Supplementary Information

Interpretable machine learning for investigating complex nanomaterial-plant-soil interactions

Hengjie Yu,[a] Zhilin Zhao,[a] Dan Luo[b] and Fang Cheng*[a]

[a] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, P.R. China

[b] Department of Biological and Environmental Engineering, Cornell University, Ithaca, New York, 14853, USA

*Contact details: E-mail: fcheng@zju.edu.cn

Tel.: +86 571 88982713

Table S1 The version numbers for the main software used in this study.

Software	Version
Python	3.8.10
scikit-learn	0.24.2
lightGBM	3.2.1.99
shap	0.39.0
PDPbox	0.2.0
imodels	1.2.5

Table S2 Final values of grid-search hyperparameters of LightGBM models for different dataset split.

	random state	min_data_in_leaf	min_sum_hessian_in_leaf	max_bin	max_depth	num_leaves	learning_rate
RCF_random split	1	16	1	9	2	4	0.070
	2	16	1	5	4	5	0.096
	3	12	1	11	5	6	0.052
	4	1	16	5	4	5	0.097
	5	1	14	6	3	4	0.088
	6	1	16	8	2	4	0.083
	7	1	17	5	3	4	0.067
	8	1	18	8	3	4	0.095
	9	1	17	15	3	4	0.087
	10	1	18	9	2	3	0.081
log(RCF)_random split	1	1	15	7	4	5	0.094
	2	1	12	13	4	6	0.075
	3	1	12	18	3	6	0.067
	4	1	14	23	3	4	0.083
	5	16	1	14	4	5	0.090
	6	12	1	12	3	6	0.048
	7	8	1	14	7	10	0.043
	8	1	17	6	3	4	0.094
	9	1	17	23	3	4	0.096
	10	11	1	14	2	4	0.044
log(RCF)_stratified shuffle split	1	1	7	22	5	9	0.013
	2	1	11	20	3	7	0.062
	3	16	1	9	4	5	0.095
	4	1	14	23	4	5	0.085
	5	1	13	11	3	6	0.080
	6	1	15	13	3	5	0.068
	7	1	17	9	3	4	0.095
	8	14	1	8	4	5	0.065
	9	12	1	6	5	6	0.090
	10	1	15	7	4	5	0.084

*Note: Hyperparameters tuning is employed in four steps for reducing computation cost, including min_data_in_leaf and min_sum_hessian_in_leaf, max_bin, max_depth

and num_leaves, and learning_rate. For all models, the hyperparameters of n_estimators, n_jobs, and max_cat_to_onehot are set to 1000, -1, and 6 respectively. Default values are used for other hyperparameters that are not listed.

Table S3 All decision rules generated by the RuleFit algorithm where categorical features had been one-hot encoded.

Type	Decision rule	Coefficient	Support	Importance
Linear	Plant subclass_Dilleniidae	0.0759	1.0000	0.0293
	Plant subclass_Rosidae	-0.0513	1.0000	0.0219
Rule	SOM (%) <= 4.025 and SOM (%) > 1.64	0.2561	0.2885	0.1160
	Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 37.5 and SOM (%) <= 3.034 and Clay (%) > 18.5 and MONP composition_CeO2 <= 0.5	-0.2436	0.2692	0.1080
	Surface charge (mV) > -43.2 and Concentration (mg/kg) <= 475.0 and MONP composition_CeO2 <= 0.5	0.2250	0.6538	0.1070
	Surface charge (mV) <= -7.705 and Concentration (mg/kg) <= 62.5	0.2432	0.2596	0.1066
	Concentration (mg/kg) > 37.5 and Exposure time (days) > 25.5 and MONP composition_ZnO <= 0.5	-0.1703	0.3077	0.0786
	Surface charge (mV) > -10.25 and Concentration (mg/kg) > 35.0 and Clay (%) <= 35.45 and MONP composition_ZnO > 0.5	0.2146	0.1442	0.0754
	Plant subclass_Rosidae > 0.5	-0.1441	0.2404	0.0616
	Surface charge (mV) <= -0.015 and Surface charge (mV) > -43.2 and Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 375.0 and Clay (%) > 18.5	-0.1604	0.1250	0.0531
	Concentration (mg/kg) > 37.5 and SOM (%) > 2.245 and Exposure time (days) > 14.0	-0.1067	0.2692	0.0473
	Size (nm) <= 64.0 and Surface charge (mV) > -43.2 and Concentration (mg/kg) <= 475.0 and Concentration (mg/kg) > 37.5 and Clay (%) <= 35.45	-0.0967	0.3173	0.0450
	Plant subclass_Rosidae > 0.5 and Concentration (mg/kg) > 35.0 and Exposure time (days) > 32.5	-0.1033	0.2019	0.0415
	Concentration (mg/kg) > 37.5 and Exposure time (days) > 82.0 and MONP composition_ZnO <= 0.5	-0.0971	0.1346	0.0332
	Surface charge (mV) > -43.2 and Concentration (mg/kg) > 37.5 and Exposure time (days) > 25.5	-0.0626	0.7019	0.0286
	Concentration (mg/kg) > 337.5 and Exposure time (days) <= 43.5	-0.0603	0.2885	0.0273
	Exposure time (days) <= 19.5 and MONP composition_ZnO <= 0.5	0.1073	0.0673	0.0269
	Concentration (mg/kg) > 37.5 and SOM (%) > 2.245 and Exposure time (days) <= 79.0 and MONP composition_CeO2 <= 0.5	0.0649	0.2019	0.0260
	Plant subclass_Dilleniidae <= 0.5 and Plant subclass_Rosidae <= 0.5 and Concentration (mg/kg) <= 475.0 and Concentration (mg/kg) > 35.0	-0.0545	0.3269	0.0256
	Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 35.0 and SOM (%) <= 1.64 and Clay (%) > 18.0	-0.0608	0.2212	0.0252
	Concentration (mg/kg) <= 337.5 and Exposure time (days) <= 43.5	0.0517	0.3654	0.0249
	Surface charge (mV) > -43.2 and Concentration (mg/kg) <= 325.0 and Concentration (mg/kg) > 35.0 and MONP composition_CeO2 > 0.5	-0.1079	0.0481	0.0231
	Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 35.0 and Clay (%) <= 28.5 and MONP composition_ZnO > 0.5	0.0719	0.1154	0.0230

Plant subclass_Rosidae <= 0.5 and Concentration (mg/kg) <= 375.0 and Clay (%) > 18.5 and MONP composition_CeO2 <= 0.5	0.0491	0.2404	0.0210
Plant subclass_Dilleniidae <= 0.5 and Concentration (mg/kg) <= 425.0 and Concentration (mg/kg) > 37.5 and MONP composition_CeO2 <= 0.5	-0.0438	0.3558	0.0210
Size (nm) <= 63.5 and Concentration (mg/kg) > 62.5 and SOM (%) > 1.64 and Exposure time (days) > 25.5	-0.0551	0.1731	0.0208
Plant subclass_Rosidae <= 0.5 and Concentration (mg/kg) <= 450.0 and Concentration (mg/kg) > 35.0 and SOM (%) <= 8.425 and Exposure time (days) <= 77.0 and Exposure time (days) > 25.5 and MONP composition_ZnO <= 0.5	0.0825	0.0673	0.0207
Concentration (mg/kg) > 35.0 and SOM (%) <= 1.41 and Exposure time (days) <= 43.5	-0.0476	0.2308	0.0201
Surface charge (mV) <= -10.415 and Concentration (mg/kg) > 200.0 and SOM (%) > 2.245	-0.0661	0.0962	0.0195
Plant subclass_Rosidae <= 0.5 and Concentration (mg/kg) <= 237.5	0.0365	0.4038	0.0179
Exposure time (days) > 19.5 and MONP composition_ZnO <= 0.5	-0.0369	0.3365	0.0175
Concentration (mg/kg) <= 475.0 and MONP composition_CeO2 <= 0.5	0.0370	0.7115	0.0168
Surface charge (mV) > -43.2 and Plant subclass_Dilleniidae > 0.5 and Concentration (mg/kg) > 35.0 and SOM (%) <= 2.245 and Clay (%) <= 35.45 and MONP composition_CeO2 <= 0.5	0.0769	0.0481	0.0164
Concentration (mg/kg) > 56.25 and Clay (%) <= 35.45 and Exposure time (days) > 32.5 and MONP composition_ZnO > 0.5	0.0379	0.1923	0.0149
Size (nm) <= 67.5 and Concentration (mg/kg) > 37.5 and Clay (%) <= 30.95 and MONP composition_ZnO > 0.5	0.0463	0.1154	0.0148
Size (nm) <= 63.5 and Concentration (mg/kg) > 56.25 and SOM (%) > 1.64	-0.0329	0.2404	0.0141
Concentration (mg/kg) > 200.0 and SOM (%) > 2.245 and Clay (%) <= 24.5	-0.0453	0.0865	0.0127
Concentration (mg/kg) > 37.5 and Exposure time (days) > 82.0	-0.0313	0.1731	0.0118
Surface charge (mV) > -7.705 and Concentration (mg/kg) <= 237.5 and Concentration (mg/kg) > 37.5	-0.0464	0.0481	0.0099
Plant subclass_Rosidae > 0.5 and Clay (%) > 34.5 and MONP composition_ZnO > 0.5	-0.0425	0.0577	0.0099
Concentration (mg/kg) > 35.0 and Clay (%) <= 35.45 and Exposure time (days) > 25.5 and MONP composition_ZnO <= 0.5	-0.0209	0.3077	0.0097
Surface charge (mV) > -43.2 and Concentration (mg/kg) <= 312.5 and Concentration (mg/kg) > 37.5 and Clay (%) <= 35.45 and MONP composition_CeO2 > 0.5	-0.0411	0.0481	0.0088
Concentration (mg/kg) <= 5.5	0.0604	0.0192	0.0083
Size (nm) > 30.0 and Plant subclass_Rosidae > 0.5 and Concentration (mg/kg) > 37.5 and SOM (%) <= 1.64 and MONP composition_CeO2 <= 0.5	-0.0304	0.0769	0.0081
Surface charge (mV) > -10.25 and Concentration (mg/kg) > 37.5 and Clay (%) <= 35.45 and MONP composition_ZnO > 0.5	0.0156	0.1442	0.0055
Concentration (mg/kg) > 37.5 and SOM (%) > 1.64 and MONP composition_ZnO <= 0.5	-0.0127	0.2500	0.0055
Surface charge (mV) <= -10.415 and Concentration (mg/kg) <= 200.0 and	0.0150	0.1250	0.0050

Concentration (mg/kg) > 37.5 and SOM (%) > 2.245			
Size (nm) <= 30.0 and Plant subclass_Rosidae > 0.5 and Concentration (mg/kg) > 37.5 and SOM (%) <= 1.64 and MONP composition_CeO2 <= 0.5	0.0213	0.0385	0.0041
Plant subclass_Rosidae > 0.5 and Concentration (mg/kg) > 56.25 and MONP composition_ZnO <= 0.5	-0.0138	0.0769	0.0037
Size (nm) <= 9.0 and Surface charge (mV) > -43.2 and Concentration (mg/kg) <= 475.0 and Concentration (mg/kg) > 35.0	-0.0128	0.0481	0.0027
Concentration (mg/kg) > 56.25 and SOM (%) <= 4.175 and Exposure time (days) > 43.5	0.0062	0.2500	0.0027
Surface charge (mV) > -43.2 and Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 37.5 and Clay (%) > 35.45	-0.0060	0.1635	0.0022
Clay (%) <= 18.5 and MONP composition_CeO2 <= 0.5 and Plant subclass_Asteridae <= 0.5	0.0043	0.3462	0.0020
Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 37.5 and Clay (%) > 35.45 and MONP composition_ZnO > 0.5	-0.0055	0.1635	0.0020
Concentration (mg/kg) > 47.5 and SOM (%) <= 1.41 and Exposure time (days) <= 44.0 and MONP composition_ZnO > 0.5	-0.0044	0.2212	0.0018
Concentration (mg/kg) > 56.25 and SOM (%) <= 1.41 and Exposure time (days) <= 43.5	-0.0021	0.2308	0.0009
Plant subclass_Commelinidae <= 0.5 and Concentration (mg/kg) > 37.5 and SOM (%) > 3.034 and Clay (%) > 18.5 and MONP composition_CeO2 <= 0.5	0.0037	0.0192	0.0005
Concentration (mg/kg) <= 325.0 and Concentration (mg/kg) > 37.5 and MONP composition_CeO2 > 0.5	-0.0022	0.0481	0.0005
Plant subclass_Commelinidae > 0.5 and Plant subclass_Rosidae <= 0.5 and Concentration (mg/kg) > 37.5 and SOM (%) <= 8.425	-0.0003	0.2596	0.0001
Concentration (mg/kg) > 37.5 and SOM (%) > 1.64 and Clay (%) <= 7.97	-0.0006	0.0288	0.0001

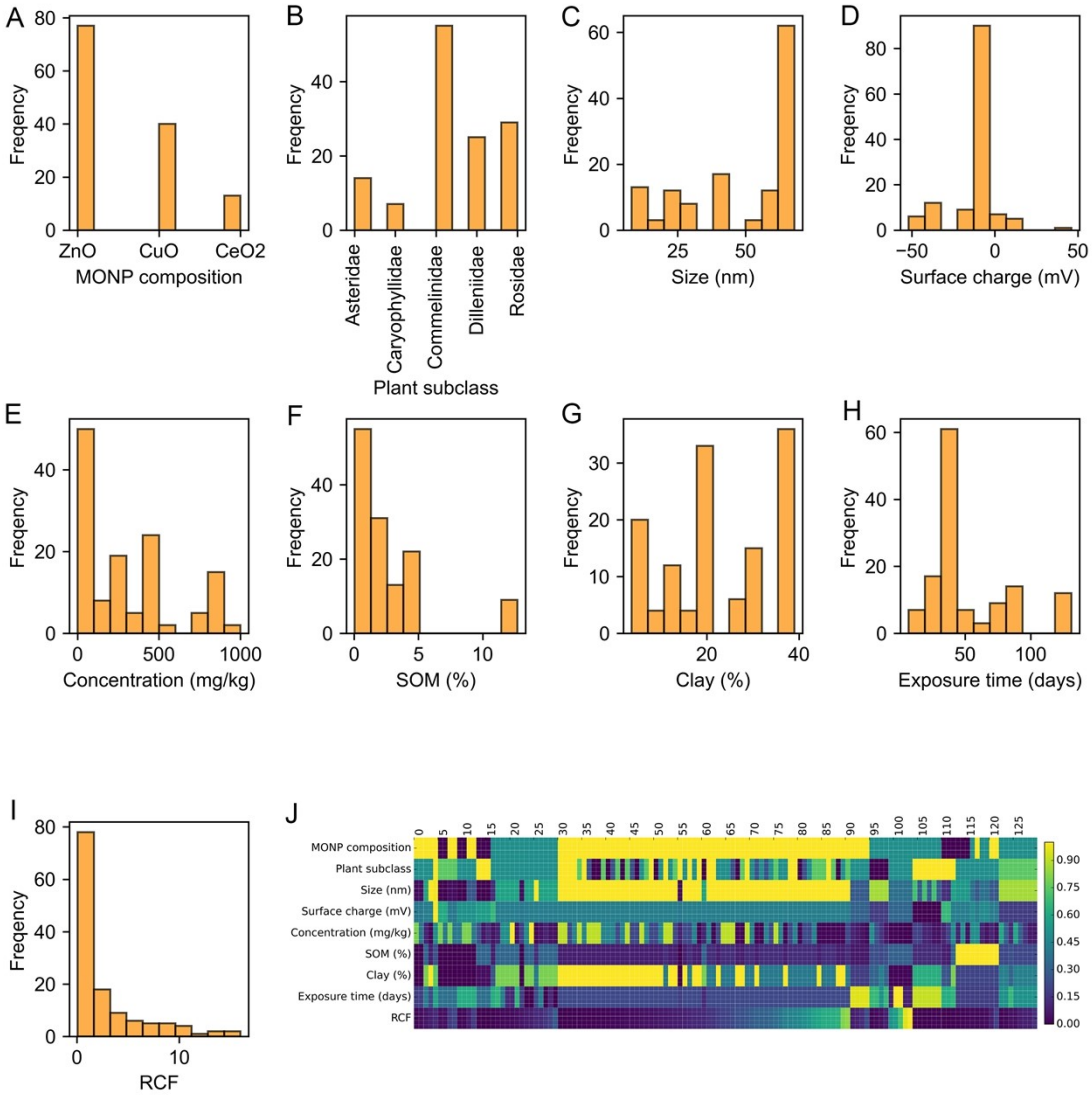

Fig. S1 Data distribution of features (A-H) and raw predicted target (I). (J) Data visualization of this dataset where the features and RCF are normalized to the 0-1 range.

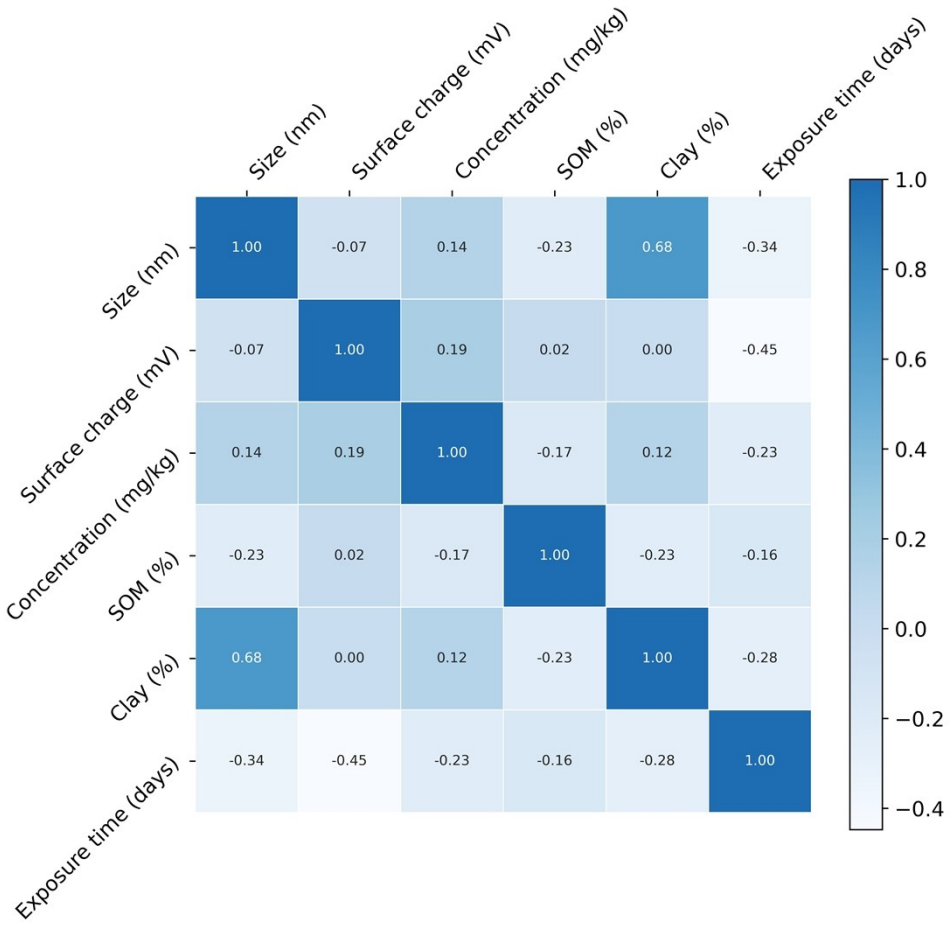

Fig. S2 Pearson correlation coefficient among numerical features.

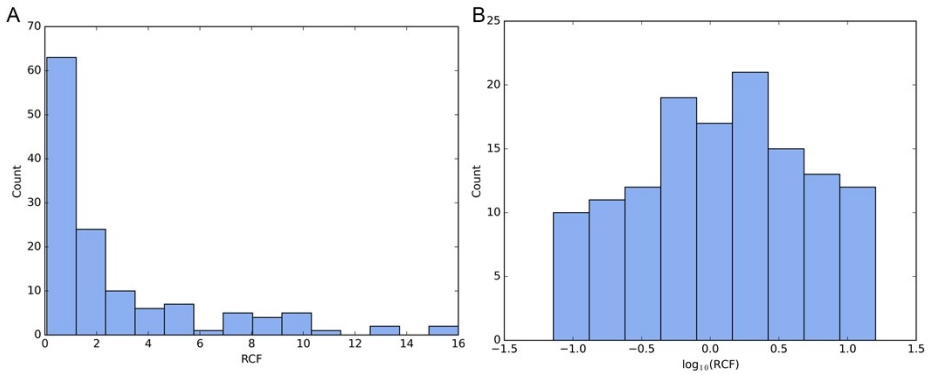

Fig. S3 Data distribution of RCF before (A) and after (B) logarithm transform.

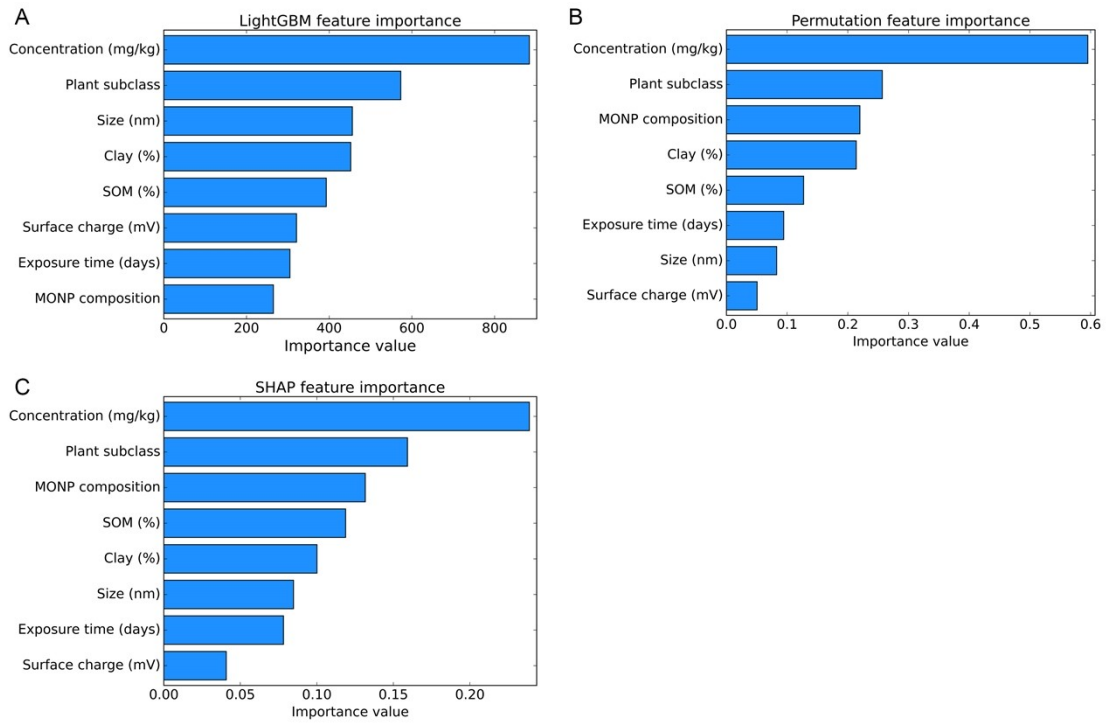

Fig. S4 Absolute feature importance of three measurement methods in the first dataset split.

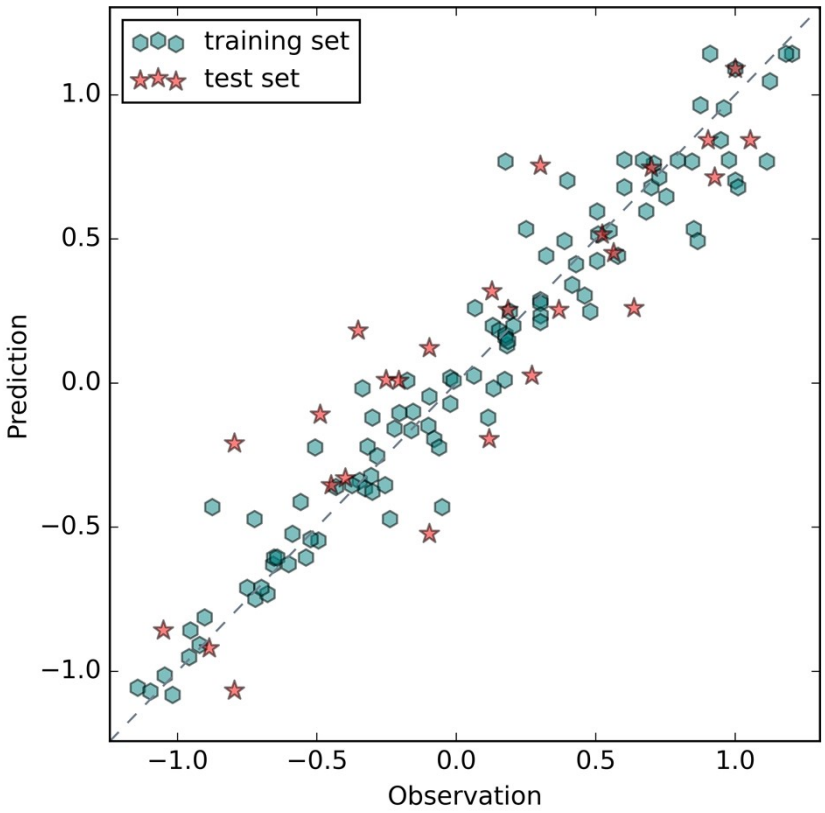

Fig. S5 Predicted versus observed logarithm transform of RCF values of the LightGBM model based on the sixth dataset split (stratified shuffle split).

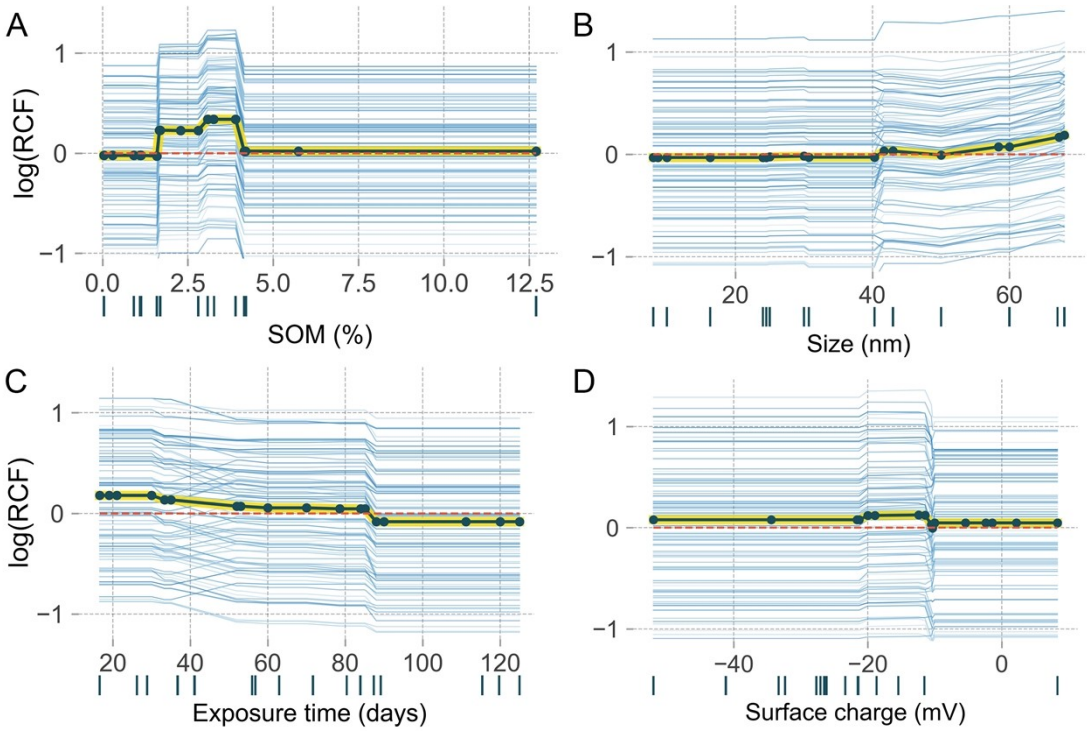

Fig. S6 PDP and ICE plots of the last four relevant features correlating RCF.

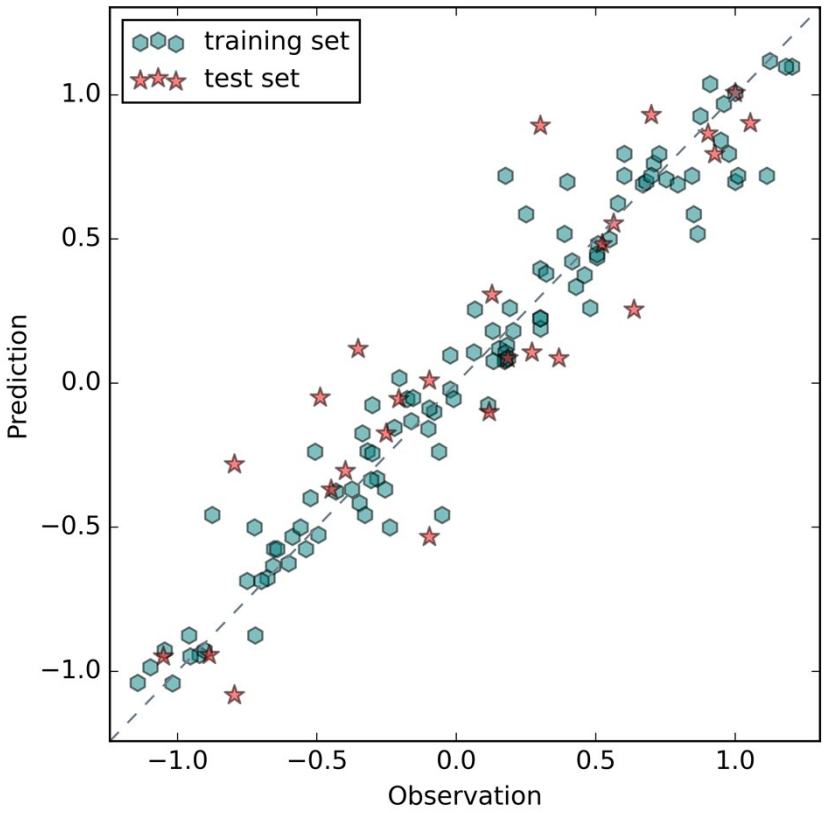

Fig. S7 Predicted versus observed logarithm transform of RCF values of the RuleFit regression based on the sixth dataset split (stratified shuffle split).

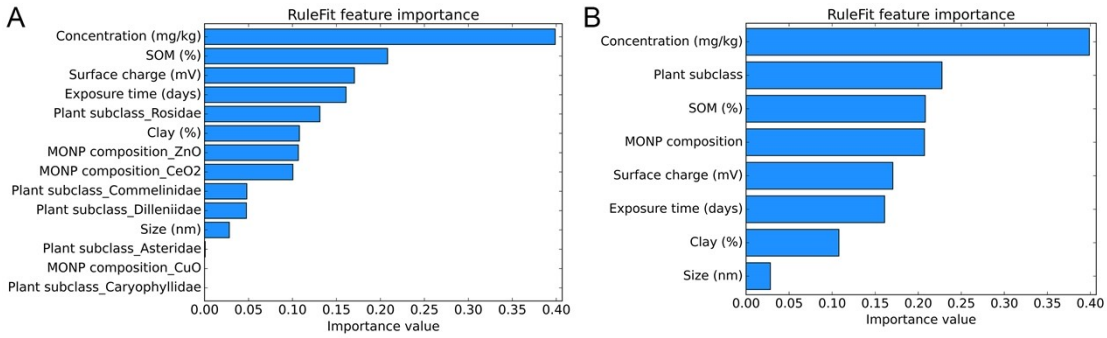

Fig. S8 (A) Absolute feature importance of input features in the RuleFit regression. (B) Absolute feature importance of features by adding the importance of one-hot encoded features.