

Uncertainty-Informed Deep Transfer Learning of PFAS Toxicity

Jeremy Feinstein¹, Ganesh Sivaraman², Kurt Picel¹, Brian Peters¹, Álvaro Vázquez-Mayagoitia³, Arvind Ramanathan², Margaret MacDonell¹, Ian Foster², Eugene Yan^{1*}

1. Environmental Science Division, Argonne National Laboratory, Lemont, IL 60439, USA
2. Data Science and Learning Division, Argonne National Laboratory, Lemont, IL 60439, USA
3. Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA

* E-mail: eyan@anl.gov

Abstract

Perfluoroalkyl and polyfluoroalkyl substances (PFAS) pose a significant hazard because of their widespread industrial uses, environmental persistence, and bioaccumulativity. A growing, increasingly diverse inventory of PFAS, including 8,163 chemicals, has recently been updated by the U.S. Environmental Protection Agency. But, with the exception of a handful of well-studied examples, little is known about their human toxicity potential because of the substantial resources required for *in vivo* toxicity experiments. We tackle the problem of expensive *in vivo* experiments by evaluating multiple machine learning (ML) methods including random forests,

deep neural networks (DNN), graph convolutional networks, and Gaussian processes, for predicting acute toxicity (e.g., median lethal dose, or LD₅₀) of PFAS compounds. To address the scarcity of toxicity information for PFAS, publicly available datasets of oral rat LD₅₀ for all organic compounds are aggregated and used to develop state-of-the-art ML source models for transfer learning. 518 fluorinated compounds containing 2 or more C-F bonds with known toxicity are used for knowledge transfer to ensembles of the best-performing source model, DNN, to generate the target models for the PFAS domain with access to uncertainty. This study predicts toxicity for PFAS with a defined chemical structure. To further inform prediction confidence, the transfer-learned model is embedded within a SelectiveNet architecture, where the model is allowed to identify regions of prediction with greater confidence and abstain from those with high uncertainty using a calibrated cutoff rate.

1 Introduction

Perfluoroalkyl and polyfluoroalkyl substances (PFAS) encompass thousands of synthetic fluorinated aliphatic compounds^{1,2}. PFAS pose a significant challenge of increasing concern because of their widespread presence, long-term persistence, extended biological half-lives (approaching nine years for some), and largely unknown toxicities. PFAS use has been identified at more than 400 U.S. military bases, and contamination has been found in the drinking-water systems of more than two dozen military sites. U.S. cleanup costs are estimated to be tens of billions of dollars, including \$2 billion for the Department of Defense alone³. The U.S. Environmental Protection Agency (EPA)'s Distributed Structure-Searchable Toxicity (DSSTox) database of PFAS structures⁴, as recently updated, contains over 8,163 PFAS chemicals. PFAS compounds can be broadly classified into polymeric and non-polymeric families². This study addresses non-polymeric PFAS,

which have a higher propensity to be absorbed via the digestive system, creating an urgent need to understand their toxicities. Their toxicities will be important determinants of target cleanup levels and associated costs as well as identification of non-toxic substitutes for future consumer products.

Traditional approaches for generating toxicity information (e.g., human epidemiological and experimental animal studies) are resource-intensive, and only limited studies have been conducted across this large set of compounds^{5,6}. The exponential growth of chemical synthesis in recent decades necessitates scalable approaches for determination of PFAS toxicities. To reduce the expense and uncertainties inherent in animal experiments, it is crucial to perform high-throughput computational toxicity predictions. Here we explore a cheminformatics approach to predicting and understanding toxicity from chemical structure.

The acceleration of computational toxicology in recent years can be attributed to 1) the development of large databases of chemical toxicities, 2) increased computing power with the advent of hardware such as Graphic Processing Units and other accelerators, and 3) advancement in machine learning (ML) that can take advantage of increased data and computational power⁷⁻¹¹. In particular, deep learning for prediction of chemical properties is becoming increasingly relevant^{12,13}. Several studies have demonstrated that deep-learning models for chemical properties and toxicity prediction can outperform traditional Quantitative Structure-Activity Relationship (QSAR) approaches such as naive Bayes, support vector machines, and random forests (RFs)¹⁴⁻¹⁸. However, a compound's toxicity is affected by multiple chemical and biological factors, adding complexity to the prediction of this crucial property¹⁹.

Acute toxicity refers to a chemical's propensity to cause adverse health effects within a short period

following exposure of a living organism. This broad definition means that there are many considerations when characterizing acute toxicity. A common nonspecific method for gauging the relative toxicity of a set of compounds without any considerations of biological pathways involved is to compare median lethal doses (LD_{50}), the minimum dose of a substance shown to cause fatality in 50% of laboratory subjects within 24 hours after initial oral or dermal exposure. Oral rat LD_{50} metrics are measured in test-substance quantity per unit mass of laboratory-rat body weight and are ranked by the EPA into four categories: I (high toxicity), II (moderate toxicity), III (low toxicity), and IV (very low toxicity). Acute oral toxicities and their respective EPA categories (defined in Table 1) provide a systematic method for classifying toxicity. However, there are only tens of PFAS compounds with reported values of oral rat LD_{50} point estimates.

Category	Toxicity	Dosage (mg/kg body weight)
I	High	≤ 50
II	Moderate	> 50 to 500
III	Low	> 500 to 5,000
IV	Very low	$> 5,000$

Table 1: EPA toxicity classes

In the face of this data scarcity, we propose an uncertainty-informed transfer-learning approach for predicting and understanding PFAS toxicities. In the context of transfer learning, we refer to oral rat LD_{50} as a property label. The transfer learning has two components: 1) Source task training, for which there is an abundance of labeled data, and 2) a target task, where there is a very small pool of labeled data, a large pool of unlabeled data (i.e., PFAS compounds are known), and high expenses

limiting access to new labels. Transfer learning enables knowledge gained from source task training to be leveraged in a related target task where sufficient labeled samples are not available for independent training²⁰. We aggregate reported values of oral rat LD₅₀ point estimates from various public data sources to create a new database that we refer to as “LDToxDB.” As a source task, we use LDToxDB to establish baselines for ML toxicology prediction. We provide a discussion of relevant literature baselines on oral rat LD₅₀ predictions and show that our source ML baselines are competitive. Then we identify 518 fluorinated compounds containing 2 or more C-F bonds within LDToxDB (which we will refer to as PFAS-like) with known LD₅₀ labels. 518 PFAS-like compounds are used with knowledge transferred from the best-performing source task, to generate the target models with access to uncertainty. The rationale is that 518 PFAS-like compounds are the closest chemical family in our database to the broader 8,163 PFAS compounds, and hence it will be important to understand how a transfer-learned target model performs on PFAS-like compounds where oral rat LD₅₀ labels are available, before attempting predictions for PFAS compounds with unknown oral rat LD₅₀. For this purpose, we review the uncertainty analysis derived from transfer-learned target models to gain insights into the quality of predictions for the PFAS-like compounds.

Finally, we temper toxicity predictions by implementing selective prediction through an abstention mechanism that forces our transfer learned target model to say “I cannot answer” when confidence in a prediction is low²¹⁻²³. When making predictions for compounds with unknown toxicity, it is extremely important to enforce an abstention mechanism as a precautionary measure against incorrectly classifying a highly toxic substance. We then apply the transfer-learned selective model to predict toxicity (or abstain) for 8,163 EPA PFAS compounds with no known oral rat LD₅₀; the details of these predictions are discussed in the results section. The added capability of a transfer-learning model to abstain from prediction for some compounds opens up the possibility of creating

a direct feedback loop to *in vivo* experiments, the details of which are further discussed in the conclusion section. We refer to the entire suite of computational toxicology tools developed as part of this study as “AI4PFAS” (Figure 1); details are discussed in the methods and results sections.

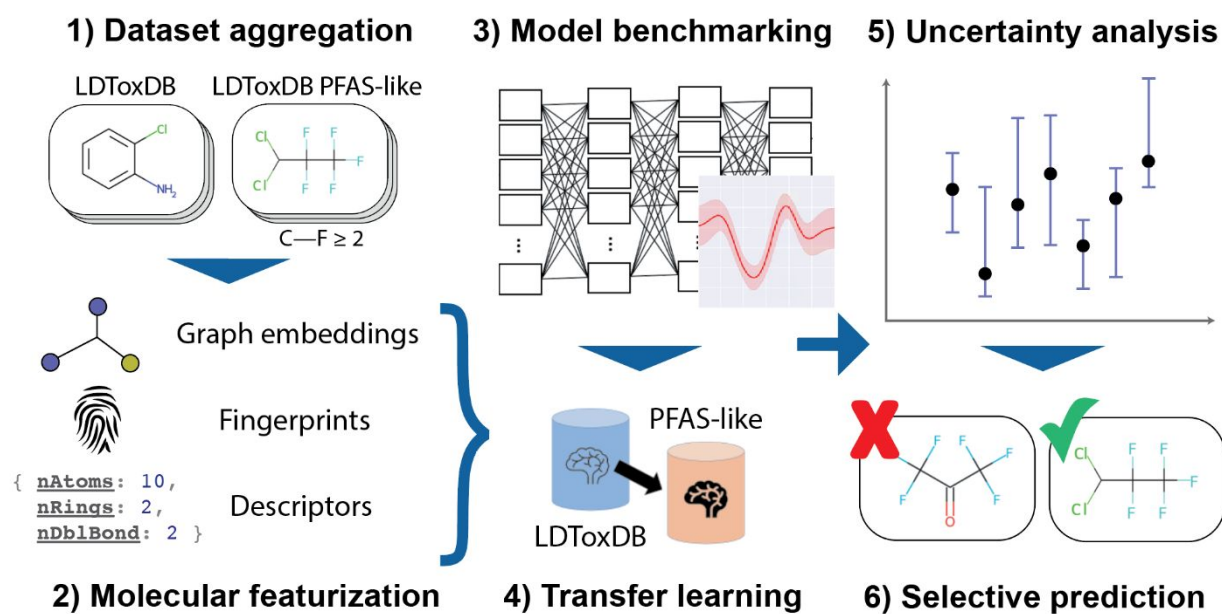


Figure 1. AI4PFAS workflow for PFAS toxicity prediction

2 Methods

Datasets

The availability of *in vivo* acute oral toxicity measurements for PFAS is limited to a handful of well-studied compounds in this family. To abate the lack of PFAS toxicity data, we constructed an expanded dataset, LDToxDB, of 13,329 unique compounds of any type with oral rat LD₅₀ measurements aggregated from the EPA Toxicity Estimation Software Tool (TEST), NIH Collaborative Acute Toxicity Modeling Suite (CATMoS), and National Toxicology Program datasets²⁴⁻²⁶. LD₅₀ point estimates provided in mg/kg were converted to units of

–log(mol/kg) to reflect per-molecule toxicity irrespective of molecular mass; the resulting histogram is shown in Figure 2. Most of these compounds were labeled as EPA toxicity class III, followed by a near-equal presence in II and IV, and lastly class I. SMILES were canonicalized using RDKit²⁷ and duplicate molecules were removed by querying each compound's hashed InChIKey.

To broadly identify PFAS-like compounds within LDToxDB, molecules with two or more C – F bonds were identified by using an RDKit SMARTS query and tagged as a PFAS-like representative subset of the labeled LDToxDB compounds. Such compounds with 2 or more C-F bonds would be polyfluorinated, likely alkyl, but may not be designated as PFAS in various databases. The resulting 518 compounds, referred to as “LDToxDB-PFAS-like” and which include 58 compounds formally labeled as PFAS, served as an important validation group to confirm that models trained on LDToxDB were able to predict toxicity of PFAS and PFAS-like compounds via chemical-structure similarity. Finally, 8,163 PFAS compounds were extracted from the EPA DSSTox database⁴, most of which have no LD₅₀^[00]; referred to as “LDToxDB-PFAS”; and ^[00]reserved for prediction. We note that 58 of these are also represented, with labels, in LDToxDB-PFAS-like. The dataset (composed of LDToxDB, LDToxDB-PFAS-like, and LDToxDB-PFAS) and the Python processing codes used to parse the data and construct the models are available at <https://github.com/AI4PFAS/AI4PFAS>.

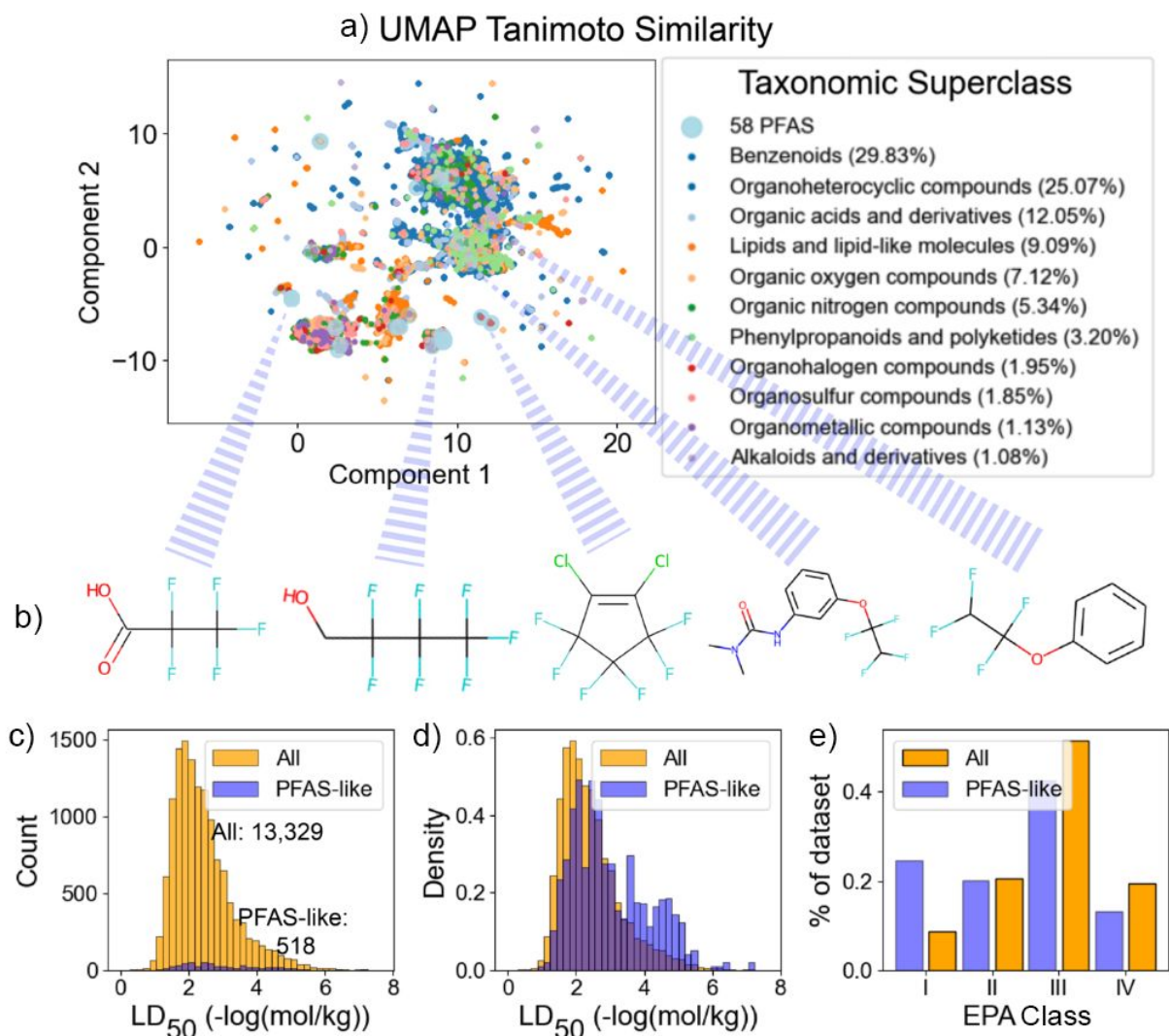


Figure 2: Visualization of datasets. a) Visual exploration of LDToxDB with dimensionality reduction performed with Uniform Manifold Approximation and Projection (UMAP) on a Tanimoto similarity matrix²⁸. Clusters are colored according to the chemical superclasses annotated by the ClassyFire server²⁹. 58 compounds identified and colored as PFAS are found by cross-referencing compounds in LDToxDB with the EPA DSSTox. b) Chemical structures for five of the 58 PFAS compounds picked from distinct clusters of the chemicals space map are shown. c) Histograms showing the LD₅₀ distribution for LDToxDB (labeled as "All"). The PFAS-like subset distribution is shown for reference. d) Normalized histograms showing the LD₅₀ distribution for LDToxDB. The PFAS-like subset distribution is shown for comparison. e) Bar plot showing percentage fraction per EPA toxicity class for LDToxDB and PFAS-like.

Chemical Featurization

Chemical featurization is the process of translating chemical attributes associated with a compound into machine-readable numeric features. We computed features for all compounds in LDToxDB for

use in supervised ML of acute oral LD₅₀ point estimates. Further, unsupervised ML can be applied to the chemical features in order to deduce chemical insights. Figure 1 shows the full set of featurizations and tools developed as a part of the AI4PFAS workflow; details are discussed below.

Chemical featurization relies primarily on encoding structural features and atom identities within a molecule. Three types of chemical featurization were considered in this study:

1) **Mordred descriptors**³⁰: We used the Mordred software package³⁰ to generate 1,800 unique molecular descriptors for each compound directly from RDKit molecules. Mordred provides quick featurization of a molecular dataset by generating a vast array of two- and three-dimensional descriptor characteristics from SMILES input. The full reference list of Mordred descriptors is available elsewhere³⁰. We trimmed down the 1,800 descriptors to 300 by using Pearson correlation coefficient (PCC) analysis to remove redundant features.

2) **Extended-connectivity fingerprinting (ECFP)**³¹ provides a mechanism for representing topological chemical space within a fixed-length bit string by iteratively measuring substructure connectivity at a provided radius around each atom. Numeric representations are created for each substructure identified in these iterations and then combined into a fixed-length bit string. Conventionally, an ECFP is described by its bit length and the maximum radius used for substructural querying: thus, for example, a 2048-bit ECFP4 has a length of 2048 bits and a maximum radius of four. Multiple bit lengths and radii were used for different purposes in this study. ECFPs are generated by using the open-source RDKit package for Python²⁷.

3) **Molecular graph encoding**^{32,33} improves on ECFP by representing molecules as graphs of arbitrary size with nodes representing atoms, and edges representing bonds. Each entity is given characteristic traits, which for nodes may include (but are not limited to) atomic identity, number of valence electrons, formal charge, and hybridization, and for edges, bond order and conjugation

status. We have adopted the graph representation and the corresponding graph convolutional neural network from the MOLAN workflow¹⁷.

4) **Non-negative matrix factorization (NMF)**^{34,35} is a dimensionality reduction technique that derives basis vectors under a non-negative constraint. We performed dimensionality reduction on ECFP to derive rich low-dimensional features. A 12-dimensional representation is found to be optimal.

Supervised machine learning

We used the following supervised ML methods to establish a baseline for the acute oral LD₅₀ prediction:

1) **Random forest regressor**³⁶: This ensemble prediction method generates a specified number of decision trees, each based on randomly initialized conditional thresholds for filtering input values. RF models provide a consensus prediction from these decision trees. This is a shallow-learning strategy, since there is no propagation algorithm or loss function with which to adjust weights³⁷. The RF regression was performed using Scikit-learn and independently featurized by ECFP, NMF-reduced ECFP and Mordred descriptors³⁷.

2) **Gaussian process (GP) regression**³⁸: This method statistically models a prediction space by constructing a joint distribution from the multivariate normal distributions of input combination pairs. We used GP approximation as the basis for a predictive model where inputs were independently featurized by 2048-bit ECFP4 and Mordred descriptors. To reduce training cost, 200 important ECFP bits and 10 important Mordred descriptors were chosen from the RF Gini feature importance. The training was performed using the GPflow package.³⁹

3) **Deep neural network**^{12,13}: Artificial neurons form the basis of a deep neural network (DNN). Composed of a linear unit and a non-linear activation function, neurons are stacked into sequential

layers where each receives as input the output from all neurons in the preceding layer. Together, these layers form a multilayer perceptron (MLP). A fully connected DNN is used to transform input chemical features into acute oral LD₅₀ predictions. The DNN is independently featurized by ECFP and Mordred descriptors. For the ECPF descriptor architecture, a single hidden layer with 2048 neurons, batch size of 512, and Adam optimizer⁴⁰ with learning rate 0.001 are found to be sufficient. Similarly, for the Mordred descriptors, four hidden layers, each with 256 neurons, batch size of 256, and Adam optimizer with learning rate of 0.01 are found to be sufficient. Property labels are normalized, and batch normalization is applied between each layer connection.

4) Graph convolutional neural network^{32,33}: Recent advances in deep learning have put Graph Convolutional Networks (GCN) at the forefront of predictive modeling with molecular graph-encoding input data. GCNs construct a 2D adjacency matrix of a graph with binary values indicating node (atom) adjacency. Inspired by the 2D convolutions on image inputs employed in convolutional neural networks, GCNs use an irregular adjacency matrix based on direct node connectivity. The aggregation function makes use of an identity matrix to normalize the parameter inputs with respect to node adjacency, rendering the weight matrices rotationally invariant with respect to the order of node embeddings in the adjacency matrix. GCNs are convenient for molecular predictive modeling because of their ability to mimic the natural structure of any substance through its atom-bond connectivity. We employed a GCN with five convolutional layers, a convolutional base size of 64, two MLP layers with dropout of 0.153, a learning rate of 0.008, and a batch size of 64. Each graph element is assigned key chemical attributes provided in Table S2 (in SI).

The performance of all the supervised ML methods was evaluated by Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R²). We use two methods for partitioning our data into 80% training and 20% testing sets: (1) random sampling and (2) stratified

sampling on binned LD₅₀ measurements. A five-fold cross-validation is employed with a random seed to ensure consistent data splits and minimize overfit bias in performance evaluations. Bayesian optimization is a powerful technique for finding optimal hyperparameters for black-box functions⁴¹. The hyperparameters of all supervised ML methods (DNN-Mordred, DNN-ECFP, GCN, and RF) are tuned using Bayesian optimization as implemented in the GPyOpt library⁴²; parameter bounds are in Table S1.

Transfer learning

In ML, repurposing knowledge from source domains for use within a target space is a powerful application of the transfer-learning concept. Low-dimensional knowledge is shared across domains and high-dimensional knowledge is trained from the basis of common understanding. This is done in practice by initially optimizing the MLP within the source domain. Prior to training on target data, the learning rates of upstream neurons are reduced relative to later ones in order to fix early neurons used for low-dimensional feature discrimination. In certain cases, no learning is allowed (i.e., the learning rate is set to zero), a process referred to as *freezing*. Downstream neurons may be reinitialized to random weights, and layers may be added. Training is then repeated within the target space, and success is indicated by positive transfer²⁰.

Uncertainty quantification

Two approaches to uncertainty were examined. The first approach, deep ensemble, employs an ensemble of deep-learning models, each using a fixed neural network architecture with different randomly initialized layer weights (prior to training) to get multiple point estimates of prediction⁴³. The variance derived from the point estimates serves as an approximation of uncertainty. The second method, a latent-space approach, relies on the distance of a prediction point to neighboring training

points in the embedded space of the final hidden layer of the neural network⁴⁴. Recent research in chemical modeling suggests that latent distance between training and inference points can effectively act as an inexpensive approximation for uncertainty. During inference, a prediction's latent-space feature representations are projected onto the training manifold approximation. The advantage of the latent-space approach is that it does not require multiple model runs as in deep ensemble, saving the exhaustive cost of training.

Learning with abstention

Selective Prediction model²¹⁻²³: ML practitioners can use uncertainty associated with individual predictions to judge their quality. In particular, predictions with high uncertainty (i.e., low confidence) could be discounted by the human practitioner. On the other hand, a standard supervised ML approach always produces an answer, even for scenarios far outside of the training region, where such models are expected to perform poorly. Hence there is a need for an Artificial Intelligence (AI) that can replicate the human-like decision to say “I can't answer” for low-confidence/high-risk scenarios. Selective prediction is a ML paradigm where the goal is to learn a prediction model that knows when it does not know. A selective prediction model performs “learning with abstention” on its own. The selective prediction model is learned jointly as a pair (f, g) , where f is a prediction function and g is a selection function which learns whether f should be allowed to predict or abstain, as described below:

$$(f,g)(x) = \begin{cases} f(x), & \text{if } g(x) \geq \tau \\ \text{Don't know,} & \text{Otherwise} \end{cases} \quad (1)$$

where $x \in X$, the input chemical feature space, and the tolerance $\tau \in (0,1)$.

In particular, we use the SelectiveNet-based selective prediction model in this study²³. This model

architecture offers easy conversion of the main body block from a reference neural network into a corresponding network with a reject option, as illustrated in Figure 3. In a SelectiveNet, the representation (last) layer will be processed by three heads (as shown in Figure 3c): 1) A prediction head ($f(x)$) for LD_{50} , 2) a selection head ($g(x)$), a classifier that decides whether the model should abstain or not, and 3) an auxiliary head ($h(x)$) that enriches the representation layer. The joint loss for (1), given k labeled samples (S_k), is written as

$$\mathcal{L}_{(f,g)} = r(f,g | S_k) + \lambda \max(0, (c - \phi(g | S_k)))^2 \quad (2)$$

where λ is the hyperparameter that controls the coupling to the squared penalty function and c is the target coverage. The empirical coverage, ϕ , is computed as the mean of the selection function output for the k input samples. The empirical selective risk, r , is defined as

$$r(f,g | S_k) = \frac{\frac{1}{k} \sum_{i=1}^k l(f(x_i), y_i) g(x_i)}{\phi(g | S_k)} \quad (3)$$

where $l(f(x_i), y_i)$ is the regressor loss for the prediction head.

Finally, the overall loss for SelectiveNet is written as the combination of (2) and the auxiliary head loss (\mathcal{L}_h), with $\alpha = 0.5$ used in this work:

$$\mathcal{L} = \alpha \mathcal{L}_{(f,g)} + (1 - \alpha) \mathcal{L}_h \quad (4)$$

An optimal selective prediction model is arrived at by optimizing the selective risk with respect to the coverage. This is done by converging the risk-coverage curve and selecting a coverage that results in minimal selective risk⁴⁵. The choice of the hyperparameter and the neural network architecture as shown in Figure 3c are further discussed in the results.

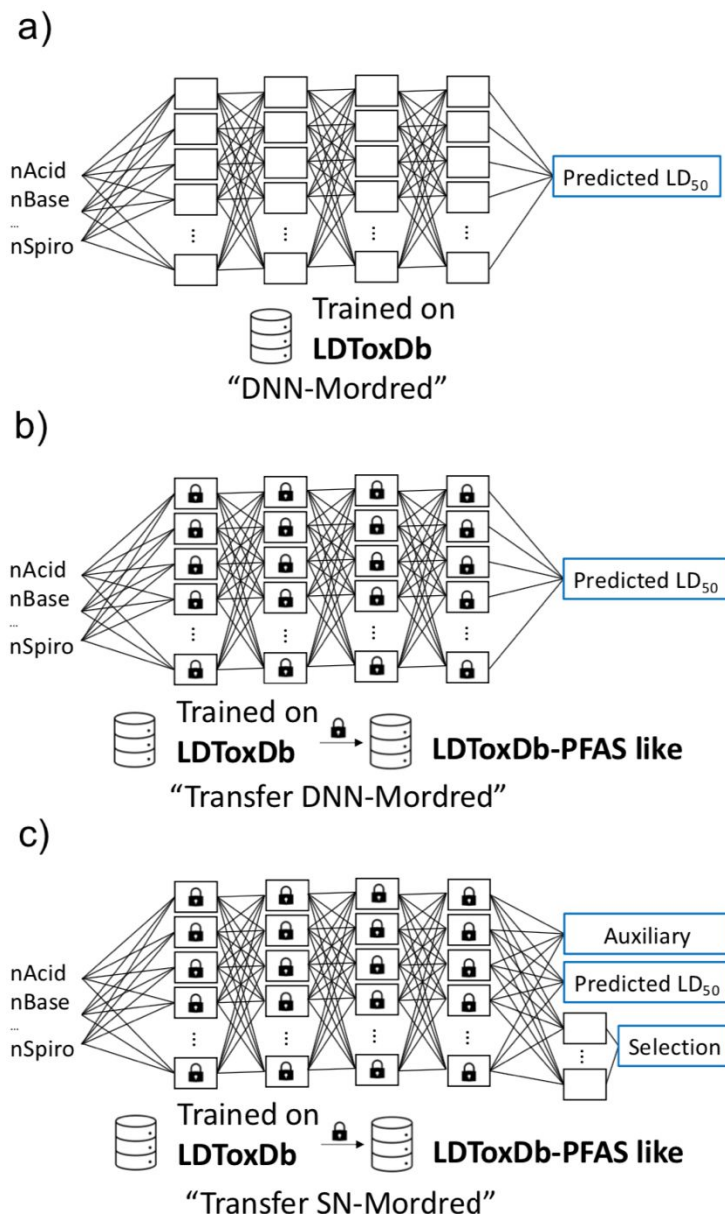


Figure 3. The three-pronged approach for machine-learning-based computational toxicology for PFAS. **a)** Source task for transfer learning. DNN-Mordred accepts molecules featured by Mordred descriptors and is trained on LDToxDB. **b)** In the transfer-learned workflow, hidden-layer weights from a) are now locked except for the final prediction layer (blue), allowing target-domain learning when training on the 518-element LDToxDB-PFAS-like dataset. **c)** In the third stage there is nearly no toxicity information for ~8,163 PFAS. Hence we now transform the DNN-Mordred from a) as the main body into a SelectiveNet architecture. The SelectiveNet architecture adds two more output heads, corresponding to an auxiliary and a decision head, for selective prediction. In this workflow, the SelectiveNet is transfer-learned using the same method used in b), except that here, uncertainty per prediction for the PFAS with unknown toxicity can be automatically converted into a decision (i.e., predict or abstain) by learning with abstention.

3 Results and Discussion

Results are organized into four subsections, each building towards the final objective of predicting the toxicity for 8,163 PFAS compounds with abstention. We first provide a review of current literature on oral rat LD₅₀ prediction, followed by results from our baseline ML benchmark models for LDToxDB. We then discuss transfer learning for the best-performing ML model on LDToxDB as the source task and LDToxDB-PFAS-like as the target task. We go on to show the benefits of uncertainty analysis for the target prediction task. Finally, we discuss the results of the SelectiveNet in predicting (or abstaining from) toxicity for 8,163 compounds from the EPA's PFAS structure list, most with no known LD₅₀ labels.

Model Baselines

Literature baselines for ML-based LD₅₀ predictions are presented in Table 2, with experiment sample size, methodology, and performance metrics for the top-performing model from each study. While variability in training datasets and testing protocols prevents a direct comparison, the best-performing models use state-of-the-art ML based on DNN. In particular, Xu et al.⁴⁶ employed a consensus based on GCN network predictions to arrive at a highly competitive model metric.

Authors	Year	Dataset	Sample size	Method	R ²	MAE	RMSE
Gadaleta et al. ²⁶	2019	CATMoS	8,448	Ab initio QSAR	0.651	0.39	0.541
Liu et al. ⁴⁷	2018	Leadscope Toxicity Db	10,363	RF regressor	0.58		0.60
Wu et al. ⁴⁸	2018	EPA ECOTOX	7,413	Consensus (RF, GBDT, ST-DNN,	0.653	0.421	0.568

Authors	Year	Dataset	Sample size	Method	R ²	MAE	RMSE
				MT-DNN)			
Xu et al. ⁴⁶	2017	admetSAR, EPA TEST, MDL	12,173	Consensus (GCN)		0.348	0.465
Bhhatarai ⁴⁹	2011	ChemIDplus	50 (PFAS only)	Linear regression. Genetic algorithm for feature selection	0.883		0.47
Zhu et al. ⁵⁰	2009	ChemIDplus	>8000	Consensus (kNN, RF, hierarchical clustering, NN)	0.71	0.39	

Table 2. Literature baselines for oral rat LD50 predictions. GBDT = Gradient Boosting Decision Tree; ST-DNN = single-task DNN; MT-DNN = multitask-DNN; kNN = k-nearest neighbors; NN = neural network; as described in the original literature. Empty cells correspond to values not reported in the same context as other metrics in their respective study.

The benchmark results from this study for the prediction of LDToxDB are presented in Table 3. Random sampling was found to give better performance compared to stratified sampling (shown in Figure S1) and is used for each model. Reported metrics (R², MAE, RMSE, and accuracy) represent average metrics computed across each testing fold for every model (i.e., five-fold cross-validation). Accuracies are provided as a supplemental metric, calculated by taking each compound’s predicted LD₅₀, converting it to mg/kg, and labeling with EPA toxicity categories. Since models used in this study are regressors, accuracies are expected to underperform in comparison to classification models present in the literature, and are hence not intended for a direct comparison with literature baselines.

From Table 3, it is observed that the ML models evaluated in this study perform in the following order, evaluating each model by the reported MAE: DNN-Mordred < RF-Mordred < GP < GCN < DNN-ECFP < RF-ECFP < RF-NMF. These results suggest that DNN with Mordred descriptors input outperforms other models with an R^2 of 0.65. While variations in datasets prevent direct one-to-one comparison to Table 2, our DNN-Mordred model yields similar performance to that reported by Zhu et al.⁵⁰ and Gadaleta et al.²⁶, justifying the evaluation of these models when further developed for the PFAS domain.

Method	Input	LDToxDB			
		R^2	MAE	RMSE	Accuracy
DNN	Mordred descriptors	0.654	0.343	0.525	0.672
DNN	2048-bit ECFP, $r=1$	0.611	0.385	0.549	0.644
GCN	Graph (node=atom, edge=bond)	0.623	0.380	0.541	0.641
GP	10 Mordred descriptors, 200 ECFP bits	0.627	0.376	0.538	0.649
RF regression	Mordred descriptors	0.647	0.372	0.523	0.65
RF regression	4096-bit ECFP, $r=2$	0.622	0.414	0.572	0.622
RF regression	NMF-reduced 4096-bit ECFP, $r=2$	0.412	0.504	0.676	0.560

Table 3. Result of 5-fold cross-validation and mean test fold metrics. Only models trained on data with random sampling are reported.

Transfer Learning on LDToxDB-PFAS-like

We next demonstrate how a DNN-Mordred model trained as the source task can be used to perform knowledge transfer within the PFAS domain. Since only 58 PFAS compounds are available with

1
2
3 reported values for oral rat LD₅₀, we use the broader 518 LDToxDB-PFAS-like subset as a measure
4
5 of whether transfer learning has a beneficial outcome. The outcome of transfer learning is directly
6
7 measured by the extent to which positive transfer occurred (i.e., no performance degradation after
8
9 transferring knowledge from source to target task)²⁰. For our comparison, we collect MAE and R²
10
11 metrics from models within the target domain both before and after transfer learning. The transfer
12
13 step involves freezing early layers trained within the source domain and reinitializing later layers to
14
15 re-train within the target domain (illustrated in Figure 3b). We refer to this model setup as “Transfer-
16
17 DNN-Mordred.” Freezing all hidden layers of DNN-Mordred and retraining the output linear layer
18
19 was found to be optimal (see Figure S2).
20
21
22
23
24
25

26 The top panel of Figure 4 shows the performance effect of transfer learning on DNN-Mordred
27
28 outcomes within LDToxDB-PFAS-like. Transfer-DNN-Mordred showed positive transfer, as seen
29
30 by the marginal decrease in error and increase in R² when looking at regression predictions,
31
32 affording greater stability in the target domain. The results are also further converted to EPA
33
34 categories (this convention will be followed through the rest of the article for a direct comparison
35
36 with EPA toxicity classes). While category IV accuracy is notably hampered, decreasing from
37
38 19.1% to 14.7%, the accuracy of category III compounds (the largest represented group in the
39
40 LDToxDB) improves from 75.9% to 76.8% and overall predictive capacity improves, as seen from
41
42 a stronger R² score.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

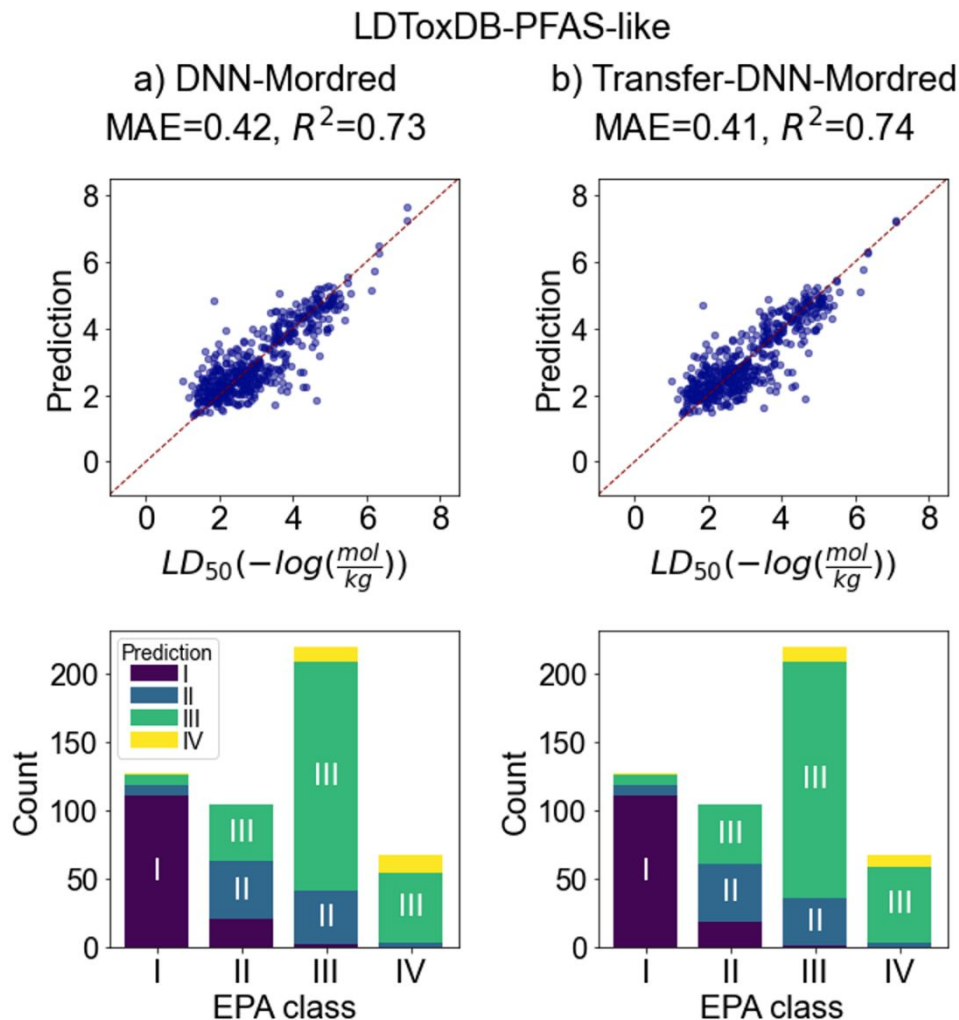


Figure 4: Comparing performances on LDToxDB-PFAS-like of original DNN-Mordred model (a) and transfer-learned DNN-Mordred model (b). Each plot presents aggregated results across five test folds. Top panels show raw regression outcomes; bottom panels convert results into corresponding EPA class. In the regression plots, horizontal axes report true labels and vertical axes are predicted.

Uncertainty Quantification and Limitations

With established evidence of positive transfer in “Transfer-DNN-Mordred” (Figure 3b), we turn our attention to calculating uncertainty per prediction. In practical settings where toxicity modeling provides consequential utility, uncertainty enables knowledgeable practitioners to discount spurious predictions. Uncertainty is evaluated here as the ability of the chosen metric to capture the model

error; in other words, a suitable measure for evaluating the efficacy of an uncertainty metric is the correlation of uncertainty with model error. We evaluate two approximations for model uncertainty, 1) deep ensemble and 2) latent space distance, and analyze the best-performing mechanism within the context of our validation set.

Literature on deep ensembles has shown that an ensemble model size as small as five is sufficient⁵¹. We evaluated the convergence of ensemble model size (Figure S3) and found that 10 DNN-Mordred models were sufficient for our purpose. To use latent space distances as a measure of uncertainty, we used a UMAP model on training-data latent space outcomes²⁸. The Euclidean distance between the latent space of inference and the nearest training point was used. PCCs grouped by superclass are provided in Table 4 and demonstrate that the deep ensemble outperforms latent space in the context of Transfer-DNN-Mordred trained on LDToxDB/LDToxDB-PFAS-like. Notably weak correlations in the largest superclasses (organoheterocyclics and benzenoids) may be explained by the high number of chemical subgroups with single members.

Superclass	Sample Size	Correlation Coeff.		Singleton Subclasses
		Deep Ensemble	Latent Space	
Organoheterocyclics	214	0.21	0.02	17
Benzenoids	168	0.38	0.05	6
Organohalogens	42	0.44	0.26	0
Lipids/lipid-likes	27	0.54	-0.17	0
Organic oxygens	23	0.57	0.59	2
Organic acids/deriv.	23	0.52	0.28	5
Organic nitrogens	10	0.47	0.20	0

Table 4. PCCs between predicted uncertainty and model error across Transfer-DNN-Mordred models on LDToxDB-PFAS-like testing folds. Compounds are grouped by taxonomic superclasses labeled by ClassyFire²⁹ to provide granularity in assessing the PCC performance. Only superclasses with greater than 10 substituents are shown. The singleton subclasses column provides the number of single-member sub-chemical classes that are present in the corresponding superclass.

As the stronger proxy metric, the standard deviations of 10 ensemble models are used to construct

95% confidence intervals (CI) representing a probabilistic forecast of the true mean of Transfer-DNN-Mordred predictions for each compound. Note that when experimental values fall outside the 95% CI, it simply means that the variance across a sample of DNN models is not high enough to accurately capture confidence with respect to the true value. We observe from Figure 5 that approximately 59.9% of experimental LD₅₀ toxicities fall within the 95% CI of Transfer-DNN-Mordred's true population mean. These results highlight two key points: (1) deep ensemble provides an appreciable mechanism for capturing model uncertainty on 59.9% of validation data; and (2) Transfer-DNN-Mordred, in its fullest capacity, conveys overconfidence (i.e., fails to accurately capture confidence) on 40.1% of validation samples.

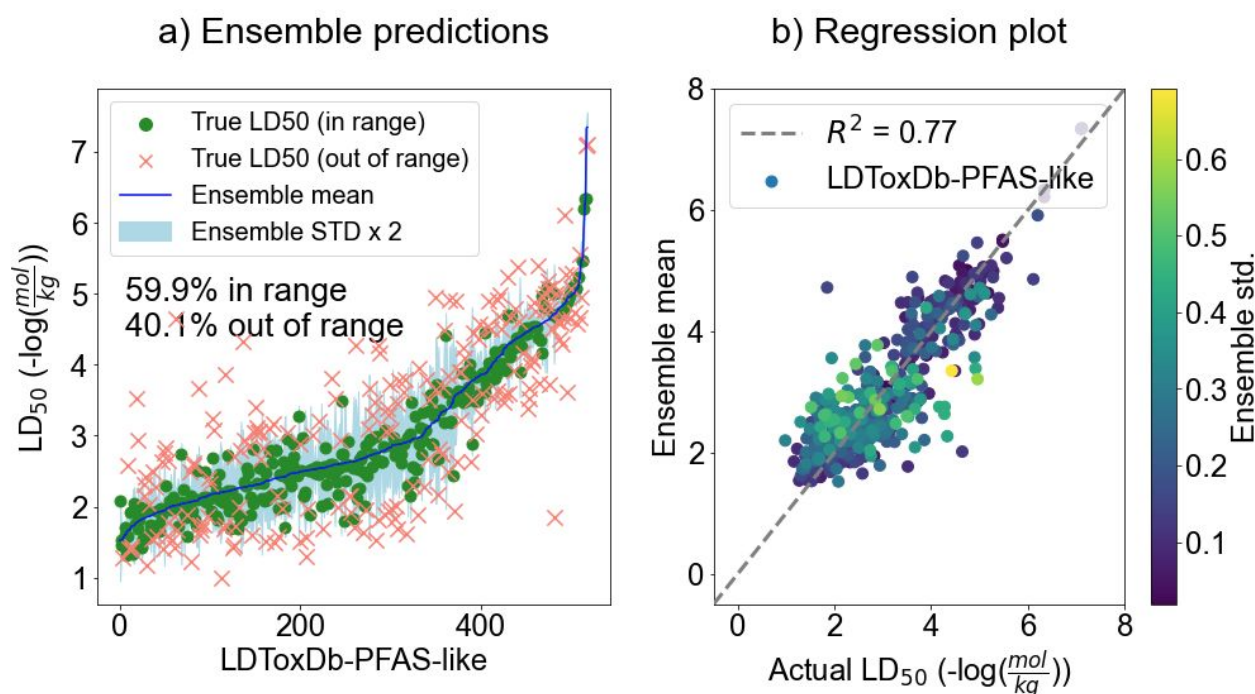


Figure 5: Uncertainty quantifications via deep ensemble for LDToxDB-PFAS-like using Transfer-DNN-Mordred models. a) Oral rat LD₅₀ indexed by the 518 LDToxDB-PFAS-like compounds. The ensemble mean is shown as a blue continuous curve, and two standard deviations as a light blue shaded region to reflect the confidence interval. Experimental oral rat LD₅₀ values that are within and outside of the confidence interval bounds are shown as green dots and red crosses, respectively. b) Experiment vs. predicted rat oral LD₅₀ corresponding to the left panel, colored by the standard value for that prediction.

The 40.1% of validation compounds that fall outside the 95% CI of the population mean of DNN-Mordred predictions invoke the larger deep-learning problem of overconfidence⁵². The deep-ensemble uncertainty fails when multiple models share a similar incorrect explanation across input space. This is an active area of research with no universal solution⁵¹. Thus, for predicting unlabeled data with a probable shift from our training set (despite efforts to isolate and transfer-learn on “PFAS-like” chemicals), we turn to an alternative in the next section: selective prediction. Using the uncertainty quantification capacity that our model does have (demonstrated on 59.9% of validation compounds), we employ a model with the means of abstaining from prediction. In practice, this approach means more cautious predictions on unlabeled data and a prioritization framework for moving forward with *in vivo* experimental trials.

Predicting PFAS compounds

In this section, we discuss predicting toxicities for unlabeled PFAS chemicals. The ensemble approach discussed in the previous section works intuitively when a clear ensemble standard deviation threshold can be used to designate compounds with high uncertainty. The definition of such a domain-dependent threshold would require some human supervision. Further, the deep-ensemble predictions can become overconfident. The prediction of a larger, unlabeled set of PFAS chemicals introduces new considerations: Can we design an AI that can understand uncertainty per prediction (when labels are not available for comparisons) and decide whether it should predict or say “I cannot predict?” Can we include an in-built safety feature in a neural network so as to minimize or avoid a catastrophic scenario? (Such a catastrophic scenario may entail a model predicting a compound as belonging to EPA class IV whereas in reality it is a highly toxic compound belonging to EPA class I.) With these considerations for prediction of PFAS compounds, the SelectiveNet architecture was implemented with DNN-Mordred operating as the main body of the

neural network (referred to as SelectiveNet Transfer DNN-Mordred, shortened to “SN-Mordred”; see Figure 3c). Transfer learning was performed as described earlier.

The optimal SN-Mordred is arrived at by minimizing the risk by constraining the coverage²³. Multiple models were trained, corresponding to coverage thresholds varying between $C=0.5$ and $C=1.0$. The selective heads of trained models were then calibrated within their respective validation sets, as recommended by Geifman et al.²³, and the total empirical risk was calculated with respect to coverage (see Table S2 in SI). A coverage threshold of 0.6 was found optimal and used to calibrate the abstention mechanism for use on LDToxDB-PFAS. Featurization of LDToxDB-PFAS by Mordred descriptors was successful for 7,058 compounds.

Figure 6a shows the distribution of selective-prediction outcomes. Since the SelectiveNet was trained with a coverage of 60%, we examine where SN-Mordred abstains by breaking down the EPA PFAS structure list by chemical superclass annotated by the ClassyFire server²⁹ (Figure 6a inset). After inference, 43 predictions are excluded as out-of-range (using water and the most toxic compound in LDToxDB as boundary limits). The dominant four superclasses within LDToxDB-PFAS (Organohalogens, benzenoids, organic acids, and benzenoids) are all predicted at a rate of approximately 75% with no trend of favorably pruning certain chemical superclasses. The most represented EPA class in LDToxDB is EPA class III (Figure 2). Consequently, it can be observed that SN-Mordred is most confident in predicting EPA class III (Figure 6b). SN-Mordred only predicted seven compounds to be in EPA class I, demonstrating considerable caution with respect to the most toxic EPA class.

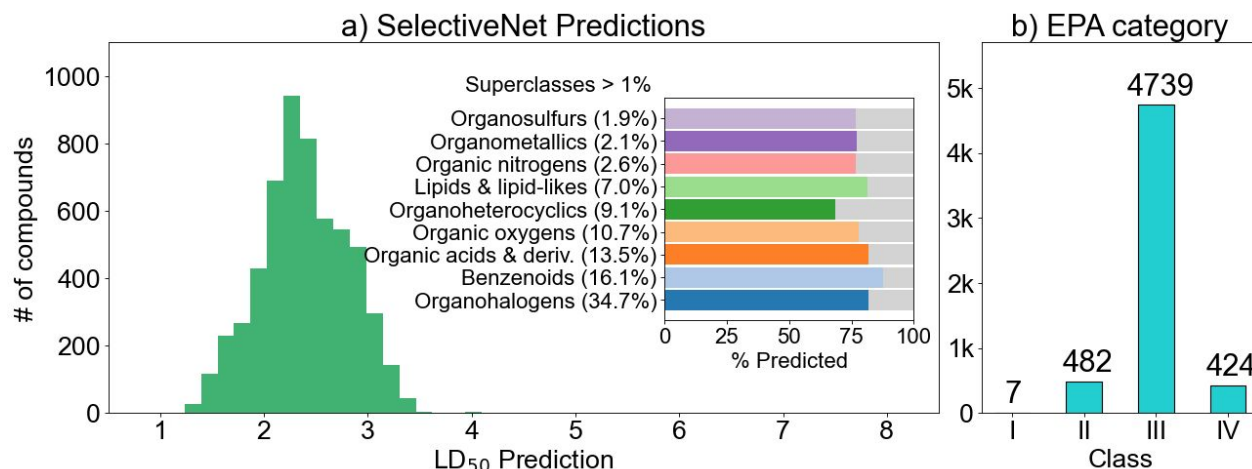
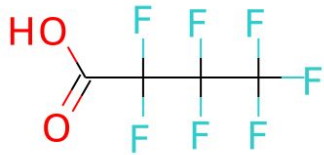
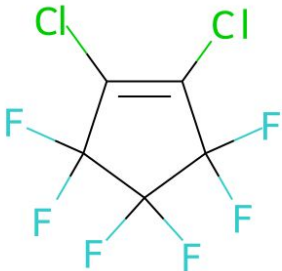
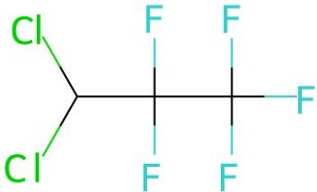
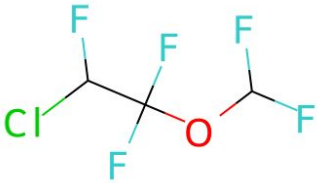
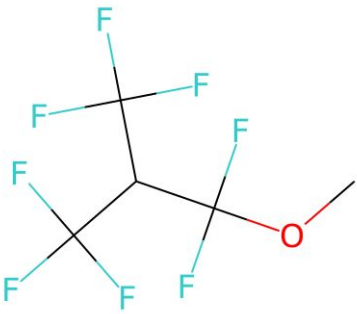
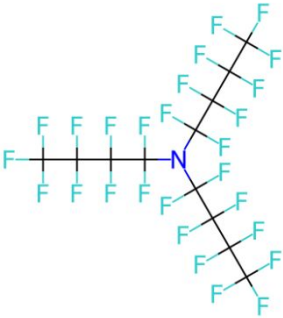
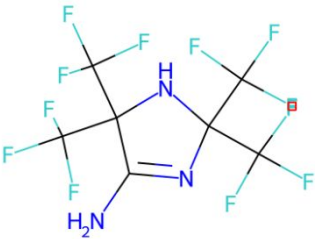
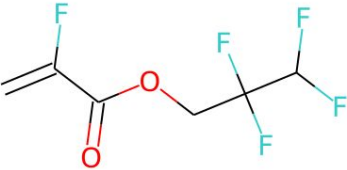
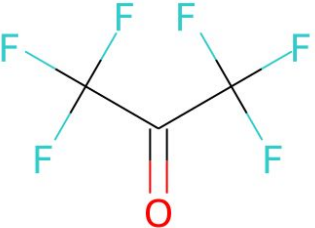


Figure 6: SelectiveNet predictions. a) Histogram of SN-Mordred predictions on EPA PFAS structure list. (Inset) Percent prediction/abstention of compounds grouped by superclass. The gray color on each bar represents the abstained fraction. The superclass label provides the percent composition of the superclass within the entire EPA PFAS structure list. b) SN-Mordred predictions categorized by EPA toxicity class. The most predicted class, level III, has 4,739 PFAS chemicals predicted to be in it.

The selective-prediction outcome for individual compounds allows us to examine how the model performs in different scenarios, particularly for the 58 PFAS compounds with known values of oral rat LD₅₀ where LDToxDB overlaps with the EPA PFAS structure list. Table 5 presents select examples of success and failure for SN-Mordred, along with scenarios where the selective mechanism refuses a prediction. Overall, the model abstained on 7 of the 58 compounds. On the 51 compounds where prediction was favorable, 47 were predicted within their actual EPA classes. For the compounds Midaflur, a highly toxic EPA class I compound, and hexafluoroacetone, a toxic EPA class II compound, the model takes a cautious approach by not predicting. A quick verification of the Transfer-DNN-Mordred model (Figure 3b) result revealed Midaflur to be incorrectly predicted as EPA class II and hexafluoroacetone as class III, thereby validating the catastrophic scenario and demonstrating the benefit of the abstention mechanism.

Chemical Structure	Chemical	Transfer SN-Mordred Inference	Actual Class
	Perfluorobutanoic acid (PFBA or HFBA)	III	III
	Cyclopentene, 1,2-dichloro- 3,3,4,4,5,5-hexafluoro-	II	II
	3,3-Dichloro-1,1,1,2,2- pentafluoropropane	IV	IV
	Enflurane	IV	IV
	1,1,1,3,3-Pentafluoro-3-methoxy-2- (trifluoromethyl)propane	II	II

Chemical Structure	Chemical	Transfer SN-Mordred Inference	Actual Class
	Perfluorotributylamine (PFTBA)	IV	IV
	Midaflur	Abstain	I
	2-Propenoic acid, 2-fluoro-, 2,2,3,3- tetrafluoropropyl ester	Abstain	II
	Hexafluoroacetone (HFA)	Abstain	II

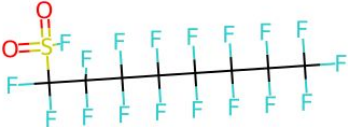
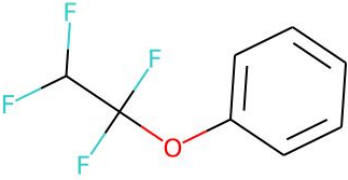
Chemical Structure	Chemical	Transfer SN-Mordred Inference	Actual Class
	Perfluorooctanesulfonyl fluoride (POSF)	II	III
	(1,1,2,2-Tetrafluoroethoxy)benzene	III	IV

Table 5. Results for chemicals successfully predicted, successfully abstained, and poorly predicted by SN-Mordred model. 2D structures were generated with RDKit²⁷ and chemical names were obtained from the EPA CompTox Chemicals Dashboard⁵³. Green shading corresponds to compounds where SN-Mordred predicted the compound in the correct EPA category. Boxes with blue shading corresponds to SN-Mordred abstention. Red corresponds to compounds that were predicted in the wrong EPA category by SN-Mordred.

To underpin the results and decisions returned using AI, future efforts could include the development of deep learning or QSAR models using molecular descriptors strongly correlated with acute toxicity⁴⁹ or by building local QSAR models from closely similar structures⁵⁴. Such efforts would provide a physical and mechanical basis grounded in molecular structure for interpreting toxicity estimates from AI. The derived relationships could further reduce the incidence of catastrophic decisions from AI predictions.

4 Conclusions

Targeted environmental cleanup of PFAS requires an understanding of PFAS toxicity. We present

a rigorous ML-based computational toxicology workflow that we use to predict toxicity for ~8,163 PFAS compounds whose toxicities are poorly understood. We achieve this result by transfer learning on knowledge of all organic compounds with known values of oral rat LD₅₀ to predict on the PFAS compound space with informed uncertainties. Learning by abstention provides an automatic mechanism for converting uncertainty per prediction into model decisions. Organ-on-a-chip systems, now possible through advancements in microfluidic technologies, have allowed for the emulation of *in vivo* physiological conditions^{55,56}. The selective prediction model can be used for deriving decisions on compounds whose toxicity values cannot be predicted reliably. The model decisions can be used to drive on-demand active learning of toxicology experiments using the organ-on-a-chip setup.

In this age of big data, neural networks have been widely adopted for applications in the chemical sciences community. We anticipate that the AI4PFAS workflow can be used for predicting many other toxic endpoints. Some of the approaches presented in this study can be used to add a layer of safety to supervised ML predictions, especially in mission-critical applications such as computational toxicology. We hope that some of the good practices presented in this study are adopted and expanded on by the wider chemical sciences community.

5 Acknowledgments

This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. J. F. and G.S. would like to thank Dr. Benjamin Sanchez-Lengeling for fruitful discussions on graph neural networks.

References

- (1) U.S. Environmental Protection Agency. PFOA, PFOS and Other PFASs: Basic Information on PFAS. No date. <https://www.epa.gov/pfas/basic-information-pfas> (accessed on Dec. 1, 2020).
- (2) Interstate Technology & Regulatory Council. *Naming Conventions and Physical and Chemical Properties of Per- and Polyfluoroalkyl Substances (PFAS)*. 2017. https://pfas-1.itrcweb.org/fact_sheets_page/PFAS_Fact_Sheet_Naming_Conventions_April2020.pdf (accessed on April 8, 2021).
- (3) Military Times website. <https://www.militarytimes.com/news/pentagon-congress/2019/03/07/2-billion-cost-to-clean-up-water-contamination-at-military-bases-defense-official-says/> (accessed on April 8, 2021).
- (4) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **2019**, *12*, 100096.
- (5) Patlewicz, G.; et al. A chemical category-based prioritization approach for selecting 75 per- and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. *Environmental Health Perspectives* **2019**, *127*(01), 014501.
- (6) Hartung, T. Toxicology for the twenty-first century. *Nature* **2009**, *460*(7252), 208–212.
- (7) Richarz, A. N. Big Data in Predictive Toxicology: Challenges, Opportunities and Perspectives. Chapter 1 in *Big Data in Predictive Toxicology*, **2019**, Royal Society of Chemistry, 1–37.
- (8) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicological Sciences* **2018**, *165*(1), 198–212.
- (9) Ciallella, H. L.; Zhu, H. Advancing computational toxicology in the big data era by artificial

- intelligence: Data-driven and mechanism-driven modeling for chemical toxicity. *Chemical Research in Toxicology* **2019**, 32(4), 536–547.
- (10) Luechtefeld, T.; Rowlands, C.; Hartung, T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicology Research* **2018**, 7(5), 732–744.
- (11) Sze, V.; Chen, Y. H.; Emer, J.; Suleiman, A.; Zhang, Z. Hardware for machine learning: Challenges and opportunities. *Proc. IEEE Custom Integrated Circuits Conference*, Austin, TX, April 30–May 3, **2017**, 1–8.
- (12) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, 521(7553), 436–444
- (13) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, 61, 85–117.
- (14) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling* **2015**, 55(2), 263–274.
- (15) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv preprint* **2014**, arXiv:1406.1231.
- (16) Mayr, A.; et al. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016**, 3, 80.
- (17) Kleinstreuer, N. C.; Tong, W.; Tetko, I. V. Computational toxicology. *Chemical Research in Toxicology* **2020**, 33, 687–688
- (18) Sivaraman, G.; Jackson, N.; Sanchez-Lengeling, B.; Vasquez-Mayagoitia, A.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. A machine learning workflow for molecular analysis: Application to melting points. *Machine Learning: Science and Technology* **2020**, 1, 025015.
- (19) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of machine

learning methods in drug toxicity prediction. *Current Topics in Medicinal Chemistry* **2018**, 18(12), 987–997.

(20) Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* **2010**, IGI global, pp. 242–264.

(21) Kompa, B.; Snoek, J.; Beam, A. L. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine* **2021**, 4(1), 1–6.

(22) Cortes, C.; DeSalvo, G.; Mohri, M. Learning with rejection. In *International Conference on Algorithmic Learning Theory* **2016**, Springer, Cham, pp. 67–82.

(23) Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *International Conference on Machine Learning* **2019**, pp. 2151–2159.

(24) U.S. Environmental Protection Agency. Toxicity Estimation Software Tool (TEST). <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test> (accessed on Dec 1, 2020).

(25) Kleinstreuer, N. C.; Karmaus, A. L.; Mansouri, K.; Allen, D. G.; Fitzpatrick, J. M.; Patlewicz, G. Predictive models for acute oral systemic toxicity: A workshop to bridge the gap from research to regulation. *Computational Toxicology* **2018**, 8, 21–24.

(26) Gadaleta, D.; et al. SAR and QSAR modeling of a large collection of LD₅₀ rat acute oral toxicity data. *Journal of Cheminformatics* **2019**, 11(1), 58.

(27) Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. <https://www.rdkit.org/> (accessed on Dec 1, 2020).

(28) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* **2018**, *arXiv:1802.03426*.

(29) Feunang, Y. D.; et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016**, 8(1), 61.

- (30) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, 10(1), 4.
- (31) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **2010**, 50(5), 742–754.
- (32) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gomez-Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Proc. of Advances in Neural Information Processing Systems 28*, Montreal, Canada, December 7–12, **2015**, pp. 2215–2223.
- (33) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *Proceedings of Machine Learning Research* **2017**, 70, 1263–1272.
- (34) Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, 5(2), 111–126.
- (35) Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, 401(6755), 788–791.
- (36) Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5–32.
- (37) Pedregosa, F.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, 12, 2825–2830.
- (38) Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning* **2003**, Springer, Berlin, Heidelberg, pp. 63–71.
- (39) Matthews, A. G. de G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrà, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research* **2017**, 18(40), 1–6.
- (40) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* **2014**, *arXiv:1412.6980*.

- (41) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **2015**, 104(1), 148–175.
- (42) González, J.; Dai, Z. GPyOpt: a Bayesian optimization framework in Python. **2016**.
<http://github.com/SheffieldML/GPyOpt> (accessed on April 8, 2021).
- (43) Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6405–16. NIPS’17. Long Beach, California, USA: Curran Associates Inc., **2017**.
- (44) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical Science* **2019**, 10(34), 7913–7922.
- (45) El-Yaniv, R.; Wiener, Y. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research* **2010**, 11, 1605–1641.
- (46) Xu, Y.; Pei, J.; Lai, L. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of Chemical Information and Modeling* **2017**, 57(11), 2672–2685.
- (47) Liu, R.; Madore, M.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. Assessing deep and shallow learning methods for quantitative prediction of acute chemical toxicity. *Toxicological Sciences* **2018**, 164(2), 512–526.
- (48) Wu, K.; Wei, G. W. Quantitative toxicity prediction using topology based multitask deep neural networks. *Journal of Chemical Information and Modeling* **2018**, 58(2), 520–531.
- (49) Bhattacharai, B.; Gramatica, P. Oral LD₅₀ toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. *Molecular Diversity* **2011**, 15(2), 467–76.
- (50) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative

structure–activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology* **2009**, 22(12), 1913–1921.

(51) Ovadia, Y.; et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *arXiv preprint* **2019**, arXiv:1906.02530.

(52) Caldeira, J.; Nord, B. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology* **2020**, 2(1), 015002.

(53) U.S. Environmental Protection Agency. CompTox Chemicals Dashboard.
<https://comptox.epa.gov/dashboard/> (accessed on Jan 18, 2020).

(54) Vukovic, K.; Gadaleta, D.; Benfenati, E. Methodology of aiQSAR: A group-specific approach to QSAR modelling. *Journal of Cheminformatics* **2019**, 11, Article 27.

(55) Cho, S.; Yoon, J. Y. Organ-on-a-chip for assessing environmental toxicants. *Current Opinion in Biotechnology* **2017**, 45, 34–42.

(56) Huh, D.; Matthews, B. D.; Mammoto, A.; Montoya-Zavala; M., Hsin; H. Y.; Ingber, D. E. Reconstituting organ-level lung functions on a chip. *Science* **2010**, 328(5986), 1662–1668.