Using Chemical Structure Information to Develop Predictive Models for In Vitro Toxicokinetic Parameters to Inform High Throughput Risk Assessment

Prachi Pradeep

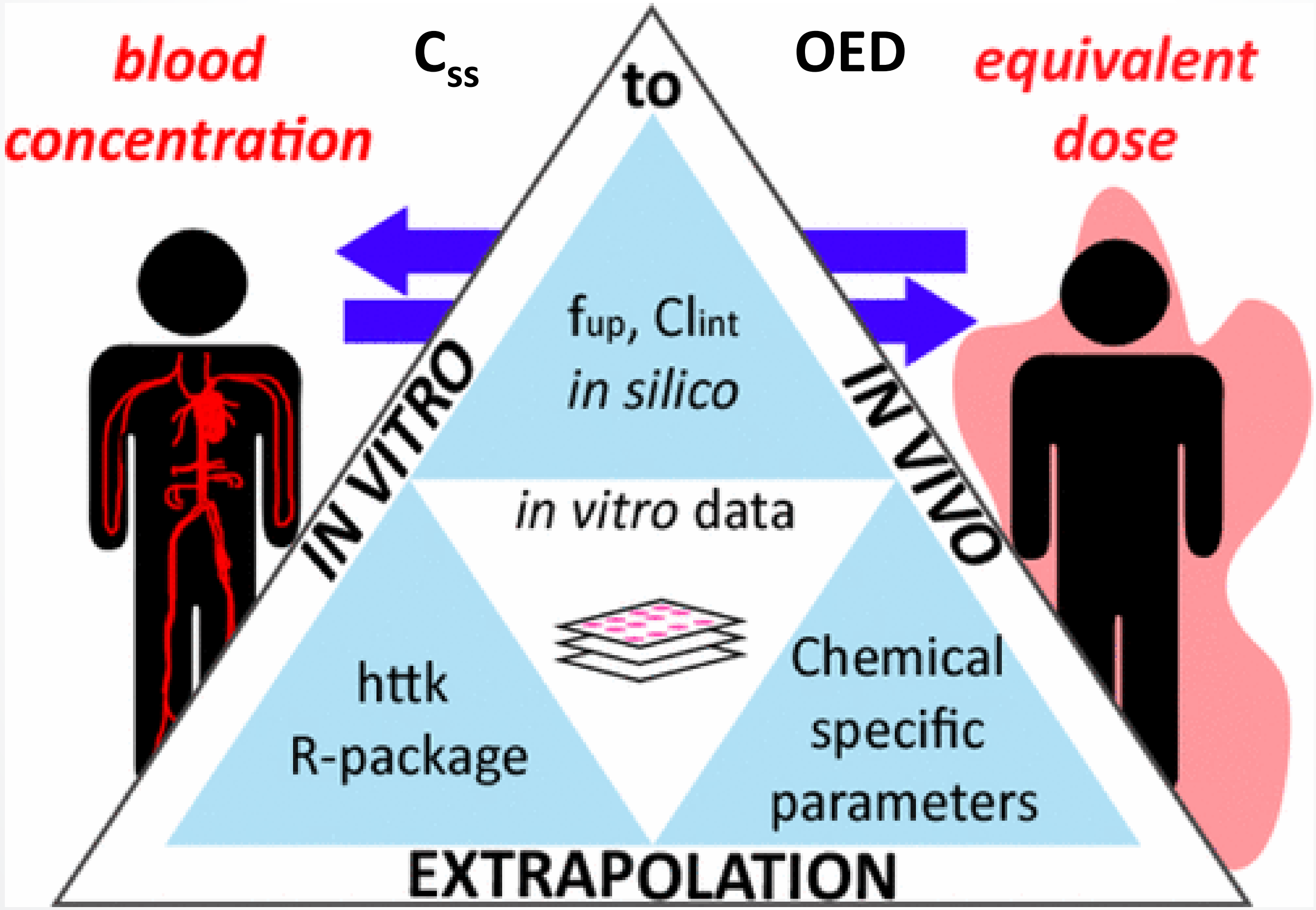
US EPA, National Center for Computational Toxicology

Society of Toxicology Annual Meeting 2019

Workshop: Predicting Metabolic Clearance Rates for Drug Leads and Chemical Risk Assessment

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Background Context



Sipes et al, 2017

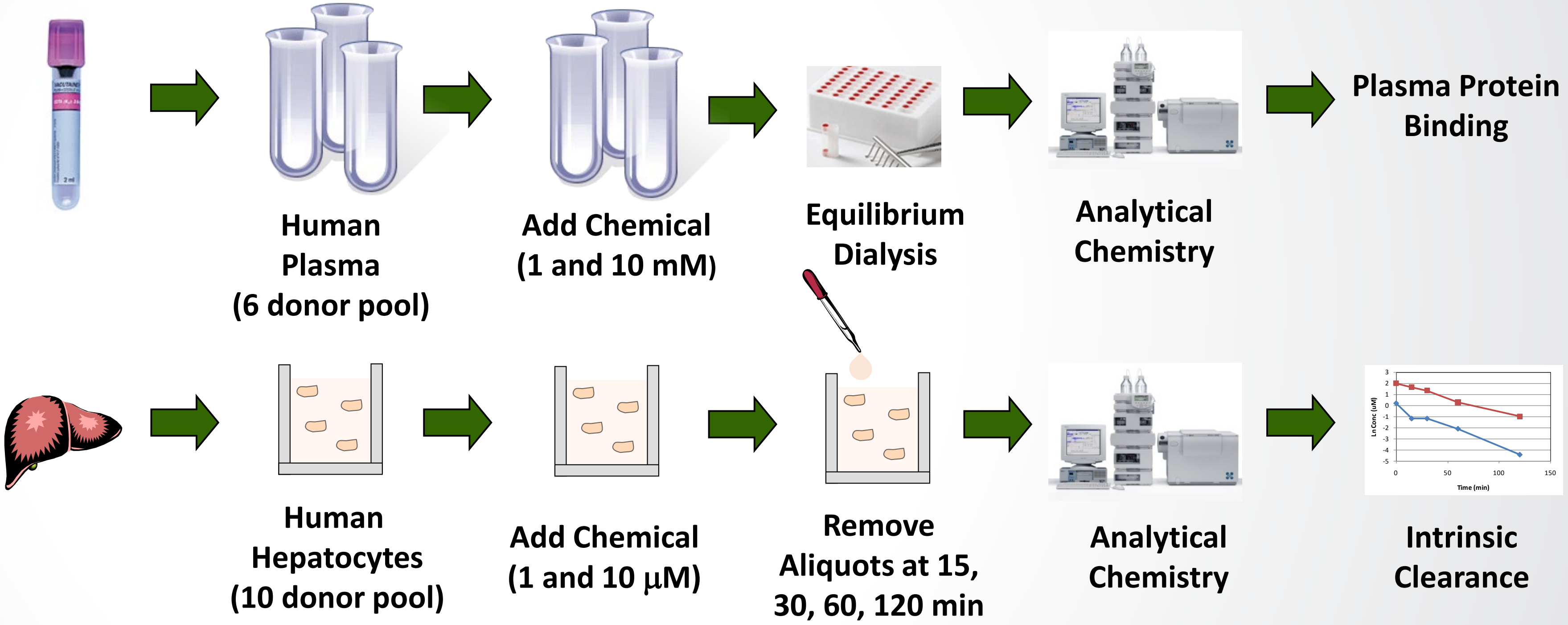
f_{up}: Fraction Unbound in Plasma
Clint: Intrinsic Clearance

Toxicokinetic Parameters

Slide courtesy: Richard Judson
John Wambaugh

1. Fraction Unbound in Plasma

2. Intrinsic Clearance

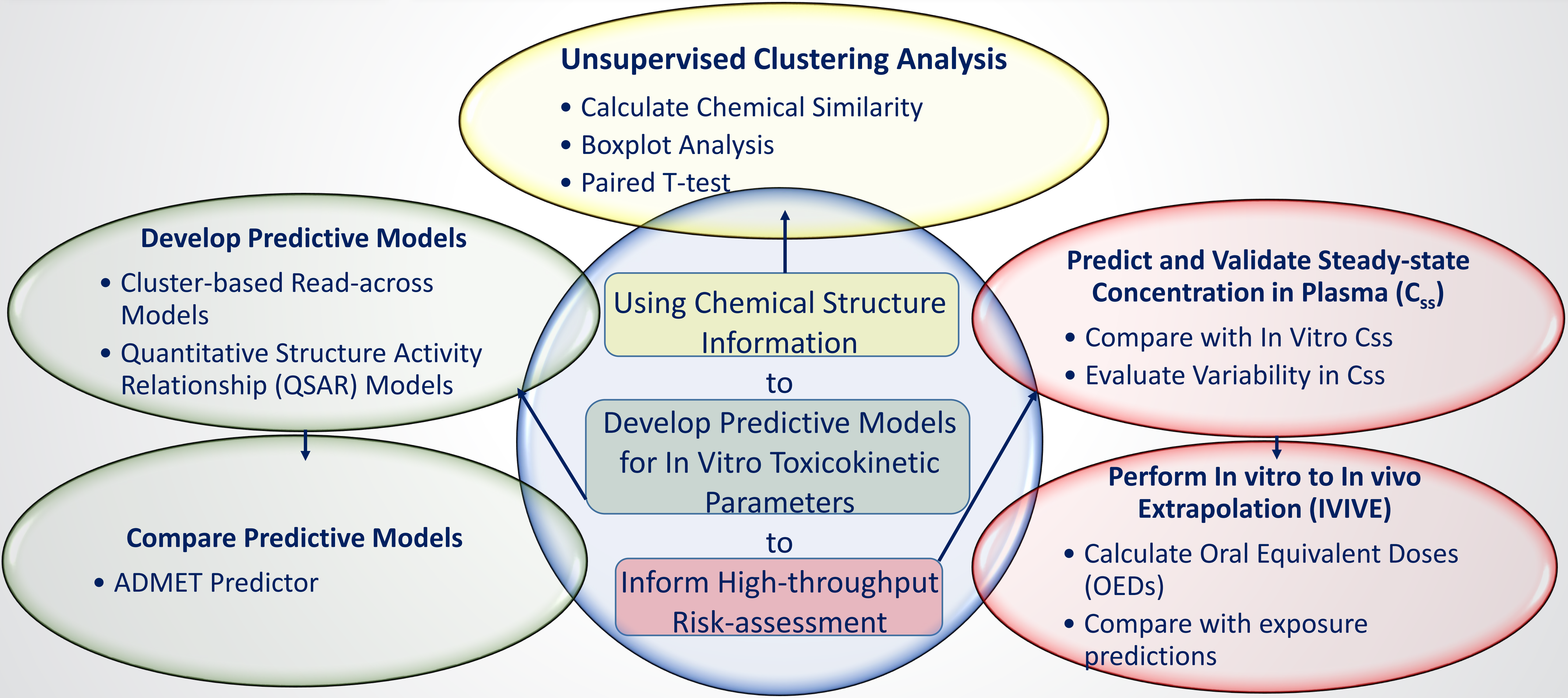


Human Plasma (6 donor pool) → Add Chemical (1 and 10 mM) → Equilibrium Dialysis → Analytical Chemistry → Plasma Protein Binding

Human Hepatocytes (10 donor pool) → Add Chemical (1 and 10 μM) → Remove Aliquots at 15, 30, 60, 120 min → Analytical Chemistry → Intrinsic Clearance

- ‼ Not high-throughput (~800 chemicals in 10yrs)
- ‼ ~7000$ per chemical

Outline



Unsupervised Clustering Analysis

- Calculate Chemical Similarity
- Boxplot Analysis
- Paired T-test

Develop Predictive Models

- Cluster-based Read-across Models
- Quantitative Structure Activity Relationship (QSAR) Models

Compare Predictive Models

- ADMET Predictor

Using Chemical Structure Information

to

Develop Predictive Models for In Vitro Toxicokinetic Parameters

to

Inform High-throughput Risk-assessment

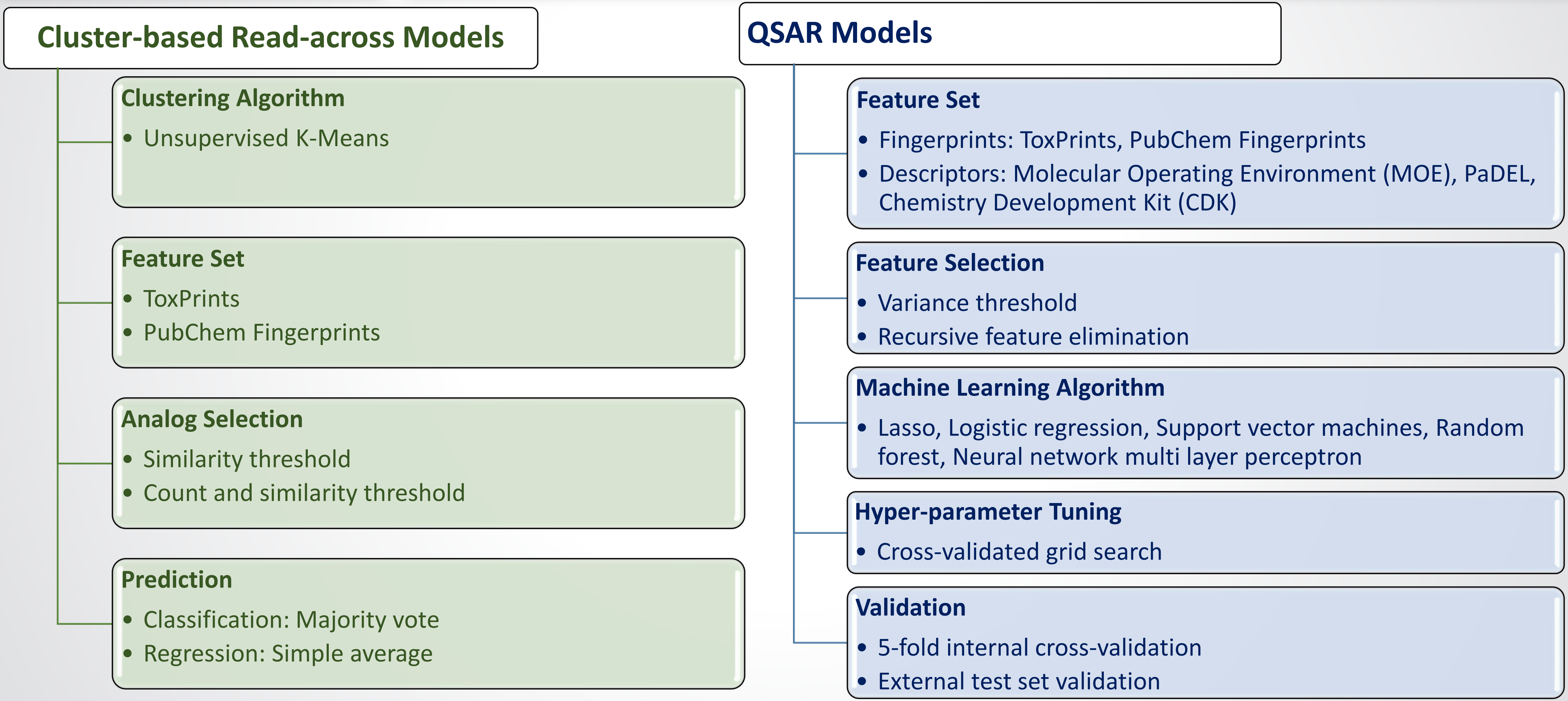
Predict and Validate Steady-state Concentration in Plasma (C_{ss})

- Compare with In Vitro Css
- Evaluate Variability in Css

Perform In vitro to In vivo Extrapolation (IVIVE)

- Calculate Oral Equivalent Doses (OEDs)
- Compare with exposure predictions

Development of Predictive Models



Cluster-based Read-across Models

Clustering Algorithm

- Unsupervised K-Means

Feature Set

- ToxPrints
- PubChem Fingerprints

Analog Selection

- Similarity threshold
- Count and similarity threshold

Prediction

- Classification: Majority vote
- Regression: Simple average

QSAR Models

Feature Set

- Fingerprints: ToxPrints, PubChem Fingerprints
- Descriptors: Molecular Operating Environment (MOE), PaDEL, Chemistry Development Kit (CDK)

Feature Selection

- Variance threshold
- Recursive feature elimination

Machine Learning Algorithm

- Lasso, Logistic regression, Support vector machines, Random forest, Neural network multi layer perceptron

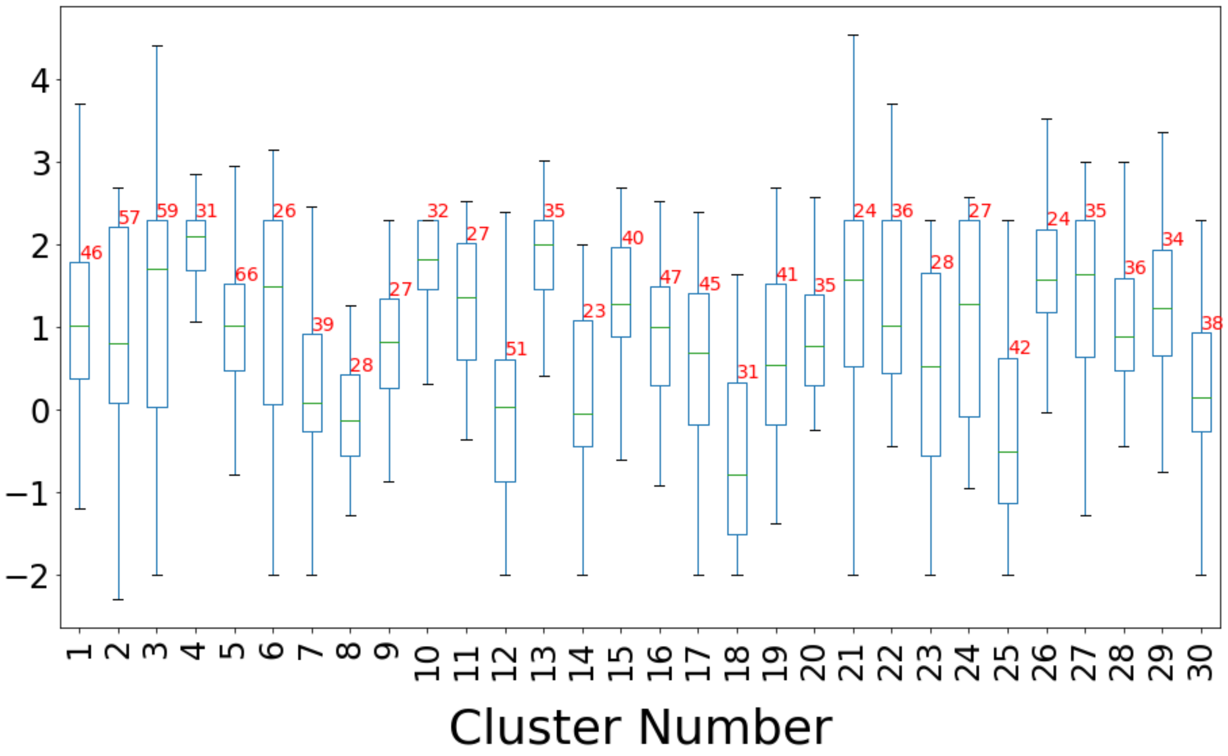
Hyper-parameter Tuning

- Cross-validated grid search

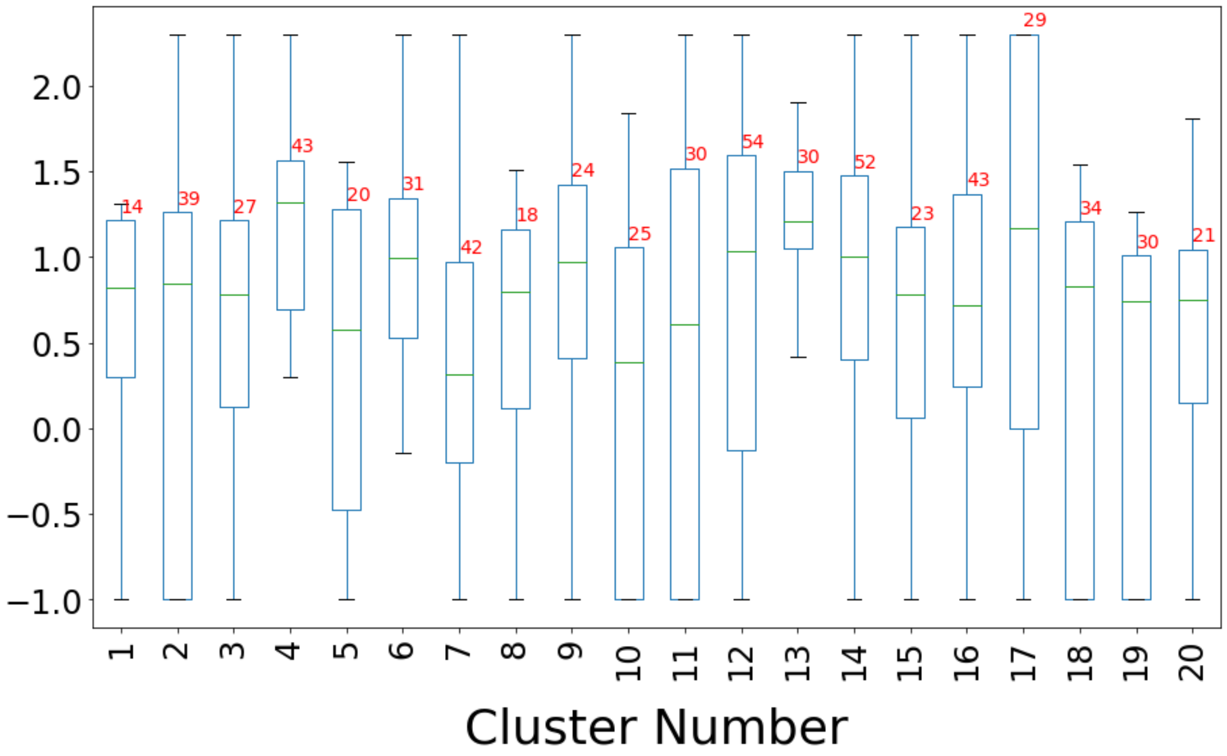
Validation

- 5-fold internal cross-validation
- External test set validation

Unsupervised Clustering Analysis



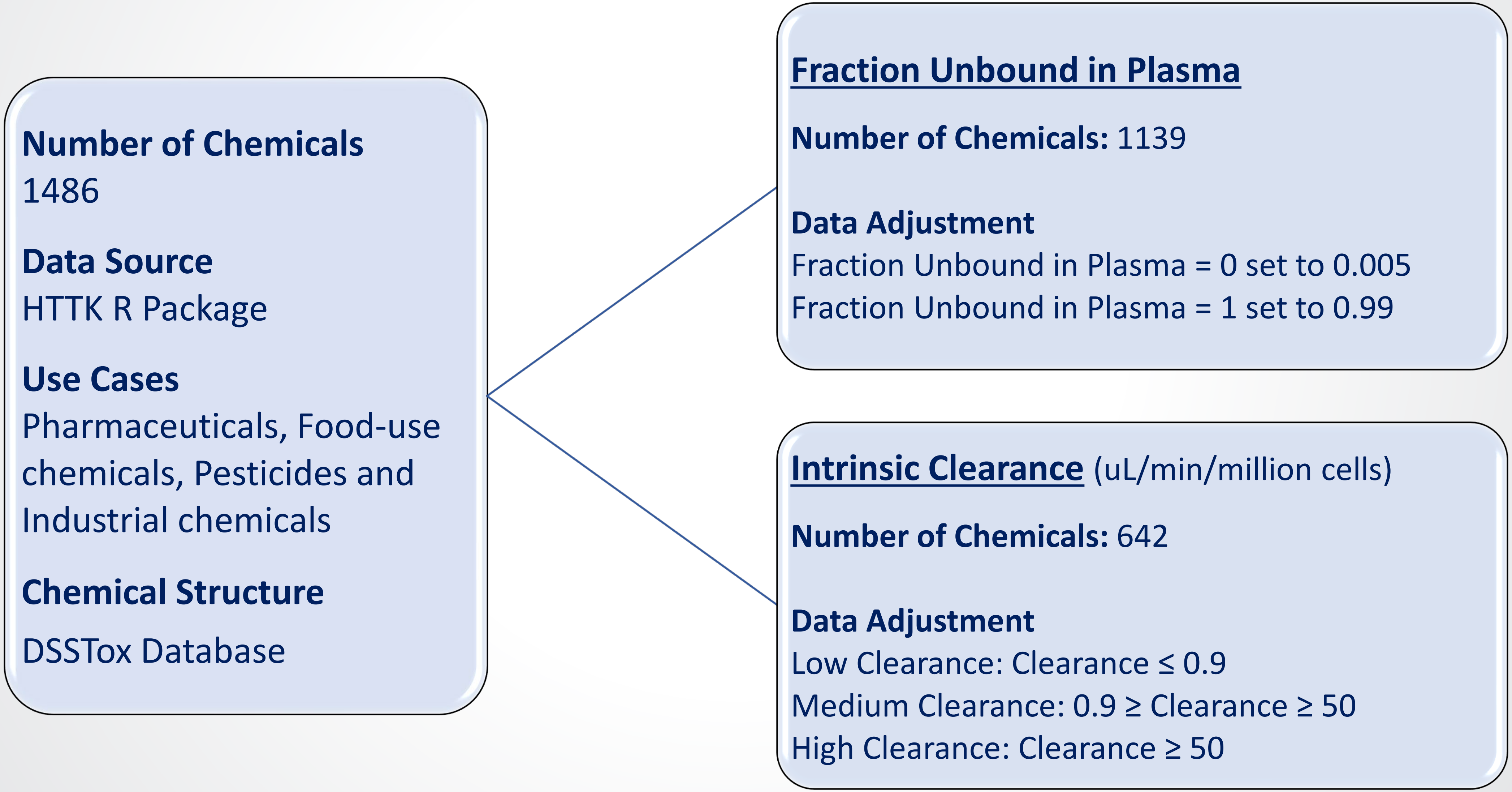
Fraction Unbound in Plasma



Intrinsic Clearance

- Range of fraction unbound in plasma is much more tightly bound across different clusters as compared to intrinsic clearance
- Paired T-test illustrates that mean fraction unbound in plasma values are more distinct across clusters as compared to intrinsic clearance

Dataset



Number of Chemicals
1486

Data Source
HTTK R Package

Use Cases
Pharmaceuticals, Food-use chemicals, Pesticides and Industrial chemicals

Chemical Structure
DSSTox Database

Fraction Unbound in Plasma

Number of Chemicals: 1139

Data Adjustment
Fraction Unbound in Plasma = 0 set to 0.005
Fraction Unbound in Plasma = 1 set to 0.99

Intrinsic Clearance (uL/min/million cells)

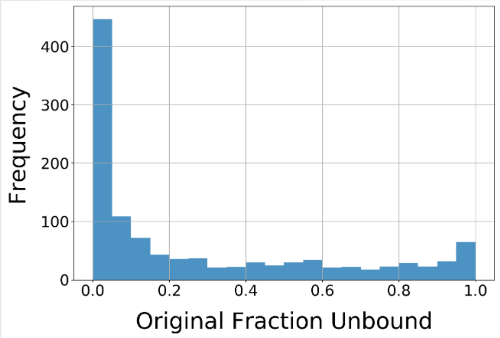
Number of Chemicals: 642

Data Adjustment
Low Clearance: Clearance ≤ 0.9
Medium Clearance: 0.9 ≥ Clearance ≥ 50
High Clearance: Clearance ≥ 50

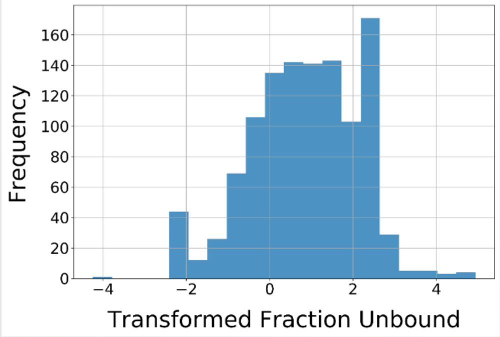
Data Transformation for Predictive Models

Fraction Unbound in Plasma

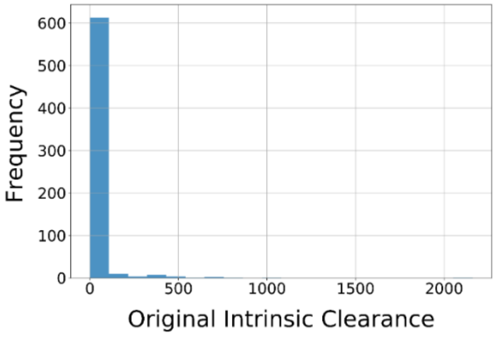
Regression Models

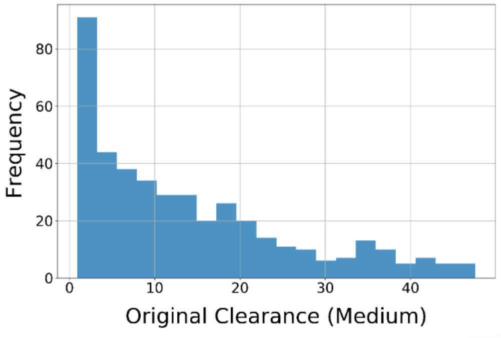


Transformation $Y' = log_{10}\frac{(1-Y)}{Y}$

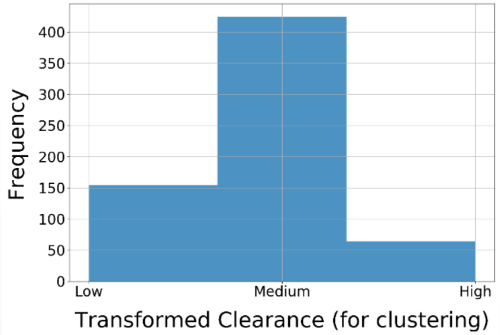


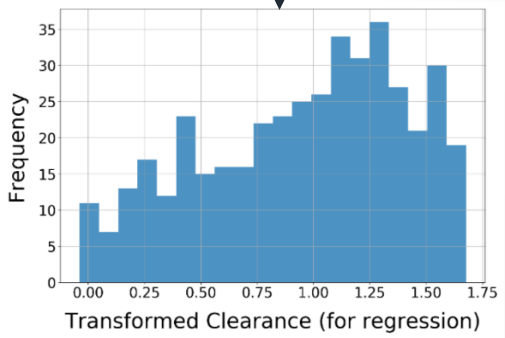
Intrinsic Clearance

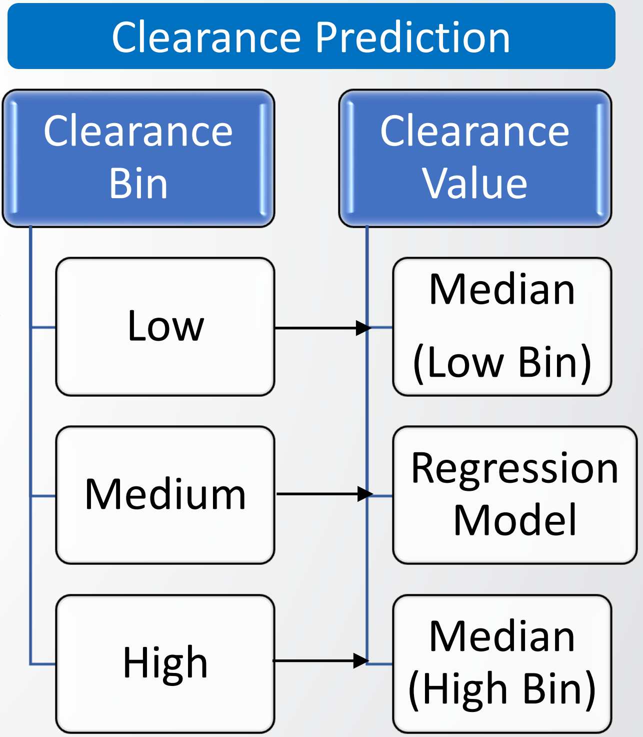




Transformation $Y' = log_{10}Y$







Clearance Prediction

Clearance Bin

Clearance Value

Low

Median (Low Bin)

Medium

Regression Model

High

Median (High Bin)

QSAR Models : Fraction Unbound in Plasma

DESCRIPTORS USED (number)	MODEL	5-FOLD INTERNAL CROSS-VALIDATION				EXTERNAL VALIDATION			
		MAE	RMSE	RMSE/σ	R^2	MAE	RMSE	RMSE/σ	R^2
Pubchem + ToxPrints (79)	Lasso regression	0.80	1.03	0.81	0.34	0.7	0.91	0.73	0.47
	Support vector regression	0.74	0.95	0.75	0.44	0.62	0.87	0.70	0.51
	Random Forest	0.75	0.97	0.76	0.42	0.65	0.89	0.71	0.49
	MLP Regression	0.76	0.98	0.78	0.40	0.68	0.89	0.72	0.48
	Consensus (SVM, RF)	0.74	0.95	0.75	0.44	0.63	0.87	0.70	0.51
Pubchem + ToxPrints (79) + MOE (3)	Lasso regression	0.68	0.90	0.72	0.48	0.69	0.89	0.68	0.54
	Support vector regression	0.62	0.84	0.67	0.55	0.66	0.86	0.66	0.57
	Random Forest	0.62	0.84	0.67	0.56	0.65	0.86	0.66	0.56
	MLP Regression	0.66	0.88	0.70	0.51	0.69	0.88	0.67	0.55
	Consensus (SVM, RF)	0.60	0.81	0.65	0.58	0.64	0.84	0.64	0.59
Pubchem + ToxPrints (79) + MOE (3) + PaDEL + CDK (10)	Lasso regression	0.66	0.87	0.70	0.51	0.70	0.90	0.68	0.53
	Support vector regression	0.59	0.82	0.65	0.57	0.64	0.84	0.64	0.59
	Random Forest	0.61	0.83	0.67	0.55	0.64	0.84	0.64	0.59
	MLP Regression	0.64	0.85	0.68	0.54	0.7	0.91	0.69	0.52
	Consensus (SVM, RF)	**0.58**	**0.80**	**0.64**	**0.59**	**0.62**	**0.82**	**0.62**	**0.61**

QSAR Models: Intrinsic Clearance (Classification)

DESCRIPTORS USED (number)	MODEL	5-FOLD INTERNAL CROSS-VALIDATION		EXTERNAL VALIDATION	
		Accuracy	F1 score	Accuracy	F1 Score
Pubchem + ToxPrints (57)	Logistic regression	67.59	[0.00, 0.81, 0.00]	61.90	[0.00, 0.76, 0.00]
	Support vector classification	69.78	[0.21, 0.82, 0.08]	64.29	[0.11, 0.78, 0.14]
	Random Forest	69.38	[0.31, 0.81, 0.40]	64.29	[0.24, 0.77, 0.13]
	MLP Classification	67.59	[0.00, 0.81, 0.00]	63.49	[0.15, 0.77, 0.00]
Pubchem + ToxPrints (57) + MOE (3)	Logistic regression	71.17	[0.38, 0.82, 0.04]	66.67	[0.29, 0.79, 0.00]
	Support vector classification	72.17	[0.43, 0.82, 0.11]	65.87	[0.31, 0.78, 0.14]
	Random Forest	**71.57**	**[0.41, 0.82, 0.38]**	**65.87**	**[0.37, 0.77, 0.13]**
	MLP Classification	68.79	[0.40, 0.80, 0.04]	61.11	[0.42, 0.73, 0.09]
Pubchem + ToxPrints (57) + MOE (3) + PaDEL + CDK (10)	Logistic regression	70.78	[0.36, 0.82, 0.00]	65.87	[0.25, 0.78, 0.00]
	Support vector classification	71.97	[0.39, 0.82, 0.18]	66.67	[0.29, 0.79, 0.14]
	Random Forest	72.37	[0.42, 0.82, 0.41]	64.29	[0.28, 0.77, 0.13]
	MLP Classification	70.78	[0.36, 0.82, 0.04]	61.11	[0.38, 0.73, 0.10]

QSAR Models: Intrinsic Clearance (Regression)

DESCRIPTORS USED (number)	MODEL	5-FOLD INTERNAL CROSS-VALIDATION				EXTERNAL VALIDATION			
		MAE	RMSE	RMSE/σ	R^2	MAE	RMSE	RMSE/σ	R^2
Pubchem + ToxPrints (53)	Lasso regression	0.38	0.44	1.00	-0.01	0.41	0.48	1.00	0.00
	Support vector regression	0.37	0.44	0.99	0.02	0.38	0.46	0.96	0.08
	Random Forest	0.37	0.45	1.01	-0.02	0.38	0.46	0.97	0.06
	MLP Regression	0.37	0.45	1.02	-0.04	0.40	0.48	1.00	0.00
	Consensus (SVM, RF)	0.37	0.44	0.99	0.02	0.38	0.46	0.96	0.09
Pubchem + ToxPrints (53) + MOE (3)	Lasso regression	0.37	0.44	0.98	0.03	0.39	0.47	0.98	0.04
	Support vector regression	0.36	0.43	0.97	0.06	0.37	0.45	0.94	0.12
	Random Forest	**0.34**	**0.42**	**0.95**	**0.09**	**0.34**	**0.43**	**0.90**	**0.20**
	MLP Regression	0.37	0.45	1.03	-0.06	0.39	0.48	1.00	0.00
	Consensus (SVM, RF)	0.35	0.42	0.94	0.11	0.36	0.44	0.92	0.15
Pubchem + ToxPrints (53) + MOE (3) + PaDEL + CDK (10)	Lasso regression	0.37	0.43	0.98	0.05	0.39	0.47	0.98	0.05
	Support vector regression	0.35	0.43	0.97	0.06	0.37	0.46	0.97	0.06
	Random Forest	0.34	0.42	0.94	0.12	0.34	0.43	0.90	0.20
	MLP Regression	0.37	0.48	1.08	-0.16	0.43	0.55	1.16	-0.34
	Consensus (SVM, RF)	0.35	0.42	0.94	0.11	0.37	0.45	0.94	0.12

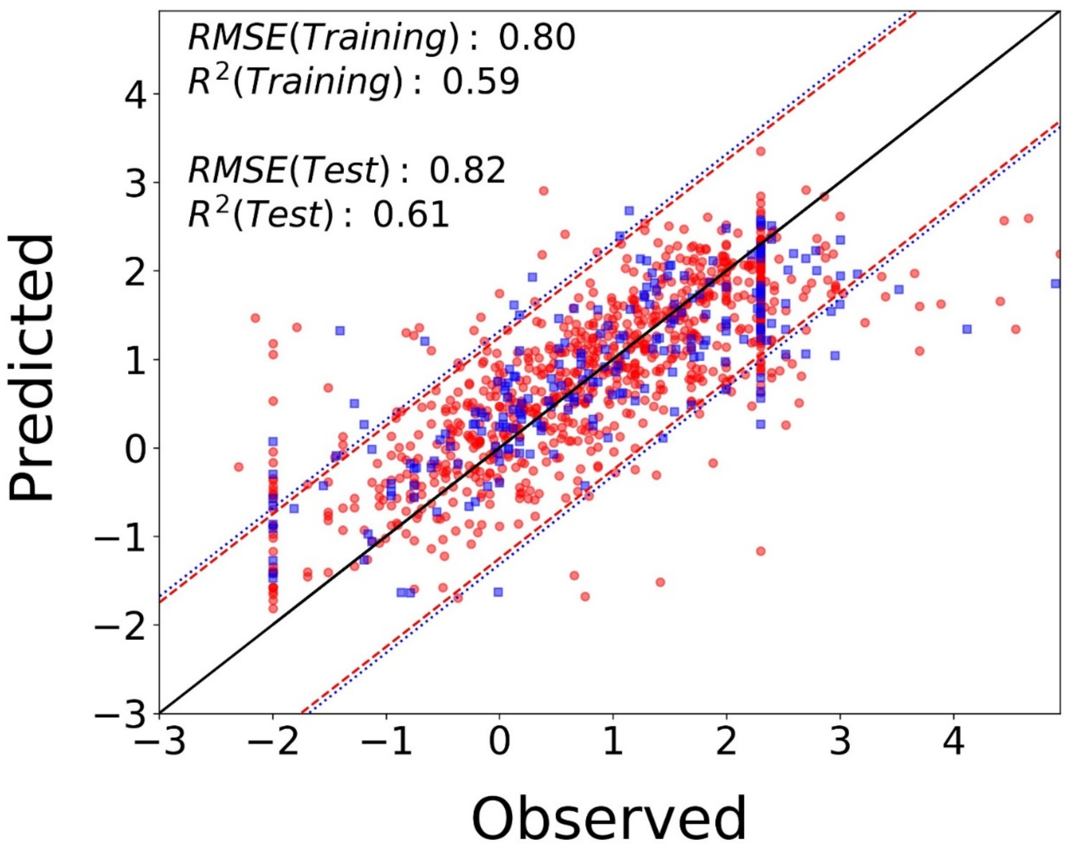
Final Model: Fraction Unbound in Plasma

Observed versus predicted fraction unbound (transformed scale) for 5-fold internal cross-validation (red dots) and external test set validation (blue squares).

Final Model
Consensus of Random Forest and Support Vector Machine

5-fold internal cross-validation
RMSE = 0.80
$R^2 = 0.59$
External test set validation
RMSE = 0.82
$R^2 = 0.61$

Black solid line: Line of perfect fit, where the predicted values would equal the experimental values.
Red dashed lines: Error margin of ±1 standard deviation of the training dataset
Blue dotted lines: Error margin of ±1 standard deviation of the test dataset.



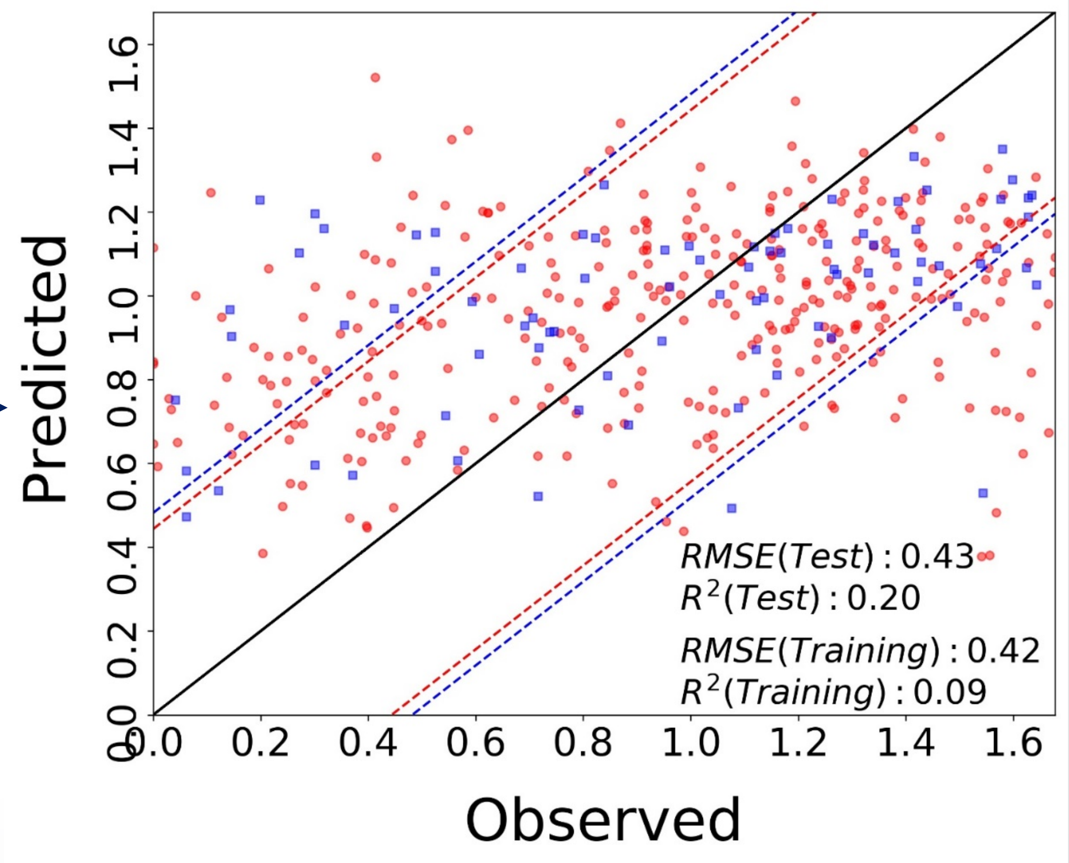
Final Model: Intrinsic Clearance

Observed versus predicted medium intrinsic clearance (transformed scale) for 5-fold internal cross-validation (red dots) and external test set validation (blue squares)

Final Model
Random Forest

5-fold internal cross-validation
RMSE = 0.42
$R^2 = 0.09$
External test set validation
RMSE = 0.43
$R^2 = 0.20$

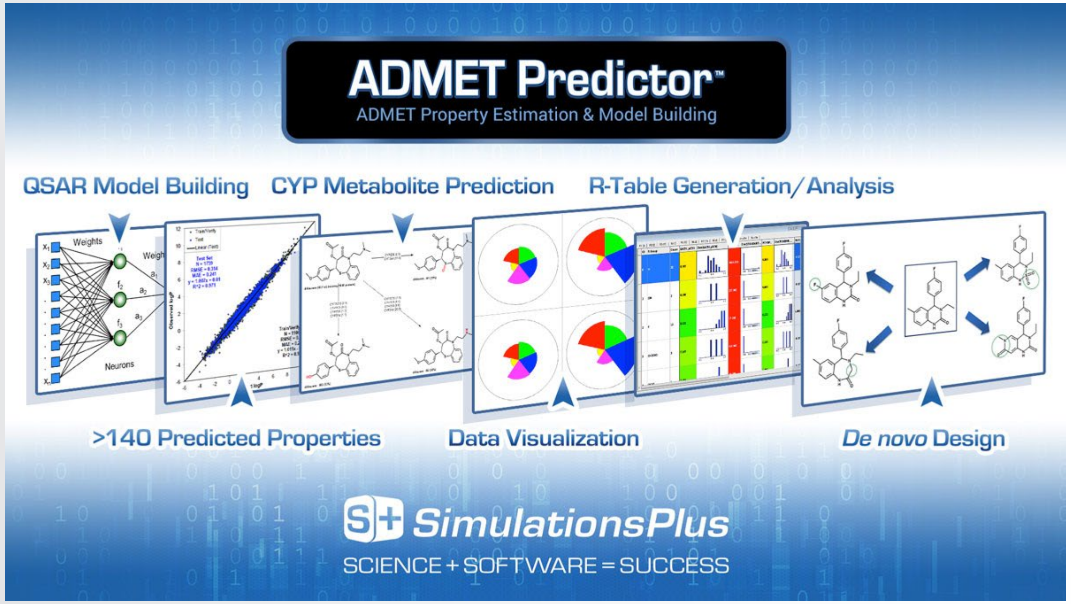
Black solid line: Line of perfect fit, where the predicted values would equal the experimental values.
Red dashed lines: Error margin of ±1 standard deviation of the training dataset
Blue dotted lines: Error margin of ±1 standard deviation of the test dataset.



Comparison with ADMET Predictor

ADMET Predictor™ 7.2

(Simulations Plus Inc., Lancaster, CA).


ADMET Predictor™
ADMET Property Estimation & Model Building
QSAR Model Building
CYP Metabolite Prediction
R-Table Generation/Analysis
>140 Predicted Properties
Data Visualization
De novo Design
SimulationsPlus
SCIENCE + SOFTWARE = SUCCESS

External dataset
1814 chemicals tested in a battery of Estrogen Receptor and Androgen Receptor assays (Kleinstreuer et al, 2017 and Judson et al, 2015)

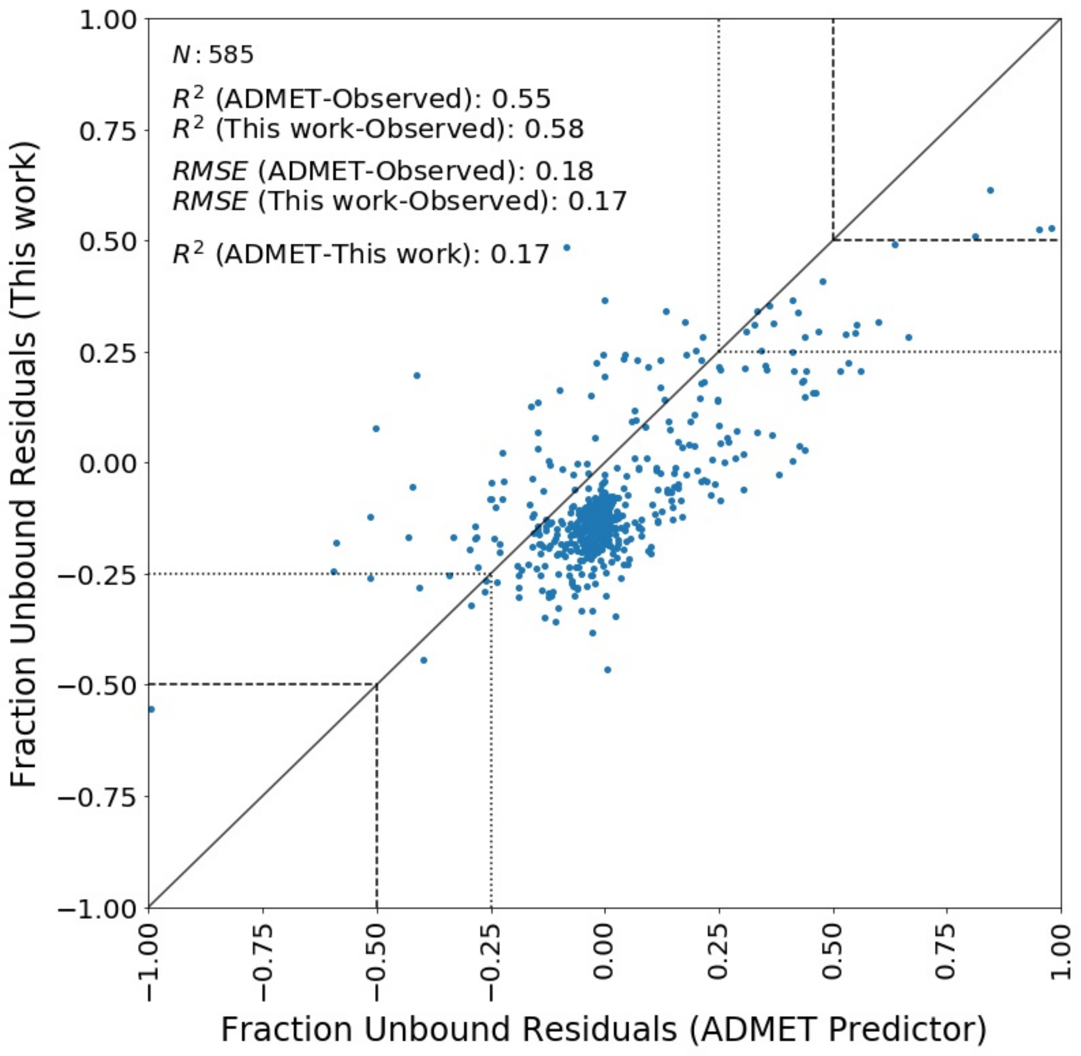
ADMET Predictions
Sipes et al, 2017

Final Common Dataset
Fraction Unbound in Plasma: 585 chemicals
Intrinsic Clearance: 515 chemicals

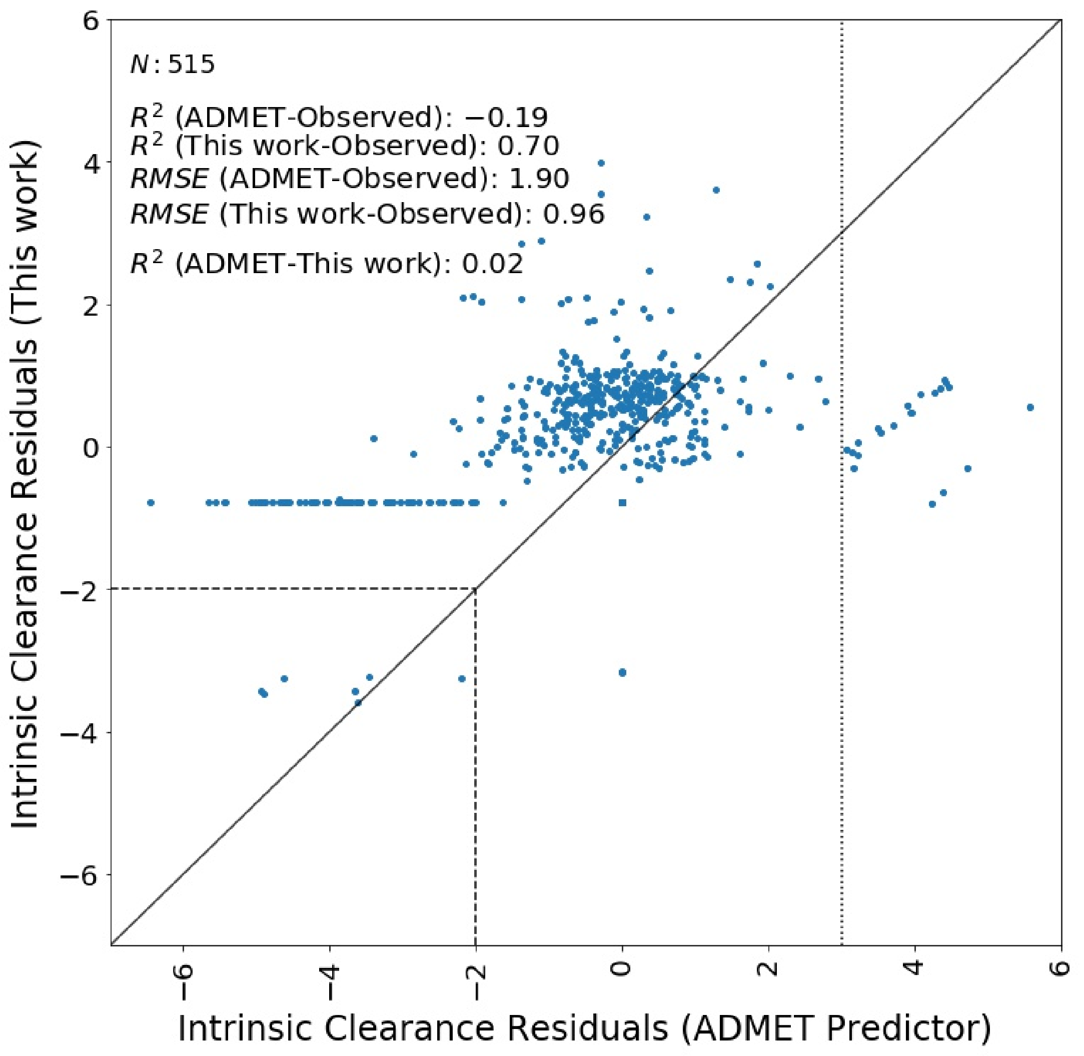
Residual Comparison Plot
Residual = Experimental – Predicted

Comparison with ADMET Predictor

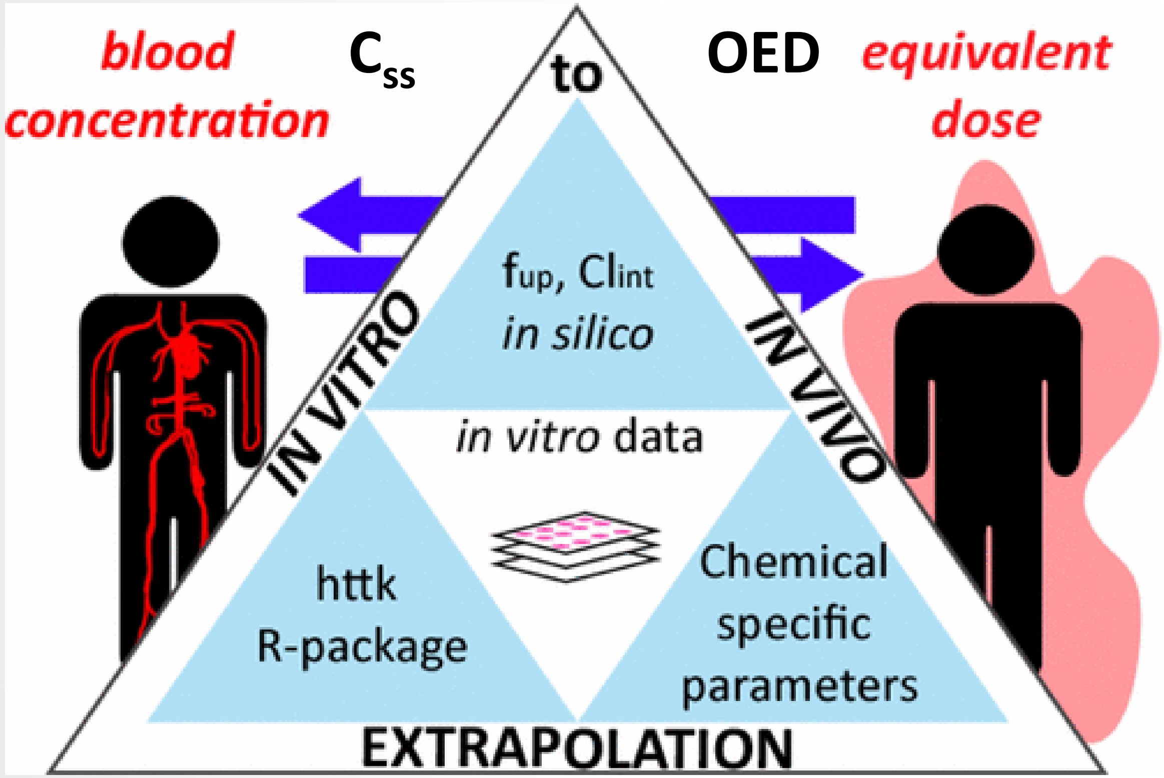
Fraction Unbound in Plasma



Intrinsic Clearance



Background Context



Sipes et al, 2017

Compare

Exposure Estimates

Input to Risk Assessment

f_{up}: Fraction Unbound in Plasma
Clint: Intrinsic Clearance

Calculation of Human Oral Equivalent Doses (OEDs) and Comparison with Exposure Predictions

Calculation of OEDs

$$OED = \frac{Activating\ Concentration\ In\ Vitro\ (ACC)}{C_{ss}}$$

where,
ACC is derived from data across 18 ER and 11 AR assays

3 estimates of OEDs

1. Conservative estimate of OED based on *in vitro* C_{ss}
2. Conservative estimate of OED based on *in silico* C_{ss}
3. Conservative estimate of OED based on variation in *in silico* C_{ss} due to physchem properties

Comparison with Exposure Predictions

EPA's ExpoCast estimates

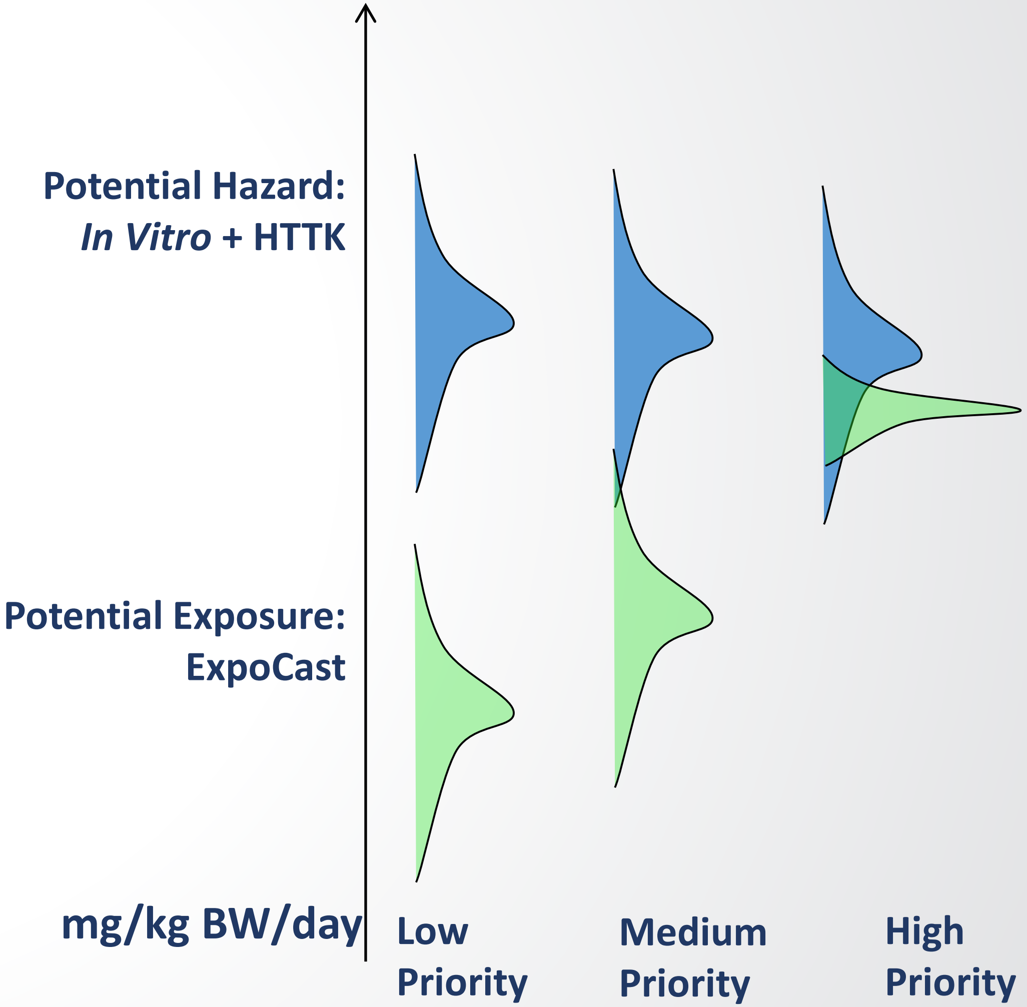


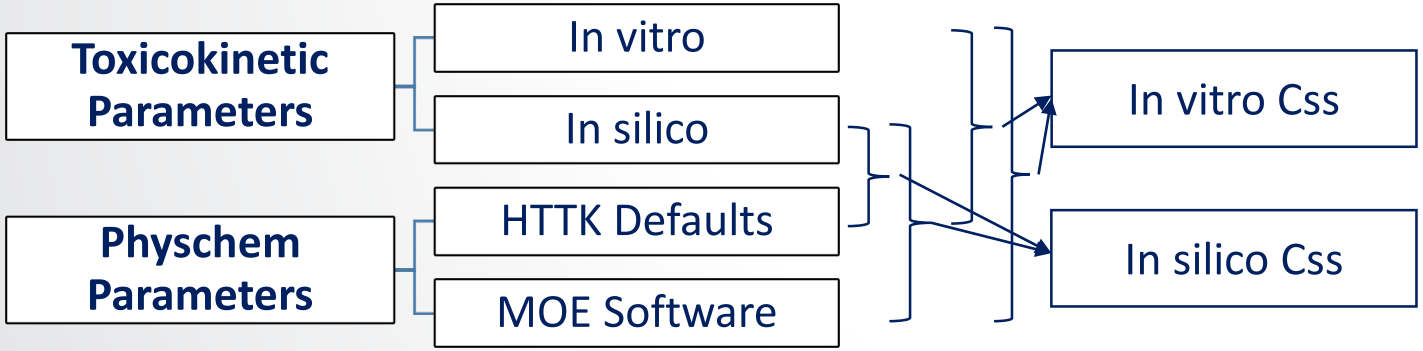
Figure Courtesy: Richard Judson

Prediction and Comparison of C_{ss}

C_{ss} is the steady-state concentration of a chemical in the plasma given a constant 1 mg/kg/day oral dose

Experimental In vitro C_{ss} Values: HTTK R Package (709 chemicals)
Predicted In Silico C_{ss} Values; HTTK R Package

4 C_{ss} values were calculated



Toxicokinetic Parameters

In vitro

In silico

Physchem Parameters

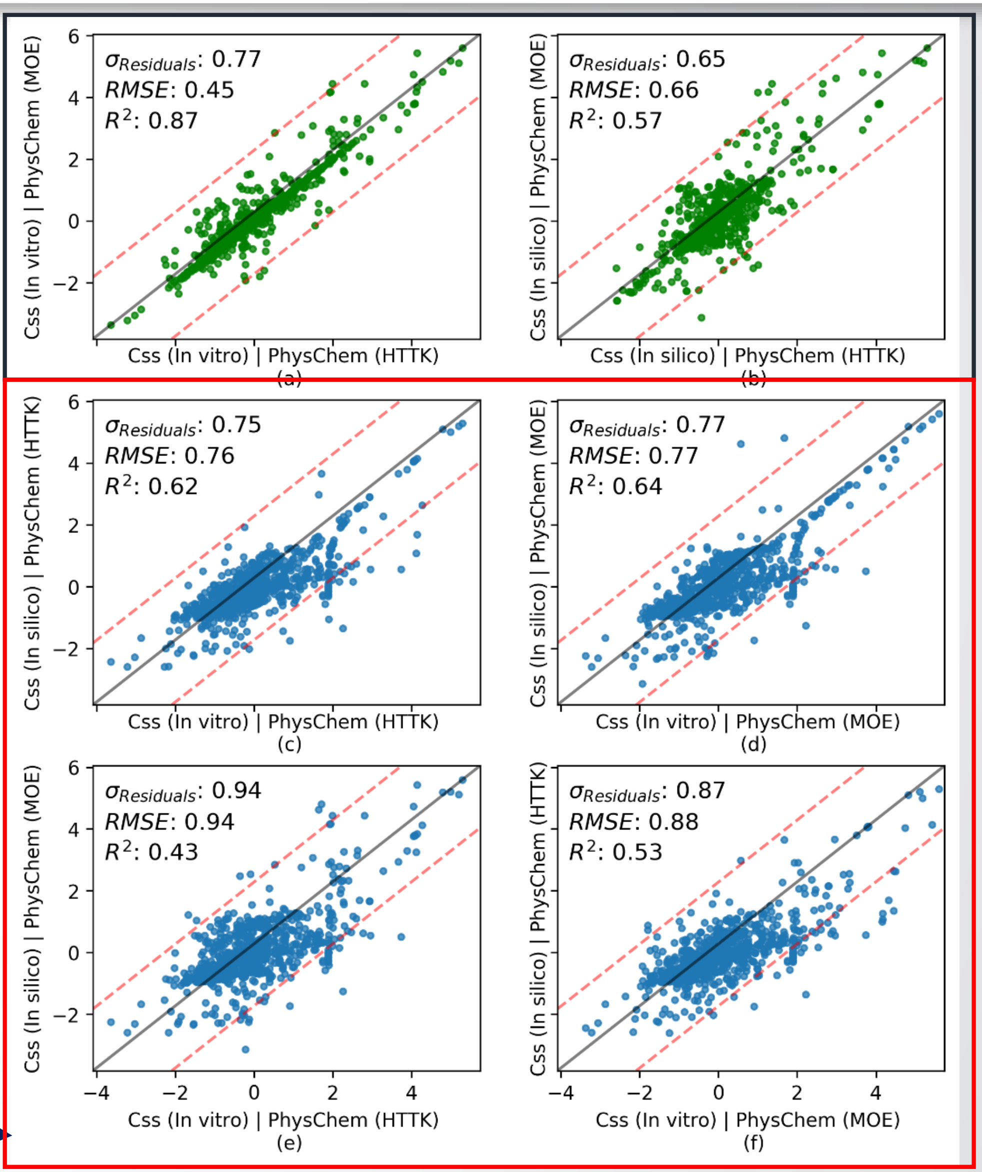
HTTK Defaults

MOE Software

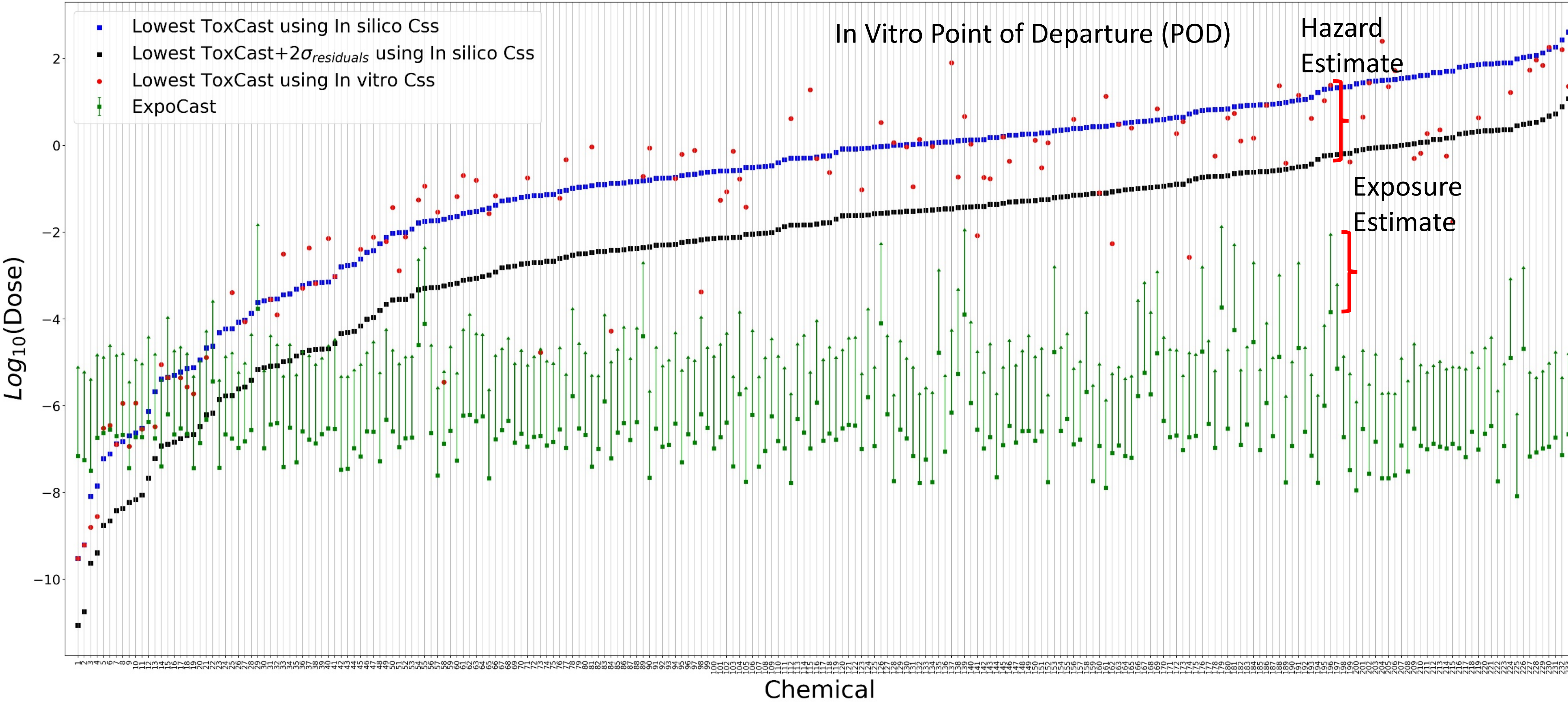
In vitro Css

In silico Css

Effect of variability in physicochemical properties values on C_{ss} calculations. The C_{ss} units are $\log_{10}$ mg/kg.



Bioactivity-exposure Ratio Plot (ER and AR Bioactivity)



Conclusions

- Unsupervised clustering analysis demonstrates that fraction unbound is structurally more predictable than intrinsic clearance
- A range of predictive models (Read-across and QSAR) were developed for fraction unbound in plasma and intrinsic clearance using a simple descriptor space and a rich chemical dataset
    - Fraction unbound: External test set RMSE = 0.82 and R^2 = 0.61
    - Intrinsic clearance (Classification): Accuracy = 65.87%
    - Intrinsic clearance (Regression): External test set RMSE = 0.43 and R^2 = 0.20
    - The models were benchmarked against commercially available ADMET software
- The model predictions were used to calculate steady-state plasma (C_{ss}) concentrations using an example dataset tested for ER and AR bioactivity
    - Variability in C_{ss} values due to variation in source of physicochemical properties was evaluated
- A range of conservative oral equivalent doses (OEDs) were calculated to allow for a conservative comparison with exposure predictions

Overall, these models and the analysis presented in this work allow prioritization of data-poor chemicals using *in silico* predictions and in vitro to in vivo extrapolation (IVIVE) methods along with high-throughput exposure predictions to facilitate rapid risk-assessment.

Acknowledgements

NCCT
Grace Patlewicz
Robert Pearce
John Wambaugh
Richard Judson

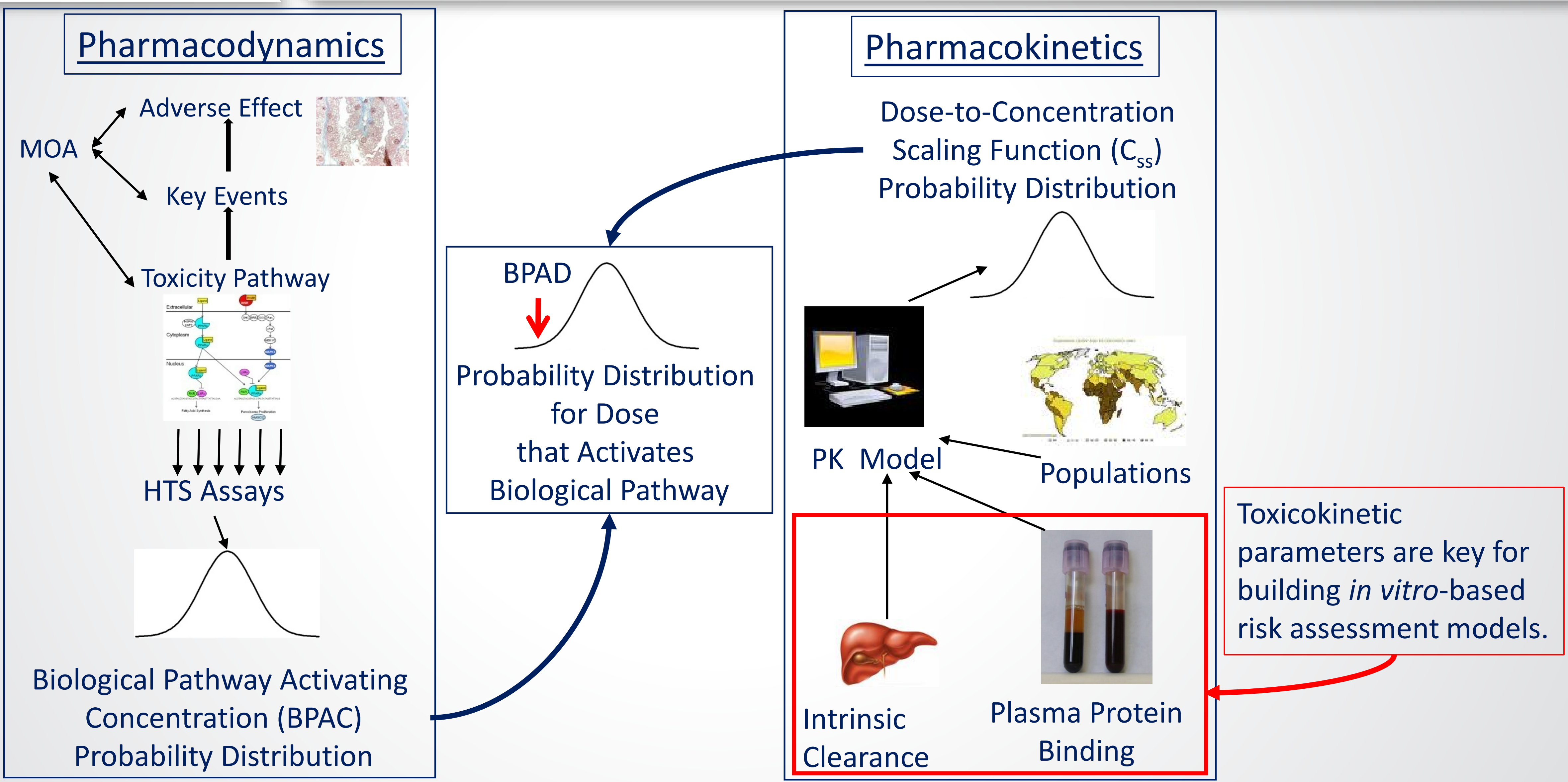
NERL
Barbara Wetmore

ORISE participant research program supported by an interagency agreement between the US EPA and DOE.

EXTRA SLIDES

High Throughput Risk Assessment

Slide courtesy: Richard Judson



Pharmacodynamics

MOA

Adverse Effect

Key Events

Toxicity Pathway

HTS Assays

Biological Pathway Activating Concentration (BPAC) Probability Distribution

BPAD

Probability Distribution for Dose that Activates Biological Pathway

Pharmacokinetics

Dose-to-Concentration Scaling Function (C_{ss}) Probability Distribution

PK Model

Populations

Intrinsic Clearance

Plasma Protein Binding

Toxicokinetic parameters are key for building *in vitro*-based risk assessment models.

Unsupervised clustering analysis

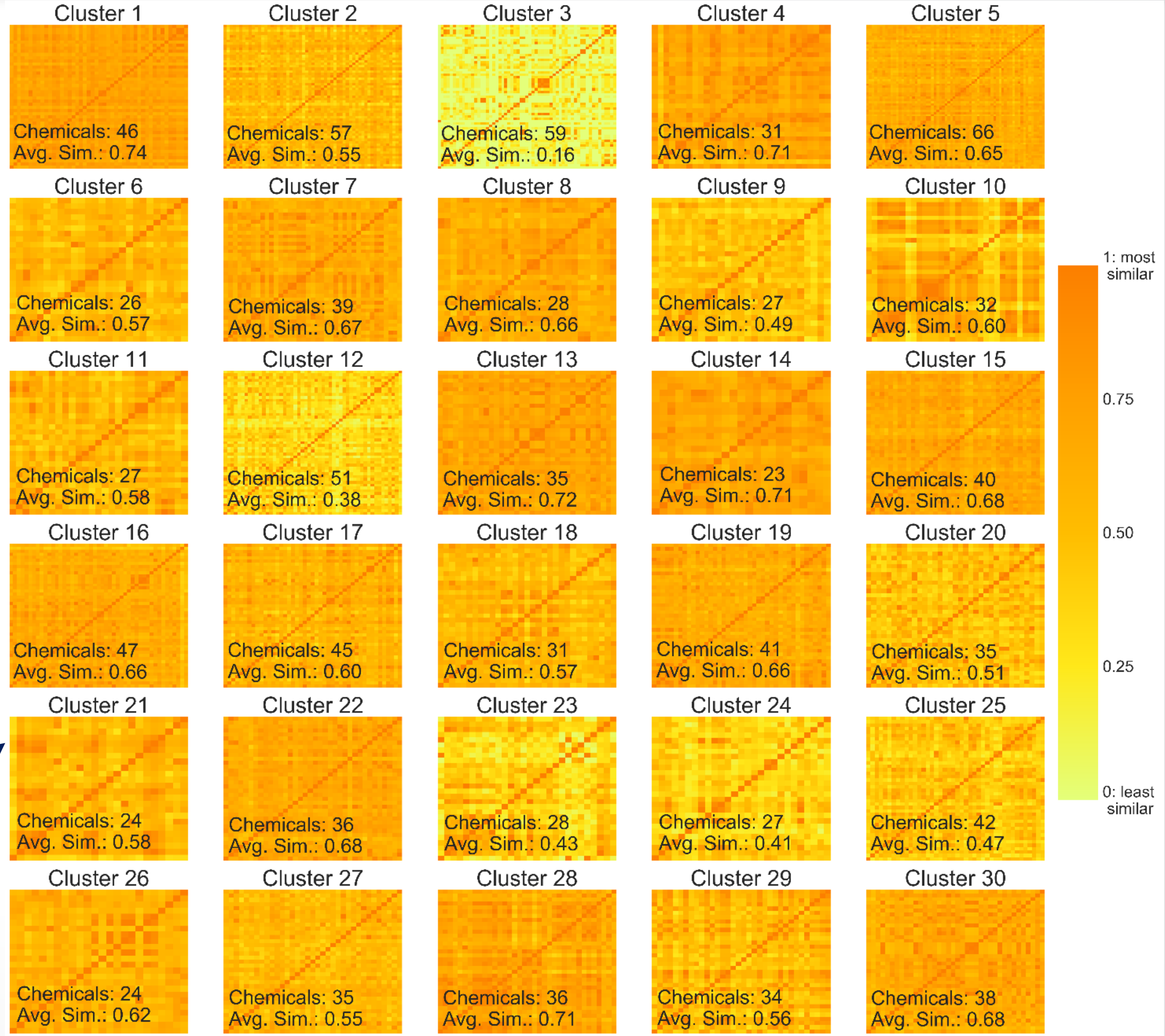
Fraction Unbound in Plasma

Algorithm: Unsupervised *k*-means
Fingerprints: ToxPrints and PubChem
Number of Clusters (Elbow Method): 30
Similarity Metric: Jaccard/Tanimoto Coefficient

Heatmaps of chemical similarity within each cluster measured using Tanimoto similarity.

Each heatmap indicates the number of chemicals and the average similarity within that cluster.

On the color-scale, darker orange means similar (Tanimoto coefficient = 1) whereas yellow means dissimilar (Tanimoto coefficient = 0).



Unsupervised clustering analysis

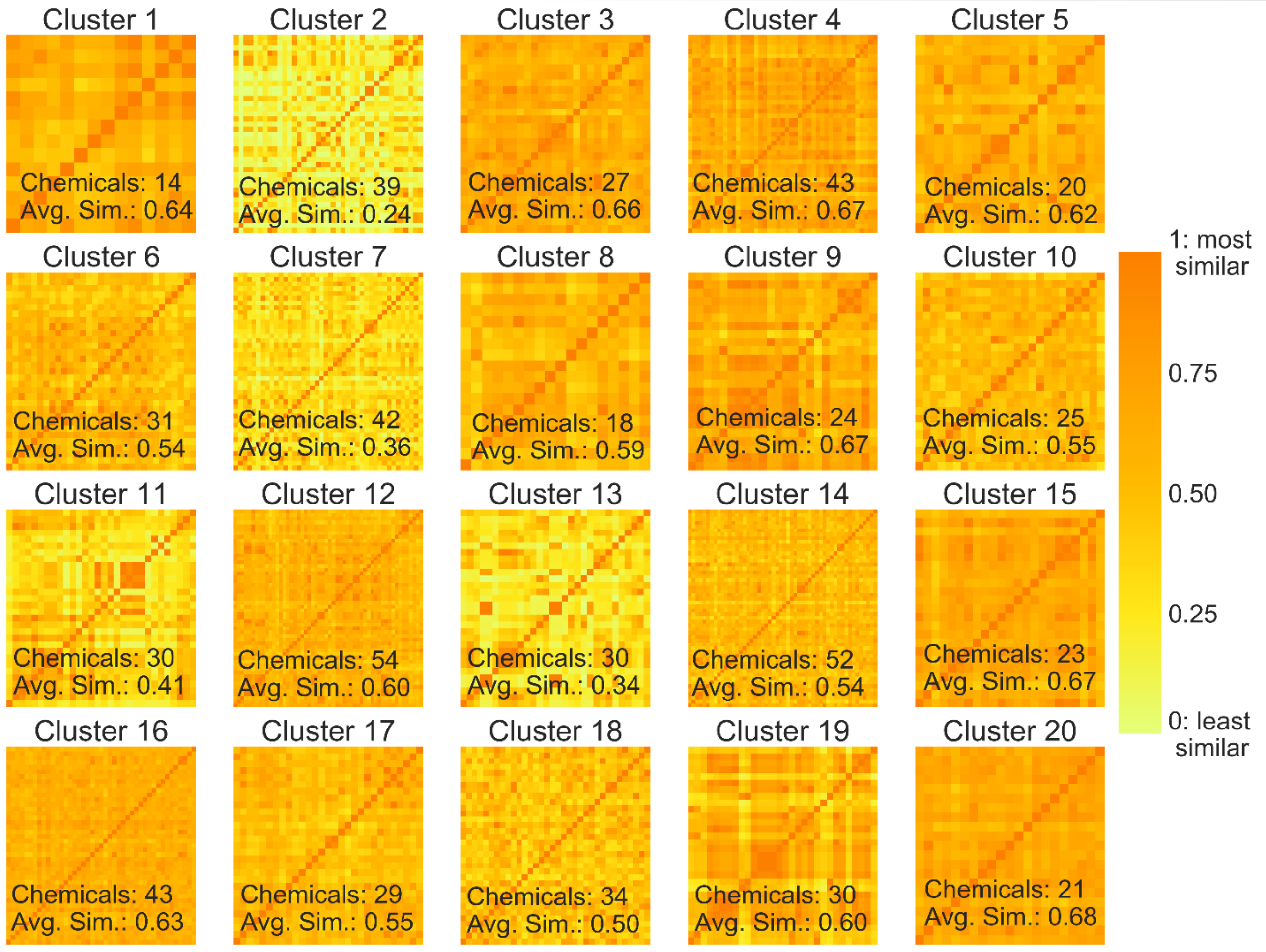
Intrinsic Clearance

Algorithm: Unsupervised *k*-means
Fingerprints: ToxPrints and PubChem
Number of Clusters (Elbow Method): 20
Similarity Metric: Jaccard/Tanimoto Coefficient

Heatmaps of chemical similarity within each cluster measured using Tanimoto similarity.

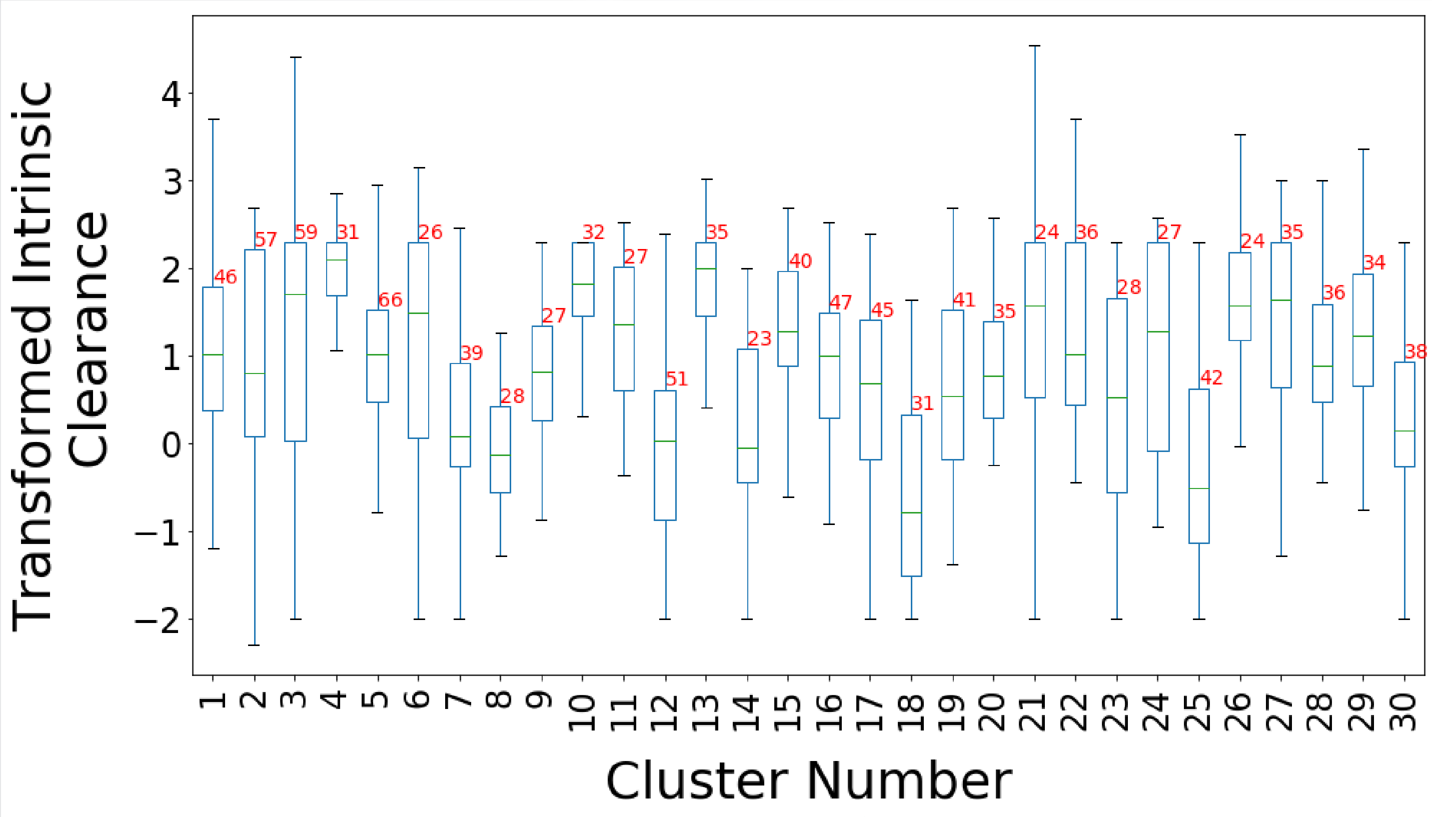
Each heatmap indicates the number of chemicals and the average similarity within that cluster.

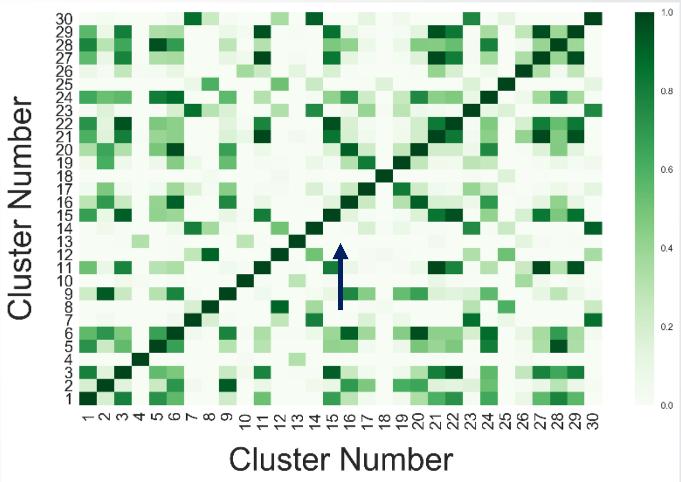
On the color-scale, darker orange means similar (Tanimoto coefficient = 1) whereas yellow means dissimilar (Tanimoto coefficient = 0).



Unsupervised Clustering Analysis

Fraction Unbound in Plasma



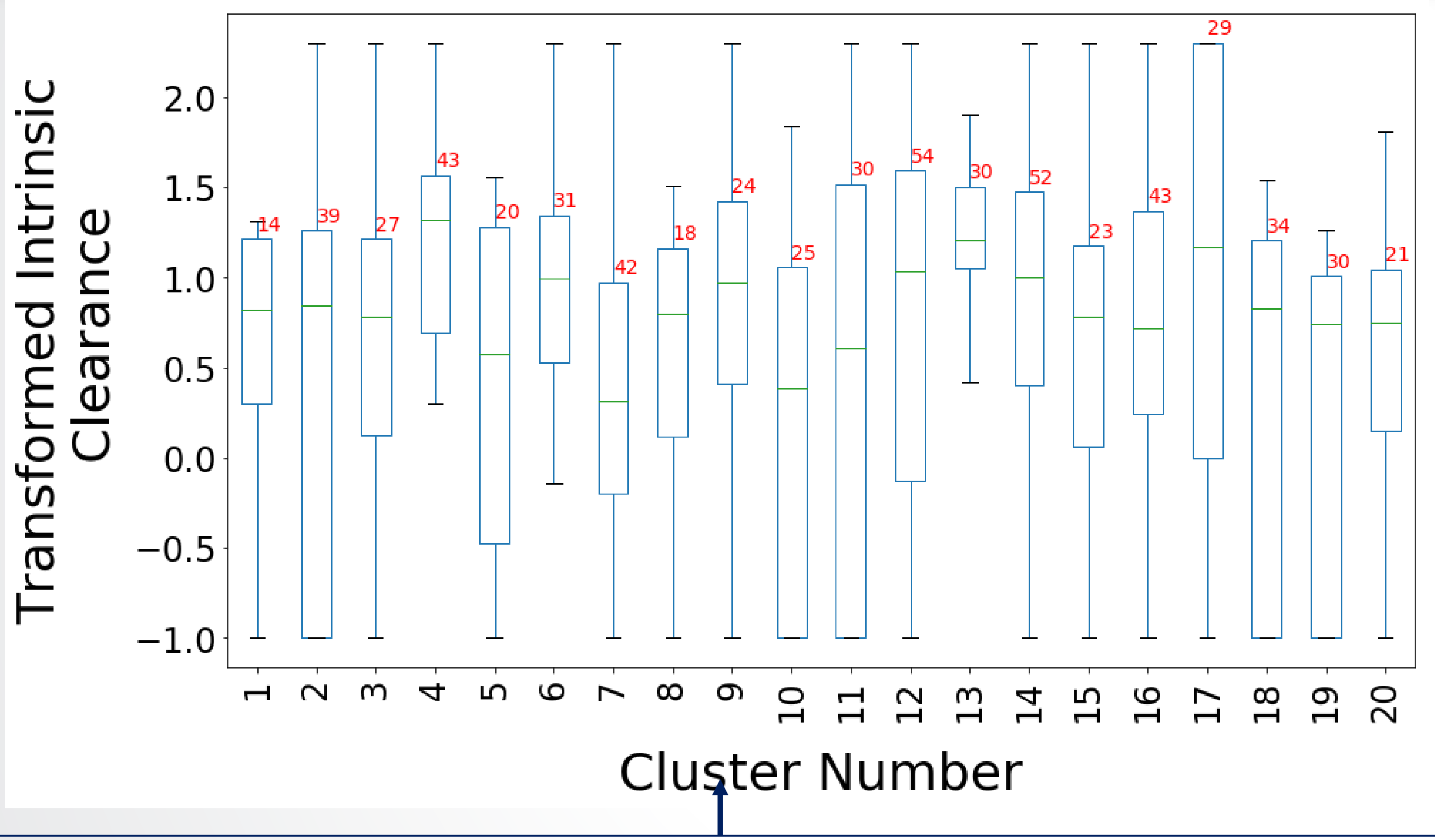


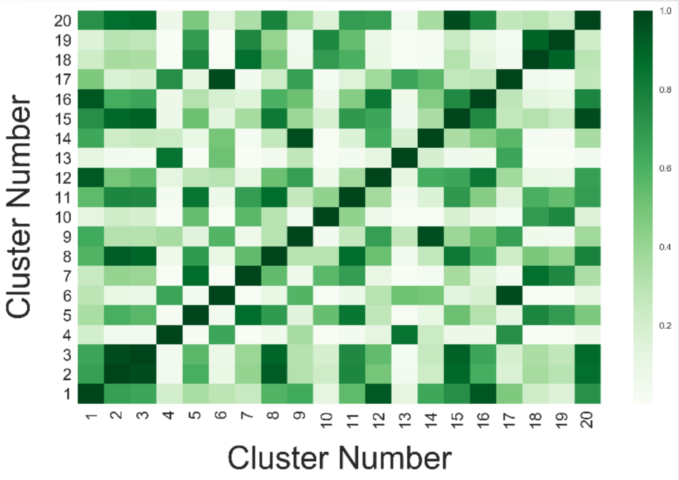
- The less structurally similar cluster have wider ranges as compared to more structurally similar clusters.
- In general, most of the clusters demonstrate a correlation between the average structural similarity in a cluster and the range of values for the chemicals

Heatmap of p-values from T-tests to determine difference between parameter mean value across each cluster.

Unsupervised Clustering Analysis

Intrinsic Clearance





Heatmap of p-values from T-tests to determine difference between parameter mean value across each cluster.

Darker green depicts higher p-value implying lesser dissimilarity between parameter mean values for a pair of cluster.

- The range of clearance values across the most structurally similar cluster (cluster number 20) and the least structurally similar cluster (cluster number 2) seem to be very similar.
- In general, the clusters do not show a strong correlation between average cluster similarity and the range of clearance values.

Read-across Models: Fraction Unbound in Plasma

DESCRIPTORS USED (number)	ANALOG SELECTION METHOD	MODEL PARAMETERS	COVERAGE	PERFORMANCE METRICS		
				MAE	RMSE	RMSE/σ
PubChem + Toxprints (49)	Similarity Threshold	Threshold = 0.7	1110	0.76	1.00	0.79
	Count and Similarity Threshold	Count = 1, Threshold = 0.7	1110	0.83	1.15	0.91
		Count = 2, Threshold = 0.7		0.77	1.04	0.83
		Count = 3, Threshold = 0.7		0.75	1.01	0.80
		Count = 4, Threshold = 0.7		0.75	1.01	0.80
		Count = 5, Threshold = 0.7		0.75	1.01	0.80

Read-across Models: Intrinsic Clearance

DESCRIPTORS USED (number)	ANALOG SELECTION METHOD	MODEL PARAMETERS	COVERAGE	PERFORMANCE METRICS		
				Classification		
				Accuracy	F1 score	
PubChem + Toxprints (49)	Similarity Threshold	Threshold = 0.7	629	64.39	[0.35, 0.76, 0.32]	
	Count and Similarity Threshold	Count = 1, Threshold = 0.7	629	58.90	[0.36, 0.71, 0.30]	
		Count = 2, Threshold = 0.7		52.65	[0.39, 0.63, 0.32]	
		Count = 3, Threshold = 0.7		61.36	[0.38, 0.73, 0.28]	
		Count = 4, Threshold = 0.7		59.09	[0.38, 0.71, 0.24]	
		Count = 5, Threshold = 0.7		64.58	[0.39, 0.76, 0.28]	
				Regression (Medium Clearance)		
				MAE	RMSE	RMSE/σ
PubChem + Toxprints (49)	Similarity Threshold	Threshold = 0.7	418	0.40	0.51	1.13
	Count and Similarity Threshold	Count = 1, Threshold = 0.7	418	0.47	0.60	1.34
		Count = 2, Threshold = 0.7		0.42	0.54	1.20
		Count = 3, Threshold = 0.7		0.41	0.52	1.17
		Count = 4, Threshold = 0.7		0.41	0.51	1.14
		Count = 5, Threshold = 0.7		0.41	0.51	1.13