JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

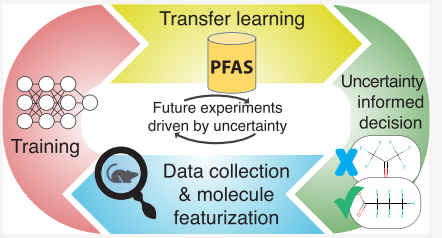
Uncertainty-Informed Deep Transfer Learning of Perfluoroalkyl and Polyfluoroalkyl Substance Toxicity

Jeremy Feinstein, Ganesh Sivaraman, Kurt Picel, Brian Peters, Álvaro Vázquez-Mayagoitia, Arvind Ramanathan, Margaret MacDonell, Ian Foster, and Eugene Yan*

Cite This: J. Chem. Inf. Model. 2021, 61, 5793−5803 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Perfluoroalkyl and polyfluoroalkyl substances (PFAS) pose a significant hazard because of their widespread industrial uses, environmental persistence, and bioaccumulation. A growing, increasingly diverse inventory of PFAS, including 8163 chemicals, has recently been updated by the U.S. Environmental Protection Agency. However, 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 LD_{50}) of PFAS compounds. To address the scarcity of toxicity information for PFAS, publicly available datasets of oral rat LD_{50} for all organic compounds are aggregated and used to develop state-of-the-art ML source models for transfer learning. A total of 519 fluorinated compounds containing two 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.



数据集

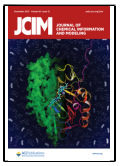
1. INTRODUCTION

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 9 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 alone.[3] The U.S. Environmental Protection Agency (EPA)'s Distributed Structure-Searchable Toxicity (DSSTox) database of PFAS structures,[4] as recently updated, contains over 8163 PFAS chemicals. PFAS compounds can be broadly classified into polymeric and nonpolymeric families.[2] This study addresses nonpolymeric 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 the identification of nontoxic 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 the 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 structures.

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 power.[7−11] In particular, deep learning for the 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

Received: October 16, 2021
Published: December 14, 2021



random forests (RFs).[14−18] The prediction performance of deep learning can be further improved by techniques such as multitasking and transfer learning.[16,19−24] However, a compound's toxicity is affected by multiple chemical and biological factors, adding complexity to the prediction of this crucial property.[25]

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 h after the 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 the reported values of oral rat LD_{50} point estimates. Consequently, only a limited number of studies have tackled computational toxicity of the PFAS compounds. Bhatarai et al.[26] use multiple linear regression on DRAGON descriptors selected by the genetic algorithm to produce 100 different models ordered and ranked by performance. Their approach exploits chemical similarity for toxicity prediction and is trained with oral mouse and rat LD_{50}s for 58 and 50 PFAS chemicals, respectively. The study of Bhatarai et al. constitutes perhaps the only "generalist" (i.e., LD_{50}) study of PFAS toxicity prediction at this time. Other computational approaches in the PFAS domain are trained on bioactivity assays; an example is a study by Cheng et al.,[27] where thousands of organic molecules were used for training and inference of 26 bioassays. Hoover et al.[28] performed in vitro cytotoxicity studies on four common PFAS compounds and in silico investigations of possible mixture combinations.

Table 1. EPA Toxicity Classes

category	toxicity	dosage (mg/kg body weight)
I	high	≤50
II	moderate	>50 to 500
III	low	>500 to 5000
IV	very low	>5000

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. Transfer learning has two components: (1) source task training, for which there is an abundance of labeled data, and (2) target task training, 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 training.[29] We aggregate the reported values of oral rat LD_{50} 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 LD_{50} predictions and show that our source ML baselines are competitive. Then, we identify 519 fluorinated compounds containing two or more C-F bonds within LDToxDB (which we will refer to as PFAS-like) with known LD_{50} labels. The 58 PFAS compounds found in the 519 PFAS-like are separated (which we will refer to as PFAS-58). The PFAS-58 is kept out of all transfer-learning model trainings and used exclusively as a global validation group. PFAS-like compounds (461) after excluding PFAS-58 are used with knowledge transferred from the best-performing source task to generate the target models with access to uncertainty. The rationale is that 461 PFAS-like compounds are the closest chemical family in our database to the broader 8163 PFAS compounds; hence, it will be important to understand how a transfer-learned target model performs on PFAS-like compounds where oral rat LD_{50} labels are available, before attempting predictions for PFAS compounds with unknown oral rat LD_{50}. 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 low.[30−32] 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 8163 EPA PFAS compounds with no known oral rat LD_{50}; 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 Section 4. 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.

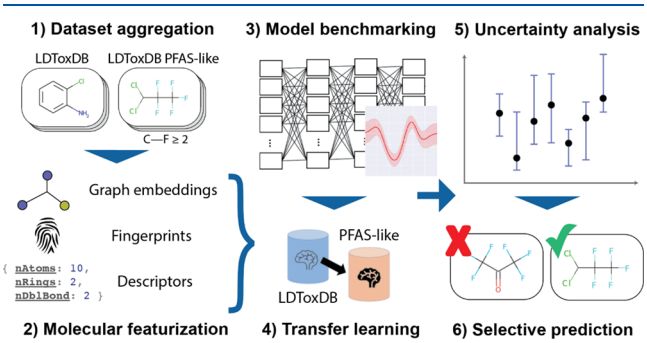


Figure 1. AI4PFAS workflow for PFAS toxicity prediction.

2. METHODS

2.1. 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 LD_{50} measurements aggregated from the EPA Toxicity Estimation

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

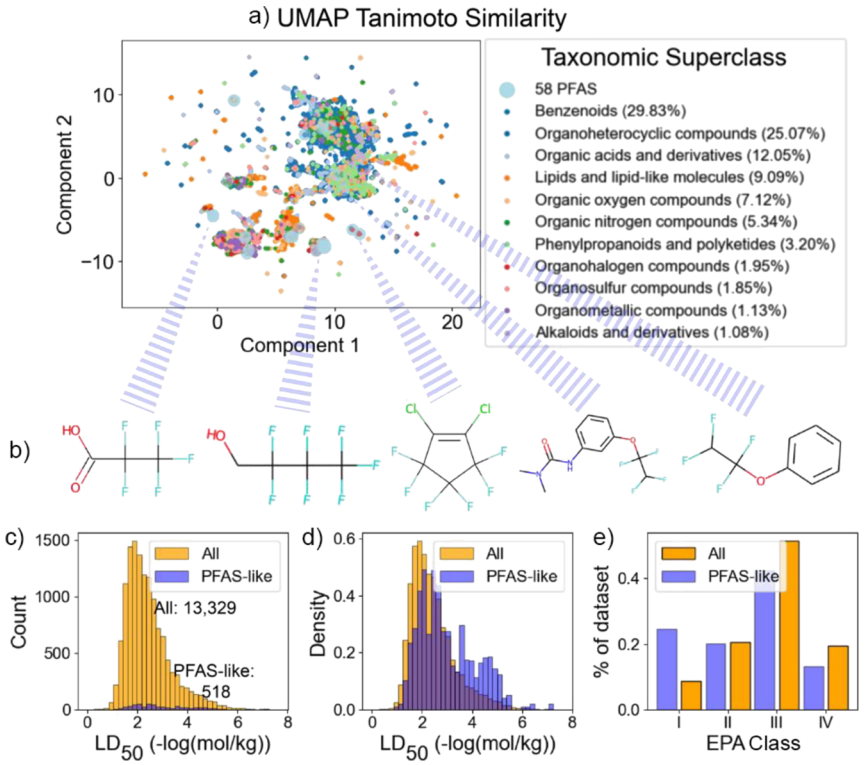


Figure 2. Visualization of datasets. (a) Visual exploration of LDToxDB with dimensionality reduction performed with uniform manifold approximation and projection (UMAP) on a Tanimoto similarity matrix.[37] Clusters are colored according to the chemical superclasses annotated by the ClassyFire server.[38] Fifty-eight 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 LD_{50} distribution for LDToxDB (labeled as "All)." The PFAS-like subset distribution is shown for reference. (d) Normalized histograms showing the LD_{50} 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.

Software Tool (TEST), NIH Collaborative Acute Toxicity Modeling Suite (CATMoS), and National Toxicology Program datasets.[33−35] LD_{50} 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 RDKit[36] and duplicate molecules were removed by querying each compound's hashed InChlKey.

To broadly identify PFAS-like compounds within LDToxDB, molecules with two or more C − F bonds were identified using an RDKit SMARTS query and tagged as a PFAS-like representative subset of the labeled LDToxDB compounds. Such compounds with two or more C−F bonds would be polyfluorinated, likely alkyl, but may not be designated as PFAS in various databases. The resulting 519 compounds, referred to as "LDToxDB-PFAS-like" and which include 58 compounds formally labeled as PFAS-58, served as an important validation group to confirm that models trained on LDToxDB were able to predict the toxicity of PFAS and PFAS-like compounds via the chemical-structure similarity. Finally, 8163 PFAS compounds were extracted from the EPA DSSTox database,[4] most of which have no LD_{50}; referred to as "LDToxDB-PFAS"; and 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.

2.2. 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 LD_{50} point estimates. Furthermore, unsupervised ML can be applied to the chemical features in order to deduce chemical insights. Figure 1 shows the full set of features and tools developed as a part of the AI4PFAS workflow; details are discussed below.

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

2.2.1. Mordred Descriptors. We used the Mordred software package to generate 1800 unique molecular descriptors for each compound directly from RDKit molecules.[39] Mordred provides quick featurization of a molecular dataset by generating a vast array of two- and three-dimensional descriptor characteristics from the SMILES input. The full reference list of Mordred descriptors is available elsewhere.[39] We trimmed down the 1800 descriptors to 300 using the Pearson correlation coefficient (PCC) analysis to remove redundant features.

2.2.2. Extended-Connectivity Fingerprinting (ECFP). ECFP provides a mechanism for representing topological chemical space by iteratively exploring substructure connectivity at a provided radius around each atom of a molecule.[40] Numeric representations are created for each substructure identified in

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

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 using the open-source RDKit package for Python.[36]

2.2.3. Molecular Graph Encoding. Molecular graph encoding improves on ECFP by representing molecules as graphs of an arbitrary size with nodes representing atoms and edges representing bonds.[41,42] Each entity is given characteristic traits, which for nodes may include (but are not limited to) atomic identity, the number of valence electrons, formal charge, and hybridization, and for edges, the bond order and conjugation status. We have adopted the graph representation and the corresponding graph convolutional neural network from the MOLAN workflow.[18]

2.2.4. Non-Negative Matrix Factorization (NMF). NMF is a dimensionality reduction technique that derives basis vectors under a non-negative constraint.[43,44] A given matrix M is decomposed into two component matrices, namely, W and H, with the condition that all three matrix elements are non-negative. We applied the NMF, as implemented in the Scikit-learn,[45] to ECFP to derive a rich low-dimensional feature matrix, W. A grid search was performed to find the optimal number of components that corresponds to the lowest reconstruction error.

2.3. Supervised ML. We used the following supervised ML methods to establish a baseline for the acute oral LD_{50} prediction:

2.3.1. RF Regressor. This ensemble prediction method generates a specified number of decision trees, each based on randomly initialized conditional thresholds for filtering input values.[46] RF models provide a consensus prediction from these decision trees. This is a shallow-learning strategy because there is no propagation algorithm or loss function with which to adjust weights.[45] The RF regression was performed using Scikit-learn and independently trained on ECFP, NMF-reduced ECFP, and Mordred descriptors.[45] For all RF models, 4096 estimators with a maximum tree depth of 32 were used.

2.3.2. Gaussian Process (GP) Regression. This method statistically models a prediction space by constructing a joint distribution from the multivariate normal distributions of input combination pairs.[47] We used GP approximation as the basis for a predictive model where inputs were independently trained on 2048-bit ECFP4 and Mordred descriptors. To reduce the 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.[48]

2.3.3. Deep Neural Network. Artificial neurons form the basis of a deep neural network (DNN).[12,13] Composed of a linear unit and a nonlinear 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 LD_{50} predictions. The DNN is independently trained on ECFP and Mordred descriptors. For the ECFP descriptor architecture, a single hidden layer with 2048 neurons, batch size of 512, and an Adam optimizer[49] with a learning rate of 0.001 are found to be sufficient. Similarly, for the Mordred descriptors, four hidden layers, each with 256 neurons, batch size of 256, and an Adam optimizer with a learning rate of 0.01 are found to be sufficient. Property labels are normalized, and batch normalization is applied between each layer connection.

2.3.4. Graph Convolutional Neural Network. Recent advances in deep learning have put graph convolutional networks (GCNs) at the forefront of predictive modeling with molecular graph-encoding input data.[41,42] GCNs construct a two-dimensional (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 a 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 the Supporting Information).

The performance of all the supervised ML methods was evaluated by the mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). We use two methods for partitioning our data into 80% training and 20% testing sets: (1) random sampling and (2) stratified sampling on binned LD_{50} measurements. A fivefold cross-validation is employed with a random seed to ensure consistent data splits and minimize overfit bias in performance evaluation. Bayesian optimization is a powerful technique for finding optimal hyperparameters for black-box functions.[50] The hyperparameters of all supervised ML methods (DNN-Mordred, DNN-ECFP, GCN, and RF) are tuned using Bayesian optimization, as implemented in the GPyOpt library;[51] parameter bounds are in Table S1.

2.4. 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 transfer.[29]

2.5. 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 obtain multiple point estimates of prediction.[52] 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 network.[53] Recent research in chemical modeling suggests that the latent distance between training and inference points can

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

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.

2.6. Learning with Abstention: Selective Prediction Model. ML practitioners can use uncertainty associated with individual predictions to judge their quality.[30−32] 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 the training region, where such models are expected to perform poorly. Hence, there is a need for 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 an 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 that learns whether f should be allowed to predict or abstain, as described below:

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

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 study.[32] 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 follows:

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

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

Finally, the overall loss for the SelectiveNet is written as the combination of (2) and the auxiliary head loss ($\mathcal{L}_h$), with α = 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 risk.[54] The choice of the hyperparameter and the neural network architecture, as shown in Figure 3c, is further discussed in the results.

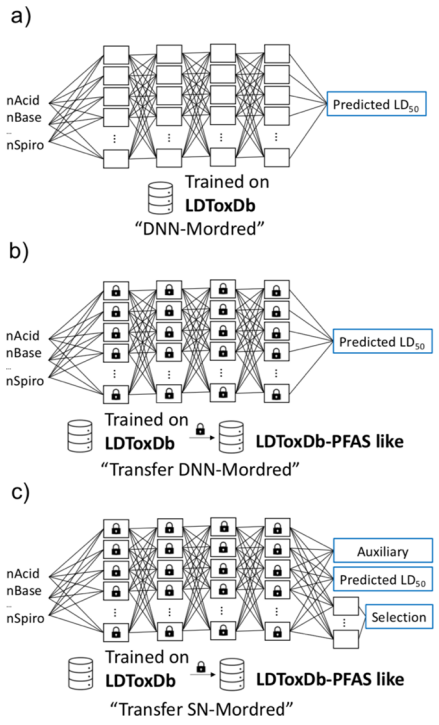


Figure 3. Three-pronged approach for ML-based computational toxicology for PFAS. (a) Source task for transfer learning. DNN-Mordred accepts molecules represented by Mordred descriptors and is trained on LDToxDB. (b) In the transfer-learned workflow, hidden-layer weights from (a) are now locked except for the final prediction layer (blue), allowing for target-domain learning when training on the 519-element LDToxDB-PFAS-like dataset. (c) In the third stage, there is nearly no toxicity information for ∼8163 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 toward the final objective of predicting the toxicity for 8163 PFAS compounds with abstention. We first provide a review of the current literature on oral rat LD_{50} 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 8163 compounds from the EPA's PFAS structure list, most with no known LD_{50} labels.

3.1. Model Baselines. Literature baselines for ML-based LD_{50} 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 the state-of-the-

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

Table 2. Literature Baselines for Oral Rat LD50 Predictions[a]

reference	year	dataset	sample size	method	R^2	MAE	RMSE
Gadaleta et al.[35]	2019	CATMoS	8448	ab initio QSAR	0.651	0.39	0.541
Liu et al.[55]	2018	Leadscope Toxicity Db	10,363	RF regressor	0.58		0.60
Wu et al.[20]	2018	EPA ECOTOX	7413	consensus (RF, GBDT, ST-DNN, MT-DNN)	0.653	0.421	0.568
Xu et al.[19]	2017	admetSAR, EPA TEST, MDL	12,173	consensus (GCN)		0.348	0.465
Bhhatarai[26]	2011	ChemIDplus	50 *(PFAS only)*	linear regression. genetic algorithm for feature selection	0.883		0.47
Zhu et al.[56]	2009	ChemIDplus	>8000	consensus (kNN, RF, hierarchical clustering, NN)	0.71	0.39	

[a]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.

Table 3. Results of Fivefold Cross-Validation and Mean Test Fold Metrics[a]

		LDToxDB			
method	input	R^2	MAE	RMSE	accuracy
DNN	Mordred descriptors	0.658	0.342	0.516	0.680
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	0.650
RF regression	Mordred descriptors	0.647	0.372	0.523	0.660
RF regression	4096-bit ECFP, r = 2	0.584	0.410	0.569	0.623
RF regression	NMF-reduced 4096-bit ECFP, r = 2	0.464	0.479	0.645	0.574

[a]Only models trained on data with random sampling are reported.

art ML based on DNN. In particular, Xu et al.[19] employed a consensus based on GCN network predictions to arrive at a highly competitive model metric.

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 (R^2, MAE, RMSE, and accuracy) represent average metrics computed across each testing fold for every model (i.e., fivefold cross-validation). Accuracies are provided as a supplemental metric, calculated by taking each compound's predicted LD_{50}, converting it to mg/kg and labeling with EPA toxicity categories. Because 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 descriptor input outperforms other models with an R^2 of 0.65. While variations in datasets prevent direct one-to-one comparison with Table 2, our DNN-Mordred model yields performance similar to that reported by Zhu et al.[56] and Gadaleta et al.,[35] justifying the evaluation of these models when further developed for the PFAS domain.

3.2. 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. For all results from this point onward, we will keep the 519 LDToxDB-PFAS-like separate from the source training set. In addition, the PFAS-58 will be excluded from all source/target training tasks and will be exclusively used in global validation. It follows that 461 LDToxDB-PFAS-like will be used for the target training tasks. Given that there are only a very few samples for the target task learning in the domain of LDToxDB-PFAS-like compounds, our goal was to ensure that there was no performance degradation after knowledge transfer was performed. The outcome of transfer learning is directly measured by the extent to which positive transfer occurred (i.e., statistically significant performance gain after transferring knowledge from the source to the target task).[29] For our comparison, we collect MAE and R^2 metrics from models within the target domain both before and after transfer learning. The hyperparameter tuning is performed independently for the transfer-learning source model to avoid numeric instabilities in training on a target task with less labeled samples. The details of these hyperparameters are discussed in Table S1 (F). The transfer step involves freezing early layers trained within the source domain and reinitializing later layers to retrain within the target domain (illustrated in Figure 3b). We refer to this model setup as "transfer-DNN-Mordred." Freezing the initial two hidden layers of DNN-Mordred and retraining all subsequent weights provided optimal positive transfer (see Figure S2).

The top panel of Figure 4 shows the performance effect of transfer learning on DNN-Mordred outcomes within LDToxDB-PFAS-like. Transfer-DNN-Mordred showed positive transfer, as seen by the stability in error and R^2 when looking at regression predictions, affording favorable inference in the target domain. The results are also further converted to EPA categories (this convention will be followed through the rest of the article for a direct comparison with EPA toxicity classes). Notable is the drastic increase in the predictive power and the improved ability to classify highly toxic level I compounds.

3.3. Uncertainty Quantification and Limitations. With established evidence of statistically significant 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

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

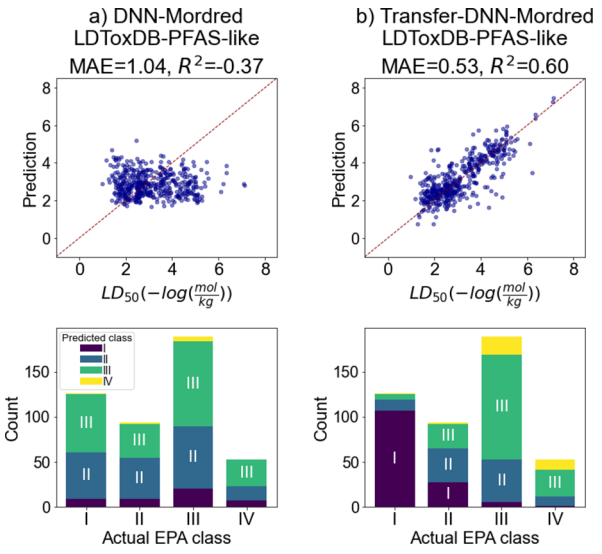


Figure 4. Comparing performances on 461 LDToxDB-PFAS-like of the original DNN-Mordred model (a) and transfer-learned DNN-Mordred model (b). Each plot presents aggregated results across five test folds. Top panels show raw regression outcomes; bottom panels convert results into corresponding EPA classes. In the regression plots, horizontal axes report true labels, and vertical axes are predicted.

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 the model error. We evaluate two approximations for the model uncertainty: (1) deep ensemble and (2) latent space distance, as well as analyze the best-performing mechanism within the context of our validation set.

The literature on deep ensembles has shown that an ensemble model size as small as five is sufficient.[57] We evaluated the convergence of the 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 the UMAP model on training-data latent space outcomes.[37] The Euclidean distance between the latent space of inference and the nearest training point was used. PCCs grouped by the superclass are provided in Table 4, and

Table 4. PCCs between the Predicted Uncertainty and Model Error across Transfer-DNN-Mordred Models on 461 LDToxDb-PFAS-Like Testing Folds[a]

superclass	sample size	correlation coeff.		singleton subclasses
superclass	sample size	deep ensemble	latent space	
organoheterocyclics	213	0.29	0.20	17
benzenoids	160	0.30	0.04	6
lipids/lipid-likes	19	0.82	0.07	0
organic oxygens	19	0.79	−0.09	1
organic acids/deriv.	18	0.11	−0.10	5
organohalogens	15	0.22	−0.18	1

[a]Compounds are grouped by taxonomic superclasses labeled by ClassyFire[38] to provide granularity in assessing the PCC performance. Only superclasses with greater than 10 substituents are shown. The singleton subclass column provides the number of single-member subchemical classes that are present in the corresponding superclass.

they 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 (organoheterocylics and benzenoids) may be explained by the high number of chemical subgroups with single members.

As a 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 the 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 83.9% of experimental LD_{50}

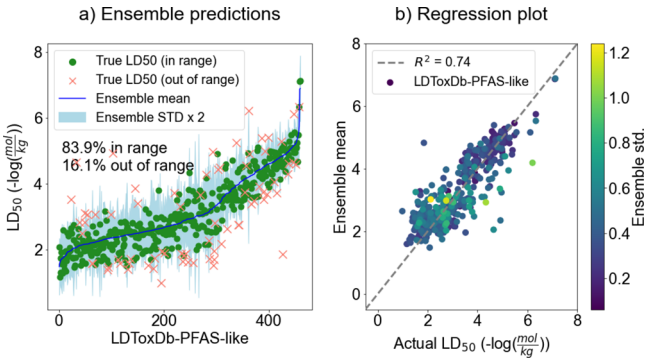


Figure 5. Uncertainty quantifications via deep ensemble for LDToxDB-PFAS-like using Transfer-DNN-Mordred models. (a) Oral rat LD_{50} indexed by the 461 LDToxDB-PFAS-like compounds. The ensemble mean is shown as a blue continuous curve, and two standard deviations as a light blue shaded region to reflect the CI. Experimental oral rat LD_{50} values that are within and outside of the CI bounds are shown as green dots and red crosses, respectively. (b) Experiment vs predicted rat oral LD_{50} corresponding to the left panel, colored by the standard value for that prediction.

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 83.9% of validation data and (2) transfer-DNN-Mordred, in its fullest capacity, conveys overconfidence (i.e., fails to accurately capture confidence) on 16.1% of validation samples.

The 16.1% of validation compounds that fall outside the 95% CI of the population mean of DNN-Mordred predictions invoke the larger deep-learning problem of overconfidence.[58] The deep-ensemble uncertainty fails when multiple models share a similar incorrect explanation across the input space. This is an active area of research with no universal solution.[57] 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 has (demonstrated on 83.9% 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.

3.4. Predicting PFAS Compounds. In this section, we discuss predicting toxicities for unlabeled PFAS chemicals. The

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

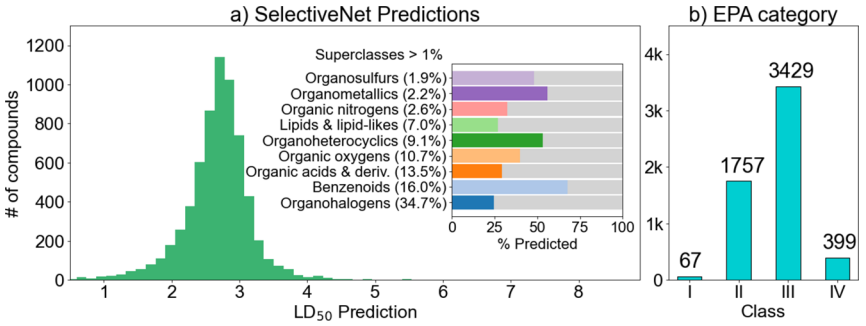


Figure 6. Consensus-style SelectiveNet predictions. (a) Histogram of SN-Mordred predictions on the EPA PFAS structure list. (Inset) Percent prediction/abstention of compounds grouped by the superclass. Gray color on each bar represents the abstained fraction. The superclass label provides the percent composition of the superclass within the entire EPA PFAS structure list. (b) SN-Mordred predictions categorized by the EPA toxicity class. The most predicted class, level III, has 3429 PFAS chemicals predicted to be in it.

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. However, the definition of such a domain-dependent threshold would require some human supervision. Furthermore, 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 the 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, except the 58 PFAS compounds found in both the LDToxDB-PFAS-like and LDToxDB-PFAS datasets were now removed from the source and target training samples and reserved as a true validation set.

The optimal SN-Mordred is arrived at by minimizing the risk by constraining the coverage.[32] 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.,[32] and the total empirical risk was calculated with respect to coverage (see Table S3 in the Supporting Information). A coverage threshold of 0.6 was found optimal and used to calibrate the abstention mechanism for use on LDToxDB-PFAS. The featurization of LDToxDB-PFAS by Mordred descriptors was successful for 7040 compounds. Consensus-type prediction is used for results where averages across each model (trained on one of five folds) are used. If ≥50% (3 or more) of models abstain, the compound is interpreted as abstained.

Figure 6a shows the distribution of selective-prediction outcomes. Because 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 server[38] (Figure 6a inset). 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 67 compounds in EPA class I, demonstrating considerable caution with respect to the most toxic EPA class.

The selective-prediction outcome for individual compounds allows us to examine how the model performs in different scenarios, particularly for the PFAS-58 compounds with the known values of oral rat LD_{50}, 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 40 of the 58 compounds. Of the 18 compounds where prediction was favorable, 12 were predicted within their actual EPA classes. For the compound Midaflur, a highly toxic EPA class I compound, the consensus was to take a cautious approach by not predicting.

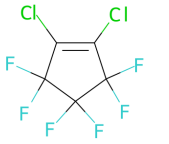
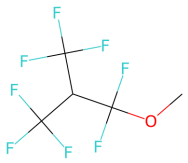
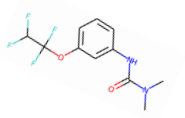
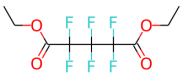
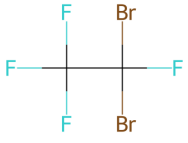
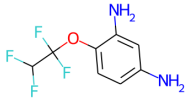
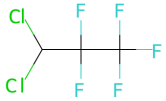
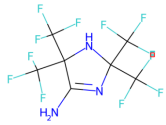
To underpin the results and decisions returned using AI, future efforts could include the development of deep-learning or QSAR models using molecular descriptors strongly correlated with acute toxicity[26] or by building local QSAR models from closely similar structures.[60] Such efforts would provide a physical and mechanical basis grounded in the molecular structure for interpreting toxicity estimates from AI. The derived relationships could further reduce the incidence of catastrophic decisions from AI predictions.

4. CONCLUSIONS

The 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 ∼8163 PFAS compounds whose toxicities are poorly understood. Results from our ML benchmark models for LDToxDB compounds demonstrate that the most competitive ML baseline is DNN-Mordred. We then evaluate transfer learning for DNN-Mordred on LDToxDB compounds as the source task and LDToxDB-PFAS-like compounds as the target task. To further identify the uncertainty of the target prediction task, two uncertainty estimation techniques, latent space and deep ensemble, are compared to illustrate the benefits of these analyses. Our findings on transfer-DNN-Mordred target prediction show that the deep ensemble provides a better uncertainty metric. However, the model is still overconfident ∼16% of the time when validated on LDToxDB-PFAS-like compounds. We then highlight crucial pitfalls of the conventional uncertainty estimation techniques

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

Table 5. Results for Chemicals Successfully Predicted, Successfully Abstained, and Poorly Predicted Using the SN-Mordred Model[a]

chemical Structure	chemical	transfer SN-Mordred inference	actual class
	cyclopentene, 1,2-dichloro-3,3,4,4,5,5-hexafluoro-	II	II
	1,1,1,3,3-pentafluoro-3-methoxy-2-(trifluoromethyl)propane	II	II
	tetrafluron	III	III
	diethyl perfluoroglutarate	III	IV
	1,1-dibromo-1,2,2,2-tetarafluoroethane	III	IV
	1,3-benzenediamine, 4-(1,1,2,2-tetrafluoroethoxy)	II	III
	enflurane	abstain	IV
	3,3 dichloro-1,1,1,2,2-pentafluoropropane	abstain	IV
	midaflur	abstain	I

[a]2D structures were generated with the RDKit package,[36] and chemical names were obtained from the EPA CompTox Chemicals dashboard.[59] Green shading corresponds to compounds where SN-Mordred predicted the compound into the correct EPA category. Blue shading corresponds to SN-Mordred abstention. Red corresponds to compounds that were predicted into the wrong EPA category by SN-Mordred.

when predicting unlabeled data that would require an expert-informed threshold. These pitfalls lead us to consider a selective prediction paradigm, where the uncertainty-informed decision to accept or abstain from a prediction is automated, deferring the uncertainty consideration to AI. Finally, we discuss the results of the SelectiveNet in predicting ~8163 compounds from the EPA's PFAS structure list, most with no known LD_{50} labels. The SelectiveNet implementation performed successful abstention on 40 out of 58 global validation groups of PFAS compounds (PFAS-58). Learning by abstention provides an automatic mechanism for converting uncertainty per prediction into model decisions. 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 and improve toxicology experiments.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01204.

Hyperparameter bounds used in Bayesian optimization of machine learning models, chemical descriptors employed for atoms and bonds in GCN, comparisons of random versus stratified sampling methods, tuning process for determining optimal transfer-DNN-Mordred, deep ensemble convergence curve, and SN-Mordred risk-coverage curve (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Eugene Yan − *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-7112-7397; Email: eyan@anl.gov.

Authors

Jeremy Feinstein − *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Ganesh Sivaraman − *Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0001-9056-9855

Kurt Picel − *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Brian Peters − *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Álvaro Vázquez-Mayagoitia − *Computational Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-1415-6300

Arvind Ramanathan − *Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Margaret MacDonell − *Environmental Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Ian Foster − *Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01204

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

The datasets used in this study including LDToxDB, LDToxDB-PFAS-like, and LDToxDB-PFAS can be accessed at https://github.com/AI4PFAS/AI4PFAS. The Python processing codes are also available at our github site for parsing the data and constructing the models that are presented in Section 2.

■ 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*. https://www.epa.gov/pfas/basic-information-pfas (accessed Dec 1, 2020).
(2) Interstate Technology and 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 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 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. *Comput. Toxicol.* **2019**, *12*, No. 100096.
(5) Patlewicz, G.; Richard, A. M.; Williams, A. J.; Grulke, C. M.; Sams, R.; Lambert, J.; Noyes, P. D.; DeVito, M. J.; Hines, R. N.; Strynar, M.; Guiseppi-Elie, A.; Thomas, R. S. A Chemical Category-based Prioritization Approach for Selecting 75 Per- and Polyfluoroalkyl Substances (PFAS) for Tiered Toxicity and Toxicokinetic Testing. *Environ. Health Perspect.* **2019**, *127*, No. 014501.
(6) Hartung, T. Toxicology for the Twenty-First Century. *Nature* **2009**, *460*, 208−212.
(7) Richarz, A. N. CHAPTER 1. Big Data in Predictive Toxicology: Challenges, Opportunities and Perspectives. *Chapter 1 in Big Data in Predictive Toxicology*; Royal Society of Chemistry, 2019; 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. *Toxicol. Sci.* **2018**, *165*, 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. *Chem. Res. Toxicol.* **2019**, *32*, 536−547.
(10) Luechtefeld, T.; Rowlands, C.; Hartung, T. Big-Data and Machine Learning to Revamp Computational Toxicology and Its Use in Risk Assessment. *Toxicol. Res.* **2018**, *7*, 732−744.
(11) Sze, V.; Chen, Y. H.; Einer, 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; pp 1−8.
(12) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 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. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(15) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. *Multi-Task Neural Networks for QSAR Predictions*, 2014, ArXiv preprint, arXiv:1406.1231.
(16) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.
(17) Kleinstreuer, N. C.; Tong, W.; Tetko, I. V. Computational Toxicology. *Chem. Res. Toxicol.* **2020**, *33*, 687−688.
(18) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vasquez-Mayagoitia, A.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. J. A Machine Learning Workflow for Molecular Analysis: Application to Melting Points. *Machine Learning: Science and Technology* **2020**, *1*, No. 025015.
(19) Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf. Model.* **2017**, *57*, 2672−2685.
(20) Wu, K.; Wei, G. W. Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 520−531.
(21) Sosnin, S.; Karlov, D.; Tetko, I. V.; Fedorov, M. V. Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **2018**, *59*, 1062−1072.
(22) Jiang, J.; Wang, R.; Wei, G. W. GGL-Tox: Geometric Graph Learning for Toxicity Prediction. *J. Chem. Inf. Model.* **2021**, *61*, 1691−1700.
(23) Abbasi, K.; Poso, A.; Ghasemi, J.; Amanlou, M.; Masoudi-Nejad, A. Deep Transferable Compound Representation across Domains and Tasks for Low Data Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59*, 4528−4539.
(24) Jain, S.; Siramshetty, V. B.; Alves, V. M.; Muratov, E. N.; Kleinstreuer, N.; Tropsha, A.; Nicklaus, M. C.; Simeonov, A.; Zakharov, A. V. Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. *J. Chem. Inf. Model.* **2021**, *61*, 653−663.
(25) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of Machine Learning Methods in Drug Toxicity Prediction. *Curr. Top. Med. Chem.* **2018**, *18*, 987−997.
(26) Bhhatarai, B.; Gramatica, P. Oral LD50 Toxicity Modeling and Prediction of Per- and Polyfluorinated Chemicals on Rat and Mouse. *Mol. Diversity* **2011**, *15*, 467−476.
(27) Cheng, W.; Carla, N. A. Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances from the OECD List. *Environ. Sci. Technol.* **2019**, *53*, 13970−13980.
(28) Hoover, G.; Kar, S.; Guffey, S.; Leszcynski, J.; Sepúlveda, M. S. In Vitro and In Silico Modeling of Perfluoroalkyl Substances Mixture Toxicity in an Amphibian Fibroblast Cell Line. *Chemosphere* **2019**, *233*, 25−33.
(29) Torrey, L.; Shavlik, J. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI global, 2010; pp 242−264.
(30) Kompa, B.; Snoek, J.; Beam, A. L. Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning. *npj Digit. Med.* **2021**, *4*, 1−6.
(31) Cortes, C.; DeSalvo, G.; Mohri, M. Learning with Rejection. In *International Conference on Algorithmic Learning Theory*; Springer: Cham, 2016; pp 67−82.
(32) 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.
(33) U.S. Environmental Protection Agency. *Toxicity Estimation Software Tool (TEST)*. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test (accessed Dec 1, 2020).
(34) 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. *Comput. Toxicol.* **2018**, *8*, 21−24.
(35) Gadaleta, D.; Vukovic, K.; Toma, C.; Lavado, G. L.; Karmaus, A. L.; Mansouri, K.; Kleinstreuer, N.; Benfenati, E.; Roncaglioni, A.

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803

SAR and QSAR Modeling of a Large Collection of LD50 Rat Acute Oral Toxicity Data. *Aust. J. Chem.* **2019**, *11*, 58.
(36) Landrum, G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*. https://www.rdkit.org/ (accessed Dec 1, 2020).
(37) McInnes, L.; Healy, J.; Melville, J. *Umap: Uniform Manifold Approximation and Projection for Dimension Reduction*; 2018, arXiv preprint, arXiv:1802.03426.
(38) Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive Computable Taxonomy. *J. Cheminf.* **2016**, *8*, 61.
(39) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Aust. J. Chem.* **2018**, *10*, 4.
(40) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(41) 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. Advances in Neural Information Processing Systems 28*, Montreal, Canada, 2015; pp. 2215−2223.
(42) 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; Vol. *70*, pp 1263−1272.
(43) Paatero, P.; Tapper, U. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* **1994**, *5*, 111−126.
(44) Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401*, 788−791.
(45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(46) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(47) Rasmussen, C. E. Gaussian Processes in Machine Learning. In *Summer School on Machine Learning*; Springer: Berlin, Heidelberg, 2003; pp. 63−71.
(48) de Matthews, A. G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrá, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian Process Library Using TensorFlow. *J. Mach. Learn. Res.* **2017**, *18*, 1−6.
(49) Kingma, D. P.; Ba, J. *Adam: A Method for Stochastic Optimization*, 2014, arXiv preprint, arXiv:1412.6980.
(50) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2015**, *104*, 148−175.
(51) González, J.; Dai, Z. *GPyOpt: A Bayesian Optimization Framework in Python*. 2016. http://github.com/SheffieldML/GPyOpt (accessed April 8, 2021).
(52) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. *Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles*. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 6405−6416. NIPS'17; Curran Associates Inc.: Long Beach, California, USA, 2017.
(53) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019**, *10*, 7913−7922.
(54) El-Yaniv, R.; Wiener, Y. On the Foundations of Noise-Free Selective Classification. *J. Mach. Learn. Res.* **2010**, *11*, 1605−1641.
(55) 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. *Toxicol. Sci.* **2018**, *164*, 512−526.
(56) 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. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.
(57) Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; Snoek, J. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift*. 2019, arXiv preprint, arXiv:1906.02530.
(58) Caldeira, J.; Nord, B. Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. *Machine Learning: Science and Technology* **2020**, *2*, No. 015002.
(59) U.S. Environmental Protection Agency. *CompTox Chemicals Dashboard*. https://comptox.epa.gov/dashboard/ (accessed Jan 18, 2020).
(60) Vukovic, K.; Gadaleta, D.; Benfenati, E. Methodology of aiQSAR: A Group-Specific Approach to QSAR Modelling. *Aust. J. Chem.* **2019**, *11*, 27.

https://doi.org/10.1021/acs.jcim.1c01204
J. Chem. Inf. Model. 2021, 61, 5793−5803