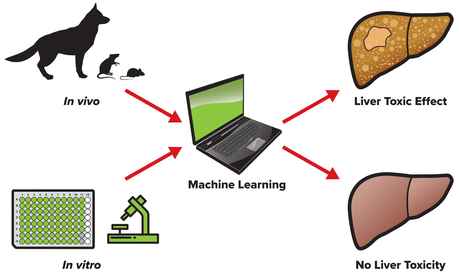
Predicting Subchronic and Chronic Animal Toxicity from *In vitro* High Content Imaging Data Using PBTK and Machine Learning

Todor Antonijevic[1,2] and Imran Shah[3]

Todor Antonijevic | ORCID ID 0000-0002-0248-8412 | tantonijevic@toxstrategies.com | ToxStrategies, Katy, TX
[1]ToxStrategies, Katy, TX. [2]ORISE, RTP, NC. [3]NCCT, US EPA, RTP, NC.

Poster #: P604
Abstract #: 2973



Background

Only a fraction of chemicals in everyday use has been subjected to thorough toxicity testing due to the infeasibility of animal experimentation. A promising alternative to the current toxicity testing paradigm is predictive toxicology based on a combination of *in vitro* high-throughput screening (HTS) assays and computational methods. HTS assays provide a rapid screening approach and access to cellular bioactivities across a large number of chemicals, whereas computational models allow extrapolation of these measurements and comparison with animal apical responses.

Here we propose a methodology to predict subchronic and chronic *in vivo* toxicity from *in vitro* HCI data using physiologically based toxicokinetics (PBTK), High Content Imaging (HCI), and Machine Learning (ML). We investigated the impact of three main factors on predictive performance:

1. choice of ML algorithm,
2. use of balanced vs imbalanced data, defined by toxic and non-toxic labels, and
3. selection of the time point at which the HCI measurements were obtained.

Objective

This study (a) outlines a methodology that combines HCI HepG2 responses, ML, and PBTK to predict mouse, rat and dog subchronic/chronic liver toxicity, and (b) identifies the algorithm, data set preparation approach and time points that provide best predictive accuracy.

Disclosure: This abstract does not reflect EPA policy.

Method

Physiologically based toxicokinetic (PBTK) modeling and machine learning (ML) were utilized to predict mouse, rat, and dog liver toxicity from high content imaging (HCI) data obtained by measuring HepG2 cell responses to 967 chemical treatments across 10 endpoints and 3 time points (1h, 24h, and 72h).

Steps performed:

- The HCI data were normalized to generate z-score data for p53, c-Jun, H2A.X, PH3, α-tubulin, mitochondrial membrane potential, mitochondrial mass, cell cycle arrest, nuclear size, and cell number.
- Lowest-observed adverse effect level (LOAEL) values for chemicals from subchronic/chronic studies in mouse (75/154), rat (161/160), and dog (69/113) were obtained from the ToxRef database.
- LOAEL values were converted to average venous concentrations using PBTK to match the *in vitro* treatment protocol.
- Each *in vitro* treatment was associated with a toxicity class as follows: nontoxic if the venous concentration corresponding to LOAEL was greater than *in vitro* concentration, and toxic otherwise.
- Five ML algorithms (k-nearest neighbors (kNN), Random forest (RF), support vector machine (SVM), decision trees (DT) and naïve Bayes (NB)) were used to evaluate the accuracy for predicting toxicity in each study type and species by each *in vitro* time point. Algorithms were evaluated using ten-fold cross validation.
- A balanced data was created using B-splines to interpolate the HCI concentration-response data at untested concentrations.

Table 1 Workflow that integrates PBTK and Machine Learning

1 ToxRefDB Liver Toxicity LOAEL

Chemical name	Species	Study type: Subcronic	Unit
Dicofol	Rat	0.64	mg/kg/day
	Dog	3.3	
	Mouse	18.2	

3 90 days PBTK (HTTK) simulation to get venous concentration

Dicofol	Rat	2.88	uM
	Dog	18.12	
	Mouse	56.99	

2 HCI Study (Features)

conc [uM]	Chemical name	time [h]	p53	SK	OS	Mt	MM	MMP	MA	CCA	NS	CN
0.39	Dicofol	1	-0.16	-0.14	0.83	-0.02	0.62	0.41	0.35	-0.02	1.00	-0.08
0.78	Dicofol	1	-0.12	-0.09	0.79	0.23	0.61	0.43	0.40	0.15	0.99	-0.29
1.56	Dicofol	1	-0.05	-0.02	0.68	0.38	0.60	0.40	0.29	0.46	0.82	-0.46
3.13	Dicofol	1	-0.02	0.01	0.45	0.24	0.49	0.28	-0.08	0.76	0.46	-0.34
6.25	Dicofol	1	-0.02	0.05	0.08	-0.04	0.23	0.06	-0.46	0.93	0.08	-0.05
12.5	Dicofol	1	0.02	0.21	-0.29	-0.20	-0.18	-0.15	-0.38	1.07	-0.14	0.08
25	Dicofol	1	0.18	0.65	-0.52	-0.28	-0.77	-0.39	0.12	1.33	-0.43	0.03
50	Dicofol	1	0.99	1.74	-0.82	-0.15	-1.35	-0.97	0.70	1.60	-1.01	-0.13
100	Dicofol	1	4.30	3.04	-3.23	2.11	0.32	-1.78	0.03	1.17	0.16	-2.06
200	Dicofol	1	8.20	3.69	-9.76	6.87	4.35	-1.83	-3.05	-0.76	4.07	-8.48

4

5 Labels

Y_{sub} mouse
0
0
0
0
0
0
0
0
1
1

6 Predict/Estimate:
1. Subchronic and chronic adverse effect observed in repeat-dose mouse, rat, and dog studies,
2. Best time point for prediction, and
3. Estimate Accuracy

7 Machine Learning by:
1. Species (rat, dog, nouse),
2. Study type (subchronic, chronic), and
3. HCI time point (1h, 24h, and 72h).

Applied Classification Algorithms:
1. k-nearest neighbors (kNN),
2. Random Forest (RF),
3. Support Vector Classifier (SVC),
4. Decision Trees (DT), and
5. Naïve Bayes (NB).

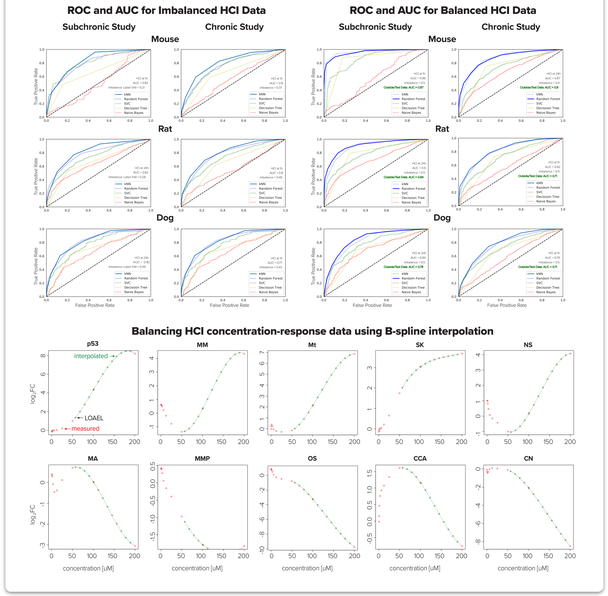
Results

Balancing the HCI data: The mean area under the receiver operating characteristic curves (AUC) for chronic and subchronic imbalanced data were 0.7 and 0.72, respectively. The predictive performance was higher for balanced datasets with a mean AUC of 0.73 and 0.79 for chronic and subchronic toxicity, respectively.

Predictions using balanced training data: RF was the best algorithm to predict subchronic liver toxicity with AUC 0.96 for mice, 0.9 for rat, and 0.84 for dog. For chronic studies, the most accurate classifiers were RF for mice (AUC 0.87), and kNN for rat (AUC 0.82) and dog (AUC 0.79). The best prediction for mouse subchronic/chronic liver toxicity was obtained from 1h/24h HepG2 data. In contrast, for rats and dogs, the highest scores were obtained at 24h/1h.

Prediction using test data: The best performing algorithms showed lower predictions with external validation data. For subchronic toxicity prediction, AUC values for the RF algorithm were 0.87, 0.84, and 0.78 for mouse, rat, and dog, respectively. Prediction of chronic toxicity with external validation data gave AUC of 0.8 for the kNN algorithm in mouse, and AUCs of 0.7 and 0.71 for RF algorithm in rat and dog, respectively. Subchronic studies have better prediction than chronic.

Figure 1 ROC and AUC for Imbalanced and balanced data and balancing process



Discussion

Previously applied ML approaches have mainly compared apical animal observations with in-vivo observations. This study uses PBTK to estimate venous concentrations that correspond to animal subchronic and chronic LOAEL doses, which serve as thresholds that assign HCI treatments into toxic and non-toxic classes.

Table 2. Summary of Results: Mouse

Algorithm	HCI time [h]	Subchronic imbalanced	Subchronic balanced	Chronic imbalanced	Chronic balanced
kNN	1	0.83	0.95	0.8	0.87
	24	0.81	0.94	0.78	0.87
	72	0.79	0.95	0.78	0.85
RF	1	0.8	0.96	0.76	0.84
	24	0.78	0.96	0.79	0.87
	72	0.78	0.94	0.75	0.84
SVM	1	0.8	0.86	0.75	0.76
	24	0.76	0.8	0.72	0.74
	72	0.72	0.8	0.72	0.72
DT	1	0.64	0.85	0.65	0.72
	24	0.67	0.85	0.66	0.74
	72	0.63	0.83	0.67	0.73
NB	1	0.51	0.58	0.56	0.59
	24	0.64	0.63	0.63	0.67
	72	0.56	0.66	0.59	0.62

Table 3. Summary of Results: Rat

Algorithm	HCI time [h]	Subchronic imbalanced	Subchronic balanced	Chronic imbalanced	Chronic balanced
kNN	1	0.81	0.88	0.8	0.82
	24	0.82	0.89	0.78	0.79
	72	0.77	0.88	0.76	0.78
RF	1	0.78	0.9	0.78	0.81
	24	0.79	0.9	0.78	0.79
	72	0.77	0.9	0.77	0.78
SVM	1	0.75	0.78	0.73	0.72
	24	0.74	0.78	0.71	0.71
	72	0.74	0.75	0.71	0.72
DT	1	0.67	0.77	0.65	0.69
	24	0.65	0.77	0.68	0.68
	72	0.66	0.76	0.68	0.66
NB	1	0.57	0.61	0.55	0.59
	24	0.64	0.68	0.62	0.63
	72	0.63	0.67	0.64	0.66

Table 4. Summary of results: Dog

Algorithm	HCI time [h]	Subchronic imbalanced	Subchronic balanced	Chronic imbalanced	Chronic balanced
kNN	1	0.77	0.81	0.77	0.79
	24	0.82	0.83	0.76	0.76
	72	0.81	0.82	0.75	0.74
RF	1	0.75	0.8	0.76	0.77
	24	0.82	0.84	0.75	0.78
	72	0.8	0.83	0.74	0.77
SVM	1	0.72	0.71	0.71	0.72
	24	0.76	0.77	0.72	0.73
	72	0.75	0.74	0.7	0.71
DT	1	0.64	0.71	0.66	0.67
	24	0.68	0.66	0.64	0.68
	72	0.67	0.71	0.65	0.67
NB	1	0.54	0.57	0.6	0.59
	24	0.66	0.68	0.62	0.64
	72	0.62	0.64	0.55	0.58

- We observed that ML gives better prediction when applied to early time point HCI measurements (1h and 24h) than at later (at 72h), which suggests that early cell responses may be more informative for toxicity prediction.
- ML prediction was better in the case of balanced HCI data. This finding highlights the utility of using concentration-response modeling to fill data gaps versus resampling alone.
- Our approach predicted subchronic toxicity better than chronic toxicity. This may be due to the the limited number of endpoints used in this study or the ability of the HepG2 model for capturing key events in adverse outcome pathways to chronic liver injury.

Conclusion

This study shows the utility of a new approach for linking *in vitro* data to *in vivo* outcomes using PBTK and machine learning. Our findings suggest that HCI data measured at early time points (within the first 24h) give better prediction than measurement obtained at later time points. Also, subchronic animal outcomes are better predicted then chronic outcomes. Additional improvement in ML performance is possible through balancing the number of 0 (non-toxic) and 1 (toxic) labels in the *in-vitro* dataset.

References

Shah I, Setzer RW, Jack J, Houck KA, Judson RS, Knudsen TB, Liu J, Martin MT, Reif DM, Richard AM, Thomas RS. Using ToxCast™ data to reconstruct dynamic cell state trajectories and estimate toxicological points of departure. Environmental health perspectives. 2016 Jul;124(7):910-9.

Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in vitro Bioactivity and Chemical Structure. Chem. Res. Toxicol. 2015, 28 (4), 738–751. https://doi.org/10.1021/tx500509h.