

A CLUSTERING-BASED QSAR MODEL FOR ACUTE ORAL SYSTEMIC TOXICITY

Haohua Wan¹, Yiwei Wang¹, Huiqin Xin¹,
Pulan Yu², Yang Li², Sean Gehen², Zhen Zhang^{2*}

University of Illinois Urbana-Champaign¹, Dow AgroSciences LLC²

* Email: zzhang13@dow.com



Dow AgroSciences

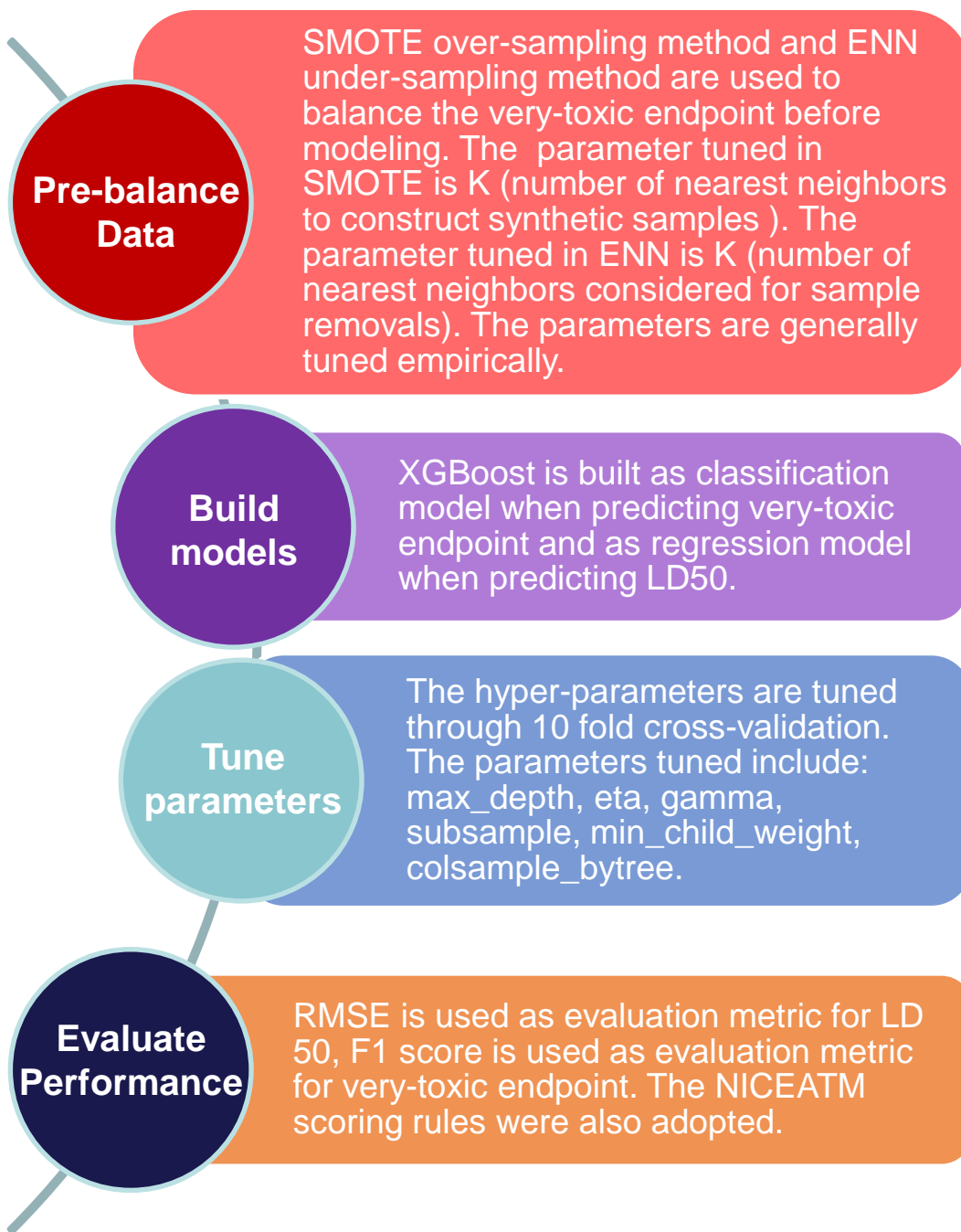
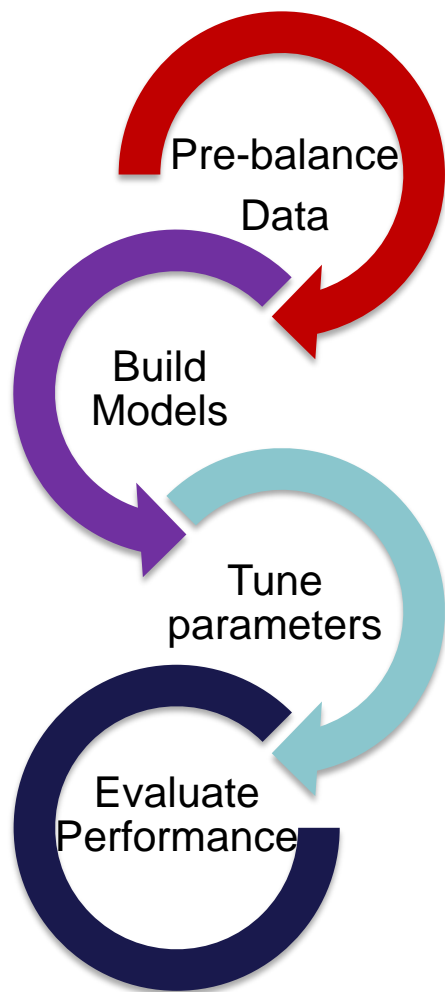
Solutions for the Growing World

Overview of methods

- For feature extraction based on the chemical structure, we used Molecular Operating Environment (MOE) and DataWarrior (DW). Only 2D structures were adopted.
- We implement XGBoost in R to implement the Extreme Gradient Boosting method, which is scalable to big data volume and high-dimensionality, and provides information gains for each variable
- For binary endpoint, the pre-balancing techniques (SMOTE, RU, ENN, etc.) were implemented for the training data in imbalanced classification.
- Tuning parameters for both pre-balancing and classifier were jointly selected by optimizing the prediction performance.

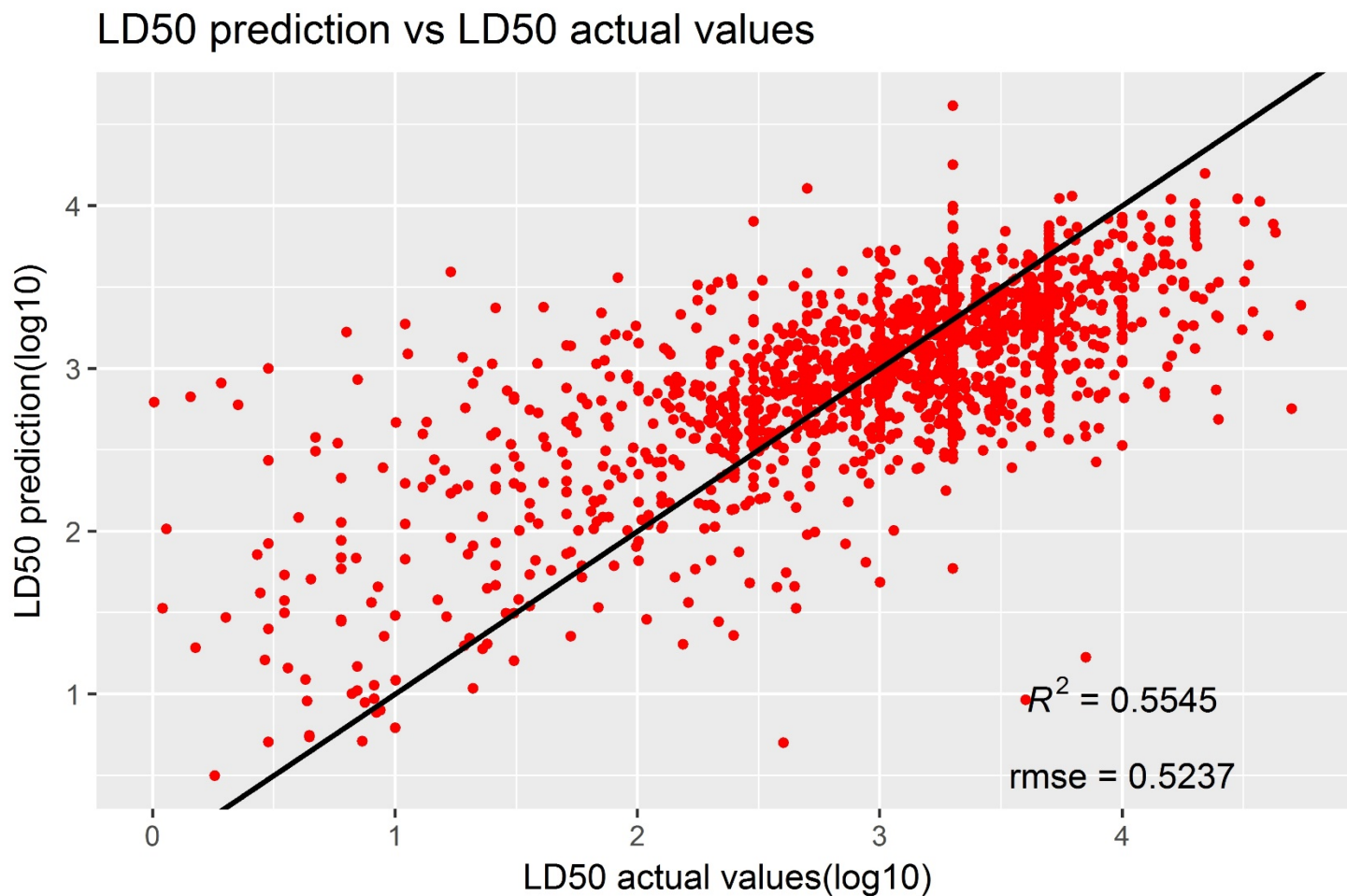
Global method

- The global model.



LD50 Prediction: global method

- The global model for predicting the LD50 endpoint has an RMSE of 0.52, and an R-squared value of 0.55.



Influential descriptors

- *Element*: a_nP (Number of phosphorus atoms), a_ICM (Atom information content), a_nN (Number of nitrogen atoms)
- *Topology*: h_pavgQ (Average total charge sum), chi1v_C (Carbon valence connectivity index) , PEOE_VSA.6.1 (van der Waals surface area)
- *logP*: GCUT_SLOGP_0 (GCUT descriptors using atomic contribution to logP. GCUT_PEOE_0, GCUT_SLOGP_1
- *Toxicity*: Druglikeness

Binary and categorical endpoints

- Train a binary classification model about the response variable `very_toxic`. 10-fold validation was used to choose the best hyper-parameters with F1 scores as follows:

No.	1	2	3	4	5	a	7	8	9	10
F1	0.553	0.576	0.601	0.597	0.575	0.646	0.672	0.623	0.611	0.625

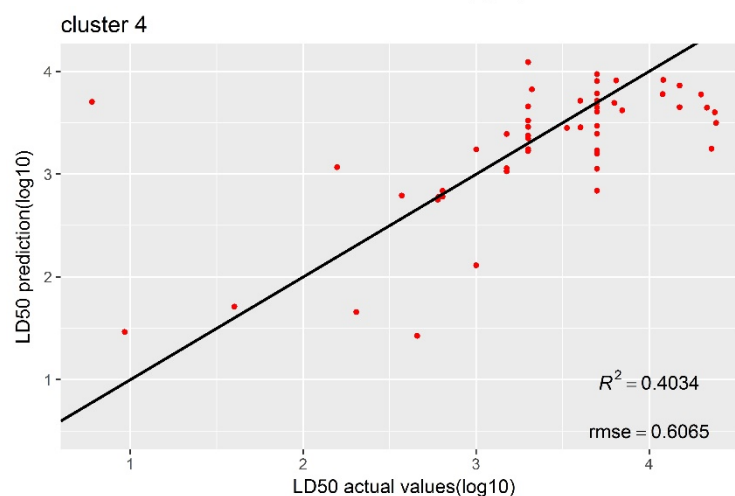
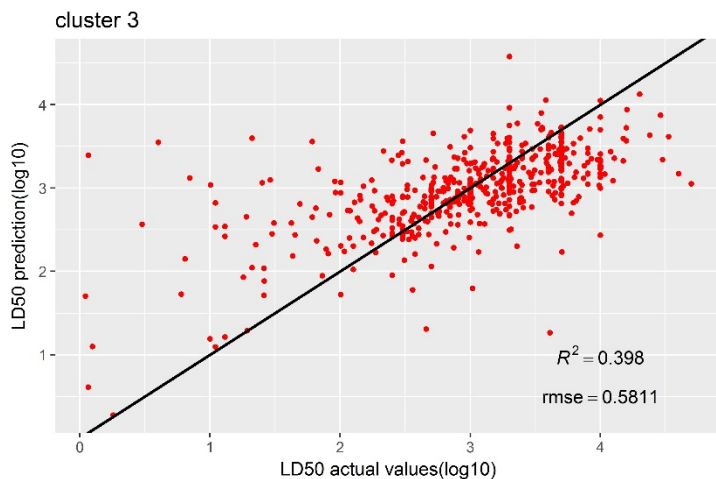
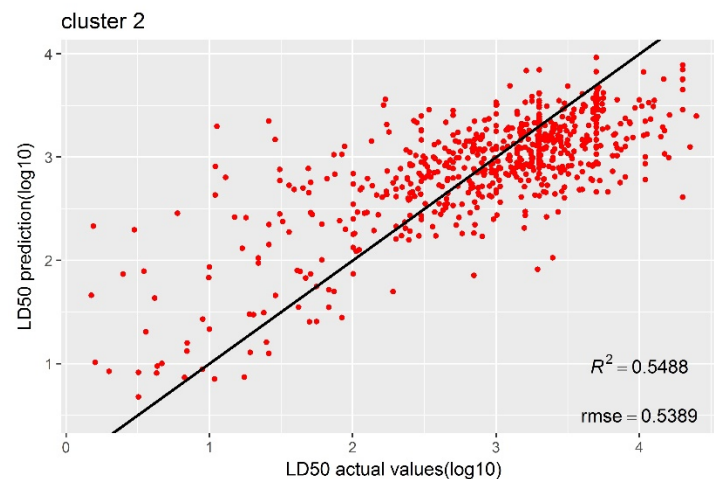
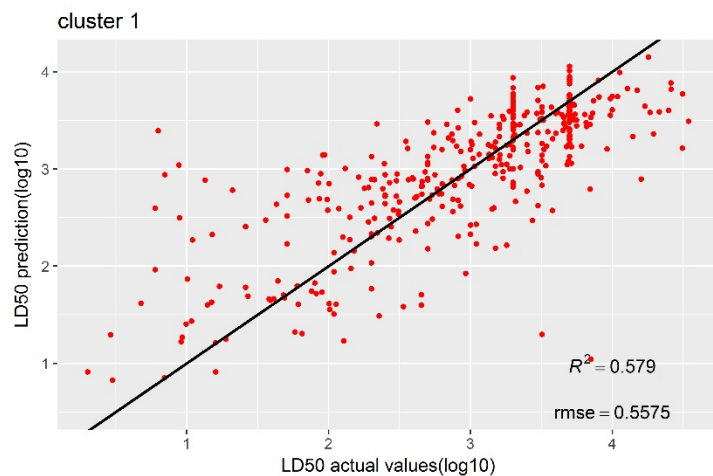
- Non-toxic, EPA, GHS prediction

	Tr_B A	TR_F 1	TR_S P	TR_S N	TR_m ean (SN- SP)	TST_ BA	TST_ F1	TST_S P	TST_S N	TST_ mean (SN- SP)	Score
Non_tox ic	0.917 0	0.926 4	0.9756	0.858 5	0.117 1	0.723 4	0.790 4	0.9267	0.5201	0.4065	0.781 6

	Tr_BA	TR_F1	TR_median(SN-SP)	TST_BA	TST_F1	TST_median(SN-SP)	score
EPA	0.9994	0.9990	0.0013	0.7617	0.5934	0.2306	0.8372
GHS	0.9997	0.9995	0.0002	0.7343	0.4873	0.3576	0.8072

LD50 Prediction: clustering-based method

- first apply *K*-means clustering method to divide the training data into 4 groups. The number of clustered is selected based on gap statistic or Elbow method, then build local XGBoost models.



Very toxicity Prediction

- Clustering based results

Cluster 1	Ref		Accuracy: 0.869
Pred	No	Yes	Kappa: 0.561
No	288	12	F1 Score: 0.638
Yes	38	44	Phi: 0.576

Cluster 3	Ref		Accuracy: 0.956
Pred	No	Yes	Kappa: 0.457
No	505	18	F1 Score: 0.478
Yes	6	11	Phi: 0.475

Cluster 2	Ref		Accuracy: 0.921
Pred	No	Yes	Kappa: 0.498
No	566	31	F1 Score: 0.541
Yes	20	30	Phi: 0.501

Cluster 4	Ref		Accuracy: 0.962
Pred	No	Yes	Kappa: 0.647
No	49	1	F1 Score: 0.667
Yes	1	2	Phi: 0.647



Dow AgroSciences

Thank you