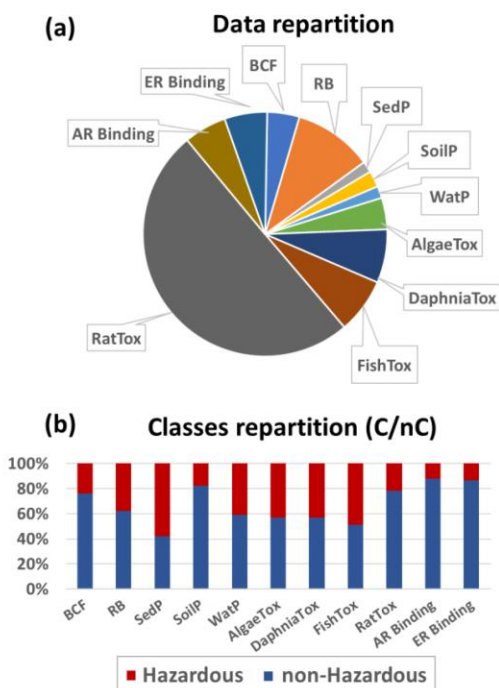


Figure 2. Datasets overview. (a) repartition of available experimental data for 11 endpoints; (b) Concern (C) and non-Concern (nC) classes repartition.

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 colour scale refers to the number of compounds populating a given area, ranging from 0 (white areas) up to 40 (yellow areas). Several zones have been selected in order to characterize the chemotypes distribution over the map.

- 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.
- 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.

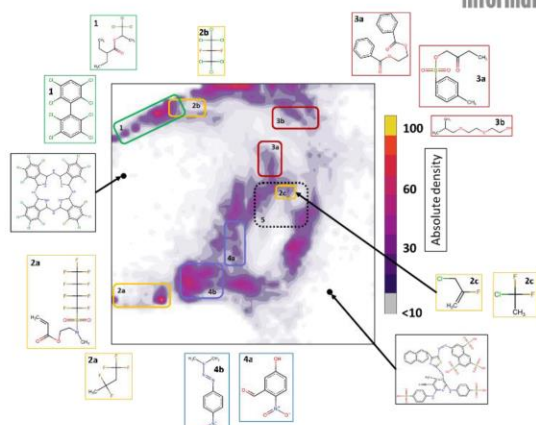


Figure 3. Density landscape of the Global dataset. The colour scale refers to the number of compounds populating a given zone. Coloured rectangles delimit map regions referred in the text. Representative structures populating selected zones of the map are shown.

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

The inventory of substances registered under the REACH-regulation (RECH-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 colours refer to zones uniquely populated by Global dataset and REACH-INV compounds, respectively whereas intermediate colours 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 (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.

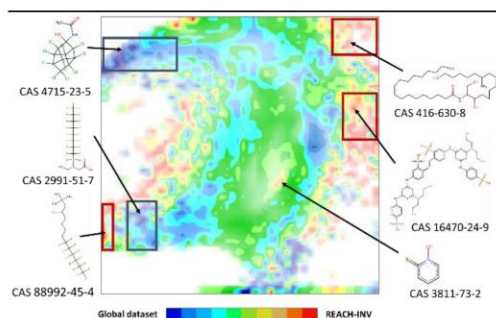


Figure 4. Global dataset and REACH-INV chemical space comparison. Blue regions are mainly populated by Global dataset

Running title

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 colours 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 S4). 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 out 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 derivatives 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 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.

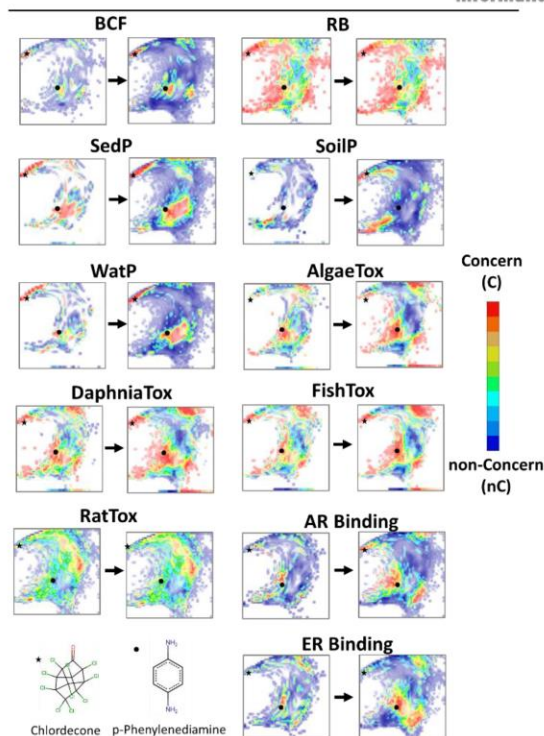


Figure 5. Class landscapes of the Global dataset for the 11 endpoints. Blue and red regions are populated by nC and C compounds, respectively. White areas correspond to unpopulated regions. Black stars and dots represent projects of two example compounds. The landscapes of the right side of the black 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 S4 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.

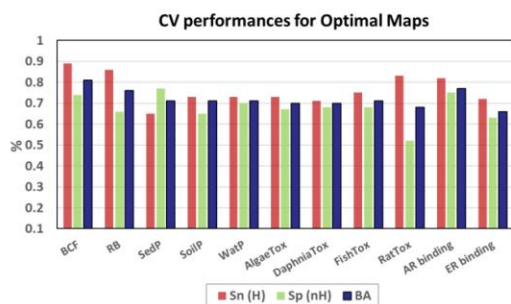


Figure 6. Optimal Maps performances.

3.6 Benchmarking studies.

Figure 7 and Table S5 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 less good 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.

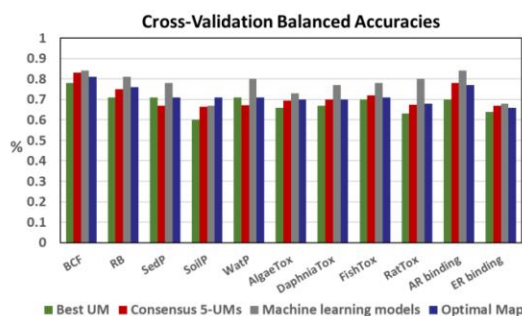


Figure 7. Results of benchmarking studies for the 11 studied endpoints. Best UM = best-performing Universal Map; Consensus 5-UMs = the top five performing Universal Maps have been ensemble in consensus; Machine-Learning models described in Section 2.10; Optimal Map is the best performing GTM model for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

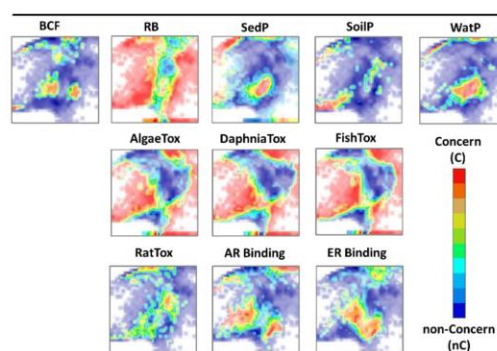


Figure 8. REACH-INV Class landscapes for the 11 endpoints predicted with the help of Universal Maps. All images in a large scale are available in SI.

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 colour 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 to 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.

Running title

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.

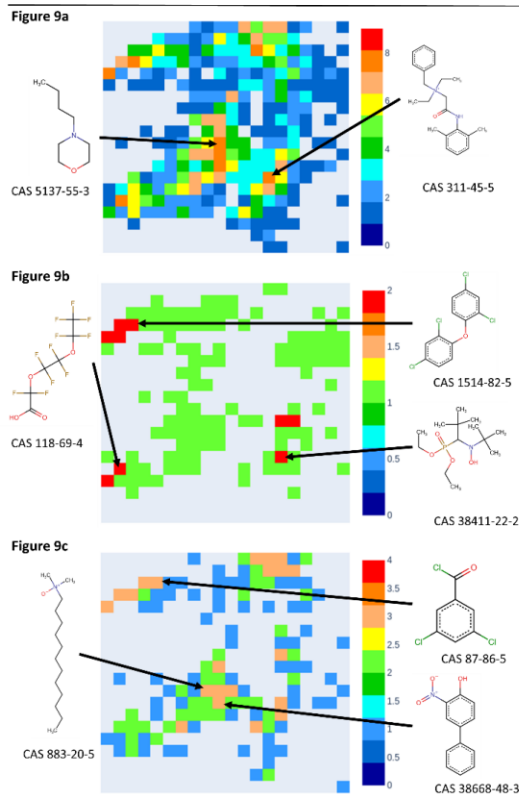


Figure 9 (a,b,c). REACH-INV profiling maps resulted from superposition of several class landscapes shown on Figure 8. Different superposition scenarios have been considered: (a) landscapes of all 11 endpoints; (b) two environmental fate endpoints; and (c) four acute aquatic toxicity endpoints. The colour 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, a perspective would be to add even more endpoints to the profiler. 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. Alberg, V. M. Alves, P. L. Andersson, C. H. Andrade, F. Bai, I. Balabin, D. Ballabio, E. Benfenati, B. Bhatarai, S. Boyer, J. Chen, V. Consonni, S. Farag, D. Fourches, A. T. Garcia-Sosa, P. Gramatica, F. Grisoni, C. M. Grulke, H. Hong, D. Horvath, X. Hu, R. Huang, N. Jeliakova, 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, "eChemPortal: Global Portal to Information on Chemical Substances", can be found under <https://www.echemportal.org/echemportal/index.action>, **2020**.
- [15] M. Berthold, N. Cebon, 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.

4.3 Part 3 – ISIDA/Predictor software implementation

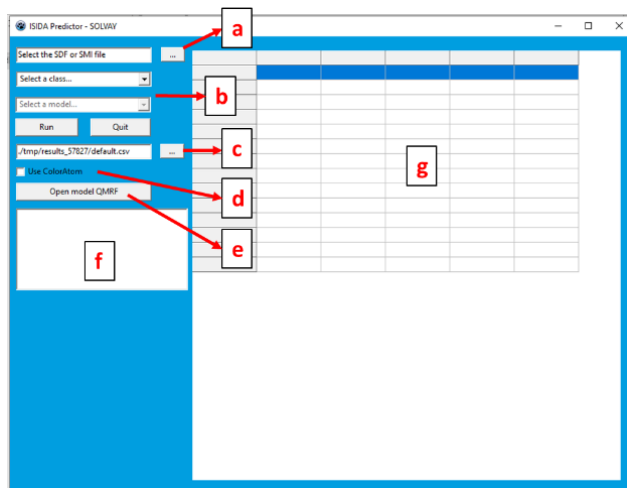
This section describes the functionalities of the ISIDA/Predictor (hereafter called “Predictor”) software. An online version is freely accessible at: http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi. All generated models have been added to the predictor together with their QMRF (QSAR Model Reporting Format) technical documentation and QPRF (QSAR Prediction Reporting Format) documents.

4.3.1 ISIDA/Predictor interface

The predictor interface looks like Figure 8, where:

- Select the file containing the molecule(s) to be predicted, either in *.sdf* format or *.smi*;
- Select the general class of the model (environmental fate, ecotoxicological or human toxicological properties) and then the specific model to be applied;
- Location where all the prediction files (including QPRFs) will be generated (one for each molecule);
- Select this option to generate the ColorAtom graphs;
- Open the QMRF file of the selected model (word document);
- General information concerning the selected model will appear here;
- Prediction for the input molecule(s) will appear here, together with their applicability domain and reliability assessment.

Figure 8. Predictor interface. See paragraph 4.3.1 for labels (a) to (g).



4.3.2 Input file

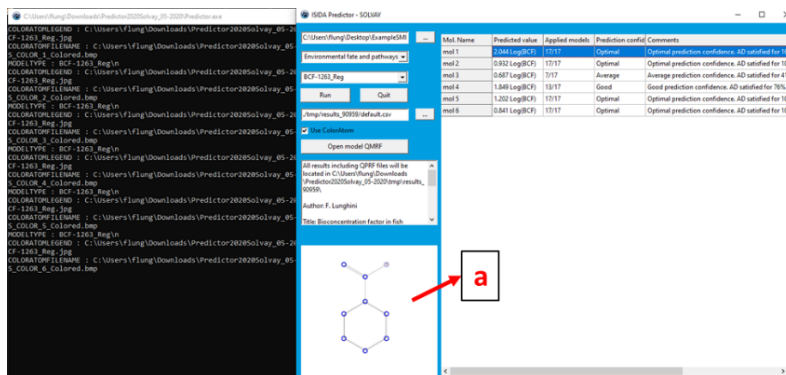
The easiest way to load molecule(s) is via the “.smi” (SMILES) file. This is simply a common text file where each line corresponds to the SMILES of a given molecule. The file can both have the “.txt” extension of the “.smi” extension. The latter is suggested as the file selection window such this type is filtered by default. Another way to load molecule(s) is with the “.sdf” (Structure-Data File) format. The advantage of this format is that it can store additional information together with the chemical structure (such as CAS, experimental properties, MW, etc.).

4.3.3 Results

Once the prediction process is done, the results will appear in the right part of the Predictor (Figure 9). Each molecule will correspond to a line, following the same order of the input file. The meaning of the columns is the following:

- Predicted value: the prediction given in standard unit ;
- Applied models: for how many individual models the query molecule fell inside AD. The higher the value, the more reliable the prediction is.
- Prediction confidence: a 4-scale score has been implemented, possible values are *OutsideAD*, *Average*, *Good*, *Optimal*.
- Comments: concise information about the quality of the prediction.
- ColorAtom: the atoms of the query molecule are color coded to render the influence of this part of the molecule on the prediction. If the “Use ColorAtom” option has not been selected, a picture of the selected molecule will appear without color code.

Figure 9. Results output. (a) illustration of a color coded molecule.



4.3.4 Reliability of the prediction

The applicability domain of each individual model is verified through the “Fragment control”: if a test molecule is found to have one fragment (i.e. a determined sequence of atoms and/or bonds) which is not present in the training set of the given individual model, that molecule is marked to be outside the applicability domain, since its structure has not been mapped entirely. In the consensus calculation, only the predictions that fell inside the AD of the individual model were taken into account.

A reliability assessment of the predictions is provided according to the number of individual models with positive AD: the following percentage thresholds of ≤ 25 , $>25 - \leq 50$; $>50 - \leq 80$; >80 were chosen to delimit reliability values: OutsideAD, Average, Good, Optimal.

Additionally, the standard deviation of the set of the individual models' predictions participating in the consensus is computed: if the st.dev. is higher than 0.5, than the prediction is considered to be OutsideAD, as the individual models are in strong disagreement.

V Conclusions and Perspectives

In this work, the following projects have been completed: (i) generation of predictive models on 11 endpoints relevant in the context of the REACH Regulation (including environmental fate, ecotoxicological and toxicological properties); (ii) application of GTM to the REACH-chemical space, which is the chemical space populated by compounds registered for REACH; (iii) multi-task model building able to profile new compounds about concerned properties; (iv) implementation of the generated models in the ISIDA/Predictor platform. The software and models have been deployed on the industrial site of Solvay.

Table 3 reports the endpoints that have been considered together with the models' performance and the link to the given dataset. 17762 unique compounds have been collected over the 11 endpoints, for a total of 29433 experimental values. This dataset has been extensively curated : (i) chemical structures have been validated by checking the correspondence between chemicals identifiers (e.g. SMILES ; CAS, IUPAC name) through automatized workflows querying online databases; (ii) only experimental values coming from reliable sources and matching the given OECD test guideline requirements were selected.

Chemical structures have been encoded by different types of substructural molecular fragments descriptors (*ISIDA Fragments*). State-of-the-art machine learning techniques (e.g. support vector machine and random forest) have been applied for model generation. For each property, a consensus model has been generated, consisting of several “*individual models*”, each one with a specific set-up of descriptors and machine learning method. Each individual model has its own applicability domain. A model prediction participates in the consensus calculation only if the compound is within the applicability domain of this model. A reliability score is defined based on the number and variance of predictions used for the consensus calculation: the higher, the more reliable the prediction is.

Generated models proved to have several advantages over already-existing tools, such as better prediction accuracy, extended applicability domain and extensive validation

on novel data coming from an industrial context. On industrial data, new models performed noticeably better, as demonstrated by their much larger applicability domain and increased accuracy (Table 1). Moreover, new industrial data has been used to update the models: as these compounds are missing from publicly-available databases, they bring new chemotypes and contribute to extend new models' applicability domains. All generated models follow the OECD Principles for QSAR models validation, which ensure that they can be used for regulatory purposes under the REACH Regulation, i.e. as valid alternatives to experimental testing.

The Generative Topographic Mapping (GTM) technique has been extensively used to compare the chemical space of public and industrial data sources. The GTM was very efficient to locate and characterize important missing chemotypes in public data sources compared to an industrial setup. A conventional strategy requiring assessment of all pairwise similarities is too time consuming whereas GTM compares data densities which is much faster procedure. GTM focalize on chemotypes, sets of compounds that share common scaffolds, rather than on individual compounds. This analysis is consistent with the results obtained during the modelling process: already-existing models (which are built on public data-only) were reported to have a restricted applicability domain when challenged to predict industrial data. The GTM approach was used to visualize the REACH-chemical space. These maps were used, in turn, as a common concept for multi-task models building able to predict all the 11 considered endpoints simultaneously. The models, termed "REACH Universal Maps", was used to profile the entire list of REACH-registered substances. We identified more concerns about acute aquatic toxicity than for environmental fate and human health. Finally, overlapping of different property landscapes provided an intuitive graphical view of regions of the chemical space populated by compounds with an undesirable (eco)toxicological profile.

All the models have been stored in the ISIDA/Predictor platform, which is freely available at http://infochim.u-strasbg.fr/cgi-bin/predictor_reach.cgi. An enhanced "desktop" version with additional features (e.g. generation of QPRF documents) has been delivered to Solvay.

5.1 Perspectives

This work can be further continued with the generation of new models on REACH-related properties. Indeed, the 11 properties considered here cover only a small fraction of a REACH dossier, constituted by tens of endpoints. The strategic bottle neck is the acquisition of good quality data. As a next step, these properties can be added to the REACH Universal Maps, improving the relevance of the profiling.

In this work, explicit labels values have been used. However, it is sometime possible to deduce an endpoint value from another endpoint. For instance, if a compound is ready biodegradable, it is automatically not persistent from the point of view of half-life assays. This strategy can be used at larger scale to supplement the existing data with new labels.

Besides, some physical properties should be investigated because they have an impact on the relevance of some predictions. For instance, aqueous solubility is of primordial importance because insoluble compounds are unlikely to have biological effect.

This PhD did not address the problems linked to metabolites and reactivity. If a compound can be considered safe, it is worth also to investigate its most likely metabolites to demonstrate that they are also safe. Reversely, the experimental outcome of an assay can be modified by the presence of metabolites. Deconvoluting the effects of metabolites and reaction products in some assays shall improve our understanding of the environmental and toxicological fate of compounds. In turn it may improve the predictive performances of models and improve chemical engineering toward better and safer chemical substances.

This work could be extended to the study of more complex objects, such as mixtures, polymers, peptides, polysaccharides or proteins. Complex systems need a lot more innovation to design a proper embedding to describe these objects.

Concerning the ISIDA/Predictor software, it can be improved with the addition of new features, such as other methods to assess the applicability domain of the models, or a read-across facility performing a similarity search of the query compound in the training set. Even more, new modules for ISIDA could be added to localize the queried compounds on a GTM of the REACH chemical space.

Another improvement could be to investigate the recent developments in artificial intelligence to the field of compounds generation, to automatically design of chemical structures modifications aiming at decreasing the hazardous potential of compound of interest.

Finally, it would be ethically better if computer-based property assessment could be used as a replacement to animal testing. Yet, this goal is still far since it would require the design of a rigorous experimental setup to assess the repeatability and transferability of measurements performed on animals and to compare it to those of computer predictions.

Table 3. Summary of generated models and collected data

Type	Endpoint (acronyme)	Tr. set	Int. Val.		Ext. Val.		Zenodo DOI :
			R ²	RMSE	R ²	RMSE	
REG	<i>Bioconcentration factor</i> (BCF)	1263	0.75	0.71	0.77	0.55	10.5281/zenodo.3228387
	<i>Algae acute toxicity</i> (AlgaTox)	1231	0.61	0.69	0.48	1.07	10.5281/zenodo.3708082
	<i>Daphnia acute toxicity</i> (DaphniaTox)	2083	0.67	0.78	0.58	0.93	10.5281/zenodo.3708082
	<i>Fish acute toxicity</i> (FishTox)	2152	0.67	0.73	0.54	0.97	10.5281/zenodo.3708082
	<i>Rat acute toxicity</i> (RodentTox)	11191	0.78	0.55	0.6	0.47	10.5281/zenodo.3300664
Type	Endpoint (acronyme)	Tr. set	Val. interne BA		Val. externe BA		
CLS	<i>Ready biodegradability</i> (RB)	3069	0.81		0.75		10.5281/zenodo.3540701
	<i>Sediment persistence</i> (SedP)	436	0.81		0.91		10.5281/zenodo.3698144
	<i>Soil persistence</i> (SoilP)	630	0.71		0.76		10.5281/zenodo.3698144
	<i>Water persistence</i> (WatP)	466	0.80		0.77		10.5281/zenodo.3698144
	<i>Androgen receptor</i> (AR binding)	1661	0.84		0.72		10.5281/zenodo.3935808
	<i>Estrogen receptor</i> (ER binding)	1661	0.68		0.60		10.5281/zenodo.3935808

Bibliography

The bibliography is divided in two parts: (i) the first part contains references to the cited literature following the citation order in the thesis, without considering those present in the published articles; (ii) the second part lists all the published papers in alphabetical order, comprising those inside the seven articles.

Part 1: References as cited in the thesis

- [1] 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.
- [2] 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.
- [3] European Commission, *On the animal testing and marketing ban and on the state of play in relation to alternative methods in the field of cosmetics*, Communication from the Commission to the European Parliament, 2013. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52013DC0135>.
- [4] F. Lunghini, G. Marcou, P. Azam, R. Patoux, M.H.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 Env. Res. 30 (2019), pp. 507–524.
- [5] F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert, A. Varnek and E.V. Miert, *Modelling of ready biodegradability based on combined public and industrial data sources*, SAR QSAR Env. Res. 31 (2020), pp. 171–186.
- [6] F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert, A. Varnek,

- E.V. Miert and A. Varnek, *Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context*, SAR QSAR Env. Res. 30 (2019), pp. 879–897.
- [7] G. Marcou, D. Horvath, F. Bonachera and A. Varnek, *Laboratoire De Chemoinformatique UMR7140 CNRS*, University of Strasbourg, Strasbourg, FR, 2019. Available at <http://infochim.ustrasbg.fr/cgi-bin/predictor.cgi>.
- [8] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou and A. Varnek, *Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison*, Mol. Inf. 31 pp. 301–312.
- [9] N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrickc and G. Patlewicz, *Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation*, Comput. Tox. 201 pp. 489–492.
- [10] OECD, *Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption*, Tech. Rep. 978-92-64-30474-1, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2018.
- [11] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek et al., *CERAPP: collaborative estrogen receptor activity prediction project*, Environ. Heal. Perspect. 124 pp. 1023–1033.
- [12] K. Mansouri, N. Kleinstreuer, A.M. Abdelaziz, D. Alberga, V.M. Alves, P.L. Andersson et al., *Compara: collaborative modeling project for androgen receptor activity*, Environ. Health Perspect. (2020), pp. 123–133.
- [13] G. Marcou, D. Horvath, V. Solov'Ev, A. Arrault, P. Vayer, A. Varnek, V. Solov'Ev, A. Arrault, P. Vayer and A. Varnek, *Interpretability of SAR/QSAR models of any complexity by atomic contributions*, Mol. Inf. 31 (2012), pp. 639–642.
- [14] *Canadian Environmental Protection Act, 1999*. Government of Canada, Canada, 2020. Available at: <https://www.canada.ca/en/services/environment/pollution-waste-management/understanding-environmental-protection-act.html>
- [15] *The Scientific Aspects of the Chemical Substances Control Law in Japan*. Ministry of International Trade and Industry, Japan, 1978. Available at: <https://www.sciencedirect.com/book/9780080220598/aquatic-pollutants>.

- [16] *US EPA*, US Environmental Protection Agency, US, 2017. Available at: <https://cfpub.epa.gov/>.
- [17] *The Globally Harmonized System (GHS) of classification and labelling of chemicals*, ST/SG/AC.10/30/Rev.4, United Nations, 2011. Available at: https://www.unece.org/fileadmin/DAM/trans/danger/publi/ghs/ghs_rev04/English/ST-SG-AC10-30-Rev4e.pdf.
- [18] *EFSA website*. European Food Safety Authority (EFSA), Parma, IT, 2020. Available at <http://www.efsa.europa.eu/it>.
- [19] E. Benfenati, A. Manganaro and G. Gini, *VEGA-qsar: ai inside a platform for predictive toxicology*, CEUR Workshop Proc. 1107 (2013), pp. 21–28.
- [20] T. Martin, P. Harten, and D. Young, *TEST (Toxicity Estimation Software Tool) v 4.1*, US Environmental Protection Agency, Washington DC, USA, 2012; software available at: <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.
- [21] *Estimation Programs Interface Suite™ for Microsoft® Windows V4.11*, US Environmental Protection Agency. Washington DC, USA.
- [22] K. Mansouri, C.M. Grulke, R.S. Judson and A.J. Williams, *OPERA models for predicting physicochemical properties and environmental fate endpoints*, J. Cheminform. 10 (2018), pp. 1–19.
- [23] C. Hansch, A. Leo and D.J. Livingstone, *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, pp. 10678–10688.
- [24] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Vol. 2, 2010.
- [25] J.L. Faulon, A. Bender, *Handbook of Chemoinformatics Algorithms*, CRC Press, 2010.
- [26] R.J.J. Larson and C.E.E. Cowan, *Quantitative application of biodegradation data to environmental risk and exposure assessments*, *environ, Toxicol. Chem.* 14 (1995), pp. 1433–1442.
- [27] K. Mansouri, C.M. Grulke, R.S. Judson and A.J. Williams, *OPERA models for predicting physicochemical properties and environmental fate endpoints*, J.

Cheminform. 10 pp. 1–19.

- [28] NITE, *Data from: Biodegradation and bioconcentration data under CSCL*, National Institute of Technology and Evaluation, 2007; dataset available at: <https://www.nite.go.jp/en/>.
- [29] NIH, *PubChem*, National Library of Medicine, National Center for Biotechnology Information, Bethesda, Maryland, 2019; available at <https://pubchem.ncbi.nlm.nih.gov/>.
- [30] OASIS, *QSAR Toolbox v 4.3*, OASIS Laboratory of mathematical chemistry, Burgas, BG, 2017; software available at <http://oasis-lmc.org/products/software/toolbox.aspx>.
- [31] M. Berthold, N. Cebon, 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.
- [32] F. Ruggiu, G. Marcou, A. Varnek and D. Horvath, *ISIDA property-labelled fragment descriptors*, Mol. Inform. 29 pp. 855–868.
- [33] D. Horvath, J. Brown, G. Marcou and A. Varnek, *An evolutionary optimizer of libsvm models*, Challenges 5 (2014), pp. 450–472.
- [34] C. Chih-Chung and L. Chih-Jen, *LIBSVM: a library for support vector machines*, ACM Trans. Intell. Syst. Technol 2 pp. 1–27.
- [35] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.
- [36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, *The WEKA machine learning workbench*, in *Data Mining: Practical Machine Learning Tools and Techniques*, I.H. Witten and E. Frank, eds., Morgan Kaufmann Publishers, San Fransisco, pp. 363–449.
- [37] NIH, *Data from: Tox21 data challenge 2014*, National Institutes of Health, dataset available at: <https://tripod.nih.gov/tox21>.

Part 2. Full bibliographic list

- 1 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.
- 2 A. Furuhashi, K. Hasunuma, T.I. Hayashi and N. Tatarazako, *Predicting algal growth inhibition toxicity: Three-step strategy using structural and physicochemical properties*, SAR QSAR Environ. Res. 27 (2016), pp. 343–362.
- 3 A. Gissi, A. Lombardo, A. Roncaglioni, D. Gadaleta, G.F. Mangiatordi, O. Nicolotti, and E. Benfenati, *Evaluation and comparison of benchmark QSAR models to predict a relevant reach endpoint: The bioconcentration factor (BCF)*, Environ. Res. 137 (2015), pp. 398–409.
- 4 A. Gissi, D. Gadaleta, M. Floris, S. Olla, A. Carotti, E. Novellino, E. Benfenati, and O. Nicolotti, *An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes*, ALTEX 31 (2014), pp. 23–36.
- 5 A. Gissi, K. Louekari, L. Hoffstadt, N. Bornatowicz and A.M. Aparicio, *Alternative acute oral toxicity assessment under REACH based on sub-acute toxicity values*, ALTEX 34 (2017), pp. 353–361.
- 6 A. Golbamaki, A. Cassano, A. Lombardo, Y. Moggio, M. Colafranceschi and E. Benfenati, *Comparison of in silico models for prediction of Daphnia magna acute toxicity*, SAR QSAR Environ. Res. 25 (2014), pp. 673–694.
- 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.
- 8 A. Levet, C. Bordes, Y. Clément, P. Mignon, C. Morell, H. Chermette, P. Marote and P. Lantéri, *Acute aquatic toxicity of organic solvents modeled by QSARs*, J. Mol. Model. 22 (2016), pp. 288–298 .
- 9 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.
- 10 A. Rybacka, C. Rudén, I. V. Tetko and P.L. Andersson, *Identifying potential endocrine disruptors among industrial chemicals and their metabolites - development and evaluation of in silico tools*, Chemosphere 139 (2015), pp. 372–378.
- 11 A. Tropsha, *Best practices for QSAR model development, validation, and exploitation*, Mol. Inform. 29 (2010), pp. 476–488.
- 12 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.
- 13 A.A. Toropov, A.P. Toropova, M. Marzo, J. Lou Dorne, N. Georgiadis and E. Benfenati, *QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL*

- software and EFSA's OpenFoodTox database, *Environ. Toxicol. Pharmacol.* 53 (2017), pp. 158–163.
- 14 A.A. Toropov, B.F. Rasulev and J. Leszczynski, *QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: Comparative analysis by MLRA and optimal descriptors*, *QSAR Comb. Sci.* 26 (2007), pp. 686–693.
 - 15 A.M. Vinggaard, J. Niemelä, E.B. Wedebye and G.E. Jensen, *Screening of 397 chemicals and development of a quantitative structure-activity relationship model for androgen receptor antagonism*, *Chem. Res. Toxicol.* 21 (2008), pp. 813–823.
 - 16 A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies et al., *The ChEMBL bioactivity database: An update*, *Nucleic Acids Res.* 42 (2014), pp. 1083-1090.
 - 17 A.P. Freidig, S. Dekkers, M. Verwei, E. Zvinavashe, J.G.M. Bessems and J.J.M. van de Sandt, *Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds*, *Toxicol. Lett.* 170 (2007), pp. 214–222.
 - 18 A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, *Coral: Quantitative models for estimating bioconcentration factor of organic compounds*, *Chemometr. Intell. Lab.* 118 (2012), pp. 70-73.
 - 19 Accelrys , (TOPKAT) *TOxicity Prediction by Komputer Assisted Technology v 3.1*, Accelrys software Inc., San Diego, CA, US, 2019; software available at: <http://www.3dsbiovia.com/>.
 - 20 ACD/Labs, ACD/Percepta Platform v 2018.1, Advanced Chemistry Development, Inc. (ACD/Labs), 2019; software available at: <http://www.acdlabs.com/>.
 - 21 C. Bertinetto, C. Duce, R. Solaro, M.R. Tiné, A. Micheli, K. Héberger, A. Miličević and S. Nikolić, *Modeling of the acute toxicity of benzene derivatives by complementary qsar methods*, *Match* 70 (2013), pp. 1005–1021.
 - 22 C. Chih-Chung and L. Chih-Jen, *LIBSVM: A library for support vector machines*, *ACM Trans. Intell. Syst. Technol.* 2 (2011), pp. 1–27.
 - 23 C. Hansch, A. Leo and D.J. Livingstone, *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, 1996.
 - 24 C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, *The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics*, *J. Chem. Inf. Comput. Sci.* 43 (2003), pp. 493-500.
 - 25 C.I. Cappelli, A. Cassano, A. Golbamaki, Y. Moggio, A. Lombardo, M. Colafranceschi and E. Benfenati, *Assessment of in silico models for acute aquatic toxicity towards fish under REACH regulation*, *SAR QSAR Environ. Res.* 26 (2015), pp. 977–999.
 - 26 C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister and R.A. Drummond, *Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)*, *Environ. Toxicol. Chem.* 16 (2005), pp. 948.
 - 27 C.L. Waller, B.W. Juma, L. Earl Gray and W.R. Kelce, *Three-dimensional quantitative structure-activity relationships for androgen receptor ligands*, *Toxicol. Appl. Pharmacol.* 137 (1996), pp. 219-227.

- 28 C.M. Bishop, M. Svensén, C.K.I. Williams and M. Svens, *The generative topographic mapping*, Neural Comput. 10 (1998), pp. 215–234.
- 29 C.M. Olsen, E.T.M. Meussen-Elholm, J.K. Hongslo, J. Stenersen and K.E. Tollefsen, *Estrogenic effects of environmental chemicals: an interspecies comparison*, Comp. Biochem. Physiol. - C Toxicol. Pharmacol. 141 (2005), pp. 267–274.
- 30 CADD Group Chemoinformatics Tools and User Services. 2019. Bethesda, Maryland, US: National Cancer Institute, Chemical Biology Laboratory. Available at: <https://cactus.nci.nih.gov/>
- 31 Canadian Environmental Protection Act, 1999. Government of Canada, Canada, 2020. Available at: <https://www.canada.ca/en/services/environment/pollution-waste-management/understanding-environmental-protection-act.html>
- 32 Catalogic: Environmental fate and ecotoxicity models v 5.11.13, OASIS Laboratory of Mathematical Chemistry, Burgas, BG, 2013; software available at: <http://oasis-lmc.org/products/software/catalogic.aspx>.
- 33 CEFIC, Data from: BCF Bioconcentration factor database, European Chemical Industry Council Long range research initiative, 2007; dataset available at: <http://cefic-lri.org/>.
- 34 ChemProp v 6.7, Helmholtz Centre for Environmental Research-UFZ, Leipzig, DE, 2018; software available at: <http://www.ufz.de/ecochem/chemprop>.
- 35 D. Ballabio, F. Biganzoli, R. Todeschini and V. Consonni, *Qualitative consensus of QSAR ready biodegradability predictions*, Toxicol. Environ. Chem. 99 (2017), pp. 1193–1216.
- 36 D. Ballabio, F. Grisoni, V. Consonni and R. Todeschini, *Integrated QSAR models to predict acute oral systemic toxicity*, preprint (2019), submitted for publication. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201800124>.
- 37 D. Ding, L. Xu, H. Fang, H. Hong, R. Perkins, S. Harris, E.D. Bearden, L. Shi and W. Tong, *The EDKB: An Established Knowledge Base for Endocrine Disrupting Chemicals*, BMC Bioinformatics 11 (2010), pp. 1471–1478.
- 38 D. Fourches, E. Muratov and A. Tropsha, *Trust but verify: On the importance of chemical structure curation in chemoinformatics and QSAR modeling research*, J. Chem. Inf. Model. 50 (2010), pp. 1189–1204.
- 39 D. Horvath, I. Baskin, G. Marcou and A. Varnek, *Generative topographic mapping of conformational space*, Mol. Inform. 36 (2017), pp. 24–36.
- 40 D. Horvath, J. Brown, G. Marcou and A. Varnek, *An Evolutionary Optimizer of libsvm Models*, Challenges 5 (2014), pp. 450–472.
- 41 D.A.R.S. Latino, J. Wicker, M. Gütlein, E. Schmid, S. Kramer and K. Fenner, *Eawag-soil in Envipath: A new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data*, Environ. Sci. Process. Impacts 19 (2017), pp. 449–464.
- 42 D.G. Lloyd, R.B. Hughes, D.M. Zisterer, D.C. Williams, C. Fattorusso, B. Catalanotti, G. Campiani and M.J. Meegan, *Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor*, J. Med. Chem. 47 (2004), pp. 5612–5615.

- 43 D.J. Dix, R.J. Kavlock, R.W. Setzer, K.A. Houck, A.M. Richard and M.T. Martin, *The toxcast program for prioritizing toxicity testing of environmental chemicals*, Toxicol. Sci. 95 (2006), pp. 5–12.
- 44 D.W. Roberts, *QSAR issues in aquatic toxicity of surfactants*, Sci. Total Environ. 109 (1991), pp. 557–568.
- 45 E. Benfenati, A. Manganaro and G. Gini, *VEGA-QSAR: AI inside a platform for predictive toxicology*, CEUR Workshop Proc. 1107 (2013), pp. 21–28.
- 46 E. Furusjö, A. Allard, S. Nilsson, M. Rahmberg and A. Svenson, *State of the art assessment of endocrine disrupters*, Tech. Rep. 070307/2009/550687/SER/D3, European Commission, DG Environment, Brussels, BE, 2012.
- 47 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.
- 48 E. Lo Piparo and A. Worth, *Review of QSAR Models and Software Tools for Predicting Developmental and Reproductive Toxicity*, Tech. Rep. EUR 24522 EN, European Commission JRC, Ispra, IT, 2010.
- 49 E. Papa and P. Gramatica, *Screening of persistent organic pollutants by QSBR classification models: A comparative study*, J. Mol. Graph. Model. 27 (2008), pp. 59–65.
- 50 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.
- 51 ECETOC, *Persistence of chemicals in the environment*, Tech. Rep 90, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, BE, 2003.
- 52 ECHA Homepage. 2019. Helsinki, FI: European Chemicals Agency. Available at: <https://echa.europa.eu/> [Accessed December 1st 2017].
- 53 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.
- 54 EFSA website. European Food Safety Authority (EFSA), Parma, IT, 2020. Available at <http://www.efsa.europa.eu/it>.
- 55 *Electronic code of federal regulations (40 cfr part 156)*. 2019. Washington DC, US: United States Government Publishing Office. Available at: <https://www.law.cornell.edu/cfr/text/40/part-156.html> [Accessed 1st May 2019].
- 56 EPA, *Data from: Endocrine disruptor screening program (EDSP) in the 21st century*, Environmental Protection Agency, dataset available at: <https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-edsp-21st-century>.
- 57 *Estimation Programs Interface Suite™ for Microsoft® Windows v 4.11*, US Environmental Protection Agency, Washington DC, USA, 2012; software available at: <https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>.
- 58 European Commission, *On the animal testing and marketing ban and on the state of play in relation to alternative methods in the field of cosmetics*, Communication from the

Commission to the European Parliament, 2013. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52013DC0135>.

- 59 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.
- 60 F. Grisoni, V. Consonni, M. Vighi, S. Villa, and R. Todeschini, *Expert QSAR system for predicting the bioconcentration factor under the reach regulation*, Environ. Res. 148 (2016), pp. 507–512.
- 61 F. Grisoni, V. Consonni, S. Villa, M. Vighi, and R. Todeschini, *QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions?*, Chemosphere 127 (2015), pp. 171–179.
- 62 F. Lunghini, G. Marcou, P. Azam, D. Horvath, R. Patoux, E. Van Miert et al., *Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context*, SAR QSAR Environ. Res. 30 (2019), pp. 879–897.
- 63 F. Lunghini, G. Marcou, P. Azam, R. Patoux, M.H. Enrici, F. Bonachera et al., *QSPR models for bioconcentration factor (BCF): Are they able to predict data of industrial interest?*, SAR QSAR Environ. Res. 30 (2019), pp. 507–524.
- 64 F. Lunghini, G. Marcou, P. Gantzer, P. Azam, D. Horvath, E. Van Miert et al., *Modelling of ready biodegradability based on combined public and industrial data sources*, SAR QSAR Environ. Res. 31 (2020), pp. 171–186.
- 65 F. Pizzo, A. Lombardo, M. Brandt, A. Manganaro and E. Benfenati, *A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH*, Environ. Int. 88 (2016), pp. 250–260.
- 66 F. Ruggiu, G. Marcou, A. Varnek and D. Horvath, *ISIDA Property-labelled fragment descriptors*, Mol. Inform. 29 (2010), pp. 855–868.
- 67 G. Gini, G. Giuseppina, T. Ferrari, A. Lombardo, A. Cassano, E. Benfenati, *A new QSAR model for acute fish toxicity based on mined structural alerts*, J. Toxicol. Risk Assess. 5 (2019), pp. 2572–2580 .
- 68 G. Marcou, D. Horvath, F. Bonachera and A. Varnek, *Laboratoire De Chimoinformatique UMR7140 CNRS*, University of Strasbourg, Strasbourg, FR, 2019. Available at <http://infochim.ustrasbg.fr/cgi-bin/predictor.cgi>.
- 69 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.
- 70 G.E. Bragin, C.W. Davis, M.H. Kung, B.A. Kelley, C.A. Sutherland and M.A. Lampi, *Biodegradation and ecotoxicity of branched alcohol ethoxylates: Application of the target*

- lipid model and implications for environmental classification*, J. Surfactants Deterg. 23 (2020), pp. 383–403.
- 71 G.E. Jensen, J.R. Niemelä, E.B. Wedeby and N.G. Nikolov, *QSAR models for reproductive toxicity and endocrine disruption in regulatory use - a preliminary investigation*, SAR QSAR Environ. Res. 19 (2008), pp. 631–641.
 - 72 G.M. Cramer, R.A. Ford and R.L. Hall, *Estimation of toxic hazard-a decision tree approach*, Food Cosmet. Toxicol. 16 (1976), pp. 255–276.
 - 73 Government of Canada, *Data from: Canadian domestic substances list (DSL), Environment and Climate Change Canada*, 1999; dataset available at: <https://www.canada.ca/en/environment-climate-change/services/canadian-environmental-protection-act-registry/substances-list/domestic.html>.
 - 74 H. Fang, W. Tong, W.S. Branham, C.L. Moland, S.L. Dial, H. Hong, Q. Xie, R. Perkins, W. Owens and D.M. Sheehan, *Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor*, Chem. Res. Toxicol. 16 (2003), pp. 1338–1358.
 - 75 H. Liu, E. Papa and P. Gramatica, *QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles*, Chem. Res. Toxicol. 19 (2006), pp. 1540–1548.
 - 76 H. Liu, X. Yang and R. Lu, *Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin*, Chemosphere 156 (2016), pp. 1–7.
 - 77 H.A. Gaspar, I.I. Baskin, G. Marcou, D. Horvath and A. Varnek, *Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge*, J. Chem. Inf. Model. 55 (2015), pp. 84–94.
 - 78 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.
 - 79 H.J.M. Verhaar, C.J. van Leeuwen and J.L.M. Hermens, *Classifying environmental pollutants*, Chemosphere 25 (1992), pp. 471–491 .
 - 80 I. Tsakovska, I. Lessigiarska, T. Netzeva and A.P. Worth, *A mini review of mammalian toxicity (Q)SAR models*, QSAR Comb. Sci. 27 (2008), pp. 41–48.
 - 81 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.
 - 82 J. He, T. Peng, X. Yang and H. Liu, *Development of QSAR models for predicting the binding affinity of endocrine disrupting chemicals to eight fish estrogen receptor*, Ecotoxicol. Environ. Saf. 148 (2018), pp. 211–219.
 - 83 J. Li and P. Gramatica, *QSAR classification of estrogen receptor binders and pre-screening of potential pleiotropic EDCS*, SAR QSAR Environ. Res. 21 (2010), pp. 657–669.

- 84 J.A. Arnot and F.A.P.C. Gobas, *A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms*, Environ. Rev. 14 (2006), pp. 257-297.
- 85 J.C. Dearden and M. Hewitt, *QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors*, SAR QSAR Environ. Res. 21 (2010), pp. 671-680.
- 86 J.F. Aranda, D.E. Bacelo, M.S. Leguizamón Aparicio, M.A. Ocsachoque, E.A. Castro, and P.R. Duchowicz, *Predicting the bioconcentration factor through a conformation-independent QSPR study*, SAR QSAR Environ. Res. 28 (2017), pp. 749-763.
- 87 J.J. Naveja, U. Norinder, D. Mucs, E. López-López and J.L. Medina-Franco, *Chemical space, diversity and activity landscape analysis of estrogen receptor binders*, RSC Adv. 8 (2018), pp. 38229-38237.
- 88 J.L. Faulon, A. Bender, *Handbook of Chemoinformatics Algorithms*, CRC Press, 2010.
- 89 J.X. Guo, J.J.-Q. Wu, J.B. Wright and G.H. Lushington, *Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: A molecular modeling study*, Chem. Res. Toxicol. 19 (2006), pp. 209-216.
- 90 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>.
- 91 K. Acharya, D. Werner, J. Dolfing, M. Barycki, P. Meynet, W. Mroziak, O. Komolafe, T. Puzyn and R.J. Davenport, *A quantitative structure-biodegradation relationship (QSBR) approach to predict biodegradation rates of aromatic chemicals*, Water Res. 157 (2019), pp. 181-190.
- 92 K. Khan and K. Roy, *Ecotoxicological QSAR modelling of organic chemicals against Pseudokirchneriella subcapitata using consensus predictions approach*, SAR QSAR Environ. Res. 30 (2019), pp. 665-681.
- 93 K. Khan, D. Baderna, C. Cappelli, C. Toma, A. Lombardo, K. Roy and E. Benfenati, *Ecotoxicological QSAR modeling of organic compounds against fish: Application of fragment based descriptors in feature analysis*, Aquat. Toxicol. 212 (2019), pp. 162-174.
- 94 K. Khan, P.M. Khan, G. Lavado, C. Valsecchi, J. Pasqualini, D. Baderna, M. Marzo, A. Lombardo, K. Roy and E. Benfenati, *QSAR modeling of Daphnia magna and fish toxicities of biocides using 2D descriptors*, Chemosphere 229 (2019), pp. 8-17.
- 95 K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A. Richard, C. 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. Kanclerla, 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, and R. Judson, *CERAPP: Collaborative estrogen receptor activity prediction project*, Environ. Health Perspect. 124 (2016), pp. 1023-1033.

- 96 K. Mansouri, C.M. Grulke, R.S. Judson and A.J. Williams, *OPERA models for predicting physicochemical properties and environmental fate endpoints*, J. Cheminform. 10 (2018), pp. 1–19.
- 97 K. Mansouri, N. Kleinstreuer, A.M. Abdelaziz, D. Alberga, V.M. Alves, P.L. Andersson et al., *CoMPARA: Collaborative modeling project for androgen receptor activity*, Environ. Health Perspect. (2020), .
- 98 K. Samghani and M. HosseinFatemi, Developing a support vector machine based QSPR model for prediction of half-life of some herbicides, Ecotoxicol. Environ. Saf. 129 (2016), pp. 10–15.
- 99 K.L.E. Kaiser and S.P. Niculescu, *Modeling acute toxicity of chemicals to Daphnia magna: A probabilistic neural network approach*, Env. Tox. Chem. 20 (2001), pp. 420–431.
- 100 K.P. Singh, S. Gupta, A. Kumar and D. Mohan, *Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology*, Chem. Res. Toxicol. 27 (2014), pp. 741–753.
- 101 L. Breiman, *Random Forests*, Mach. Learn. 45 (2001), pp. 5–32.
- 102 L. Zhang, A. Sedykh, A. Tripathi, H. Zhu, A. Afantitis, V.D. Mouchlis, G. Melagraki, I. Rusyn and A. Tropsha, *Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure based virtual screening approaches*, Toxicol. Appl. Pharmacol. 272 (2013), pp. 67–76.
- 103 L.B. Salum, I. Polikarpov and A.D. Andricopulo, *Structure-based approach for the study of estrogen receptor binding affinity and subtype selectivity*, J. Chem. Inf. Model. 48 (2008), pp. 2243–2253.
- 104 M. Berthold, N. Cebon, F. Dill, T. Gabriel, T. Kötter, T. Meinl et al., *KNIME - The Konstanz information miner: Version 2.0 and Beyond*, SIGKDD Explor. 11 (2009), pp. 26–31.
- 105 M. Cassotti, D. Ballabio, R. Todeschini and V. Consonni, *A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas)*, SAR QSAR Environ. Res. 26 (2015), pp. 217–243.
- 106 M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I. V. Tetko and R. Todeschini, *Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method*, ATLA Altern. to Lab. Anim. 42 (2014), pp. 31–41.
- 107 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, *The Weka Workbench*, in *Data Mining: Practical Machine Learning Tools And Techniques*, I.H. Witten, eds., Morgan Kaufman Publishers, San Fransisco, 2005, 363–449.
- 108 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, *The WEKA machine learning workbench*, in *Data Mining: Practical Machine Learning Tools and Techniques*, I. H. Witten and E. Frank, eds., Morgan Kaufmann Publishers, San Fransisco, 2005, pp. 363–449.
- 109 M. Hrovat, H. Segner and S. Jeram, *Variability of in vivo fish acute toxicity data*, Regul. Toxicol. Pharmacol. 54 (2009), pp. 294–300.

- 110 M. Nendza and M. Müller, *Screening for low aquatic bioaccumulation (1): Lipinski's 'rule of 5' and molecular size*, SAR QSAR Environ. Res. 21 (2010), pp. 495-512.
- 111 M. Pavan and A.P. Worth, *Review of estimation models for biodegradation*, QSAR Comb. Sci. 27 (2008), pp. 32-40.
- 112 M. Pavan, A.P. Worth and T.I. Netzeva, *Comparative assessment of QSAR models for aquatic toxicity*, Tech. Rep. 21750EN, Joint Research Centre, Ispra, IT, 2005.
- 113 M. Pavan, T.I. Netzeva, and W. Andrew P, *Review of literature-based quantitative structure - activity relationship models for bioconcentration*, QSAR Comb. Sci. 27 (2008), pp. 21-31.
- 114 M. Salahinejad, E. Zolfonoun and J.B. Ghasemi, *Predicting degradation half-life of organophosphorus pesticides in soil using three-dimensional molecular interaction fields*, Int. J. Quant. Struct. Relationships 2 (2017), pp. 27-35.
- 115 M. Todorov, E. Mombelli, S. Aït-Aïssa and O. Mekenyan, *Androgen receptor binding affinity: a QSAR evaluation*, SAR QSAR Environ. Res. 22 (2011), pp. 265-291.
- 116 M.I. Petoumenou, F. Pizzo, J. Cester, A. Fernández, and E. Benfenati, *Comparison between bioconcentration factor (BCF) data provided by industry to the European chemicals agency (ECHA) and data derived from QSAR models*, Environ. Res. 142 (2015), pp. 529-534.
- 117 M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology*, Nucleic Acids Res. 44 (2016), pp. 1045-1053.
- 118 M.O. Taha, M. Tarairah, H. Zalloum and G. Abu-Sheikha, *Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and in silico search for new hits*, J. Mol. Graph. Model. 28 (2010), pp. 383-400.
- 119 M.S. McLachlana, *Can the stockholm convention address the spectrum of chemicals currently under regulatory scrutiny? Advocating a more prominent role for modeling in POP screening assessment*, Environ. Sci. Proces. Impacts 20 (2018), pp. 32-37.
- 120 N. Andersson, M. Arena, D. Auteri, S. Barmaz, E. Grignard, A. Kienzler, P. Lepper, A.M. Lostia, S. Munn, J.M. Parra Morte, F. Pellizzato, J. Tarazona, A. Terron and S. Van der Linden, *Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009*, EFSA J. 16 (2018), pp. 1-135.
- 121 N. Burden, S.K. Maynard, L. Weltje and J.R. Wheeler, *The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: A retrospective validation approach*, Regul. Toxicol. Pharmacol. 80 (2016), pp. 241-246.
- 122 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.
- 123 N.C. Kleinstreuer, A.L. Karmaus, K. Mansouri, D.G. Allen, J.M. Fitzpatrick and G. Patlewicz, *Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation*, Comput. Tox. 201 (2018), pp. 489-492.

- 124 NIH, *Data from: Tox21 data challenge 2014*, National Institutes of Health, dataset available at: <https://tripod.nih.gov/tox21>.
- 125 NIH, *PubChem*, National Library of Medicine, National Center for Biotechnology Information, Bethesda, Maryland, 2019; available at <https://pubchem.ncbi.nlm.nih.gov/>.
- 126 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/>.
- 127 NPIC, *Data from: OSU extension pesticide properties database*. National Pesticide Information Center; dataset available at: <http://npic.orst.edu/ingred/ppdmove.htm> [Accessed 1 January 2019].
- 128 OASIS, *QSAR Toolbox v 4.3*, OASIS Laboratory of mathematical chemistry, Burgas, BG, 2017; software available at <http://oasis-lmc.org/products/software/toolbox.aspx>.
- 129 OECD, *Data from: eChemPortal: Global portal to information on chemical substances*, Organisation for Economic Co-operation Development, 2017; dataset available at: <https://www.echemportal.org/echemportal/index.action>.
- 130 OECD, *Guidance document on aquatic toxicity testing of difficult substances and mixtures*, Tech. Rep. JT03442844, Organisation for Economic Co-operation Development, Paris, FR, 1992.
- 131 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.
- 132 OECD, *Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption*, Tech. Rep. 978-92-64-30474-1, Organisation for Economic Cooperation and Development (OECD), Paris, FR, 2018.
- 133 OECD, *Test No. 301: Ready biodegradability*, Tech. Rep. 9789264070349, Organisation for Economic Co-operation Development, Paris, FR, 1992.
- 134 OECD, *Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure*, Tech Rep. 9264185291, Organisation for Economic Co-operation Development, Paris, FR, 2012.
- 135 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.
- 136 P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.
- 137 P. Gramatica, S. Cassani, and A. Sangion, *PBT assessment and prioritization by PBT Index and consensus modeling: Comparison of screening results from structural models*, Environ. Int. 77 (2015), pp. 25–34.
- 138 P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap and E. Papa, *QSAR modeling is not “Push a button and find a correlation”: A case study of toxicity of (Benzo-)triazoles on Algae*, Mol. Inform. 31 (2012), pp. 817–835.

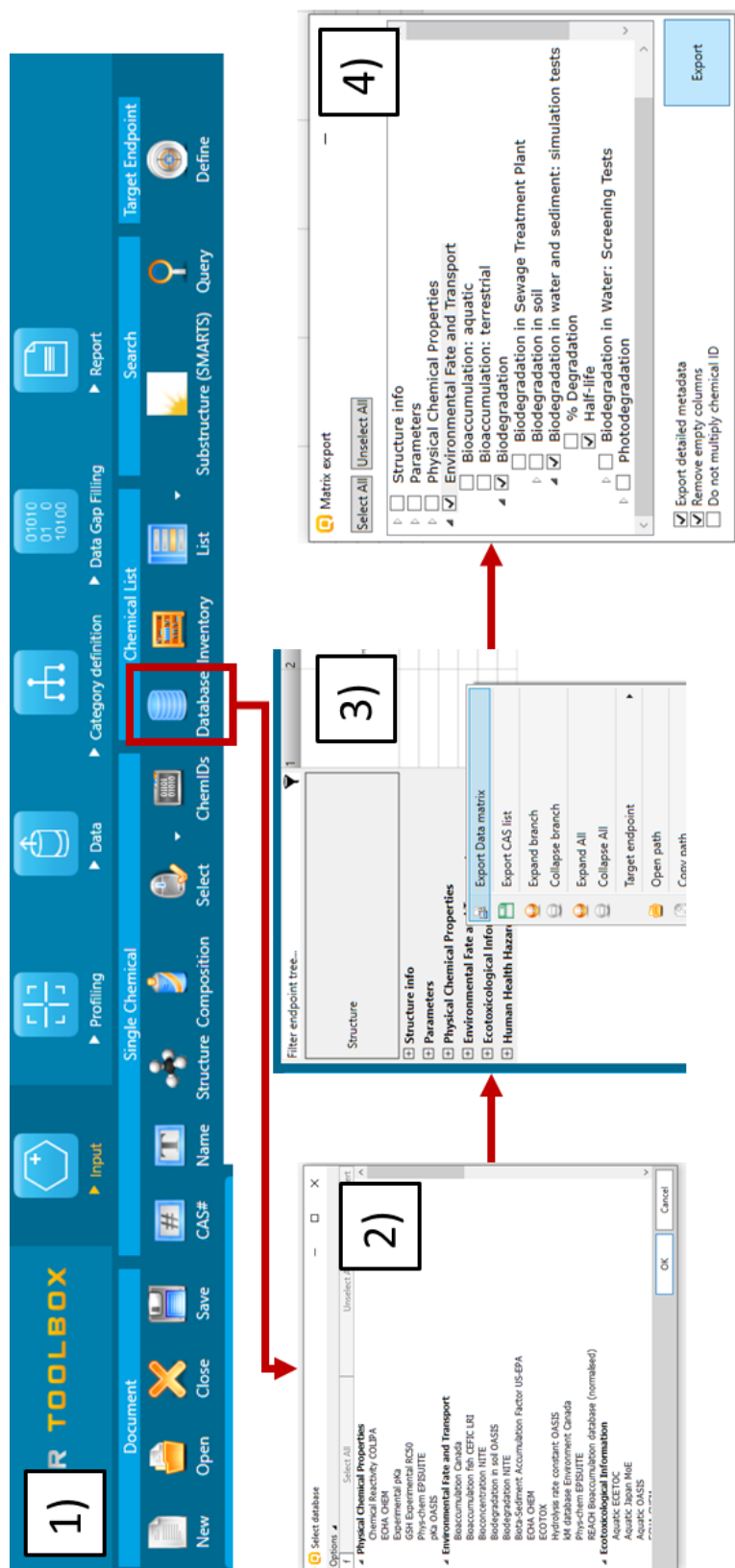
- 139 P.C. Thomas, P. Bicherel and F.J. Bauer, *How in silico and QSAR approaches can increase confidence in environmental hazard and risk assessment*, Integr. Environ. Assess. Manag. 15 (2019), pp. 40–50.
- 140 R. Boethling, *Comparison of ready biodegradation estimation methods for fragrance materials*, Sci. Total Environ. 497–498 (2014), pp. 60–67.
- 141 R. Kavlock, K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla, D. Reif, A. Richard, D. Rotroff, N. Sipes and D. Dix, *Update on EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management*, Chem Res Toxicol 7 (2012), pp. 1287–1302.
- 142 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.
- 143 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley, 2010.
- 144 R.E. Carhart, D.H. Smith, and R. Venkataraghavan, *Atom pairs as molecular features in structure-activity studies: Definition and applications*, J. Chem. Inf. Comput. Sci. 25 (1985), pp. 64–73.
- 145 R.J. Larson and C.E. Cowan, *Quantitative application of biodegradation data to environmental risk and exposure assessments*, Environ. Toxicol. Chem. 14 (1995), pp. 1433–1442.
- 146 R.R. Tice, C.P. Austin, R.J. Kavlock and J.R. Bucher, *Improving the human hazard characterization of chemicals: A Tox21 update*, Environ Health Perspect 7 (2013), pp. 756–765.
- 147 R.S. Boethling and J. Costanza, *Domain of EPI Suite biotransformation models*, SAR QSAR Environ. Res. 21 (2010), pp. 415–443.
- 148 R.S. Boethling, E. Sommer and D. DiFiore, *Designing small molecules for biodegradability*, Chem. Rev. 107 (2007), pp. 2207–2227.
- 149 R.S. Judson, F.M. Magpantay, V. Chickarmane, C. Haskell, N. Tania, J. Taylor et al., *Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor*, Toxicol. Sci. 148 (2015), pp. 137–154.
- 150 RIVM, *Pesticides: Benefaction or Pandora's Box? A synopsis of the environmental aspects of 243 pesticides*, Tech. Rep. 679101014, National Institute of Public Health and Environmental Protection, Bilthoven, NL, 1994.
- 151 S. Adler, D. Basketter, S. Creton, O. Pelkonen, J. Van Benthem, V. Zuang et al., *Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010*, Arch. Toxicol. 85 (2011), pp. 367–485.
- 152 S. Bhatia, T. Schultz, D. Roberts, J. Shen, L. Kromidas and A. Marie Api, *Comparison of cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgment*, Regul. Toxicol. Pharmacol. 71 (2015), pp. 52–62.

- 153 S. Cassani, S. Kovarich, E. Papa, P.P. Roy, L. van der Wal and P. Gramatica, *Daphnia and fish toxicity of (benzo)triazoles: Validated qsar models, and interspecies quantitative activity-activity modelling*, J. Hazard. Mater. 258–259 (2013), pp. 50–60.
- 154 S. Dimitrov, *Base-line model for identifying the bioaccumulation potential of chemicals*, SAR QSAR Environ. Res. 16 (2005), pp. 531–554.
- 155 S. Kamata, T. Matsui, N. Haga, M. Nakamura, K. Odaguchi, T. Itoh, T. Shimizu, T. Suzuki and M. Ishibashi, *Aldosterone antagonists. 2. synthesis and biological activities of 11,12-dehydropregnane derivatives*, J. Med. Chem. 30 (1987), pp. 1647–1658.
- 156 S. Kar and K. Roy, *QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors*, J. Hazard. Mater. 177 (2010), pp. 344–351.
- 157 S. Lozano, E. Lescot, M.P. Halm, A. Lepailleur, R. Bureau and S. Rault, *Prediction of acute toxicity in fish by using QSAR methods and chemical modes of action*, J. Enzyme Inhib. Med. Chem. 25 (2010), pp. 195–203.
- 158 S.J. Enoch, M. Hewitt, M.T.D. Cronin, S. Azam and J.C. Madden, *Classification of chemicals according to mechanism of aquatic toxicity: An evaluation of the implementation of the Verhaar scheme in Toxtree*, Chemosphere 73 (2008), pp. 243–248.
- 159 Simulations Plus Inc, *ADMET Predictor*, Simulations Plus Inc., Lancaster, US, 2019; software available at: <http://www.simulations-plus.com/>.
- 160 T. Guin, I. Cousins and D. Mackay, *Comparison of two methods for obtaining degradation half-lives*, Chemosphere 56 (2004), pp. 531–535.
- 161 T. Junker, A. Coors and G. Schüürmann, *Compartment-specific screening tools for persistence: Potential role and application in the regulatory context*, Integr. Environ. Assess. Manag. 15 (2019), pp. 470–481.
- 162 T. Martin, P. Harten, and D. Young, *TEST (Toxicity Estimation Software Tool) v 4.1*, US Environmental Protection Agency, Washington DC, USA, 2012; software available at: <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.
- 163 T. Sheffield and R.S. Judson, *Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure*, Environ. Sci. Technol. 53 (2019), pp. 2793–2802.
- 164 T.A. Grese, S. Cho, D.R. Finley, A.G. Godfrey, C.D. Jones, C.W.L. Iii et al., *Modifications to the 2-arylbenzothiophene core of raloxifene*, 2623 (1997), pp. 146–167.
- 165 T.I. Netzeva, A.G. Saliner and A.P. Worth, *Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory*, Environ. Toxicol. Chem. 25 (2006), pp. 1223–1230.
- 166 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.
- 167 TerraBase, *TerraQSAR biological effect programs*, TerraBase Inc., 2006; software available at: <http://www.terrabase-inc.com/>.

- 168 The European Parliament and the Council of the European Union, *Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006*, Off. J. Eur. Union 353 (2008), pp. 1–1389.
- 169 *The Globally Harmonized System (GHS) of classification and labelling of chemicals*, ST/SG/AC.10/30/Rev.4, United Nations, 2011. Available at: https://www.unece.org/fileadmin/DAM/trans/danger/publi/ghs/ghs_rev04/English/ST-SG-AC10-30-Rev4e.pdf.
- 170 *The Scientific Aspects of the Chemical Substances Control Law in Japan*. Ministry of International Trade and Industry, Japan, 1978. Available at: <https://www.sciencedirect.com/book/9780080220598/aquatic-pollutants>.
- 171 UH, *Data from: PPDB: Pesticide properties database*. Agriculture & Environment Research Unit (AERU), University of Hertfordshire [Accessed 1 January 2019],
- 172 US EPA, *Data from: ECOTOX Knowledgebase*, US Environmental Protection Agency, 2017; dataset available at <https://cfpub.epa.gov/ecotox/>.
- 173 V. Aruoja, M. Moosus, A. Kahru, M. Sihtmäe and U. Maran, Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga *Pseudokirchneriella subcapitata*, *Chemosphere* 96 (2014), pp. 23–32.
- 174 V.P. Solov'ev, A. Varnek, and G. Wipff, *Modeling of ion complexation and extraction using substructural molecular fragments*, *J. Chem. Inf. Comput. Sci.* 40 (2000), pp. 847–858.
- 175 W. Fu, A. Franco, and S. Trapp, *Methods for estimating the bioconcentration factor of ionizable organic chemicals*, *Environ. Tox. Chem.* 28 (2009), pp. 1372–1379.
- 176 X. Wu, Q. Zhang and J. Hu, *QSAR study of the acute toxicity to fathead minnow based on a large dataset*, *SAR QSAR Environ. Res.* 27 (2016), pp. 147–164.
- 177 Y. Wang, M. Zheng, J. Xiao, Y. Lu, F. Wang, J. Lu, X. Luo, W. Zhu, H. Jiang and K. Chen, *Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow*, *SAR QSAR Environ. Res.* 21 (2010), pp. 559–570.
- 178 Y.C. Wei, *PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints*, *J. Comp. Chem.* 32 (2010), pp. 1466–1474.
- 179 Y.S. Choe, P.J. Lidstroem, D.Y. Chi, T.A. Bonasera, M.J. Welch and J.A. Katzenellenbogen, *Synthesis of 11.beta.-[18f]fluoro-5.alpha.-dihydrotestosterone and 11.beta.-[18f]fluoro-19-nor-5.alpha.-dihydrotestosterone: preparation via halofluorination-reduction, receptor binding, and tissue distribution*, *J. Med. Chem.* 38 (1995), pp. 816–825.

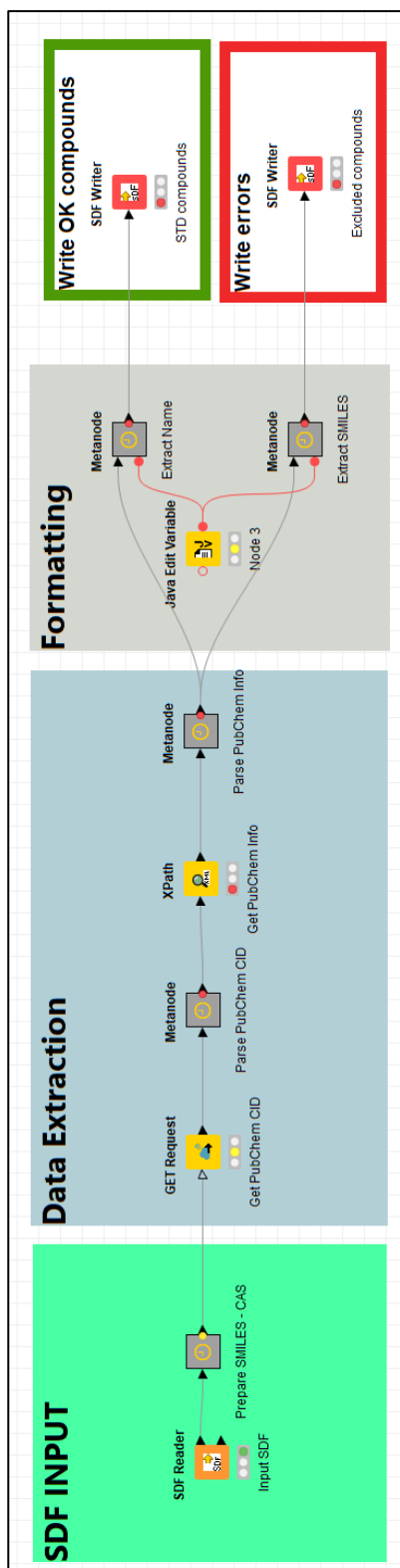
Appendix I. KNIME workflows

Appendix 1.1 – Data extraction from QSAR Toolbox



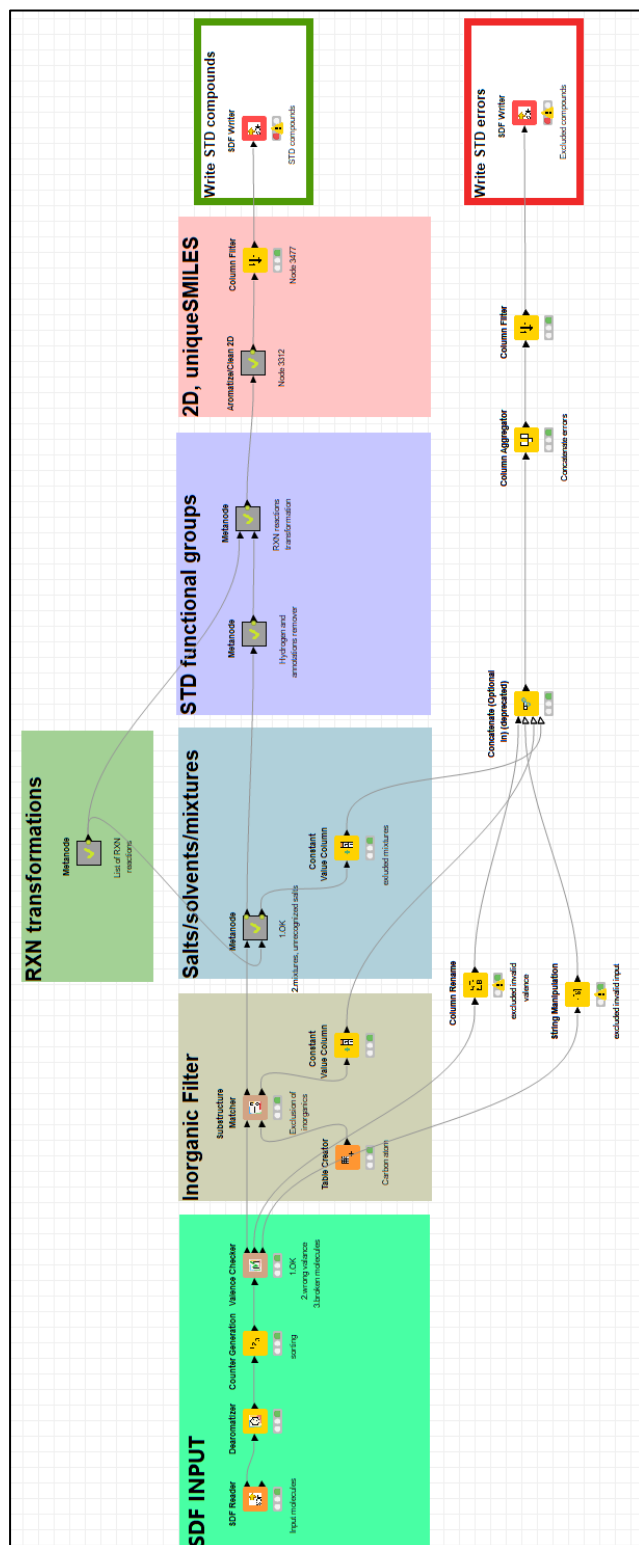
- 1) Select the *database*;
- 2) Chose the desired database to query from the list (right click and select *about* for obtaining information about concerned endpoints);
- 3) Right click anywhere under the *structure* section and select *export data matrix*;
- 4) Chose the desired endpoints and tick the option *export detailed metadata*.

Appendix 1.2 – Data extraction from Pubchem



The workflow requires an SDF file with chemical structures of query compounds and, eventually, their respective CAS number. A query is performed using the PubChem webservises in order to get SMILES and/or CAS numbers from the input structure. For each compound, the SMILES, CAS are then extracted. In addition, further general information on substance's, such as physicochemical properties and main uses, are downloaded as well. Two types of output are generated: the first lists all entries for which either SMILES or CAS numbers were found; the latter where no information was available, either due to an error or because no entry was available in PubChem.

Appendix 1.3 - Structures standardization workflow



The workflow requires an SDF file with the chemical structures to be standardized. A first check is done in order to verify whether some structures are corrupted. The first filter is the removal of inorganic compounds, salts and mixtures. Then the standardization rules, defined as list of RXN reactions, are applied, including, neutralization, removal of explicit hydrogens, aromatic representation for benzene rings, removal of stereo information and transformation of -nitro and -sulpho containing groups into canonical notation. Finally, unique SMILES are generated.