Chemosphere 184 (2017) 514–523

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Endocrine-disrupting activity of per- and polyfluoroalkyl substances: Exploring combined approaches of ligand and structure based modeling

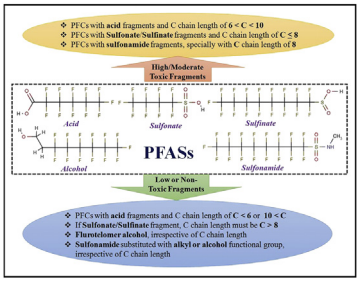
Supratik Kar [a], Maria S. Sepúlveda [b], Kunal Roy [c], Jerzy Leszczynski [a, *]

[a] Interdisciplinary Center for Nanotoxicity, Department of Chemistry and Biochemistry, Jackson State University, Jackson, MS, 39217, USA
[b] Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN, 47907, USA
[c] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

HIGHLIGHTS

- PFASs can compete with T4 for binding to TTR in resulting endocrine disruption.
- The T4-TTR competing potency of 24 PFCs was modeled employing QSAR.
- The docking study corroborates evidence of binding interactions with TTR.
- Important structural attributes of PFASs for endocrine disruption were identified.
- Developed models may serve as an efficient tool for screening of large databases.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 4 April 2017
Received in revised form
5 June 2017
Accepted 7 June 2017
Available online 9 June 2017

Handling Editor: I. Cousins

Keywords:
Docking
Endocrine disruption
PFASs
QSAR
Thyroid hormone
Thyroxine
Transport protein transthyretin

ABSTRACT

Exposure to perfluorinated and polyfluoroalkyl substances (PFCs/PFASs), endocrine disrupting halogenated pollutants, has been linked to various diseases including thyroid toxicity in human populations across the globe. PFASs can compete with thyroxine (T4) for binding to the human thyroid hormone transport protein transthyretin (TTR) which may lead to reduce thyroid hormone levels leading to endocrine disrupting adverse effects. Environmental fate and endocrine-disrupting activity of PFASs has initiated several research projects, but the amount of experimental data available for these pollutants is limited. In this study, experimental data for T4-TTR competing potency of 24 PFASs obtained in a radioligand-binding assay were modeled using classification- and regression-based quantitative structure-activity relationship (QSAR) tools with simple molecular descriptors obtained from chemical structure of these compounds in order to identify the responsible structural features and fragments of the studied PFASs for endocrine disruption activity. Additionally, docking studies were performed employing the crystal structure complex of TTR with bound 2′, 6′-difluorobiphenyl-4-carboxylic acid (PDB: 2F7I) in order to constitute the receptor model for human TTR. The results corroborate evidence for these binding interactions and indicate multiple high-affinity modes of binding. The developed *in silico* models therefore advance our understanding of important structural attributes of these chemicals and

* Corresponding author. Department of Chemistry and Biochemistry, Jackson State University, 1400 Lynch Street, Jackson, MS, 39217-0510, USA.
E-mail address: jerzy@icnanotox.org (J. Leszczynski).

http://dx.doi.org/10.1016/j.chemosphere.2017.06.024
0045-6535/

may provide important information for the design of future synthesis of PFASs as well as may serve as an efficient query tool for virtual screening of large PFAS databases to check their endocrine toxicity profile.

1. Introduction

Perfluorinated and polyfluoroalkyl substances (PFCs/PFASs), belonging to an emerging class of endocrine disrupting halogenated pollutants, and extensively used as surfactants, non-adhesives, waterproof fabrics, and fire-fighting foams, have been linked to thyroid toxicity in humans (Dreyer et al., 2009; Fromme et al., 2009). Their widespread use coupled with high environmental persistence and long-half life (4–7 years) has resulted in their ubiquitous presence in almost all types of environmental media and biota including humans For instance, in the United States general population, perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA) are detected in >95% of individuals at an average serum concentration of 3 and 9 ppb, respectively (Kato et al., 2015). Occupational and accidental environmental exposure, however, can result in serum average concentrations two to three orders of magnitude higher (Olsen, 2015).

Several lines of evidence point to PFASs as endocrine disruptors capable of altering the hypothalamus-pituitary-thyroid (HPT) axis of vertebrates. In rodents and monkeys, PFOS cause depression in thyroid hormone (TH) levels (Lau et al., 2007). Epidemiological studies have reported abnormal TH levels and thyroid diseases in children living close to a chemical plant and in the general population (Lopez-Espinosa et al., 2012). PFASs are highly persistent, bioaccumulative and toxic; and have been identified as chemicals of emerging concern at existing superfund and hazardous waste sites. They have been detected in aquatic and terrestrial animals, including amphibians, across the globe (Kannan et al., 2005). However the mechanisms leading to thyroid toxicity are not well understood.

Accepted mechanisms reported so far from *in vitro* and *in vivo* studies using different animal models suggest that PFASs interfere with the binding of THs to the thyroid hormone receptor (THR) (Ren et al., 2015) and to serum carrier proteins such as transthyretin (TTR), thyroid binding globulin (TBG) and albumin (Weiss et al., 2009). Laboratory studies have corroborated that PFASs can compete with thyroxine (T4) for binding to the human TTR which may lead to reduce thyroid hormone levels leading to endocrine disrupting activity. Exposure to PFASs occurs in the form of mixtures composed of compounds that vary in number of halogenated carbons, which directly impact half-lives and potency. Maternal transfer is a major pathway of PFAS exposure to embryos, fetuses and newborns. In summary, there is a real need to determine the mode of action and adverse outcomes elicited by PFASs during early life stage development. Globally, the chemical industry and Regulatory Agencies such as the United States Environmental Protection Agency (US EPA) spend millions of dollars in testing and assessing the health risks associated with chemicals. But, it is difficult to make decisions regarding exposure guidelines for environmental contaminants when insufficient experimental data are available regarding systemic toxicity (US EPA, 2009).

In such a situation, *in silico* approaches, explicitly quantitative structure-activity relationship (QSAR) models along with molecular docking studies have the capability to predict such health hazards potential from chemical exposure. These *in silico* studies not only save time but also precious resources which could be endowed more sensibly (Roy et al., 2015a, 2015b). Not only that, escalating demands from social and economic surroundings to cut out the use of animal testing is another reason to develop *in silico* models which also support 3Rs (replacement, refinement and reduction of animals in research) and Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) policies (Benigni and Giuliani, 2003; Williams et al., 2009). The European Chemical Bureau supports the exploitation of predictive models in a regulatory framework for making regulatory decisions and for risk screening assessments. Predictive classification-based QSAR models are used by US Food and Drug Administration (USFDA) to curtail false negatives and false positives (Benigni and Zito, 2004). Nevertheless, the models should be validated according to Organization for Economic Cooperation and Development (OECD) principles for dependable predictions (Roy and Kar, 2015a, 2015b).

A very limited number of QSAR models have been developed with PFAS. Simple descriptors from 2D molecular structures of PFASs were employed to make QSAR equations for the competitive potencies of the TTR-binding compounds by Weiss et al. (2009). Thereafter, Kovarich et al. (2012) developed local classification models for PFASs to predict their T4-TTR competing potency and encoded the key structural features involved in TTR binding affinity. In another study, Ren et al. (2015) investigated the binding interactions of 16 structurally diverse PFASs with human TTR and performed QSAR studies which revealed that fluorinated alkyl chain length longer than ten and an acid end group were optimal for TR binding. Additionally, docking analysis revealed that most of the tested PFASs efficiently fit into the T3-binding pocket in the THR. These studies are the examples how to overcome the problems of data gaps and to prove that *in silico* predictive techniques can be applied confidently with a consequent reduction of costs and numbers of tested animals. Although a handful of models have been developed, *in silico* approaches that combine QSAR and docking of PFASs with interacting amino acid residues in TTR have not yet been performed. Therefore, this study is the first of its kind where ligand and structure-based approaches are combined to explore the endocrine-disrupting activity of PFASs.

The objectives of this study were first to perform a comprehensive *in silico* study by employing a classification-based QSAR through linear discriminant analysis (LDA) to identify the discriminatory features between active and inactive endocrine disrupting PFASs; followed by a regression-based QSAR model to identify the structural fragments of PFASs responsible for the mentioned toxicity. Further, docking studies were also carried out to evaluate the binding sites of PFASs with TTR for predicting endocrine disruption toxicity. The complete computational framework is illustrated in Fig. 1. Note that the objective of the QSAR studies have been to identify the fragments responsible for the toxicity and their quantitative contributions while the docking study was performed to understand the interacting residues in the receptor site. Thus, these two approaches are complimentary.

2. Materials and methods

2.1. Dataset

Experimental data for T4-TTR competing potency of PFASs, measured *in vitro* by Weiss et al. (2009) were employed to develop

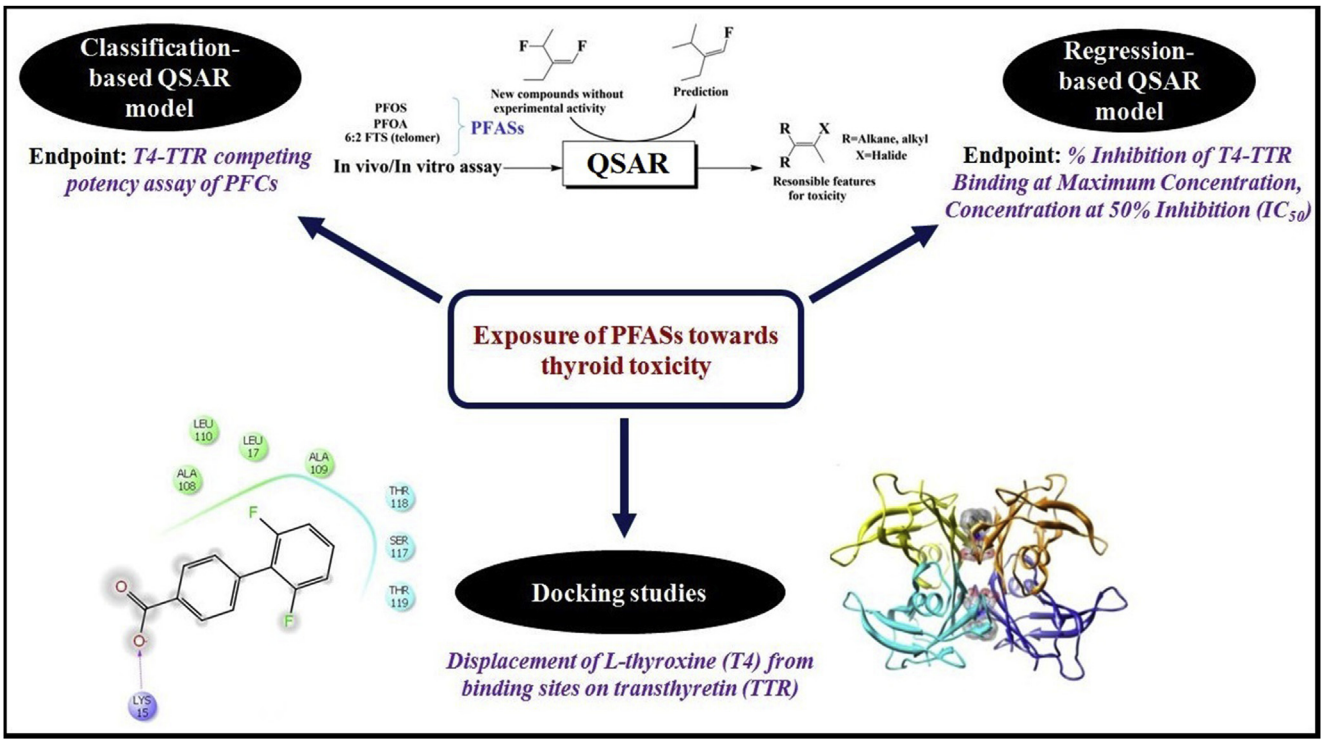


Fig. 1. Schematic diagram of complete computational framework employed in the study.

the QSAR models. The data set consisted of 24 PFASs including several perfluorinated alkyl acids (PFAA), sulfonates (PFOS), sulfonamides and telomere alcohols (FTOH) (See Table 1). Considering T4-TTR % competing potency at maximal concentration, 15

Table 1
Experimental and prediction assay values from the QSAR analysis along with the binding energies from the docking.

ID	Chemical	T4-TTR Binding %	Experimental Classification	Predicted classification[a]	Experimental pIC_{50} (mM)	Predicted pIC_{50} (mM) [b]	Binding energy (kcal/mol) [c]
1[d]	Perfluorobutyric acid	106	Inactive (N)	Active (Y)	n.d.	n.d.	−3.28
2[d,e]	Perfluorohexanoic acid	43	Active (Y)	Active (Y)	2.09	2.17	−4.24
3	Perfluoroheptanoic acid	7	Active (Y)	Active (Y)	2.81	2.62	−2.83
4	Perfluorooctanoic acid	4	Active (Y)	Active (Y)	3.02	2.77	−2.36
5[d]	Perfluorononanoic acid	18	Active (Y)	Active (Y)	2.56	2.57	−2.24
6[e]	Perfluorodecanoic acid	46	Active (Y)	Active (Y)	2.05	2.19	ND
7	Perfluoroundecanoic acid	74	Active (Y)	Active (Y)	1.67	1.82	ND
8	Perfluorododecanoic acid	91	Active (Y)	Active (Y)	1.33	1.47	ND
9[d,e]	Perfluorotetradecanoic acid	71	Active (Y)	Active (Y)	1.54	0.86	ND
10	7H-Perfluoroheptanoic acid	45	Active (Y)	Active (Y)	2.06	2.56	−3.81
11[d]	2H-Perfluoro-2-octenoic acid (6:2)	47	Active (Y)	Active (Y)	2.05	1.99	−3.10
12	Perfluorobutane sulfonate	69	Active (Y)	Active (Y)	1.71	1.44	−4.23
13[d]	Perfluorohexane sulfonate	3	Active (Y)	Active (Y)	3.14	3.13	−4.37
14[e]	Perfluorooctane sulfonate	1	Active (Y)	Active (Y)	3.03	3.02	−2.59
15	Perfluorodecane sulfonate	122	Inactive (N)	Active (Y)	n.d.	n.d.	ND
16[e]	Perfluorooctane sulfinate	6	Active (Y)	Active (Y)	2.76	2.57	−2.04
17	2-Perfluorohexyl ethanol	117	Inactive (N)	Inactive (N)	n.d.	n.d.	−1.10
18	2-Perfluorooctyl ethanol	117	Inactive (N)	Inactive (N)	n.d.	n.d.	−1.30
19[e]	2-(*N*-methylperfluoro-1-octane sulfonamido) ethanol	119	Inactive (N)	Inactive (N)	n.d.	n.d.	−1.48
20	2-(*N*-ethylperfluoro-1-octane sulfonamido) ethanol	122	Inactive (N)	Inactive (N)	n.d.	n.d.	ND
21	Perfluorooctane sulfonamide	32	Active (Y)	Active (Y)	2.21	2.21	−2.39
22[d]	*N*,*N*-dimethyl perfluorooctane sulfonamide	115	Inactive (N)	Inactive (N)	n.d.	n.d.	ND
23	*N*-methyl perfluorooctane sulfonamide	114	Inactive (N)	Inactive (N)	n.d.	n.d.	−0.94
24	*N*-ethyl perfluorooctane sulfonamide	112	Inactive (N)	Inactive (N)	n.d.	n.d.	−1.50

n.d.-Not detected, ND-Not docked.
[a] Based on classification-based QSAR equation.
[b] Based on regression-based QSAR equation.
[c] Based on docking study.
[d] Test set compounds for classification model.
[e] Test set for regression-based model.

out of 24 PFASs were identified as active/toxic (% binding value < 100, denoted as Y) and 9 as inactive/non-toxic (% binding value > 100, denoted as N). Therefore, all 24 PFASs were considered for the classification-based QSAR model. In order to develop regression-based QSAR models, T4-TTR competing potency measured as the ratio concentration resulting in 50% inhibition of T4 (IC_{50}) were transformed to pIC_{50} using a negative logarithmic function. For the docking study, we selected the crystal structure complex of TTR with bound 2,′6′-difluorobiphenyl-4-carboxylic acid (PDB: 2F7I) to constitute the receptor model for human TTR. To check the binding energy, all 24 PFASs were docked in the mentioned 2F7I protein.

2.2. Descriptor calculation

The structures of the studied PFASs were drawn using HyperChem (HyperChem[TM], Hypercube, Inc., USA) software and minimized to their lowest energy conformation using the semi-empirical AM1 method. The output file from AM1 was also employed to calculate basic electronic properties such as dipole-moment, highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, *etc*. Further, optimized geometrical structures were employed to calculate topological, structural, and constitutional, functional group counts, as well as several indices including information, extended topochemical atom (ETA), atom-centered fragments, and atom-type E-state molecular descriptors employing DRAGON software (DRAGON, 2011).

2.3. Dataset splitting

Selection of the training and test sets plays a crucial role in the construction of a statistically significant QSAR model. The selection should be such that the test set molecules lie within the chemical space occupied by the training set molecules. In this study, splitting of the dataset into training and test sets was performed based on the activity sorted response. As a result, two-thirds of the compounds were incorporated into the training set and the rest were incorporated into the test set ($N_{Training}$: 16 and 10, N_{Test}: 8 and 5 for classification and regression-based QSAR, respectively).

2.4. Model development

2.4.1. QSAR models

The classification-based QSAR model was developed using the LDA technique (Mitteroecker and Bookstein, 2011) employing STATISTICA 7.0 (STATISTICA, STATSOFT Inc., USA). The regression-based QSAR model was built using genetic function approximation (GFA) followed by multiple linear regression (MLR) (Darlington, 1990; Fan et al., 2001) employing Genetic Algorithm tool 4.1 and MLR Plus Validation GUI 1.3 (DTC Lab Software Tools).

2.4.2. Docking protocol

A docking study was performed in order to identify the important interactions between the protein structure complex of TTR with bound 2′, 6′-difluorobiphenyl-4-carboxylic acid (PDB: 2F7I) and the ligand (i.e., PFASs). Prior to docking, it is important to prepare both protein and ligand for glide docking. The receptor was pre-processed using the protein preparation wizard tool available in multi step Schrodinger Suite (2012). The steps involved are as followed:

- Hydrogen was added to the receptor and minimized using OPLS 2005 force field.
- All the water molecules were removed except the active site water.
- The ligands were prepared using ligprep option available under Schrodinger software.
- Ligands were docked to the receptor using Glide.
- After ensuring that protein and ligands were in the correct form for docking, receptor grid files were generated using the grid-receptor generation program.
- A 10 Å size was used for the bounding and enclosing boxes.
- A grid box was generated at the centroid of the ligand bound to the active site of the receptor. Using extra precision (XP) mode of Glide tool, docking studies was performed.

Initially, the performance of the docking method was evaluated by re-docking the crystal ligand and correlating the docking scores with the protein. Thereafter, all the studied ligands were docked one by one. The docked co-crystal ligand showing key interacting residues similar to those present in the PDB was used to validate the docking methodology, the RMSD between them is 0.255 (Fig. 2).

2.5. Model validation

To assess the classifier model performance and classification capability, a number of statistical tests were employed. Such tests include computation of Wilks λ statistics, Canonical index (R_c), Matthews correlation coefficient (MCC), squared Mahalanobis distance and plotting of Receiver Operating Characteristic (ROC) curve, chi-square (χ^2) *etc*. Details are discussed elsewhere (Roy and Kar, 2015a).

In case of regression-based QSAR model, the goodness-of-fit of the equation was judged by the quality metric determination coefficient (R^2), as well as using the following internal validation metric Q^2_{LOO} and external validation metric $R^2_{pred.}$ The r^2_m metrics, namely $\overline{r^2_m}$ and Δr^2_m, developed by Roy et al. (2013) for internal, external and overall validation of models were also employed. The external predictivity of the model was further assessed in terms of error based metrics, such as the mean absolute error (MAE) based criteria (Roy et al., 2016). The models were also subjected to additional validation tests like $Q^2_{ext(F2)}$ (Schüürmann et al., 2008) and Golbraikh and Tropsha's (2002) criteria to check model reliability.

2.6. Y-randomization

The model randomization was performed 100 times via shuffling the dependent variables while maintaining the original independent variables. The average R^2 of 100 random models was computed and defined as R_r^2 followed by calculation of the ${}^cR_p^2$ parameter (Roy and Kar, 2015a; QSAR Model Development Using DTC Lab Software Tools) that penalizes model R^2 for small differences in the values of R^2 and R_r^2.

$$ {}^cR_p^2 = R \times \sqrt{R^2 - R_r^2} \quad (1) $$

For an acceptable model, the value of ${}^cR_p^2$ should be greater than 0.5.

2.7. Applicability domain study

According to OECD principle 3, the applicability domain (AD) of the QSAR model was checked using two different approaches: a) the Euclidean distance approach (Roy and Kar, 2015b) and b) the standardization based technique (Roy et al., 2015c).

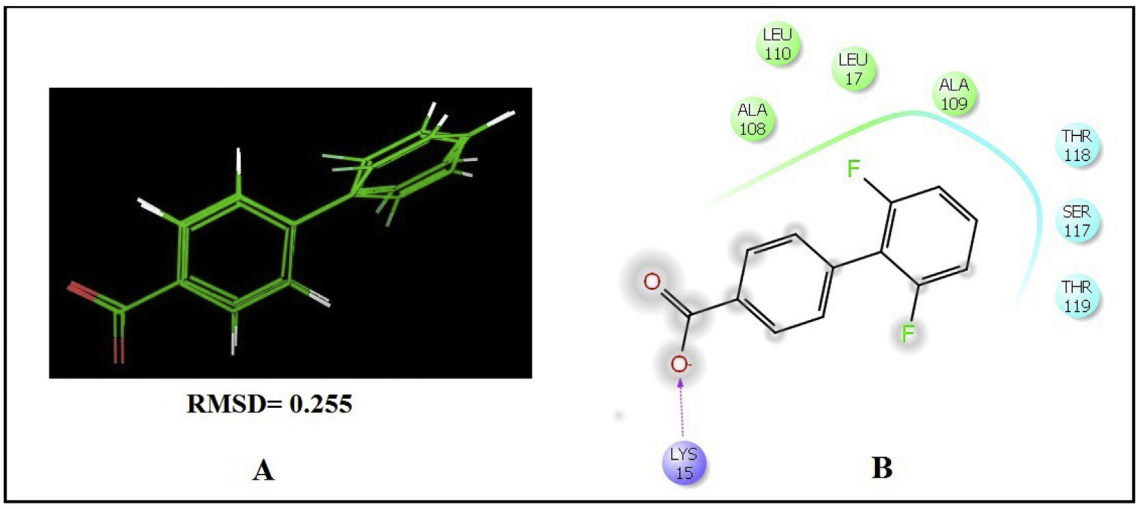


Fig. 2. (**A**) Superposition of the co-crystal on its docked position. (**B**) Amino acid pockets and binding interaction of bounded ligand in the 2F7I protein.

3. Results and discussion

3.1. Classification-based QSAR model

A two-group classification-based QSAR model was developed in this study. Considering the % binding threshold value, 15 PFASs were recognized as actives (Y) and 9 as inactives (N). Among the 16 training set molecules, 10 belonged to the Y class and 6 to the N group. Correspondingly, 5 and 3 compounds belonging to the Y and N groups respectively constituted the test set. LDA was executed by means of a stepwise method of variable selection with objective function F = 4 for inclusion, F = 3.9 for exclusion, and a tolerance value of 0.001, followed by discriminant analysis. The calculated and predicted classification categories of the discrimination and test sets respectively were determined at 50% probability level. The discriminant function ΔP is represented with the following equation:

$$\Delta P = 1.405 + 0.868 \times Me + 0.853 \times nCsp2 + 1.609 \times H - 050 \quad (2)$$

The LDA equation contains only 3 independent variables. The obtained metric based on the confusion matrix developed using the LDA technique satisfactorily suggested that the developed model is highly robust and predictive (Table 2). The predicted classification of studied PFASs is reported in Table 1.

The area under the ROC curve (AUROC) was also determined in order to check for the performance of the classification model for both the discrimination and test sets. The calculated values of AUROC for discrimination and test sets were 0.95 and 1.00 respectively, both values being much higher than the acceptable limit of 0.5. Thus, the AUROC also strongly supports the reliability of the developed discrimination model. The ROC curves for the discrimination and test sets are presented in Fig. S1 in Supplementary material section. Our model showed a value of 0.804 for ROCED, which corresponds to a good quality of the ROC analysis. ROCFIT was calculated by dividing the ROCED with Wilks λ value, and an acceptable value of 2.089 was obtained. The present study also showed acceptable values for the MCC: 0.87 for the training and 0.75 for the test sets, suggesting good classification-based model.

Analyzing the classification-based equation, we have found discriminating properties between Y and N classes. The identified important properties are mean atomic Sanderson electronegativity for scaled on carbon atom (Me) (Todeschini and Consonni, 2000), number of sp2 hybridized carbon atoms (nCsp2) (Todeschini and Consonni, 2000) and hydrophobicity of H atoms attached to heteroatom (H-050) (Ghose et al., 1998). Although the classification based equation may identify important discriminating properties, a contribution plot allows for the identification of the type and degree of contribution of each discriminating property towards toxicity. The contribution plot clearly suggested that hydrophobicity of the H atom attached to heteroatom is the most important discriminating feature between Y and N (Fig. 3a); a higher value of this property contributes towards the Y class of compounds. Two additional properties were also contributory towards the Y class of compounds, but their degree of contribution was much lower than H-050 descriptor. Compounds like perfluorohexanoic acid (ID: 2), 7H-perfluoroheptanoic acid (ID: 10), perfluorobutane sulfonate (ID: 12), perfluorohexane sulfonate (ID: 13), and perfluorooctane sulfonamide (ID: 21) all have contributions from the H-050 descriptor, therefore they are grouped into the active or Y class. On the contrary, molecules like 2-(*N*-methylperfluoro-1-octane sulfonamide) ethanol (ID: 19), *N*,*N*′-dimethyl perfluorooctane sulfonamide (ID:

Table 2
Qualitative prediction of classification-based QSAR.

Parameter/Metrics	Training set	Test set
Number of compounds	16	8
λ	0.256	–
F (df = 3, 12)	11.59145, p < 0.0007	–
Rc	0.862	–
Squared Mahalanobis Distance	10.818	–
$\chi^2(df = 3)$	17.005, p < 0.0007	–
Sensitivity	100.00%	100.00%
Specificity	83.33%	100.00%
Precision	90.91%	100.00%
Accuracy	93.75%	100.00%
F-measure	95.24%	100.00%
AUROC	0.95	1.00
G-means	0.91	0.82
Cohen's k	0.86	0.71
MCC	0.87	0.75
ROCED	0.78	
ROCFIT	3.04	

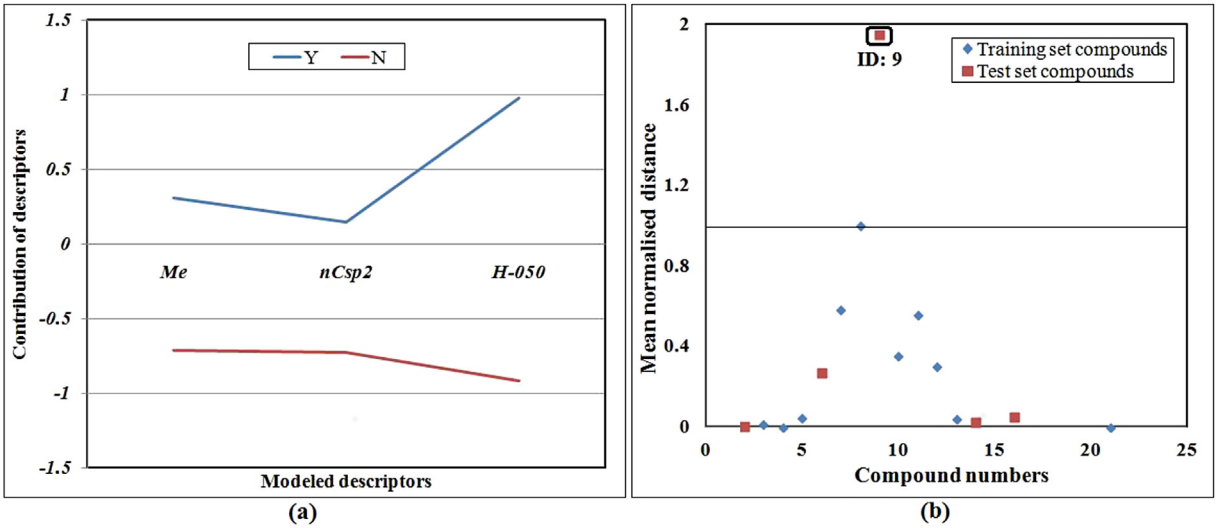


Fig. 3. (**a**) Contribution plot for the discriminating properties between toxic (Y) and non-toxic (N) PFASs based on classification-based QSAR model. (**b**) Euclidean distance plot of the developed regression-based QSAR model.

22), *N*-methyl perfluorooctane sulfonamide (ID: 23), and *N*-ethyl perfluorooctane sulfonamide (ID: 24) have no contribution from the mentioned descriptors resulting into their grouping as inactive or N class compounds. Four compounds (ID: 9, 15, 17 and 18) do not follow the above classification, and although they have contributions from the mentioned descriptors, they belong to the N class. Therefore, we performed further analysis through regression-based QSAR and docking study to get a better insight.

3.2. Regression-based QSAR model

A statistically significant regression based QSAR model was developed using GFA followed by MLR. The developed equation is as follows:

$$pIC_{50} = -87.472\ (\pm 16.854) + 4.307\ (\pm 0.659) \times IC3 + 114.379\ (\pm 23.119) \times \sum \beta'_S$$
$$n_{Training} = 10, R^2 = 0.86, R^2_a = 0.82, Q^2_{LOO} = 0.73, \overline{r^2_{m(LOO)Scaled}} = 0.65, \Delta r^2_{m(LOO)Scaled} = 0.06$$
$$n_{Test} = 5, Q^2_{F1} = R^2_{pred} = 0.64, Q^2_{F2} = 0.64, \overline{r^2_{m(test)Scaled}} = 0.56, \Delta r^2_{m(test)Scaled} = 0.17 \quad (3)$$
$$\overline{r^2_{m(overall)Scaled}} = 0.754,\ \Delta r^2_{m(overall)Scaled} = 0.033,\ MAE = 0.107,\ 'Good\ \ prediction'$$

Acceptable values of all statistical metrics for the model indicate that the obtained model has statistical reliability and good internal as well as external predictive potential. Least possible deviations of the predicted activity data from the corresponding observed ones is further implied from the satisfactory values of all the r^2_m metrics. The scatter plot (Fig. S2 in Supplementary material section) of experimentally determined (observed) versus predicted assay (employing equation (3)) value of PFASs confirmed goodness-of-fit and predictability of the model. Predicted pIC_{50} value (in mM scale) of studied PFASs is presented in Table 1. The QSAR model also satisfied the statistical validation criteria set forth by Golbraikh and Tropsha (2002). The results are as follow:

$$Q^2 = 0.73 > 0.5,\ \ Passed$$
$$r^2 = 0.87 > 0.6,\ \ Passed$$
$$\left|r^2_0 - r'^2_0\right| = 0.17 < 0.3,\ \ Passed$$
$$\frac{r^2 - r^2_0}{r^2} = 0.25 < 0.1,\ \ or,\ \ \frac{r^2 - r'^2_0}{r^2} = 0.06 < 0.1,\ \ Passed$$
$$0.85 \le k = 1.02 \le 1.15\ \ or,\ \ 0.85 \le k' = 0.96 \le 1.15,\ \ Passed$$

To further check the quality and predictability of the models, we employed another stringent test which measures the error based judgment of test set predictions. The obtained MAE value was 0.107 with a standard deviation (SD) of the absolute errors of 0.078. These MAE-based criteria results suggest that the model is able to produce 'GOOD' predictions.

The contribution of the descriptors to the inhibitory activity was in the following order:

$$IC3 > \sum \beta'_S$$

In equation (3), IC3 is denoted as the information content index (neighborhood symmetry of 3-order) (Todeschini and Consonni, 2000) which has a positive contribution towards inhibitory activity. Again $\sum \beta'_S$, an ETA index (Roy and Das, 2011), also has a positive contribution towards the inhibitory activity. $\sum \beta'_S$ signifies sigma average VEM count (the mobile valence electronic environment) described with the following equation:

$$\sum \beta'_S = \sum \beta_S \Big/ N_V \quad (4)$$

where, $\sum \beta_S$ is sum of β_s values of all non-hydrogen vertices of a molecule and Nv is vertex count (excluding hydrogen).

Before analyzing the descriptors, one has to remember that all 15 PFASs considered for the regression-based model are toxic or active ones. So, the analysis was performed among high, moderate and low toxic compounds. A balanced effect of both descriptors is crucial for the prediction of toxic effect. Compounds 13, 3 and 4 possess higher values for both descriptors, resulting in the top three toxic compounds among the modeled compounds. The lowest values of IC3 and moderate values of $\sum \beta'_S$ make molecules 12, 7 and 8 the least toxic. Moderate values for both descriptors in molecules 5, 10 and 21 place them into the group of moderate toxicity. If we scrutinize all these compounds minutely, PFASs belonging to the high or moderate toxic classes have a carbon chain length between 6 and 10. On the other hand, PFASs with a C chain length below 6 or above 10, fall in the group of low toxicity. Therefore, there is a strong connection between carbon chain length as well as functional groups present in PFASs with these two descriptors.

A Y-randomization (Fig. S3 in Supplementary material section) study was performed to ensure that the model was not the outcome of mere chance alone. The values of average R^2 and Q^2_{LOO} for 100 random models were 0.24 and −0.69, respectively. The lack of chance correlation in the QSAR model is also well reflected from the value of $^cR^2_p$ (0.76) which is higher than the acceptable threshold value of 0.5.

Employing the standardization based AD determination technique, no test compounds were found outside the AD. Additionally, in the case of the Euclidean distance AD validation approach (i.e., employing a Euclidean distance measure), one compound from the test set (ID: 9, perfluorotetradecanoic acid) was found to lie outside the domain defined by the training set (Fig. 3b). Therefore, the combined results obtained from the AD study with the Euclidean distance approach and the standardization based technique indicate that out of 5, 4 test compounds were inside the AD and their predictions are completely reliable. Therefore, we can confidently predict the toxicity of 80% of the test compounds.

3.3. Docking study

The bound ligand present in the crystal structure was extracted from the receptor and docked first to validate the docking method using the glide module of Schrodinger Suite (2012) by adopting the XP extra precision method as described in the materials and method section. Thereafter, all 24 PFASs were docked maintaining the same docking algorithm. Docking interaction diagrams of PFASs with human-TTR showed the crucial amino acid residues responsible for toxicity response. Interestingly, of the 24 PFASs, 7 did not dock (interestingly, 3 molecules were inactive and 3 had very low activity, Table 1). Meticulous analyses of the ligand-receptor interactions were carried out, and final coordinates of the ligand and receptor determined. The binding energy of all compounds were calculated and presented in Table 1.

The majority of the toxic compounds studied showed high binding energy. Considering the experimental toxic potency and obtained high-binding energy from the docking study, interaction diagrams for four of the PFASs studied are presented in Fig. 4. Compound 13 (perfluorohexane sulfonate), the most active compound considering its pIC_{50} and T4-TTR % binding potency values, showed the highest binding energy (−4.37 kcal*mol-1) by fitting in the active protein sites via hydrophobic interactions with Leu17, Ala108, Ala109, Leu110 and Val121; polar interactions with Thr118 and Thr119; and hydrogen bonding with positively charged Lys15 (Fig. 4). The other three compounds fitted exactly in the identical active sites making interactions with Leu17, Ala108, Ala109, Leu110, Thr118, Thr119 and Lys15, but not with Val121. Captivatingly, in all four compounds, a hydrogen bond formed with the positively charged Lys15 and the negatively charged oxygen atom of sulfonate or acid functional groups. Among the mentioned four compounds, perfluorobutane sulfonate (ID: 12) showed high binding energy though its pIC_{50} and T4-TTR % binding potency thresholds were moderate because of its small C chain length (C = 4) which is responsible for a much higher diverse structural orientation inside the protein compared to other PFASs.

In Fig. 5, the binding interactions of four inactive compounds based on T4-TTR % binding potency are presented for further analyses. Interestingly, their experimental assay values were not detected in measured scales making them non-toxic/lower toxic among the studied PFASs. The lower values of binding energy among the docked compounds also strongly support the claim of experimentalists about their non/lower toxic profile (Weiss et al., 2009). For these PFASs, no interaction was identified with Lys15. Interestingly, molecules 17 and 18 are fluorotelomer alcohols, making H-bonds with Ser117 and Ala109, respectively. Both molecules are surrounded with almost identical amino acids like Lys15, Leu17, Ala108, Ala 109, Leu110, Ser117, Thr119 and Val 121. Hydrogen bonding with different amino acids occurred due to 2 carbon chain length between 2-perfluorooctyl ethanol (ID: 18) and 2-perfluorohexyl ethanol (ID: 17) leading to different structural orientation in the protein pocket. Out of the remaining two PFASs, 2-(*N*-methylperfluoro-1-octane sulfonamide) ethanol (ID: 19) is also a substituted PFAS with an alcohol functional group making hydrogen bonding with Ser117 and surrounded with similar amino acid residues. Although *N*-methyl perfluorooctane sulfonamide (ID: 23) did not form any hydrogen bonding with the protein, identical amino acid residues were observed compared with the other three PFASs resulting in the lowest observed binding energy among all the studied molecules.

According to the docking studies and QSAR results, it is clear that the choice of functional groups is very significant in order to avoid interaction with Lys15 amino acid and that carbon chain length plays a crucial role in the observed toxicity. Therefore, for better understanding of outcome from all methods, a combined interpretation is provided in the Overview section. It is important to mention that a higher binding energy suggests strong interaction of the ligand with the TTR protein active pocket and *vice versa*. In the case of the docking study, only interactions between ligand and TTR were computed. Therefore, it is important to mention that in a competitive inhibition study, a ligand must be strong enough to replace T4 from the TTR binding site.

The present study can only partially address this issue. Therefore for future studies, we recommend conducting ligand binding assays in which we can quantify the competitive binding potency of ligands by replacing T4 from the human TTR protein. These experimental assays will be supported with multiple docking studies in our future work.

4. Overview and derived chemical rules

Satisfyingly, the results obtained from the regression based and classification based QSAR models and docking studies were quite complementary to each other with respect to the requisite features and structural fragments that enable to discriminate among active/

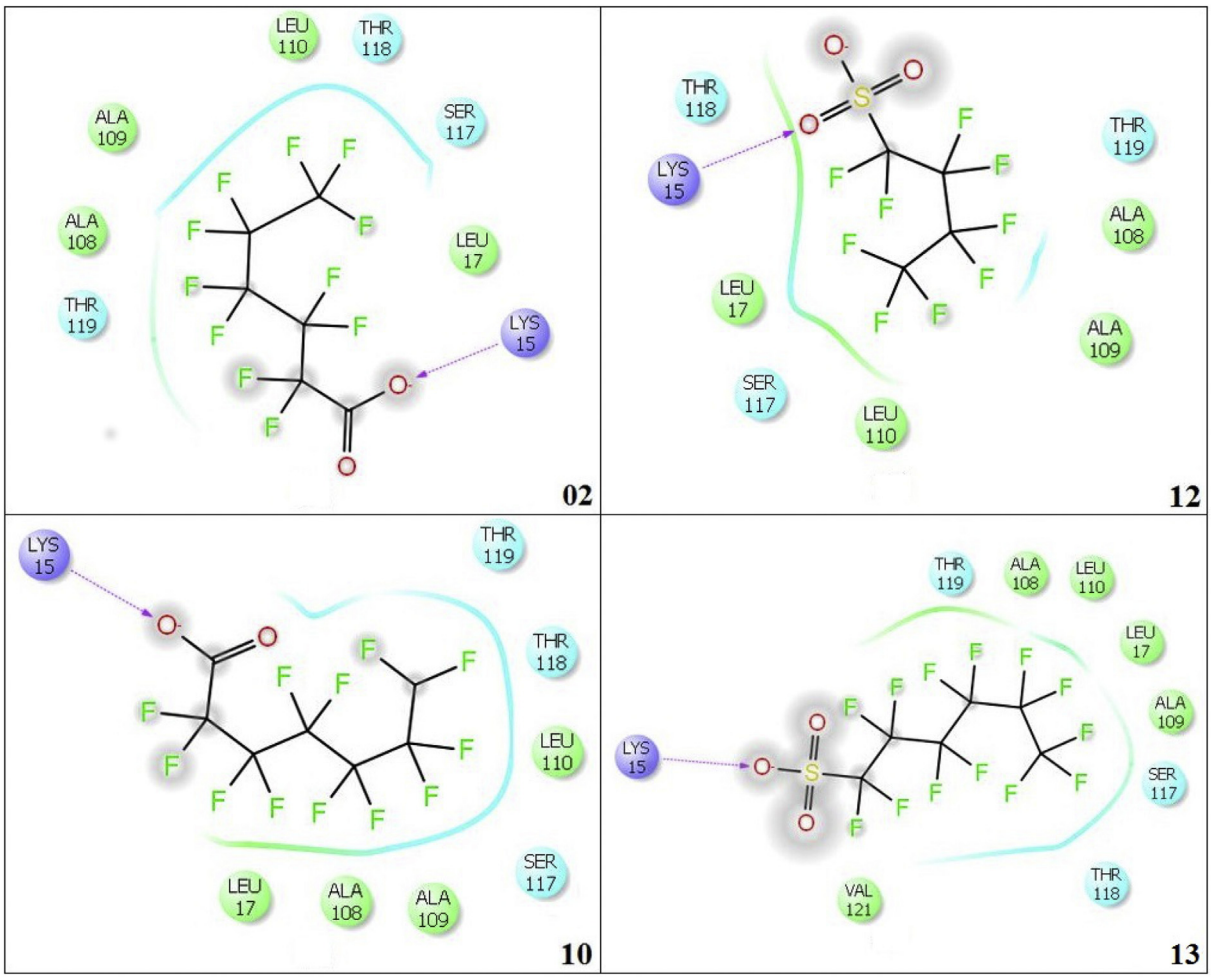


Fig. 4. Interactions observed for the active/toxic PFASs with 2F7I protein from docking analyses.

toxic and inactive/non-toxic T4-TTR binding potency of the studied compounds. Martin et al. (2003) already showed experimentally that C chain length of PFASs has a large impact in toxicity. Interestingly, our modeling results support this observation, as well as show the prominent role of type of functional group on potential endocrine disruption. In summary, the final interpretations are discussed based on the modeled descriptors, amino acid interactions with PFASs and their structural pattern for the studied dataset as follows:

a) All of the studied PFASs containing acid functional groups are toxic, with the exception of perfluorobutyric acid (ID: 1). Interestingly, carbon chain length has a major role to play in determining the toxicity potency. If the carbon chain length is between 6 and 10, PFASs will be highly toxic or active ones. Toxicity will be lower or non-existent for PFASs containing a C chain length greater than 10 or below 6, respectively.
b) PFASs containing sulfonate or sulfinate functional groups are active or toxic. Compounds will be inactive or show little toxicity if the carbon chain length is over 8 atoms and containing any of the abovementioned functional groups (Example: perfluorodecane sulfonate is inactive or non-toxic). So, it is suggested that sulfonate or sulfinate groups should be avoided for PFASs containing less than 8 C.
c) Fluorotelomer alcohols are inactive, irrespective of their carbon chain length.
d) Interestingly, PFASs containing sulfonamide functional groups are toxic when C chain length = 8. But, substitution of sulfonamide groups with alkyl or alcohol group leads to the production of inactive compounds irrespective of their carbon chain length. Therefore, for C8 PFAS sulfonamide or sulfonamido functional groups should be replaced with either alkyl or alcohol functional groups.

5. Conclusion

We have presented classification- and regression-based QSAR models and a docking study for the prediction of toxicity and identification of the required structural features and fragments of a diverse class of PFASs based on their binding affinity to human TTR. These models provide an understanding of important structural attributes of this group of contaminants and can also provide important information for the design of safer PFASs for synthesis of future molecules with diminished systemic toxicity. Moreover, the developed models may serve as an efficient query tool for screening of large databases and identifying PFASs molecules bearing the detrimental structural features associated with their toxicity potential. It is conceivable that our *in silico* approach offers an efficient

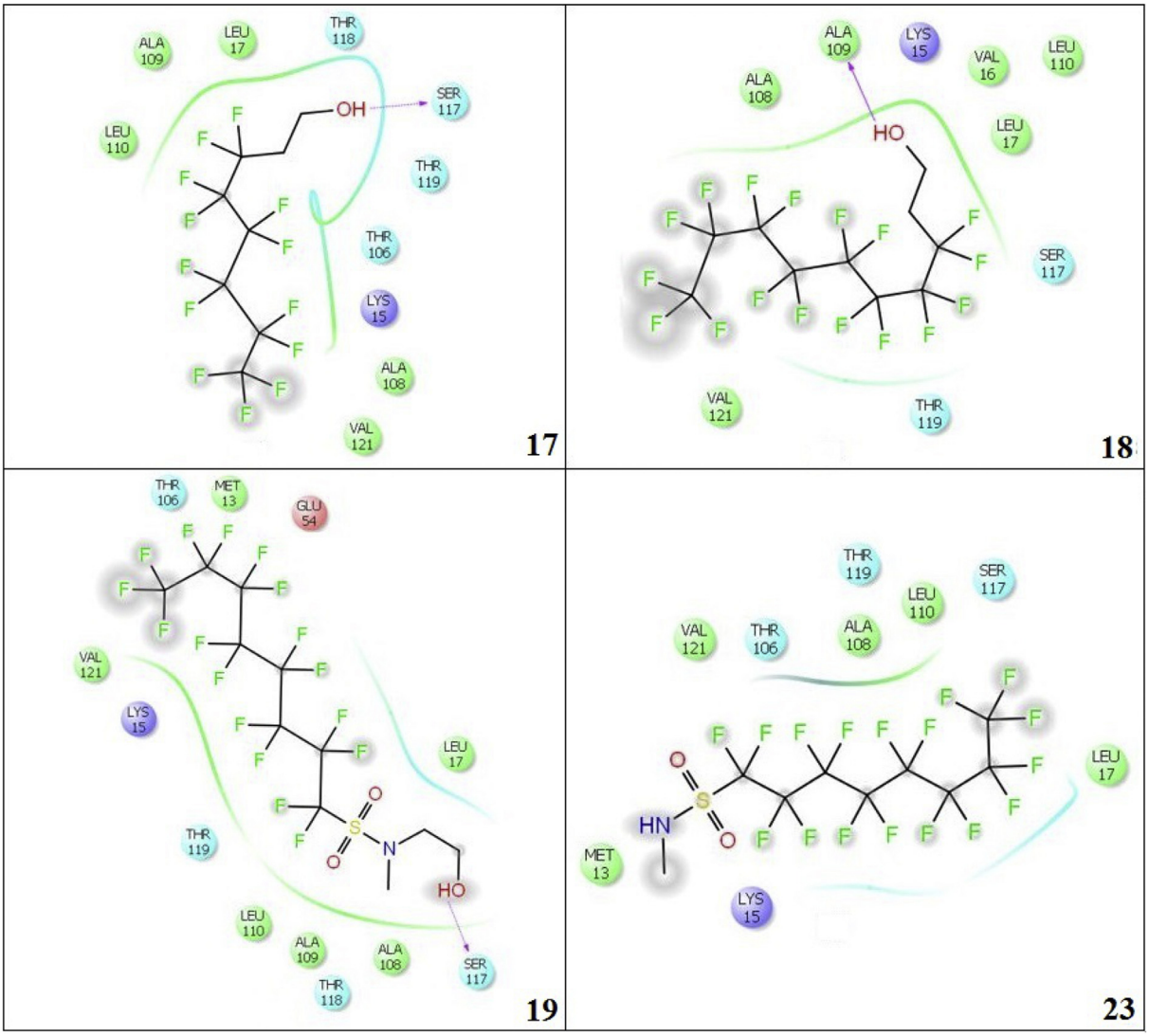


Fig. 5. Interactions observed for the inactive/non-toxic PFASs with 2F7I protein from the docking.

screening method for toxicity and can be used in integration with existing as well as future *in vivo*/*in vitro* methods to increase overall predictivity.

Acknowledgements

S.K and J.L thank the National Science Foundation (NSF/CREST HRD-1547754, and EPSCoR (Award#: 362492-190200-01/NSFEPS-0903787) for financial support. K.R. thanks the UGC, New Delhi for funding the MAESTRO software under the UPEII programme. The authors want to acknowledge the Mississippi Center for Supercomputing Research (MCSR) for providing state-of-the-art high performance computing facilities and outstanding services for supporting this research.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemosphere.2017.06.024.

References

Benigni, R., Giuliani, A., 2003. Putting the predictive toxicology challenge into perspective: reflections on the results. Bioinformatics 19, 1194–1200. http://dx.doi.org/10.1093/bioinformatics/btg099.

Benigni, R., Zito, R., 2004. The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. Mutat. Res. 566, 49–63. http://dx.doi.org/10.1016/S1383-5742(03)00051-6.

Darlington, R.B., 1990. Regression and Linear Models. McGrawHill, New York.

DRAGON Version 6.0, 2011, http://www.talete.mi.it/.

Dreyer, A., Weinberg, I., Temme, C., Ebinghaus, R., 2009. Perfluorinated compounds in the atmosphere of the atlantic and southern oceans: evidence of a global distribution. Environ. Sci. Technol. 43, 6507–6514. http://dx.doi.org/10.1021/es9010465.

DTC Lab Software Tools: http://teqip.jdvu.ac.in/QSAR_Tools/.

Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y., Weinstein, J.N., 2001. Quantitative structure–antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. J. Med. Chem. 44, 3254–3263. http://dx.doi.org/10.1021/jm0005151.

Fromme, H., Tittlemier, S.A., Völkel, W., Wilhelm, M., Twardella, D., 2009. Perfluorinated compounds – exposure assessment for the general population in western countries. Int. J. Hyg. Environ. Health 212, 239–270. http://dx.doi.org/10.1016/j.ijheh.2008.04.007.

Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., 1998. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental Methods: an analysis of ALOGP and CLOGP methods. J. Phys. Chem. A 102, 3762–3772. http://dx.doi.org/10.1021/jp980230o.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Model 20, 269–276. http://dx.doi.org/10.1016/S1093-3263(01)00123-1.
HyperChem(TM), Hypercube, Inc, 1115 NW 4th Street, Gainesville, Florida 32601, USA.
Kannan, K., Tao, L., Sinclair, E., Pastva, S.D., Jude, D.J., Giesy, J.P., 2005. Perfluorinated compounds in aquatic organisms at various trophic levels in a Great Lakes food chain. Arch. Environ. Toxicol. 48, 559–566. http://dx.doi.org/10.1007/s00244-004-0133-x.
Kato, K., Ye, X., Calafat, A.M., 2015. PFASs in the General Population. in: Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances, Molecular and Integrative Toxicology. Springer International Publishing Switzerland, pp. 51–76.
Kovarich, S., Papa, E., Li, J., Gramatica, P., 2012. QSAR classification models for the screening of the endocrine-disrupting activity of perfluorinated compounds. Sar. QSAR Environ. Res. 23, 207–220. http://dx.doi.org/10.1080/1062936X.2012.657235.
Lau, C., Amitole, k., Hodes, C., Lai, D., Pfahles-Hutchens, A., Seed, J., 2007. Perfluoroalkyl acids: a review of monitoring and toxicological findings. Toxicol. Sci. 99, 366–394. http://dx.doi.org/10.1093/toxsci/kfm128.
Lopez-Espinosa, M.J., Mondal, D., Armstrong, B., Bloom, M.S., Fletcher, T., 2012. Thyroid function and perfluoroalkyl acids in children living near a chemical plant. Environ. Health Perspect. 120, 1036–1041. http://dx.doi.org/10.1289/ehp.1104370.
Martin, J.W., Mabury, S.A., Solomon, K.R., Muir, D.C.G., 2003. Bioconcentration and tissue distribution of perfluorinated acids in rainbow trout (*Oncorhynchus mykiss*). Environ. Toxicol. Chem. 22, 196–204. http://dx.doi.org/10.1002/etc.5620220126.
Mitteroecker, P., Bookstein, F., 2011. Linear discrimination, ordination, and the isualization of selection gradients in modern morphometrics. Evol. Biol. 38, 100–114. http://dx.doi.org/10.1007/s11692-011-9109-8.
Olsen, G.W., 2015. PFAS Biomonitoring in Higher Exposed Populations. in: Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances, Molecular and Integrative Toxicology. Springer International Publishing Switzerland, pp. 77–125.
Ren, X.-M., Zhang, Y.-F., Guo, L.-H., Qin, Z.-F., Lv, Q.-Y., Zhang, L.-Y., 2015. Structure-activity relations in binding of perfluroalkyl compounds to human thyroid hormone T3 receptor. Arch. Toxicol. 89, 233–242. http://dx.doi.org/10.1007/s00204-014-1258-y.
Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., 2013. Some case studies on application of "r^2_m" metrics for judging quality of QSAR predictions: emphasis on scaling of response data. J. Comput. Chem. 34, 1071–1082. http://dx.doi.org/10.1002/jcc.23231.
Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom. Intell. Lab. Sys 152, 18–33. http://dx.doi.org/10.1016/j.chemolab.2016.01.008.
Roy, K., Kar, S., 2015a. How to judge predictive quality of classification and regression based QSAR models? In: Haq, Z.U., Madura, J. (Eds.), Frontiers of Computational Chemistry. Bentham, pp. 71–120. http://ebooks.benthamscience.com/book/9781608059782/chapter/128894/.
Roy, K., Kar, S., 2015b. Importance of applicability domain of QSAR models. In: Roy, K. (Ed.), Quantitative Structure-activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment. IGI Global, pp. 180–211. http://dx.doi.org/10.4018/978-1-4666-8136-1.ch005.
Roy, K., Kar, S., Ambure, P., 2015c. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145, 22–29. http://dx.doi.org/10.1016/j.chemolab.2015.04.013.
Roy, K., Kar, S., Das, R.N., 2015a. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press. http://books.google.co.uk/books/reader?id=bkFOBQAAQBAJ&printsec=frontcover&output=reader&source=gbs_atb&pg=GBS.PA23.
Roy, K., Kar, S., Das, R.N., 2015b. A Primer on QSAR/QSPR Modeling: Fundamental Concepts (SpringerBriefs in Molecular Science). Springer. http://www.springer.com/gp/book/9783319172804.
Roy, K., Das, R.N., 2011. On extended topochemical atom (ETA) indices for QSPR studies. In: Castro, E.A., Hagi, A.K. (Eds.), Advanced Methods and Applications in Chemoinformatics: Research Progress and New Applications. IGI Global, PA, pp. 380–411. http://www.igi-global.com/bookstore/chapter.aspx?titleid=56464.
Schrodinger Suite (2012): http://www.schrodinger.com/glide.
Schüürmann, G., Ebert, R.U., Chen, J., Wang, B., Kühne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient-test set activity mean vs. training set activity mean. J. Chem. Inf. Model 48, 2140–2145. http://dx.doi.org/10.1021/ci800253u.
STATISTICA is a Statistical Software of STATSOFT Inc., USA. Available from <http://www.statsoft.com/>.
Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors, vol. 11. Wiley-VCH, Weinheim, pp. 1–668.
US EPA, 2009. Integrated Risk Information System. US Environmental Protection Agency. National Center for Environmental Assessment, Washington, DC. Available from: http://www.epa.gov/iris/ (Accessed 01 January 2011).
Weiss, J.M., Andersson, P.L., Lamoree, M.H., Leonards, P.E.G., van Leeuwen, S.P.J., Hamers, T., 2009. Competitive binding of poly- and perfluorinated compounds to the thyroid hormone transport protein transthyretin. Toxicol. Sci. 109, 206–216. http://dx.doi.org/10.1093/toxsci/kfp055.
Williams, E.S., Panko, J., Paustenbach, D.J., 2009. The European Union's REACH regulation: a review of its history and requirements. Crit. Rev. Toxicol. 39, 553–675. http://dx.doi.org/10.1080/10408440903036056.