Leveraging machine learning and artificial intelligence to enhance toxicological sciences and physiologically based pharmacokinetic modeling

Zhoumeng Lin, BMed, PhD, DABT, CPH, ERT

Center for Environmental and Human Toxicology (CEHT) & Center for Pharmacometrics and Systems Pharmacology (CPSP)  
Department of Environmental and Global Health, College of Public Health and Health Professions (primary appointment)  
Department of Physiological Sciences, College of Veterinary Medicine (jointly appointed)  
Department of Pharmaceutics, College of Pharmacy (jointly appointed)  
University of Florida, Gainesville, FL 32610
Outline

Introduction
- Research overview
- Background and terminology (PBPK, QSAR, ML, AI)

Different Applications
- AI in predicting ADME of chemicals
- AI in predicting toxicity of chemicals
- AI-assisted PBPK model for nanoparticles

Summary
- Summary and Discussion
- Acknowledgements
Research Program in Computational Toxicology and Pharmacology

We develop computational models to address nanomedicine, animal-derived food safety, human health risk assessment issues!

- Nanomaterials
- Environmental chemicals
- Drugs
Machine Learning (ML) and Artificial Intelligence (AI)

- Artificial intelligence (AI) is a rapidly developing subdiscipline of computer science with the goal of designing and creating machines or computational models that can perform a variety of cognitive tasks at a level comparable or even exceed human intelligence.

- In this presentation, it mainly refers to the applications of various machine learning methods in the prediction and evaluation of chemical toxicokinetic (i.e., absorption, distribution, metabolism, and excretion [ADME]) and toxicity properties.

- Machine learning (ML) is a subarea of artificial intelligence, and it refers to mathematical or computer algorithms designed to teach or train a computational model to solve a problem or perform complex tasks based on some input parameters.

What is PBPK Modeling?
What is QSAR Modeling?

- Quantitative structure activity relationship analysis (QSAR): the study of the relationship between chemical structure and biological properties of substances.
- QSAR has long been used by researchers to predict pharmacokinetics and toxicity properties of chemicals and to develop new products or therapeutic agents with desirable properties.

Applying ML and AI in Different Subject Areas of Toxicology

- Physiologically based pharmacokinetic (PBPK) modeling
- Quantitative structure-activity relationship (QSAR) modeling
- Adverse outcome pathway (AOP) analysis
- High-content image-based screening
- Toxicogenomics

CONTEMPORARY REVIEW

Machine Learning and Artificial Intelligence in Toxicological Sciences

Zhoumeng Lin [*],† and Wei-Chun Chou [*],†

*Department of Environmental and Global Health, College of Public Health and Health Professions, University of Florida, Gainesville, Florida 32610, USA; and †Center for Environmental and Human Toxicology, University of Florida, Gainesville, Florida 32608, USA

To whom correspondence should be addressed at Department of Environmental and Global Health, College of Public Health and Health Professions, University of Florida, 1225 Center Drive, Gainesville, FL 32610, USA. E-mail: linzhoumeng@ufl.edu.

## Commonly Used Machine Learning Methods in Toxicology

<table>
<thead>
<tr>
<th>Method</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised linear methods</td>
<td></td>
</tr>
<tr>
<td>Multiple linear regression</td>
<td>Use multiple explanatory variables to predict the outcome of a response variable with a multivariate linear equation.</td>
</tr>
<tr>
<td>Naive Bayes classifier</td>
<td>Based on Bayes’ theorem with strong assumptions of conditional independence among molecular descriptors (i.e., explanatory variables).</td>
</tr>
<tr>
<td>Supervised nonlinear methods</td>
<td></td>
</tr>
<tr>
<td>k-nearest neighbors</td>
<td>Classify a test chemical by looking for the training chemicals with the nearest distance to it.</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>Map molecular descriptor vectors into a higher dimensional feature space to build a maximal margin hyperplane to distinguish active (toxic) from inactive (nontoxic) chemicals.</td>
</tr>
<tr>
<td>Decision trees</td>
<td>Each model is a series of rules organized in the format of a tree containing a single root node and any number of internal nodes and several leaf nodes. The path from the root to a leaf stands for a sequence of classification rules predicting a toxicity endpoint for a given chemical.</td>
</tr>
<tr>
<td>Ensemble learning</td>
<td>Combine several base models into a more predictive one. Popular types of ensemble modeling include bagging, random spaces, boosting, and stacking.</td>
</tr>
<tr>
<td>Random forest</td>
<td>Combine the bagging with the random spaces approaches in application to decision trees base models.</td>
</tr>
<tr>
<td>Artificial neural networks</td>
<td></td>
</tr>
<tr>
<td>Backpropagation neural networks</td>
<td>All neurons are divided into 3 layers, with information flowing from the first layer of input neurons to the second layer of hidden neurons, and then to the third layer of output neurons.</td>
</tr>
<tr>
<td>Bayesian-regularized neural networks</td>
<td>Apply Bayesian methods to perform regularization so that the model complexity is balanced against the accuracy of reproducing training data.</td>
</tr>
<tr>
<td>Associative neural networks</td>
<td>Apply ensemble learning to backpropagation neural networks.</td>
</tr>
<tr>
<td>Deep neural networks</td>
<td>Artificial neural networks with multiple hidden layers (also called deep learning).</td>
</tr>
<tr>
<td>Unsupervised methods</td>
<td></td>
</tr>
<tr>
<td>Principle component analysis</td>
<td>Reduce the dimensionality of the data to only the first few principal components while preserving as much of the data’s variation as possible.</td>
</tr>
<tr>
<td>Kohonen’s self-organizing maps</td>
<td>Map molecules from the original descriptor space onto a 2D grid of neurons. Similar molecules will be mapped to the same closely located neurons in the grid.</td>
</tr>
</tbody>
</table>

This table is based on the book chapter by Baspin (2018). Please refer to Baspin (2018) for detailed description about each of the listed machine learning algorithms.
List of Studies using ML in QSAR Modeling to Predict Toxicity

<table>
<thead>
<tr>
<th>Best Machine learning Method</th>
<th>Training Dataset</th>
<th>Endpoint</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning (ie, DeepTox)</td>
<td>11,764 chemicals from Tox21</td>
<td>12 bioassays</td>
<td>Mayr et al. (2016)</td>
</tr>
<tr>
<td>Ensemble extreme gradient boosting</td>
<td>1003 chemicals</td>
<td>Carcinogenicity</td>
<td>Zhang et al. (2017)</td>
</tr>
<tr>
<td>Random forest</td>
<td>Over 866,000 chemical properties/hazards</td>
<td>Acute oral and dermal toxicity, eye and skin irritation, mutagenicity, and skin sensitization</td>
<td>Luechtfeld et al. (2018)</td>
</tr>
<tr>
<td>Ensemble support vector machine</td>
<td>400 chemicals</td>
<td>Aquatic acute toxicity</td>
<td>Ai et al. (2019)</td>
</tr>
<tr>
<td>Multitask neural networks and graph convolutional networks</td>
<td>1012 PFAS</td>
<td>Bioactivity on 26 bioassays</td>
<td>Cheng and Ng (2019)</td>
</tr>
<tr>
<td>Extra trees</td>
<td>Over 1000 chemicals from different databases</td>
<td>Various toxicities</td>
<td>Fu et al. (2019)</td>
</tr>
<tr>
<td>Ensemble model</td>
<td>7385 chemicals</td>
<td>Acute toxicity in rats</td>
<td>Russo et al. (2019)</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>482 chemicals</td>
<td>Acute toxicity in fathead minnow</td>
<td>Chen et al. (2020)</td>
</tr>
<tr>
<td>Deep learning (ie, CapsCarcin)</td>
<td>1003 chemicals from CFDB</td>
<td>Carcinogenicity</td>
<td>Wang et al. (2020)</td>
</tr>
<tr>
<td>Kernel-weighted local polynomial approach</td>
<td>Hundreds of chemicals depending on the species</td>
<td>Acute aquatic toxicity</td>
<td>Gajewicz-Skretta et al. (2021)</td>
</tr>
<tr>
<td>Meta ensembling of multitask deep learning models (ie, QuantitativeTox)</td>
<td>Hundreds to thousands of compounds depending on the endpoint</td>
<td>LD_50 and LC_50</td>
<td>Karim et al. (2021)</td>
</tr>
<tr>
<td>Deep learning-based model-level representations (ie, DeepCarc)</td>
<td>692 chemicals</td>
<td>Carcinogenicity</td>
<td>Li et al. (2021)</td>
</tr>
<tr>
<td>Extra trees</td>
<td>Over 18,600 drug-bacteria interactions</td>
<td>Gut bacterial growth</td>
<td>McCoubrey et al. (2021)</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>676 pesticides</td>
<td>Acute contact toxicity on honey bees</td>
<td>Xu et al. (2021)</td>
</tr>
<tr>
<td>A consensus model based on 4 algorithms</td>
<td>1244 chemicals</td>
<td>Prenatal developmental toxicity</td>
<td>Ciallella et al. (2022)</td>
</tr>
<tr>
<td>Deep learning</td>
<td>31 chemicals with known or suspected clinical skin toxicity</td>
<td>Skin toxicity</td>
<td>Hu et al. (2022)</td>
</tr>
<tr>
<td>Random forest</td>
<td>1476 food contact chemicals</td>
<td>Carcinogenicity</td>
<td>Wang et al. (2022)</td>
</tr>
</tbody>
</table>

CPDB, Carcinogenic Potency Database. LD_50 and LC_50 refer to the compound concentrations that kill half the members of the tested animal population, respectively.

<table>
<thead>
<tr>
<th>References</th>
<th>N</th>
<th>Predict Target</th>
<th>Descriptor Types</th>
<th>Modeling Method</th>
<th>Performance[^a]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Absorption</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Agatonovic-Rustin et al. (2001) | 86 | HIA | 0D-3D theoretical descriptors | ANN, RBF, GNN | Training set: $R^2 = 0.82; RMSE = 0.59$  
Test set: RMSE = 0.90  
Whole data set: RMSE = 0.93 |
| Decoeneck et al. (2007) | 67 | HIA | 1D-3D theoretical descriptors; one of Abraham’s solvation parameters | MARS | Training set: RMSE = 6.5  
Test set: RMSE = 22.8 |
| Niwa (2003) | 86 | HIA | 0D-1D theoretical descriptors | GRNN, PNN | Training set: RMSE = 0.9  
Test set: RMSE = 0.8; RMSE = 0.18  
Whole data set: RMSE = 0.21 |
| Talevi et al. (2011) | 120 | HIA | 0D-3D Dragon theoretical descriptors | MLR, AN, SVM | Training set: RMSE = 0.77; RMSE = 0.16  
Whole data set: $R^2 = 0.77; RMSE = 16$ |
| Yan et al. (2008) | 52 | HIA | Adriana Code and Cerius2 0D-2D theoretical descriptors | GA, PLS, SVM | Training set: Q = 98.5%  
Test set: Q = 99%  
Whole data set: $R = 0.84-0.85$ |
| Shen et al. (2010) | 1593 | HIA | 1D-2D theoretical descriptors | SVM | Training set: SE = 0.89; SP = 0.85; Q = 0.89  
Test set: SE = 0.88; SP = 0.81; Q = 0.87 |
| Kamiya et al. (2011b) | 184 | Papp | Chemical descriptors (not specific descriptions) | SVM, PLS, RBF | Training set: RMSE = 15.44  
Test set: RMSE = 23.84 |
| Ghafourian et al. (2012) | 310 | HIA | A total of 215 descriptors (not specific descriptions) | MLR | Training set: $R^2 = 0.973$  
Test set: $R^2 = 0.98$ |
| Hou et al. (2007) | 648 | HIA | 0D-2D theoretical descriptors | MARS, GA | Training set: SE = 0.89; SP = 0.85; Q = 0.89  
Test set: SE = 0.88; SP = 0.81; Q = 0.87 |
| Wang et al. (2017) | 970 | HIA | 2D-3D descriptors, molecular fingerprints, and structural fragments | RF | Test set: $AFE = 0.96$ (Cmax); $0.89$ (AUC), $0.69$ (Vd); $AAFE = 1.2$ (Cmax); $1.30$ (AUC), $1.71$ (Vd); $R^2 = 0.99$ (Cmax), $0.98$ (AUC), $0.99$ (Vd) |
| **Distribution** |    |                |                  |                |                 |
| Antontsev et al. (2021) | 21 | Kp | Not explained in the study | BIOISIM | Test set: $AHE = 0.96$ (Cmax), $0.89$ (AUC), $0.69$ (Vd); $AAFE = 1.2$ (Cmax), $1.30$ (AUC), $1.71$ (Vd); $R^2 = 0.99$ (Cmax), $0.98$ (AUC), $0.99$ (Vd) |
| Golmohammadi et al. (2012) | 310 | Kp | 3D descriptors and molecular structural information | SVM; GA, PLS | Training set: $R^2 = 0.98$, RMSE = 0.117  
Test set: $R^2 = 0.98$, RMSE = 0.118  
Whole data set: $R^2 = 0.974$, RMSE = 0.0289 |
| Liu et al. (2005) | 208 | Kp | Constitutional, topological, geometrical, electrostatic and quantum chemical descriptors | SVM | Test set: $R^2 = 0.97$, RMSE = 0.02 |
| Yun et al. (2014) | 122 | Kp | LogP, pKa, fu | DT; RF | Test set: $R^2 = 0.974$, RMSE = 0.0289 |

[^a]: RMSE: Root Mean Square Error; $R^2$: Coefficient of determination; SE: Sensitivity; SP: Specificity; AHE: Area under the Effective Concentration curve; AAFE: Area under the Absolute Effective Concentration curve; Cmax: Maximum effective concentration; AUC: Area under the curve; Vd: Volume of distribution; AU: Absolute uncertainty; T: Time.
### Studies That Used ML/AI to Predict ADME for Pharmaceutical Compounds

<table>
<thead>
<tr>
<th>References</th>
<th>N</th>
<th>Predict Target</th>
<th>Descriptor Types</th>
<th>Modeling Method</th>
<th>Performance*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metabolic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Athemuch et al. (2013)</td>
<td>15</td>
<td>Classify the metabolic pathways of test compounds</td>
<td>PCA, PLS</td>
<td>Whole data set: R² = 0.96, Q = 77.5%</td>
<td></td>
</tr>
<tr>
<td>Baranwal et al. (2020)</td>
<td>6669</td>
<td>Classify the metabolic pathways of test compounds</td>
<td>RF and GCN</td>
<td>Test set: Q = 98.999%</td>
<td></td>
</tr>
<tr>
<td>Jia et al. (2020)</td>
<td>5682</td>
<td>Classify the metabolic pathways of test compounds</td>
<td>RF</td>
<td>Whole data set: Q = 94%</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>44</td>
<td>V_{max}, K_m</td>
<td>ANN</td>
<td>Whole data set: R² = 0.6-0.9 (K_m), R² = 0.6-0.7 (V_{max}), RMSE = 0.3-0.5 (K_m), RMSE = 0.4-0.7 (V_{max})</td>
<td></td>
</tr>
<tr>
<td>Sarigannis et al. (2017)</td>
<td>54</td>
<td>V_{max}, K_m</td>
<td>ANN, NLR</td>
<td>Test set: R² = 0.82 (K_m), R² = 0.99 (V_{max})</td>
<td></td>
</tr>
<tr>
<td><strong>Elimination</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsiao et al. (2013)</td>
<td>244</td>
<td>C_{tot}</td>
<td>PLS, RF, PCA</td>
<td>Whole data set: R² = 0.96; Q = 48%</td>
<td></td>
</tr>
<tr>
<td>Iwata et al. (2021)</td>
<td>748</td>
<td>C_{tot}</td>
<td>DL</td>
<td>Test data set: GMFE = 2.68</td>
<td></td>
</tr>
<tr>
<td>Kosugi and Hosea (2020)</td>
<td>1114</td>
<td>C_{tot}, 2D SMARTS-based descriptors</td>
<td>RF, RBF</td>
<td>Whole data set: R² = 0.55, RMSE = 0.332</td>
<td></td>
</tr>
<tr>
<td>Patne et al. (2010)</td>
<td>349</td>
<td>C_{tot}</td>
<td>RF</td>
<td>Training set: R² = 0.93, RMSE = 0.32</td>
<td></td>
</tr>
<tr>
<td>Paixao et al. (2010)</td>
<td>112</td>
<td>C_{tot}</td>
<td>ANN</td>
<td>Training set: R² = 0.953, RMSE = 0.236</td>
<td></td>
</tr>
<tr>
<td>Wang et al. (2019)</td>
<td>1352</td>
<td>C_{tot}</td>
<td>SVM, GBM, XGBoost</td>
<td>Test set: R² = 0.804, RMSE = 0.544</td>
<td></td>
</tr>
<tr>
<td>Gombar and Hall (2013)</td>
<td>525</td>
<td>C_{tot}</td>
<td>SVM, MLR</td>
<td>Test set: R² = 0.875, RMSE = 0.103, Q = 0.7</td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations: AAUE, absolute average fold error; AFE, absolute fold error; ANN, artificial neural networks; C_{tot}, intrinsic metabolic clearance; C_{renal}, renal clearance; C_{plasm}, total plasma clearance; DL, deep learning; DT, decision tree; GA, genetic algorithm; GBM, gradient boosting machine; GCN, graphical convolutional network; GMFE, geometric mean fold error; GNN, general neural network; GRNN, general regression neural network; F, oral bioavailability; HIA, human intestinal absorption; K_m, Michaelis constant; MARS, multivariate adaptive regression splines; MLR, multiple linear regression; NLR, nonlinear regression; Papp, apparent membrane permeability coefficient; PCA, principle component analysis; PLS, partial least squares; PNN, probabilistic neural network; Q, prediction accuracy; R², squared Pearson's correlation coefficient; RF, random forest; RMSE, root-mean-square error; SVM, support vector machine; V_{max}, maximal reaction rate; XGBoost, extreme Gradient Boosting.

*The performance from the best model.
### Table 3. A List of Representative Studies That Used Machine Learning and Artificial Intelligence Approaches in the Predictions of Toxicokinetic Parameters for Nonpharmaceutical Compounds

<table>
<thead>
<tr>
<th>References</th>
<th>N</th>
<th>Predict Target</th>
<th>Descriptor Types</th>
<th>Modeling Method</th>
<th>Performance&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wambaugh et al. (2015)</td>
<td>271</td>
<td>Transporter affinity</td>
<td>NA</td>
<td>RF</td>
<td>Training set: $R^2 = 0.82$; RMSE = 0.59; Test set: $R^2 = 0.51$; RMSE = 0.218</td>
</tr>
<tr>
<td>Ingle et al. (2016)</td>
<td>1651</td>
<td>Fub</td>
<td>2D molecular descriptors</td>
<td>kNN, SVM, RF</td>
<td>Whole data set: $R^2 = 0.728$; RMSE = 0.145</td>
</tr>
<tr>
<td>Watanabe et al. (2018)</td>
<td>2738</td>
<td>Fub</td>
<td>2D molecular descriptors</td>
<td>kNN, SVM, RF, PLS</td>
<td>Whole data set: $R^2 = 0.80$, RMSE = 0.62</td>
</tr>
<tr>
<td>Papa et al. (2018)</td>
<td>1000</td>
<td>$C_{int}$</td>
<td>2–3D molecular descriptors</td>
<td>PLS</td>
<td></td>
</tr>
<tr>
<td>Pradeep et al. (2020)</td>
<td>1487</td>
<td>Fub, $C_{int}$</td>
<td>0–3D molecular descriptors</td>
<td>SVM, RF, ANN</td>
<td>Fub: Training set: $R^2 = 0.56$, RMSE = 0.82; Test set: $R^2 = 0.57$, RMSE = 0.80</td>
</tr>
<tr>
<td>Dawson et al. (2021)</td>
<td>6484</td>
<td>Fub, $C_{int}$</td>
<td>1–3D molecular descriptors</td>
<td>RF</td>
<td>$C_{int}$: Training set: $R^2 = 0.99$, RMSE = 0.46; Test set: $R^2 = 0.16$, RMSE = 0.40</td>
</tr>
<tr>
<td>Yun et al. (2021)</td>
<td>818</td>
<td>Fub</td>
<td>2D molecular descriptors</td>
<td>kNN, SVM, RF, PLS</td>
<td>Fub: Training set: $R^2 = 0.584$, RMSE = 0.206; Test set: $R^2 = 0.591$, RMSE = 0.187</td>
</tr>
</tbody>
</table>

<sup>a</sup>The performance from the best model.

**Abbreviations:** ANN, artificial neural networks; $C_{int}$, intrinsic metabolic clearance; PLS, partial least squares; PNN, probabilistic neural network; Q, prediction accuracy; $R^2$, squared Pearson’s correlation coefficient; RF, random forest; RMSE, root mean square error; SVM, support vector machine.
### A List of Databases That Contains PK Data for Machine Learning Analysis

<table>
<thead>
<tr>
<th>Database name</th>
<th>Number of compounds</th>
<th>PK parameters</th>
<th>Description</th>
<th>Website</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK-DB</td>
<td>676</td>
<td>Cl, t$<em>{1/2}$, AUC, C$</em>{max}$, Kel and PK time-courses data</td>
<td>PK-DB is a comprehensive database, which contains data from human clinical trials and provides curated PK information on characteristics of studied patient cohorts, applied interventions, PK parameters, and PK time-courses data.</td>
<td><a href="https://pk-db.com">https://pk-db.com</a></td>
<td>Grzegorzekski et al. (2021)</td>
</tr>
<tr>
<td>PK/DB</td>
<td>1203</td>
<td>HIA, F, fu, BBB, Vd, Cl, t$_{1/2}$</td>
<td>PK/DB is a robust database for PK studies and in silico ADME prediction.</td>
<td><a href="http://www.pkdb.ifsc.usp.br">www.pkdb.ifsc.usp.br</a></td>
<td>Moda et al. (2008)</td>
</tr>
<tr>
<td>PKKB</td>
<td>1685</td>
<td>HIA, fu, Vd, Cl, LD50</td>
<td>Pharmacokinetic Knowledge Base (PKKB) is a comprehensive database of PK and toxic properties for drugs.</td>
<td><a href="http://cadd.suda.edu.cn/admet">http://cadd.suda.edu.cn/admet</a></td>
<td>Cao et al. (2012)</td>
</tr>
<tr>
<td>e-Drug3D</td>
<td>1852</td>
<td>Vd, Cl, t$<em>{1/2}$, PPB, F, C$</em>{max}$, and T$_{max}$</td>
<td>e-Drug3D is a database of 1852 FDA-approved drugs with 3-D chemical structures and information on PK parameters.</td>
<td><a href="https://cheminfo.ipmc.cnrs.fr/MOLDB/index.php">https://cheminfo.ipmc.cnrs.fr/MOLDB/index.php</a></td>
<td>Pihan et al. (2012)</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>&gt;1M</td>
<td>Not available</td>
<td>Open-access database containing ADME and toxic information for numerous drug-like compounds</td>
<td><a href="http://www.ebi.ac.uk/chembl/">www.ebi.ac.uk/chembl/</a></td>
<td>Gaulton et al. (2012)</td>
</tr>
<tr>
<td>Lombardo’s database</td>
<td>1352</td>
<td>Vd, Cl, MRT, fu, t$_{1/2}$</td>
<td>A human intravenous PK data set derived from the literature.</td>
<td>Not available</td>
<td>Lombardo et al. (2018)</td>
</tr>
<tr>
<td>Wang’s database</td>
<td>970</td>
<td>HIA</td>
<td>A human intestinal absorption data set consists of 970 compounds, and 9 different types of descriptors.</td>
<td>Not available</td>
<td>Wang et al. (2017)</td>
</tr>
<tr>
<td>CvT</td>
<td>144</td>
<td>PK time-course data</td>
<td>A public database of chemical time-series concentration data for 144 environmentally relevant chemicals and their metabolites</td>
<td><a href="https://github.com/USEPA/CompTox-PK-CvTdb">https://github.com/USEPA/CompTox-PK-CvTdb</a></td>
<td>Sayre et al. (2020)</td>
</tr>
</tbody>
</table>

**Abbreviations:** AUC, area under curve; BBB, blood brain barrier; Cl, clearance; Cmax, maximum concentration; F, oral bioavailability; fu, fraction unbound in plasma; HIA, human intestinal absorption; Kel, elimination rate; LD, lethal dose; MRT, mean residence time; PK, pharmacokinetic; PPB, plasma protein binding; t1/2, terminal half-life; Tmax, time to peak drug concentration; Vd, volume of distribution.

---

A List of Databases Relevant to Computational Toxicology

<table>
<thead>
<tr>
<th>Database</th>
<th>Data Size*</th>
<th>Data Type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACToR</td>
<td>Over 800 000 compounds and 500 000 assays</td>
<td>In vitro and in vivo toxicity</td>
<td>Judson et al. (2008)</td>
</tr>
<tr>
<td>Bioassays list</td>
<td>726 chemical pollutants</td>
<td>Concentration data in bioassays</td>
<td>Richman et al. (2022)</td>
</tr>
<tr>
<td>CERS</td>
<td>Over 1 000 compounds and 8000 studies</td>
<td>Gene expression data</td>
<td>Les et al. (2017)</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>1.1 million bioassays, 1.8 million compounds, over 10 million activities</td>
<td>Literature data on binding, function, and toxicity of drugs and drug-like chemicals</td>
<td>Gauton et al. (2012)</td>
</tr>
<tr>
<td>Connectivity map</td>
<td>Around 1300 compounds and 7000 genes</td>
<td>Gene expression data</td>
<td>Subramanian et al. (2017)</td>
</tr>
<tr>
<td>CTD</td>
<td>Over 14 000 compounds, 42 000 genes, 6000 diseases</td>
<td>Relationships among compounds, genes, and diseases</td>
<td>Davis et al. (2021)</td>
</tr>
<tr>
<td>DrugMatrix</td>
<td>Around 600 drug molecules and 10 000 genes</td>
<td>Gene expression data</td>
<td>Gartner et al. (2005)</td>
</tr>
<tr>
<td>GID</td>
<td>Over 4300 subdata sets</td>
<td>Microarray, next-generation sequencing, and other forms of high-throughput functional genomics data</td>
<td>Barrett et al. (2013)</td>
</tr>
<tr>
<td>eNanoMapper</td>
<td>Over 700 types of nanomaterials</td>
<td>Diverse data types on nanomaterials physicochemical properties and safety</td>
<td>Jeliazkova et al. (2015)</td>
</tr>
<tr>
<td>MoleculeNet</td>
<td>Over 700 000 compounds</td>
<td>Quantum mechanics, physical chemistry, biophysics, and physiology</td>
<td>Wu et al. (2018)</td>
</tr>
<tr>
<td>Open TG-GATEs</td>
<td>170 compounds</td>
<td>Gene expression data and metadata</td>
<td>Igarashi et al. (2015)</td>
</tr>
<tr>
<td>PubChem</td>
<td>Over 77 million compounds, 1.39 million bioassays, and 293 million bioactivity data points</td>
<td>Toxicology, genomics, pharmacology, and literature data</td>
<td>Kim et al. (2021)</td>
</tr>
<tr>
<td>Pubvinas</td>
<td>11 types of nanomaterials with 705 unique nanomaterials</td>
<td>Up to 6 physicochemical properties and/or bioactivities</td>
<td>Yan et al. (2020)</td>
</tr>
<tr>
<td>REACH</td>
<td>21,405 unique substances with information from 89,905 dossiers</td>
<td>Data submitted in European Union chemical legislation</td>
<td>Luechtefeld et al. (2016)</td>
</tr>
<tr>
<td>RepDose</td>
<td>364 compounds investigated in 1017 studies, resulting in 6,002 specific effects</td>
<td>Repeat-dose study data in dogs, mice, and rats</td>
<td>Bitche et al. (2006)</td>
</tr>
<tr>
<td>SIURAT</td>
<td>Over 5500 cosmetic-type compounds in the current COSMOS database web portal</td>
<td>Animal toxicity data</td>
<td>Vincken et al. (2012)</td>
</tr>
<tr>
<td>Toxicod</td>
<td>231 chemicals</td>
<td>Toxicogenomic data</td>
<td>Nair et al. (2020)</td>
</tr>
<tr>
<td>Toxnet</td>
<td>Over 50 000 environmental chemicals from 56 resources</td>
<td>In vitro and in vivo toxicity data</td>
<td>Fouger et al. (2008)</td>
</tr>
</tbody>
</table>

*On the basis of live web counts or most recent literature publications as of March 2022. ACToR, Aggregated Computational Toxicology Resource; CTD, Comparative Toxicogenomics Database; CEBS, Chemical Effects in Biological Systems; GEO, Gene Expression Omnibus; Open TG-GATEs, a large-scale toxicogenomic database; REACH, Registration, Evaluation, Authorization, and Restriction of Chemicals; SEURAT, Safety Evaluation Ultimately Replacing Animal Testing; ToxNET, Toxicology Data Network. This table was adapted from Ciallella and Zhu (2019) with permission from the publisher.
Case Study 1: AI in Predicting ADME Properties
Background, Objective, and Rationale of Case Study 1

**Background**

- ~94.4 million
- ~72.9 million
- ~5.23 million
- ~2.62 million
- ~8.54 billion
- ~238 million

**Objective and Rationale**

- To develop models to predict the withdrawal time of drugs following extralabel use in food animals in order to protect safety of animal-derived food products, such as meat, milk, and eggs.

**Research Program**

Food Animal Residue Avoidance Databank
(A component of the Food Animal Residue Avoidance & Depletion Program)

Overview and Timeline of Our PK/PBPK Models (KSU + UF)

2014-2016
- Established methodology
- Created PBPK models for drugs in an average animal
- Ceftiofur, enrofloxacin, flunixin, sulfamethazine
- Swine and Cattle

Lin et al. 2015. J Pharm Sci

2016-2018
- Improved the methodology
- Monte Carlo simulation
- Created PBPK models for drugs in a diverse population of animals
- Penicillin G
- Swine, beef cattle, dairy cows

Lin et al. 2017. Toxicol Sci
Li et al. 2017. Food Chem Toxicol
Li et al. 2018. Toxicol Sci

2018-2022
- Graphical user interface (GUI)
- Population PBPK models
- Penicillin G, flunixin, florfenicol, and oxytetracycline
- Physiological parameter database: cattle, swine, chickens, turkeys, sheep, goats
- Other quantitative methods from FDA & EMA

Li et al. 2019. Arch Toxicol

Smith et al. 2020. Front Vet Sci
Riad et al. 2021. Toxicol Sci
Chou et al. 2022. Toxicol Sci
Yuan et al. 2022. Food Chem Toxicol
Yuan et al. 2022. RTP
Role of AI and PBPK in Animal-Derived Food Safety Assessment

- Long-term: Integration of AI with PBPK and/or QSAR/QSPR to predict PK properties of drugs
- Short-term: Build an AI-QSAR model to predict plasma half-life of animal drugs

Extract Pharmacokinetic (PK) Data
- Plasma and tissue half-lives
- Clearance
- Other pharmacokinetic parameters
- Dosing regimens

Extract Cheminformatics data
- Molecular descriptors
- Fingerprints

Data Processing
Input layer: All data except half-lives
Output layer: Half-lives

Machine Learning and Artificial Intelligence Methods
- Long-term: Integration of AI with PBPK and/or QSAR/QSPR to predict PK properties of drugs
- Short-term: Build an AI-QSAR model to predict plasma half-life of animal drugs

QSAR: Quantitative structure-activity relationships
QSPR: Quantitative structure-property relationships
Schematic Workflow of AI-based QSAR Modeling

Wu PY, et al., unpublished results from the Lin Lab at UF.
Wu PY, et al., unpublished results from the Lin Lab at UF.
Case Study 2: AI in Predicting Toxicity
Case Study 2: Machine Learning Models to Predict Chemical Carcinogenicity

- Carcinogenicity testing plays an important role in identifying carcinogens in drug development and environmental chemical risk assessment.

- Traditionally, the carcinogenic potency is evaluated with a 2-year carcinogenicity study in rodents, but this process is very time-consuming and resource-intensive.

- Chemical carcinogenicity assessment is required to be conducted in at least 2 species.
Case Study 2: Machine Learning Models to Predict Chemical Carcinogenicity

- DeepCarc model to predict carcinogenicity for small molecules using deep learning-based model-level representations. The DeepCarc model was developed with a dataset of 692 chemicals and evaluated with a test set consisting of 171 chemicals.

- The data were obtained from the National Center for Toxicological Research liver cancer database and involved both rats and mice.

- The authors also compared performance of the DeepCarc model with other deep learning models that were based on molecule-level representations, including Text Convolutional neural network from DeepChem, Convolutional Neural Network Fingerprint, Edge Attention-based Multi-relational Graph Convolutional Networks, and Chemistry Chainer-Neural Fingerprint.

- This DeepCarc model provides an early screening nonanimal-based tool to assess potential carcinogenicity of new chemicals and is useful for prioritizing chemicals on their potential carcinogenic risk.
TABLE 2 | The model performance of DeepCarc and four advanced DNN models on the test set.

<table>
<thead>
<tr>
<th>Models</th>
<th>MCC</th>
<th>Accuracy</th>
<th>AUC</th>
<th>F1</th>
<th>BA</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepCarc</td>
<td>0.432</td>
<td>0.754</td>
<td>0.776</td>
<td>0.828</td>
<td>0.688</td>
<td>0.910</td>
<td>0.467</td>
</tr>
<tr>
<td>DC-TEXTCNN</td>
<td>0.392</td>
<td>0.735</td>
<td>0.719</td>
<td>0.829</td>
<td>0.627</td>
<td>0.982</td>
<td>0.271</td>
</tr>
<tr>
<td>CH-NFP</td>
<td>0.353</td>
<td>0.725</td>
<td>0.776</td>
<td>0.814</td>
<td>0.639</td>
<td>0.928</td>
<td>0.350</td>
</tr>
<tr>
<td>EAGCNG</td>
<td>0.328</td>
<td>0.713</td>
<td>0.682</td>
<td>0.800</td>
<td>0.641</td>
<td>0.883</td>
<td>0.400</td>
</tr>
<tr>
<td>CNF</td>
<td>0.185</td>
<td>0.673</td>
<td>0.636</td>
<td>0.796</td>
<td>0.541</td>
<td>0.982</td>
<td>0.100</td>
</tr>
</tbody>
</table>
Case Study 3: AI-assisted PBPK Model for Nanoparticle Risk Assessment Cancer Nanomedicine
Case Study 3: AI-assisted PBPK Model for Nanoparticles

Delivery efficiency of NPs to tumors based on studies published each year

Critical barriers to progress in this field

- Nanotoxicology: lack of robust computational tools to assess risk
- Nanomedicine: low delivery efficiency (<1%) to target tissues (i.e., tumor)

Case Study 3: AI-assisted PBPK Model for Nanoparticles

PBPK Structure in tumor-bearing mice

Note: other cell types not shown here

Nano-Tumor Database

- Literature search
  - Database: PubMed
  - Time range: 2015/1/1 – 2018/9/4
  - Keywords:
    - Nanoparticle delivery
    - Nanomaterial delivery
    - Biodistribution; Pharmacokinetics
    - Mice; Rats
    - Tumor; Tumour
  - Language: English
  - Document type: Peer-reviewed journals
- Included (n = 200)
  - CNR database [90]
  - Newly incorporated (110)
- Excluded (n = 938)
  - Duplicate with CNR database
  - Other nanomaterials not used for cancer therapeutics or tumor delivery
  - e.g., vaccine
  - Other administration routes, e.g., oral, intraperitoneal, subcutaneous, intratracheal, intratumoral, and etc.
  - Biodistribution data not reported in the units of μg/g, %ID/g or %ID
  - Tumor-bearing animals other than rodents (mice or rats)
  - Pharmacokinetic or biodistribution data from healthy rodents
- Included (n = 393)
  - Articles screened on the basis of title and abstract
  - Manuscript review and application of inclusion criteria
- Combined results (n = 1331)

Note: currently, this database contains 376 datasets from 200 studies published from 2005 to 2018.

Representative Results in Tumor-Bearing Mouse Model

Case Study 3: AI-assisted PBPK Model for Nanoparticles


Summary and Discussion

• By leveraging machine learning and artificial intelligence approaches, now it is possible to:

  (1) Develop in silico models to predict ADME properties of hundreds of chemicals with acceptable accuracies;
  (2) Develop PBPK models for hundreds of chemicals efficiently;
  (3) Create in silico models to predict toxicity for a large number of chemicals with similar accuracies compared with in vivo animal experiments;
  (4) Analyze a large amount of different types of data (toxicogenomics, high-content image data, etc.) to generate new insights into toxicity mechanisms rapidly, which was difficult by manual approaches in the past.

• Several challenges should be considered:

  (1) Not all machine learning models are equally useful for a particular type of toxicology data, and thus it is important to test different methods to determine the optimal approach;
  (2) Current toxicity prediction is mainly on bioactivity classification (yes/no), so additional studies are needed to predict the intensity of effect or dose-response relationship;
  (3) As more data become available, it is crucial to perform rigorous data quality check and develop infrastructure to store, share, analyze, evaluate, and manage big data;
  (4) It is important to convert machine learning models to user-friendly interfaces to facilitate their applications by both computational and bench scientists.
Lab members:
Zhoumeng Lin
Wei-Chun Chou
Qiran Chen
Xue Wu
Malek Hussein Hajjawi
Pei-Yu Wu
Chi-Yun Chen
Venkata Nithin Kamineni
Yashas Kuchimanchi
Zhicheng Zhang

Former members:
Miao Li
Yi-Hsien Cheng
Long Yuan
Md Mahbubul Huq Riad
Dongping Zeng
Trevor Elwell-Cuddy
Paula Solar; Sichao Mao
Yilei Zheng; Yi-Jun Lin
Ning Xu; Yu Shin Wang
Jake Willson
Gabriel (Guanyu) Tao

Collaborators:
ICCM/NICKS/KSU
EGH/CEHT/UF
CPSP/UF
FARAD Team

Advisors:
Dr. Jim E. Riviere
Dr. Nancy A. Monteiro-Riviere
Dr. Nikolay M. Filipov
Dr. Jeffrey W. Fisher
Dr. Ronette Gehring

Active Funding:
- NIH/NIBIB Grant #: R01EB031022
- USDA/NIFA Award #: 2023-41480-41035
- USDA/NIFA Award #: 2022-41480-38137
- USDA/NIFA Award #: 2021-41480-35271
- USDA/NIFA Award #: 2021-67015-34084
- UF PHHP PhD Fellowship in Artificial Intelligence

Completed Funding:
- USDA/NIFA Award #: 2020-41480-32497
- USDA/NIFA Award #: 2020-67015-31456
- USDA/NIFA Award #: 2020-67030-31479
- CHOP subaward #: FP37698_SUB01_01
- NIH/NIBIB Grant #: R03EB026045
- USDA/NIFA Award #: 2019-41480-30296
- USDA/NIFA Subaward #: A20-2028-S002
- NIH/NIBIB Grant #: R03EB025566
- USDA/NIFA Award #: 2018-41480-28805
- USDA/NIFA Award #: 2017-68003-26499
- USDA/NIFA Award #: 2017-41480-27310
- USDA/NIFA Award #: 2016-41480-25729
- AASV Foundation Grant Award #: A00-1103-001
- K-State CVM SUCCESS-FYI Program
- K-State Mark Derrick Canine Research Fund
- K-State Global Campus Internal Grant Program
- K-State Mentoring Fellowship
- K-State University Small Research Grant (USRG)