Green Chemistry

ROYAL SOCIETY OF CHEMISTRY

CRITICAL REVIEW

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Green Chem.*, 2023, **25**, 1261

Towards higher scientific validity and regulatory acceptance of predictive models for PFAS†

Anita Sosnowska, *[a] Natalia Bulawska, [a] Dominika Kowalska [a] and Tomasz Puzyn *[a,b]

Received 17th November 2022,
Accepted 4th January 2023

DOI: 10.1039/d2gc04341f

rsc.li/greenchem

Per- and polyfluoroalkyl substances (PFAS) are widespread in the environment. Properly developed QSAR/QSPR models can be used to assess the impact of these chemicals on humans and the environment. This work assesses 38 *in silico* models developed for this group of compounds, which mainly show physicochemical (22), and also toxic (8) and ecotoxic (8) properties. The evaluation of the models was carried out based on the (Q)SAR Model Reporting Format (QMRF), which was found in the QSAR Database (5) or was prepared manually, according to the information contained in scientific publications based on the QMRFEditor-v3.0.0 format (33). We based our evaluations on an individual assessment of each of the OECD principles described in the document and then summing up everything together. During the analysis, we identified 22 models as scientifically valid and could be used in the prediction of new compounds. Twelve of them contained all the information necessary to reproduce the model, and another 10, despite the lack of some information, are still reproducible. The other 16 models do not contain enough information to reproduce them and therefore they are scientifically invalid. The present work allows identifying the remaining gaps, needs, and recommendations that should be considered in further development of predictive models in the PFAS area.

1. Introduction

1.1. PFAS – "forever chemicals"

Per- and polyfluoroalkyl substances (PFAS) form an extensive family of fluorinated chemicals that have been in use since the late 40s. According to the Organization for Economic Cooperation and Development (OECD), there are nearly 5000 PFAS that have been registered and/or produced.[1] Many of those compounds have numerous applications in various areas of life. PFAS are effective surfactants or surface protectors; they reduce surface tension in an aqueous environment including processing aids to produce fluoropolymers, water-film forming coatings, and aqueous film-forming foams (AFFFs) used to fight fires involving highly flammable fluids.[2] They are also used in cosmetics which contribute to easier spreading on the skin.[3] Moreover they are applied in food packaging production, which prevents fat seepage, or in the production of pots and pans preventing food from sticking to the pan. PFAS find a wide range of applications due to their unusual properties like chemical and thermal stability and hydrophobic and lipophobic nature. What is more, all PFAS contain carbon–fluorine bonds – one of the strongest bonds found in organic chemistry.[2] This means that they are extremely resistant to degradation during use and after release into the environment. In addition, most PFAS easily spread into the environment, traveling long distances from the site of their release. They can enter the environment within direct (*e.g.* industrial facilities using PFAS), and indirect ways (*e.g.* during the use of consumer products like cosmetics, clothing, and food packets). The continuous emission of PFAS leads to the accumulation of levels in the environment and an increased probability of causing adverse effects.

On one hand, the unique properties of PFAS make them possible to be used in numerous applications; however at the same time these properties make PFAS dangerous for the environment and humans. PFAS may affect the immune,[4] digestive,[5] metabolic,[6,7] endocrine,[8,9] and nervous systems.[10] They contribute to the maturation change and increase the risk of developing breast, kidney, testis, prostate, and ovary cancer.[11] PFAS may also act as endocrine disruptors by influencing for example thyroid hormone levels. Studies are indicating that PFAS can lower a woman's chances of getting pregnant, as well as affect the growth, learning, and behavior of infants and older children. From the environmental point of view, PFAS are not eliminated by natural barriers of terrestrial and aquatic ecosys-

[a]QSAR Lab, ul. Trzy Lipy 3, Gdańsk, Poland. E-mail: a.sosnowska@qsarlab.com, t.puzyn@qsarlab.com
[b]University of Gdansk, Faculty of Chemistry, Wita Stwosza 63, 80-308 Gdansk, Poland. E-mail: tomasz.puzyn@ug.edu.pl

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2gc04341f

tems and therefore can concentrate in water, nutrients, *etc*.[12] Due to the solubility of many PFAS in water and their low potential for absorption onto particles, it is very difficult to remove PFAS from the aquatic environment, including drinking water sources, using conventional methods. That is why they are known also as "Forever Chemicals".

All reports and the literature pointed out that PFAS may trigger adverse effects, and that they are very persistent, mobile (PM), and difficult to remove from the environment prompting the regulatory authorities to take a closer look at this group of chemicals. In fact, in recent years the production and use of several groups of PFAS are restricted under REACH Regulations. Two of the most known and frequently used PFAS have been included in the Stockholm International Convention – perfluorooctane sulfonic acid (PFOS)[13] and its derivatives since 2009, and perfluorooctanoic acid (PFOA),[14] its salts and PFOA related compounds since 2020. Several European countries (mainly Norway, Germany, Netherlands, Denmark, and Sweden) constantly raise initiatives that lead to proposed restrictions on the different groups of PFAS.

Based on these initiatives, European Commission decided that from February 2023 the next group of PFAS – perfluorinated carboxylic acids (C9-14 PFCAs),[15] their precursors and salts will be restricted in the EU/EEA. It seems that soon such restrictions will affect more and more groups of those chemicals, considering that many of them are also on the REACH candidate list of substances of very high concern (SVHC), which classify them as carcinogens, mutagens, and reprotoxicants (CMRs), and also persistent, bioaccumulative and toxic/very persistent and very bioaccumulative (PBTs/vPvBs) chemicals. PFAS are the subject of interest of the European Green Deal initiative and Zero pollution action plan[16] which assumes the reduction of levels of pollutants in air, water, and soil and creates a toxic-free environment. There are three European Horizon 2020 projects founded (PROMISCES (101036449), SCENARIOS (101036756), ZeroPM (101037509)) dealing with this topic and proposing new strategies to protect the environment and human health from PFAS and PM chemicals.

Even though PFAS are produced for more than 80 years now, the fact that they are composed of a huge group (about 5000) of compounds shows that only a small number of them have been fully tested and their properties are known. It is impossible to test experimentally every substance; therefore, ECHA recommended the holistic group approach in the regulatory assessment and risk management based on the EU strategy for PFAS.[17] In this respect, computational methods for deriving the activities/properties of PFAS can be widely applied to replace/complete experimental methods. Using *in silico* methods, data analysis, and machine learning, it is possible to determine the toxic potential/physicochemical properties of a large set of compounds based only on the small, experimental set of available data.

1.2. *In silico* methods used for regulatory purpose

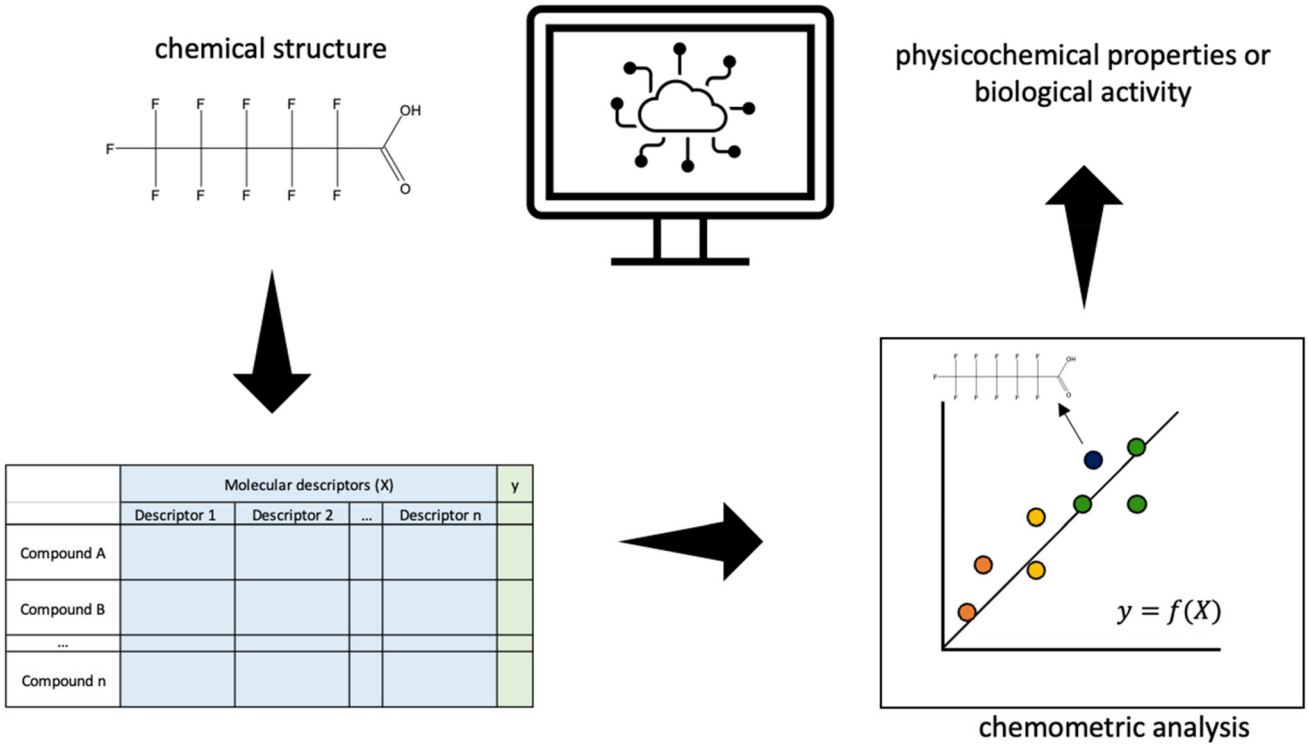
The *in silico* toxicity and safety assessment methods can be used as an alternative to testing on animals for the REACH Regulation.[18] These methods are based on the assumption that there is a relationship between the chemical structure of a compound and its properties, including biological activity. Moreover, structurally similar compounds may have similar properties.

One of the groups of methods which is based on the similarity of relationships, recommended by REACH for supporting the substance registration process, is quantitative structure–activity(property) relationship models (QSAR/QSPR models). The QSAR/QSPR models relate the set of descriptors (X) with the response variables (Y). The chemical structure (variable (X)) is represented numerically through descriptors such as molar mass, number of atoms, number of bonds, number of aromatic rings, hydrophobicity, *etc*. (Scheme 1).[19–22] The choice of the appropriate modeling method depends both on the nature of the modeled quantity and the nature of the relationship between the descriptors and the predicted value (linear and non-linear). If the modeled variable is quantified, then linear and nonlinear regression techniques can be used for modeling. When the data is qualitative then the selection of modeling methods is limited to the classification one. The credibility of the models is confirmed by appropriate statistical parameters. A properly developed QSAR model should be characterized by a good fit to the training set, robustness, and defined predictive ability.[19] The model developed in this way, in addition to providing information about the properties of chemical compounds, should also support the understanding of the mechanisms related to the biological activity of substances.

In silico methods are an attractive and faster alternative compared to time-consuming laboratory and clinical research methods. They are also used to support experimental data or to support prioritization in the absence of experimental data for regulatory purpose. Based on the existing registration dossiers, the European Chemicals Agency (ECHA) carried out an analysis of the used method for obtaining information on the properties of the substances, which showed that the alternative methods to animal testing specified and recommended in REACH are successfully used by registrants. Annex XI in REACH regulation allows for the application of the (Q)SAR models as a standard mode of research.[20] However, to use such models for predictions supporting the substance registration process, certain conditions have to be fulfilled. The model used for prediction should follow the golden standards established for QSAR models in 2004.[21] In accordance with these rules "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".[21,22]

Only models which fulfill all the above-mentioned requirements can be used for predicting biological activity/physical parameters to support the registration of the substance. What



Scheme 1 Quantitative structure–activity/property relationship – QSAR/QSPR – *in silico* approach.

is more, it is recommended that a well-documented description of the applied model should be attached to the registration dossier. For this purpose, QMRF ((Q)SAR Model Reporting Format) was proposed in 2007.[22]

1.3. (Q)SAR model reporting format

The QMRF framework refers to a comprehensive summary of the QSAR/QSPR models, which also includes their appropriate validation.[22,23] This format has been designed so that it can be easily checked whether the developed/applied model complies with the principles of good practice for developing QSAR models created by the OECD. The EC Joint Research Centre established a freely-accessible inventory of evaluated QSARs (QSAR Database)[24] which contained uploaded and valid QSAR/QSPR models for regulatory purposes (in QMRF format). Moreover, there is a shared application (QMRFEditor-v3.0.0) for the creation, storage, and download of QMRFs, and a web-based interface for retrieval QMRFs and transforming them into the submission (*i.e.* excel format) of QMRF.[25]

The structure of the report is divided into 10 sections (Scheme 2) which refer to different aspects of QSAR/QSPR models required for regulatory purposes.[23] Below we specified what should be indicated in each section of the QMRF report.

Section 1 (QSAR identifier) is related to the description of the model, where the title, other related models, and software coding of the model should be indicated.

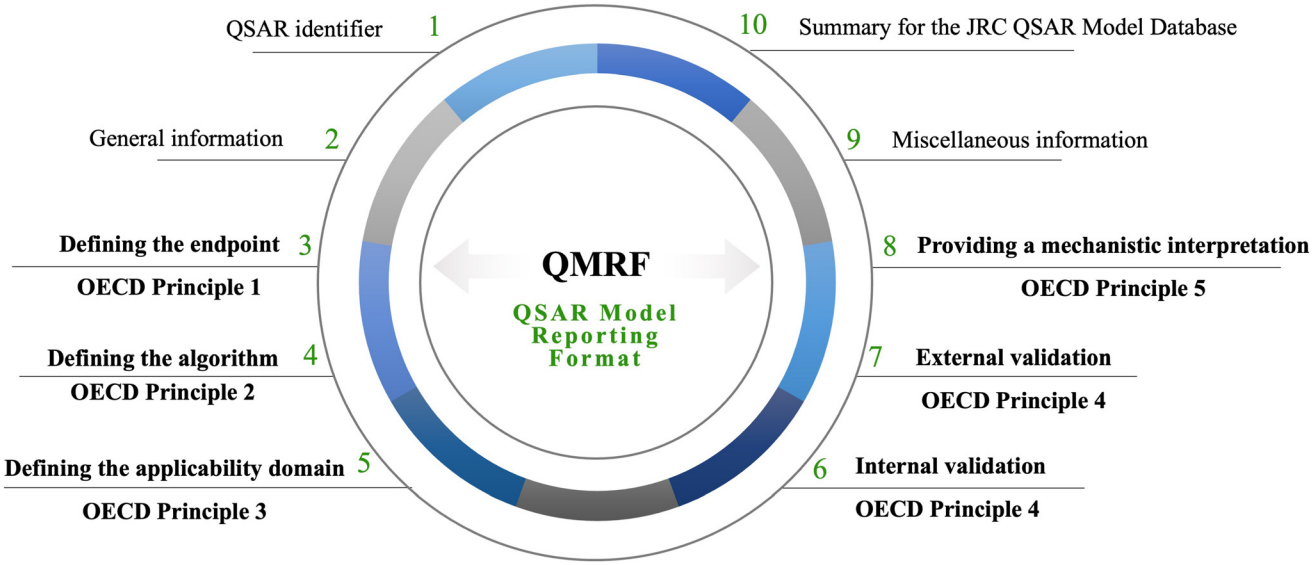
Section 2 (general information) provides general information about the developed model, *e.g.*, the date and authors who prepared QMRF, and the authors of the model and referring publication, available information of the model (*e.g.* training and external validation sets, source code, and algorithm) and information if there exist other QMRF documents for the exact model.

Section 3 (defining the endpoint – OECD principle 1) should specify the endpoint (physicochemical, biological, or environmental effects – from the pre-defined classification)[23] species and units for the endpoint, the experimental protocol followed by the collection of the experimental sets, and the information on the data quality assessment and the relationship of the modeled (dependent variable) and measured endpoints (*e.g.* transformation, *etc.*).

Section 4 (defining the algorithm – OECD principle 2) refers to the type of the model (*e.g.* SAR, QSAR, Expert-based Systems, and Neural Networks) and all information connected with the developed relationship. In particular, it should be indicated which descriptors are used in the model, how they were estimated, how the selection of the descriptors was performed, and what is the ratio of the descriptors used in the model to chemicals in the training set. Moreover, the method and software used to derive the relationship (algorithm) should be specified in this section.

Section 5 (defining the applicability domain – OECD principle 3) provides detailed information about the chemical space in which the model predicts properly. Here the comments on methods and software used for defining the applicability domain, and their limitations should be mentioned.

Section 6 (defining goodness-of-fit and robustness – OECD principle 4) contains notes about the statistical analysis that



Scheme 2 The structure of the QMRF document.

should be performed to establish the performance of the model, consisting also of the internal validation (*i.e.*, measures of goodness-of-fit and robustness). In this section, there is a need to give information about the availability of the training set, with all specifications, *e.g.*, CAS number, SMILES, Mol file, *etc*. Moreover, the data for an endpoint for the modeled values and the descriptors values for chemicals in the training set should be included here. The authors should indicate if the model is developed based on the rare data, or if any transformation was applied. All statistics describing the goodness-of-fit ((r^2, r^2 adjusted, standard error, sensitivity, specificity, false negatives, false positives, predictive values, *etc*.),[26,27] and robustness (*e.g.* leave-one-out and leave-many-out cross-validation,[28] Y-scrambling,[29] bootstrap[29] or any other corresponding statistics) should be reported.[30]

Section 7 (defining predictivity – OECD principle 4) is associated with the external validation of the model and determination of the model's predictive power, which is the measure that describes how well the models predict endpoints for new chemicals, which was not considered to develop the model. In this section the following information should be provided: the availability of the validation set, with all specifications *e.g.*, CAS number, SMILES, Mol file, *etc*. Data for each descriptor and dependent variable for the external validation set and the information on how the validation set was defined (*e.g.*, randomly, using a specific algorithm, searching in the literature, *etc*.) need to be presented. Moreover, all statistics obtained by external validation[28,31–33] and predictivity assessment (discussion on the magnitude of the validation set and if it is sufficient and the representative of the applicability domain) should be specified.[30]

Section 8 (providing the mechanistic interpretation – OECD principle 5) refers to the mechanistic interpretation of the presented model. Here, information on the mechanistic basis of the model should be provided. The description of the structural features that are responsible for the modeled properties should be demonstrated. Also, if possible, the physicochemical interpretation of the used descriptors should be explained. It should be pointed out if the mechanistic interpretation was determined *a priori* (before modeling, and the training set and descriptors were fitted to the already known statements) or posteriori (after modeling, and it was the result of the interpretation of the obtained relationship).

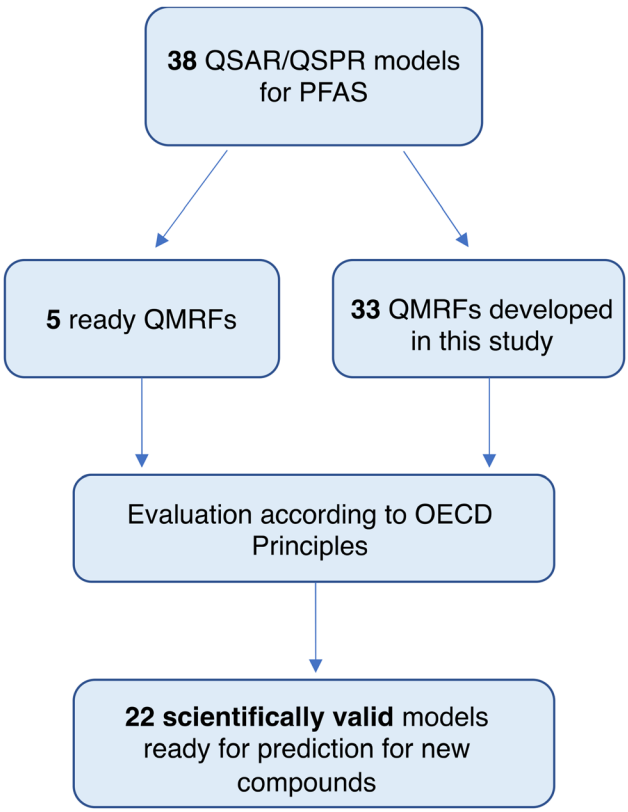
Section 9 (miscellaneous information) includes any other relevant and useful comments not indicated above, a bibliography (references not strictly associated with the developed model), and the ESI† (if it is attached to the QMRF, the ESI† may include the training and test sets submitted in defined file formats).

Section 10 (summary for the JRC QSAR model database)[24] is a summary section specified for the JRC Database. Here the QMRF number is generated and the publication date, keywords, and comments relevant to the publication of the QMRF in the JRC Database (*e.g.*, updates) should be reported.

The QSAR models are playing an increasingly important role in defining the properties for the hazard and risk assessment of chemicals. Using this method, it is possible to search for compounds that are safe for the environment and humans but still exhibit certain desired properties. New compounds can be registered based on the QSAR models validated in the form of QMRF, only then documentation is standardized and predictions with these models are reliable.[34]

In light of the considerations above, the present work attempts on summarizing the previous QSAR/QSPR studies of the PFAS and verifying whether the models developed so far for predicting the physicochemical properties and biological activity are scientifically valid and could be easily applied to

predict the properties for new (safer) compounds. What is more, this review will allow one to highlight the remaining gaps in this field and define further challenges related to applying the computational methods for predicting the activity and properties of PFASs.



Scheme 3 Workflow on searching and evaluation of the available QSAR/QSPR models.

2. Review methodology

In the first step, the literature was searched for the availability of QSAR/QSPR models that contained per- and polyfluoroalkyl compounds in the training set (Scheme 3). The keywords used for searching were as follows: (PFAS), (perfluoroalkyl compounds), (QSAR), (QSPR), (modeling), and (predictive models). We focused only on the models where PFASs were present in the training set.

In this way, we collected 38 models: 22 for predicting physicochemical properties, 8 for toxicological, and 8 models for ecotoxicological endpoints for PFASs (Table 1). Among those 38 models, five have ready-made QMRF documents available in the QSAR Database[35] (VP2, S2, LC50(1) – acute inhalation toxicity, LC50(2) – acute inhalation toxicity, and LD50(1) – acute oral toxicity), while one of them (LC50(1)) has also entered the JRC database.[24] It is worth mentioning that 5 ready-made QMRF documents (listed above) are also implemented in the QSARINS-Chem software, which is dedicated to the development and validation of QSAR models.[36] For the rest of the collected models, the missing QMRFs have been prepared based on the format from QMRFEditor-v3.0.0.[25] Ten sections (see the Introduction section) of the QMRF document were completed based on the information provided in the original papers and their supplements. All prepared QMRF documents are presented in ESI 2.† In the next step, all developed QMRFs were evaluated in terms of the availability of information on each of the OECD principle sections (sections 3–8), Table 2. The presence/absence of important information for the correctness of the particular model was assessed using +/−, whereas the presence/absence of additional information (comments) was marked using ✓/✗. Next, the results were analyzed in terms of the suitability of the developed QMRFs for regulatory purposes and the possibility of easily repeating the models and predicting missing values for new compounds. For the comparison between the collected models, we considered only the important information on the QMRF (bold in

Table 1 Endpoint list of available QSAR/QSPR models for PFASs

Physicochemical endpoints	Vapor pressure	VP1,[37] VP2,[38] VP3,[39] VP4[40]
	Water solubility	S1,[37] S2,[38] S3[39]
	Octanol–water partition	K_{OW}[39]
	Air–water partition	K_{AW}[39]
	Octanol–air partition coefficient	K_{OA}[39]
	Fluid–fluid interfacial adsorption coefficient	K_i[41]
	Melting point	MP1,[42] MP2,[42] MP3,[42] MP4[42]
	Boiling point	BP1,[42] BP2,[42] BP3,[42] BP4[42]
	Critical micelle concentration	CMC[37]
	Defluorination factor	DF[43]
	C–F bond dissociation energy	CFDE[44]
Toxicological endpoints	T4-TTR binding (TTR)	IC50(3)[45]
	Acute inhalation toxicity (*Rattus*, *Mus musculus*)	LC50(1),[38] LC50(2),[38] LC50(3),[46] LC50(4)[46]
	Acute oral toxicity (*Rattus*, *Mus musculus*)	LD50(1),[38] LD50(2),[47] LD50(3)[47]
Ecotoxicological endpoints	Cytotoxicity (*Xenopus tropicalis*)	IC50(1)[48]
	Developmental toxicity (*Danio rerio*)	IC50(2)[49]
	Toxic effect on root elongation (*Lactuca sativa, Pseudokirchneriella subcapitata*)	EC50(1),[50] EC50(2)[50]
	Acute toxicity (*Pseudokirchneriella subcapitata, Chlorella vulgaris, Daphnia magna, Danio rerio*)	EC50(3),[51] EC50(4),[51] LC50(5),[51] LC50(6)[51]

Table 2), not additional (comments). The OECD principles were evaluated one by one and five color-coded classes were established according to the percentages of the presence of necessary information in QMRFs. The thresholds of these classes were as follows: green (100–80%), gray (79–60%), yellow (59–40%), light orange (39–20%), and dark orange (19–0%). In the next step, we compile the principles together and conclude about the suitability of the available models for repeating and applying them in prediction for new compounds.

3. Evaluation of existing QSAR/QSPR models for PFASs

When preparing QMRF which may be useful in making predictions for new compounds, they must contain all key information about the developed model that will allow its reconstruction. The aim of this work was to verify if the available developed QSAR/QSPR models are scientifically valid and can be easily applied for predicting PFAS's properties. Considering on one side the great potential in the application of the PFAS in different fields and on the other side the possibilities of adverse effects on the environment and humans the predictive models can help to understand how properties/activities are related to its structure composition, and therefore allow one to verify which structures have similar properties, but at the same time (with the use of appropriate models) will be less or non-toxic. Here, we have collected 38 QSAR/QSPR models. Most of them relate to the physicochemical properties of PFASs compared to predictive (eco)toxicological modeling studies.

The QMRF format was proposed in 2007, and all models collected in this work were developed after 2007, however surprisingly, only 5 of them have provided QMRF documents available in the *QSAR Database*[35] (models for the following endpoints: VP2, S2, LC50(1) – acute inhalation toxicity, LC50 (2) – acute inhalation toxicity, and LD50(1) – acute oral toxicity). In addition, the QMRF document for LC50(1) has also entered the JRC database.[24] The analysis itself shows that in most developed models the authors did not consider their application for regulatory purposes, but rather would like to explain the processes and relationships between the structure of PFASs and their properties. However, for the model to be used to predict new compounds, QMRFs must be available. Therefore, for the 33 collected models, we have completed the QMRFs using the information provided in the publications. Next, we evaluated each QMRF in terms of fulfillment of the OECD principles and verified if all necessary information to repeat the model is available in the paper.

3.1. OECD principle 1

The first OECD principle (section 3 in the QMRF document) contains all information related to the predicted endpoint. The aim of this principle is "to ensure clarity in the endpoint being predicted by a given model since a given endpoint could be determined by different experimental protocols and under different experimental conditions". For regulatory purpose the endpoint should follow the pre-defined classification proposed by OECD.[23] All models collected in this work have clearly defined species (if required) and endpoints, and their units are provided (Table 2). Only one model[44] has no information on the processing of the experimental raw data (*e.g.*, a transformation of the endpoint). More than half of the collected models have indicated the experimental protocol providing important information about experimental conditions which could affect the measurements and thereby predictions. The majority of the available models (33 out of 38) have provided details on the endpoint data quality and variability, which allows an end-user to judge the quality of the experimental data. Overall, 21 of the 38 models completely fulfill the first OECD principle and 16 have one missing element (experimental protocol or data quality and variability) – green color in Table 2, whereas one model (CFDE) contains no additional information except endpoint and its unit – yellow color in Table 2. Summing up, the vast majority of the available models fulfill the first OECD principle established for the QSAR models.

3.2. OECD principle 2

The second OECD principle (section 4 in the QMRF document) concerns transparency in the description of the model algorithm. All necessary information for the estimation of the endpoint values and the reproduction of the model should be highlighted in the model's description. All collected models have indicated the type of modeling method applied in the study and the description of the models (please refer to Table S1, in ESI 1†). In the majority of these models, the authors used simple linear regression or multiple linear regression methods (MLR);[52] four of them (VP4, MP2, BP2, IC50(3)) are based on the partial least squares regression (PLS) method,[53] whereas three (CFDE, MP4, BP4) used Random Forest,[54] LASSO Regression, or Feed-forward Neural Networks models.[44] Only one model (IC50(3) – acute inhalation toxicity) does not contain an explicit definition of the algorithm including definitions of all descriptors. All models have provided detailed justification on how descriptors were selected. In the case of algorithm and descriptor generation, the majority of models contained all necessary information. The most frequently used method employed for obtaining the lowest energy conformation for PFASs was the semi-empirical AM1 method. The overwhelming majority of models used 1D, and 2D descriptors for describing the relationship between the activities/properties and the structure of the PFASs; however in several cases 3D descriptors were also taken into account (*e.g.*, total energy, or lowest unoccupied molecule orbital energy). Descriptors were calculated using different softwares (PaDEL,[55] DRAGON,[56] Mopac 2012[57] and 2016,[58] Chemaxon,[59] ACD/Labs,[60] JBSMM software,[61] Molecular Operating Environment,[62] OCHEM,[63] and EPI Suite[64]). 37 models have indicated the ratio of the number of chemicals (in the training set) to a number of descriptors. Analyzing the fulfillment of the second OECD principle, it could be concluded that similarly to the first principle the majority of the available QSAR/

Table 2 Results of QMRF evaluation of the collected models

		VP1	VP2	VP3	VP4	S1	S2	S3	Kow	Kaw	Koa	Ki	MP1	MP2	MP3	MP4	BP1	BP2	BP3	BP4	CMC	DF	CFDE	IC50 (3)	LC50 (1)	LC50 (2)	LC50 (3)	LC50 (4)	LD50 (1)	LD50 (2)	LD50 (3)	IC50 (1)	IC50 (2)	EC50 (1)	EC50 (2)	EC50 (3)	EC50 (4)	LC50 (5)	LC50 (6)
OECD principle 1	**1.1 Species**	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	1.2 Endpoint	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	1.3 Comment on endpoint	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	1.4 Endpoint units	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	1.5 Dependent variable	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	1.6 Experimental protocol	+	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–	–	+	+	–	+	–	+	–	–	+	+	+	–	–	+	+	+	+	+	+	+
	1.7 Endpoint data quality and variability	+	–	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+
OECD principle 2	**2.1 Type of model**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	2.2 Explicit algorithm	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	2.3 Descriptors in the model	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	2.4 Descriptor selection	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	2.5 Algorithm and descriptor generation	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	–	+	+	+	+	+	+	+	+	+	–	–	+	+	+
	2.6 Software name and version for descriptor generation	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	+	+	+
	2.7 Chemicals/ descriptors ratio	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
OECD principle 3	**3.1 Description of the applicability domain of the model**	+	+	–	+	+	+	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	–	–	–	–
	3.2 Method used to assess the applicability domain	+	+	–	+	+	+	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	–	–	–	–	–
	3.3 Software name and version for applicability domain assessment	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
	3.4 Limits of applicability	+	+	–	+	+	+	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
OECD principle 4	**4.1 Availability of the training set**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	4.2 Available information for the training set	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	4.3 Data for each descriptor variable for the training set	+	+	+	+	+	+	+	+	+	+	–	–	–	–	+	–	–	–	–	+	–	–	+	+	+	+	+	+	+	+	+	–	+	+	+	+	+	+

Table 2 *(Contd.)*

	VP1	VP2	VP3	VP4	S1	S2	S3	Kow	Kaw	Koa	Ki	MP1	MP2	MP3	MP4	BP1	BP2	BP3	BP4	CMC	DF	CFDE	IC50 (3)	LC50 (1)	LC50 (2)	LC50 (3)	LC50 (4)	LD50 (1)	LD50 (2)	LD50 (3)	IC50 (1)	IC50 (2)	EC50 (1)	EC50 (2)	EC50 (3)	EC50 (4)	LC50 (5)	LC50 (6)
4.4 Data for the dependent variable for the training set	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4.5 Other information about the training set	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
4.6 Pre-processing of data before modelling	+	+	+	+	+	+	+	+	+	+	+	−	+	−	−	−	+	−	−	+	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4.7 Statistics for goodness-of-fit	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4.8 Robustness – statistics obtained by leave-one-out cross-validation	+	+	+	−	+	+	+	+	+	+	−	+	+	+	−	+	+	+	−	+	+	−	−	+	+	+	+	+	+	+	+	−	−	+	+	+	+
4.9 Robustness – statistics obtained by leave-many-out cross-validation	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	+	−	−	−	−	−	−	−	−	−
4.10 Robustness – statistics obtained by Y-scrambling	+	+	−	−	+	+	−	−	−	−	−	+	−	−	−	+	−	+	−	+	+	−	−	+	+	+	+	+	+	+	+	−	−	−	−	−	−
4.11 Robustness – statistics obtained by bootstrap	+	−	−	−	+	−	−	−	−	−	−	+	−	−	−	+	−	−	−	+	−	−	−	−	−	+	+	−	+	+	−	−	−	−	−	−	−
4.12 Robustness – statistics obtained by other methods	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
4.13 Availability of the external validation set	+	+[a]	−	+	+	+[a]	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+[a]	+[a]	+	+	+[a]	+	+	+	+	−	−	−	−	−	−
4.14 Available information for the external validation set	✓	✓[a]	✗	✓	✓	✓[a]	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓[a]	✓[a]	✓	✓	✓[a]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
4.15 Data for each descriptor variable for the external validation set	+	+[a]	−	+	+	+[a]	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+[a]	+[a]	+	+	+[a]	+	+	+	−	−	−	−	−	−	−
4.16 Data for the dependent variable for the external validation set	+	+[a]	−	+	+	+[a]	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	+[a]	+[a]	+	+	+[a]	+	+	+	+	−	−	−	−	−	−
4.17 Other information about the external validation set	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
4.18 Experimental design of test set	+	+	−	−	+	+	−	−	−	−	+	+	+	+	+	+	+	+	+	−	−	−	−	+	+	+	+	+	+	+	+	+	−	−	−	−	−

Table 2 (Contd.)

		VP1	VP2	VP3	VP4	S1	S2	S3	Kow	Kaw	Koa	Ki	MP1	MP2	MP3	MP4	BP1	BP2	BP3	BP4	CMC	DF	CFDE	IC50 (3)	LC50 (1)	LC50 (2)	LC50 (3)	LC50 (4)	LD50 (1)	LD50 (2)	LD50 (3)	IC50 (1)	IC50 (2)	EC50 (1)	EC50 (2)	EC50 (3)	EC50 (4)	EC50 (5)	LC50 (6)
	4.19 Predictivity – statistics obtained by external validation	+	+		+	+	+	–	–	–	–	–	+	+	+	+	+	+	+	+	–	+	–	–	+	+	+	+	+	+	+	+	–	–	–	–	–	–
	4.20 Predictivity – assessment of the external validation set	–	+	–	–	–	+	–	–	–	–	–	+	–	–	–	–	–	–	+	–	–	–	–	+	–	–	+	–	–	–	–	–	–	–	–	–	–
	4.21 Comments on the external validation of the model	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
OECD principle 5	5.1 Mechanistic basis of the model	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	5.2 A priori or a posteriori mechanistic interpretation	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	+	+	+	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	5.3 Other information about the mechanistic interpretation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
	SUMMARY	SV	SV	NS	SV	SV	SV	NS	NS	NS	NS	SV	SV	SV	SV	NS	SV	SV	SV	NS	SV	NS	NS	NS	SV	SV	SV	SV	SV	SV	SV	SV	SV	NS	NS	NS	NS	NS	NS

Using +/–, the presence/absence of important information for the correctness of the particular model was assessed. Using ✓/✗, the presence/absence the additional information (comments) was marked; — not required. [a] Information included in ready QMRF but not available in the publication. The fulfilment of each of the OECD principles: color green (100–80%), yellow (79–60%), gray (59–40%), light orange (39–20%), and dark orange (19–0%). The SV – model is scientifically valid. The SN – model is not scientifically valid.

QSPR models indicate necessary information regarding the unambiguous algorithm. 32 of the 38 models provided all information, and 3 models have one missing element (descriptors in the model or algorithm and descriptor generation or chemical/descriptor ratio) – green color in Table 2, another three of them have no information about two elements (algorithm and descriptor generation and explicit algorithm or software name and version for descriptor generation) – gray color in Table 2.

3.3. OECD principle 3

The third OECD principle (section 5 in the QMRF document) states that every suitable QSAR model should have a defined domain of applicability. The applicability domain (AD) describes the boundaries of a theoretical chemical space where the predictions are plausible. Since each QSAR model is calibrated with a defined number of compounds, the quality of predictions for new compounds largely depends on their structural similarity to the compounds used to calibrate the model. AD is an area where the predictions for new chemical compounds are the result of interpolation and are therefore considered much more precise. Models with defined AD will satisfy the regulatory requirements. In the present work, 25 of the 38 models have described the AD; therefore an end-user can decide whether the model is applicable to a specific chemical of interest or not. Similarly, 25 collected models have also indicated the method used in defining AD. In most cases (18/25) there were the leverage approach and Williams plot;[28] however Euclidean-based AD and Standardization-based AD methods (IC50(1) – cytotoxicity and IC50(2) – developmental toxicity),[65] Residual Standard Deviation (the Euclidean distance) and the leverage (the Mehalanobis distance) (MP2, BP2), standard deviation of ensembles of neural network models (MP4, BP4) and distance approach (IC50(3) – T4-TTR binding) were also applied. The information on the software for defining the AD and the limits of the applicability of the model was provided for 22 and 32 models, respectively. Summarizing the third OECD principle, it is clearly seen that in more than half of the verified models defining the applicability domain was the intended purpose. 25 out of 38 models have provided all necessary information of the performance of the model – green in Table 2, 3 models have no information about the method used to assess the applicability domain – gray in Table 2, and other 10 collected models have no defined AD or have very rudimentary information – light and dark orange in Table 2.

3.4. OECD principle 4

Following the 4th OECD principle (sections 6 and 7 in the QMRF document), every suitable QSAR model should have performed a statistical validation to establish the performance of the model. In this step, the internal and external validation should be carried out and the appropriate measures of goodness-of-fit and robustness (internal performance) and predictivity (external performance) should be presented. Two sections of the QMRF document relate to this principle (section 6

– internal validation and section 7 – external validation, please see ESI 2†). In the internal validation section, all information about the training set and measures of goodness-of-fit should be specified, whereas the external validation section is related to the validation set and predictivity of the developed model. One of the most important criteria for evaluating QSAR models is the diversity and size of training and validation sets and their similarity to each other. We have analyzed training and validation sets of the gathered models in this work. We have provided the numbers of compounds in each set and the variety of compounds found there (group of PFAS). All gathered information is collected in Table S1 in ESI 1.† Analyzing the QMRFs developed in this study, almost all of the collected models (36 indeed) have indicated the list of PFASs available in the training set simultaneously dependent variable for these compounds but only 27 contain data for descriptors. In some cases, there were indicated sets of data but with no assignment of compounds to training and validation sets – in such a situation we marked that in this model data were not available (Table 2 section 4.12). Therefore, the availability of data for an external validation set was very small – only in 15 cases the list of external validation compounds was available, and all of them contain information on the modeled endpoints; however, only 13 indicated the values of the descriptors used in the models. It is very surprising because such data are necessary to reproduce and apply the QSAR/QSPR model to estimate the endpoint values for new compounds. The measure of goodness-of-fit and robustness were the next parameters that we evaluated. In almost all models (except one IC50(3) – T4-TTR binding) the determination coefficient (R^2) on the prediction for the training set was calculated. In the majority of analyzed models (30 models) internal validation was performed with the cross-validation leave-one-out (LOO) technique[50,51] and the robustness of the model was expressed by the cross-validation coefficient (Q^2_{CV}). Additionally, in five cases the leave-many-out method[28] was applied, whereas 18 models used also the Dependent Variable Scrambling[66] test to reduce the possibility of correlation by chance and to confirm the statistical significance of the developed models. Summing up, 32 collected QSAR models provided the necessary information to fulfill the 4th OECD principle (green, gray, and yellow color in Table 2). These models can be easily reproduced because they have available training and validation data and all statistical parameters. Four models (acute toxicity: EC50(3), EC50(4), LC50(5), and LC50(6)) can be reproduced (data for the training set are available) but the models were not externally validated. The rest of the models cannot be easily repeated (dark orange in Table 2).

3.5. OECD principle 5

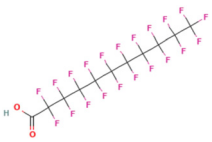
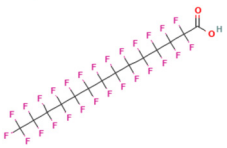
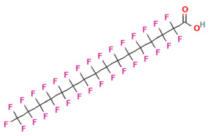
The OECD principle 5 for validation of (Q)SARs says that every developed QSAR model should have a mechanistic interpretation, if possible. Therefore, the structural features of chemical compounds used in the model should be described in the context of their influence on a given (modeled) property. The purpose of this principle is therefore to ensure that the mechanistic relationships between the descriptors used in the model and the predicted endpoint are assessed and to document each assessment. In our analysis, only two models do not contain any information on the interpretation of the modeled relationship (MP4 and BP4). All other models have provided the mechanistic basis of the model or mechanistic interpretation and fulfill the 5th OECD principle – green color in Table 2. Most QSAR/QSPR models found by us in the literature are based on statistical dependences and they give only the physicochemical interpretation of the descriptors used in the model. They indicated mostly the relationship between the structural features of the compounds and the modeled properties/activities; however, they do not focus on indicating the relationship between the structure of the PFAS and the molecular initiating events (MIE) related to the adverse outcome pathway (AOP). Only Zhang *et al.* 2021[51] provided a deeper interpretation and explained the mechanism of acute toxicity (MOA). We covered this information in Table 2 (point 5.3 other information about the mechanistic interpretation).

Summarizing the evaluation of all available QSAR/QSPR models in terms of fulfilling five OECD principles it could be stated that 6 out of 38 (VP2, S2, acute inhalation toxicity – LC50(1–2), acute oral toxicity LD50(1), cytotoxicity IC50(1)) are scientifically valid – they contained all information necessary to reproduce the model and predict endpoints values for the new compounds. 16 (VP1, VP4, S1, Ki, MP(1–3), BP(1–3), CMC, LC50(3–4) – acute inhalation toxicity, LD50(2–3) – acute oral toxicity, and IC50(2) – developmental toxicity) did not have some information, but the models are also reproducible. In the case of the other 16 models, they do not have details on many more items, and therefore, probably should not be used to predict the value of these endpoints for the new compounds. The presence of QSAR models built on PFAS mixtures is worth mentioning here. This model for mixtures[48] cannot be used for the prediction of the new single PFASs. However, this paper aimed to review possibly all QSAR/QSPR models related to PFAS, and evaluate their possibility of reproduction, and therefore we also included it. This model is very important in terms of mixtures which are more and more often considered in the PFAS assessment and is the subject of many currently conducted research studies in European projects (*e.g.* in PARC, PROMISCESS).

3.6. External testing of the predictivity of the gathered models

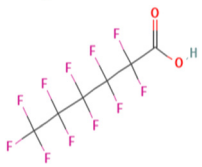
To verify the predictive ability of the gathered models and point out their limitations, the external testing of available models using the new compounds was performed. First, we searched for additional experimental data in the literature and databases (Norman database[67] and ITRC PFAS Team[68]). Although there are about 5000 PFAS, there are additional experimental data for a very small number of them. Concerning the endpoints for the available PFAS QSAR/QSPR models, we have found the external experimental data only for two endpoints (water solubility and vapor pressure) for three compounds (Tables 3 and 4). We used the equations proposed

Table 3 Results of external testing of the model for water solubility (S1)

Acronim	CAS number	Structure	Carbon chain length	Exp. water solubility[b] log (mg L^{-1}) 20 °C ± 0.5	Pred. water solubility[c] log (mg L^{-1})[a]
PFUnA	2058-94-8		11	−0.22	−2.0282875
PFTeDA	376-06-7		14	−0.53	−5.678205
PFHxDA	67905-19-5		16	−0.82	−8.95235

[a] $\log Aq S = -0.418(\pm 1.940) - 0.003(\pm 0001)T(F..F) + 5.185(\pm 3.849)SIC1$. [b] Inoue *et al.*[70]. [c] Bhhatarai and Gramatica[37].

Table 4 Results of external testing of the model for vapor pressure (VP1)

Acronim	CAS number	Structure	Carbon chain length	Exp. vapor pressure log (Pa)	Pred. vapor pressure log (Pa)[a,d]
PFPrA	422-64-0		3	3.59[b]	3.31
PFPeA	2706-90-3		5	3.43[b]	2.6
PFHxA	307-24-4		6	1.1[c]	2.18

[a] $\log VP\ (mmHg) = 7.97 - 0.16 \times F03[C\text{–}F] - 3.16 \times ACC - 0.64 \times nDB$. [b] Kwan[71]. [c] Zhang *et al.*[72]. [d] Bhhatarai and Gramatica[37].

by the authors (please refer to ESI 2†) and predicted endpoint values for three new chemicals, which were not included in the training sets. Predictions of water solubility obtained by Bhhatarai and Gramatica (S1)[37] have been compared with experimental data from the literature for three PFASs, which are not included in the training set of the model. It is worth mentioning that the training set of the model contains only three carboxylic acids (8–10 carbon atoms), whereas the external compounds consist of 11–16 carbon atoms; this difference may affect the predictions. The model predicts water solubility for PFASs at 25 °C; however, the external experimental data for CAS 2058-94-8, CAS 376-06-7, and CAS 67905-19-5 were determined according to OECD 103 at 20 °C ± 0.5. Despite the temperature differences, the descending trend can be seen while the chain length of carboxylic acid increases. In summary, the differences between the predicted and experimental endpoints, (found in the literature) are due to the fact that the data from the literature differ significantly from the data selected by the authors for building the model.

Similarly, as in the case of water solubility, the model for predicting vapor pressure (VP1) was implemented for three compounds that are not included in the training set (CAS 422-64-0, CAS 2706-90-3, and CAS 307-24-4). Both experimental data found for three perfluorocarboxylic acids and data used by Bhhatarai and Gramatica[37] to build the model were obtained at 25 °C. However, the differences in the log(Pa) values may be due to the different methods and conditions of data collection. Despite these differences, a descending trend in vapor pressure can be observed as a function of the increasing number of carbon atoms in the PFAS main chain. External

compounds consist of 3–5 carbon atoms; however, eight carboxylic acids in the training set of the model include 2–12 carbon atoms in the main chain, which leads to correct prediction.

4. Future directions and perspective

The presented review aimed to verify if the QSAR/QSPR models for PFASs available in the literature are ready to be used for new compound prediction and are scientifically valid. This review was based on the 38 QSAR/QSPR models available in the literature (for 33 we prepared QMRF documents because they were not available). The unavailability of QMRF documents shows that the authors of the developed models wanted to use them to explain the mechanism/dependence of a given property on the chemical structure rather than to use models themselves to predict the properties/activities of new compounds that can be further registered. The evaluation of individual QMRFs was mainly based on the assessment of the availability of the necessary information on individual OECD principles. The analysis showed that there are no major problems with the fulfillment of the first and second OECD principles. The modeled endpoint and dependent variable are correctly defined. The authors of the models mostly provide the model equation and descriptors (which and how they were generated) which are used to determine the property/activity-structure relationship. Similarly, in the majority of collected models, the authors defined the preliminary mechanistic basis and interpreted the statistical relationships, thus fulfilling the 5th OECD principle previously established in 2004. However, nowadays, with the increasing need for knowing the mechanisms causing the adverse outcomes (AO), the developed QSAR models predicting the toxicological endpoints should also follow this trend and try to give a deeper mechanistic interpretation, which is in line with being developed/developed AOPs. Therefore, it may be necessary in the future to update the 5th OECD principle and indicate exactly what interpretation should be included in a properly developed QSAR model.

The 3rd and 4th OECD principles showed the most shortcomings. Regarding the 3rd OECD principle only nearly half of the prepared QMRFs described the applicability domain of the developed models. This is very surprising since each good QSAR/QSPR model should have a defined space of validity. In another way, the model could be used to predict chemicals for which the predictions could be unacceptably unreliable. Moreover, considering the 4th OECD principle many collected models have not been correctly validated or have not provided all required parameters. A variety of statistics validation techniques are available to assess the robustness and predictability of models, and different parameters are now routinely used to express these aspects of model performance. They are the standards that the developed QSAR models should follow. Another issue is related to the availability of the data with which the model was calibrated and validated. Providing the information on endpoints and descriptors values for training and validation sets is required to reproduce the model and properly predict the values for new compounds. It is an unwritten standard in QSAR/QSPR model building. However, more than half of the available models for PFASs contain this information.

Summing up, more than half (22) of the collected models are scientifically valid based on the OECD principles and are ready to be used to predict the properties of new compounds. The rest of the models can be used to gain knowledge about the studied phenomenon, but they cannot be used to register new compounds, *e.g.*, derivatives of PFAS, which would not have a negative impact on the environment and human life, simultaneously maintaining the desired properties. The scientific validity of the QSAR/QSPR model is the condition *sine qua non* for regulatory acceptance for using such a model for the prediction of the new compounds.

The present study shows two major issues when analyzing the available predictive models dedicated to PFAS. Firstly, existing models are very limited in helping to characterize or assess the environmental fate and transport of PFAS. They do not focus on the relevant physicochemical endpoints in this field. Because PFASs are widely used in commercial and industrial products they have been frequently detected in industrialized and developing countries in drinking water, surface water, and groundwater across. Their solubility in water (especially short chain PFASs) is high, they are often persistent during degradation and treatment, and the understanding of their degradation products and toxicity is limited. Therefore, an innovative integrated modeling approach to predict the transport pathways and fate of PFASs in different environmental compartments (*e.g.*, soil, sediment, groundwater, and surface water) is needed. *In silico* predictive models (especially QSPR approaches) can be here helpful in generating inputs to the fate and transport models. However, the model endpoints here should follow the needs required by the fate and transport models. For example, there are available QSPR models dedicated to vapor pressure, whereas Henry's Law may be of more appropriate and environmental relevance. There is still a need to develop predictive, scientifically valid models for all partition coefficients (octanol–water, air–water, air–octanol), degradation rates (biodegradation, abiotic degradation, and photodegradation), soil toxicity and bioconcentration factor for the whole group of PFASs, so that it is possible to model the fate and transport of those groups.

The other issue connected with the development of predictive models is the availability of experimental data for model training. Although PFASs are a very large group of compounds (+5000), experimental data for relevant environmental and toxicity endpoints are available only for a small group of them, which does not represent equally different groups/classes of these compounds. In such a case, the solution is to develop local QSAR/QSPR models for a single class of compounds, where it is possible to use the quantum chemistry method to simulate the physicochemical endpoint's values for model training. Of course, in the second approach also several experimental data are needed to validate/support theoretical calculations. Moreover, experimental data are also needed for exter-

nal validation of available models, to determine their predictivity in new chemical space. It is very important due to the fact that a majority of available models are dedicated only to the PFAS compounds (the training sets contain also other pollutants); therefore they cannot be expected to work properly for every PFAS. In fact, it should be verified for which groups of PFASs the available models work. In this case, it would be necessary to select a truly external set of compounds, several PFASs belonging to different groups, then conduct experimental studies and compare them with the results obtained by applying QSAR/QSPR models. In this way, it will be possible to show how predictive these models are, and what are the limitations vis a vis the type of compound in the external dataset.

Secondly, the QSAR models available in the literature focused mainly on acute toxicity (EC50 or LC50), and they do not indicate a clear relationship between the structure of the PFAS and the adverse outcome pathway (AOP) and molecular initiating events (MIE). In fact, they only explain the basic structure/activity relationship (statistical approach) but do not indicate the real mechanism of action. Recent studies[69] indicate that exposure to PFASs may have a negative impact on all components of metabolic syndrome. This was proved not only on individual PFASs but also mixtures of these compounds. Taking into account these studies, an appropriate method of modeling mixtures should be developed. Such steps were taken in the newly established Partnership for the Assessment of Risk from Chemicals (PARC) which is an innovative research program to support EU and national institutions involved in chemical risk assessment and risk management. All the above-mentioned issues should be considered in further development of the predictive models to be valuable and applicable in the human and risk assessment of these compounds.

Author contributions

A. S. and T. P. formulated the research problem and main hypotheses. A. S. designed the study. N. B. and D. K. conducted the research. All authors contributed to the analysis and discussion. A. S., N. B., and D. K. wrote the manuscript. All authors edited the manuscript and accepted its final version.

Conflicts of interest

The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors would like to express gratitude to Dr Andrew Worth and Dr Cecile Valsecchi from the EU Joint Research Center, Ispra, Italy, and three Anonymous Reviewers for valuable comments that helped to improve the manuscript. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101036449.

References

1 https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO(2018)7&doclanguage=en.
2 E. Kissa, *Fluorinated Surfactants and Repellents*, 2001.
3 H. D. Whitehead, M. Venier, Y. Wu, E. Eastman, S. Urbanik, M. L. Diamond, A. Shalin, H. Schwartz-Narbonne, T. A. Bruton, A. Blum, Z. Wang, M. Green, M. Tighe, J. T. Wilkinson, S. McGuinness and G. F. Peaslee, *Environ. Sci. Technol. Lett.*, 2021, **8**, 538–544.
4 M. M. Peden-Adams, J. M. Keller, J. G. EuDaly, J. Berger, G. S. Gilkeson and D. E. Keil, *Toxicol. Sci.*, 2008, **104**, 144–154.
5 Y. Xu, Y. Li, K. Scott, C. H. Lindh, K. Jakobsson, T. Fletcher, B. Ohlsson and E. M. Andersson, *Environ. Res.*, 2019, **181**, 108923.
6 K. Roth, Z. Imran, W. Liu and M. C. Petriello, *Front. Toxicol.*, 2020, **2**, 601149.
7 G. Liu, K. Dhana, J. D. Furtado, J. Rood, G. Zong, L. Liang, L. Qi, G. A. Bray, L. DeJonge, B. Coull, P. Grandjean and Q. Sun, *PLoS Med.*, 2018, **15**, e1002502.
8 L. G. Kahn, C. Philippat, S. F. Nakayama, R. Slama and L. Trasande, *Lancet Diabetes Endocrinol.*, 2020, **8**, 703–718.
9 P. A. Behnisch, H. Besselink, R. Weber, W. Willand, J. Huang and A. Brouwer, *Environ. Int.*, 2021, **157**, 106791.
10 R. Foguth, M. S. Sepúlveda and J. Cannon, *Toxics*, 2020, **8**, 42.
11 K. Steenland and A. Winquist, *Environ. Res.*, 2020, **194**, 110690.
12 T. Reemtsma, U. Berger, H. P. H. Arp, H. Gallard, T. P. Knepper, M. Neumann, J. B. Quintana and P. de Voogt, *Environ. Sci. Technol.*, 2016, **50**, 10308–10315.
13 SC-4/17: Listing of perfluorooctane sulfonic acid, its salts and perfluorooctane sulfonyl fluoride.
14 SC-9/12: Listing of perfluorooctanoic acid (PFOA), its salts and PFOA-related compounds.
15 Commission Regulation (EU) 2021/1297 of 4 August 2021 amending Annex XVII to Regulation (EC) No 1907/2006 of the European Parliament and of the Council as regards perfluorocarboxylic acids containing 9 to 14 carbon atoms in the chain (C9–C14 PFCAs), their salts and C9–C14 PFCA-related substances.
16 Communication from The Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of The Regions. Pathway to a Healthy Planet for All EU Action Plan: "Towards Zero Pollution for Air, Water and Soil" (COM (2021) 400 final).
17 https://www.regjeringen.no/contentassets/1439a5cc9e82467385ea9f090f3c7bd7/fluor–eu-strategy-for-pfass–december-19.pdf.

18 Final draft: Fourth report under Article 117(3) of the REACH Regulation (2020) (europa.eu).
19 M. T. D. Cronin and T. W. Schultz, *J. Mol. Struct.: THEOCHEM*, 2003, **622**, 39–51.
20 European Chemicals Agency, *The use of alternatives to testing on animals for the REACH Regulation*, European Chemicals Agency, 2021, **https://data.europa.eu/doi/10.2823/092305**.
21 OECD, *OECD Principles for the Validation, for Regulatory Purposes, of Quantitative Structure-activity Relationship Models*, Paris, 2004.
22 OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. OECD Series on Testing and Assessment No. 69. ENV/JM/MONO(2007)2. Organisation for Economic 133 CHAPTER R.6 – QSARS AND GROUPING OF CHEMICALS 134 Cooperation and Development*, Paris, France, 2007a, 154 pp. **https://www.oecd.org/document/30/0,2340,-en26493436519166381111,00.html**.
23 European Chemicals Agency, *Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals*, 2008.
24 **https://ec.europa.eu/jrc/en/scientific-tool/jrc-qsar-model-database**.
25 **https://qmrf.sourceforge.net**.
26 R. Bro, K. Kjeldahl, A. K. Smilde and H. A. Kiers, *Anal. Bioanal. Chem.*, 2008, **390**, 1241–1251.
27 V. Consonni, D. Ballabio and R. Todeschini, *J. Chemom.*, 2010, **24**, 194–201.
28 P. Gramatica, *QSAR Comb. Sci.*, 2007, **26**, 694–701.
29 S. Wold, L. Eriksson and S. Clementi, *Chemometrics Methods in Molecular Design*, 1995, pp. 309–318.
30 P. Gramatica, *Int. J. Quant. Struct.-Prop. Relat.*, 2020, **5**, 61–97.
31 K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das and H. Kabir, *J. Chem. Inf. Model.*, 2012, **52**, 396–408.
32 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2011, **51**, 2320–2335.
33 R. Todeschini, D. Ballabio and F. Grisoni, *J. Chem. Inf. Model.*, 2016, **56**, 1905–1913.
34 G. Piir, I. Kahn, A. T. García-Sosa, S. Sild, P. Ahte and U. Maran, *Environ. Health Perspect.*, 2018, **126**(12), 126001.
35 **https://qsardb.org**.
36 P. Gramatica, N. Chirico, E. Papa, S. Cassani and S. Kovarich, *J. Comput. Chem.*, 2013, **34**, 2121–2132.
37 B. Bhhatarai and P. Gramatica, *Environ. Sci. Technol.*, 2010, **45**, 8120–8128.
38 P. Gramatica, S. Cassani and N. Chirico, *J. Comput. Chem.*, 2014, **35**, 1036–1044.
39 M. Kim, L. Y. Li, J. R. Grace and C. Yue, *Environ. Pollut.*, 2015, **196**, 462–472.
40 G. Ding, M. Shao, J. Zhang, J. Tang and W. J. G. M. Peijnenburg, *Atmos. Environ.*, 2013, **75**, 147–152.
41 M. L. Brusseau, *Water Res.*, 2019, **152**, 148–158.
42 B. Bhhatarai, W. Teetz, T. Liu, T. Öberg, N. Jeliazkova, N. Kochev, O. Pukalov, I. V. Tetko, S. Kovarich, E. Papa and P. Gramatica, *Mol. Inf.*, 2011, **30**, 189–204.
43 Z. Cheng, Q. Chen, Z. Liu, J. Liu, Y. Liu, S. Liu, X. Gao, Y. Tan and Z. Shen, *Environ. Sci. Technol. Lett.*, 2021, **8**, 645–650.
44 A. Raza, S. Bardhan, L. Xu, S. S. R. K. C. Yamijala, C. Lian, H. Kwon and B. M. Wong, *Environ. Sci. Technol. Lett.*, 2019, **6**, 624–629.
45 J. M. Weiss, P. L. Andersson, M. H. Lamoree, P. E. G. Leonards, S. P. J. van Leeuwen and T. Hamers, *Toxicol. Sci.*, 2009, **109**, 206–216.
46 B. Bhhatarai and P. Gramatica, *Chem. Res. Toxicol.*, 2010, **23**, 528–539.
47 B. Bhhatarai and P. Gramatica, *Mol. Diversity*, 2011, **15**, 467–476.
48 G. Hoover, S. Kar, S. Guffey, J. Leszczynski and M. S. Sepúlveda, *Chemosphere*, 2019, **233**, 25–33.
49 S. Kar, S. Ghosh and J. Leszczynski, *Chemosphere*, 2018, **210**, 588–596.
50 G. Ding, M. Wouterse, R. Baerselman and W. J. G. M. Peijnenburg, *Arch. Environ. Contam. Toxicol.*, 2011, **62**, 49–55.
51 J. Zhang, M. Zhang, H. Tao, G. Qi, W. Guo, H. Ge and J. Shi, *Molecules*, 2021, **26**, 6574.
52 A. M. Pires and I. M. Rodrigues, *State Med.*, 2007, **26**, 2901–2918.
53 A. L. Boulesteix and K. Strimmer, *Briefings Bioinf.*, 2007, **8**, 32–44.
54 J. L. Speiser, M. E. Miller, J. Tooze and E. Ip, *Expert Syst. Appl.*, 2019, **134**, 93–101.
55 C. W. Yap, *J. Comput. Chem.*, 2010, **32**, 1466–1474.
56 R. Todeschini, V. Consonni, A. Mauri and M. Pavan, **https://www.talete.it**.
57 J. J. P. Stewart, *MOPAC 2012. Stewart Computational Chemistry*, 2012, **http://OpenMOPAC.net**.
58 J. J. P. Stewart, *MOPAC 2016. Stewart Computational Chemistry*, 2016, **https://OpenMOPAC.net**.
59 **https://chemaxon.com**.
60 **https://www.acdlabs.com**.
61 JBSMM software. N. Kochev, O. Pukalov, Software system for molecular modeling–JBSMM; Scientific Researches of the Union of Scientists in Bulgaria – Plovdiv, Series C: Technics and Technologies, Vol. V., Balkan Conference of Young Scientists; Plovdiv, 16–18 June, 2005, pp. 315–320.
62 MOE (The Molecular Operating Environment), software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7. **https://www.chemcomp.com**.
63 **https://ochem.eu/home/show.do**.
64 EPI Suite™ – Estimation Program Interface, **https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface**.
65 K. Roy, S. Kar and P. Ambure, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22–29.
66 S. Wold and L. Eriksson, *Chemometric Methods in Molecular Design*, VCH and Weinheim, 1995.
67 Norman Database System, **https://www.norman-network.com/nds/**.

68 https://pfas-1.itrcweb.org.
69 A.-M. Kaiser, M. Z. Jeddi, M. Uhl, F. Jornod, M. F. Fernandez and K. Audouze, *Toxics*, 2022, **10**, 449.
70 Y. Inoue, N. Hashizume, N. Yakata, *et al.*, Unique Physicochemical Properties of Perfluorinated Compounds and Their Bioconcentration in Common Carp *Cyprinus carpio* L., *Arch. Environ. Contam. Toxicol.*, 2012, **62**, 672–680.
71 W. C. Kwan, *Physical property determination of perfluorinated surfactants*, PhD diss., National Library of Canada=Bibliothèque nationale du Canada, 2001.
72 M. Zhang, K. Yamada, S. Bourguet, J. Guelfo and E. M. Suuberg, Vapor Pressure of Nine Perfluoroalkyl Substances (PFASs) Determined Using the Knudsen Effusion Method, *J. Chem. Eng. Data*, 2020, **65**, 2332–2342.