JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

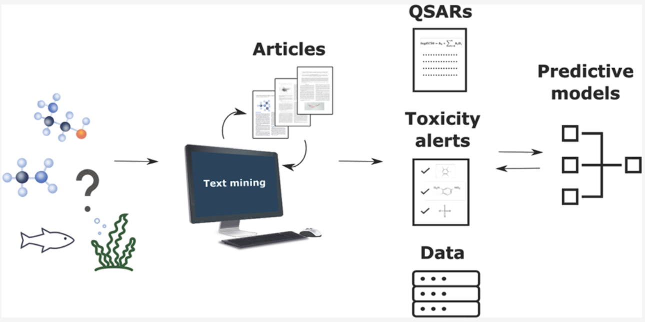
# Prior Knowledge for Predictive Modeling: The Case of Acute Aquatic Toxicity

Gulnara Shavalieva, Stavros Papadokonstantakis,* and Gregory Peters

Cite This: J. Chem. Inf. Model. 2022, 62, 4018−4031 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Early assessment of the potential impact of chemicals on health and the environment requires toxicological properties of the molecules. Predictive modeling is often used to estimate the property values *in silico* from pre-existing experimental data, which is often scarce and uncertain. One of the ways to advance the predictive modeling procedure might be the use of knowledge existing in the field. Scientific publications contain a vast amount of knowledge. However, the amount of manual work required to process the enormous volumes of information gathered in scientific articles might hinder its utilization. This work explores the opportunity of semiautomated knowledge extraction from scientific papers and investigates a few potential ways of its use for predictive modeling. The knowledge extraction and predictive modeling are applied to the field of acute aquatic toxicity. Acute aquatic toxicity is an important parameter of the safety assessment of chemicals. The extensive amount of diverse information existing in the field makes acute aquatic toxicity an attractive area for investigation of knowledge use for predictive modeling. The work demonstrates that the knowledge collection and classification procedure could be useful in hybrid modeling studies concerning the model and predictor selection, addressing data gaps, and evaluation of models' performance.

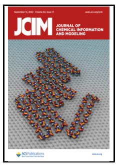


## INTRODUCTION

Environmental hazard, risk, and life-cycle (LCA) assessments of existing and newly developed chemicals for various industrial processes are highly dependent on the availability of chemical property data, which are often challenging to obtain. For instance, data on toxicological properties have been traditionally obtained through *in vivo* testing, resulting in the death of many animals and significant financial expenses.[1] From early on, it was realized that the need for experimental testing could be reduced by applying *in silico* methods assisting in the prediction of chemical property data required for the chemicals' safety assessment. *In silico* or nontesting methods for obtaining chemical property data include quantitative structure−activity relationships (QSARs), pharmacophores, and molecular modeling and data analysis tools, including machine learning (ML), data mining (DM) algorithms, and network analysis.[2] The methods are constantly improved, and new tools are developed to enhance their performance and reliability.

*In silico* approaches are often used for prescreening a vast number of chemical alternatives to select potentially better options before proceeding with more rigorous and more resource intensive approaches, typically including experimental evaluation. They are constantly gaining ground over human expertise-based brainstorming, especially in early phases of chemical process and product design. The information on the potential impact on health and the environment caused by the production and use of chemicals, including the development of new compounds, has been shown to facilitate the design of greener alternatives from early on, before their synthesis, commercialization, and use. For instance, *in silico* methods are largely used in the automated computer-aided molecular design (CAMD) in the form of predictive models of physicochemical properties based on molecular structure. To this end, property prediction techniques that describe broad classes of compounds are desired to improve computational efficiency (i.e., to avoid use of segregated information and look-up tables), often in the cost of more accurate predictions. This is however generally accepted in these early design phases, given the complexity and broadness of the task (e.g., often many thousands of chemical structures are screened). Clearly, when the CAMD algorithm compares some hundreds or thousands of various molecular alternatives, the relative information on properties is of greater importance than the exact values. Thus, the accuracy of the methods must be sufficient to result in a meaningful final list of the candidate

**Received:** September 6, 2021
**Published:** August 23, 2022



https://doi.org/10.1021/acs.jcim.1c01079

molecules.[3] While the accuracy of the thermodynamic property prediction models (e.g., boiling point, viscosity, heat capacity) is typically high, the predictions of the sustainability-related properties are subjected to a lower accuracy due to the lack of models, often as a result of lack of data required to construct the model.[4] However, despite the uncertainties introduced by the prediction models, it is beneficial to incorporate sustainability related indices during CAMD to widen the multicriteria nature of the screening,[4] rather than completely ignore these less accurate sustainability related property predictions only to perform rigorous sustainability assessment (e.g., based on more solid experimental evidence) in later phases of design for very few selected compounds. Thus, researchers apply the models but find ways to account for the prediction errors by, for example, relaxing the property constraints, running a sensitivity,[5] uncertainty,[6] or reliability[7] analysis. The most common indices incorporated into CAMD are health, safety, and environmental indicators,[7−11] computed using such properties like acute oral toxicity and permissible exposure limits, flammability and explosiveness, and aquatic toxicity and bioconcentration, respectively. A limited number of studies integrated also LCA indices. For example, Weis and Visco (2010)[11] and Heintz et al. (2014)[12] integrated a single LCA score computed by quantitative−structure property relationships (QSPRs) constructed using the data on 46 frequently used solvents. Papadopoulos et al. (2020)[4] applied a ML-based FineChem model with accuracy 20−40% to estimate the LCA values from fragments constituting the molecules. It should be clear that all these attempts to incorporate sustainability related properties into CAMD do not have the same rigor as, for instance, the one required in occupational health and safety reports or hazard and operability studies where the domain of interest is orders of magnitude smaller with respect to the number of chemicals.

The ML models show better performance and increased reliability in predicting the property values of chemicals when provided with a larger amount and better-quality training data. However, such data are often limited, especially for the newly designed chemical structures. One way to deal with insufficient training data is the development of models integrating data and scientific knowledge already existing in a certain field.[13−15] Such models use both data and field-specific information, also called prior knowledge (PK) (e.g., in the form of certain generalizations and rules from a relevant field). The integration of prior knowledge improves robustness[15] and interpretability of the model outputs.[14] The approach has been successfully applied for image recognition, weather and climate modeling, medicine, bioinformatics, etc. For instance, Diligenti et al. (2017)[16] have demonstrated an accuracy increase of a state-of-the-art deep neural network applied for image classification with the integration of prior knowledge. Faghmous and Kumar (2014)[17] and Kashinath et al. (2021)[18] have highlighted the importance of introducing the scientific theory and first-principles constraints to avoid dubious findings made by models built on large volumes of climate data. Culos et al. (2020)[19] reported improved predictions for clinically relevant outcomes when immunological knowledge was incorporated into predictive models. Xuan et al. (2019)[20] have integrated knowledge about drugs and diseases and sparse characteristics of drug-disease associations into predictive models to capture drug-related disease indications. The use of prior knowledge has also been shown to benefit the performance of chemical property prediction models. Palomba et al. (2012)[21] have developed quantitative structure−property relationship (QSPR) models estimating blood-to-liver partition coefficients (log P(liver)) for volatile organic compounds. A hybrid approach combining the ML method with descriptor selection based on expert knowledge yielded higher accuracy models. Xu et al. (2017)[22] have applied knowledge of correlated molecular activities to improve a multitask deep neural network (DNN) model's predictive performance.

In all these cases of prior knowledge incorporation into predictive models, expertise in the field is of great importance. The enormous volumes of information exist for almost every domain. Researchers have been striving to reduce the amount of manual work required to process this information,[23] which is a challenging process in itself.[24] One of the sources of the domain's prior knowledge is scientific publications. Extraction of knowledge existing in scientific articles is challenging,[25] but the need for making such knowledge more accessible to researchers and nonprofessional users is growing. For example, Zhang et al. (2019)[26] have proposed a Solution-oriented Knowledge Repository framework that provides scientific solutions mined from academic articles to the given research problems. Pandi et al. (2020)[27] have described a text-mining approach to extract pharmacogenomics associations. Guo et al. (2021)[28] have presented a method for extracting reactions from the chemical literature. Recent commercial software, IRIS,[29] an AI engine for scientific text understanding, has been developed to facilitate literature review and data extraction from scientific publications. However, to our knowledge, there are no studies examining knowledge mining for predictive modeling.

This work aims to explore knowledge existing in the field of acute aquatic toxicity in a semiautomated way and evaluates a few potential ways of its use in predictive models for initial screening of chemicals. Acute aquatic toxicity testing is an essential element of environmental hazard and risk assessments frameworks and is included in EU chemical legislation.[30] It plays a critical role in designing molecules with reduced persistency, bioaccumulation in the environment, and toxicity (PBT). There is also a need to develop models that could be integrated into the automated prescreening of chemicals in such applications as, for example, CAMD.[4] Such applications might require the computation of acute aquatic toxicity values for new chemical structures not empirically tested, thus demanding the existence of quantitative structure−activity relationship (QSAR) models with adequately populated training sets. Therefore, the development of reliable, preferably easy to interpret, *in silico* acute aquatic toxicity prediction models is required to perform an environmental hazard assessment of compounds early in the design phase. The most commonly used *in silico* approaches are QSA(P)R (from now on referred to as QSARs) and read-across (observed similarity of molecular properties between structurally similar compounds) methods. The recent models strive to improve the prediction accuracy by applying machine learning algorithms[31−33] and consensus modeling based on the prediction made by several models.[31,34] There has also been increasing interest in developing advanced data-driven methods, e.g., models taking advantage of the knowledge transfer from related tasks[35,36] to alleviate the molecular data scarcity problem.

On the one hand, the vast amount of research in the field of acute aquatic toxicity makes it an attractive topic for the development of predictive models with the use of knowledge

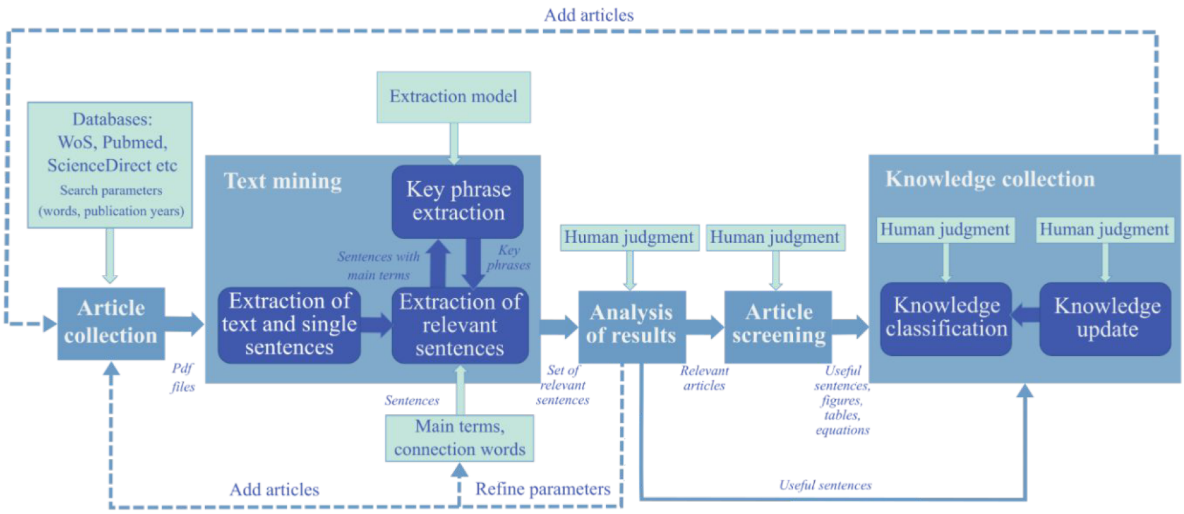


**Figure 1.** Knowledge extraction method for a specific domain.[37]

existing in the field. On the other hand, the area is more specific than a more general topic like "environmental safety". Limiting the subject to a more particular subdomain of the field might aid identification of the relevant research.

The current work aims to address the following questions: How can (semi)automated literature review accelerate knowledge mining? How can this be applied to reveal key factors influencing acute aquatic toxicity as an important safety-related metric for chemicals? In which way could the extracted knowledge be used in predictive modeling? A semiautomated knowledge extraction and the knowledge utilization methods in predictive modeling are proposed in the Methods section to answer these questions. The methods' implementation is evaluated in the Results and Discussion section for the knowledge extraction part and the assessment of various ways to use this knowledge to hybridize predictive models. The Conclusions section summarizes the main findings and potential extensions of the work.

## ■ METHODS

**Knowledge Extraction.** The method applied for knowledge extraction from scientific articles presented in Figure 1 combines automated and manual text processing.[37] The knowledge extraction starts with article collection, followed by text mining, analysis of the obtained results, additional article screening, and knowledge collection. The text mining part was automated, and the results of the automated part provided information that guided the manual processing of the extracted knowledge. The method is best applied in a specific domain; this helps guide the automated knowledge extraction and the manual process requiring human judgment.[37]

The knowledge extraction was performed on scientific articles gathered from the ScienceDirect, PubMed, and Web of Science databases. "Aquatic toxicity" and a period covering 21 years (2000−2020) were used as the search parameters.[37] Only the articles with titles related to predictive ecotoxicity, QSARs, information on the aquatic toxicity of the separate chemical classes (groups), and modes of action (MoA) were collected. Studies on inorganics, metals and metallorganic compounds, ionic liquids, epoxides, peroxides, and mixtures were excluded. The exclusion of certain groups of chemicals is a standard practice in the domain due to the software's inability to compute descriptors and/or read SMILEs (simplified molecular-input line-entry systems) of specific chemical classes. For instance, domain knowledge indicates that chemicals with rapidly degrading groups, such as peroxides and epoxides, are very reactive under environmental conditions, and it is recommended to consider the breakdown products instead.[38] The article collection step resulted in the identification of around 400 publications, which were then used for the text mining.[37] The bibliometric information on the collected articles is analyzed in Figures S1 and S2 in Supporting Information, S1.

The automated text mining (Figure 1) consisted of three main parts: extraction of article text and single sentences, key phrase extraction, and extraction of the relevant sentences.[37] First, the text was recognized and precleaned, such that the title, abstract, and references were removed, as well as extra spaces that appeared during the text recognition. The complete sentences of the text were used for the extraction of relevant sentences. The relevant sentences were identified based on the reader-provided input, namely the presence of preselected "main terms", and "connection words". The main terms included words "toxicity", "acute", "LC50", and "EC50". The following words served as the connection words (as complete words or lemmas): "increase", "decreas", "relat", "correlate", "structure", "fragment", "class", "significant", "high", "affect", "low", "link", "reason", "determin", "predict", "influence", "severe", and "depend". The text mining generated a list of relevant sentences for every article.[37] A python-based package "*knowmine*" was developed to automate the text mining step. A more detailed description of the *knowmine* package can be found in Supporting Information, S2.

The extracted set of the relevant sentences was then evaluated manually to identify useful sentences.[37] In this study, a sentence was considered useful if it contained information that could be used in predictive modeling (i.e., the sentence referred to aspects influencing acute aquatic toxicity values). The useful sentences were directly collected as knowledge or used to find articles and parts of the text for additional manual screening. The extra screening was performed to extract tables, figures, and equations as well as if the extracted sentence's information was insufficient or needed clarification for future use (i.e., predictive modeling). The information retrieved in this step was used for structuring the knowledge via the development of a classification and update scheme.[37] The information extracted from the articles steered the development of the knowledge classification and update schemes.

The update mechanism (Figure S3) was required for cases when the newly extracted information competed or complemented the previously classified information. A detailed description of the method and development of the classification scheme and update mechanism is presented in Supporting Information, S1.

The knowledge classification scheme is generic but in the present study fitted the purpose of identifying and understanding various types of knowledge of the aquatic toxicity field. It should also be remembered that the useful sentences, and thus the article screening and the knowledge collection, were set to contain information describing aspects influencing the chemicals' acute aquatic toxicity value. This means that the obtained knowledge classification scheme is intended to assist efforts in developing predictive models with the use of the extracted knowledge.

In this sense, it could also be argued that the proposed knowledge extraction procedure is perhaps better described by the term "knowledge distillation", with respect to the "purity" and loss of retrieved information and the oriented purpose of predictive modeling. However, the term knowledge extraction will be kept in the rest of the paper to describe the proposed approach as it is predominantly used in the relevant scientific literature.

**Predictive Modeling.** A complete procedure for designing and implementing a predictive model based on an ML algorithm consists of several steps: task formulation, construction, training on the data, and evaluation of the model and inference regarding its use.[39] The process does not necessarily follow such a chronological order but is rather iterative. Prior knowledge can be incorporated anywhere in this process.[39] Some generic examples of the strategies for using prior knowledge in predictive modeling are described in Supporting Information, S3.

*Data Set.*[37] The aquatic toxicity data used in this study were retrieved from the PBT (persistency, bioaccumulation potential, toxicity) database collated by Strempel (2012).[40] The original database created by Strempel (2012) contains 94,483 chemicals. Chemicals identified as inorganics, epoxides, and peroxides and those with molecular weight > 1,000 were excluded to avoid the errors encountered with prediction tools such as ECOSAR. The ECOSAR databases and the corresponding models were used to obtain the acute aquatic toxicity values for most of the chemicals in the original database. For approximately 2,000 chemicals, the toxicity data were obtained from the Aquire ECOTOX, Canadian Domestic Substance list, and EnviChem databases.[40] "The most-sensitive species" approach was followed, i.e., the lowest effect concentration with LC50 and EC50 and a duration of either 96 h for fathead minnow or 48 h for daphnia (*D. magna*) was selected. For those chemicals for which no data were available in ECOSAR (7,783 molecules), the baseline toxicity was calculated on the basis of the octanol−water partition coefficient.[37]

Due to the variability of the data sources, data quality, and the absence of an indication of the origin of every value, the data set is associated with some uncertainties and inaccuracies.[37] For instance, the reported accuracy level (i.e., when the estimated LC50 falls within the same regulatory category, high, moderate, low, no hazard, as the measured LC50) of ECOSAR predictions is only around 60%.[37,41] Moreover, this data set is characterized by "mixed data" in terms of species, duration class, study type, etc. In this study, instead of targeting homogeneous clusters of data (i.e., where biological activity is measured for all compounds under the same conditions) to formulate a set of QSARs, the modeling approach follows the general concept (i.e., not the methodology) of Sheffield and Judson,[42] namely it prioritizes having a large amount of data over having a pure data set, based on the assumption that chemical structure is the principal driver of end point variation. This approach avoids limiting the training data by experiment type and leaving the decision for the application and extension of the models to the user; instead it incorporates all commonly available data and "allows the model to adjust its predictions accordingly".[42]

Another type of applying mixed data for toxicity inference of chemical substances is proposed by the "Guidance on Information Requirements and Chemical Safety Assessment"[43] published by the European Chemicals Agency, where short-term aquatic toxicity data based on acute test or QSARs can be used for the screening of molecules. The substance is assessed as very toxic if L(E)C50 for algae, daphnia, or fish < 0.01 mg/L. In this case, a definitive conclusion can be drawn that the substance fulfills the T (Toxic) criterion without further testing. If EC50 or LC50 < 0.1 mg/L, the substance is considered as a Potential T (PT) candidate. If EC50 or LC50 ≥ 0.1 mg/L, long-term or chronic aquatic toxicity data is required for a more definitive assessment. Therefore, despite the mixed species approach and uncertainties, the current data set is suitable for use in the study to evaluate in which ways knowledge can be applied to improve the first screening of the substances (Toxic, Potentially Toxic, or Not Toxic) by means of predictive models.[43]

Only saturated aliphatic compounds that contain C, O, H, and N were considered for the study. This reduced data set of 2106 molecules enabled a decrease in the computational time, while testing several ways of prior knowledge use in predictive modeling. The data set is provided as a separate Excel file in the Supporting Information.

*Predictive Models.* In this work, a k-nearest neighbors (kNN) ML approach was chosen as an exemplary algorithm for constructing predictive models with and without prior knowledge. Models that were constructed without using prior knowledge were designated as "standard" models, and the models that did utilize prior knowledge were termed "hybrid" models.[37] The standard models were developed for purposes of comparison, to evaluate the impact of the knowledge incorporation on the performances of the models. The hybrid models used the extracted knowledge that was relevant to the data set (e.g., concerning the toxicities of aliphatic compounds for fish and crustaceans).[37]

The kNN is an easy-to-implement algorithm without a highly intensive training procedure. This enables quicker testing of different ways of knowledge use in predictive modeling. The kNN estimates a missing property value using the molecules that are structurally most-similar (nearest neighbors) with the known property values.[37] The nearest neighbors were identified in two ways: (i) a Tanimoto similarity[44] between molecular fingerprints and (ii) the Manhattan distance[44] between the molecular descriptor vectors (93 descriptors) representing the chemicals.[37] The molecular fingerprints and descriptors were computed with the help of the open-source cheminformatics tools RDKit[45] and PaDELPy.[46] The optimal number of neighbors was determined

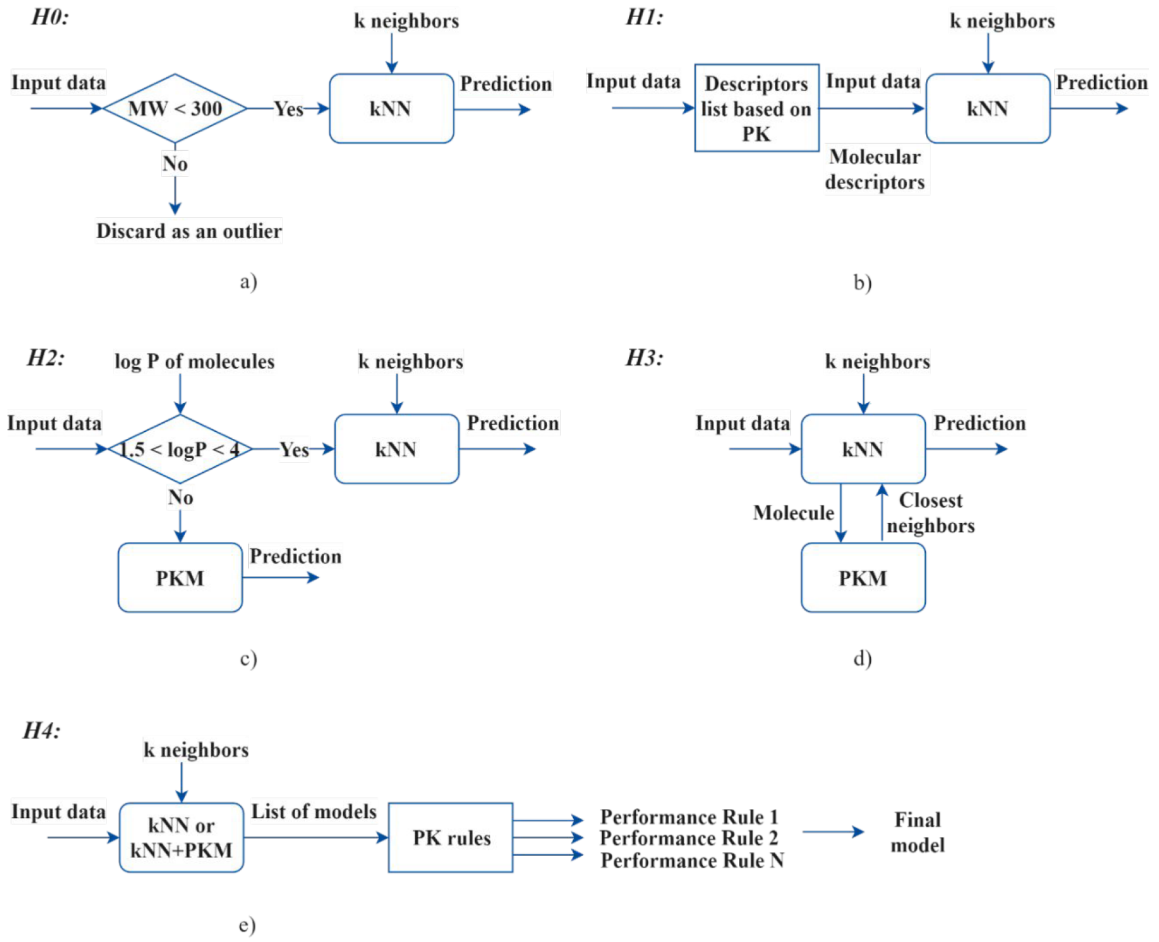


**Figure 2.** Schematic representation of the use of prior knowledge in the development of hybrid models: PK, prior knowledge; PKM, prior knowledge model (GC+QSAR)[37]

by a cross-validation (CV) procedure for validation to training data ratios ranging from 5 to 30%. Out of the numerous CV runs the number of neighbors leading to the highest performance was selected for the final standard models: descriptor (DESC with the number of neighbors equal to 2, 4, 5, 6, and 8) and fingerprint-based (FPN with 2, 5, 7, 12, and 14 neighbors). More details on the development of the standard models can be found in Supporting Information S3.

Prior knowledge was applied before, during, and after the kNN algorithm approach, according to the different schemes[37] presented in Figure 2. For hybrid model H0 (Figure 2a), a concept of outlier detection based on single descriptor molecular weight (MW) was tested. The hypothesis behind this concept implied the presence of shared outliers between most of the descriptors used for the prediction. For instance, according to prior knowledge, the toxicity of a chemical increases with an increase in its MW. As seen in Figure 3, a correlation between the MW and toxicity values can be observed. However, the correlation is weak for most of the molecules with MW > 300 g/mol. In this way, 187 molecules were removed from the data set as outliers.[37] A more detailed analysis of the outliers and the tested hypothesis is presented in Supporting Information, S4.

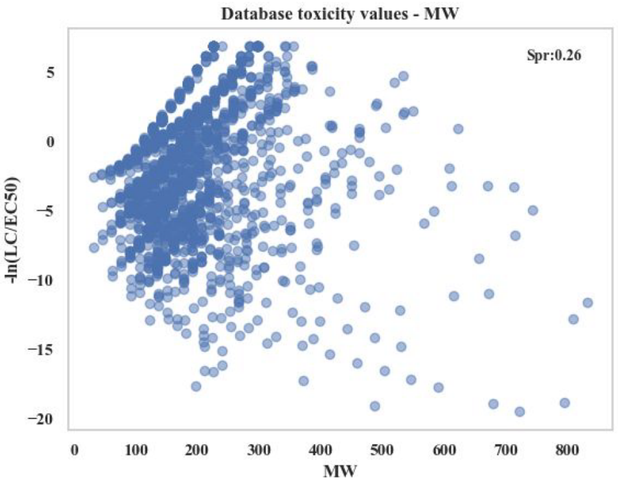


**Figure 3.** Correlation of data set toxicity (−ln(LC50/EC50)) values with molecular weight.

For hybrid model H1 (Figure 2b), a descriptor (predictor) selection was performed, such that from a list of potential molecular descriptors, those descriptors identified by prior knowledge as having a high influence on aquatic toxicity were selected to represent the molecular structures.[37] It should be noted that these descriptors not only refer to descriptors used in previous QSAR modeling attempts but also incorporate information based on qualitative knowledge extraction. It should also be noted that this hybridization technique could generally involve a more thorough search for subsets of descriptors toward optimality, which however lies outside the

Table 1. Set of Rules for Evaluation of the Performances of the Models[37]

| main toxicity trends | expressed in descriptors |
|---|---|
| Toxicity increases with hydrophobicity.[49,50] | Toxicity increases with an increase of MolLogP (RDkit). |
| Toxicity increases with polarizability.[31,50,51] | Toxicity increases with an increase of molar refractivity MR (RDkit). |
| | Toxicity decreases with an increase of GATS 1p (PaDELPy). |
| | Toxicity increases with an increase of AATSC0p (PaDELPy). |
| Toxicity has a negative correlation with topological polar surface area.[50,52] | Toxicity decreases with an increase in TPSA (RDkit). |
| Most of the toxic compounds act as hydrogen-bonding acceptors, while the least toxic compounds act mainly as hydrogen-bonding donors.[53,54] | Toxic compounds have lower SHBd (PaDELPy). |
| | Toxic compounds have lower maxHBint2 (PaDELPy). |
| There is a positive effect of unsaturation and electronegative atom count.[55] | Toxicity decreases with an increase of ETA_dEpsilon_A (PaDELPy). |
| Toxicity decreases with increase in ionization potential.[31,51] | Toxicity decreases when Mi (PaDELPy) increases. |
| | The larger the "GATS1i" (PaDELPy), the less likely the compound will be to react and generate toxicity. |
| Molecular size and bulk have positive influences on toxicity.[34,50,55,56] | With an increase of MW (RDkit), the toxicity increases. |
| | Toxicity is higher for higher values of ETA_Alpha (PaDELPy). |
| There is an inverse effect of branching on toxicity.[50,52,55,57] | Toxicity decreases with an increase of ETA_EtaP_B (PaDELPy). |
| Toxicities of primary, secondary, and dimethyl tertiary amines increase with increasing chain length.[58] | Toxicity of molecules containing N or amine group increases if the number of carbon atoms increases. |
| Toxicity increases with increasing alkyl chain length in ethoxylates.[59] | Toxicity of molecules containing the methoxy group increases if the number of carbon atoms increases. |
| Substitution of H atom with a methyl group (−CH3) on the N atom reduces the toxicity of amine surfactants.[60] | The toxicity of molecules decreases with the number of N−CH3 fragments. |

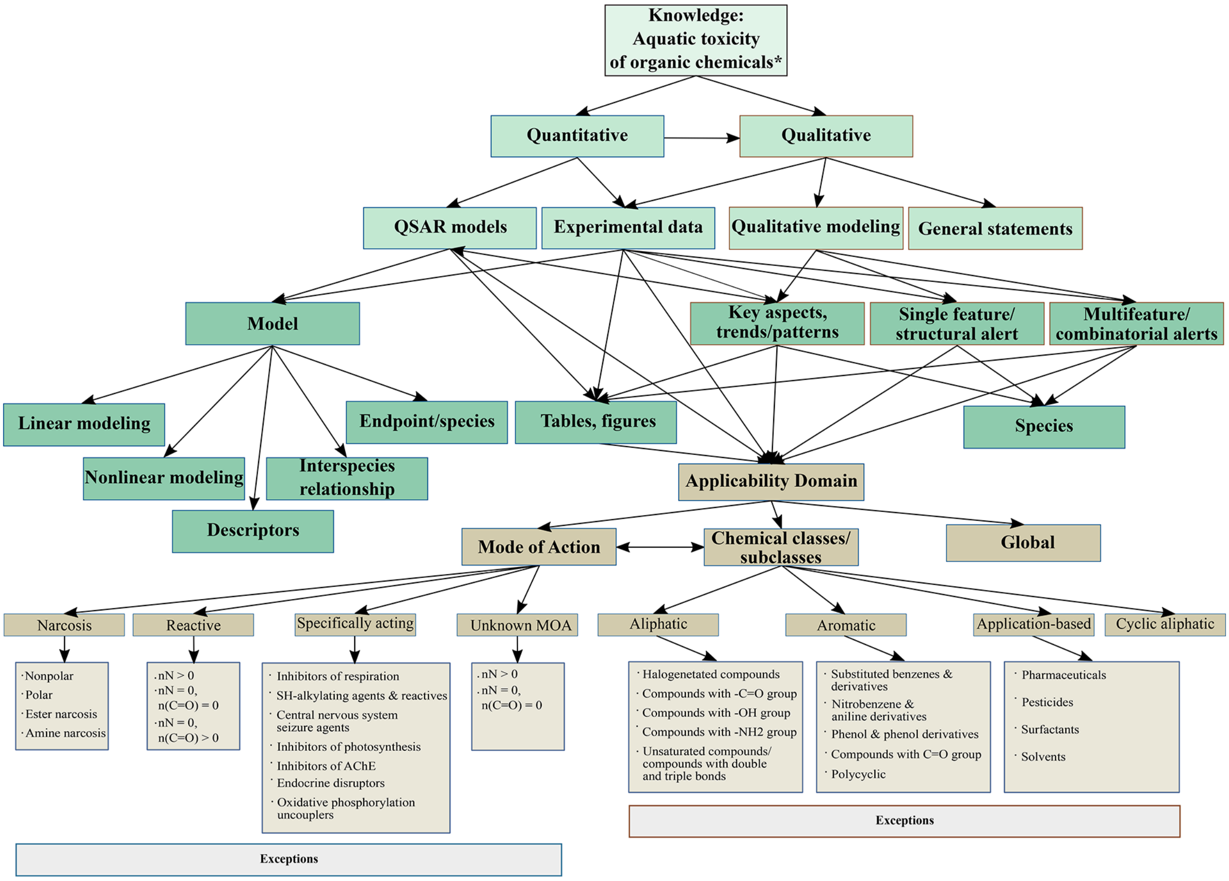

Knowledge:
Aquatic toxicity
of organic chemicals*
Quantitative
Qualitative
QSAR models
Experimental data
Qualitative modeling
General statements
Model
Key aspects, trends/patterns
Single feature/ structural alert
Multifeature/ combinatorial alerts
Linear modeling
Endpoint/species
Tables, figures
Species
Nonlinear modeling
Interspecies relationship
Applicability Domain
Descriptors
Mode of Action
Chemical classes/ subclasses
Global
Narcosis
Reactive
Specifically acting
Unknown MOA
Aliphatic
Aromatic
Application-based
Cyclic aliphatic
·Nonpolar
·Polar
·Ester narcosis
·Amine narcosis
.nN > 0
·nN = 0, n(C=O) = 0
·nN = 0, n(C=O) > 0
· Inhibitors of respiration
· SH-alkylating agents & reactives
· Central nervous system seizure agents
· Inhibitors of photosynthesis
· Inhibitors of AChE
· Endocrine disruptors
· Oxidative phosphorylation uncouplers
.nN > 0
.nN = 0, n(C=O) = 0
· Halogenetated compounds
· Compounds with -C=O group
· Compounds with -OH group
· Compounds with -NH2 group
· Unsaturated compounds/ compounds with double and triple bonds
· Substituted benzenes & derivatives
· Nitrobenzene & aniline derivatives
· Phenol & phenol derivatives
· Compounds with C=O group
· Polycyclic
· Pharmaceuticals
· Pesticides
· Surfactants
· Solvents
Exceptions
Exceptions

**Figure 4.** Knowledge classification scheme for aquatic toxicities of chemicals (*excluding inorganics, metals and metalloorganic compounds, ionic liquids, epoxides, peroxides, and mixtures).[37]

scope of this work which aims to exemplify the potential of the knowledge extraction in hybridization methods. In this way, the descriptors "LogP", "AATSC0p", "TPSA", "ETA_dEpsilon_A", "SHBd", and "Mi" were selected to represent some of the most frequently mentioned toxicity trends. The trends are relevant for the data set molecules and exhibit a good correlation with the toxicity values (Table 1). In the case of H2 hybridization (Figure 2c), the toxicity of the molecules was predicted using either a kNN algorithm or a prior knowledge-based model (PKM), depending on the value of octanol−water partition coefficients of the molecules log P (i.e., estimated by RDkit[45]). For molecules with log P between 1.5 and 4.0 (811

https://doi.org/10.1021/acs.jcim.1c01079

**Table 2. Examples of the Quantitative Knowledge Extracted as Part of the Classification Scheme**

| applicability domain | model type | end point | descriptors | performance[a] | ref |
|---|---|---|---|---|---|
| global | linear modeling (MLR+GA) | −logLC50 *Pimephales* promelas | AlogP, $E_{LUMO}$, S2K, nRNH2 | $n = 408$, $R^2 = 0.80$<br>$Q^2_{LOO} = 0.80$, $Q^2_{Boostr} = 0.80$, $Q^2_{ext} = 0.72$ | Pavan et al. (2006)[61] |
| MoA, specifically acting chemicals | linear modeling | log(1/LC50) *Poecilia* reticulata | $E_a$(max), $\sum C_a$, Nv1 | $n = 31$, $R^2 = 0.77$ | Raevsky et al. (2009)[62] |

[a] Selected performance indicators from the respective original article.

molecules), the kNN fingerprint-based model was used for estimation of their toxicity. The rest of the toxicity values were predicted by the PKM combining a GC[47] model for fish fathead minnow with an interspecies equation for daphnia (*D. magna*) fish[48] toxicity. The daphnia model was chosen as a better fit for the data set since daphnia is generally a more sensitive species than fish (See Data set subchapter). The specific range of log P was selected after the analysis of the standard models' performance revealing larger prediction errors for molecules with log P in the 1.5−4.0 range. The selected GC approach has been shown to sufficiently cover both areas log P < 1 and log P > 4 based on the list of the used compounds for the regression.[37]

The daphnia PKM model has also been used for the development of the H3 hybrid models.[37] The models (Figure 2d) applied the PKM model to assist the kNN algorithm in selecting the closest neighbors; molecules with the closest estimations (smallest difference between the predictions) to those obtained by the PKM model were considered the nearest neighbors. Prior knowledge in the form of rules was also applied as a postassessment method for screening the developed models (H4, Figure 2e) to evaluate their performance.[37] The rules are presented in Table 1, and full names of the descriptors introduced in the rules can be found in Table S16.

## ■ RESULTS AND DISCUSSION

This section is divided into two parts. The first part briefly introduces the knowledge extraction and classification results, and the second part presents the usage of prior knowledge for predictive modeling.

**1. Knowledge Extraction.** The main advantage of the partly automated literature review was a significant reduction (Table S1, Supporting Information) of the text for initial reading (>85%).[37] Most of the sentences extracted by the automated text-mining procedure were useful, in that they could be used for predictive modeling or pointed out specific parts of the initial article for the subsequent manual screening. The method did not seem to require extensive knowledge of the field, as only some prior understanding was needed to define the main terms and connection words that would guide the search for relevant information.[37] However, it might be useful to run a sensitivity analysis on a limited number of articles to adjust the search parameters if there is no field expertise or it is wished to limit the amount of the extracted information further.

The knowledge collected from the scientific articles could be classified under two main categories: quantitative and qualitative information (Figure 4). The quantitative category comprises quantitative structure−activity relationship (QSAR) models and experimental data. The QSAR models differ by the type of modeling (linear, nonlinear) and descriptors used to develop the models. The QSARs define the organism-aquatic toxicity end point relationship for certain species or establish the correlation between end points of two different species (interspecies quantitative activity−activity relationships (QAARs)). The qualitative category contains qualitative modeling and general statements. Qualitative modeling comprises information on key toxicity aspects, trends, patterns, and single feature/structural or multifeature/combinatorial alerts. Compared to other categories, the general statements do not give detailed information about the descriptors or models but make a more generic description of the toxicological properties, models, or data quality. Qualitative and quantitative information could also be presented or analyzed in graphs, figures, and tables. Tables 2 and 3 present examples of prior knowledge collected under quantitative and qualitative categories. The knowledge collected for each category of the classification scheme depicted in Figure 4 is discussed in the following paragraphs in more detail.

**Table 3. Examples of the Qualitative Knowledge Extracted as Part of the Classification Scheme**

| applicability domain | species, end point | extracted knowledge | ref |
|---|---|---|---|
| substituted benzenes | *Tetrahymena pyriformis*<br>pIGC50 | positive correlation with end point:<br>- MW,<br>- nAtomP,<br>- TopoPSA<br>negative correlation<br>- SHdsCH,<br>lipoaffinity index | Gupta et al. (2015)[63] |
| pharmaceuticals | *D. magna*, fish<br>LC50 | higher toxicity to *D. magna*:<br>- keto group<br>- aasC fragment<br>higher toxicity to fish:<br>- keto group,<br>- X=C=X fragment,<br>- R−C(=X)−X fragment,<br>R−C≡X fragment | Kar et al. (2018)[64] |

*QSAR Models.* The QSAR models collected during the knowledge acquisition were developed for prediction of the toxicity values or classification of chemicals according to different toxicity levels and MoA classes.[37] Most of the QSARs applied linear modeling (e.g., multilinear regression, principal component analysis, linear partial least-squares, ordinary least-square method) due to its simplicity and interpretability. The nonlinear models often exhibited higher levels of accuracy than the linear models built using the same set of chemicals. A general outcome from the QSAR studies was that ensemble or consensus models that combined several methods outperformed the models based on a single method.[65−68] Improved performance was also observed when similar chemicals were grouped based on MoA[69] or other similarity criteria[69,70] before developing the prediction model.[37]

A wide variety of descriptors were used in the collected QSAR models.[37] The descriptors with the highest impact on the acute toxicity value were associated with hydrophobic features (i.e., log Kow, logP, logD, Crippen logP, B08[C−C]), electrophilicity (i.e., ELUMO, Amax), polarizability (i.e., $\alpha$, GATS 1p), acceptors and donors of hydrogen bonds (i.e., Ca, NHdon Hacc, polar groups descriptors), molecular size and branching (i.e., Vm, ElipVol, RDCHI), and polar surface area (i.e., TPSA). According to Gramatica et al. (2018),[71] the nX (number of halogen atoms) and nBondsM (number of multiple bonds) descriptors, which are related to halogen substitution and unsaturation, increase the PBT behaviors of chemicals. The descriptors linked to the decrease in the PBT index were MAXDP2 (maximal electrotopological positive variation) and nHDonLipinski (number of donor atoms for H bonds). These two descriptors encode the ability of a chemical to form electrostatic and dipole−dipole interactions, as well as hydrogen bonds in the surrounding environment. Additionally, descriptors connected to the ionization potential were reported as an important parameter influencing the toxicity of the compounds.[31,72−74] Ionization was shown to affect the biouptake and mechanisms of interaction with the macromolecule at the target sites.[72,73] Hossain and Roy (2018)[75] and Önlü and Saçan (2018)[76] have developed QSAR models for Contaminants of Emerging Concern (CECs), including for instance pharmaceuticals, personal care products, pesticides, and surfactants. They have determined that the toxicities of CECs are mostly related to hydrophobicity,[75,76] aromaticity,[75] polarizability, and molecular size and shape.[76]

Various species were used to obtain toxicity information.[37] The most commonly used species in the QSAR and other types of studies were the algal *Tetrahymena pyriformis (T. pyriformis)* (IGC50), crustacean *Daphnia magna (D. magna)* (EC50, LC50), and fish *Pimephales promelas* (fathead minnow) (LC50). The interspecies QSARs can be regarded as a separate class of the QSAR models. These models are typically based on a small volume of data and have a linear functional form with a few predictor variables.[37] The fish-based model was recognized to be superior for predicting lacking toxicity data (i.e., for *T. pyriformis* and *D. magna*).[48] The collected QSA(A)Rs, including details on the corresponding models and symbols, can be found in Tables S3−S7.

*Sensitivity of Species.* The smaller species like bacteria, algae, or crustaceans were found to be more sensitive than fish. However, the sensitivity of species varied depending on the type of chemicals they were exposed to. For instance, *Vibrio fisheri* was sensitive to parabens,[77] nitrates,[78] benzoic acids,[79] and alkoxy-substituted benzenes.[78] *Chlorella vulgaris* was very sensitive to haloalkanes,[80] while *Pseudokirchneriella subcapitata (P. subcapitata)* showed increased sensitivity to nitriles[81] and, in general, to organic pollutants.[82] High sensitivity to aromatic amines and highly lipophilic compounds[34] was observed for daphnids.[83,84] The skin and lipid content of multicellular organisms like daphnia and fish could prevent the biouptake for ionizable compounds; thus, the toxicity effect would be decreased.[51]

The algal *T. pyriformis* showed less sensitivity than other species, which might imply less experimental uncertainty of toxicity data available for this species.[85] Although fish is frequently used for tests, Rawlings et al. (2019)[86] argue that it is advisible to invest in algal and daphnids testing resulting in more conservative predictions than any fish.

*Identified Toxicity Alerts, Trends, and Patterns.* The collected knowledge[37] suggests a consensus among researchers that acute toxicity is defined by the mode of toxicological action and the chemical characteristics.[87] The higher toxicity values have often been associated with increased lipophilicity.[50,88,89] The most toxic compounds were hydrophobic and acted as hydrogen-bonding or electron acceptors (e.g., hydrophobic nitroaromatic compounds with halogen and amino substituents[52,53,81,90,91]). Khan et al. (2019)[92] have advised that if a hydrophobic group is necessary during the design of a drug compound, a higher polarity substitution should be preferred. Voutchkova et al. (2011)[93] have suggested keeping logPo/w < 2 and $\Delta E$ (LUMO−HOMO) > 9 eV to increase the likelihood of designing a compound with low aquatic toxicity.[37]

Specific functional groups, such as cyano,[87] isothiocyanate,[94] and halogens[88,95−97] enhance the toxicities of molecules.[37] However, the extent of the increase appears to be dependent upon the molecular structure and position of the group in the molecule. Among the other reported toxicity alerts were amino groups, the presence of additional (one or more) aromatic rings with highly electronegative substituents close to each other (5−7 Å apart),[98] nitro group, nitrile, disulfide, phosphoric acid derivatives, pyrazolyl group, and formamide groups,[99] ring aromaticity, sulfur, aromatic esters, and vinyl moiety,[52] double and triple bonds, and acrylate groups,[100] to name the most frequently encountered.[37] Table S2 of the Supporting Information presents examples of molecular features reported to increase or decrease toxicity. The extended version of the table containing the information collected under the qualitative category is available on request from the corresponding author.

*Applicability Domain.* Affinity for a specific chemical class or MoA was often seen as a critical determinant for predicting and understanding chemical toxicity,[101−104] with MoA being more challenging to determine.[37,105] The most covered applicability domain in MoA seemed to be nonpolar and polar narcosis, followed by specifically acting chemicals.[37] The chemical classes that were most highly represented in the collected knowledge base were nitrobenzene and phenol derivatives, pesticides, pharmaceuticals, and halogenated aliphatics. Other chemical classes such as aliphatic alcohols, amines, amides, and acids were represented to a lesser extent, probably because their toxicity effects are instead studied in the context of a particular MoA. Compounds with double and triple bonds, such as vinyl/allyl group-containing chemicals, nitriles, propargyl alcohols, carbonyl-containing $\alpha,\beta$-unsaturated chemicals, carbamates, and quinones, have often been examined separately, likely because of their reactive nature.[57,97,106−108] Despite the clear benefits of assigning compounds to certain chemical classes or MoAs, many researchers strive to develop "global" models[37] that were not limited by chemical class or MoA.[109]

The academic field of aquatic toxicity is diverse and extensive, both from the quantitative and qualitative perspectives.[37] On the one hand, this diversity might foster the identification of relevant information, which could be used for predictive modeling. For example, the extracted knowledge can guide predictor (variable) selection and prioritization. Moreover, the experimental data collected from various studies can be used for training or external validation of the developed models. Interspecies correlations could help to close the data gaps present for data sets of certain species. Information on the species' sensitivity and outliers discovered during the construction of the QSAR models could be used to explain the results or observed deviations in the predictions. The QSAR equations and alerts could be directly integrated into the training phase of the data science models as additional constraints or applied for model refinement. The discovered aquatic toxicity trends and patterns could be helpful for the analysis of the results obtained by the developed models, thus contributing to evaluation and selection of the optimal models. On the other hand, the wide variety of descriptors used in the studies, different quality of the toxicity data, applicability domain limitations, to name some important factors, make it quite challenging to apply directly the knowledge without analyzing the available information and constraints associated with its use. Thus, mapping and evaluating domain knowledge before its application could be helpful to facilitate navigation of the data.[37] The results of some strategies for using prior knowledge in predictive modeling are presented in the next section.

**2. Predictive Modeling.** The summary of the performances of the final standard and hybrid models can be seen in Table 4. It is evident that the hybridization improves the coefficient of determination ($R^2$) and Spearman's correlation coefficient (Spr_m) of the standard descriptor-based models.[37] The fingerprint-based models (both standard and hybrid) show lower values of $R^2$ and Spr_m. Additionally, classification metrics were applied for the performance analysis. For the classification metrics, the data set toxicity values and all the predictions made by the models were divided into three categories: T, PT, and NT depending on the values (<0.01 mg/L, <0.1 mg/L, and ≥0.1 mg/L respectively). Then the classified predictions were assessed via balanced accuracy, recall, and precision. All the hybrid models except H3 improve the results of the classification, compared to the standard models' performance. It should, however, be considered that the hybridization of the descriptor-based models was performed using the same data set as for the standard models. The QSAR-based model used for hybridizing the fingerprint-based models was built on a different set of data, which to some extent explains the lower performance. The method, however, might be used as a way to analyze and deal with the uncertainty of the data set. Moreover, there may be a trade-off between the level of prediction and classification accuracies for a specific data set and qualitative assessment of a model based on the identified prior knowledge trends, as will be discussed next.

Table 4. Summary of the Performances of the Models[a]

| model | $R^2$ | Spr_m | accuracy | precision | recall |
|---|---|---|---|---|---|
| DESC_2 | 0.83 | 0.94 | 0.87 | 0.96 | 0.96 |
| DESC_4 | 0.85 | 0.95 | 0.87 | 0.96 | 0.96 |
| DESC_5 | 0.85 | 0.95 | 0.88 | 0.97 | 0.97 |
| DESC _6 | 0.86 | 0.95 | 0.87 | 0.96 | 0.96 |
| DESC _8 | 0.86 | 0.95 | 0.85 | 0.96 | 0.96 |
| DESC_H0_3 | **0.95** | **0.98** | **0.91** | **0.98** | **0.98** |
| DESC_H0_7 | **0.95** | **0.98** | **0.90** | **0.97** | **0.97** |
| DESC_H1_2 | **0.92** | **0.98** | **0.91** | **0.98** | **0.98** |
| FPN_2 | 0.70 | 0.84 | 0.56 | 0.85 | 0.84 |
| FPN_5 | 0.74 | 0.86 | 0.53 | 0.84 | 0.85 |
| FPN_7 | 0.74 | 0.86 | 0.50 | 0.83 | 0.84 |
| FPN_12 | 0.73 | 0.86 | 0.51 | 0.84 | 0.85 |
| FPN_14 | 0.73 | 0.86 | 0.50 | 0.84 | 0.85 |
| FPN_H2_2 | 0.43 | 0.80 | **0.74** | **0.90** | **0.87** |
| FPN_H2_7 | 0.45 | 0.80 | **0.74** | **0.90** | **0.88** |
| FPN_H2_12 | 0.46 | 0.80 | **0.74** | **0.90** | **0.88** |
| FPN_H3_12 | 0.52 | 0.76 | 0.54 | 0.87 | 0.88 |
| FPN_H3_14 | 0.53 | 0.76 | 0.53 | 0.86 | 0.88 |

[a] DESC and FPN specify the descriptor-based and fingerprint-based models, respectively. H0, H1, H2, and H3 are the applied types of hybridization models. The designations _2 to _14 indicate the numbers of closest neighbors used for the prediction. The classification metrics uses three labels (T, PT, and NT) and considers the label imbalance.

The heatmap in Figure 5 depicts how closely the predicted toxicity values follow the rules presented in Table 1. It is clear that some of the rules are better followed by the data set toxicity values (Data) than others. The rules related to maxHBint2 (hydrogen bonding) MW, ETA_EtaP_B (branching), and the methoxy group (nC methoxy) are supported to a lesser extent.[37] Poor performance of the data set regarding the methoxy group-containing molecules might imply a deviant behavior of these molecules compared to the remainder of the data set. The low diversity of the data set in terms of branching might be the reason for a very low correlation of toxicity values with the ETA_EtaP_B descriptor. The predictions made by the standard descriptor-based models (DESC_2 and DESC_8) (except nC methoxy) show similar trends to the Data. The fingerprint-based models (FPN_2 and FPN_7) show lower levels of compliance with the rules than the descriptor-based models and the Data.[37]

The values for the hybrid models often show a better correlation with the rules.[37] Interestingly, the highest variation between the results of the standard and the hybrid models is observed for the fingerprint-based models, which are the models with the lower prediction accuracy compared to the standard or descriptor-based hybrid models. Since different data and model parameters were used to develop the prior knowledge model, the hybrid models reveal trends not observed in the data set. A much higher influence of molecular size (MW, ETA_Alpha, partly MR) on the toxicity is noticed for the hybrid models. The toxicity values predicted by the models H2 and H3 exhibit better correlations with the increasing chain length in amines and methoxy group-containing molecules. However, models H2 and H3 seem to be inferior in considering the electronegativity and topology of molecules (lower correlation for GATS 1p, TPSA, GATS1i, ETA_dEpsilon_A) compared to the standard models. The models H2 and H3 do not capture the influence of N−CH3 fragments on the toxicity as well, probably due to the absence of this fragment's toxicity contribution in the PKM GC method. The performance of H1 models is similar to or slightly lower than the rest of the descriptor-based models and Data. This suggests that a limited number of descriptors can fully represent the molecules in the data set.[37]

The analysis (Supporting Information, S4) of the outliers identified on the basis of MW (model H0) showed that these molecules also differ in logP, MR, TPSA, and ETA_Alpha values compared to the rest of the data set. The set of the outliers seems to consist of highly hydrophobic molecules exhibiting high toxicity and the low toxicity compounds with the reduced ability to permeate the cells of the living organisms due to their larger TPSA and size.

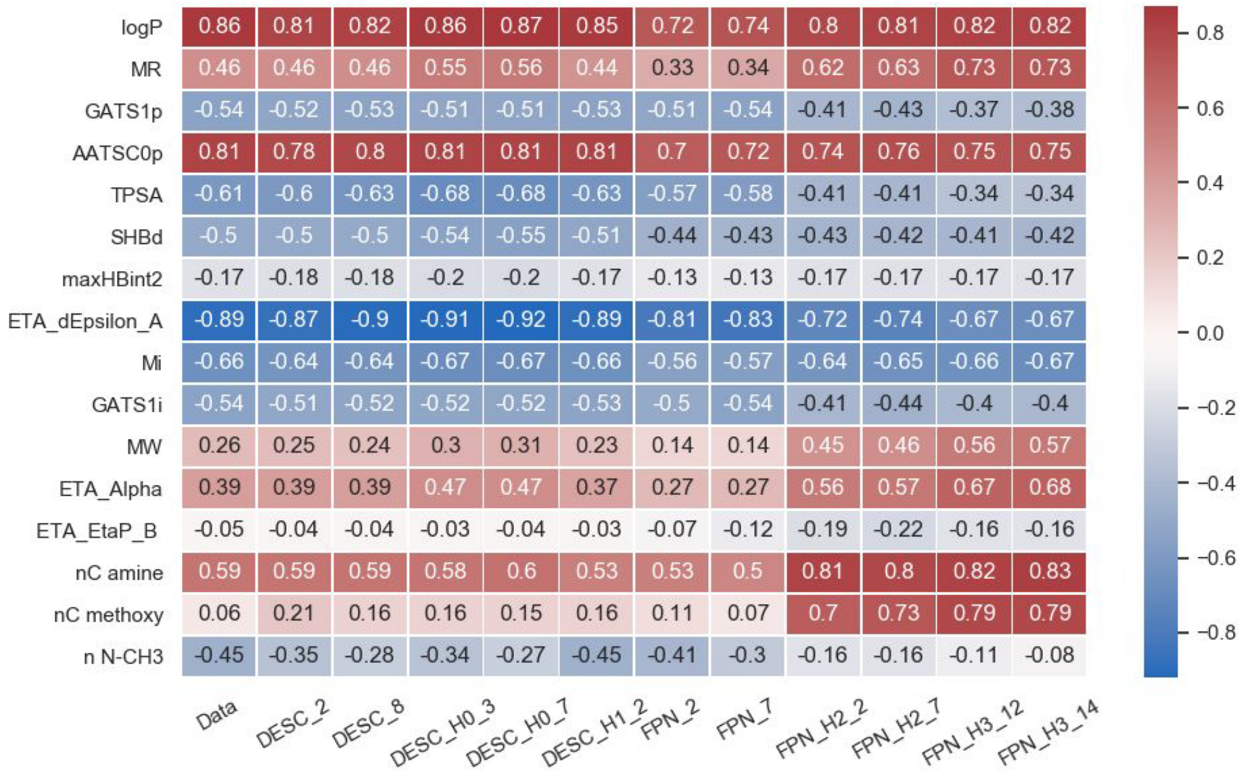


**Figure 5.** Spearman's correlation coefficient between the descriptors (Table 1) and toxicity predictions made by the models. Red: positive correlation with toxicity, blue: negative. Only two descriptor and fingerprint-based models (best and worst) are shown due to the similar performance of the rest of the standard models to the presented ones.[37]

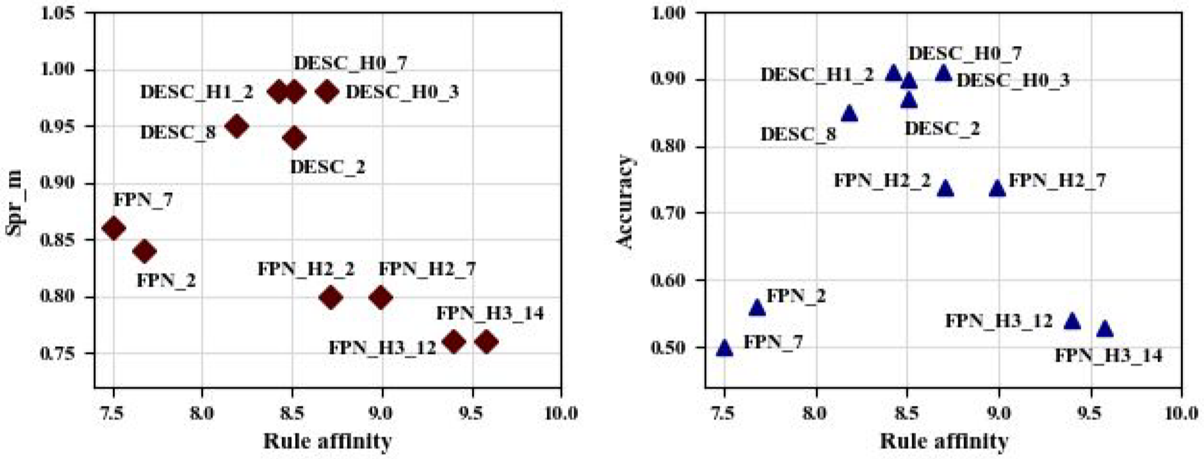


**Figure 6.** Spearman correlation coefficient (Spr_m) (left) and accuracy scores (right) vs Rule affinity for the standard and hybrid models.

Figure 6 presents the performances of the models in terms of the Spearman correlation coefficients, classification accuracies, and overall compliance of the models with the prior knowledge-based rules (Table 1 and Figure 5). The compliance with the rules, or rule affinity, was computed by first min-max normalization of the correlation coefficients within the same descriptor category followed by the summing of the normalized values over all the descriptors. In this multidimensional assessment it can be seen that the hybrid models DESC_H0_3 (part of the molecules are removed as outliers) and FPN_H2_7 (prediction is made by either standard or knowledge-based model depending on the value of log P) present some interesting optimality characteristics. Model DESC_H0_3 shows the best affinity with the rules compared to the rest of descriptor-based models and has the highest Spearman coefficient and accuracy. Model FPN_H2_7 demonstrates a good compromise between the correlation coefficient and accuracy and compliance with the rules.

The assessment presented in Figures 5 and 6 can generate ideas on developing hybrid models further toward an optimization-based approach (i.e., populating the Pareto-front of Figure 6 by more models, etc.). For instance, a combination of the knowledge-based approaches (i.e., H0 and H2, using MW and log P values) might be a better alternative to predict the toxicity values of the molecules of the data set.

The models are subjected to uncertainty due to the use of the mixed data set. However, the models are useful to compare the aquatic toxicity of a vast number of molecular structures (e.g., generated by CAMD) on the basis of a unified modeling framework and thus systematically reduce the number of chemical alternatives for later screening stages where rigorous experimentation and testing can be applied. Furthermore, the models illustrate the application of prior knowledge for the development and evaluation of hybrid models, which as a concept and methodology is valid independently from the nature of the mixed data set.

## CONCLUSIONS

This work demonstrated a systematic, semiautomated extraction and classification of knowledge in the field of acute aquatic toxicity and tested some ways of its integration into predictive models, which can be particularly useful in early design phases, where a vast number of chemicals should be screened, as in the case of CAMD approaches.

The semiautomated procedure of knowledge extraction significantly reduced the manual work required to process a large number of scientific articles while extracting generic and case-specific models, statements, and alerts. The automated text extraction might lead to the loss of valuable information; thus, the combination of the automated procedure with the manual text mining safeguards for the critical loss of relevant knowledge. The semiautomated knowledge extraction can be of assistance in interdisciplinary research when quick knowledge acquisition is required for different purposes (e.g., impact assessment). However, it should be noted that the semiautomated approach may still be introducing biases (although in a less subjective way than purely human-centric approaches). The bias can be introduced either through the standardization approaches when screening the vast amount of textual information or by the human-machine interaction in the form of keywords as input for the text mining approach and analysis of the extracted information and knowledge classification.

The knowledge collection and classification procedure can be useful in hybrid modeling studies concerning the model and predictor selection, prioritization, and constraints, addressing data gaps, and validating and interpreting model performance. The study demonstrated how the incorporation of prior knowledge improved the performance of the predictive models either in prediction accuracy or compliance with previously observed trends from the extracted prior knowledge. Furthermore, it was shown that the knowledge could be used in a variety of ways not only during the development of the models but also before and after for data analysis, model selection, evaluation, and adjustment.

The presented knowledge extraction method and approaches for knowledge incorporation into predictive models are generic and can be used in many other knowledge domains. The knowledge extraction method can easily incorporate more resources (in terms of amount and type), while the classified knowledge allows for more hybrid alternatives, also depending on the machine learning approach used (i.e., neural networks and deep learning approaches, classification trees, random forest regression, etc.). Thus, the presented hybridization methods should only be considered as examples of integrating the results of semiautomated knowledge extraction in the concept of hybrid modeling, and certainly diverse hybridization approaches can extend the presented concepts (e.g., using semiquantitative toxicity information not directly suitable for model calibration but possibly providing information to steer the calibration in the proper direction and test the results). Additionally, the knowledge extraction introduces a secondary model assessment beyond the prediction accuracy, namely the degree of compliance with different trends previously observed in the investigated knowledge domain. This makes it possible to apply multiobjective optimization in the development of predictive models, either with or without hybridization. Thus, more insights into model selection can be provided leading to more robust model development.

## DATA AND SOFTWARE AVAILABILITY

The knowmine package is available for installation via pip; the source files can be retrieved from https://github.com/GulnaraSh/Knowledge-mining-python. The titles of the input articles used for the knowledge mining are available by request from the corresponding author.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01079.

Details on knowledge extraction, collected QSA(A)Rs, and description of text mining knowmine package and models' development (PDF)

Data set of 2106 molecules containing smiles and acute aquatic toxicity values (XLS)

## AUTHOR INFORMATION

### Corresponding Author

**Stavros Papadokonstantakis** − *Department of Space, Earth and Environment, Division of Energy Technology, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden; Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, 1060 Vienna, Austria;* orcid.org/0000-0002-7774-0192;
Email: stavros.papadokonstantakis@tuwien.ac.at

### Authors

**Gulnara Shavalieva** − *Department of Space, Earth and Environment, Division of Energy Technology, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden;* orcid.org/0000-0001-6424-8505

**Gregory Peters** − *Department of Technology Management and Economics, Chalmers University of Technology, SE-411 33 Gothenburg, Sweden;* orcid.org/0000-0001-8319-168X

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.1c01079

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

## REFERENCES

(1) Rim, K. T. In Silico Prediction of Toxicity and Its Applications for Chemicals at Work. *Toxicol. Environ. Health Sci.* **2020**, *12*, 191−202.

(2) Ekins, S.; Mestres, J.; Testa, B. In Silico Pharmacology for Drug Discovery: Methods for Virtual Ligand Screening and Profiling. *Br. J. Pharmacol.* **2007**, *152*, 9−20.

(3) Apostolakou, A.; Adjiman, C. S. Optimization Methods in CAMD − II. Computer Aided Chemical Engineering; 2003; pp 63−93.

(4) Papadopoulos, A. I.; Shavalieva, G.; Papadokonstantakis, S.; Seferlis, P.; Perdomo, F. A.; Galindo, A.; Jackson, G.; Adjiman, C. S. An Approach for Simultaneous Computer-Aided Molecular Design with Holistic Sustainability Assessment: Application to Phase-Change CO2 Capture Solvents. *Comput. Chem. Eng.* **2020**, *135*, 106769.

(5) Chang, S. S. L.; Kong, Y. L.; Lim, W. X.; Ooi, J.; Ng, D. K. S.; Chemmangattuvalappil, N. G. Design of Alternate Solvent for Recovery of Residual Palm Oil: Simultaneous Optimization of Process Performance with Environmental, Health and Safety Aspects. *Clean Technol. Environ. Policy* **2018**, *20*, 949−968.

(6) Ten, J. Y.; Ng, L. Y.; Hassim, M. H.; Ng, D. K. S.; Chemmangattuvalappil, N. G. Managing Uncertainty on the Integration of Safety and Health Indexes in Computer-Aided Molecular Design. *Ind. Eng. Chem. Res.* **2017**, *56*, 10413−10427.

(7) Vanderveen, J. R.; Patiny, L.; Chalifoux, C. B.; Jessop, M. J.; Jessop, P. G. A Virtual Screening Approach to Identifying the Greenest Compound for a Task: Application to Switchable-Hydrophilicity Solvents. *Green Chem.* **2015**, *17*, 5182−5188.

(8) Khor, S. Y.; Liam, K. Y.; Loh, W. X.; Tan, C. Y.; Ng, L. Y.; Hassim, M. H.; Ng, D. K. S.; Chemmangattuvalappil, N. G. Computer Aided Molecular Design for Alternative Sustainable Solvent to Extract Oil from Palm Pressed Fibre. *Process Saf. Environ. Prot.* **2017**, *106*, 211−223.

(9) Keßler, T.; Kunde, C.; Linke, S.; McBride, K.; Sundmacher, K.; Kienle, A. Computer Aided Molecular Design of Green Solvents for the Hydroformylation of Long-Chain Olefines. Computer Aided Chemical Engineering; 2020; pp 745−750,.

(10) Ooi, J.; Promentilla, M. A. B.; Tan, R. R.; Ng, D. K. S.; Chemmangattuvalappil, N. G. A Systematic Methodology for Multi-Objective Molecular Design via Analytic Hierarchy Process. *Process Saf. Environ. Prot.* **2017**, *111*, 663−677.

(11) Weis, D. C.; Visco, D. P. Computer-Aided Molecular Design Using the Signature Molecular Descriptor: Application to Solvent Selection. *Comput. Chem. Eng.* **2010**, *34*, 1018−1029.

(12) Heintz, J.; Belaud, J. P.; Pandya, N.; Teles Dos Santos, M.; Gerbaud, V. Computer Aided Product Design Tool for Sustainable Product Development. *Comput. Chem. Eng.* **2014**, *71*, 362−376.

(13) Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating Physics-Based Modeling with Machine Learning: A Survey. *arXiv*. 2020. https://arxiv.org/abs/2003.04919 (accessed 2022-07-29).

(14) Beckh, K.; Müller, S.; Jakobs, M.; Toborek, V.; Tan, H.; Fischer, R.; Welke, P.; Houben, S.; von Rueden, L. Explainable Machine Learning with Prior Knowledge: An Overview. *arXiv* 2021. https://arxiv.org/abs/2105.10172 (accessed 2022-07-29).

(15) Von Rueden, L.; Mayer, S.; Beckh, K.; Georgiev, B.; Giesselbach, S.; Heese, R.; Kirsch, B.; Pfrommer, J.; Pick, A.; Ramamurthy, R.; Walczak, M.; Garcke, J.; Bauckhage, C.; Schuecker, J. Informed Machine Learning - a Taxonomy and Survey of Integrating Knowledge into Learning Systems. *arXiv*. 2019. https://arxiv.org/abs/1903.12394 (accessed 2022-07-29).

(16) Diligenti, M.; Roychowdhury, S.; Gori, M. Integrating Prior Knowledge into Deep Learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*; IEEE: 2017; Vol. 2017-Dec, pp 920−923, DOI: 10.1109/ICMLA.2017.00-37.

(17) Faghmous, J. H.; Kumar, V. A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data* **2014**, *2*, 155−163.

(18) Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J. L.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A.; Manepalli, A.; Chirila, D.; Yu, R.; Walters, R.; White, B.; Xiao, H.; Tchelepi, H. A.; Marcus, P.; Anandkumar, A.; Hassanzadeh, P.; Prabhat. Physics-Informed Machine Learning: Case Studies for Weather and Climate Modelling. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2021**, *379*, 20200093.

(19) Culos, A.; Tsai, A. S.; Stanley, N.; Becker, M.; Ghaemi, M. S.; McIlwain, D. R.; Fallahzadeh, R.; Tanada, A.; Nassar, H.; Espinosa, C.; Xenochristou, M.; Ganio, E.; Peterson, L.; Han, X.; Stelzer, I. A.; Ando, K.; Gaudilliere, D.; Phongpreecha, T.; Marić, I.; Chang, A. L.; Shaw, G. M.; Stevenson, D. K.; Bendall, S.; Davis, K. L.; Fantl, W.; Nolan, G. P.; Hastie, T.; Tibshirani, R.; Angst, M. S.; Gaudilliere, B.; Aghaeepour, N. Integration of Mechanistic Immunological Knowledge into a Machine Learning Pipeline Improves Predictions. *Nat. Mach. Intell.* **2020**, *2*, 619−628.

(20) Xuan, P.; Cao, Y.; Zhang, T.; Wang, X.; Pan, S.; Shen, T. Drug Repositioning through Integration of Prior Knowledge and Projections of Drugs and Diseases. *Bioinformatics* **2019**, *35*, 4108−4119.

(21) Palomba, D.; Martïnez, M. J.; Ponzoni, I.; Dïaz, M. F.; Vazquez, G. E.; Soto, A. J. QSPR Models for Predicting Log Pliver Values for Volatile Organic Compounds Combining Statistical Methods and Domain Knowledge. *Molecules* **2012**, *17*, 14937−14953.

(22) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490−2504.

(23) Jonnalagadda, S. R.; Goyal, P.; Huffman, M. D. Automating Data Extraction in Systematic Reviews: A Systematic Review. *Syst. Rev.* **2015**, *4*, 78.

(24) Shahid, A.; Afzal, M. T.; Abdar, M.; Basiri, M. E.; Zhou, X.; Yen, N. Y.; Chang, J.-W. Insights into Relevant Knowledge Extraction Techniques: A Comprehensive Review. *J. Supercomput.* **2020**, *76*, 1695−1733.

(25) Talib, R.; Hanif, M. K.; Ayesha, S.; Fatima, F. Text Mining : Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications* **2016**, *7*, 414−418.

(26) Zhang, Y.; Wang, M.; Saberi, M.; Chang, E. From Big Scholarly Data to Solution-Oriented Knowledge Repository. *Front. Big Data* **2019**, *2*, 1−10.

(27) Pandi, M.-T.; van der Spek, P. J.; Koromina, M.; Patrinos, G. P. A Novel Text-Mining Approach for Retrieving Pharmacogenomics Associations From the Literature. *Front. Pharmacol.* **2020**, *11*, 1−9.

(28) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Automated Chemical Reaction Extraction from Scientific Literature. *J. Chem. Inf. Model.* **2021**, *62*, 2035.

(29) IRIS. AI. Powerful Tools for Your Research. https://iris.ai/ (accessed 2022-07-29).

(30) European Commission. Aquatic toxicity. https://joint-research-centre.ec.europa.eu/eu-reference-laboratory-alternatives-animal-testing/alternative-methods-toxicity-testing/validated-test-methods-health-effects/aquatic-toxicity_en (accessed 2021-07-17).

(31) Wang, Y.; Chen, X. A Joint Optimization QSAR Model of Fathead Minnow Acute Toxicity Based on a Radial Basis Function Neural Network and Its Consensus Modeling. *RSC Adv.* **2020**, *10*, 21292−21308.

(32) Pu, L.; Naderi, M.; Liu, T.; Wu, H. C.; Mukhopadhyay, S.; Brylinski, M. EToxPred: A Machine Learning-Based Approach to Estimate the Toxicity of Drug Candidates 11 Medical and Health Sciences 1115 Pharmacology and Pharmaceutical Sciences 03 Chemical Sciences 0305 Organic Chemistry 03 Chemical Sciences 0304 Medicinal and Biomol. *BMC Pharmacol. Toxicol.* **2019**, *20*, 1−15.

(33) Chen, X.; Dang, L.; Yang, H.; Huang, X.; Yu, X. Machine Learning-Based Prediction of Toxicity of Organic Compounds towards Fathead Minnow. *RSC Adv.* **2020**, *10*, 36174−36180.

(34) Khan, K.; Benfenati, E.; Roy, K. Consensus QSAR Modeling of Toxicity of Pharmaceuticals to Different Aquatic Organisms: Ranking and Prioritization of the DrugBank Database Compounds. *Ecotoxicol. Environ. Saf.* **2019**, *168*, 287−297.

(35) Gupta, V.; Choudhary, K.; Tavazza, F.; Campbell, C.; Liao, W. k.; Choudhary, A.; Agrawal, A. Cross-Property Deep Transfer Learning Framework for Enhanced Predictive Analytics on Small Materials Data. *Nat. Commun.* **2021**, *12*, 6595.

(36) Liu, S.; Qu, M.; Zhang, Z.; Cai, H.; Tang, J. *Multi-Task Learning with Domain Knowledge for Molecular Property Prediction* **2021**, 1−11.

(37) Shavalieva, G. *Environmental, Health, and Safety Assessment of Chemical Alternatives during Early Process Design : The Role of Predictive Modeling and Streamlined Techniques*, Ph.D. thesis, Chalmers University of Technology, Sweden, 2022.

(38) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics among Chemicals in Commerce. III: Byproducts, Impurities, and Transformation Products. *Environ. Sci. Technol.* **2013**, *47*, 5259−5266.

(39) Deng, C.; Ji, X.; Rainey, C.; Zhang, J.; Lu, W. Integrating Machine Learning with Human Knowledge. *iScience* **2020**, *23*, 101656.

(40) Strempel, S. *PBT Assessment under REACH by Means of Chemoinformatics*, Ph.D. thesis, ETH Zurich, Switzerland, 2012.

(41) Melnikov, F.; Kostal, J.; Voutchkova-Kostal, A.; Zimmerman, J. B.; Anastas, P. T. Assessment of Predictive Models for Estimating the Acute Aquatic Toxicity of Organic Chemicals. *Green Chem.* **2016**, *18*, 4432−4445.

(42) Sheffield, T. Y.; Judson, R. S. Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Lethal Concentrations and Points of Departure. *Environ. Sci. Technol.* **2019**, *53*, 12793−12802.

(43) ECHA. *Guidance on Information Requirements and Chemical Safety Assessment*; 2017; DOI: 10.2823/128621.

(44) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations. *J. Cheminform.* **2015**, *7*, 1−13.

(45) Landrum, G. *RDKit Documentation. Release 2017.03.1*; 2017.

(46) Kessler, T. *PaDELPy: A Python Wrapper for PaDEL-Descriptor Software*; 2021.

(47) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC50) of Organic Compounds to the Fathead Minnow (Pimephales Promelas) Using a Group Contribution Method. *Chem. Res. Toxicol.* **2001**, *14*, 1378−1385.

(48) Bouhedjar, K.; Benfenati, E.; Nacereddine, A. K. Modelling Quantitative Structure Activity−Activity Relationships (QSAARs): Auto-Pass-Pass, a New Approach to Fill Data Gaps in Environmental Risk Assessment under the REACH Regulation. *SAR QSAR Environ. Res.* **2020**, *31*, 785−801.

(49) Amini, A.; Muggleton, S. H.; Lodhi, H.; Sternberg, M. J. E. A Novel Logic-Based Approach for Quantitative Toxicology Prediction. *J. Chem. Inf. Model.* **2007**, *47*, 998−1006.

(50) Cassotti, M.; Ballabio, D.; Consonni, V.; Mauri, A.; Tetko, I. V.; Todeschini, R. Prediction of Acute Aquatic Toxicity toward Daphnia Magna by Using the GA- k NN Method. *Altern. to Lab. Anim.* **2014**, *42*, 31−41.

(51) Qin, W. C.; Su, L. M.; Zhang, X. J.; Qin, H. W.; Wen, Y.; Guo, Z.; Sun, F. T.; Sheng, L. X.; Zhao, Y. H.; Abraham, M. H. Toxicity of Organic Pollutants to Seven Aquatic Organisms: Effect of Polarity and Ionization. *SAR QSAR Environ. Res.* **2010**, *21*, 389−401.

(52) Khan, K.; Baderna, D.; Cappelli, C.; Toma, C.; Lombardo, A.; Roy, K.; Benfenati, E. Ecotoxicological QSAR Modeling of Organic Compounds against Fish: Application of Fragment Based Descriptors in Feature Analysis. *Aquat. Toxicol.* **2019**, *212*, 162−174.

(53) Katritzky, A. R.; Slavov, S. H.; Stoyanova-Slavova, I. S.; Kahn, I.; Karelson, M. Quantitative Structure-Activity Relationship (QSAR) Modeling of EC 50 of Aquatic Toxicities for Daphnia Magna. *J. Toxicol. Environ. Heal. - Part A Curr. Issues* **2009**, *72*, 1181−1190.

(54) Liu, L.; Yang, H.; Cai, Y.; Cao, Q.; Sun, L.; Wang, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Chemical Aquatic Toxicity for Marine Crustaceans via Machine Learning. *Toxicol. Res. (Camb).* **2019**, *8*, 341−352.

(55) Roy, K.; Das, R. N. QSTR with Extended Topochemical Atom (ETA) Indices. 15. Development of Predictive Models for Toxicity of Organic Chemicals against Fathead Minnow Using Second-Generation ETA Indices. *SAR QSAR Environ. Res.* **2012**, *23*, 125−140.

(56) Nendza, M.; Müller, M. Discriminating Toxicant Classes by Mode of Action: 2. Physico-Chemical Descriptors. *Quant. Struct. Relationships* **2000**, *19*, 581−598.

(57) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I.; Aptula, A. O. Structure−Toxicity Relationships for Aliphatic Chemicals Evaluated with Tetrahymena Pyriformis. *Chem. Res. Toxicol.* **2002**, *15*, 1602−1609.

(58) Toropova, A. P.; Toropov, A. A.; Veselinović, A. M.; Veselinović, J. B.; Leszczynska, D.; Leszczynski, J. Monte Carlo−Based Quantitative Structure−Activity Relationship Models for Toxicity of Organic Chemicals to Daphnia Magna. *Environ. Toxicol. Chem.* **2016**, *35*, 2691−2697.

(59) Morrall, D. D.; Belanger, S. E.; Dunphy, J. C. Acute and Chronic Aquatic Toxicity Structure-Activity Relationships for Alcohol Ethoxylates. *Ecotoxicol. Environ. Saf.* **2003**, *56*, 381−389.

(60) Liu, W.; Wang, X.; Zhou, X.; Duan, H.; Zhao, P.; Liu, W. Quantitative Structure-Activity Relationship between the Toxicity of Amine Surfactant and Its Molecular Structure. *Sci. Total Environ.* **2020**, *702*, 134593.

(61) Pavan, M.; Netzeva, T. I.; Worth, A. P. Validation of a QSAR Model for Acute Toxicity. *SAR QSAR Environ. Res.* **2006**, *17*, 147−171.

(62) Raevsky, O. A.; Grigor'ev, V. Y.; Tikhonova, O. V. Molecular-Biological Problems of Drug Design and Mechanism of Drug Action Development of Structure-Toxicity Relationship Models of Chemicals With Respect To Guppy. *Pharm. Chem. J.* **2009**, *43*, 125−129.

(63) Gupta, S.; Basant, N.; Singh, K. P. Predicting Aquatic Toxicities of Benzene Derivatives in Multiple Test Species Using Local, Global and Interspecies QSTR Modeling Approaches. *RSC Adv.* **2015**, *5*, 71153−71163.

(64) Kar, S.; Roy, K.; Leszczynski, J. Impact of Pharmaceuticals on the Environment: Risk Assessment Using QSAR Modeling Approach. In *Computational Toxicology: Methods and Protocols, Methods in Molecular Biology, vol.1800*; 2018; Vol. *1800*, DOI: 10.1007/978-1-4939-7899-1_19.

(65) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatical, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766−784.

(66) Bašic, I.; Lučié, B.; Nikolić, S.; Papeš-Šokčević, L.; Nadramija, D. Improvement of Ensemble of Multi-Regression Structure-Toxicity Models by Clustering of Molecules in Descriptor Space. *AIP Conf. Proc.* **2009**, *1148* (2), 408−411.

(67) Grigor'ev, V. Y.; Razdol'skii, A. N.; Zagrebin, A. O.; Tonkopii, V. D.; Raevskii, O. A. QSAR Classification Models of Acute Toxicity of Organic Compounds with Respect to Daphnia Magna. *Pharm. Chem. J.* **2014**, *48*, 242−245.

(68) Cassotti, M.; Consonni, V.; Mauri, A.; Ballabio, D. Validation and Extension of a Similarity-Based Approach for Prediction of Acute Aquatic Toxicity towards Daphnia Magna. *SAR QSAR Environ. Res.* **2014**, *25*, 1013−1036.

(69) Martin, T. M.; Young, D. M.; Lilavois, C. R.; Barron, M. G. Comparison of Global and Mode of Action-Based Models for Aquatic Toxicity. *SAR QSAR Environ. Res.* **2015**, *26*, 245−262.

(70) Gini, G.; Craciun, M. V.; König, C.; Benfenati, E. Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897−1902.

(71) Gramatica, P.; Papa, E.; Sangion, A. QSAR Modeling of Cumulative Environmental End-Points for the Prioritization of Hazardous Chemicals. *Environ. Sci. Process. Impacts* **2018**, *20*, 38−47.

(72) Su, L.; Fu, L.; He, J.; Qin, W.; Sheng, L.; Abraham, M. H.; Zhao, Y. H. Comparison of Tetrahymena Pyriformis Toxicity Based on Hydrophobicity, Polarity, Ionization and Reactivity of Class-Based Compounds. *SAR QSAR Environ. Res.* **2012**, *23*, 537−552.

(73) Ren, Y. Y.; Zhou, L. C.; Yang, L.; Liu, P. Y.; Zhao, B. W.; Liu, H. X. Predicting the Aquatic Toxicity Mode of Action Using Logistic Regression and Linear Discriminant Analysis. *SAR QSAR Environ. Res.* **2016**, *27*, 721−746.

(74) Finizio, A.; Di Nica, V.; Rizzi, C.; Villa, S. A Quantitative Structure-Activity Relationships Approach to Predict the Toxicity of Narcotic Compounds to Aquatic Communities. *Ecotoxicol. Environ. Saf.* **2020**, *190*, 110068.

(75) Hossain, K. A.; Roy, K. Chemometric Modeling of Aquatic Toxicity of Contaminants of Emerging Concern (CECs) in Dugesia Japonica and Its Interspecies Correlation with Daphnia and Fish: QSTR and QSTTR Approaches. *Ecotoxicol. Environ. Saf.* **2018**, *166*, 92−101.

(76) Önlü, S.; Saçan, M. T. Toxicity of Contaminants of Emerging Concern to Dugesia Japonica: QSTR Modeling and Toxicity Relationship with Daphnia Magna. *J. Hazard. Mater.* **2018**, *351*, 20−28.

(77) Terasaki, M.; Makino, M.; Tatarazako, N. Acute Toxicity of Parabens and Their Chlorinated By-Products with Daphnia Magna and Vibrio Fischeri Bioassays. *J. Appl. Toxicol.* **2009**, *29*, 242−247.

(78) Li, J. J.; Zhang, X. J.; Yang, Y.; Huang, T.; Li, C.; Su, L.; Zhao, Y. H.; Cronin, M. T. D. Development of Thresholds of Excess Toxicity for Environmental Species and Their Application to Identification of Modes of Acute Toxic Action. *Sci. Total Environ.* **2018**, *616*−*617*, 491−499.

(79) Zhang, X. J.; Qin, H. W.; Su, L. M.; Qin, W. C.; Zou, M. Y.; Sheng, L. X.; Zhao, Y. H.; Abraham, M. H. Interspecies Correlations of Toxicity to Eight Aquatic Organisms: Theoretical Considerations. *Sci. Total Environ.* **2010**, *408*, 4549−4555.

(80) Yen, J.-H.; Lin, K.-H.; Wang, Y.-S. Acute Lethal Toxicity of Environmental Pollutants to Aquatic Organisms. *Ecotoxicol. Environ. Saf.* **2002**, *52*, 113−116.

(81) Huang, C. P.; Wang, Y. J.; Chen, C. Y. Toxicity and Quantitative Structure-Activity Relationships of Nitriles Based on Pseudokirchneriella Subcapitata. *Ecotoxicol. Environ. Saf.* **2007**, *67*, 439−446.

(82) Tsai, K. P.; Chen, C. Y. An Algal Toxicity Database of Organic Toxicants Derived by a Closed-System Technique. *Environ. Toxicol. Chem.* **2007**, *26*, 1931−1939.

(83) Dom, N.; Knapen, D.; Benoot, D.; Nobels, I.; Blust, R. Aquatic Multi-Species Acute Toxicity of (Chlorinated) Anilines: Experimental versus Predicted Data. *Chemosphere* **2010**, *81*, 177−186.

(84) Chen, C. Y.; Ko, C. W.; Lee, P. I. Toxicity of Substituted Anilines to Pseudokirchneriella Subcapitata and Quantitative Structure-Activity Relationship Analysis for Polar Narcotics. *Environ. Toxicol. Chem.* **2007**, *26*, 1158−1164.

(85) Li, J. J.; Wang, X. H.; Wang, Y.; Wen, Y.; Qin, W. C.; Su, L. M.; Zhao, Y. H. Discrimination of Excess Toxicity from Narcotic Effect: Influence of Species Sensitivity and Bioconcentration on the Classification of Modes of Action. *Chemosphere* **2015**, *120*, 660−673.

(86) Rawlings, J. M.; Belanger, S. E.; Connors, K. A.; Carr, G. J. Fish Embryo Tests and Acute Fish Toxicity Tests Are Interchangeable in the Application of the Threshold Approach. *Environ. Toxicol. Chem.* **2019**, *38*, 671−681.

(87) Sánchez-Bayo, F. Comparative Acute Toxicity of Organic Pollutants and Reference Values for Crustaceans. I. Branchiopoda, Copepoda and Ostracoda. *Environ. Pollut.* **2006**, *139*, 385−420.

(88) Lunghini, F.; Marcou, G.; Azam, P.; Enrici, M. H.; Van Miert, E.; Varnek, A. Consensus QSAR Models Estimating Acute Toxicity to Aquatic Organisms from Different Trophic Levels: Algae, Daphnia and Fish. *SAR QSAR Environ. Res.* **2020**, *31*, 655−675.

(89) Khan, K.; Roy, K. Ecotoxicological QSAR Modelling of Organic Chemicals against Pseudokirchneriella Subcapitata Using Consensus Predictions Approach. *SAR QSAR Environ. Res.* **2019**, *30*, 665−681.

(90) Tinkov, O. V.; Ognichenko, L. N.; Kuz'Min, V. E.; Gorb, L. G.; Kosinskaya, A. P.; Muratov, N. N.; Muratov, E. N.; Hill, F. C.; Leszczynski, J. Computational Assessment of Environmental Hazards of Nitroaromatic Compounds: Influence of the Type and Position of Aromatic Ring Substituents on Toxicity. *Struct. Chem.* **2016**, *27*, 191−198.

(91) Galimberti, F.; Moretto, A.; Papa, E. Application of Chemometric Methods and QSAR Models to Support Pesticide Risk Assessment Starting from Ecotoxicological Datasets. *Water Res.* **2020**, *174*, 115583.

(92) Khan, K.; Kar, S.; Sanderson, H.; Roy, K.; Leszczynski, J. Ecotoxicological Modeling, Ranking and Prioritization of Pharmaceuticals Using QSTR and i-QSTTR Approaches: Application of 2D and Fragment Based Descriptors. *Mol. Inform.* **2019**, *38*, 1800078.

(93) Voutchkova, A. M.; Kostal, J.; Steinfeld, J. B.; Emerson, J. W.; Brooks, B. W.; Anastas, P.; Zimmerman, J. B. Towards Rational Molecular Design: Derivation of Property Guidelines for Reduced Acute Aquatic Toxicity. *Green Chem.* **2011**, *13*, 2373−2379.

(94) Schultz, T. W.; Yarbrough, J. W.; Pilkington, T. B. Aquatic Toxicity and Abiotic Thiol Reactivity of Aliphatic Isothiocyanates: Effects of Alkyl-Size and -Shape. *Environ. Toxicol. Pharmacol.* **2007**, *23*, 10−17.

(95) Roy, K.; Ghosh, G. QSTR with Extended Topochemical Atom Indices. 3. Toxicity of Nitrobenzenes ToTetrahymena Pyriformis. *QSAR Comb. Sci.* **2004**, *23*, 99−108.

(96) Matveieva, M.; Cronin, M. T. D.; Polishchuk, P. Interpretation of QSAR Models: Mining Structural Patterns Taking into Account Molecular Context. *Mol. Inform.* **2019**, *38*, 1800084.

(97) Netzeva, T. I.; Pavan, M.; Worth, A. P. Review of (Quantitative) Structure−Activity Relationships for Acute Aquatic Toxicity. *QSAR Comb. Sci.* **2008**, *27*, 77−90.

(98) Stoyanova-Slavova, I. B.; Slavov, S. H.; Pearce, B.; Buzatu, D. A.; Beger, R. D.; Wilkes, J. G. Partial Least Square and K-Nearest Neighbor Algorithms for Improved 3D Quantitative Spectral Data-Activity Relationship Consensus Modeling of Acute Toxicity. *Environ. Toxicol. Chem.* **2014**, *33*, 1271−1282.

(99) Li, F.; Fan, D.; Wang, H.; Yang, H.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Pesticide Aquatic Toxicity with Chemical Category Approaches. *Toxicol. Res. (Camb).* **2017**, *6*, 831−842.

(100) Martin, T. M.; Grulke, C. M.; Young, D. M.; Russom, C. L.; Wang, N. Y.; Jackson, C. R.; Barron, M. G. Prediction of Aquatic Toxicity Mode of Action Using Linear Discriminant and Random Forest Models. *J. Chem. Inf. Model.* **2013**, *53*, 2229−2239.

(101) Kienzler, A.; Barron, M. G.; Belanger, S. E.; Beasley, A.; Embry, M. R. Mode of Action (MOA) Assignment Classifications for Ecotoxicology: An Evaluation of Approaches. *Environ. Sci. Technol.* **2017**, *51*, 10203−10211.

(102) Fogel, G. B.; Cheung, M. Derivation of Quantitative Structure-Toxicity Relationships for Ecotoxicological Effects of Organic Chemicals: Evolving Neural Networks and Evolving Rules. *2005 IEEE Congr. Evol. Comput. IEEE CEC 2005. Proc.* **2005**, *1*, 274−281.

(103) Nendza, M.; Müller, M.; Wenzel, A. Classification of Baseline Toxicants for QSAR Predictions to Replace Fish Acute Toxicity Studies. *Environ. Sci. Process. Impacts* **2017**, *19*, 429−437.

(104) Colombo, A.; Benfenati, E.; Karelson, M.; Maran, U. The Proposal of Architecture for Chemical Splitting to Optimize QSAR Models for Aquatic Toxicity. *Chemosphere* **2008**, *72*, 772−780.

(105) Papa, E.; Villa, F.; Gramatica, P. Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales Promelas (Fathead Minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256−1266.

(106) Valavanidis, A.; Vlahogianni, T.; Dassenakis, M.; Scoullos, M. Molecular Biomarkers of Oxidative Stress in Aquatic Organisms in Relation to Toxic Environmental Pollutants. *Ecotoxicol. Environ. Saf.* **2006**, *64*, 178−189.

(107) Chen, C. Y.; Kuo, K. L.; Fan, J. W. Toxicity of Propargylic Alcohols on Green Alga - Pseudokirchneriella Subcapitata. *J. Environ. Monit.* **2012**, *14*, 181−186.

(108) Payne, M. P.; Button, W. G. Prediction of Acute Aquatic Toxicity in Tetrahymena Pyriformis - "Eco-Derek", a Knowledge-Based System Approach£. *SAR QSAR Environ. Res.* **2013**, *24*, 439−460.

(109) Wu, X.; Zhang, Q.; Hu, J. QSAR Study of the Acute Toxicity to Fathead Minnow Based on a Large Dataset. *SAR QSAR Environ. Res.* **2016**, *27*, 147−164.