Predicting Environmental Chemical Toxicity using a New Hybrid Deep Machine Learning Method.

Sarita Limbu[1], Cyril Zakka[2] & Sivanesan Dakshanamurthy[1,*]

[1]Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC USA.
[2]Faculty of Medicine, American University of Beirut Medical Center, Lebanon.

*Corresponding author:
Dr. Sivanesan Dakshanamurthy,
Professor, & Director of Computational Chemistry Shared Resources,
3970 Reservoir Road,
Lombardi Comprehensive Cancer Center,
Georgetown University Medical center,
Washington DC 20057 USA
Email: sd233@georgetown.edu

ABSTRACT

Humans are exposed to thousands of potentially toxic chemicals including environmental chemicals. Approximately 300,000 such chemicals are currently in use, unfortunately little is known about their potential toxicity. Determining human toxicity potential of chemicals remains a challenge due to a substantial resource required to assess a chemical in-vivo, and only a few thousand single chemicals in commercial use has been evaluated. In this study, to predict the environmental chemical toxicity, we developed a new hybrid neural network (HNN) deep learning model consisting of a Convolutional Neural Network (CNN) and multilayer perceptron (MLP) type feed forward neural network (FFNN). We developed several deep learning binary and multiclass categorical toxicity models on the thousands of datasets obtained from US NLM ChemIDplus, Toxin Target database (T3DB), and Environment Protection Agency (EPA). The performance of our HNN deep learning models was compared with models developed using other machine learning methods including Random Forest (RF), Bootstrap Aggregation (Bagging), and Adaptive Boosting (AdaBoost). We analyzed the machine learning model performance dependency on the varying features and dataset size. Compared to other methods, our HNN deep learning model trained on 22,000 chemicals with known acute toxicity (LD50), maintained its predictive ability even after reducing the descriptor size from 318 to 51. The average accuracy was 84.96% and 84.11% and the average AUC were 0.897 and 0.887 for HNN models based on 318 and 51 descriptors respectively. To our knowledge this study is the first to report a large-scale prediction of environmental chemical toxicity. Our hybrid HNN deep learning models can be used in predicting chemical toxicity with high accuracy for a diverse set of chemicals, has a wide applicability in the prediction of chemical toxicity and its mixtures, and greatly minimize the need for costly and unethical animal-based toxicity predictions.

Key Words. Environmental chemical toxicity; Chemical toxicity; Machine learning method; Deep learning; Binary toxicity; Multiclass categorical toxicity.

INTRODUCTION

Humans are exposed to thousands of potentially toxic chemicals including environmental chemicals such as industrial wastes, food products, solvents, air pollutants, fertilizers, pesticides, insecticides, carcinogens, drugs, metals/metalloids, and other industrial chemicals. Factors affecting the degree of toxicity are route of exposure (oral, dermal, inhalation and injection), duration of exposure, dose of chemical, age, gender and health condition. The conventional *in vivo* and *in vitro* tests for finding the toxicity of large number of chemicals are expensive in terms of both time and money. In this context, *in silico* toxicity prediction for chemical substances are gaining popularity as a quicker and an inexpensive alternative that also eliminates the need for further animal testing which is controversial due to ethical concerns.

Computational models are developed upon using different machine learning algorithms for various toxicity endpoints[1,2]. Toxicity predictions could be quantitative[3], predicting the quantity of chemical required for the adverse outcome or qualitative, predicting binary endpoint (such as whether the chemical is toxic or nontoxic) or ordinary, predicting the categorical endpoint (such as high, moderate, and low). To predict oral acute toxicity (LD50), regression models are developed with a confined applicability domain to improve prediction accuracy[1,4]. Chavan et al. used k-Nearest Neighbor (KNN) method to predict the acute toxicity of chemicals with 79.17% accuracy[5]. Cherkasov et al. predicted the antibacterial activity of chemicals with 93% accuracy using the Artificial Neural Network (ANN) method for Quantitative Structure-Activity Relationship (QSAR) model[6]. Zhang et al. reported 70% accuracy on binary models for predicting chemical carcinogenicity using ensemble of Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) machine learning methods[7]. Tanabe et al. reported an average accuracy of 70% using SVM modeling and improved the accuracy to 80% by developing models on the chemical subgroups based on their structure[8]. Li et al. developed multiclassification models using SVM, RF, decision tree, k Nearest Neighbor (kNN), and Naïve Bayes for predicting the categorical toxicity of chemicals with overall accuracy ranging from 42% to 83% and from 25.1% to 89.9% for external validation set I and II respectively[9].

Deep neural network (DNN) architecture is the ANN with more than one hidden layer, has been successfully applied in many areas including speech recognition[10] and image recognition[11]. Convolutional Neural Network (CNN) is a class of deep neural network more suitable for image processing tasks such as visual data recognition[12]. The strength of DNN has been demonstrated in toxicity prediction by Mayr et al, the winning team of the Tox21 challenge[13] by achieving AUCs between 0.79 to 0.94 for different cell-based assay training data set. Dahl et al, the winning team of the QSAR competition sponsored by Merck also used DNN for compound activity prediction[14]. Beside the competitions, deep learning has also been used in the image-based toxicity prediction[15,16], drug-induced liver injury prediction[17]. Deep learning based multiclass model developed by Xu et al. predicted with accuracy of 95.5% and 96.3% on test sets I and II respectively[2]. DNN performs better with larger datasets[18]. The more dimensions or diversification in the model is desired, larger the dataset is required for training otherwise results in overfitting. The success of deep learning proves that it has a great potential in the field of toxicity prediction if significantly large dataset can be obtained to train the model. Thus, need of a highly reliable model to predict toxicity for a diverse chemical can be contented with the application of deep learning methods.

To harness the potential of deep learning method in the field of toxicity prediction, in this study, we have developed various DNN based hybrid neural network (HNN) models consisting of a CNN and a feed-forward neural network (FFNN) for toxicity prediction. We constructed hybrid HNN models with a large chemical domain coverage based on the input training data set obtained from ChemIDplus, Toxin and Toxin Target Database (T3DB), and Environmental Protection Agency (EPA). We have developed models on the dataset of various sizes and tested the effect of increasing the dataset size with varying descriptors. The performance of the deep learning HNN model was compared with other machine learning algorithms including Random Forest (RF), Bootstrap Aggregation (Bagging), and Adaptive Boosting (AdaBoost). The effects of dataset size and class imbalance on the prediction capability of the deep learning model were studied. The HNN model developed on the ChemIDplus data presented the best predictive performance in comparison to the models developed on a smaller dataset. The HNN models can be used in predicting chemical toxicity with high accuracy for a diverse set of chemicals. This will greatly minimize the need for costly and unethical animal-based toxicity predictions.

MATERIALS AND METHODS

Training and Test Data set and Feature attributes calculations. HNN models were developed on various types of data set collected from different number of samples, attributes, and sources. We obtained datasets from the following sources: i) ChemIDplus, ii) Toxin and Toxin Target database (T3DB), iii) Environmental Protection Agency (EPA) and iv) Tox21 Challenge. Separate models were developed on these datasets and their prediction capability was assessed. The T3DB, and National Toxicology Program (NTP) data sets were used as external validation set to test the predictive ability of the models.

1) ChemIDplus Data Set

CAS registry numbers (CASRN) were obtained from the ChemIDplus database at ftp://ftp.nlm.nih.gov/nlmdata/.chemidlease/. These CASRN were used to retrieve the LD50 data from the ChemIDplus available online using openStream() method of URL class in the Java's java.net package. Total of 386,620 chemicals with CASRN were retrieved. Only 92,322 chemicals annotated with LD50 values, SMILES and other physico-chemical properties were used. The structconvert utility in Schrodinger software was used to convert the SMILES of the 92,322 chemicals to 2D structures in .sdf format. We filtered the chemicals further by removing metal containing compounds and obtained a final set of 59,373 chemicals. 3D minimization application in Schrodinger's Canvas module was used to convert the 2D structures to .sdf file containing 3D structures. We calculated 51 physicochemical property descriptors for 59,373 chemicals using QikProp application in the Schrodinger suite.

Data set 1a) IP/IV/Subcutaneous/Oral- All animals: Out of 59,373 chemicals, 55,856 chemicals were annotated with LD50 values obtained for all animals treated via IP, IV, subcutaneous and oral route of exposure. We annotated 26,923 as nontoxic and 28,933 as toxic chemicals based on the toxic-nontoxic cutoff set at LD50 value of 500 mg/kg. Randomly selected 5,000 chemicals were used as a test set while the remaining chemicals were used as a training set during each prediction.

Data set 1b) Oral- Rat/Mouse: Out of 59,373 chemicals, 22,808 chemicals with LD50 values were obtained by filtering rat and mouse species via oral route of exposure. We calculated 31 ADMET (absorption, distribution, metabolism, excretion and toxicity) properties for the 22,792 chemicals using ADMETlab platform[19]. Additionally, 12 physicochemical properties, 224 topological properties and 155 MACCS fingerprints were calculated using the Canvas application of Schrodinger suite. Total of 318 descriptors and 155 fingerprints were calculated for 22,792 chemicals. We then annotated 16,311 chemicals as non-toxic and 6,481 as toxic. Randomly selected 4,500 chemicals were used as the test set while the remaining chemicals as the training set for each prediction. We used the Toxin and Toxin Target database (T3DB) data set of 636 rat and mouse oral dataset as an external test set to validate the model based on the ChemIDplus dataset.

2) National Toxicology Program Data set

We used the predictive toxicity models project data set provided by the National Toxicology Program[20] as external validation set. This dataset consisted of LD50 values for rat acute oral toxicity and were classified as toxic if LD50 values were >500 mg/kg. We calculated 51 property descriptors using QikProp.

NTP data as external validation set for ChemIDplus Oral Rat/Mouse data: The duplicate chemicals in NTP data that also existed in the Oral dataset from ChemIDplus (described in section 1b) were removed and the final list of 1703 chemicals was obtained as the external validation set. The 1703 chemicals were used as external validation set for the models built on Oral ChemIDplus training set with 51 QikProp descriptors.

NTP as external validation set for ChemIDplus IP/IV/Sub/Oral data: The duplicate chemicals in NTP data that also existed in the IP/IV/Sub/Oral dataset from ChemIDplus (described in section 1a) were removed and the final list of 1648 chemicals was obtained as the external validation set. The 1648 chemicals were used as the external validation set for the models built on IP/IV/Sub/Oral ChemIDplus training set with 51 QikProp descriptors.

3) Toxin and Toxin Target database (T3DB) data set.
The toxins data with 3,673 chemicals was curated from the T3DB database[21]. 778 chemicals matched in the 62 descriptors file from our local dataset (**Supplementary Table S1**).

3a) Oral data: 687 chemicals with 62 descriptors were separated as Oral data.
3b) IP/IV/Sub/Oral data: 752 chemicals were separated as IP/IV/Subcutaneous/Oral data.

Binary Classification Models: Data set were annotated with toxicity classification at various LD50 cutoffs (250 mg/kg, 500 mg/kg, 750 mg/kg, and 1000 mg/kg) to determine the chemical toxicity and compare the impact of various cutoffs and resulting class imbalance on the predictive performance of the models.

Multiclass Classification Models: 687 Oral data in 3a were also classified into 4 categories: a) LD50 < 50 mg/kg, b) 50 mg/kg ≤ LD50 < 500 mg/kg, c) 500 ≤ LD50 < 1000, and d) LD50 ≥ 100 to determine the categorical toxicity of the chemicals.

T3DB data as external validation set for ChemIDplus Oral Rat/Mouse data: The duplicate chemicals in T3DB data (3a) that also existed in the Oral dataset from ChemIDplus (described in section 1b) were removed and the final list of 636 chemicals was obtained as the external validation set. Data set were annotated with toxicity classification at LD50 cutoff set at 500 mg/kg. Total of 318 descriptors and 155 fingerprints (as described for dataset in section 1b) were calculated and the data were used as external validation set for the models built on Oral ChemIDplus training set described in section 1b.

4) Animal Toxicity Data from EPA.
The animal toxicity data of 980 chemicals was downloaded from EPA's website. The lowest effect level (lel) dose of the chemicals was considered for determining the toxicity. Various threshold values for toxicity considered were 250 mg/kg, 500 mg/kg, 750 mg/kg, and 1000 mg/kg.

5) Tox21 Challenge Data set.
Tox21 Challenge dataset of 12 different assays and their corresponding 801 descriptors were obtained from the online resource (http://bioinf.jku.at/research/DeepTox/tox21.html) of Mayr et al., the winning team of the Tox21 challenge[13].

SMILES Preprocessing.

SMILES were used as one of the key chemical attributes and is used in our hybrid deep learning model. Raw texts cannot be directly used as an input for the deep learning models but should be encoded as numbers. The entire list of SMILES strings was first fit onto the tokenizer to create a dictionary of the set of all the possible characters in the SMILES string and their corresponding index. We assumed that a dictionary D is created where, D = {'C': 1, '=': 2, '(': 3, ')': 4, '#': 5, 'N': 6, … , ' ': M }. This results in every character in the SMILES string being assigned a unique integer value which is the index of the character in the dictionary. The SMILES entry for every chemical is then converted to one-hot encoded 2-D matrix. As an example, acrylonitrile-d3 with SMILES string C=CC#N is one-hot encoded as:

$$\begin{bmatrix} C \\ = \\ C \\ C \\ \# \\ N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & & 0 \\ & & & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

A 3-D matrix of size K x L x M is obtained eventually where K is the number of chemicals, L is the maximum length of the SMILES string, and M is the number of all possible characters in the SMILES string from K chemicals (number of entries in the created dictionary). One-hot encoding means converting the integer value of each character in the SMILES to its equivalent binary vector of length M.

The Hybrid Neural Network Model

The Hybrid Neural Network (HNN) model is developed in python using the Keras API with Tensorflow in the backend. The model consists of a Convolutional Neural Network (CNN) for deep learning on the basis of structure attribute (SMILES) and multilayer perceptron (MLP) type feed forward neural network (FFNN) for learning on the basis of remaining attributes of the chemicals (**Figure 1**). Basic layers of a CNN include are convolutional layer, non-linearity layer, pooling or sub sampling layer and fully-connected layer. Convolution layer learns and extracts features from the input array computing dot product between the weights and a small region of the input matrix. The weights are represented by a matrix, called kernel or filter, smaller in size than the input matrix. In order to represent the real-world data, non-linearity layer applies one of the various available activation functions such as Rectified Linear Unit (ReLU), sigmoid and tanh to introduce non-linearity in the model. Activation function ReLU represented mathematically as max(0, x) is used in the model that replaces all the negative values with zeros. The derivative of ReLU is always 1 for positive input that counteracts the vanishing gradient problem during the backpropagation. Pooling or sub sampling reduces the dimension of the data while retaining the important information in the data. Max pooling is used in the model for sub sampling. Fully connected layer does the final classification implementing softmax activation function in case of multiclass classification and sigmoid activation function in case of binary classification. In fully connected layer, every neuron in the current layer is connected to every neuron in the previous layer. A multilayer perceptron (MLP) type feed forward neural network (FFNN) contains one or more hidden layers. The 3-D array of one-hot encoded SMILES strings was the input for the CNN and chemical descriptors was the input for the FFNN. The output of the pooling layer of the CNN was merged with the final fully connected layer of FFNN to perform the classification task.

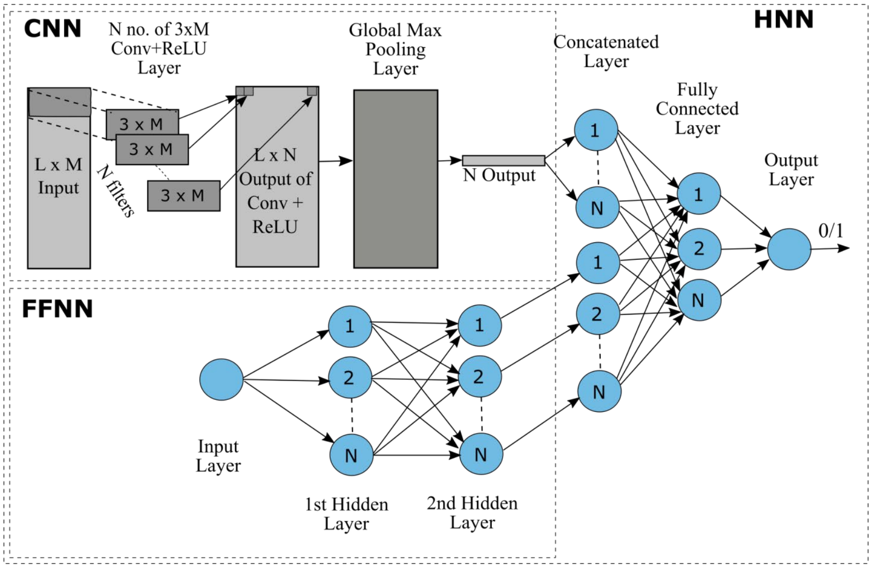


Figure 1: Schematic diagram of the Hybrid Neural Network (HNN) consisting of Convolutional Neural Network (CNN) and Feed Forward Neural Network (FFNN). L, length of the SMILES string; M, ; N, number of filters (possibly different at each layer).

Parameter Tuning.

We implemented Hyperparameter tuning to improve the performance of the model. Hyperopt package from python that uses Tree-structured Parzen Estimator (TPE) method was used for hyperparameter optimization. This approach requires defining the objective function that fmin() function minimizes, the parameter space over which the search is performed and the number of experiments to run. The Area Under the receiver operating characteristic Curve (AUC) was the metric used for evaluating the performance of each model. Class imbalance is a common problem while modeling toxicity data and AUC is a better metric to be optimized than accuracy.

Other Machine Learning Algorithms

To test performance of the HNN model, and to create an ensemble model, we developed number of other machine learning models based on the other machine learning algorithms such as Random Forest, Bootstrap Aggregation (Bagging) using Bagged Decision Tree, and Adaptive Boosting (AdaBoost).

Ensemble Model

To optimize performance of the model, ensemble of model predictions is considered a good option. Random Forest, Bagging method and Adaboost were used for making ensemble predictions to boost the overall performance of the HNN. The ensemble method derived by Mayr et. al.[13] for calculating ensemble probabilities was used in making the final prediction from the ensemble model [**Supplementary Equation S1**].

Model Performance Evaluation

All the results presented are the average of 10 simulation run repeats for ChemIDplus data, 30 simulation repeats for T3DB data, and Tox21 Challenge data. 20% of the data set were separated randomly each time as the test set, remaining data were used as the training set which is similar to 10-fold cross validation except that the test sets were randomly selected each time. The performance of each model was evaluated based on the accuracy, and the Area Under the receiver operating characteristic Curve (AUC). The AUC gives the probability of positive outcome being ranked before the negative outcome and is a better metric for evaluating a binary classifier than accuracy[22,23]. The models were also evaluated for the sensitivity, specificity and precision [**Supplementary Equation S2**].

The overall workflow presented in this paper is shown in **Supplementary Figure S1**. Data from various databases were downloaded, processed to select them based on their route of exposure and determine whether they are toxic or non toxic based on the experimental data. Their molecular descriptors along with their SMILES strings are computed and the hybrid neural network models are developed to make the predictions. Other machine learning algorithms Random Forest, Bagging method and Adaboost were used for comparisons and ensemble predictions to improve overall performance of the HNN.

RESULTS AND DISCUSSION

We developed a new hybrid HNN model by combining the CNN and FFNN. To examine and compare performance of the HNN model, several other machine learning models were developed. To test the dependency and performance variations of the neural network models, we also used various training data set sizes, descriptors and fingerprints. We first developed various machine learning based binary classification models to predict chemical toxicity. To predict the degree of toxicity (categorical) of these chemicals, we also developed multiclass classification models and classified the chemicals into different categories based on their toxicity level and the results are presented in later section.

Toxicity Prediction using Binary Classification.

ChemIDplus Data Predictive Toxicity Analysis (Oral Toxicity: rat and mouse). Our goal is to build highly generalized model with high predicting capability. Larger data size is proven to be helpful in improving accuracy in many scenarios of machine learning such as overfitting models with high variance[24]. We collected the largest chemicals dataset associated with experimentally determined LD50 from the ChemIDplus online server. The 318 descriptors and 155 MACCS fingerprints were calculated for 22,792 chemicals and the predictive models were developed with a toxicity threshold set at LD50 value 500 mg/kg. For the HNN, the FFNN was developed based on the 318 descriptors and two CNNs were developed based on the 155 fingerprints and SMILES string separately. Independently, the RF, Bag, and Ada models were developed based on the 318 descriptors. The models based on the HNN, RF and Bag exhibit similar accuracy and AUC whereas the HNN model sensitivity to correctly identify the positives i.e. the toxic chemicals, was significantly higher compare to other models (**Figure 2**). The ensemble model improved the performance of the model by providing high accuracy of 86.50% and AUC of 91.65%.

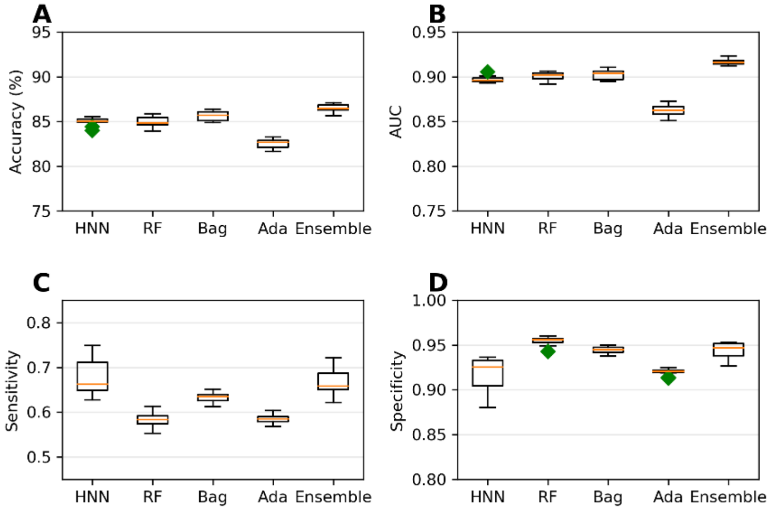


Figure 2. A) Accuracy percentage, B) AUC, C) Sensitivity, and D) Specificity for the ChemIDplus Oral data as given by HNN, RF, Bagging, AdaBoost and the Ensemble methods with additional descriptors from ADMETlab and Canvas.

The prediction accuracy achieved with a qualitative binary toxicity prediction model by Sharma et al[25]. was 93%. In their model, the training set consisted of the chemicals obtained from the T3DB database as positive dataset and human metabolites as negative dataset. Such high accuracy achieved is likely because of different type of compounds in the toxic and non-toxic group since compounds in T3DB database were compositionally distinct from the metabolites as revealed by their compositional analysis. The toxic and non-toxic compounds in our dataset were obtained from the same source and represent the real-life chemicals. Thus, our model is more generalized with diverse set of chemicals in both the toxic and non-toxic groups.

To investigate the effect of descriptors and SMILES on the performance, models were developed with 51 Schrodinger QikProp descriptors (instead of 318 descriptors) for RF, Bagging, and Ada and HNN in addition to the SMILES for the CNN (but no fingerprints). With many descriptors missing such as ADMET properties, the accuracy and the AUC of RF, Bagging, and Ada are reduced significantly (**Table 1**). This means that the absence of 277 descriptors, play a significant role in the prediction. But the performance of the HNN model was not affected significantly. This may be due to the HNN is using additional SMILES as input feature that enriched the model to learn based on the structure of the compound and enabled the model to compensate for the missing descriptors in the toxicity prediction.

	No of Desc	HNN	RF	Bag	Ada
Accuracy %	318	84.96	84.94	85.62	82.53
	51	84.11	82.07	82.05	76.24
AUC	318	0.897	0.901	0.902	0.862
	51	0.887	0.865	0.861	0.765

Table 1. Accuracy % and AUC for the ChemIDplus oral data with 51 descriptors and 318 descriptors.

Validation of the ChemIDplus Oral data models.

To ensure the prediction accuracy of the models developed on the ChemIDplus dataset, performance of the models was evaluated by making predictions for external data that was not used to develop the models. The models were tested on the T3DB and NTP dataset as the external validation dataset.

A) T3DB data as external validation set

The common chemicals present in both the T3DB and the ChemIDplus datasets were removed from the ChemIDplus dataset but not from the validation set so that we have significant number of chemicals to perform the model validation test. The training set included 22,438 rat and mouse oral toxicity data from ChemIDplus whereas the test set included 636 rat and mouse oral toxicity data from T3DB were used for the validation of the model. The models predicted with an average accuracy of 76.94%, 75.69%, 72.76%, 74.37%, and 75.28% and the average AUC of 0.833, 0.815, 0.794, 0.808, and 0.842 for the HNN, RF, Bag, Ada, and Ensemble respectively (**Supplementary Figure S3**). These results showed that the models can make predictions for the external dataset with high accuracy and AUC which proves their predictive ability. The HNN model displayed the best average performance here in terms of accuracy, AUC, and sensitivity. The sensitivity of the HNN model was significantly higher than other models which proves the better ability of the model to correctly identify the toxic chemicals. The ensemble method predicted the toxicity of external validation set with an accuracy of 75.28% and AUC of 0.842.

B) NTP data as external validation set

We next used the NTP dataset as the external validation dataset. Models were developed with 51 QikProp descriptors. For the external validation dataset consisting of rat acute oral toxicity, the models predicted with an average accuracy percentage of 73.29, 75.23, 75.39, 69.82 and 75.44 for the HNN, RF, Bagging, AdaBoost and Ensemble respectively (**Supplementary Figure S4**). The average AUCs of the models were 0.766, 0.783, 0.779, 0.705 and 0.789. The training set is a mix of the rat and mice data but the validation set comprises of the rat data only. Thus, the training set is not very specific with rat

LD50 values which could be the reason for decrease in the performance of these models. Taken together, the two external validation sets (T3DB and NTP) results demonstrate the robust toxicity predicting capability of these models.

ChemIDplus Data Predictive Toxicity Analysis (IP/IV/Subcutaneous/Oral toxicity: all animals/birds). Extending the domain of the data enhances the applicability of the model. Thus, we sought to apply our models to a more generalized larger data set. The LD50 data for all animals/birds obtained via IP/IV/Subcutaneous/Oral route of exposure were selected from the ChemIDplus data. Molecular descriptors were calculated using Schrodinger QikProp tool for 55,856 chemicals.

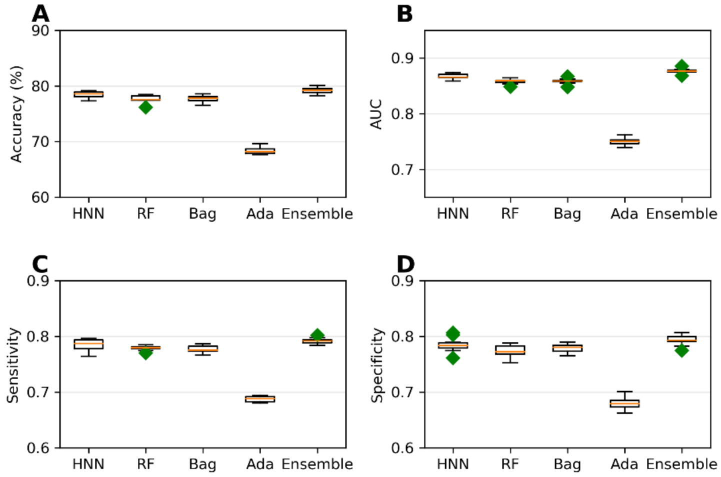


Figure 3: A) Accuracy percentage, B) AUC, C) Sensitivity, and D) Specificity for the ChemIDplus IP/IV/Sub/Oral data as given by HNN, RF, Bagging, AdaBoost and the Ensemble methods

The predictive performance of the model developed on this more generalized large dataset when compared to the models developed on oral data for rat and mouse only, the accuracy decreased from 84.96% (**Figure 2**) to 78.43% (**Figure 3**) for HNN but the decrease in AUC was very small (from 0.887 to 0.866) which means the model still has the similar ability of ranking the toxic chemicals higher than the nontoxic chemicals. The new model's sensitivity increased from 0.677 to 0.784 whereas specificity decreased from 0.917 to 0.784.

Validation of the ChemIDPlus IP/IV/Sub/Oral data models.

The prediction capability of the models was evaluated on the NTP data as external validation dataset. For the validation set, the models predicted with AUC 0.72 (**Supplementary Figure S6**). The reason for the poor performance of the models with external validation set compare to test set, because the training and the test set comprises of all animals, birds data whereas the validation data comprises of rat only.

T3DB dataset predictive toxicity analysis via Oral route of exposure. The T3DB toxin data were processed and annotated with the LD50 values. The data set obtained by chemical administration via only oral route of exposure were separated. This dataset includes 687 chemicals with 62 descriptors were computed (data 3a of Data Section in Materials and Methods). Models were developed on these toxins data using 250 mg/kg, 500 mg/kg, 750 mg/kg and 1000 mg/kg LD50 values as four different cutoffs. The average accuracy is highest when the LD50 threshold is set at 250 mg/kg whereas, the average

AUC was lowest in this category (**Figure 4,** and **5**). This is due to higher imbalance in data when the LD50 threshold is 250 mg/kg (ratio of 3.37:1 for NonToxic:Toxic) in comparison to those obtained for the LD50 threshold value higher than 250 mg/kg (ratio of 1.41:1 to 2:05 for NonToxic:Toxic). The model lost its ability to accurately rank toxic substances over the nontoxic which resulted in lower AUC at 250 mg/kg threshold. Imbalanced dataset comprises of majority of samples being classified into one class while very few samples are classified into the other class. The model trained on such dataset predicts the test samples as belonging to the majority class more often ignoring the minority class. If the test set also comprises of imbalanced data, this results in an apparently very high accuracy. However, the ability of the model to rank the samples to be predicted as 1s ie. toxic before the samples to be predicted as 0s ie. non-toxic decreases which is demonstrated by the lower AUC (**Figure 9**, LD50 threshold 250 mg/kg).

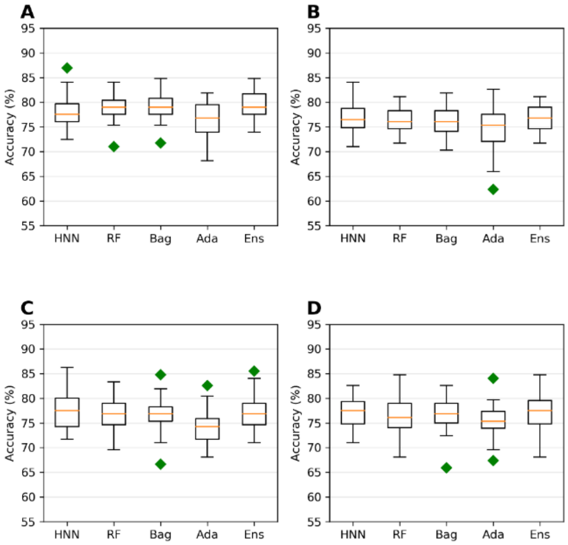


Figure 4: Accuracy percentage for Toxins LD50 data obtained via Oral route of exposure with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

T3DB dataset predictive toxicity analysis via IP/IV/Subcutaneous/Oral route of exposure.

To study how the data obtained for chemicals administered via various route, affect the toxicity prediction, the toxins data from T3DB were separated to include Intraperitoneal (IP), Intravenous (IV), Subcutaneous and Oral route of exposure (IPIVSubOral) data (**Supplementary Table S2**). The models were developed on the data for the combined route which included 752 chemicals with 62 descriptors (data 3b of Data Section in Materials and Methods). The change in toxicity determination method based on the route of chemical administration changed the overall toxic:nontoxic ratio of data for the four LD50 threshold values. Here, the category with 1000 mg/kg threshold value of LD50 has the most imbalanced data with ratio of 3.2:1 and the accuracy is highest with 78.96%, 80.38%, 79.73%, 74.71% and 80.64% for HNN, RF, Bagging, Ada and Ensemble methods respectively in this category but the AUCs are lower for the categories with 1000 mg/kg and 250 mg/kg thresholds (**Supplementary Figure S7 and S8**). The AUCs are higher with 0.853, 0.855, 0.856, 0.799 and 0.866 for HNN, RF, Bagging, Ada and Ensemble at 500 mg/kg threshold and 0.861, 0.859, 0.855, 0.796 and 0.869 for HNN, RF, Bagging, Ada and Ensemble at 750 mg/kg threshold when the data are more balanced.

There is no significant change in the accuracy or the AUC of the models developed with samples from IP, IV and subcutaneous route of chemical administration when compared to the performance of the models developed on the data from oral route of exposure. The accuracy and AUC were changed in the ratio of toxic to non-toxic samples in both the cases (Oral and IPIVSubOral). Adding additional data (non-oral) but from different route of chemical administration did not increase the accuracy or AUC of the model (**Supplementary Figure 8**). This is because additional non-oral route of administration data is not contributing any robust toxicity information to the oral-route training data set. Further, the performance of a model depends on the quality of data and the class proportion (i.e. ratio of the two classes: toxic and non-toxic).

Combined T3DB and EPA Dataset

To investigate if achieving larger dataset by combining data from different sources affects the model's predictive ability, the toxins data from T3DB with LD50 values and animal toxicity data from EPA with lel dose values were combined to form a single dataset (**Supplementary Table S4**). Both the datasets were not separated based on the route of chemical administration. The 778 chemicals from T3DB were combined with 427 chemicals from EPA to obtain 1054 unique chemicals. The model from this combined dataset exhibited highest accuracy percentage when the threshold value was set at 1,000 mg/kg. The average accuracy was 88.16%, 88.41%, 88.02%, 85.95% and 88.51% for HNN, RF, Bagging, AdaBoost and ensemble methods respectively (**Figure 6**). Their AUCs were lowest with the values of 0.751, 0.781, 0.77, 0.701 and 0.784 respectively. Very high accuracy percentage of greater than 88% were achieved but with low AUC on this dataset. The accuracy percentage was lowest for the threshold value of 250 mg/kg. The combined dataset is highly imbalanced with ratio of 8.32 to 1 for the threshold value of 1000 mg/kg. The AUCs were high and similar in the case of datasets with 250 mg/kg, 500 mg/kg and 750 mg/kg whose dataset ratios of Toxic:NonToxic varied from 1.44 to 2.15. The

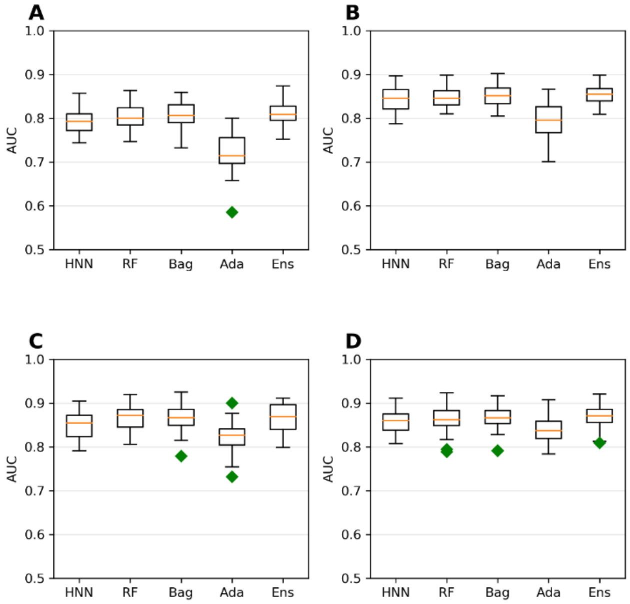


Figure 5: AUC for Toxins LD50 data obtained via Oral route of exposure with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

dataset with 750 mg/kg threshold (nonToxic:Toxic data ratio of 2.15:1) yielded the highest average AUC (approximately 0.881) when the results of all the algorithms were considered (**Figure 7C**). The dataset's average accuracy percentage was approximately 86% (**Figure 6C**). These models showed that optimal AUC can be achieved if the class is balanced (ratio between 1 and 2) for smaller datasets. Highly imbalanced data resulted in overfitting that increases the accuracy apparently but the model's ability to rank the toxic substances before the nontoxic substances reduces as shown by their corresponding AUC.

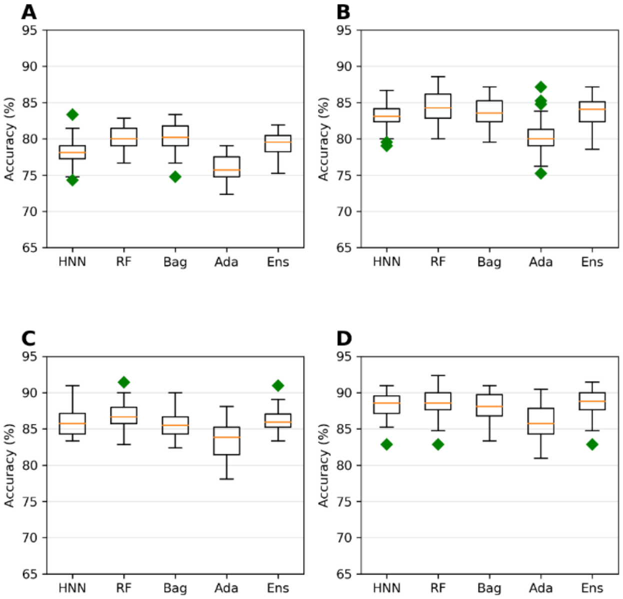


Figure 6. Accuracy percentage for the combined data (Toxins data from T3DB + Animal Toxicity data from EPA) with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

Considering the dataset with highest average AUC i.e. 1000 mg/kg threshold dataset in case of T3DB Oral data (**Figure 5D**) and 750 mg/kg threshold dataset in case of T3DB IPIVSubOral data (**Supplementary Figure 8C**), their highest average AUCs are 0.858 and 0.848 with an average accuracy of 77% (**Figure 4D and Supplementary Figure 7C**). Increasing the data size by adding additional animal toxicity data from EPA increased the accuracy of the model significantly from 77% to 86% (**Figure 6C**) for the group with highest AUCs and average AUC also increased to 0.881 (**Figure 7C**). The highest average accuracy achieved was approximately 88% with the highly imbalanced dataset for 1000 mg/kg threshold (**Figure 6D**) due to overfitting however the AUC significantly reduced to 0.75 **(Figure 7D**). Hence, augmenting the training set with additional data can increase the accuracy while maintaining the AUC if the dataset does not become highly imbalanced.

Tox21 Challenge Dataset.

To assess the performance of our models, the Tox21Challenge 12 different experimental assay data were obtained from the National Institute of Health database and their 801 molecular descriptors were obtained from the Mayr et al[13]. The models were developed using the Tox21 challenge dataset to predict whether the chemicals are active or not. The overall performance of each model was good as can be seen in the **Table 2**. The AUC of the HNN model were similar and sometimes even better than the RF and Bagging method. The AUC of the HNN model was comparable to the DeepTox [winner of the Tox21 challenge, Mayr et al. 2016] model AUC values in many assays in spite of not increasing the performance of the models with additional data. The ensemble method improved the accuracy and the AUC most of the assay data set.

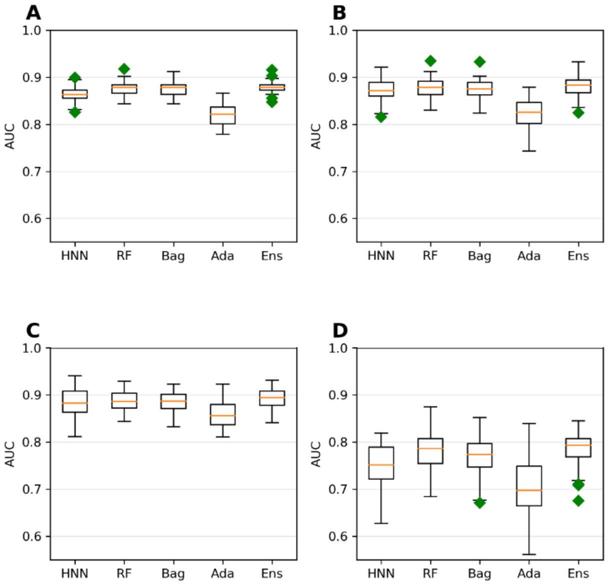


Figure 7. AUC for the combined data (Toxins data from T3DB + Animal Toxicity data from EPA) with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

Table 2. Accuracy percentage and AUC for Tox21 challenge data with 801 descriptors & SMILES for 12 different assays.

		AhR	**AR**	**ARE**	**AR-LBD**	**Aromatase**	**ATAD5**	**ER**	**ER-LBD**	**HSE**	**MMP**	**P53**	**PPAR-gamma**
Accuracy	**HNN**	89.1	97.8	83.2	98.2	92.4	93.7	90.8	96.7	96.6	89.6	93.1	94.7
	RF	90.4	97.9	83.5	98.3	92.8	93.9	91.9	97.1	96.5	90.5	93.3	94.7
	Bag	89.8	97.9	83.1	98.4	92.8	94.6	91.0	96.4	96.8	91.1	93.3	94.7
	Ens	90.3	97.9	83.6	98.4	92.8	93.9	91.2	97.1	96.6	90.7	93.4	94.7
AUC	**HNN**	.886	.783	.775	.756	.769	.782	.751	.741	.774	.917	.828	.733
	RF	.895	.718	.768	.709	.772	.759	.769	.757	.767	.919	.784	.699
	Bag	.897	.741	.759	.718	.761	.749	.751	.766	.776	.908	.765	.711
	Ens	.905	.767	.786	.720	.788	.781	.767	.781	.792	.928	.801	.730
	Deep Tox	.928	.807	.840	.879	.834	.793	.810	.814	.865	.942	.862	.861

Multiclass classification of T3DB oral data.

Degree of toxicity can vary from substance to substance and these substances can be categorized based on their toxicity severity level. To predict the toxicity level of substances the toxins from T3DB database were classified into 4 categories: a) LD50 < 50 mg/kg, b) 50 mg/kg ≤ LD50 < 500 mg/kg, c) 500 ≤ LD50 < 1000, and d) LD50 ≥ 1000 (**Supplementary Table S5**). Multiclass classification models based on HNN, RF, Bagging and SVM algorithms were developed on the toxin dataset and the classes of the test data were predicted. The average accuracy and the corresponding micro AUC are shown in **Figure 8**. In multiclass classification with imbalanced dataset, micro averaging of any metric is preferred when compared to macro averaging. Micro averaging involves calculating the AUC by converting the data in multiple classes to binary classes in contrast to macro averaging which involves averaging the AUC obtained from each class by giving them equal weight. Giving equal weights to each class while calculating the average produces results biased towards the majority class.

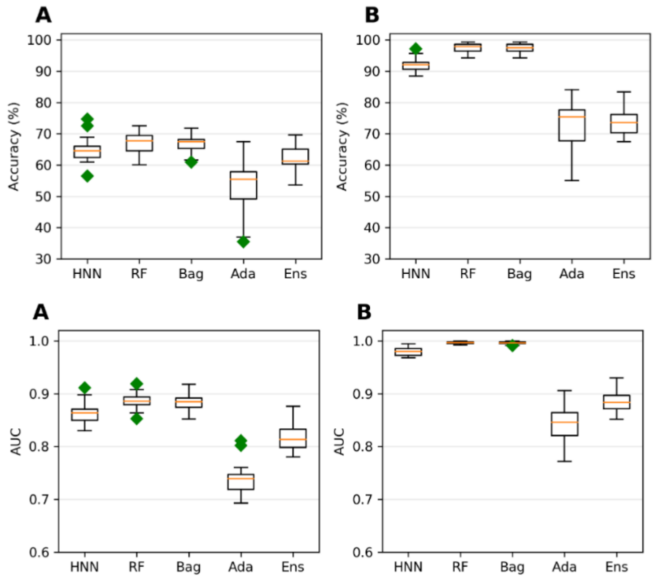


Figure 8. Accuracy percentage and AUC for the multiclass classification of A) Toxins Oral data and B) Toxins Oral data+Oversampling by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

Class imbalance is one of the major problems encountered with toxicity data. This problem is more apparent in case of multiclass classification as separating chemicals into multiple categories increases the possibility of highly imbalanced classes with insufficient number of samples in some classes for training purpose. Hence, the decrease in the accuracy of the model developed on T3DB oral data in case of the multiclass classification is possibly due to these minority classes with smaller training set. The random oversampling method, a simple and competitive method when compared with other complex oversampling methods[26], was applied to overcome the poor performance of the model caused by imbalanced dataset. 100 samples were randomly selected as test set and the remaining chemicals in classes 1, 2 and 3 were duplicated to increase the number of samples in the training set to 100, 200, and 100 respectively drawing them randomly with replacement. By sampling with replacement for multiclass classification, not just the model accuracy but their average AUC also improved significantly (**Figure 8**). Oversampling results in more data to train on in each class and improves the performance of the models.

ACKNOWLEDGMENTS

The author Sivanesan Dakshanamurthy (SD) wishes to acknowledge the support in part by the United States Department of Defense (DOD) grant CA140882. The author Sivanesan Dakshanamurthy wishes to acknowledge the support by the GUMC Lombardi Comprehensive Cancer Center, the CCSG grant P30 CA051008/CA/NCI NIH HHS/United States, and the GUMC Computational Chemistry Shared Resources (CCSR).

REFERENCES

(1) Lei, T.; Li, Y.; Song, Y.; Li, D.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery: 15. Accurate Prediction of Rat Oral Acute Toxicity Using Relevance Vector Machine and Consensus Modeling. *J. Cheminformatics* **2016**, *8*. https://doi.org/10.1186/s13321-016-0117-7.

(2) 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* (11), 2672–2685. https://doi.org/10.1021/acs.jcim.7b00244.

(3) Wu, K.; Wei, G.-W. Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 520–531. https://doi.org/10.1021/acs.jcim.7b00558.

(4) Lu, J.; Peng, J.; Wang, J.; Shen, Q.; Bi, Y.; Gong, L.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; Chen, K. Estimation of Acute Oral Toxicity in Rat Using Local Lazy Learning. *J. Cheminformatics* **2014**, *6*, 26. https://doi.org/10.1186/1758-2946-6-26.

(5) Chavan, S.; Friedman, R.; Nicholls, I. A. Acute Toxicity-Supported Chronic Toxicity Prediction: A k-Nearest Neighbor Coupled Read-Across Strategy. *Int. J. Mol. Sci.* **2015**, *16* (5), 11659–11677. https://doi.org/10.3390/ijms160511659.

(6) Cherkasov, A. Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks. *Int. J. Mol. Sci.* **2005**, *6* (1), 63–86. https://doi.org/10.3390/i6010063.

(7) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* **2017**, *7* (1), 2118. https://doi.org/10.1038/s41598-017-02365-0.

(8) Tanabe, K.; Lučić, B.; Amić, D.; Kurita, T.; Kaihara, M.; Onodera, N.; Suzuki, T. Prediction of Carcinogenicity for Diverse Chemicals Based on Substructure Grouping and SVM Modeling. *Mol. Divers.* **2010**, *14* (4), 789–802. https://doi.org/10.1007/s11030-010-9232-y.

(9) Li, X.; Chen, L.; Cheng, F.; Wu, Z.; Bian, H.; Xu, C.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods. *J. Chem. Inf. Model.* **2014**, *54* (4), 1061–1069. https://doi.org/10.1021/ci5000467.

(10) Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; Kingsbury, B. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29* (6), 82–97. https://doi.org/10.1109/MSP.2012.2205597.

(11) Traore, B. B.; Kamsu-Foguem, B.; Tangara, F. Deep Convolution Neural Network for Image Recognition. *Ecol. Inform.* **2018**, *48*, 257–268. https://doi.org/10.1016/j.ecoinf.2018.10.002.

(12) Lin, T.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-Grained Visual Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*; 2015; pp 1449–1457. https://doi.org/10.1109/ICCV.2015.170.

(13) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*. https://doi.org/10.3389/fenvs.2015.00080.

(14) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. *ArXiv14061231 Cs Stat* **2014**.

(15) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *J. Chem. Inf. Model.* **2018**, *58* (8), 1533–1543. https://doi.org/10.1021/acs.jcim.8b00338.

(16) Jimenez-Carretero, D.; Abrishami, V.; Fernández-de-Manuel, L.; Palacios, I.; Quílez-Álvarez, A.; Díez-Sánchez, A.; Pozo, M. A. del; Montoya, M. C. Tox_(R)CNN: Deep Learning-Based Nuclei Profiling Tool for Drug Toxicity Screening. *PLOS Comput. Biol.* **2018**, *14* (11), e1006238. https://doi.org/10.1371/journal.pcbi.1006238.

(17) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55* (10), 2085–2093. https://doi.org/10.1021/acs.jcim.5b00238.

(18) Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*; 2014; pp 1701–1708. https://doi.org/10.1109/CVPR.2014.220.

(19) Dong, J.; Wang, N.-N.; Yao, Z.-J.; Zhang, L.; Cheng, Y.; Ouyang, D.; Lu, A.-P.; Cao, D.-S. ADMETlab: A Platform for Systematic ADMET Evaluation Based on a Comprehensively Collected ADMET Database. *J. Cheminformatics* **2018**, *10* (1), 29. https://doi.org/10.1186/s13321-018-0283-x.

(20) 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. https://doi.org/10.1016/j.comtox.2018.08.002.

(21) Lim, E.; Pon, A.; Djoumbou, Y.; Knox, C.; Shrivastava, S.; Guo, A.; Neveu, V.; Wishart, D. T3DB: A Comprehensively Annotated Database of Common Toxins and Their Targets. *Nucleic Acids Res* **2010**. https://doi.org/10.1093/nar/gkp934.

(22) Bradley, A. P. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognit.* **1997**, *30* (7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

(23) Jin Huang; Ling, C. X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17* (3), 299–310. https://doi.org/10.1109/TKDE.2005.50.

(24) Banko, M.; Brill, E. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*; ACL ’01; Association for Computational Linguistics: USA, 2001; pp 26–33. https://doi.org/10.3115/1073012.1073017.

(25) Sharma, A. K.; Srivastava, G. N.; Roy, A.; Sharma, V. K. ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front. Pharmacol.* **2017**, *8*. https://doi.org/10.3389/fphar.2017.00880.

(26) Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor Newsl* **2004**, *6* (1), 20–29. https://doi.org/10.1145/1007730.1007735.

Supplementary Materials:

Equation S1. Equation to calculate the ensemble probability derived by Mayr et al.

$$\frac{\prod_{i=1}^{n} p(t = 1 \mid y_i)}{\prod_{i=1}^{n} p(t = 1 \mid y_i) + \prod_{i=1}^{n} p(t = 0 \mid y_i)}$$

Where, n is the number of models,y_i is the prediction score by the model *i*, $p(t = 1|y_i)$ is the predicted probability for class 1 and $p(t = 0|y_i)$ is the predicted probability for class 0.

Equation S2. Equations to calculate the evaluation metrics.

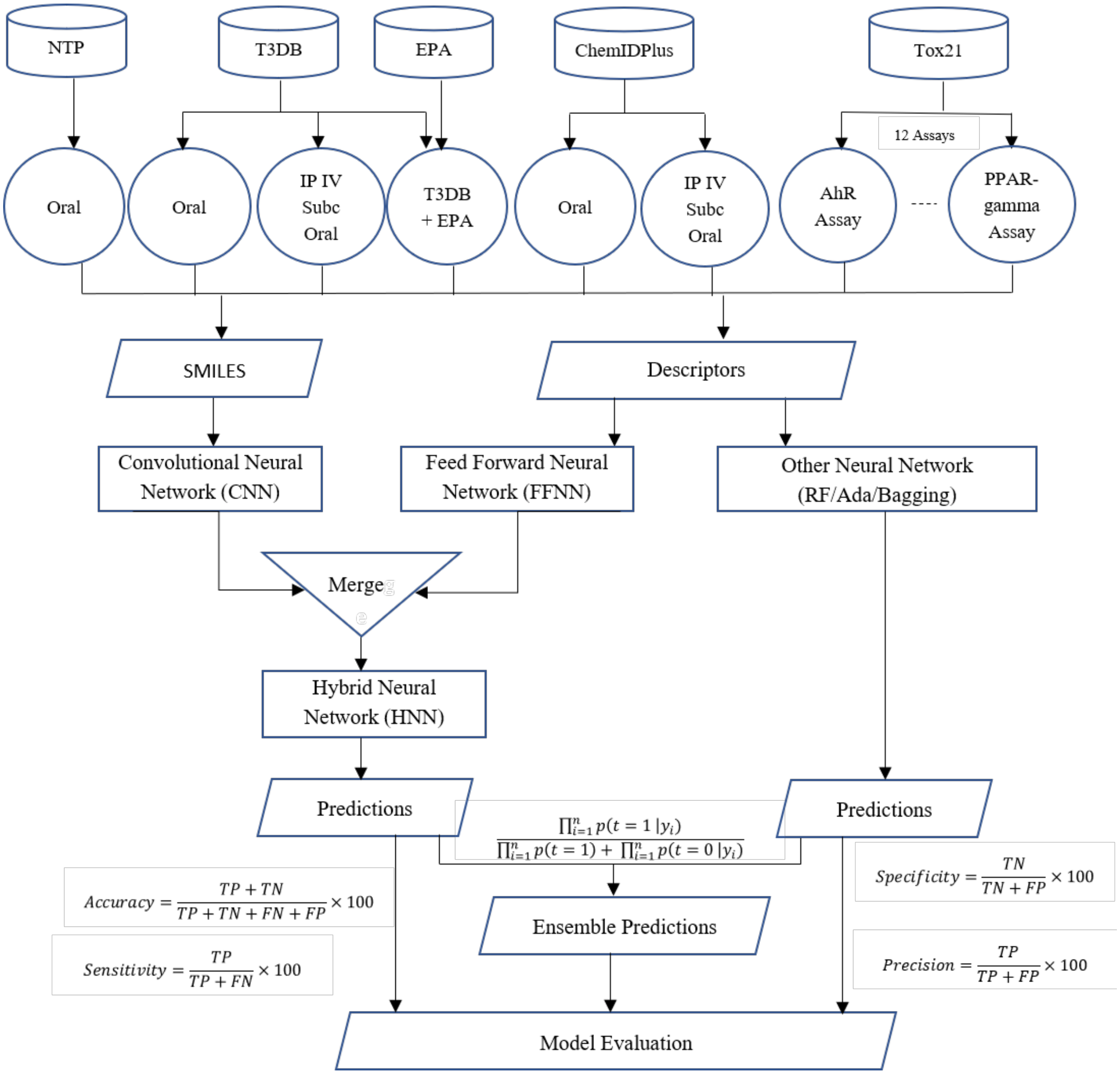
$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100$$

$$Sensitivity(truepositiverate) = \frac{TP}{TP + FN} \times 100$$

$$Specificity(truenegativerate) = \frac{TN}{TN + FP} \times 100$$

$$Precision = \frac{TP}{TP + FP} \times 100$$

Where, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.


NTP
T3DB
EPA
ChemIDPlus
Tox21
12 Assays
Oral
Oral
IP IV Subc Oral
T3DB + EPA
Oral
IP IV Subc Oral
AhR Assay

PPAR-gamma Assay
SMILES
Descriptors
Convolutional Neural Network (CNN)
Feed Forward Neural Network (FFNN)
Other Neural Network (RF/Ada/Bagging)
Merge
Hybrid Neural Network (HNN)
Predictions
$\frac{\prod_{i=1}^{n} p(t=1\,|y_i)}{\prod_{i=1}^{n} p(t=1) + \prod_{i=1}^{n} p(t=0\,|y_i)}$
Predictions
$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \times 100$
$Specificity = \frac{TN}{TN+FP} \times 100$
$Sensitivity = \frac{TP}{TP+FN} \times 100$
$Precision = \frac{TP}{TP+FP} \times 100$
Ensemble Predictions
Model Evaluation

Figure S1. The overall flowchart of the data and models involved. Data from ChemIDplus, T3DB, EPA, NTP, and Tox21 Challenge are obtained, preprocessed, appropriate descriptors calculated. HNN, RF, Bagging, AdaBoost and Ensemble models are developed on data from ChemIDplus, T3DB, EPA, and Tox21 Challenge, toxicity predicted and the predictive performance of the models are evaluated using various statistical metrics. The data from NTP and T3DB were used as external validation set for the models developed on ChemIDplus data.

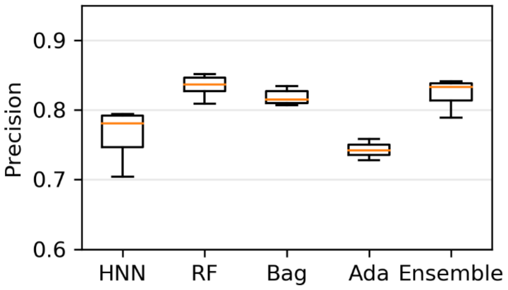


Figure S2. Precision for the ChemIDplus Oral data as given by HNN, RF, Bagging, AdaBoost and the Ensemble methods with additional descriptors from ADMETlab and Canvas.

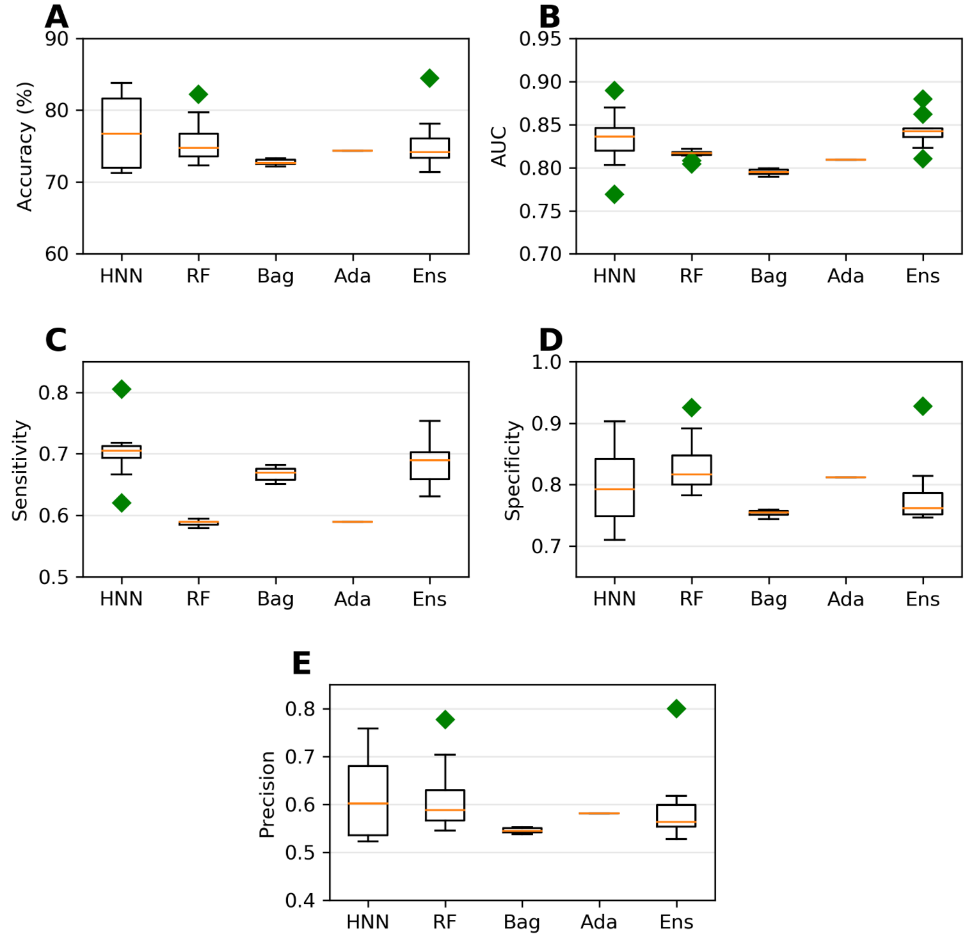


Figure S3: A) Accuracy percentage, B) AUC, C) Sensitivity, D) Specificity and E) Precision for the T3DB external validation dataset by HNN, RF, Bagging, AdaBoost and the Ensemble to validate the models built on ChemIDplus Oral data.

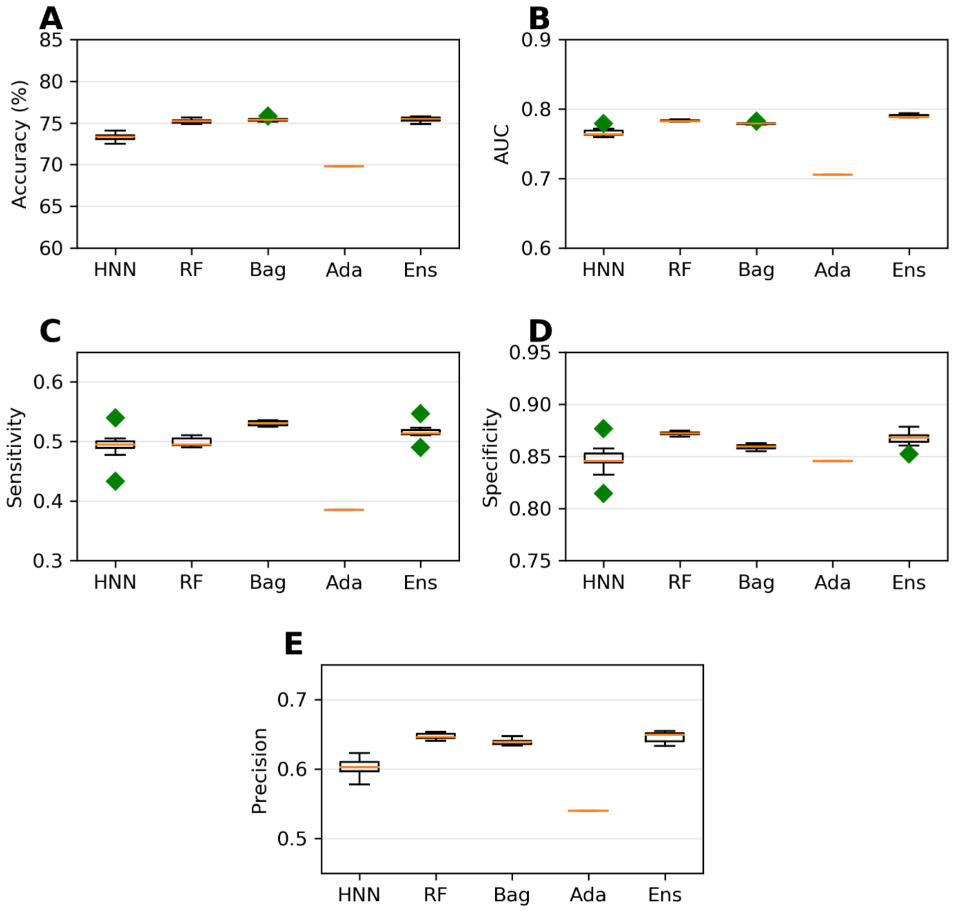


Figure S4: A) Accuracy percentage, B) AUC, C) Sensitivity, D) Specificity and E) Precision for the external validation dataset from NTP by HNN, RF, Bagging, AdaBoost and the Ensemble methods to validate the models built on ChemIDplus Oral data.

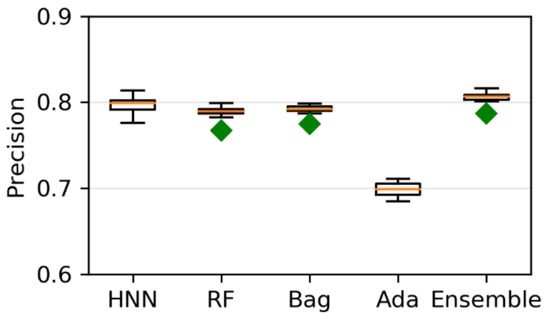


Figure S5: Precision for the ChemIDplus IP/IV/Sub/Oral data as given by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

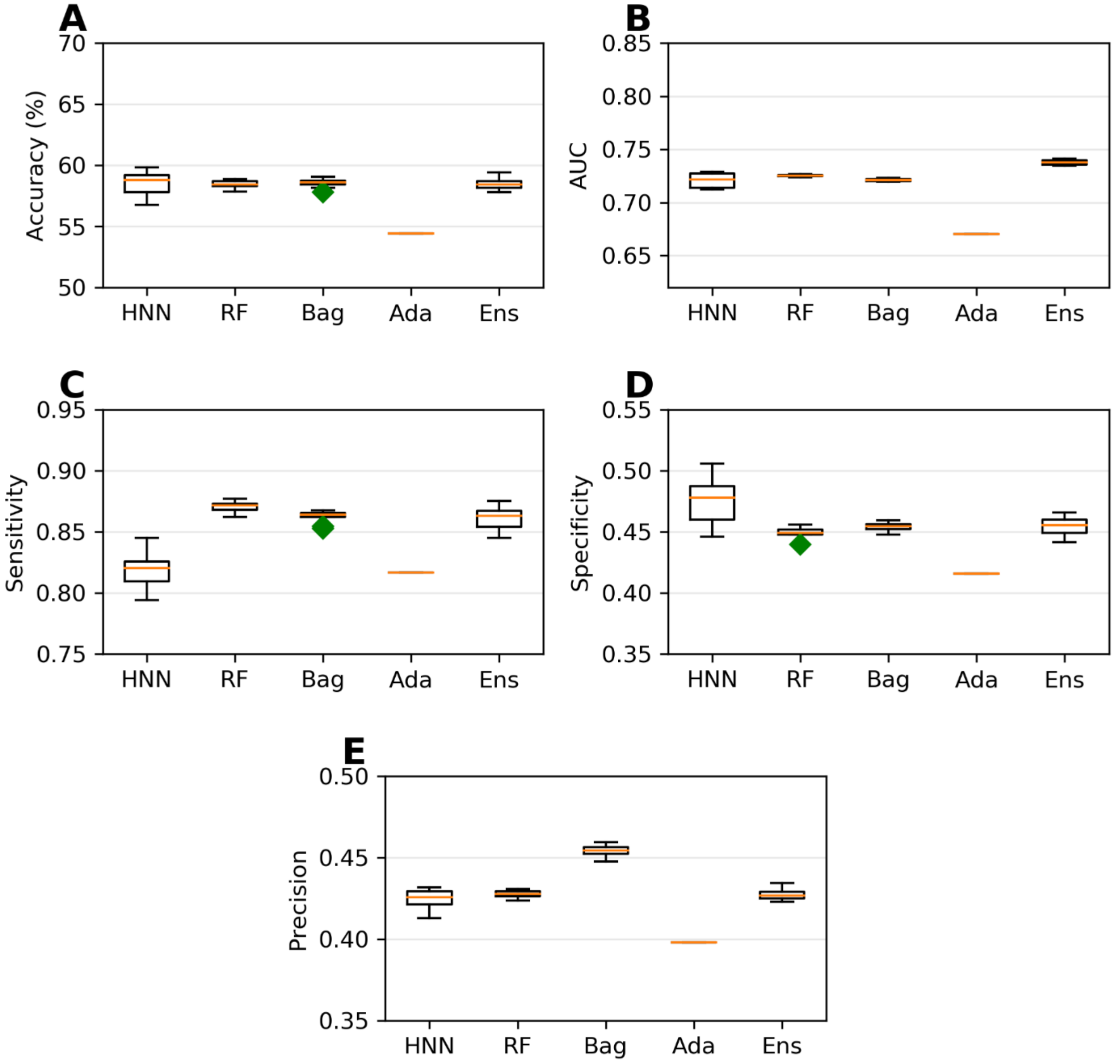


Figure S6: A) Accuracy percentage, B) AUC, C) Sensitivity, D) Specificity and E) Precision for the NTP external validation dataset by HNN, RF, Bagging, AdaBoost and the Ensemble methods to validate the models built on ChemIDplus IP/IV/Sub/Oral data.

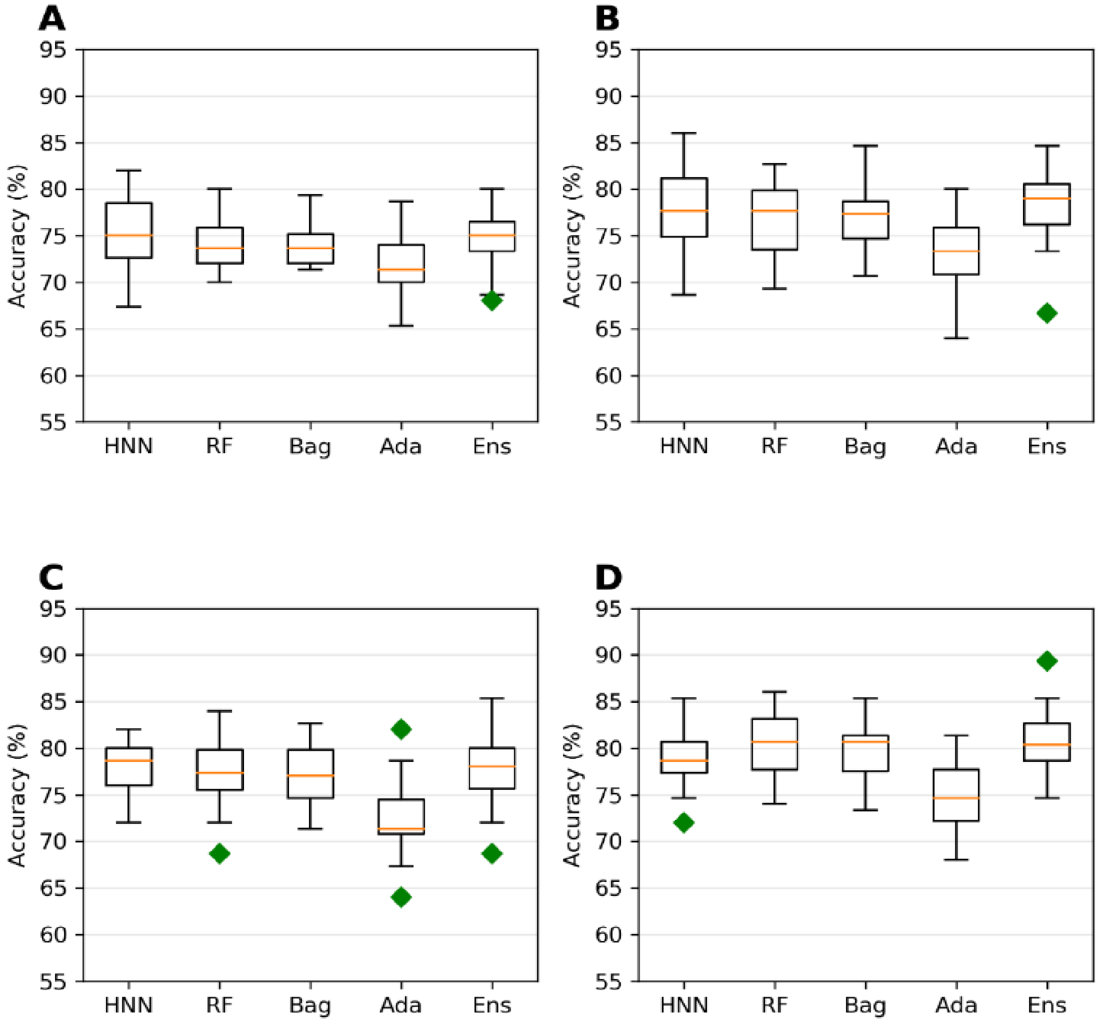


Figure S7: Accuracy percentage for Toxins LD50 data obtained via IP, IV, Subcutaneous and Oral route of exposure with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

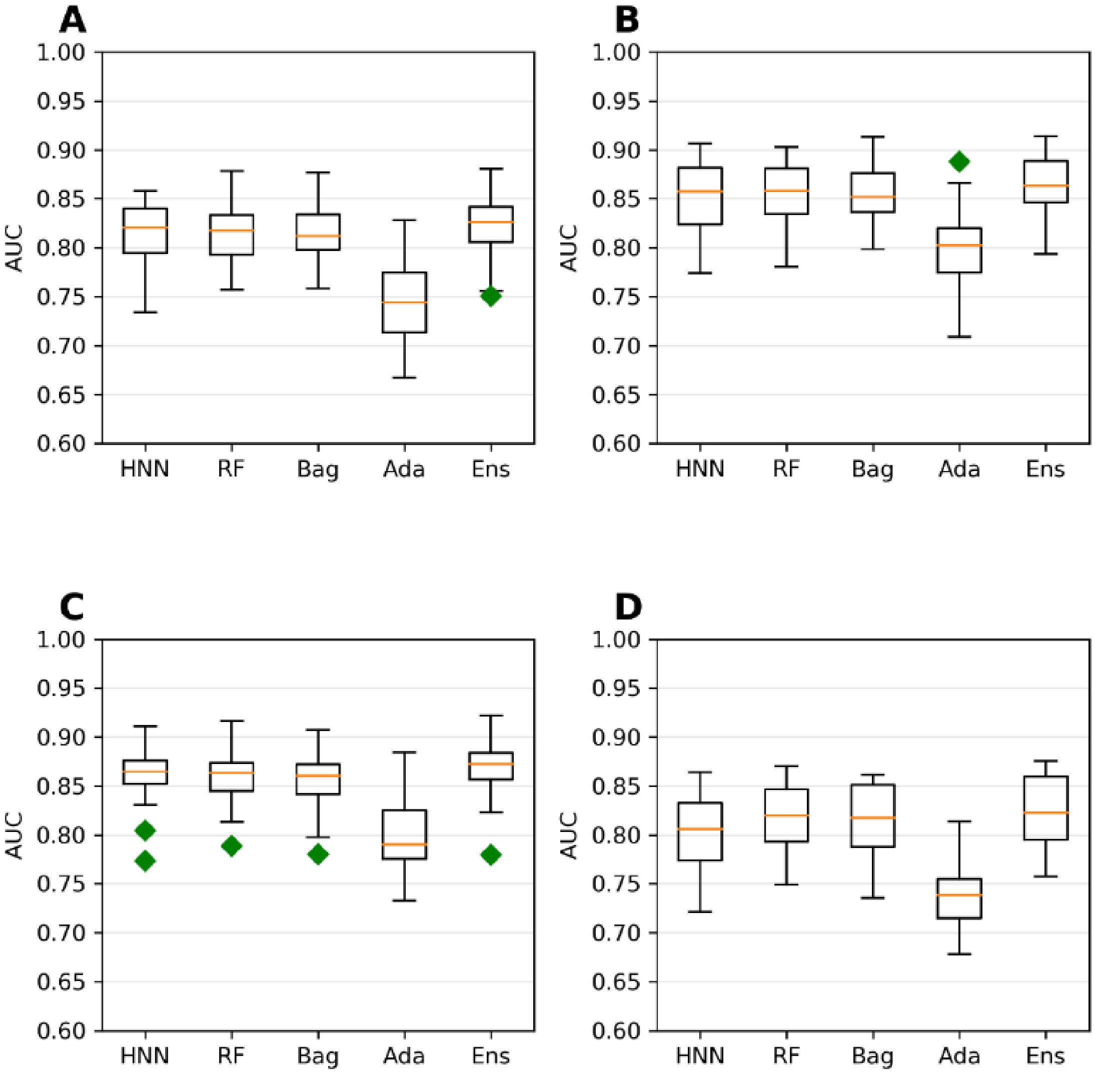


Figure S8: AUC for Toxins LD50 data obtained via IP, IV, Subcutaneous and Oral route of exposure with cutoffs at A) 250 mg/kg, B) 500 mg/kg, C) 750 mg/kg and D) 1000 mg/kg by HNN, RF, Bagging, AdaBoost and the Ensemble methods.

Table S1. Toxins data from T3DB.

LD50 threshold	T3DB's Toxins	
	62 desc (778)	
	Toxic	Non Toxic
<1000	632	146
<750	404	374
<500	363	415
<250	285	493

Table S2. Toxins data from T3DB IP/IV/Subcutaneous/Oral route of exposure.

LD50 threshold	Oral		IP, IV, Subcutaneous, Oral	
	62 desc (687)		62 desc (752)	
	Toxic	Non T	Toxic	Non T
<1000	285	402	573	179
<750	262	425	345	407
<500	225	462	307	445
<250	157	530	238	514

Table S3. Animal toxicity data from EPA.

LD50 threshold	EPA's Animal Toxicity	
	62 desc (427)	
	Toxic	Non T
<1000	418	9
<750	418	9
<500	412	15
<250	404	23

Table S4. Unique data after merging animal toxicity data from EPA (Table S1) and toxins data from T3DB (Table S3).

LD50 threshold	**EPA’s Animal Toxicity+T3DB**	
	62 desc (1054)	
	Toxic	**Non T**
<1000	941	113
<750	720	334
<500	687	367
<250	623	431

Table S5. Distribution of Toxins Oral data among the 4 classes in multiclass classification.

Threshold (mg/kg)	**Class**	**No. of Chem**
LD50<50	3	68
50≤ LD50<500	2	157
500≤LD50<1000	1	60
LD50≥1000	0	402