A CLUSTERING-BASED QSAR MODEL FOR ACUTE ORAL SYSTEMIC TOXICITY

Haohua Wan[1], Yiwei Wang[1], Huiqin Xin[1], Pulan Yu[2], Yang Li[2], Sean Gehen[2], Zhen Zhang[2*]

University of Illinois Urbana-Champaign[1], Dow AgroSciences LLC[2]

[*] 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.

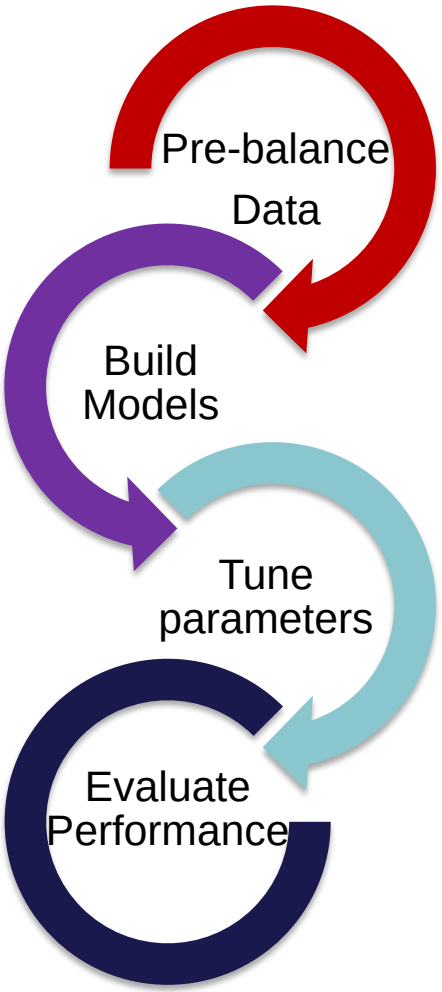
- We implement XGBoost in R to implement the Extreme Gradient Boosting method, which is scalable to big data volume and high-dimensionality, and provides information gains for each variable

- For binary endpint, 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.



Pre-balance Data

SMOTE over-sampling method and ENN under-sampling method are used to balance the very-toxic endpoint before modeling. The parameter tuned in SMOTE is K (number of nearest neighbors to construct synthetic samples). The parameter tuned in ENN is K (number of nearest neighbors considered for sample removals). The parameters are generally tuned empirically.

Build models

XGBoost is built as classification model when predicting very-toxic endpoint and as regression model when predicting LD50.

Tune parameters

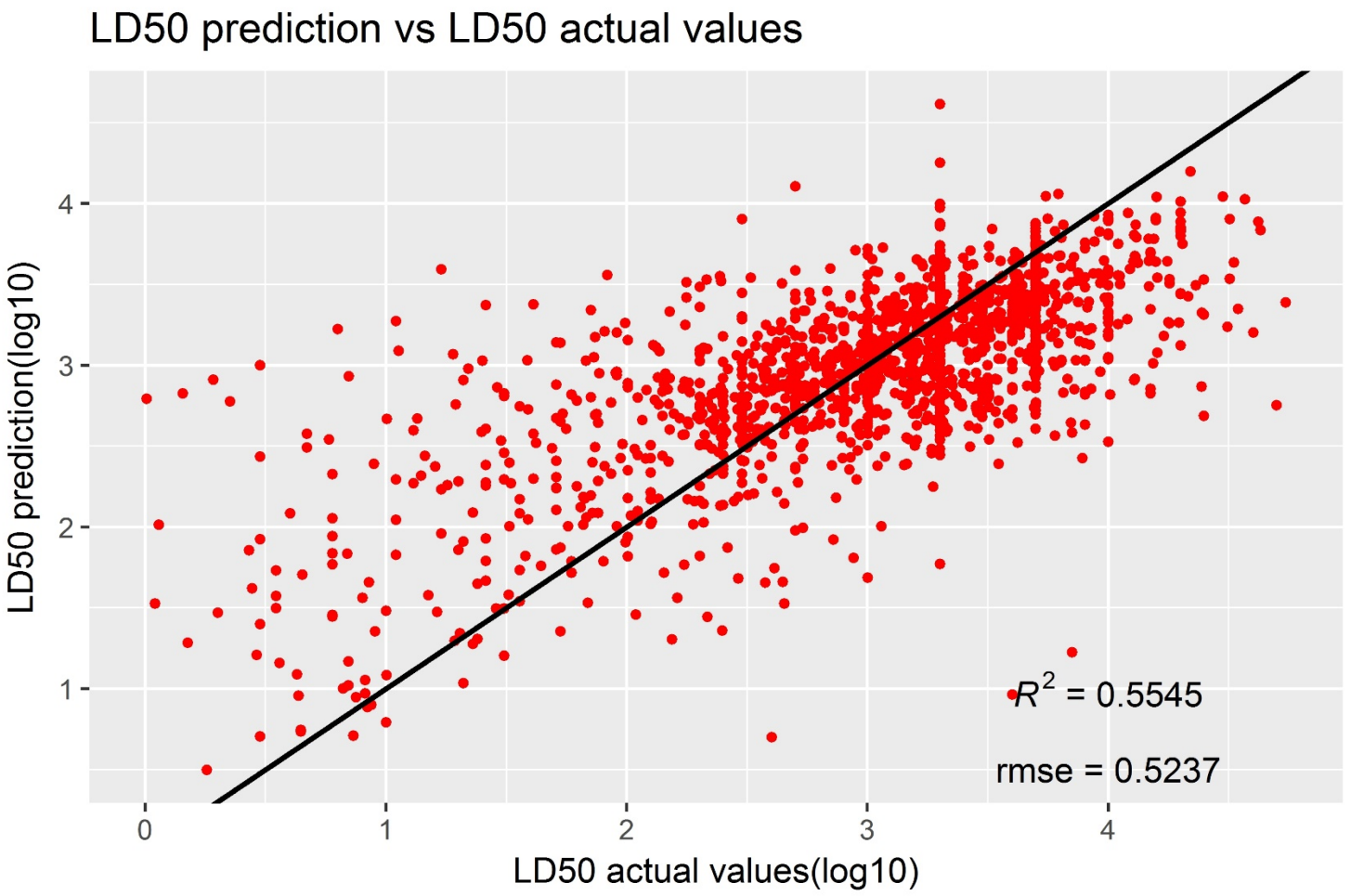
The hyper-parameters are tuned through 10 fold cross-validation. The parameters tuned include: max_depth, eta, gamma, subsample, min_child_weight, colsample_bytree.

Evaluate Performance

RMSE is used as evaluation metric for LD 50, F1 score is used as evaluation metric for very-toxic endpoint. The NICEATM scoring rules were also adopted.

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

- Toxcity: 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

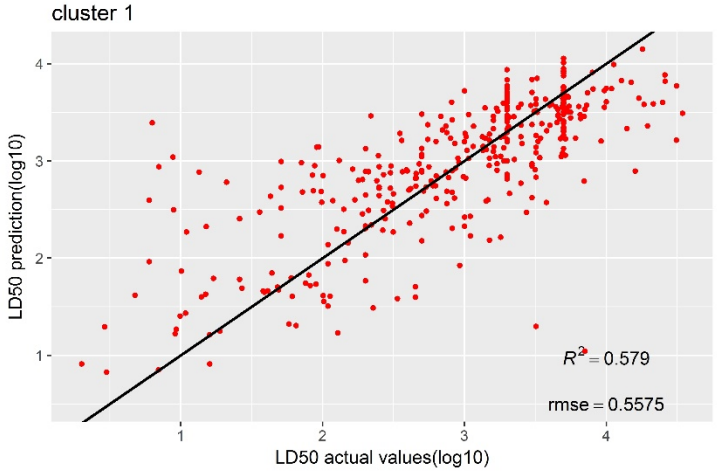
- Non-toxic, EPA, GHS prediction

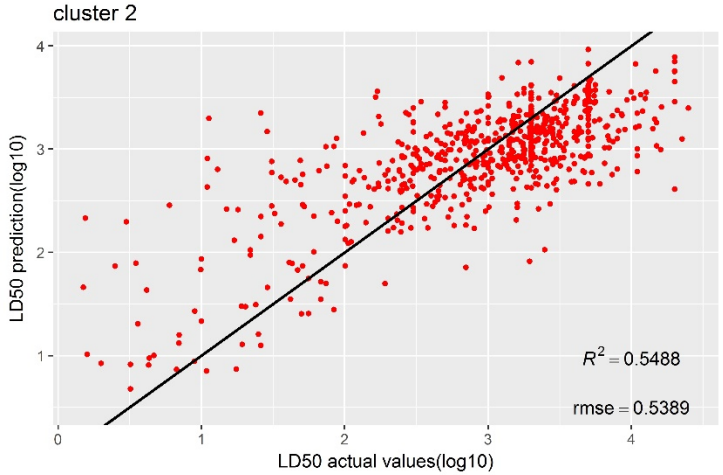
	Tr_BA	TR_F1	TR_SP	TR_SN	TR_mean (SN-SP)	TST_BA	TST_F1	TST_SP	TST_SN	TST_mean (SN-SP)	Score
Non_toxic	0.9170	0.9264	0.9756	0.8585	0.1171	0.7234	0.7904	0.9267	0.5201	0.4065	0.7816

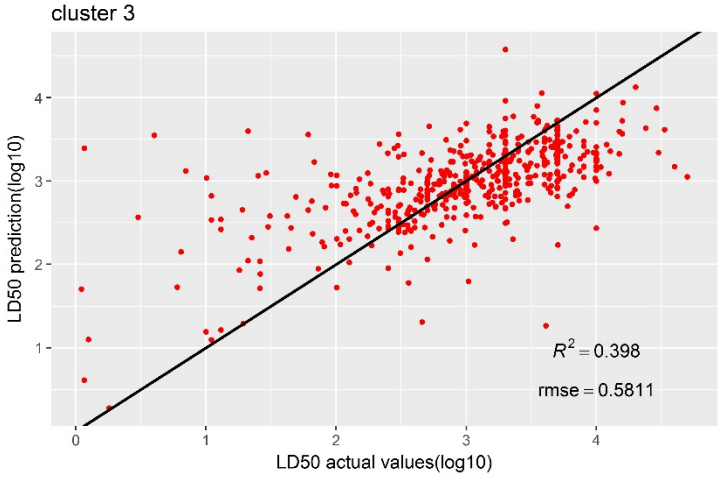
	Tr_BA	TR_F1	TR_median(SN-SP)	TST_BA	TST_F1	TST_median(SN-SP)	score
EPA	0.9994	0.9990	0.0013	0.7617	0.5934	0.2306	0.8372
GHS	0.9997	0.9995	0.0002	0.7343	0.4873	0.3576	0.8072

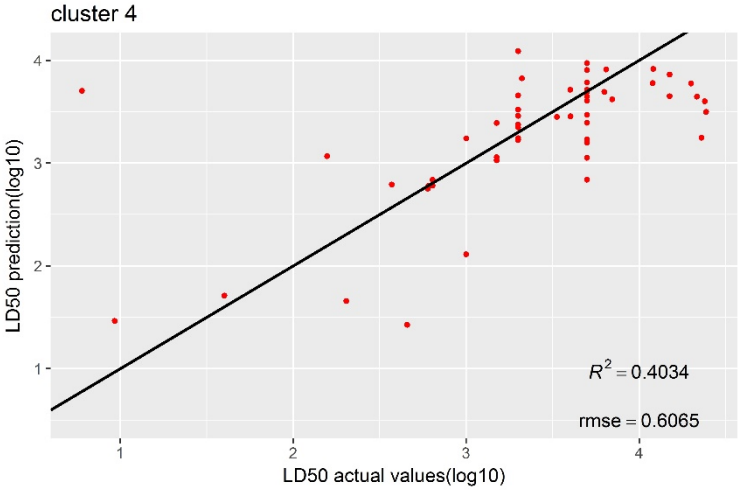
LD50 Prediction: clustering-based method

- 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 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 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 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