

Investigating the mechanisms of emerging and legacy contaminants bioconcentration in fish tissues using machine learning models



THE UNIVERSITY OF TEXAS
AT ARLINGTON

Shashwat Dhayade^{1,2}, Dhruvilkumar Ashokbhai Chodvadiya³, Yihan Fan⁴, Feng Gao⁴, Yike Shen¹

¹Department of Earth and Environmental Sciences, University of Texas at Arlington

²Division of Data Science, University of Texas at Arlington

³Department of Computer Science and Engineering, University of Texas at Arlington

⁴Department of Environmental Health Sciences, University of California Los Angeles

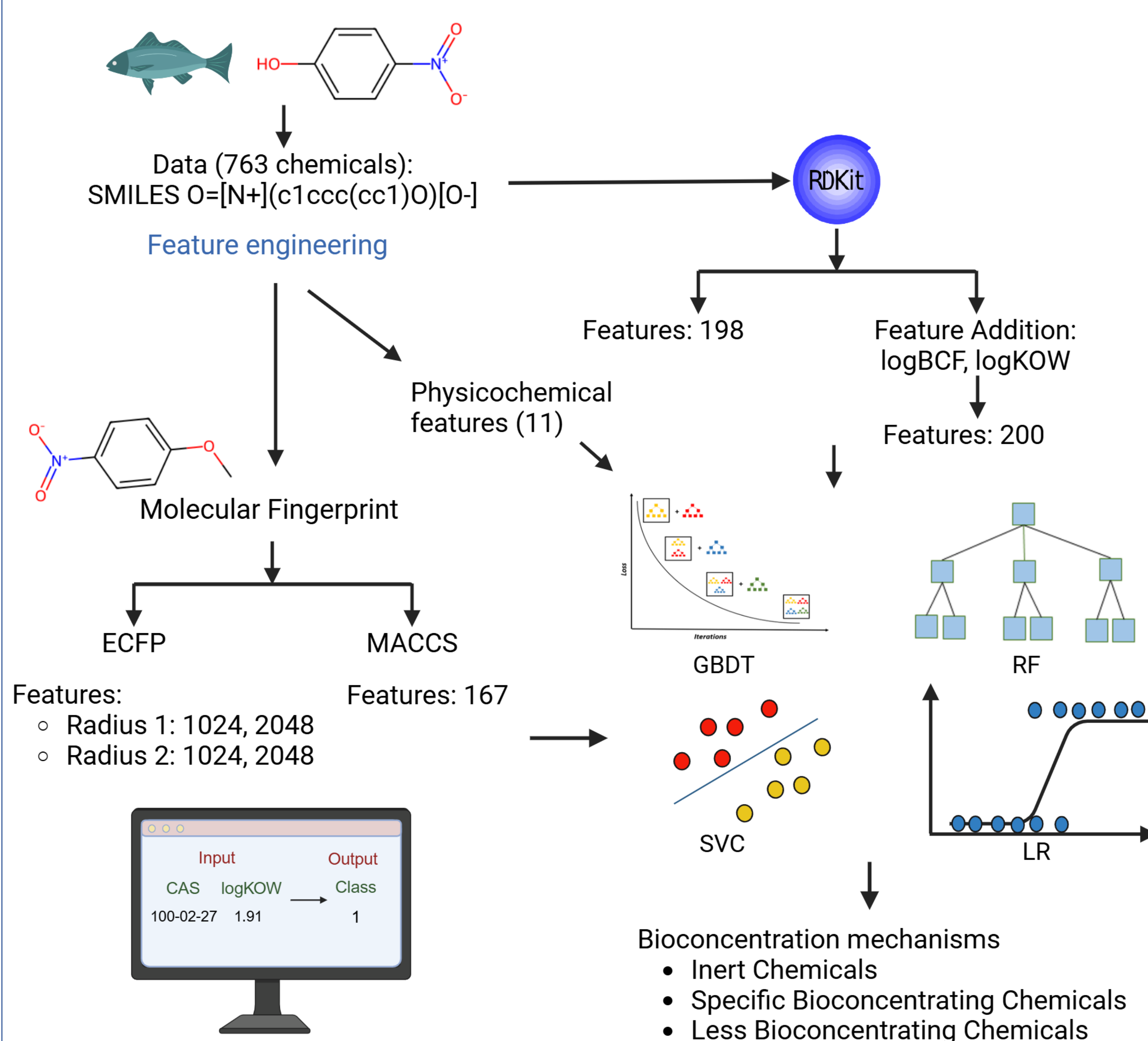
ssd6515@mavs.uta.edu



Background & Objective

- Bioconcentration of contaminants in aquatic organisms poses environmental and public health risks as toxins accumulate in fish and transfer through the food chain.
- In this study, we examined different chemical feature representations (i.e., physicochemical properties, molecular fingerprints including ECFP and MACCS, and RDKit-generated molecular descriptors) to classify fish bioconcentration mechanisms.

Methods



All models are run 25 times using five-fold cross-validation and five repeats. All hyperparameters are fine-tuned using 70% training, 10% validation, and 20% test data.

Results and Discussion

GBDT model achieved a high prediction performance of 0.895 for physicochemical features and 0.863 for RDKit-generated features plus logKow and logBCF, possibly due to its ability of step-by-step learning that strengthens positive class impact on difficult samples.

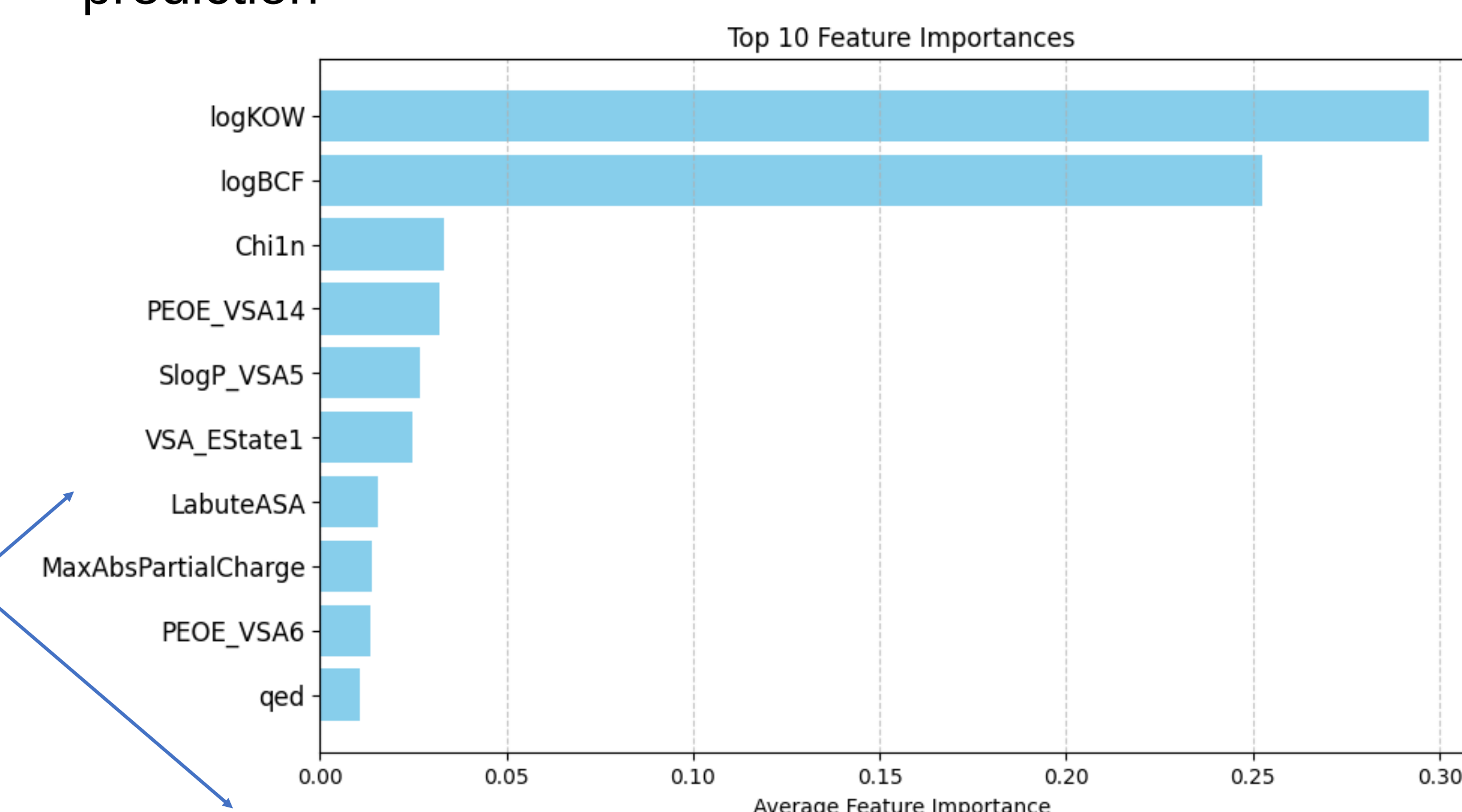
Model	Accuracy	Recall	Precision	F1 (Weighted)
PanelA: Physicochemical properties features				
GBDT	0.895 ± 0.022	0.826 ± 0.054	0.904 ± 0.033	0.893 ± 0.023
SVC	0.900 ± 0.026	0.824 ± 0.049	0.910 ± 0.037	0.897 ± 0.027
LR	0.849 ± 0.025	0.861 ± 0.040	0.764 ± 0.024	0.857 ± 0.025
RF	0.784 ± 0.039	0.631 ± 0.058	0.836 ± 0.059	0.769 ± 0.043
PanelB: RDKit generated features				
GBDT	0.682 ± 0.033	0.554 ± 0.065	0.693 ± 0.087	0.666 ± 0.037
SVC	0.599 ± 0.035	0.388 ± 0.036	0.533 ± 0.156	0.481 ± 0.047
LR	0.279 ± 0.106	0.333 ± 5.551e-17	0.093 ± 0.035	0.134 ± 0.067
RF	0.696 ± 0.039	0.559 ± 0.054	0.721 ± 0.074	0.678 ± 0.042
PanelC: RDKit generated features + logKow + logBCF				
GBDT	0.863 ± 0.024	0.767 ± 0.052	0.869 ± 0.048	0.858 ± 0.026
SVC	0.599 ± 0.036	0.377 ± 0.028	0.579 ± 0.183	0.474 ± 0.046
LR	0.239 ± 0.116	0.333 ± 5.55e-17	0.080 ± 0.039	0.108 ± 0.069
RF	0.700 ± 0.044	0.567 ± 0.047	0.745 ± 0.056	0.686 ± 0.047

Molecular fingerprints are less competitive than physicochemical features in this dataset

Model	Accuracy	Recall	Precision	F1 (Weighted)
PanelD: ECFP features				
GBDT	0.644 ± 0.036	0.472 ± 0.052	0.641 ± 0.105	0.601 ± 0.046
SVC	0.654 ± 0.032	0.470 ± 0.042	0.610 ± 0.138	0.623 ± 0.035
LR	0.602 ± 0.041	0.527 ± 0.046	0.525 ± 0.049	0.606 ± 0.039
RF	0.644 ± 0.039	0.442 ± 0.044	0.681 ± 0.128	0.583 ± 0.053
MLP	0.622 ± 0.036	0.457 ± 0.078	0.516 ± 0.103	0.580 ± 0.057
PanelE: MACCS features				
GBDT	0.632 ± 0.034	0.469 ± 0.039	0.652 ± 0.102	0.603 ± 0.040
SVC	0.630 ± 0.043	0.465 ± 0.049	0.571 ± 0.109	0.606 ± 0.043
LR	0.526 ± 0.042	0.533 ± 0.068	0.470 ± 0.045	0.541 ± 0.039
RF	0.648 ± 0.038	0.463 ± 0.048	0.629 ± 0.135	0.613 ± 0.047

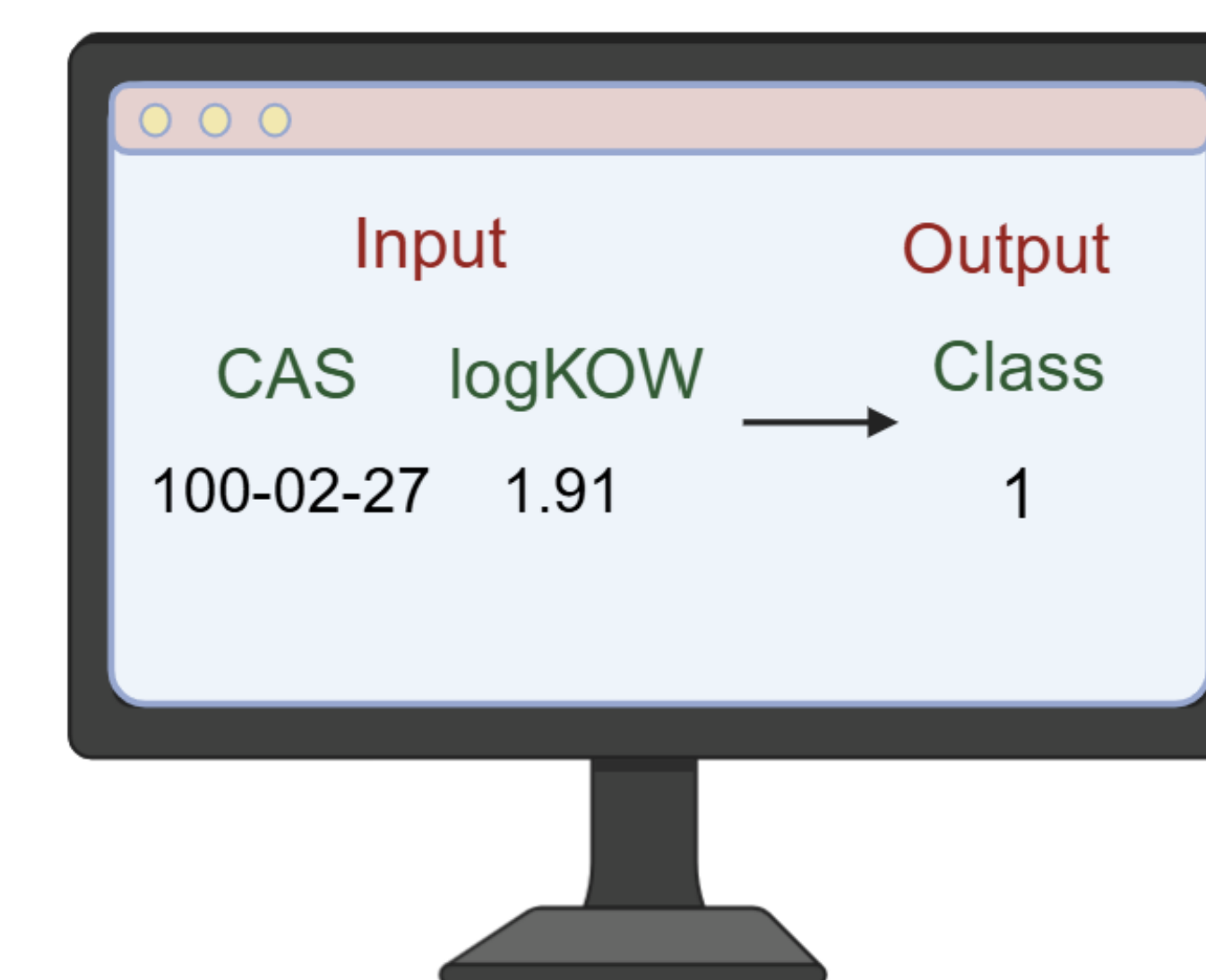
GBDT: gradient boosted decision tree; SVC: support vector classifier; LR: logistic regression; RF: random forest; MLP: Multi-layer perceptron.

RDKit features combined with key physicochemical features can assist users and enable the development of a one-stop solution for bioconcentration mechanism prediction



Webtool development

Coming soon!



Conclusions

- Accurate prediction of bioconcentration mechanisms is important for chemical risk assessment.
- Our GBDT model achieved a high prediction performance of 0.895 ± 0.022 for physicochemical features and 0.863 ± 0.024 for RDKit-generated features plus logKow and logBCF.
- Performance varied with feature representations, highlighting the importance of selecting feature sets for each dataset and model.