

INTRODUCTION

Toxicokinetic (TK) data are key for relating exposure and internal dose when building *in vitro* based high-throughput risk assessment models. The two key TK parameters that are critical to relating oral dose to an internal steady state concentration are: (1) Fraction unbound in plasma, and (2) Intrinsic clearance. However, *in vitro* TK studies are not high-throughput, and have time and cost limitations. Consequently, *in vitro* TK data is available only for a limited set of chemicals. In the absence of experimental data, read-across and quantitative structure-activity relationships (QSARs) can be used to predict these two parameters *in silico*.

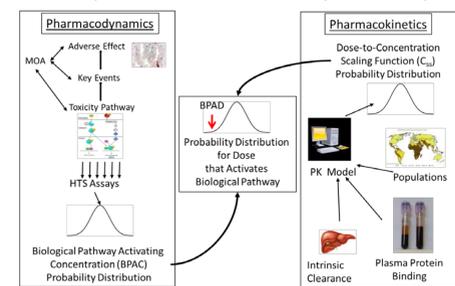


Figure 1: High-throughput risk assessment framework. Judson et. al. 2011.

OBJECTIVES:

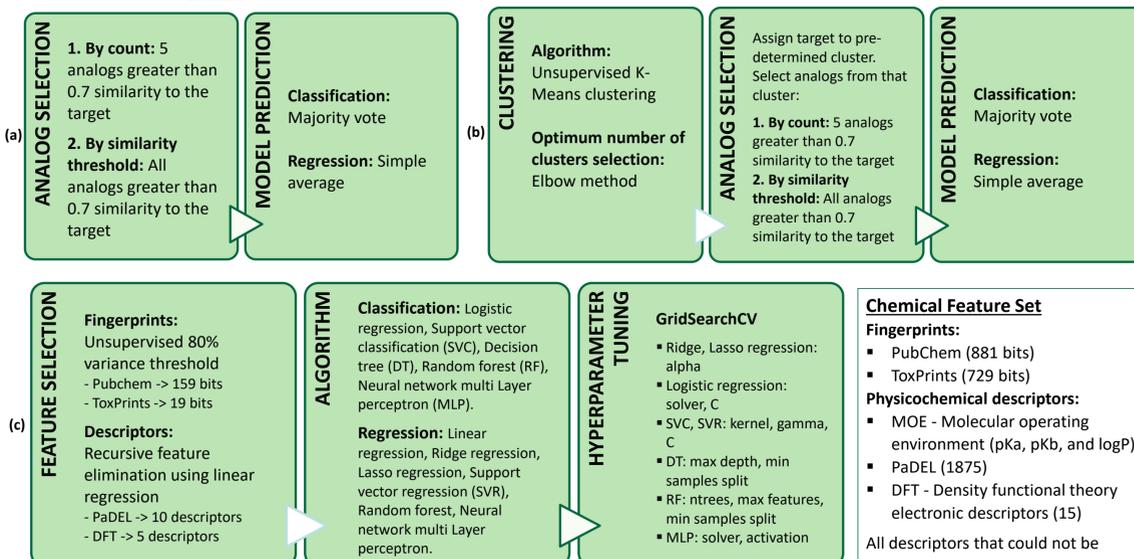
1. Perform unsupervised clustering of chemicals with human *in vitro* TK parameters (fraction unbound in plasma and intrinsic clearance) data to determine whether structurally related chemicals had similar TK parameters.
2. Develop read-across and QSARs models for the two *in vitro* TK parameters.

The analysis relied upon two datasets: (1) 874 chemicals that had human *in vitro* fraction unbound, and (2) 540 chemicals that had human *in vitro* intrinsic clearance data.

METHODS

1. Unsupervised K-Means clustering was performed with PubChem and MOE, and ToxPrints and MOE features. These clusters were further used in the cluster-based read across models.
2. Regression models were developed for fraction unbound in plasma. Two types of models were developed for intrinsic clearance: (1) classification models to predict low, medium and high clearance, and (2) regression models for medium clearance. All models were evaluated using 5-fold cross-validation

Figure 2: Modeling workflow. (a) Simple read-across models, (b) Cluster-based read-across models, and (c) QSAR models.



DATA TRANSFORMATION

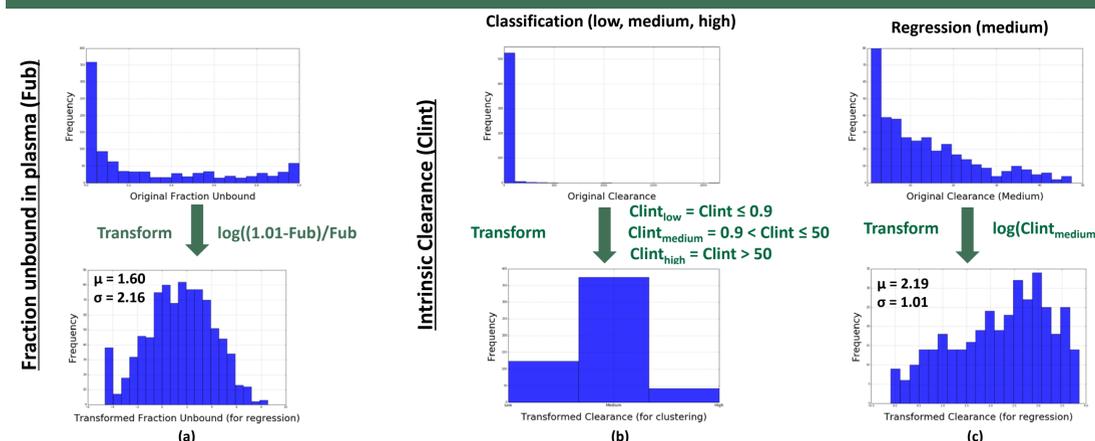


Figure 3: The data was transformed for modeling both the parameters. (a) Regression data for fraction unbound. The mean and standard deviation of the transformed data are: $\mu = 1.60$ and $\sigma = 2.16$. (b) Classification data for clearance. The three bins represent low, medium and high clearance. (c) Regression data for clearance. The mean and standard deviation of the transformed data are: $\mu = 2.19$ and $\sigma = 1.01$.

UNSUPERVISED CLUSTERING RESULTS

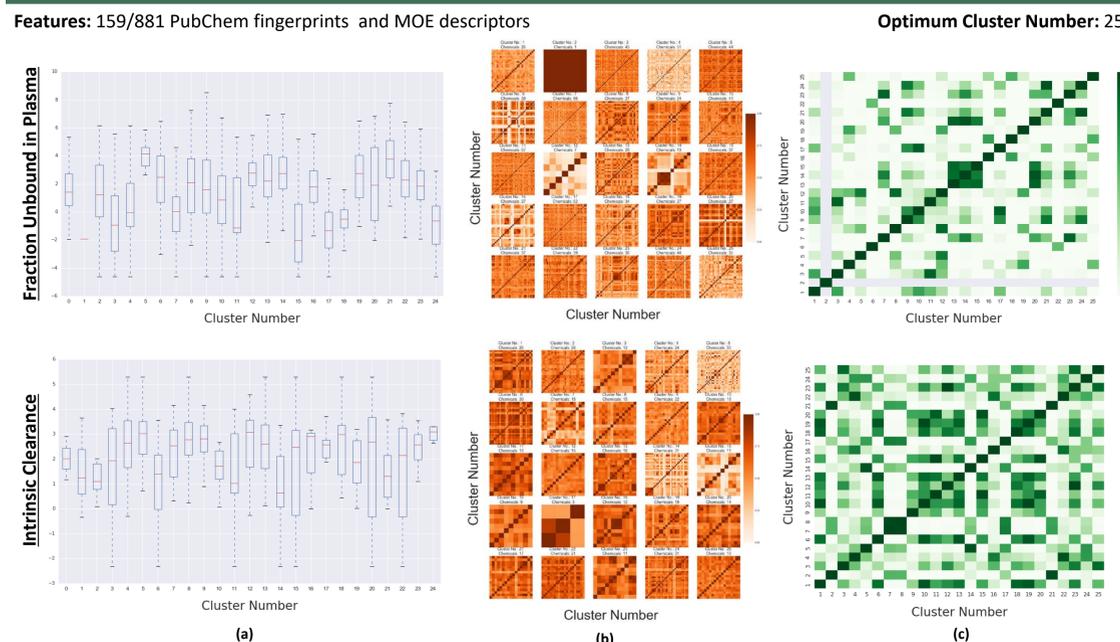


Figure 4: Clustering results using PubChem fingerprints. (a) Boxplots showing distribution of parameter values in different clusters. Fraction unbound has higher inter-cluster variance in parameter value as compared to clearance. (b) Heatmaps of chemical similarity within each cluster measured using Tanimoto similarity. Fraction unbound has more structurally similar chemicals in the clusters as compared to clearance (darker orange shade depicts higher similarity). (c) p-values from T-tests to determine difference between parameter mean value across each cluster. Fraction unbound has more clusters with statistically different parameter means than clearance (darker green depicts higher p-value implying lesser dissimilarity between parameter mean values for a pair of clusters). In terms of fingerprints, PubChem clusters were better than ToxPrints clusters.

READ-ACROSS AND QSAR RESULTS

METHOD	COVERAGE	MODEL	MAE	RMSE	RMSE/ σ	R ²
READ-ACROSS Similarity Threshold: 0.7 Features: PubChem + MOE	712	Simple	1.69	2.28	0.85	0.19
	712	By count	1.67	2.25	0.84	0.20
	573	By similarity threshold	1.68	2.23	0.83	0.17
	524	Cluster-based	1.65	2.21	0.82	0.20
QSAR Features: PubChem + Toxprints + MOE	874	Linear regression	1.63	2.13	0.79	0.37
	874	Ridge regression	1.63	2.13	0.79	0.37
	874	Lasso regression	1.73	2.20	0.82	0.33
	874	SVR	1.35	1.81	0.67	0.55
874	RF	1.37	1.84	0.69	0.53	
QSAR Features: PubChem + Toxprints + MOE + Padel	874	Linear regression	1.68	2.32	0.86	0.23
	874	Ridge regression	1.63	2.17	0.81	0.35
	874	Lasso regression	1.74	2.24	0.83	0.31
	874	SVR	1.35	1.81	0.67	0.55
874	RF	1.39	1.88	0.70	0.51	
QSAR Features: PubChem + Toxprints + MOE + DFT	322	Linear regression	3.47	5.08	1.89	-3.01
	322	Ridge regression	1.74	2.31	0.86	0.17
	322	Lasso regression	1.43	1.94	0.72	0.41
	322	SVR	1.39	1.89	0.74	0.45
322	RF	1.37	1.85	0.69	0.47	

Table 1: Performance metrics for read-across and QSAR models for fraction unbound. The best model is highlighted in red.

METHOD	COVERAGE	CLASSIFICATION MODELS		REGRESSION MODELS FOR $Cl_{int,medium}$		
		MODEL	ACCURACY (%)	COVERAGE	MODEL	RMSE
READ-ACROSS Similarity Threshold: 0.7 Features: PubChem + MOE	432	Simple	66.0	275	Simple	1.16
	432	By count	66.7	275	By similarity threshold	1.16
	373	By similarity threshold	66.8		Cluster-based	1.09
	376	Cluster-based	68.1		By similarity threshold	1.14
QSAR Features: PubChem + Toxprints + MOE	540	Logistic regression	72.4	375	Ridge regression	1.17
	540	SVR	74.1	375	Lasso regression	0.98
	540	DT	62.2	375	SVR	0.96
	540	RF	72.8	375	RF	0.95
540	MLP classifier	72.4	375	MLP regression	0.96	
QSAR Features: PubChem + Toxprints + MOE + Padel	540	Logistic regression	72.6	375	Ridge regression	1.18
	540	SVC	73.7	375	Lasso regression	0.98
	540	DT	65.0	375	SVR	0.96
	540	RF	73.0	375	RF	0.96
540	MLP classifier	70.4	375	MLP regression	0.98	
QSAR Features: PubChem + Toxprints + MOE + DFT	390	Logistic regression	74.6	278	Ridge regression	1.08
	390	SVC	75.6	278	Lasso regression	0.92
	390	DT	66.9	278	SVR	0.91
	390	RF	74.9	278	RF	0.90
390	MLP classifier	74.4	278	MLP regression	0.93	

Table 2: Performance metrics for read-across and QSAR classification and regression models for clearance. The best models are highlighted in red. MAE: Mean-average-error, RMSE: Root-mean-squared-error, Coverage: Number of chemicals predicted by the model

SUMMARY

Unsupervised Clustering

The difference in parameter mean values between clusters, as measured with a T-test, suggest that fraction unbound in plasma is more structurally predictable than intrinsic clearance.

Fraction Unbound in Plasma Models

The best read-across and QSAR regression models for transformed fraction unbound in plasma have RMSE of 2.21 and 1.81, respectively. The prediction errors are on the order of one standard deviation of the modeled data, which is comparable to the errors in the underlying data (~5%. Wambaugh 2016, in prep).

Intrinsic Clearance Models

The best read-across and QSAR classification models for clearance have accuracy of 68.1% and 74.1%, respectively. The best medium clearance regression model has RMSE of 0.90, which is on the order of one standard deviation of the modeled data. This endpoint is the result of a complex set of biological processes and is expected to be difficult to model using just structural features. However, these are the first set of models for clearance based on a large dataset, and will serve as a baseline for improved models.

The model results show that adding more descriptors to the structural fingerprints does not improve the results to a significant extent for either parameter. However, the clustering results demonstrate that some clusters are more predictable than others. Future work will include building cluster-based QSAR models for both the parameters.

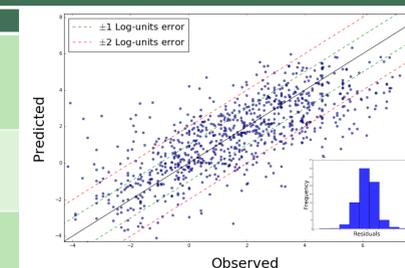


Figure 5: Observed versus predicted fraction unbound in plasma (transformed scale) for the best model: SVR (kernel: 'rbf', gamma: 0.01, C: 10) - highlighted in red in Table 1. The figure in inset shows the distribution of residuals.

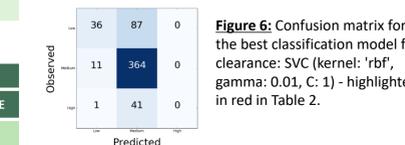


Figure 6: Confusion matrix for the best classification model for clearance: SVC (kernel: 'rbf', gamma: 0.01, C: 1) - highlighted in red in Table 2.

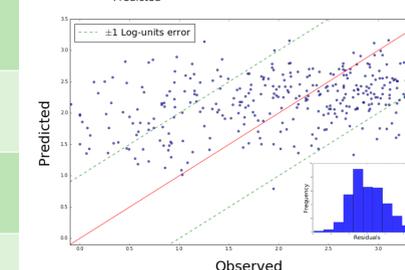


Figure 7: Observed versus predicted medium clearance (transformed scale) for the best model: RF (ntrees: 250, max_features: 'auto', min_samples_split: 5) - highlighted in red in Table 2. The figure in inset shows the distribution of residuals.