UF College of Public Health and Health Professions
Department of Environmental and Global Health
UNIVERSITY of FLORIDA

CPSP UF CENTER FOR PHARMACOMETRICS & SYSTEMS PHARMACOLOGY

UF Center for Environmental & Human Toxicology
College of Veterinary Medicine

Applications of machine learning and AI approaches to develop PBPK and QSAR models to predict ADMET properties to aid chemical safety assessment

---- Workshop on "Integrating AI into Chemical Safety Assessment", October 16-17, 2024, Sophia Antipolis

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

- Overview of the application of machine learning and AI approaches in the field of toxicological sciences, especially in PBPK modeling

ADMET

- Applications of machine learning and AI approaches to predict ADMET properties of chemicals

PBPK

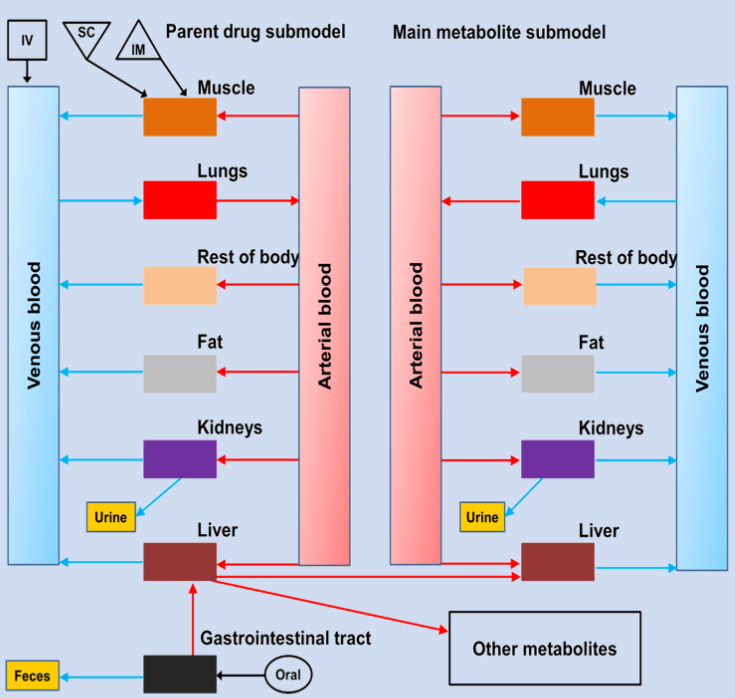
- Applications of machine learning and AI approaches to develop PBPK models for xenobiotics in a high-throughput manner

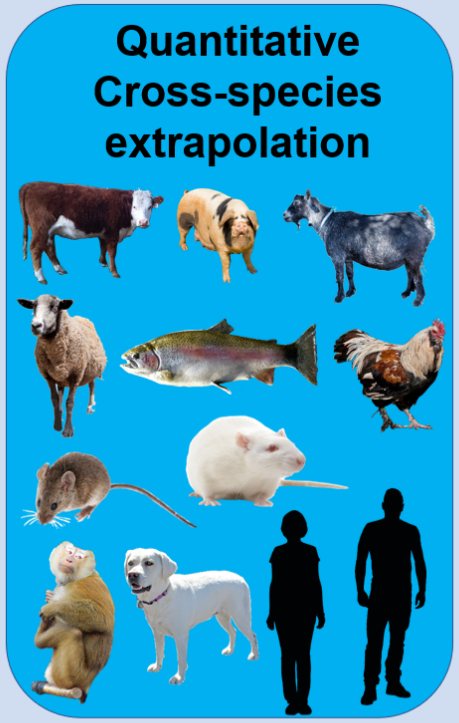
Conclusion

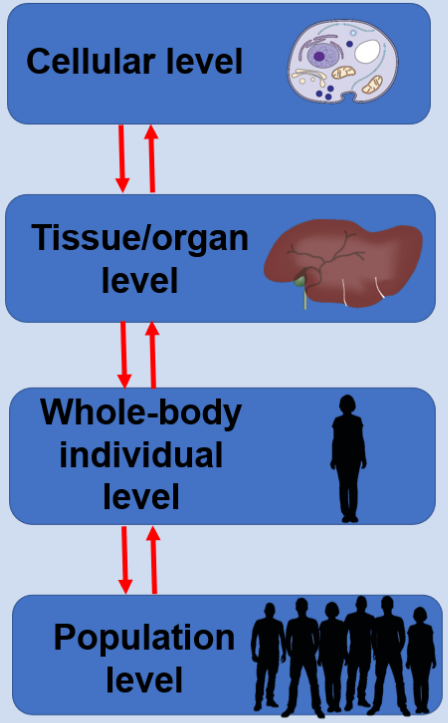
- Summary and Discussion
- Acknowledgements

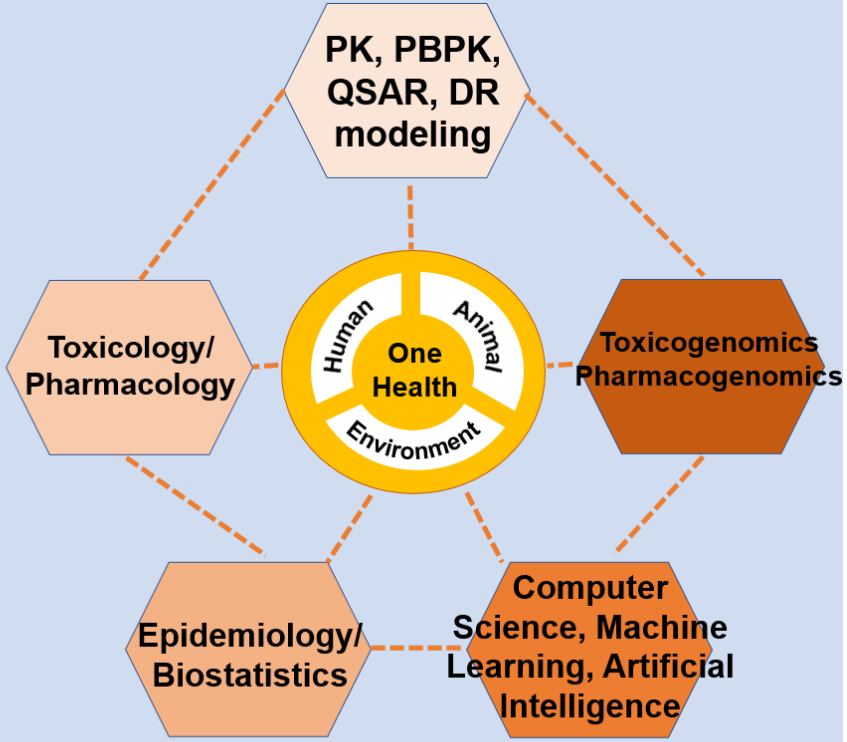
ADMET: Absorption, distribution, metabolism, excretion, and toxicity
PBPK: Physiologically based pharmacokinetic modeling

Research Program in Computational Toxicology and Pharmacology







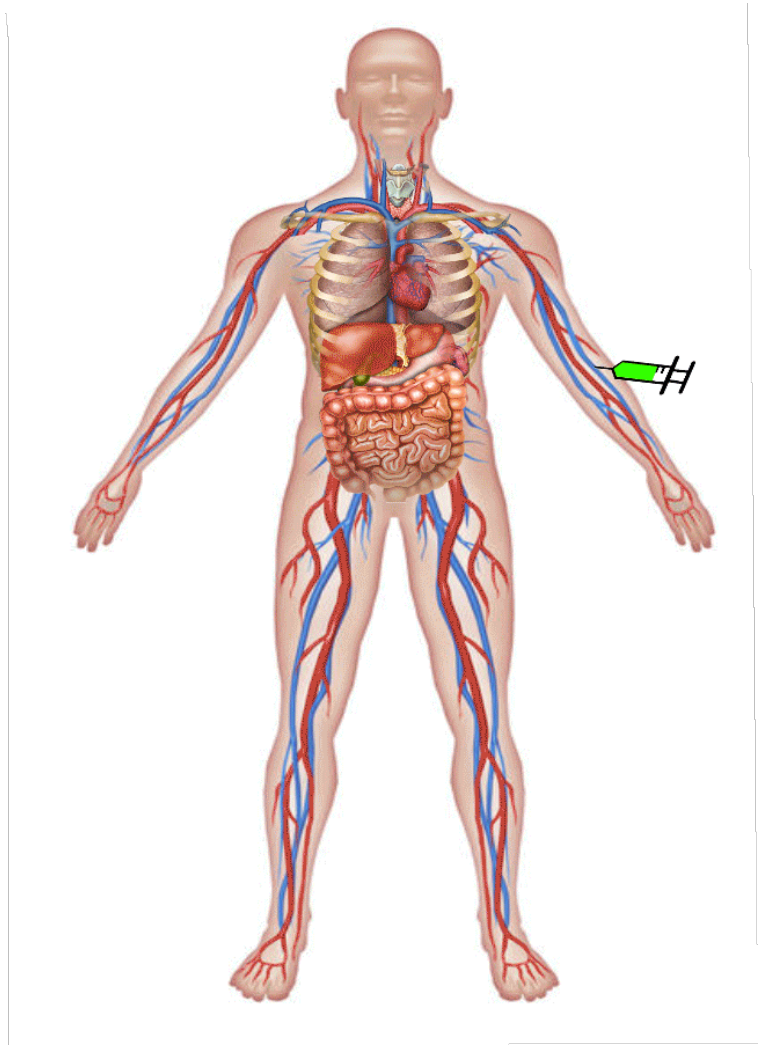


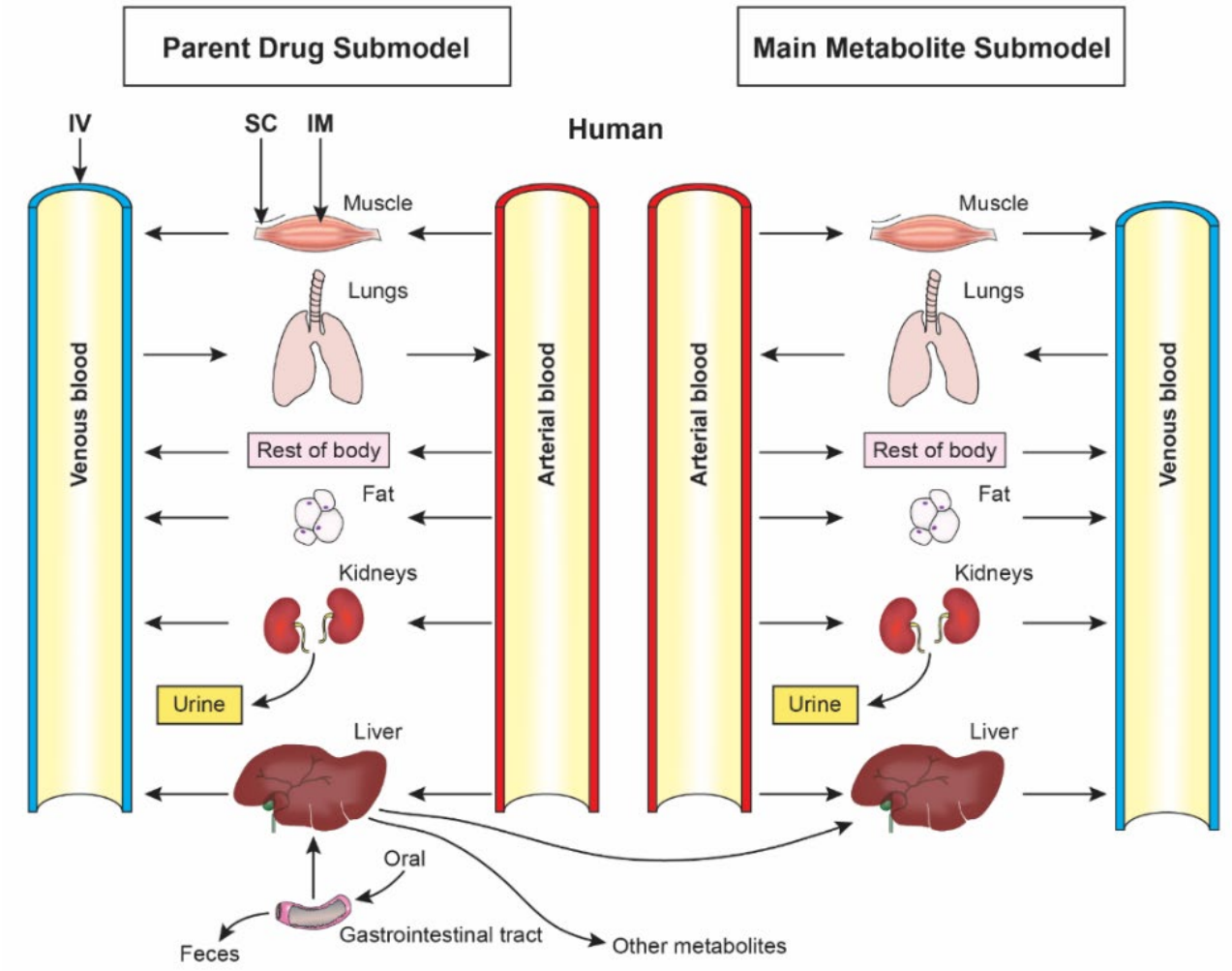
We develop computational models to address nanomedicine, animal-derived food safety, human health risk assessment issues!

- Nanomaterials
- Microplastics & nanoplastics
- Environmental chemicals
- Drugs

PBPK: Physiologically based pharmacokinetic modeling
QSAR: Quantitative structure-activity relationship modeling

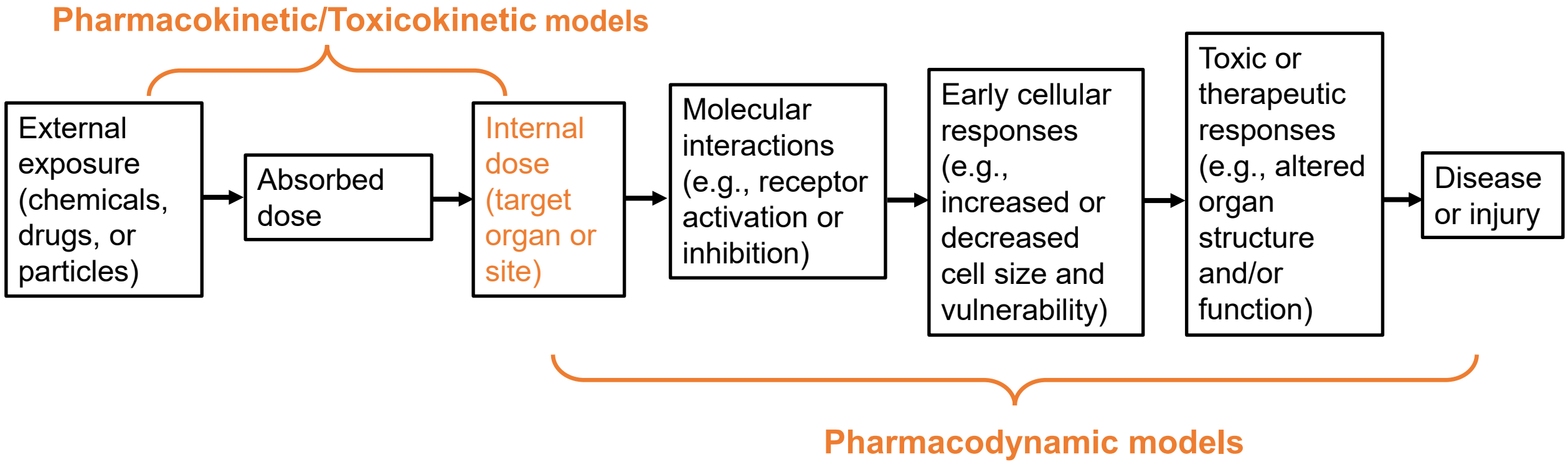
What is Physiologically Based Pharmacokinetic (PBPK) modeling?





Why is PBPK modeling important?

- From the Perspective of Toxicokinetics and Risk Assessment:



- A fundamental tenet: both beneficial and adverse responses to compounds are related to the concentrations of active chemicals reaching target tissues rather than the amounts of chemicals at the site of absorption

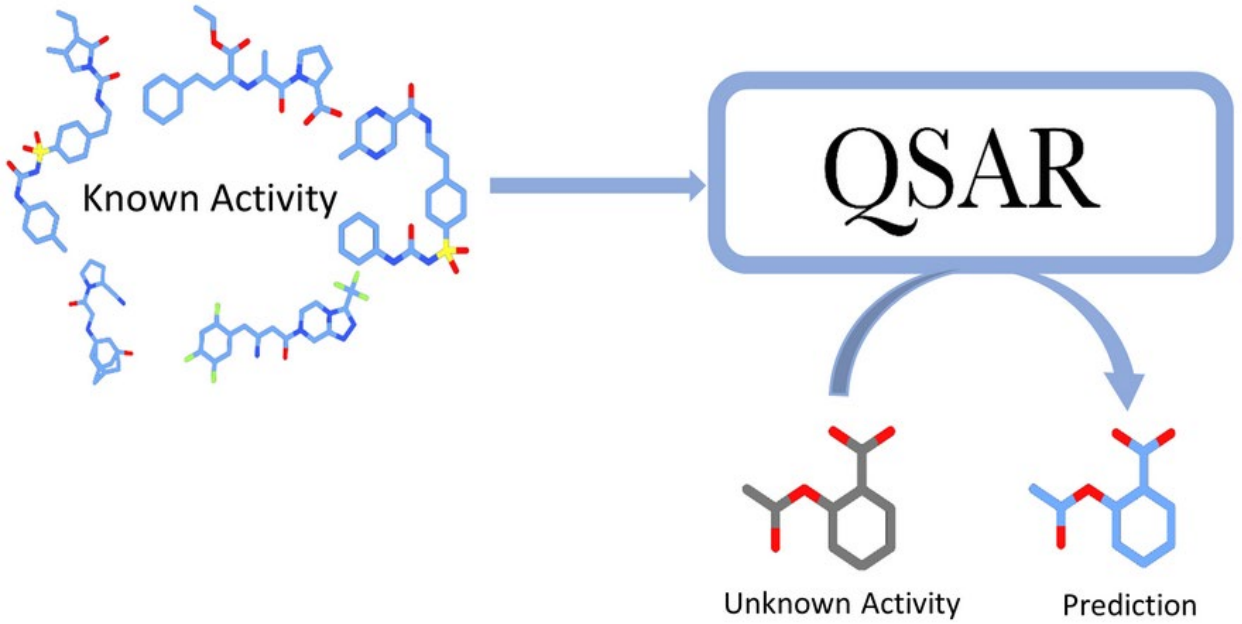
- Need a tool that can relate internal concentrations of active compounds at their target sites with the external doses of chemicals that an animal or human is exposed to

What is Quantitative Structure-Activity Relationship (QSAR) modeling?

- Quantitative structure activity relationship analysis (QSAR): the study of the relationship between chemical structure and biological properties of substances.
- These activities include absorption, distribution, metabolism, and excretion, as well as toxicity (ADMET) properties.

OECD, 2017. https://www.oecd-ilibrary.org/environment/fundamental-and-guiding-principles-for-q-sar-analysis-of-chemical-carcinogens-with-mechanistic-considerations_9789264274792-en

Why is QSAR modeling important?



- 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.
- Toxicology and risk assessment
- Drug discovery and development
- Screening, activity ranking, and prioritization of chemicals
- Alternative to animal experimentation

Image credit: https://protoqsar.com/en/qsar-models/

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.

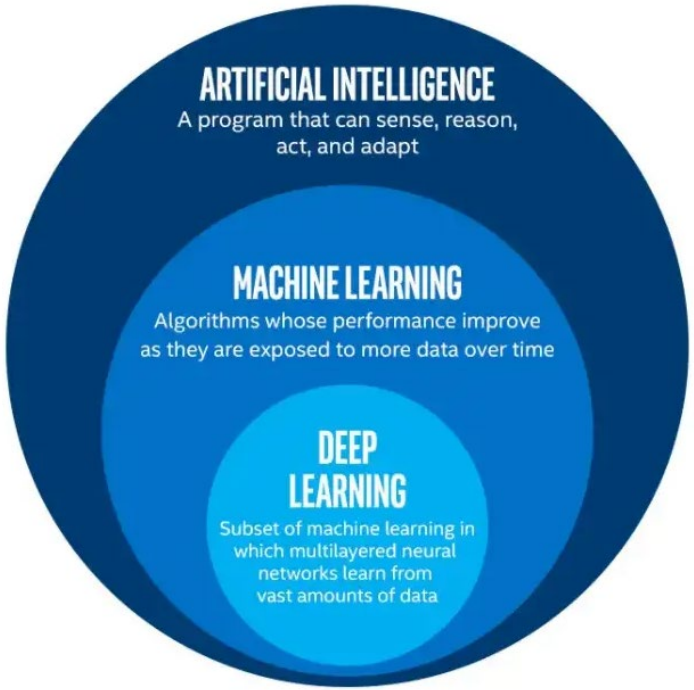


Image source: https://towardsdatascience.com/cousins-of-artificial-intelligence-dda4edc27b55

Applications of ML and AI in different subject areas of toxicology

CONTEMPORARY REVIEW

Machine Learning and Artificial Intelligence in Toxicological Sciences

Zhoumeng Lin *,†,1 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

[1]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.

OXFORD SOT Society of Toxicology academic.oup.com/toxsci

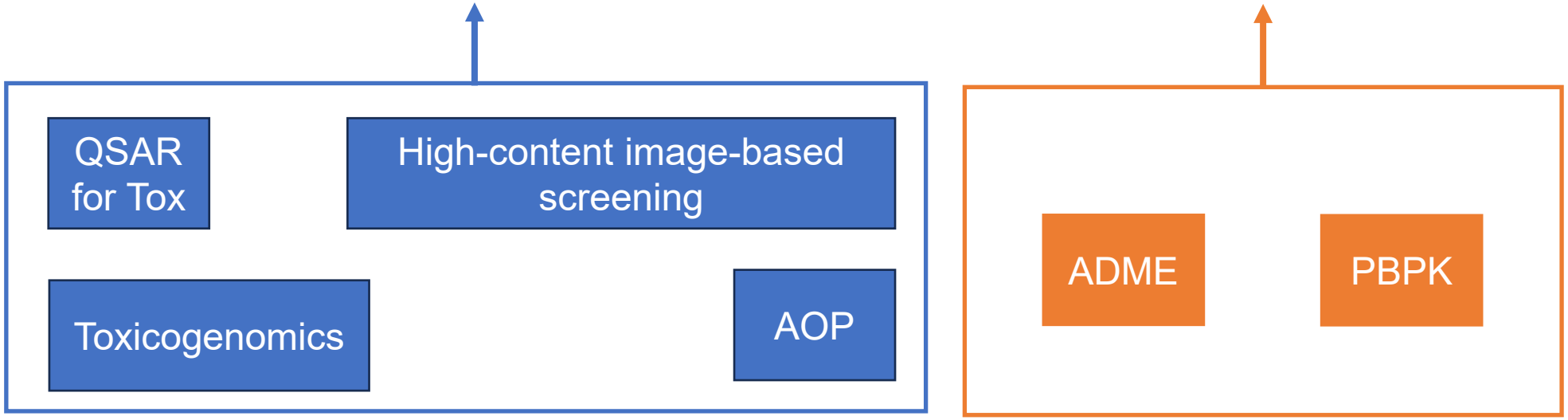
Toxicological Sciences, 2023, **191(1)**, 1–14
https://doi.org/10.1093/toxsci/kfac101
Advance Access Publication Date: 26 September 2022
Contemporary Review

Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling

Wei-Chun Chou [1,2], Zhoumeng Lin [1,2,*]

[1]Department of Environmental and Global Health, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610, USA
[2]Center for Environmental and Human Toxicology, University of Florida, Gainesville, FL 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.



Lin Z, Chou WC. (2022). Machine learning and artificial intelligence in toxicological sciences. *Toxicological Sciences*, 189(1):7-19.

Chou WC, Lin Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences*, 191(1):1-14.

Commonly used machine learning methods in toxicology

Table 1. A List of Machine Learning Methods Commonly Used in Toxicological Research

Method	Brief Description
Supervised linear methods	
Multiple linear regression	Use multiple explanatory variables to predict the outcome of a response variable with a multivariate linear equation
Naïve Bayes classifier	Based on Bayes' theorem with strong assumptions of conditional independence among molecular descriptors (ie, explanatory variables)
Supervised nonlinear methods	
k-nearest neighbors	Classify a test chemical by looking for the training chemicals with the nearest distance to it
Support vector machine	Map molecular descriptor vectors into a higher dimensional feature space to build a maximal margin hyperplane to distinguish active (toxic) from inactive (nontoxic) chemicals
Decision trees	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
Ensemble learning	Combine several base models into a more predictive one. Popular types of ensemble modeling include bagging, random spaces, boosting, and stacking.
Random forest	Combine the bagging with the random spaces approaches in application to decision trees base models
Artificial neural networks	
Backpropagation neural networks	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
Bayesian-regularized neural networks	Apply Bayesian methods to perform regularization so that the model complexity is balanced against the accuracy of reproducing training data
Associative neural networks	Apply ensemble learning to backpropagation neural networks
Deep neural networks	Artificial neural networks with multiple hidden layers (also called deep learning)
Unsupervised methods	
Principle component analysis	Reduce the dimensionality of the data to only the first few principal components while preserving as much of the data's variation as possible
Kohonen's self-organizing maps	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

This table is based on the book chapter by Baskin (2018). Please refer to Baskin (2018) for detailed description about each of the listed machine learning algorithms.

Lin Z, Chou WC. (2022). Machine learning and artificial intelligence in toxicological sciences. *Toxicological Sciences*, 189(1):7-19.

List of studies using ML in QSAR modeling to predict toxicity

Table 2. Representative Studies Integrating Machine Learning Approaches With Quantitative Structure-Activity Relationship Modeling

Best Machine learning Method	Training Dataset	Endpoint	Reference
Deep learning (ie, DeepTox)	11 764 chemicals from Tox21	12 bioassays	Mayr *et al.* (2016)
Ensemble extreme gradient boosting	1003 chemicals	Carcinogenicity	Zhang *et al.* (2017)
Random forest	Over 866 000 chemical properties/hazards	Acute oral and dermal toxicity, eye and skin irritation, mutagenicity, and skin sensitization	Luechtefeld *et al.* (2018)
Ensemble support vector machine	400 chemicals	Aquatic acute toxicity	Ai *et al.* (2019)
Multitask neural networks and graph convolutional networks	1012 PFAS	Bioactivity on 26 bioassays	Cheng and Ng (2019)
Extra trees	Over 1000 chemicals from different databases	Various toxicities	Pu *et al.* (2019)
Ensemble model	7385 chemicals	Acute toxicity in rats	Russo *et al.* (2019)
Support vector machine	482 chemicals	Acute toxicity in fathead minnow	Chen *et al.* (2020)
Deep learning (ie, CapsCarcino)	1003 chemicals from CPDB	Carcinogenicity	Wang *et al.* (2020)
Kernel-weighted local polynomial approach	Hundreds of chemicals depending on the species	Acute aquatic toxicity	Gajewicz-Skretna *et al.* (2021)
Meta ensembling of multitask deep learning models (ie, QuantitativeTox)	Hundreds to thousands of compounds depending on the endpoint	LD_{50} and LC_{50}	Karim *et al.* (2021)
Deep learning-based model-level representations (ie, DeepCarc)	692 chemicals	Carcinogenicity	Li *et al.* (2021)
Extra trees	Over 18 600 drug-bacteria interactions	Gut bacterial growth	McCoubrey *et al.* (2021)
Support vector machine	676 pesticides	Acute contact toxicity on honey bees	Xu *et al.* (2021)
A consensus model based on 4 algorithms	1244 chemicals	Prenatal developmental toxicity	Ciallella *et al.* (2022)
Deep learning	31 chemicals with known or suspected clinical skin toxicity	Skin toxicity	Hu *et al.* (2022)
Random forest	1476 food contact chemicals	Carcinogenicity	Wang *et al.* (2022)

CPDB, Carcinogenic Potency Database. LC_{50} and LD_{50} refer to the compound concentrations that kill half the members of the tested animal population, respectively.

Lin Z, Chou WC. (2022). Machine learning and artificial intelligence in toxicological sciences. *Toxicological Sciences*, 189(1):7-19.

Studies that used ML/AI to predict ADME for **Pharmaceutical** compounds

Table 2. A List of Representative Studies That Used Machine Learning and Artificial Intelligence Approaches in the Predictions of Absorption, Distribution, Metabolism, and Excretion Properties for Pharmaceutical Compounds

References	N	Predict Target	Descriptor Types	Modeling Method	Performance[a]
Absorption					
Agatonovic-Kustrin *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
Deconinck *et al.* (2007)	67	HIA	1D–3D theoretical descriptors plus one of Abraham's solvation parameters	MARS	Whole data set: RMSE = 7.2%; Whole data set: $R^2 = 0.93$
Niwa (2003)	86	HIA	0D–1D theoretical descriptors	GRNN, PNN	Training set: RMSE = 6.5 Test set: RMSE = 22.8
Talevi *et al.* (2011)	120	HIA	0D–3D Dragon theoretical descriptors	MLR, ANN, SVM	Training set: R^2= 0.8; RMSE = 0.18 Test set: R^2= 0.66; RMSE = 0.21
Yan *et al.* (2008)	52	HIA	Adriana Code and Cerius2 0D–2D theoretical descriptors	GA, PLS, SVM	Training set: R^2= 0.66; RMSE = 12.5 Test set: R^2= 0.77; RMSE = 16
Shen *et al.* (2010)	1593	HIA	1D–2D theoretical descriptors	SVM	Training set: Q = 98.5% Test set: Q = 99%
Kamiya *et al.* (2021b)	184	P_{app}	Chemical descriptors (not specific descriptions)	SVM, PLS, RBF	Whole data set: R = 0.84–0.85
Ghafourian *et al.* (2012)	310	HIA	A total of 215 descriptors (not specific descriptions)	MLR	Training set: RMSE = 14.54 Test set: RMSE = 23.84
Hou *et al.* (2007)	648	HIA	0D–2D theoretical descriptors	MARS, GA	Training set: R^2= 0.97.3 Test set: R^2= 0.98
Wang *et al.* (2017)	970	HIA	2D–3D descriptors, molecular fingerprints, and structural fragments	RF	Training set: SE = 0.89; SP = 0.85; Q = 0.89 Test set: SE = 0.88; SP = 0.81; Q = 0.87
Distribution					
Antontsev *et al.* (2021)	21	Kp	Not explained in the study	BIOiSIM	Test set: AFE = 0.96 (C_{max}), 0.89 (AUC), 0.69 (Vd); AAFE = 1.2 (C_{max}), 1.30 (AUC), 1.71 (Vd); R^2 = 0.99 (C_{max}), 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
Liu *et al.* (2005)	208	Kp	Constitutional, topological, geometrical, electrostatic and quantum chemical descriptors	SVM	Training set: $R^2 = 0.97$, RMSE = 0.02 Test set: $R^2 = 0.974$, RMSE = 0.0289
Yun *et al.* (2014)	122	Kp	LogP, pKa, fu	DT; RF	Whole data set: Q = 72%

Chou WC, Lin Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences*, 191(1):1-14.

Studies that used ML/AI to predict ADME for Pharmaceutical compounds

Table 2.. (continued)

References	N	Predict Target	Descriptor Types	Modeling Method	Performance[a]
Metabolic					
Athersuch et al. (2013)	15		Classify the metabolic pathways of test compounds	PCA, PLS	Whole data set: $R^2 = 0.96$, Q = 77.5%
Baranwal et al. (2020)	6669		Classify the metabolic pathways of test compounds	RF and GCN	Test set: Q = 98.99%
Jia et al. (2020)	5682		Classify the metabolic pathways of test compounds	RF	Whole data set: Q = 94%
Zhang et al. (2008)	44	V_{max}, K_m	Molecular fingerprints	ANN	Whole data set: $R^2 = 0.6$–0.9 (K_m), $R^2 = 0.6$–0.7 (V_{max}), RMSE = 0.3–0.5 (K_m), RMSE = 0.4–0.7 (V_{max})
Sarigiannis et al. (2017)	54	V_{max}, K_m	Physicochemical properties based on Abraham's solvation equation	ANN, NLR	Test set: $R^2 = 0.82$ (K_m), $R^2 = 0.99$ (V_{max})
Elimination					
Hsiao et al. (2013)	244	Cl_{int}	Molecular fingerprints, physicochemical properties, and 3D quantum chemical descriptors	PLS, RF, PCA	Whole data set: $R^2 = 0.96$; Q = 48%
Iwata et al. (2021)	748	Cl_{total}	The chemical structure was represented as graph data	DL	Test data set: GMFE = 2.68
Kosugi and Hosea (2020)	1114	Cl_{total}	2D SMARTS-based descriptors	RF, RBF	Whole data set: $R^2 = 0.55$, RMSE= 0.332
Paine et al. (2010)	349	Cl_{renal}	195 descriptors	RF	Training set: $R^2 = 0.93$, RMSE = 0.32 Test set: $R^2 = 0.63$, RMSE = 0.63
Paixao et al. (2010)	112	Cl_{int}	233 molecular descriptors	ANN	Training set: $R^2 = 0.953$, RMSE = 0.236 Test set: $R^2 = 0.804$, RMSE = 0.544
Wang et al. (2019)	1352	Cl_{total}	2D and 3D descriptors, and 49 fingerprints.	SVM, GBM, XGBoost	Training set: $R^2 = 0.882$, RMSE = 0.239 Test set: $R^2 = 0.875$, RMSE = 0.103
Gombar and Hall (2013)	525	Cl_{tot}	89 descriptors calculated from electro-topological state (E-state) fingerprints	SVM, MLR	Test set: $R^2 = 0.70$

Abbreviations: AAFE, absolute average fold error; AFE, absolute fold error; ANN, artificial neural networks; Cl_{int}, intrinsic metabolic clearance; Cl_{renal}, renal clearance; Cl_{total}, total plasma clearance; DL, deep learning; DT, decision tree; GA, generic algorithm; GBM, gradient boosting machine; GCN, graphical conventional 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 coefficients; PCA, principle component analysis; PLS, partial least squares; PNN, probabilistic neural network; Q, prediction accuracy; R^2, squared Pearson's correlation coefficient; RBF, radial basis function; RF, random forest; RMSE, root-mean-square error; SVM, support vector machine; V_{max}, maximal reaction rate; XGBoost, eXtreme Gradient Boosting.

[a]The performance from the best model.

Chou WC, Lin Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences*, 191(1):1-14.

Studies that used ML/AI to predict ADME for Nonpharmaceutical compounds

Table 3. A List of Representative Studies That Used Machine Learning and Artificial Intelligence Approaches in the Predictions of Toxicokinetic Parameters for Nonpharmaceutical Compounds

References	N	Predict Target	Descriptor Types	Modeling Method	Performance[a]
Wambaugh *et al.* (2015)	271	Transporter affinity	NA	RF	NA
Ingle *et al.* (2016)	1651	Fub	2D molecular descriptors	kNN, SVM, RF	Training set: $R^2 = 0.82$; RMSE = 0.59 Test set: $R^2 = 0.51$; RMSE = 0.218
Watanabe *et al.* (2018)	2738	Fub	2D molecular descriptors	kNN, SVM, RF, PLS	Test set: $R^2 = 0.728$; RMSE = 0.145
Papa *et al.* (2018)	1000	Cl_{int}	2–3D molecular descriptors	PLS	Whole data set: $R^2 = 0.80$, RMSE= 0.62
Pradeep *et al.* (2020)	1487	Fub, Cl_{int}	0–3D molecular descriptors	SVM, RF, ANN	Fub: Training set: $R^2 = 0.56$, RMSE = 0.82; Test set: $R^2 = 0.57$, RMSE = 0.80 Cl_{int}: Training set: $R^2 = -0.00$, RMSE = 0.46; Test set: $R^2 = 0.16$, RMSE = 0.40
Dawson *et al.* (2021)	6484	Fub, Cl_{int}	1–3D molecular descriptors	RF	Fub: Training set: $R^2 = 0.584$, RMSE = 0.206; Test set: $R^2 = 0.591$, RMSE = 0.187 (Environment chemicals from ToxCast) Clint: Test set: Q = 0.55 (Class 1), 0.12 (Class 2), 0.90 (Class 3)
Yun *et al.* (2021)	818	Fub	2D molecular descriptors	kNN, SVM, RF, PLS	Test set: $R^2 = 0.52$, Mean absolute error = 12.6

Abbreviations: ANN, artificial neural networks; Cl_{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]The performance from the best model.

Chou WC, Lin Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences*, 191(1):1-14.

A list of databases that contains PK data for machine learning analysis

Table 1. A list of databases that contain pharmacokinetic data for machine learning analyses

Database name	Number of compounds	PK parameters	Description	Website	References
PK-DB	676	Cl, $t_{1/2}$, AUC, C_{max}, Kel and PK time-courses data	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.	https://pk-db.com	Grzegorzewski *et al.* (2021)
PK/DB	1203	HIA, F, fu, BBB, Vd, Cl, $t_{1/2}$	PK/DB is a robust database for PK studies and *in silico* ADME prediction.	www.pkdb.ifsc.usp.br	Moda *et al.* (2008)
PKKB	1685	HIA, fu, Vd, Cl, LD50	Pharmacokinetic Knowledge Base (PKKB) is a comprehensive database of PK and toxic properties for drugs.	http://cadd.suda.edu.cn/admet	Cao *et al.* (2012)
e-Drug3D	1852	Vd, Cl, $t^1/_2$, PPB, F, C_{max}, and Tmax	e-Drug3D is a database of 1852 FDA-approved drugs with 3-D chemical structures and information on PK parameters	https://che-moinfo.ipmc.cnrs.fr/MOLDB/index.php	Pihan *et al.* (2012)
ChEMBL	>1M	Not available	Open-access database containing ADME and toxic information for numerous drug-like compounds	www.ebi.ac.uk/chembl/	Gaulton *et al.* (2012)
Lombardo's database	1352	Vd, Cl, MRT, fu, $t_{1/2}$	A human intravenous PK data set derived from the literature.	Not available	Lombardo *et al.* (2018)
Wang's database	970	HIA	A human intestinal absorption data set consists of 970 compounds, and 9 different types of descriptors.	Not available	Wang *et al.* (2017)
CvT	144	PK time-course data	A public database of chemical time-series concentration data for 144 environmentally relevant chemicals and their metabolites	https://github.com/USEPA/CompTox-PK-CvTdb	Sayre *et al.* (2020)

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.

Chou WC, Lin Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicological Sciences*, 191(1):1-14.

A list of databases relevant to computational toxicology

Table 3. A List of Databases Relevant to Computational Toxicology

Database	Data Size[a]	Data Type	Reference
ACToR	Over 800 000 compounds and 500 000 assays	*In vitro* and *in vivo* toxicity	Judson *et al.* (2008)
Biosolids list	726 chemical pollutants	Concentration data in biosolids	Richman *et al.* (2022)
CEBS	Over 11 000 compounds and 8000 studies	Gene expression data	Lea *et al.* (2017)
ChEMBL	1.1 million bioassays, 1.8 million compounds, over 15 million activities	Literature data on binding, function, and toxicity of drugs and drug-like chemicals	Gaulton *et al.* (2012)
Connectivity map	Around 1300 compounds and 7000 genes	Gene expression data	Subramanian *et al.* (2017)
CTD	Over 14 000 compounds, 42 000 genes, 6000 diseases	Relationships among compounds, genes, and diseases	Davis *et al.* (2021)
DrugMatrix	Around 600 drug molecules and 10 000 genes	Gene expression data	Ganter *et al.* (2005)
GEO	Over 4300 subdata sets	Microarray, next-generation sequencing, and other forms of high-throughput functional genomics data	Barrett *et al.* (2013)
eNanoMapper	Over 700 types of nanomaterials	Diverse data types on nanomaterial physicochemical properties and safety	Jeliazkova *et al.* (2015)
MoleculeNet	Over 700 000 compounds	Quantum mechanics, physical chemistry, biophysics, and physiology	Wu *et al.* (2018)
Open TG-GATEs	170 compounds	Gene expression data and metadata	Igarashi *et al.* (2015)
PubChem	Over 111 million compounds, 1.39 million bioassays, and 293 million bioactivity data points	Toxicology, genomics, pharmacology, and literature data	Kim *et al.* (2021)
Pubvinas	11 types of nanomaterials with 705 unique nanomaterials	Up to 6 physicochemical properties and/or bioactivities	Yan *et al.* (2020)
REACH	21,405 unique substances with information from 89,905 dossiers	Data submitted in European Union chemical legislation	Luechtefeld *et al.* (2016)
RepDose	364 compounds investigated in 1017 studies, resulting in 6,002 specific effects	Repeat-dose study data in dogs, mice, and rats	Bitsch *et al.* (2006)
SEURAT	Over 5500 cosmetic-type compounds in the current COSMOS database web portal	Animal toxicity data	Vinken *et al.* (2012)
ToxicoDB	231 chemicals	Toxicogenomic data	Nair *et al.* (2020)
ToxNET	Over 50 000 environmental chemicals from 16 resources	*In vitro* and *in vivo* toxicity data	Fonger *et al.* (2000)

[a] 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.

Lin Z, Chou WC. (2022). Machine learning and artificial intelligence in toxicological sciences. *Toxicological Sciences*, 189(1):7-19.

Progress and significant data gaps and challenges in this field

- Integrating ML/AI approaches with traditional computational toxicological models is promising and has substantially driven toxicological sciences forward.

- Existing ML-based QSAR studies to predict ADME properties are focused on rodents and humans, but not food-producing animals

- Existing AI-based PBPK models are mostly for small molecular drugs, but fewer for environmental chemicals, and even very limited for nanomaterials.

Application 1:
Applications of machine learning and AI approaches to predict ADMET properties of chemicals

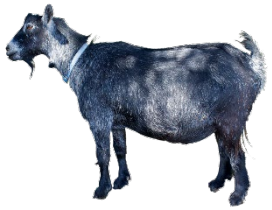
Application 1 Background: Food Animal Residue Avoidance Databank (FARAD)

Background

~94.4 million

~72.9 million

~5.23 million



~2.62 million

~8.54 billion

~238 million

Terminology

The meat withdrawal period or milk discard time is the interval between the time of the last administration of a new animal drug and the time when the animal can be safely slaughtered for food or the milk can be safely consumed.

The tolerance (or maximum residue limit [MRL]) is the maximum concentration of a marker residue, or other residue indicated for monitoring, that can legally remain in a specific edible tissue of a treated animal.

Extralabel drug use (ELDU) describes the use of an approved drug in a manner that is not in accordance with the approved labeling, yet meets the conditions set forth by the Animal Medicinal Drug Use Clarification Act of 1994 (AMDUCA) and U.S. Food and Drug Administration (FDA) regulations.

We use the term "withdrawal interval" when a drug is used extralabel.

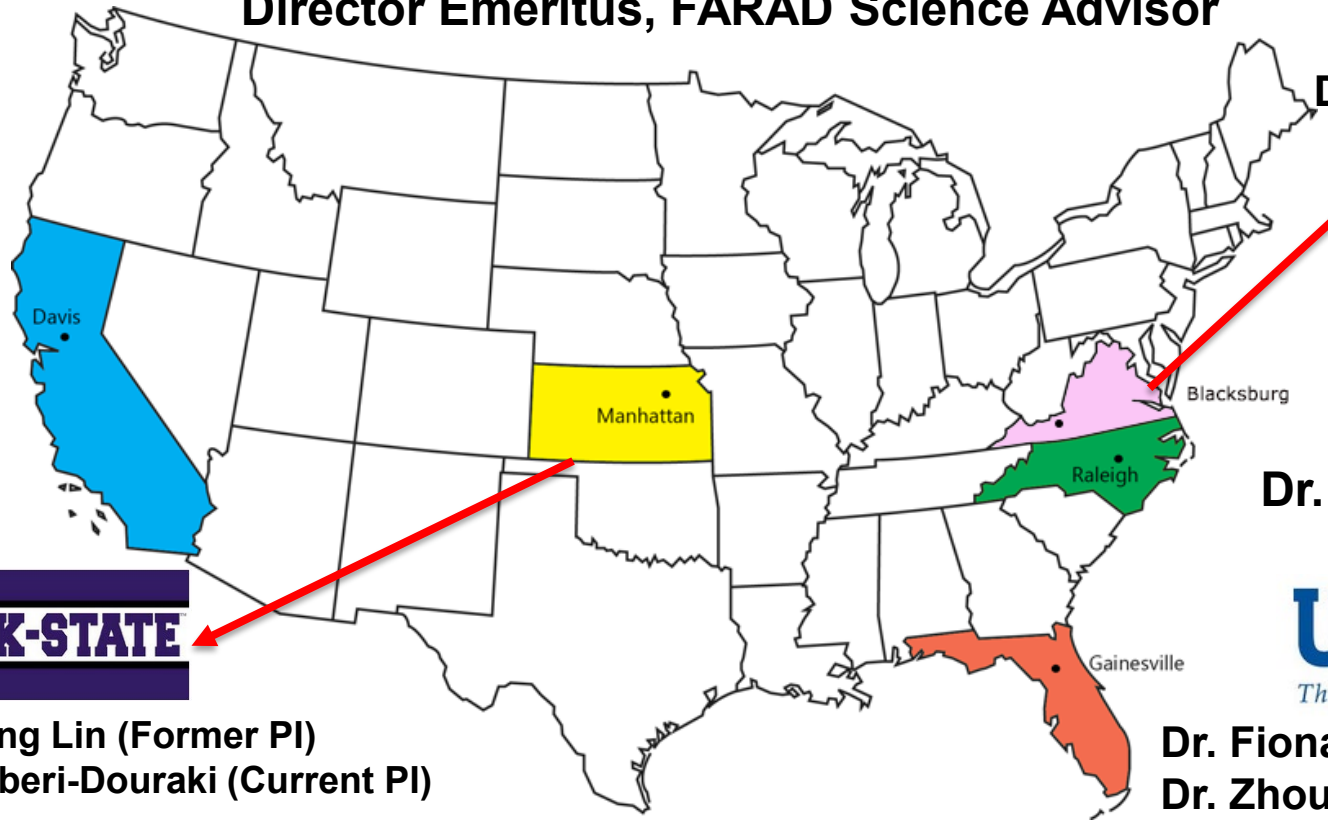
The challenge in this field is how to calculate withdrawal interval after extralabel drug use.

USDA National Residue Sample Results "Red Book": https://www.fsis.usda.gov/wps/portal/fsis/topics/data-collection-and-reports/chemistry/Residue-Chemistry
USDA Economic Research Service Statistics & Information: https://www.ers.usda.gov/topics/animal-products/cattle-beef/statistics-information.aspx

Application 1 Background: Food Animal Residue Avoidance Databank (FARAD)

FARAD's primary mission is to help producers and veterinarians prevent or mitigate illegal or harmful residues of drugs, pesticides, biotoxins and other chemical agents that may contaminate foods of animal origin.

Dr. Jim E. Riviere, Co-Founder,
Director Emeritus, FARAD Science Advisor



Dr. Jennifer L. Davis

Dr. Lisa A. Tell

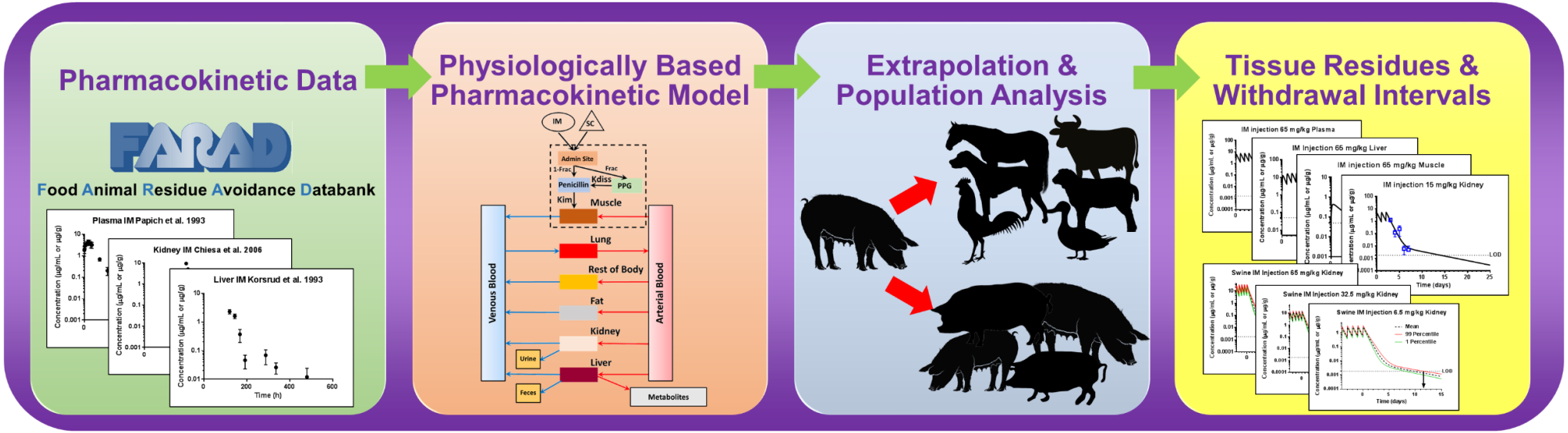
Dr. Ronald E. Baynes

Dr. Zhoumeng Lin (Former PI)
Dr. Majid Jaberi-Douraki (Current PI)

Dr. Fiona P. Maunsell (Current PI)
Dr. Zhoumeng Lin (Current Co-PI)

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

PK/PBPK Component of FARAD at UF



Objective: To develop web-based computational models/platforms that allow FARAD responders to easily calculate withdrawal intervals for drugs or other chemicals in different food animal species

Specific responsibilities:

- Develop PBPK and QSAR models and web-based interfaces
- Provide pharmacokinetic and toxicokinetic support to other regional centers
- Provide advice on withdrawal intervals and potential food safety risk
- Provide training to FARAD responders on how to calculate withdrawal intervals

Overview and Timeline of Our PK/PBPK/AI-QSAR 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
Lin et al. 2016
Lin et al. 2016

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
Li et al. 2017
Li et al. 2018

2018-2023

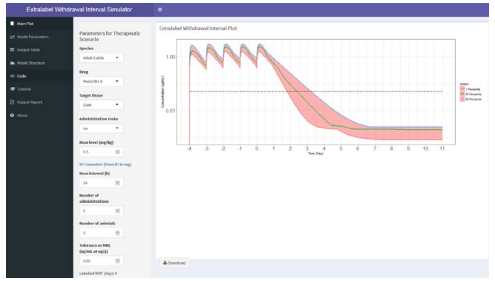
- Graphical user interface (GUI)
- Population PBPK: penicillin G, flunixin, florfenicol, oxytetracycline, PFAS
- Physiological parameter database: cattle, swine, chickens, turkeys, sheep, goats
- Quantitative methods from FDA & EMA

Li et al. 2019a
Li et al. 2019b
Bates et al. 2020
Wang et al. 2021
Lin et al. 2019
Lin et al. 2020
Smith et al. 2020
Li et al. 2021
Riad et al. 2021
Chou et al. 2022
Yuan et al. 2022a
Yuan et al. 2022b
Chou et al. 2023
Wu et al. 2023

2021-present

- Incorporate machine learning and AI approaches into our PBPK/QSAR models

Lin and Chou, 2022
Chou and Lin, 2023
Wu et al. 2024



FDA U.S. FOOD & DRUG ADMINISTRATION

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

http://www.thecis.co.uk/theCIS/images/ciscows_slider.jpg
http://www.nature.com/polopoly_fs/7.11560.1374594725!/image/Pigs.jpg_gen/derivatives/landscape_630/Pigs.jpg

Goals of our AI-based PBPK/QSAR in predicting ADME of chemicals

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

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

FARAD
Food Animal Residue Avoidance Databank

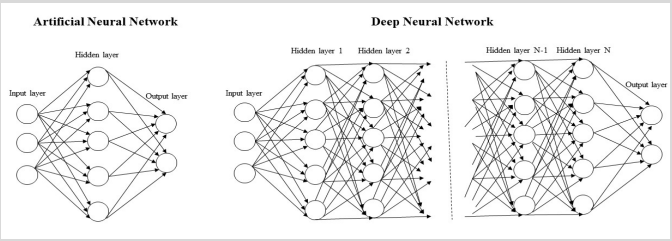
Drugs/active ingredients with reported half-lives

Data Processing

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

Extract Cheminformatics data
- Molecular descriptors
- Fingerprints

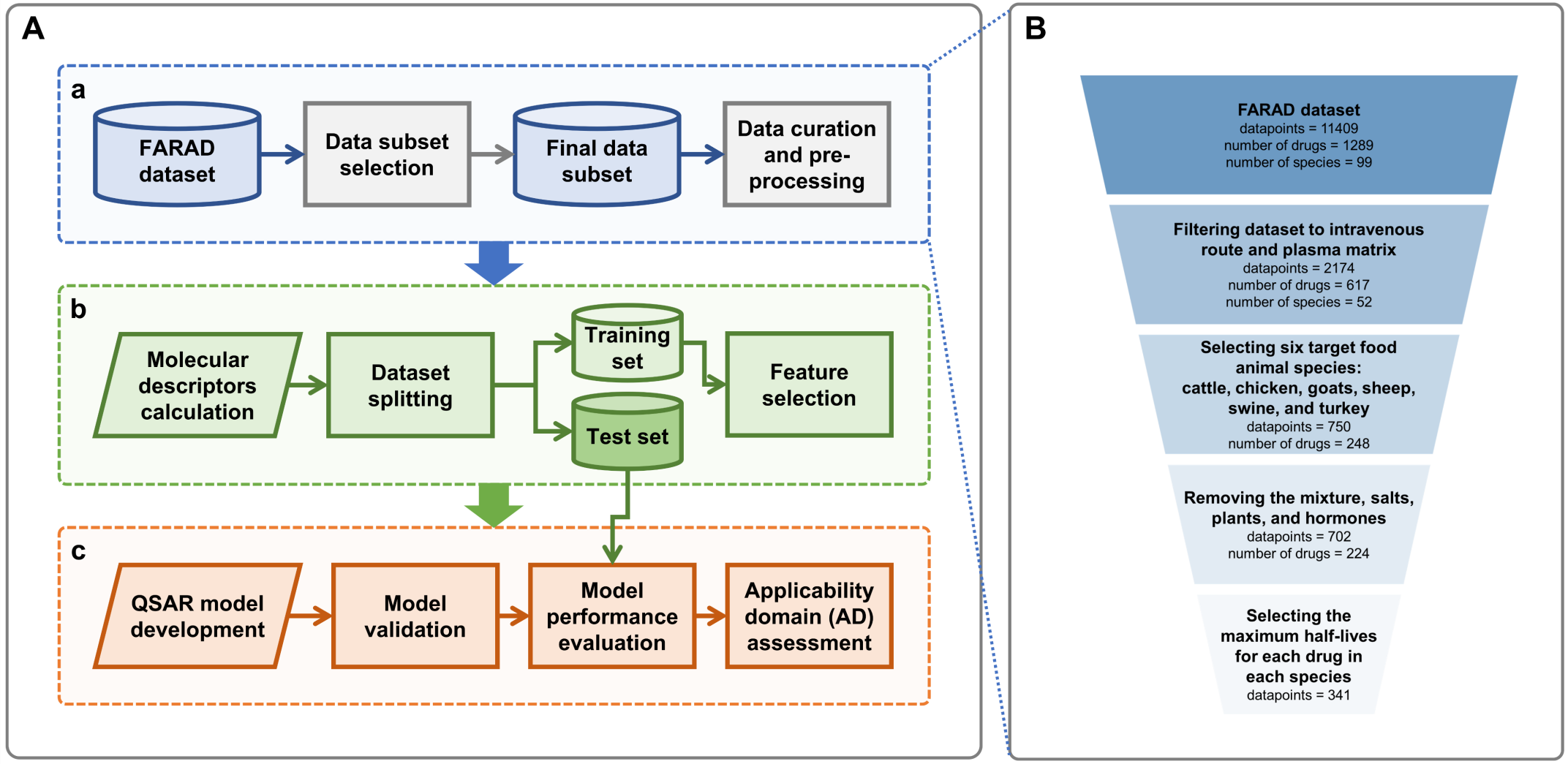
Machine Learning and Artificial Intelligence Methods



QSAR: Quantitative structure-activity relationships
QSPR: Quantitative structure-property relationships

A ML-based QSAR model to predict the plasma half-life of drugs in food animals

Schematic Workflow of ML-based QSAR Modeling



Wu PY, Chou WC, Wu X, Kamineni VN, Kuchimanchi Y, Tell LA, Maunsell FP, Lin Z. (2024). Development of Machine Learning-Based Quantitative Structure-Activity Relationship Models for Predicting Plasma Half-Lives of Drugs in Six Common Food Animal Species. *Toxicological Sciences*, in press.

A ML-based QSAR model to predict the plasma half-life of drugs in food animals

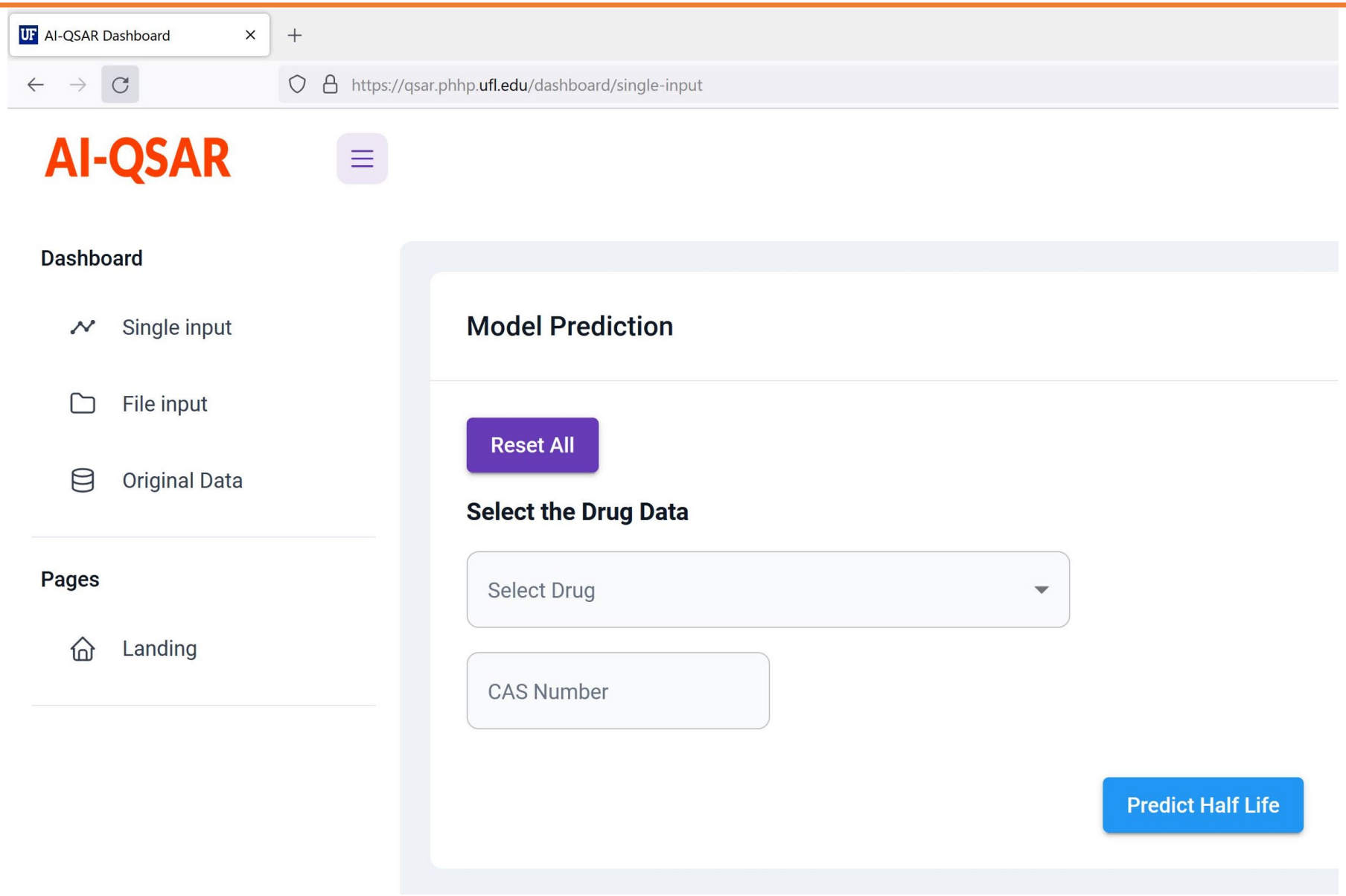
Representative Results

Descriptor / Model	All		RDKit		ECFP		FCFP		MACCS	
	5-fold CV	Test	5-fold CV	Test	5-fold CV	Test	5-fold CV	Test	5-fold CV	Test
KNN										
R^2	0.21±0.25	0.21	0.09±0.11	0.09	0.15±0.15	0.24	0.16±0.16	0.25	0.01±0.07	0.11
RMSE	35.26±27.47	26.49	36.50±26.28	28.50	35.62±26.09	26.10	35.60±26.27	25.90	37.32±25.39	28.12
RF										
R^2	0.05±0.10	0.12	0.01±0.07	0.12	0.05±0.06	0.12	0.09±0.10	0.17	0.04±0.05	0.20
RMSE	36.36±24.79	28.04	36.77±24.81	28.07	36.84±25.80	28.08	37.02±25.18	27.23	36.93±24.78	26.27
SVM										
R^2	0.25±0.26	0.09	0.23±0.27	0.21	0.33±0.31	0.09	0.34±0.31	0.09	0.35±0.29	0.16
RMSE	34.35±26.82	28.45	34.25±26.25	26.53	32.87±27.14	28.53	32.80±27.07	28.46	32.54±26.85	27.35
DNN										
R^2	0.82±0.19	0.67	0.85±0.21	0.40	0.46±0.31	0.44	0.82±0.24	0.49	0.61±0.23	0.43
RMSE	13.53±8.21	17.23	11.87±10.73	23.24	28.46±13.39	22.30	11.01±8.98	21.31	22.91±8.86	22.66

CV: cross-validation
ECFP: extended-connectivity fingerprints, FCFP: functional-class fingerprints, MACCS: molecular ACCess system
kNN: k-nearest neighbors, RF: random forest, SVM: support vector machine, DNN: deep neural network

Wu PY, Chou WC, Wu X, Kamineni VN, Kuchimanchi Y, Tell LA, Maunsell FP, Lin Z. (2024). Development of Machine Learning-Based Quantitative Structure-Activity Relationship Models for Predicting Plasma Half-Lives of Drugs in Six Common Food Animal Species. *Toxicological Sciences*, in press.

A ML-based QSAR model to predict the plasma half-life of drugs in food animals



Wu PY, Chou WC, Wu X, Kamineni VN, Kuchimanchi Y, Tell LA, Maunsell FP, Lin Z. (2024). Development of Machine Learning-Based Quantitative Structure-Activity Relationship Models for Predicting Plasma Half-Lives of Drugs in Six Common Food Animal Species. *Toxicological Sciences*, in press.

A ML-based QSAR model to predict plasma half-life of PFAS in rodents/humans

toxics

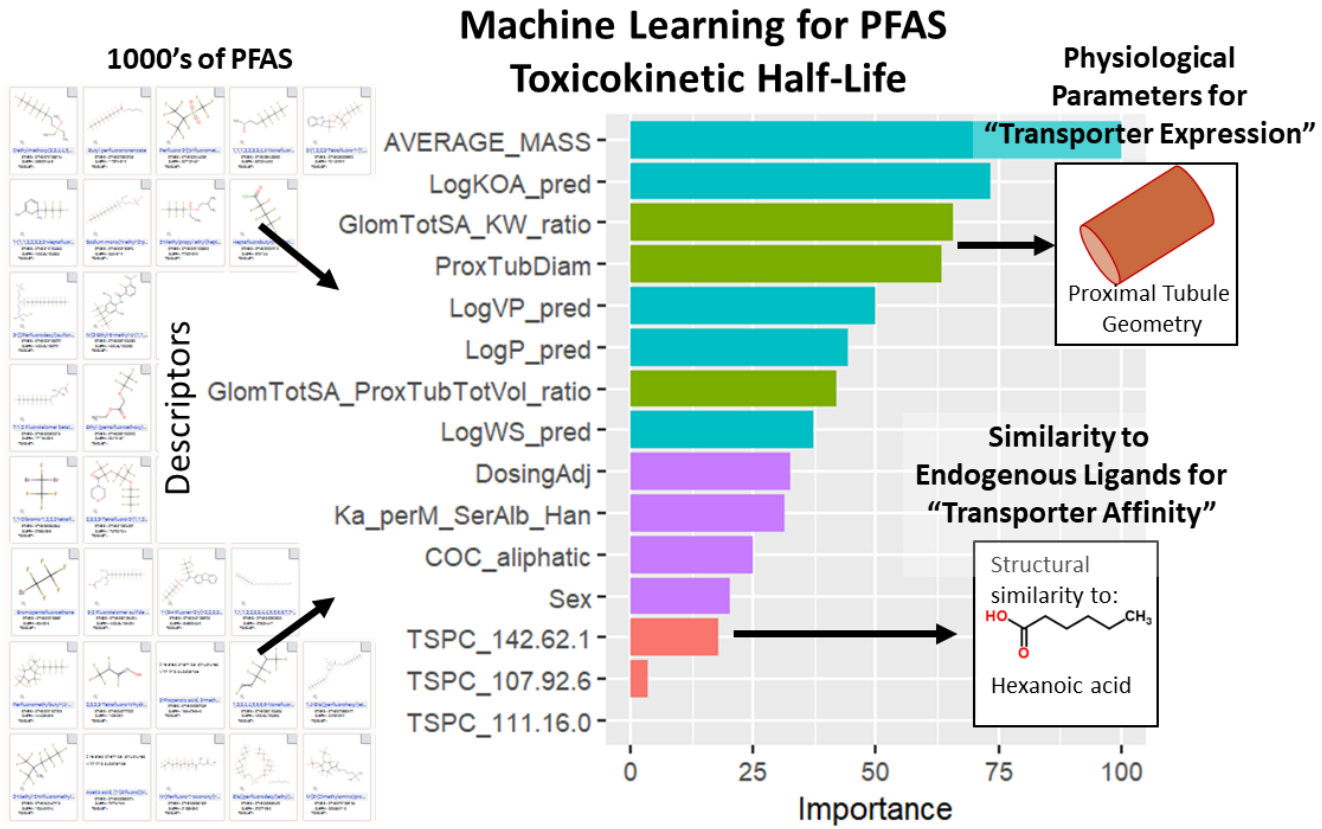
MDPI

Article

A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per- and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species

Daniel E. Dawson [1], Christopher Lau [2], Prachi Pradeep [1,3], Risa R. Sayre [1], Richard S. Judson [1], Rogelio Tornero-Velez [1] and John F. Wambaugh [1,*]

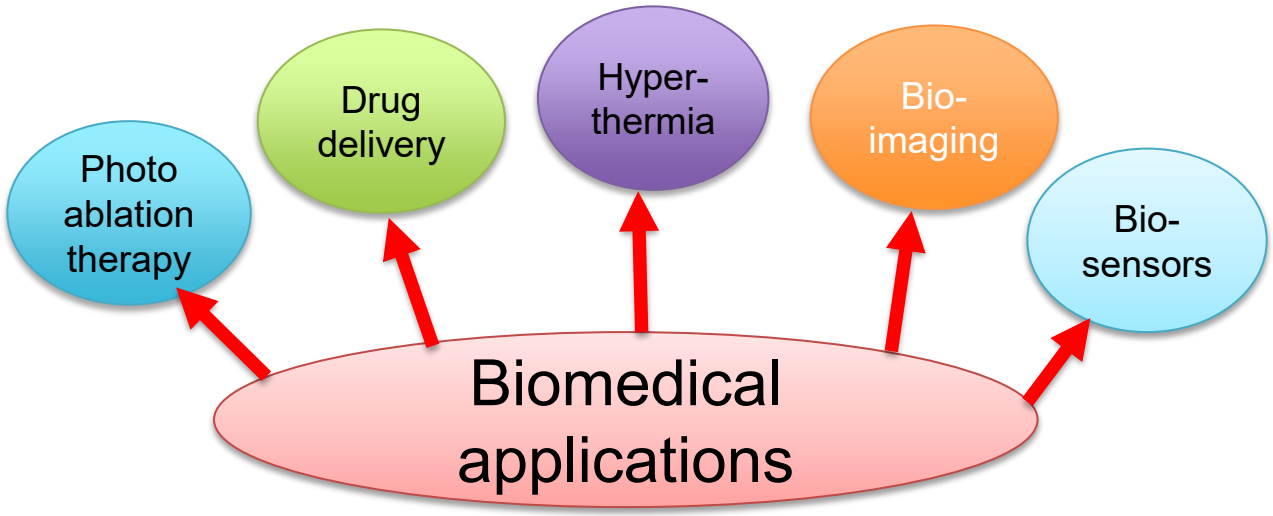
[1] U.S. Environmental Protection Agency, Office of Research and Development, Cente
Toxicology and Exposure, 109 T.W. Alexander Drive, Research Triangle Park, NC 2
[2] U.S. Environmental Protection Agency, Office of Research and Development, Cente
Environmental Assessment, 109 T.W. Alexander Drive, Research Triangle Park, NC
[3] Oak Ridge Institutes for Science and Education, Oak Ridge, TN 37830, USA
* Correspondence: wambaugh.john@epa.gov; Tel.: +1-919-541-7641



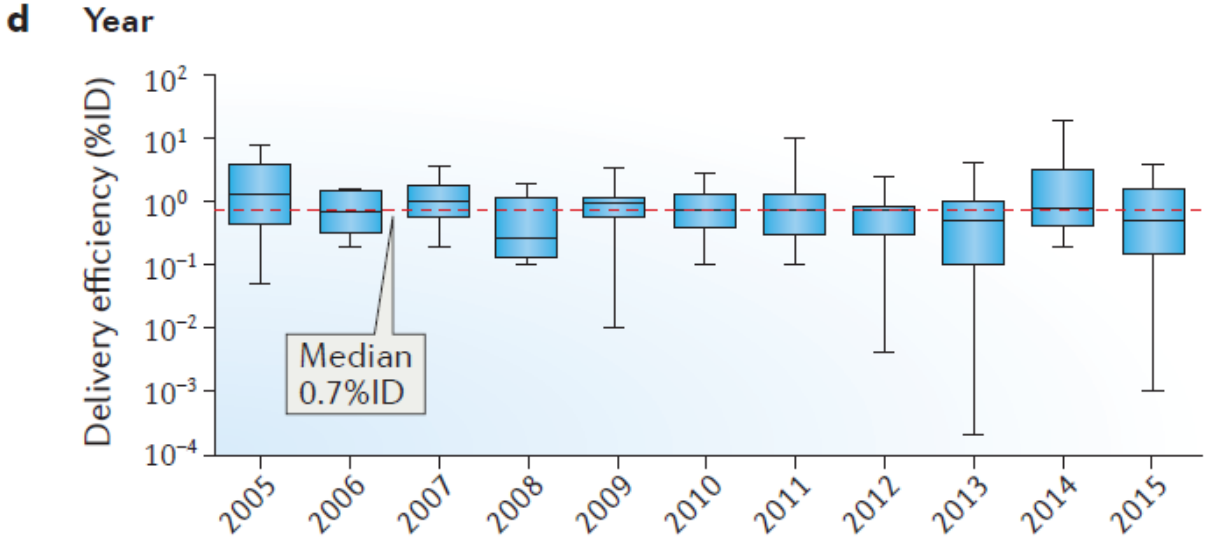
Dawson DE, Lau C, Pradeep P, Sayre RR, Judson RS, Tornero-Velez R, Wambaugh JF. A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per- and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species. *Toxics*. 2023; 11(2):98.

Application 2:
Applications of machine learning and AI approaches to develop PBPK models for xenobiotics in a high-throughput manner

Application 2: AI-enabled 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)

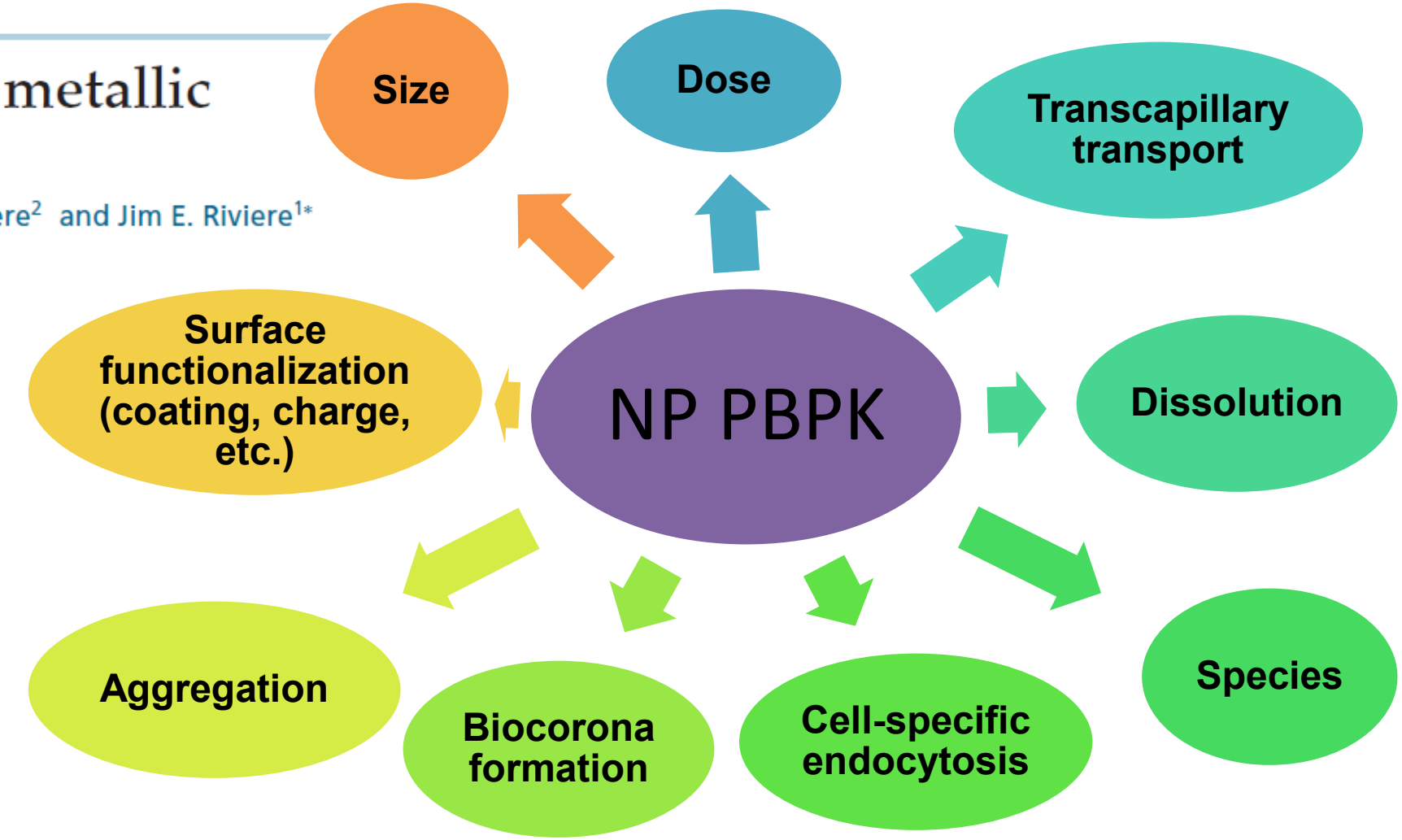
- Wilhelm S, Tavares AJ, Dai Q, Ohta S, Audet J, Dvorak HF, Chan WCW. Analysis of nanoparticle delivery to tumours. 2016. Nature Reviews Materials, 1, 16014.
- Cheng YH, He C, Riviere JE, Monteiro-Riviere NA, Lin Z. Meta-Analysis of Nanoparticle Delivery to Tumors Using a Physiologically Based Pharmacokinetic Modeling and Simulation Approach. ACS Nano. 2020;14(3):3075-3095. (**Best Paper Award of the Year 2020 – Honorable Mention presented by Society of Toxicology Biological Modeling Specialty Section in 2021**)
- Chen Q, Riviere JE, Lin Z. Toxicokinetics, dose-response, and risk assessment of nanomaterials: Methodology, challenges, and future perspectives. WIREs Nanomed Nanobiotechnol. 2022 Nov;14(6):e1808.

PK and PBPK Modeling of Nanoparticles

Overview

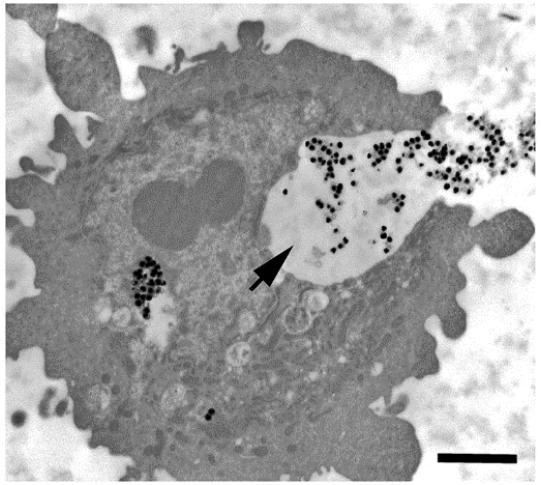
Pharmacokinetics of metallic nanoparticles

Zhoumeng Lin,[1] Nancy A. Monteiro-Riviere[2] and Jim E. Riviere[1]*

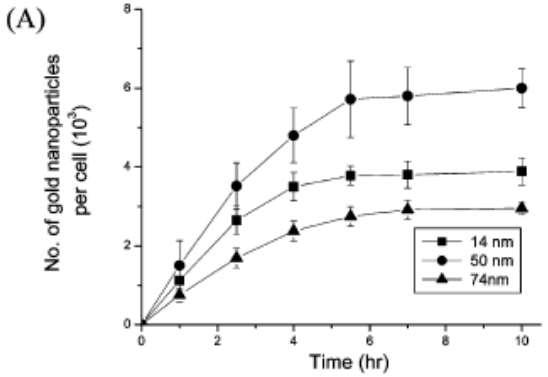


Lin Z, Monteiro-Riviere NA, Riviere JE. Pharmacokinetics of metallic nanoparticles. *WIREs Nanomed Nanobiotechnol*, 2015, 7:189-217.

PBPK Modeling of Nanoparticles vs. Small Molecules



Monteiro-Riviere et al. 2013. Toxicology Letters



Chithrani et al. 2006. Nano Letters

- Partition Coefficient vs. Time-dependent Uptake
- Hill function to simulate endocytosis of gold nanoparticles

$$K_{up\_liver} = \frac{K_{\max\_liver} \times T^{n_{liver}}}{K_{50\_liver}^{n_{liver}} + T^{n_{liver}}}$$

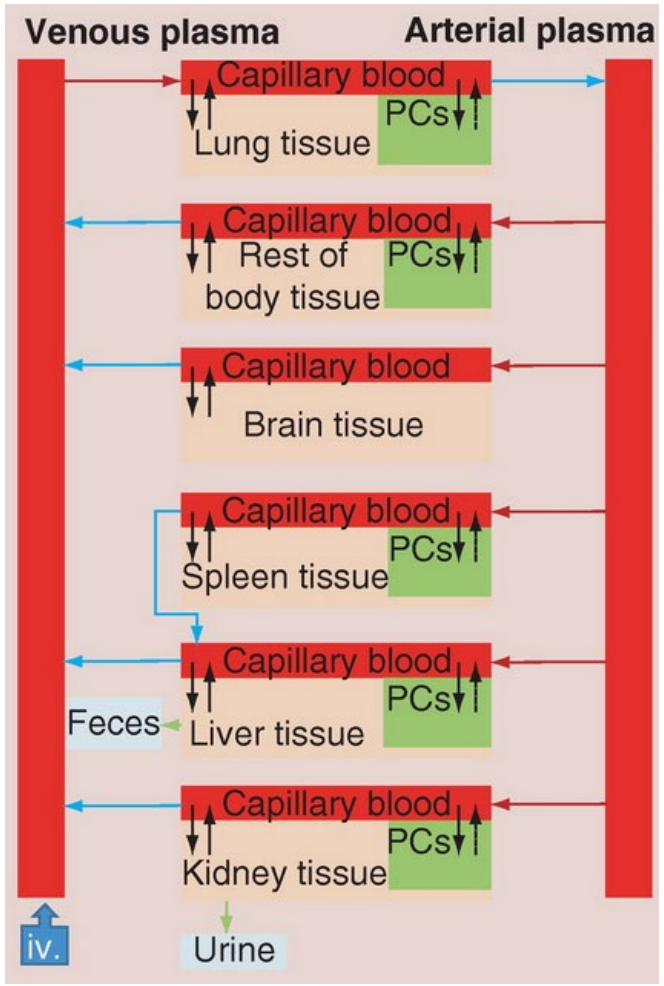
K_{max_liver}: maximum uptake rate

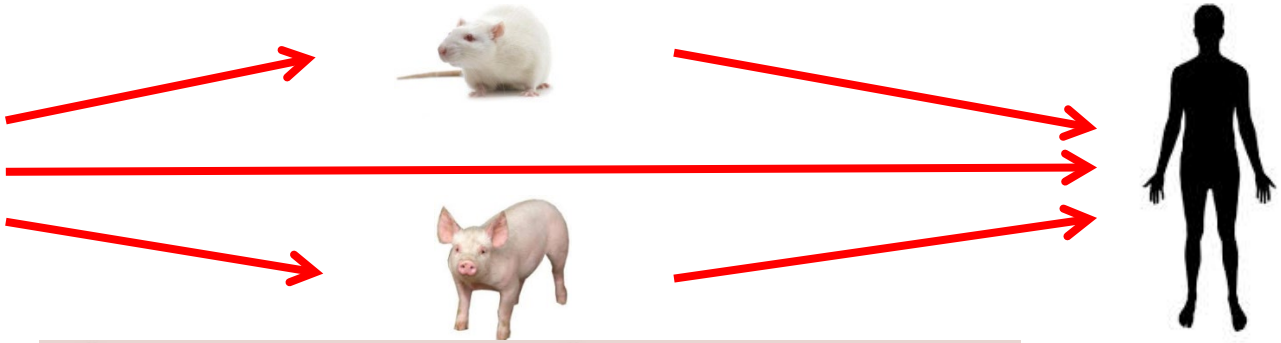
K_{50_liver}: time reaching half maximum rate

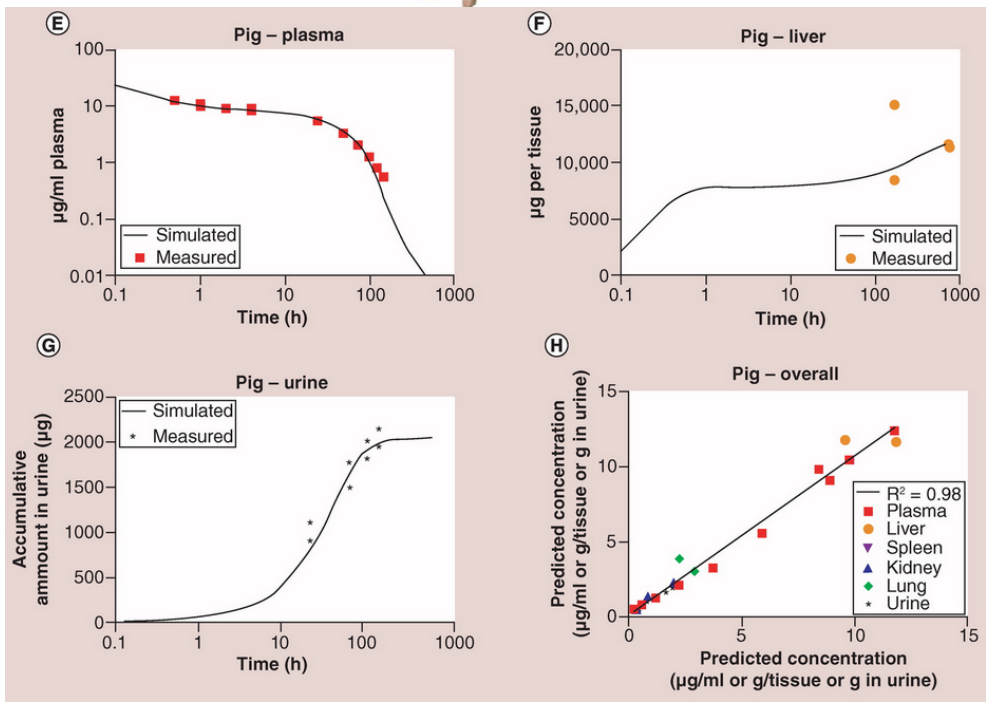
n_{liver}: Hill coefficient

- Lin Z, Monteiro-Riviere NA, Riviere JE. A physiologically based pharmacokinetic model for polyethylene glycol-coated gold nanoparticles of different sizes in adult mice. *Nanotoxicology*. 2016;10(2):162-72. (**Best Paper Award [Honorable Mention], 2016 Society of Toxicology Biological Modeling Specialty Section**)

A PBPK Model for Gold Nanoparticles (AuNPs)



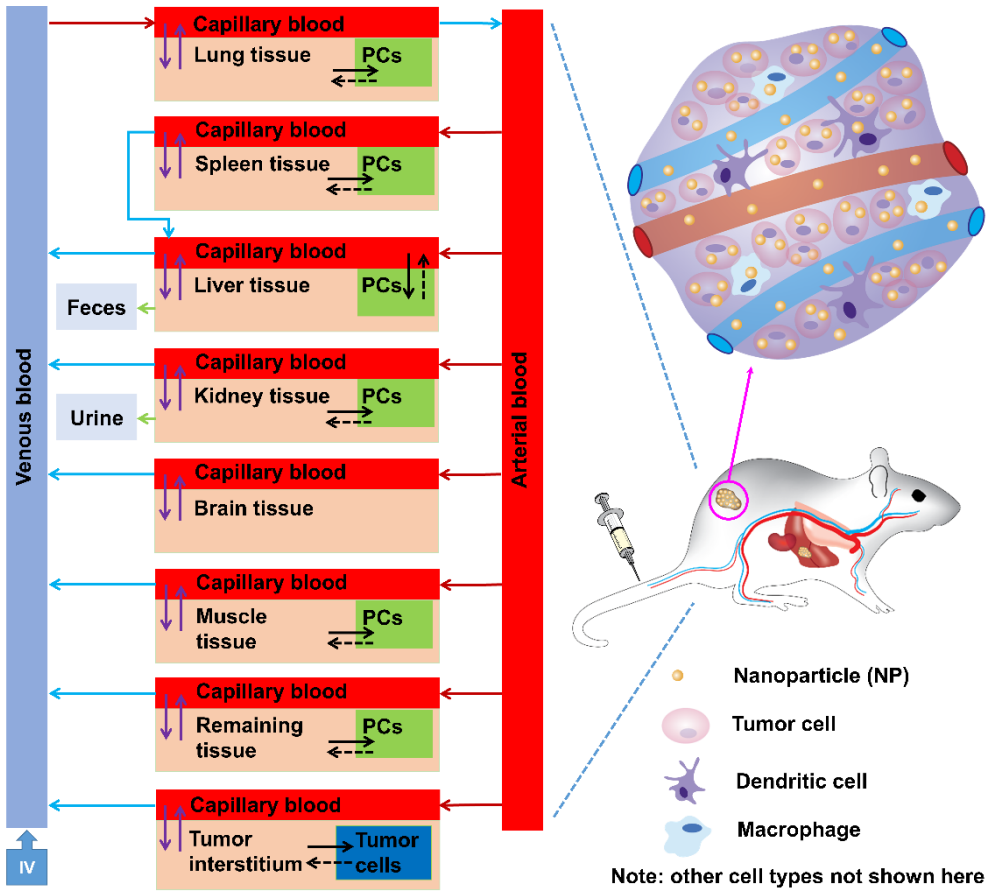




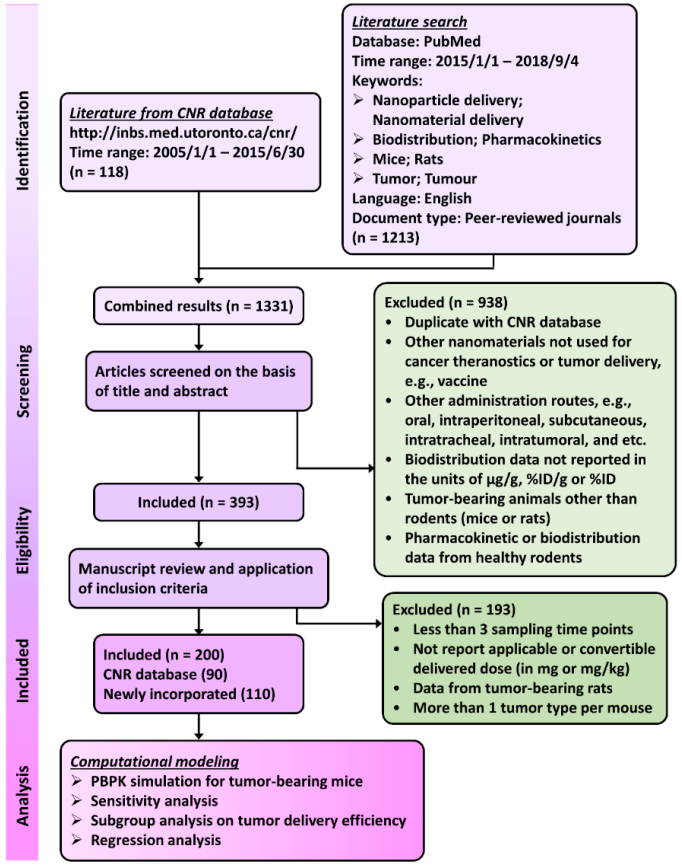
- Lin Z, Monteiro-Riviere NA, Kannan R, Riviere JE. A computational framework for interspecies pharmacokinetics, exposure and toxicity assessment of gold nanoparticles. *Nanomedicine (Lond)*. 2016 Jan;11(2):107-19.

From Healthy Animals to Tumor-Bearing Mice

PBPK Structure in tumor-bearing mice



Nano-Tumor Database

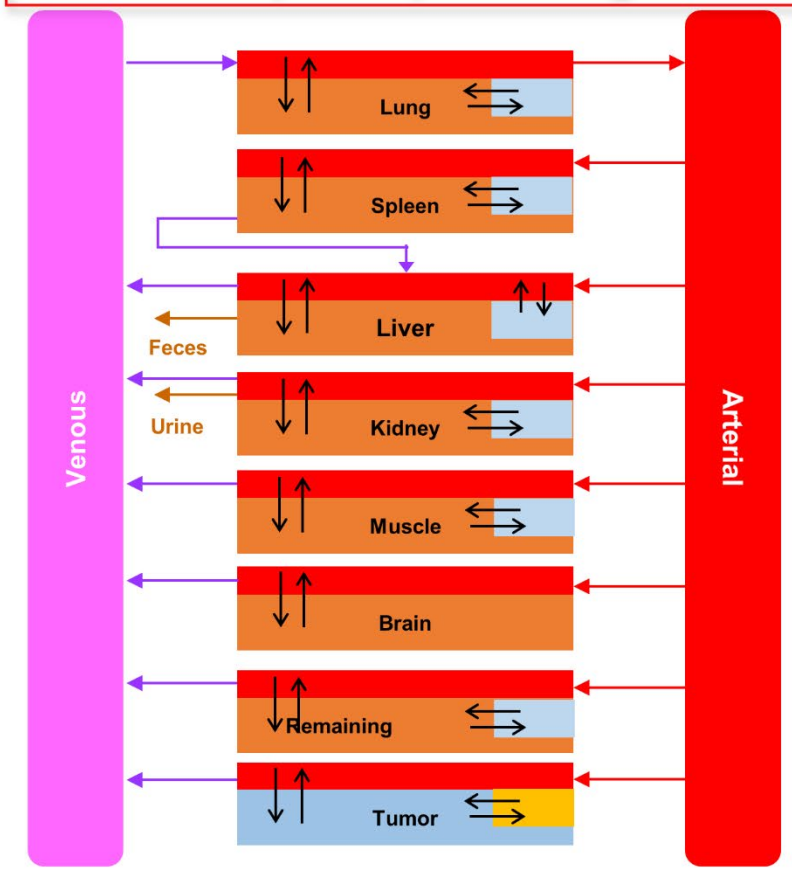


Phase 1: 376 datasets from 200 studies published from 2005 to 2018 (Cheng et al., 2020).
Phase 2: 534 datasets from 297 studies published from 2005 to 2021 (Chen et al., 2023).

- Cheng YH, He C, Riviere JE, Monteiro-Riviere NA, Lin Z*. (2020). Meta-analysis of nanoparticle delivery to tumors using a physiologically based pharmacokinetic modeling and simulation approach. *ACS Nano*, 14(3): 3075-3095. (**Best Paper Award of the Year 2020 – Honorable Mention presented by Society of Toxicology Biological Modeling Specialty Section in 2021**)
- Chen Q, Yuan L, Chou WC, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. Meta-analysis of nanoparticle distribution in tumors and major organs in tumor-bearing mice. *ACS Nano*. 2023;17(20):19810-19831.

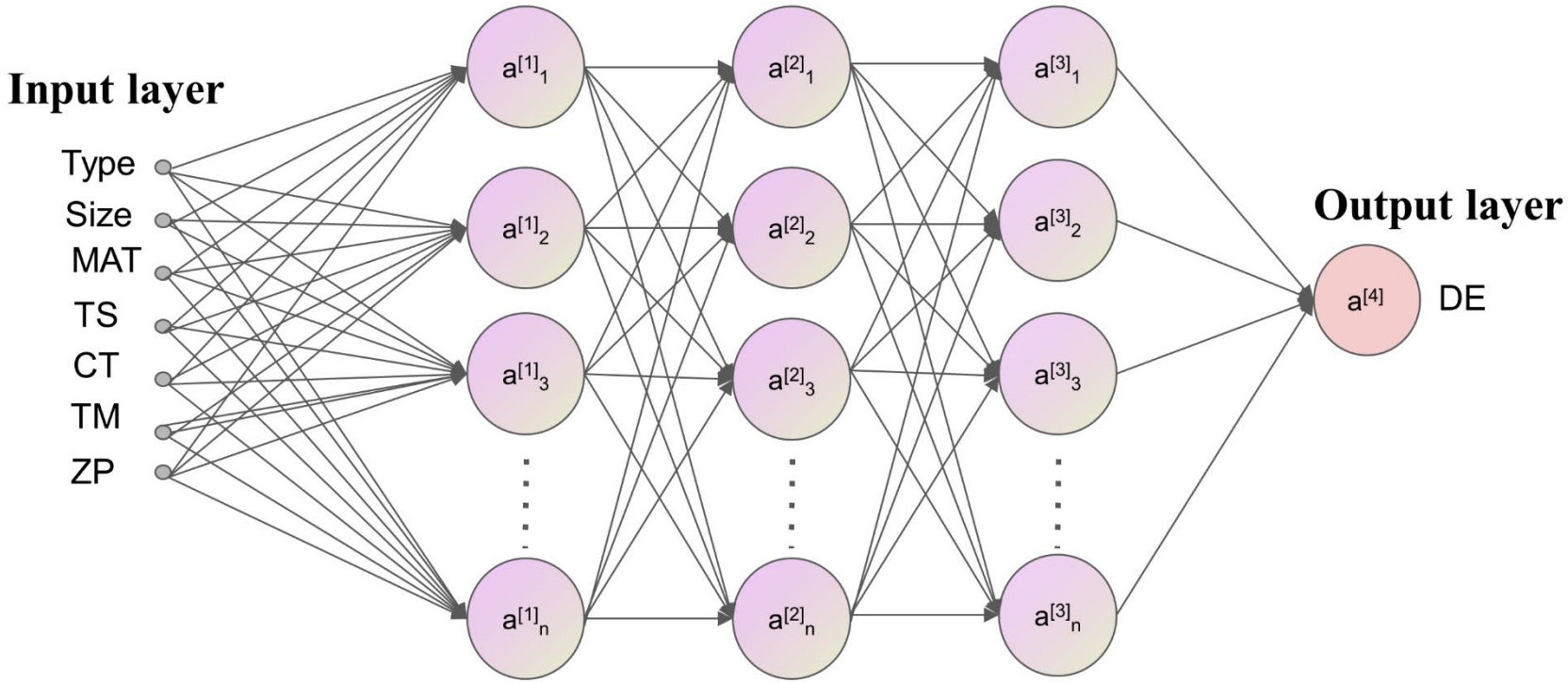
A Data-Driven Approach

Physiologically Based Pharmacokinetic (PBPK) Modeling



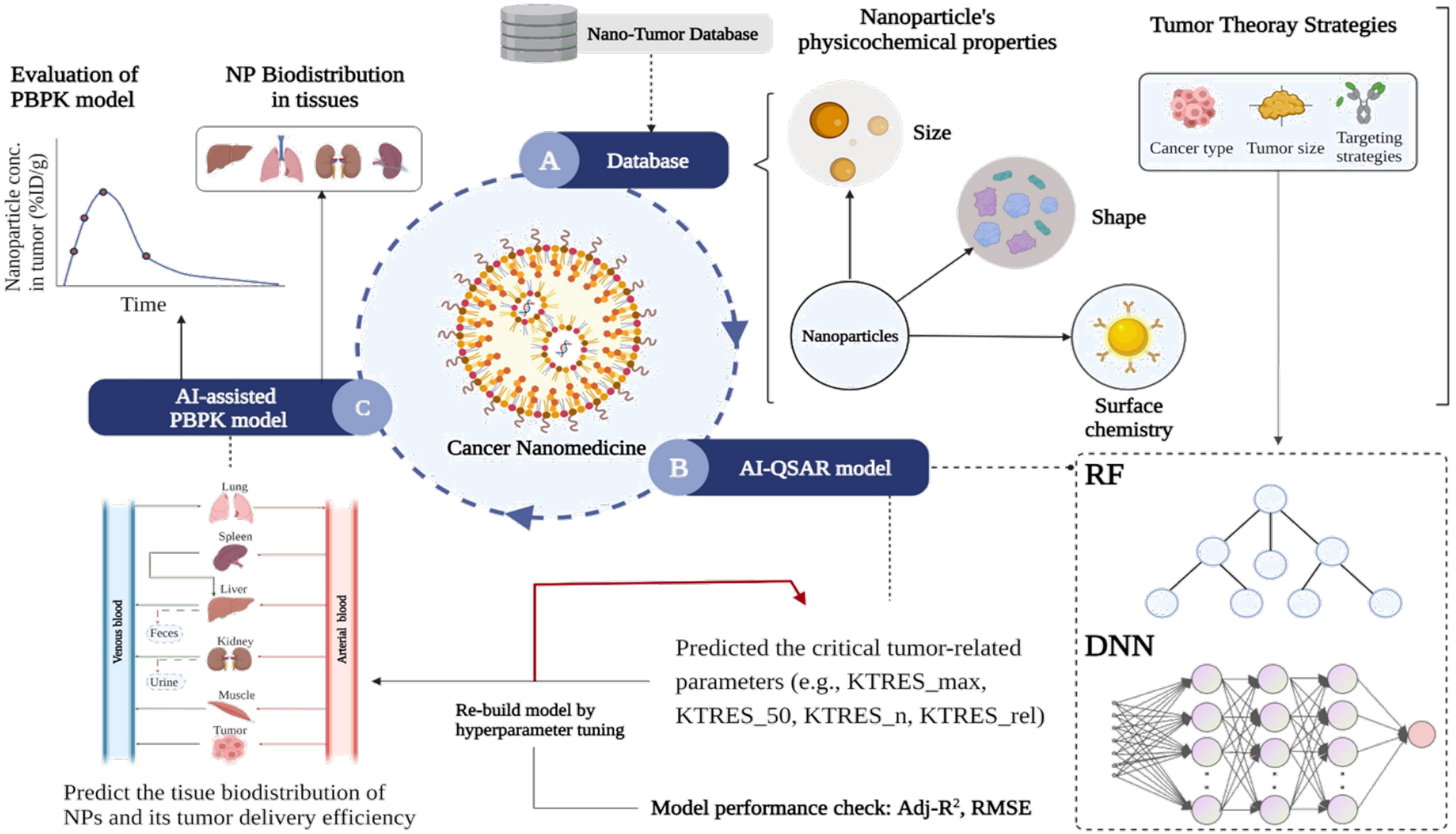
Machine Learning and Artificial Intelligence

Cancer Nanomedicine



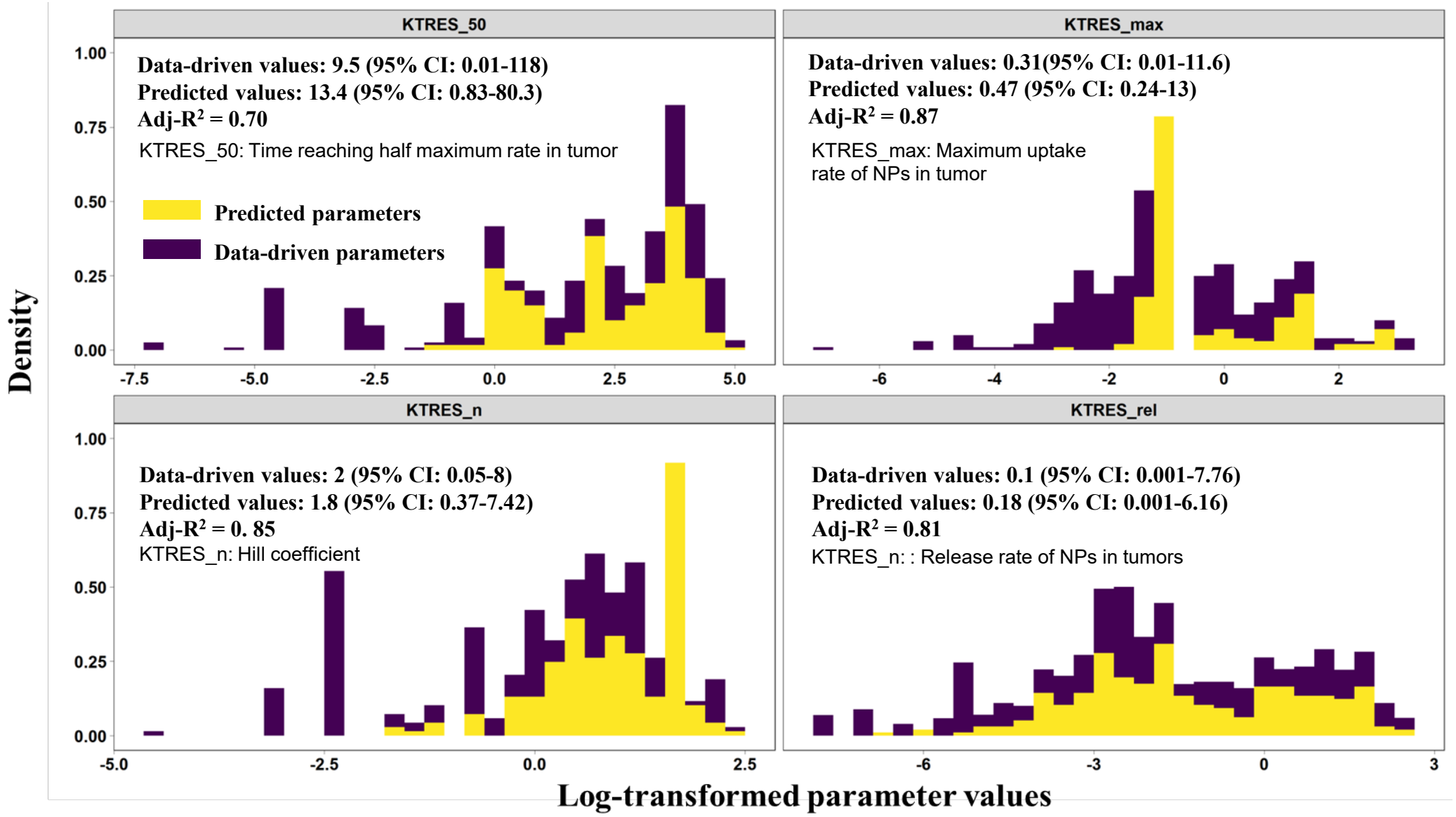
Lin Z, Chou WC, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE. (2022). Predicting Nanoparticle Delivery to Tumors Using Machine Learning and Artificial Intelligence Approaches. *International Journal of Nanomedicine*, 17: 1365-1379.

A Hybrid Approach



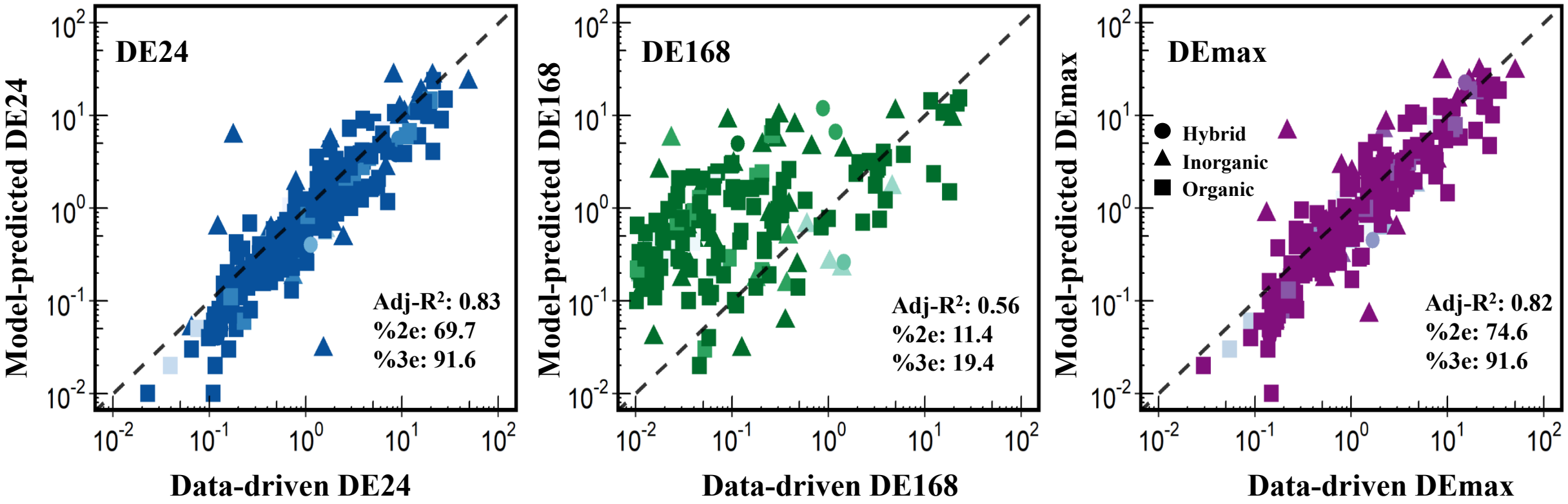
Chou WC, Chen Q, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2023). An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. *Journal of Controlled Release*, 361:53-63. [**Best Paper Award of the Year 2023 by Society of Toxicology Biological Modeling Specialty Section in 2024**]

Similarity between Predicted and Data-Driven Parameters



Chou WC, Chen Q, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2023). An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. *Journal of Controlled Release*, 361:53-63. [**Best Paper Award of the Year 2023 by Society of Toxicology Biological Modeling Specialty Section in 2024**]

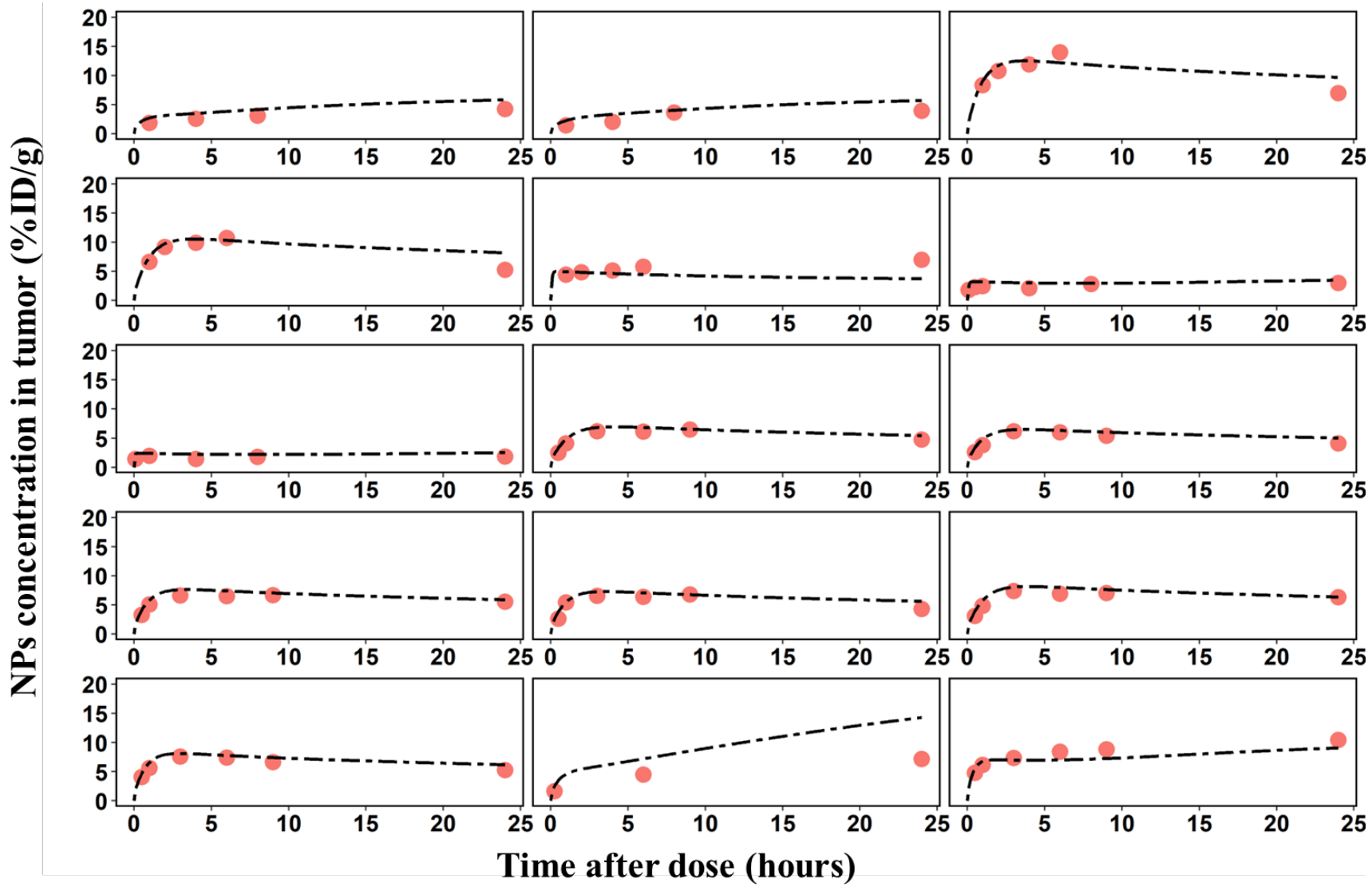
Evaluation Results of AI-PBPK Model-Predicted Tumor Delivery Efficiency



Abbreviation: DE, delivery efficiency; DE24, delivery efficiency at 24 hours; DE168, delivery efficiency at 168 hours; DEmax, maximum of DE; %2e, percentage of 2-fold error range %3e, percentage of 3-fold error range

Chou WC, Chen Q, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2023). An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. *Journal of Controlled Release*, 361:53-63. [**Best Paper Award of the Year 2023 by Society of Toxicology Biological Modeling Specialty Section in 2024**]

Representative Evaluation Results of the AI-PBPK Model



Chou WC, Chen Q, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2023). An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. *Journal of Controlled Release*, 361:53-63. [**Best Paper Award of the Year 2023 by Society of Toxicology Biological Modeling Specialty Section in 2024**]

Limitations of the Previous Studies

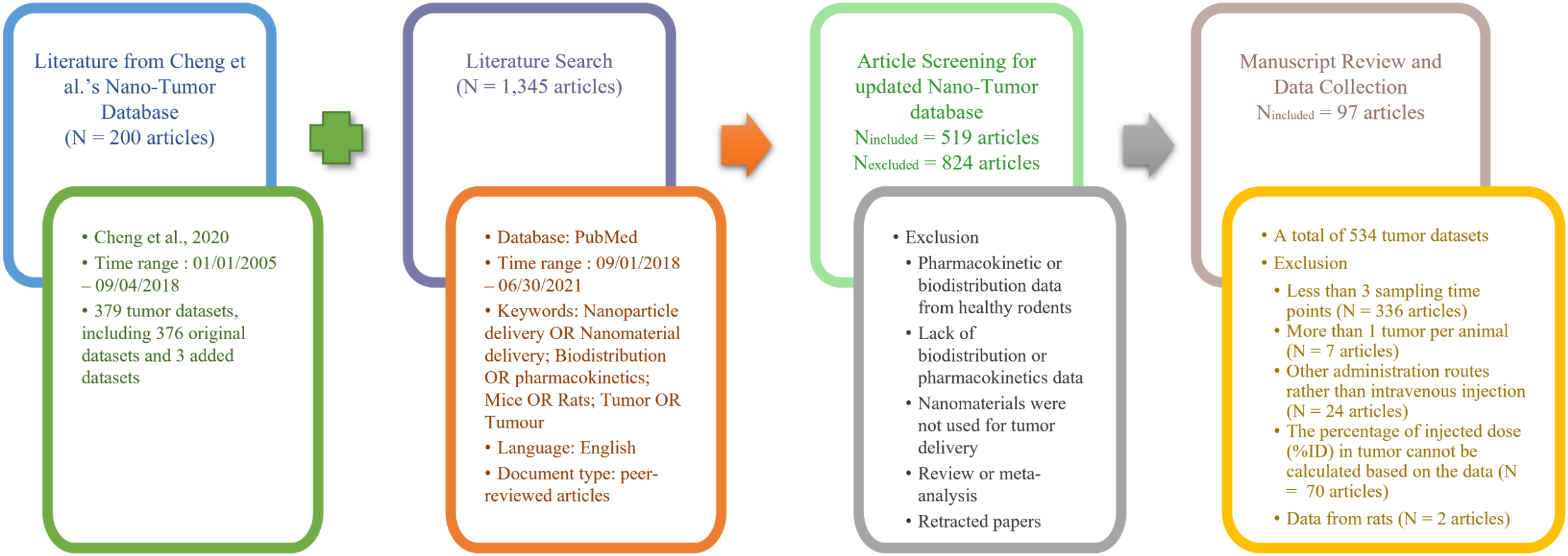
Previous Studies

Tumors only

New Studies

Tumors and major organs (i.e., liver, kidneys, spleen, lungs, and heart)

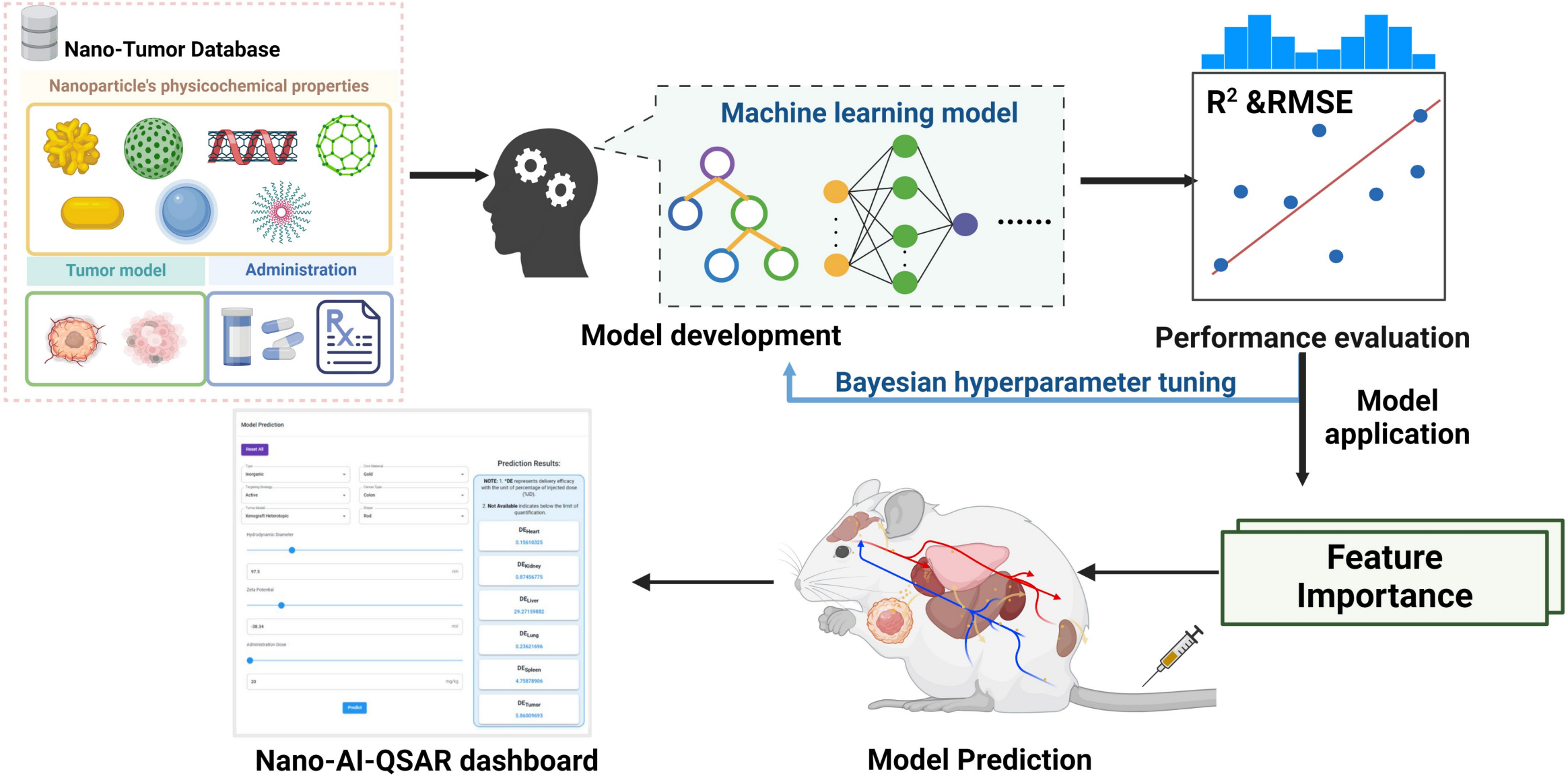
Updated Nano-Tumor Database



In addition to 534 tumor datasets with 2345 data points, the current database also includes 1972 datasets for five major organs (i.e., liver, spleen, lungs, heart, and kidneys) with 8461 data points.

Chen Q, Yuan L, Chou WC, Cheng YH, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. Meta-analysis of nanoparticle distribution in tumors and major organs in tumor-bearing mice. *ACS Nano*. 2023;17(20):19810-19831.

Machine Learning Models to Predict Tissue Distribution and Tumor Delivery



Mi K, Chou WC, Chen Q, Yuan L, Kamineni VN, Kuchimanchi Y, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2024). Predicting tissue distribution and tumor delivery of nanoparticles in mice using machine learning models. *Journal of Controlled Release*, 374:219-229. (**Selected as the Front Cover Paper**)

Machine Learning Models to Predict Tissue Distribution and Tumor Delivery

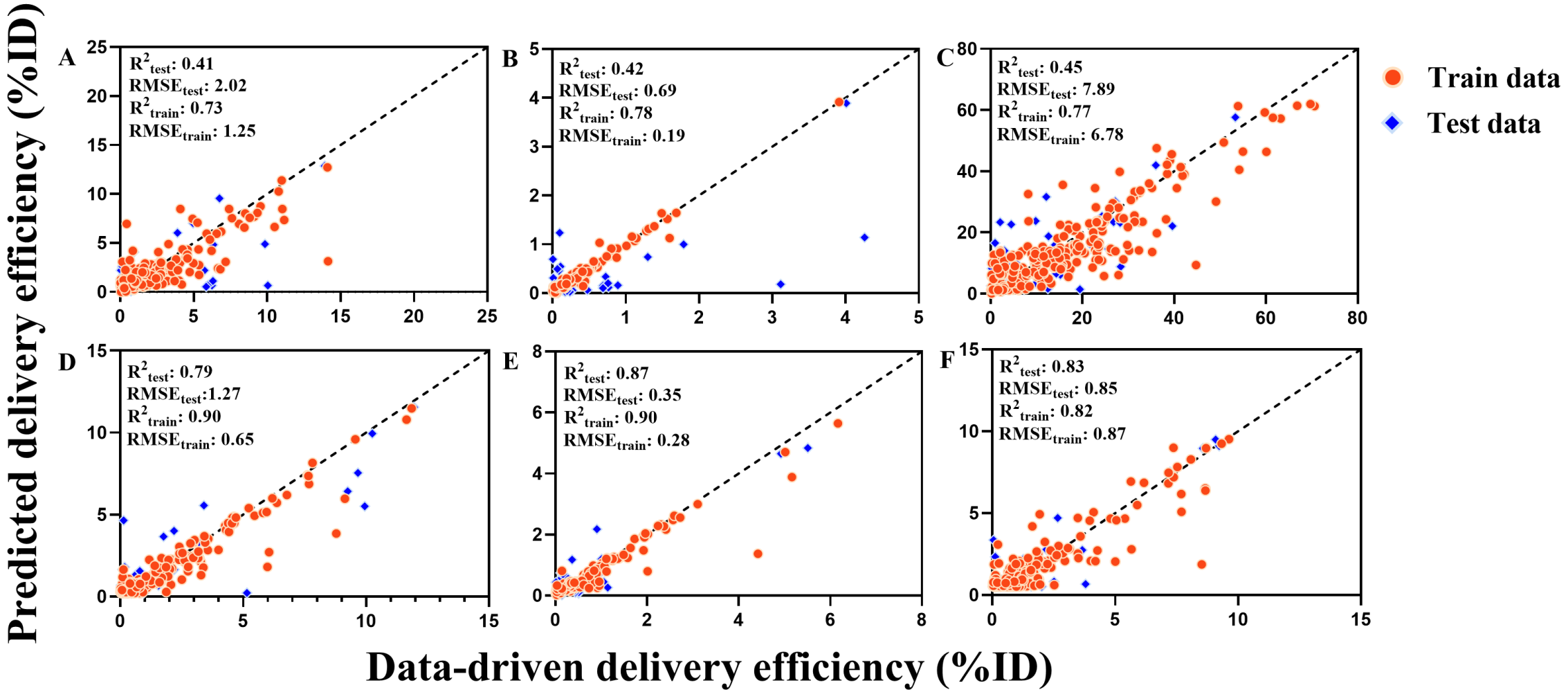


Figure 3. Correlation between observed vs. model-predicted delivery efficiency to tumor (A), 715 heart (B), liver (C), spleen (D), lung (E), and kidney (F) by the deep neural networks (DNN) model.

Mi K, Chou WC, Chen Q, Yuan L, Kamineni VN, Kuchimanchi Y, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2024). Predicting tissue distribution and tumor delivery of nanoparticles in mice using machine learning models. *Journal of Controlled Release*, 374:219-229. (**Selected as the Front Cover Paper**)

Feature Importance Analysis

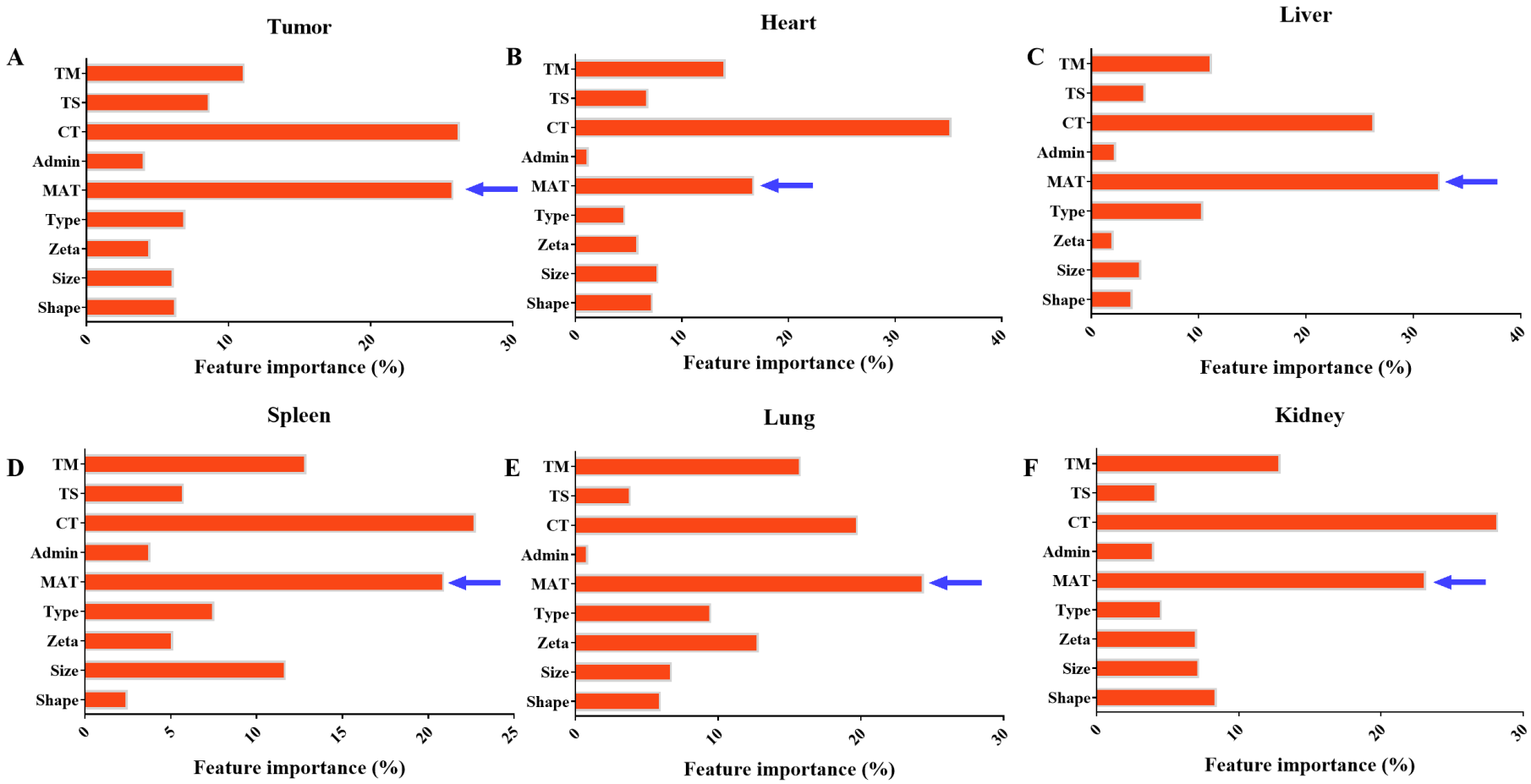
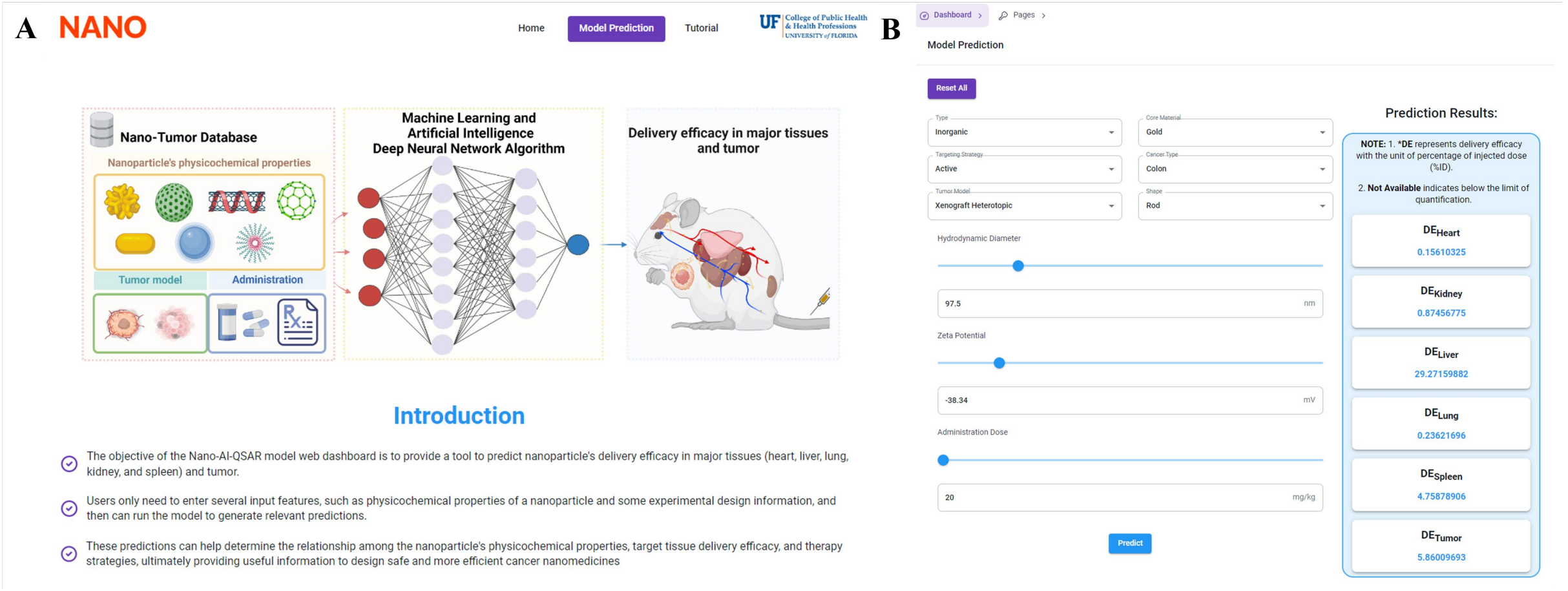


Figure 4. Feature importance of the DNN model in tumor (A), heart (B), liver (C), spleen (D), 719 lungs (E), and kidneys (F). Bar plots represent the final SHAP values. Blue arrows represent the most important contributor to the model predictions among nanoparticles' physicochemical properties. TM, tumor model; TS, targeting strategy; CT, cancer type; Admin, administration dose; MAT, core material of nanoparticles; Type, type of nanoparticles; Zeta, zeta potential; Size, log-transformed value of the hydrodynamic size; Shape, shape of nanoparticles.

Mi K, Chou WC, Chen Q, Yuan L, Kamineni VN, Kuchimanchi Y, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2024). Predicting tissue distribution and tumor delivery of nanoparticles in mice using machine learning models. *Journal of Controlled Release*, 374:219-229. (**Selected as the Front Cover Paper**)

Nano-AI-QSAR Web Dashboard



Mi K, Chou WC, Chen Q, Yuan L, Kamineni VN, Kuchimanchi Y, He C, Monteiro-Riviere NA, Riviere JE, Lin Z. (2024). Predicting tissue distribution and tumor delivery of nanoparticles in mice using machine learning models. *Journal of Controlled Release*, 374:219-229. (**Selected as the Front Cover Paper**)

Summary and Discussion

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

  (1) AI-enabled QSAR models to predict ADMET properties of hundreds of chemicals

  (2) AI-assisted PBPK models for hundreds of chemicals and nanoparticles

  (3) Analyze a large amount of different types of data to generate new insights into toxicity mechanisms rapidly, which was difficult by manual approaches in the past.

- Several challenges and scientific gaps should be considered:

  (1) Rigorous data curation, quality check, and infrastructure to store, share, analyze, evaluate, and manage big data

  (2) Evaluate different machine learning methods to determine the optimal model

  (3) Bioactivity classification (yes/no) vs. the intensity of effect or dose-response relationship

  (4) User-friendly interfaces to facilitate applications of AI-QSAR/PBPK models

  (5) Existing studies are mostly based on data from rodents and humans, with very few in food animals

  (6) There are many studies for small molecular drugs, but fewer for environmental chemicals, and even very limited for nanomaterials

  (7) Education and training the next generation of toxicology students with AI expertise

Acknowledgements

Lab members:

Zhoumeng Lin
Qiran Chen
Xue Wu
Kun Mi
Pei-Yu Wu
Chi-Yun Chen
Venkata Nithin Kamineni
Zhicheng Zhang
Xinyue Chen

Former members:

Wei-Chun Chou
Yi-Hsien Cheng
Miao Li, 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, Ethan Cecil
Gabriel (Guanyu) Tao
Malek Hussein Hajjawi
Yashas Kuchimanchi

Collaborators:

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

Advisors:

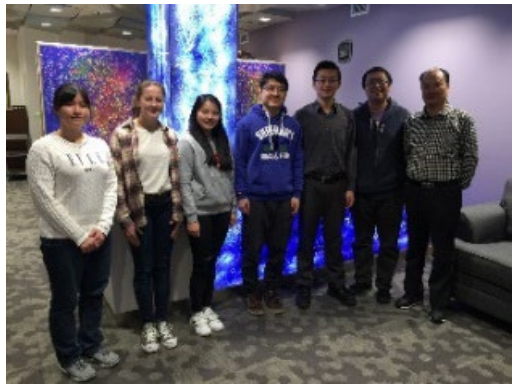
Dr. Jim E. Riviere
Dr. Nancy A. Monteiro-Riviere
Dr. Nikolay M. Filipov
Dr. Jeffrey W. Fisher
Dr. Ronette Gehring

Funding Agencies:

NIH NIBIB
USDA

UGA 2013

KSU 2017



KSU Lab 2019

UF Lab 2023

National FARAD 2023