In Silico Toxicity Assessment by QSAR Model: Bridging the Knowledge Gap Using Standardized Electronic Submission Data

Md Aminul Islam Prodhan[1,2], C M Sabbir Ahmed[1], Kevin Snyder[1]

[1]U.S. Food and Drug Administration (FDA), Silver Spring, MD; [2]Oak Ridge Institute for Science and Education, Oak Ridge, TN.

BACKGROUND

Quantitative structure activity relationship (QSAR) modeling is currently used in CDER to qualify the safety of drug substance impurities with respect to genetic toxicology; however, reliable use of QSAR with respect to general toxicology endpoints has yet to be robustly demonstrated. CDER has recently started to receive electronic SEND datasets along with study reports from in vivo general toxicology studies, and automated analyses of these SEND datasets can be used to build QSAR models to predict the toxicological profile of novel compounds with structures similar to those present in the database.

Body weight was selected as the first general toxicology study endpoint to be addressed primarily because it is a relatively simple endpoint to normalize and analyze in the context of QSAR modeling. Although the mechanistic causes of drug-induced decreases in body weight may be diverse, it is possible that the model may be able to reliably predict the absence of this effect for certain classes of chemical structures. Modeling of complex, multi-endpoint toxicities, e.g. hepatotoxicity, will be conducted in the future.

METHODS

BODY WEIGHT Z-SCORE CALCULATION

The weight of each animal at the end of the dosing period in each study was initially normalized by subtracting their baseline weights on the first day of dosing. These values then were normalized via Z-score (Equation 1) to the respective control groups for each study.

$$Z_{s,i} = \frac{x_{s,i} - \mu_{s,c}}{\sigma_{s,c}}$$

Equation 1. Z-score equation for data normalization where x is the endpoint value being observed, μ is the mean value of that endpoint, σ is the standard deviation of that endpoint, s is the study, i is an individual animal from that study, and c is the control-treated group of animals from that study.

MOLECULAR DESCRIPTOR CALCULATION

Python version 3.9.12, scikit-learn, SQLite3 and R version 4.3.1 was used for data analysis and implementing various machine learning algorithms. Mordred, a python-based descriptor calculator package, was used for molecular descriptor calculation. Mordred is a python-based software that can calculate more than 1800 two- and three-dimensional descriptors. SMILES molecular notation was used for descriptor calculation. As SEND datasets do not natively contain SMILES strings, SMILES strings were retrieved from the GSRS dataset in combination with the SEND dataset based on application number. After removing highly correlated descriptors, 194 final descriptors were used for the machine leaning classifier model building.

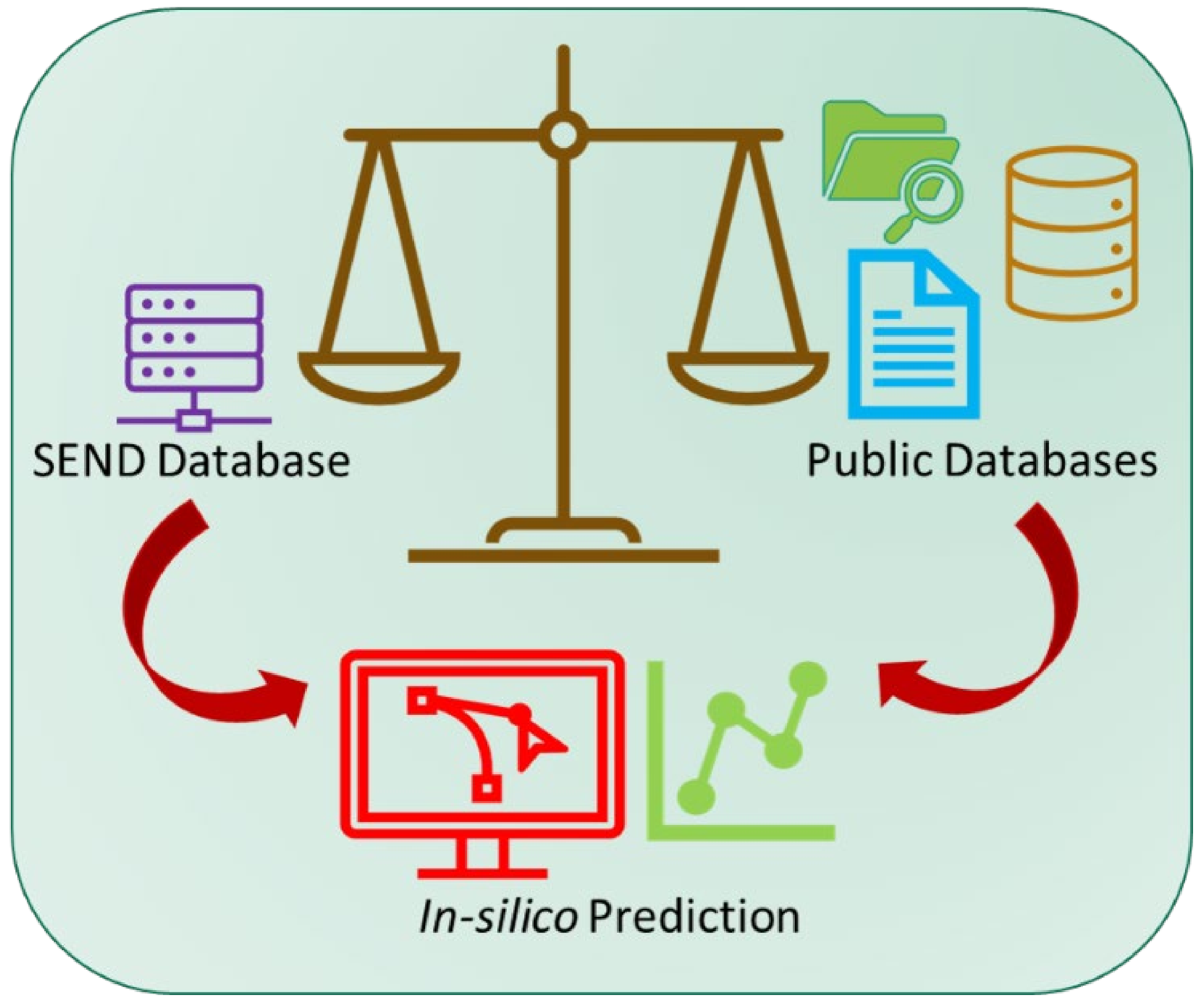

Figure 1. Schematic of QSAR modeling

QSAR MODEL BUILDING

Animal body weight z-score data and the corresponding calculated molecular descriptors were used to build the machine learning classifier and predict the probability of toxicity. Z-score values were binned into different classes and used as the response variable for the molecular descriptors. Body weight z-score data were respectively binned into two categories: toxic, and non-toxic based on the following threshold values z-score ≤ -2 , and z-score > -2.

The data were split into training data (80%) and test data (20%) . The models were validated by k-fold-cross validation. In addition, We also applied a feature selection technique to remove the highly correlated features. The hyper-parameters of the models were optimized by the grid search method. In addition to chemical descriptor, species and SEX of the animal were also used as a features by converting them in numerical values by encoding technique.

MODEL PERFORMANCE EVALUATION

Four metrics (precision, recall, f1-score, and accuracy) were used to assess model performance as described by the traditional equations containing the true positive, false positive, true negative, and false negative cases.

RESULTS

Classifier	Precision	Recall	F1-score
RandomForest	0.6136	0.5420	0.5214
Kneighbors	0.6019	0.5924	0.5956
GradientBoosting	0.5873	0.5462	0.5368
AdaBoost	0.5406	0.5308	0.5282
XGB	0.5263	0.5210	0.5185
LGBM	0.5238	0.5126	0.4947
CatBoost	0.5217	0.5042	0.4500
DecisionTree	0.5217	0.5238	0.5209
ExtraTrees	0.5048	0.5028	0.4862

Table 1. ML performance metrics

Classifier	Accuracy	Precision	F1 Score
RandomForest	0.6889	0.7051	0.7971
Kneighbors	0.5778	0.6447	0.7206
GradientBoosting	0.6444	0.6750	0.7714
AdaBoost	0.5667	0.6479	0.7023
XGB	0.6333	0.6753	0.7591
LGBM	0.6222	0.6625	0.7571
CatBoost	0.6333	0.6588	0.7724
DecisionTree	0.6000	0.6875	0.7097
ExtraTrees	0.6667	0.6875	0.7857

Table 2. performance metrics on prediction

The data analysis provided 1988 unique rows of 447 unique compounds. The rows were unique based on the compounds, species and sex value. From this 1988 rows, 447 unique rows were selected based on unique SMILES notation. Among 447 rows, 357 rows were used for model building and 90 rows used for toxicity prediction. From the training data set, the resulting classifiers showed moderate overall accuracy (balanced accuracy, average 55%) with other metrics as described in Table 1. The testing data set (Table 2) showed an overall average accuracy of around 62% on toxicity prediction. To improve the model, we also applied the features reduction method and highly correlated features were removed and then model were built and tested . However, there were no significant improvement in terms of toxicity prediction.

DISCUSSION AND FUTURE DIRECTIONS

It was not altogether surprising that model performance was not sufficient to produce reliable predictions of drug-induced decreased in body weight as this endpoint could be driven by a wide variety of pharmacological mechanisms; however, the infrastructure that was developed to build models to predict this endpoint can be easily repurposed to predict other types of toxicity, e.g. hepatotoxicity, cardiotoxicity, nephrotoxicity. Although the integration of study endpoints related to these types of toxicity into a single response variable score that can be modeled and predicted will be challenging, it is likely that the relationship between chemical features and these more distinct, mechanistically-driven toxicities will be able to be modeled with greater reliably.

In the future, incorporation of structural alert data from the chemical structure of the compounds may also be implemented to improve the model. Additionally, the model may be able to be enhanced by incorporation of data from publicly available sources, e.g. ToxRefDB.

ACKNOWLEDGEMENT

This project was supported in part by an appointment to the Research Fellowship Program at the Immediate Office/OND/CDER, U.S. Food and Drug Administration, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and FDA..