Supporting Information

DiZyme: The Ultimate Resource for Nanozyme Multiple Catalytic Activity Prediction

Author Information

Affiliations

International Institute "Solution Chemistry of Advanced Materials and Technologies", ITMO University, Saint-Petersburg, 191002, Russian Federation

Julia Razlivina, Andrei Dmitrenko, and Vladimir Vinogradov

Advanced Engineering School, Almetyevsk, Russia

Julia Razlivina

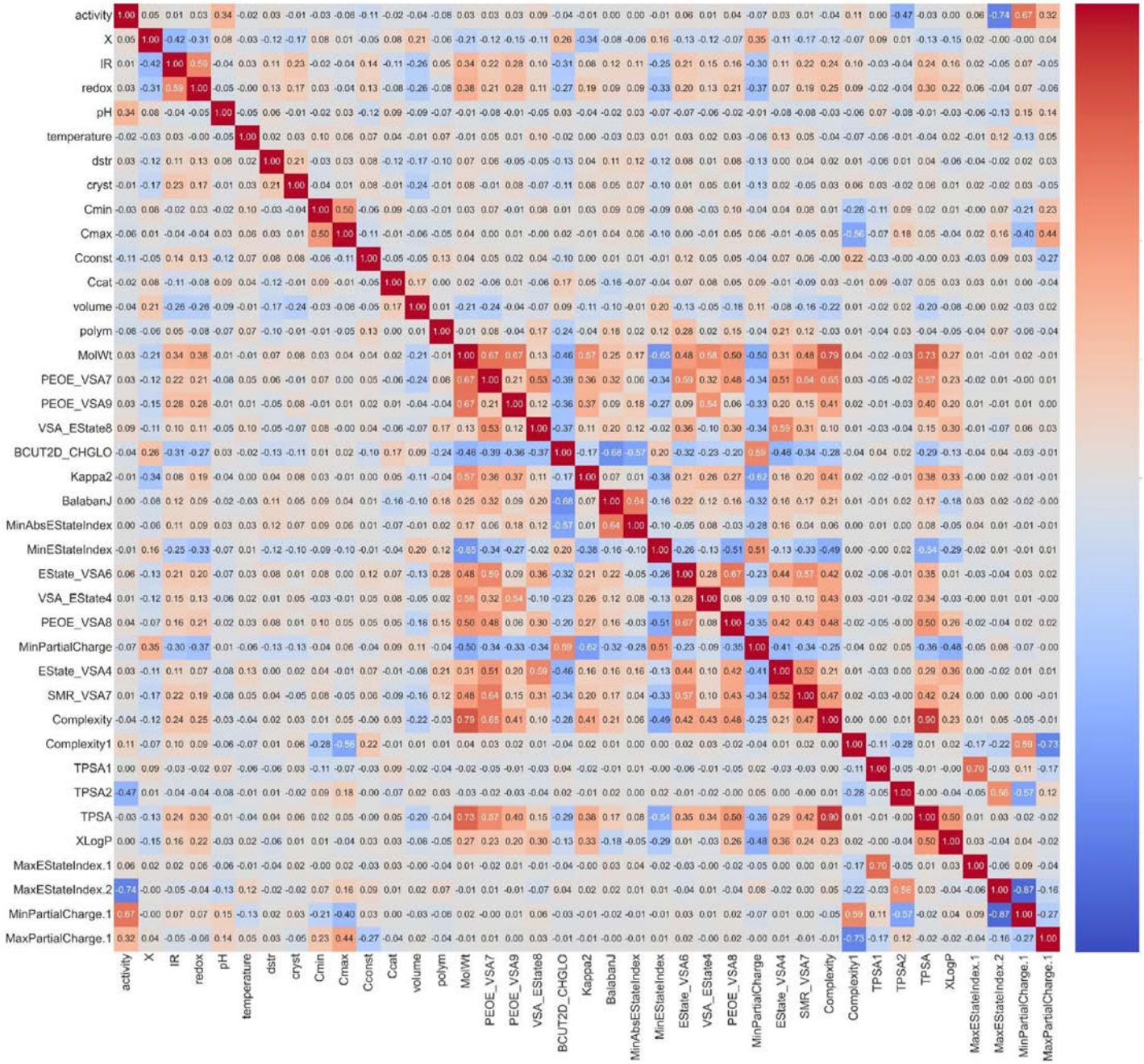

Figure S1. Final multiple catalytic activity dataset heatmap on the covariance matrix colored by Pearson's correlation coefficient.

Variable group	Code assignment	Variable description
target	Km	Michaelis-Menten constant, mM
	Vmax	Maximum velocity of reaction, mM/s
nanoparticle	activity	Type of activity
	cryst	Cristal system
	dstr	Dimensionality, calculated from nanoparticle sizes
	volume	cell volume, nm^3
	IR	The average ionic radius of nps, ang
	X	The average electronegativity of nps, mV
	redox	The average RedOx potential (to solid transition) of metals in nps
conditions	ph	ph
	temp	Temperature, ^oC
	Cmin	min concentration of substrate, mM
	Cmax	max concentration of substrate, mM
	Cconst	concentration of co-substrate, mM
	Ccat	concentration of nps, mg/ml
surface coating and dopants	Surface	Type of surface modification: 0-naked, 0.2-in synthesis, 1-coating
	polym	The polymerization number, calculated from coating molecular mass and its monomer
	MolWt	The molecular mass of one molecule of coating
	MinEStateIndex	Minimum EState index
	MinPartialCharge	Minimal partial charge
	MinAbsEStateIndex	Maximum absolute EState index
	TPSA	topological polar surface area
	BCUT2D_CHGLO	The Burden parameters are based on a combination of the atomic number for each atom and a description of the nominal bond-type for adjacent and nonadjacent atoms.

surface coating and dopants	EState_VSA6	MOE-type descriptors using EState indices and surface area contributions
	SMR_VSA7	MOE-type descriptors using MR contributions and surface area contributions
	EState_VSA4	MOE-type descriptors using EState indices and surface area contributions
	MaxEStateIndex	Maximum EState index
	Kappa2	Hall-Kier Kappa2 value
	BalabanJ	Balaban's J value
	PEOE_VSA7	MOE-type descriptors using partial charges and surface area contributions
	PEOE_VSA9	MOE-type descriptors using partial charges and surface area contributions
	XLogP	XLogP
	VSA_EState8	MOE-type descriptors using EState indices and surface area contributions
	VSA_EState4	MOE-type descriptors using EState indices and surface area contributions
	Complexity	The complexity rating of a compound is a rough estimate of how complicated the structure is, seen from the point of view of both the elements contained and the displayed structural features, including symmetry.
substrate	MaxEStateIndex.1	Maximum EState index
	MinPartialCharge.1	Minimal partial charge
	MaxPartialCharge.1	Maximal partial charge

substrate	TPSA.1	topological polar surface area
	Complexity.1	The complexity rating of a compound is a rough estimate of how complicated the structure is, seen from the point of view of both the elements contained and the displayed structural features, including symmetry.
co-substrate	MaxEStateIndex.2	Maximum EState index
	TPSA.2	topological polar surface area

Table S1. Final dataset features with descriptions.

ML model	Hyperparameters for Km prediction	R^2 for Km	MAE for Km	Hyperparameters for Vmax prediction	R^2 for Vmax	MAE for Vmax
Extra Trees Regressor	'max_depth': 16, 'max_features': None, 'min_samples_leaf': 2, 'n_estimators': 200	0.73	0.51	'max_depth': 64, 'max_features': None, 'min_samples_leaf': 2, 'n_estimators': 600	0.65	0.58
LGBM Regressor	'max_depth': 12, 'n_estimators': 50, 'num_leaves': 30, 'reg_alpha': 1.0	0.71	0.54	'class_weight': 'balanced', 'max_depth': 12, 'n_estimators': 50, 'num_leaves': 20, 'reg_alpha': 0.1	0.59	0.69
Random Forest Regressor	'max_depth': 80, 'max_features': None, 'min_samples_leaf': 2, 'n_estimators': 600	0.74	0.51	'max_depth': 32, 'max_features': None, 'min_samples_leaf':2, 'n_estimators': 200	0.66	0.60
CatBoost Regressor	'depth': 8, 'iterations': 500, 'l2_leaf_reg': 3, 'learning_rate': 0.05	0.74	0.52	'depth': 8, 'iterations': 500, 'l2_leaf_reg': 7, 'learning_rate': 0.1	0.73	0.57
Hist Gradient Boosting Regressor	'learning_rate': 0.08, 'max_bins': 200, 'max_depth': 12	0.74	0.50	'learning_rate': 0.23, 'max_bins': 100, 'max_depth': 12, 'min_samples_leaf': 10	0.76	0.55
AdaBoost Regressor	'learning_rate': 0.1, 'loss': 'exponential', 'n_estimators': 100	0.58	0.69	'learning_rate': 1, 'loss': 'square', 'n_estimators': 100	0.54	0.80

Table S2. Table of models hyperparameters and performance for multiple catalytic activity prediction.

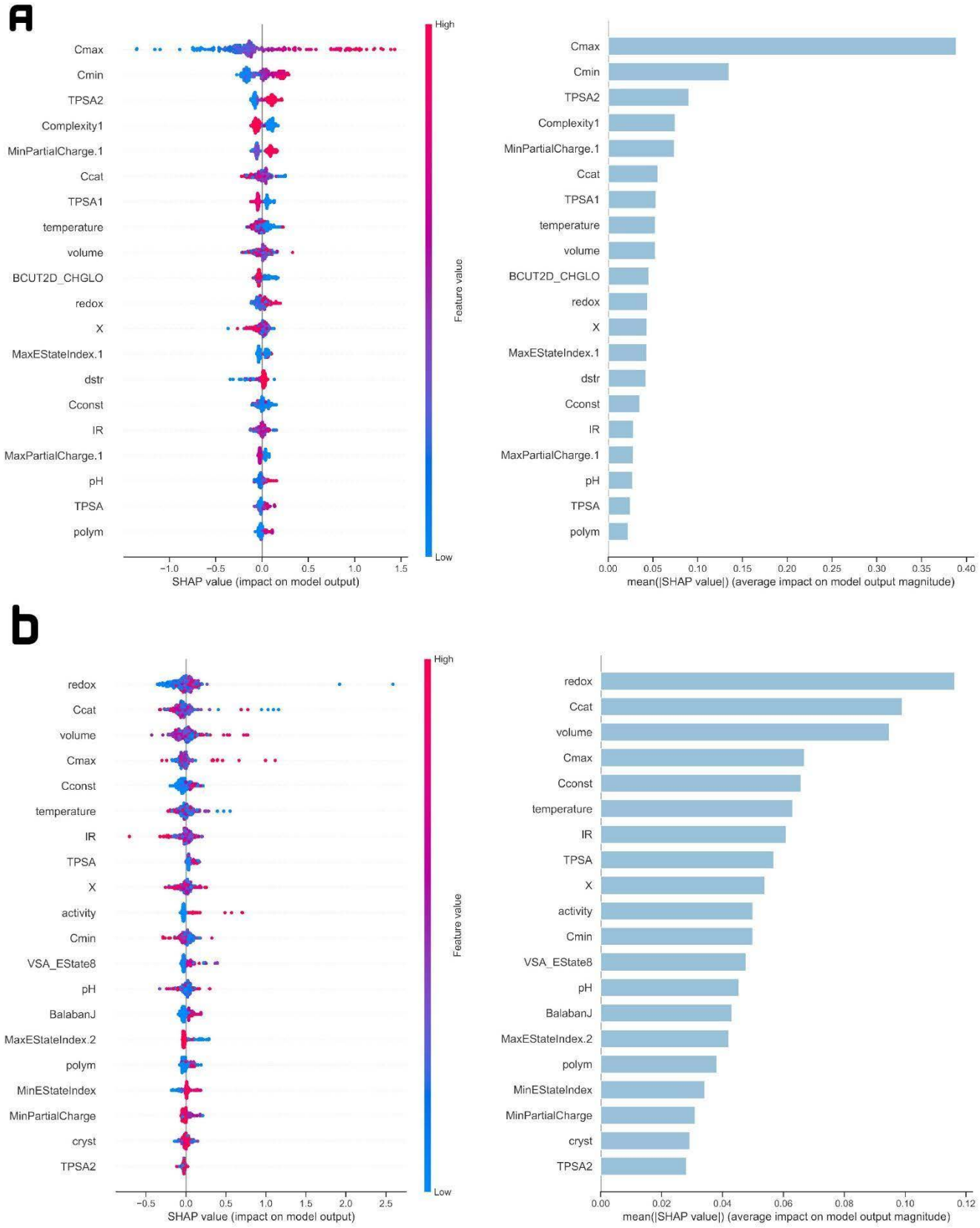

Figure S2. Feature importance SHAP-diagrams for categorical boosting regression interpretation of prediction: a) lgKm, b) lgVmax.

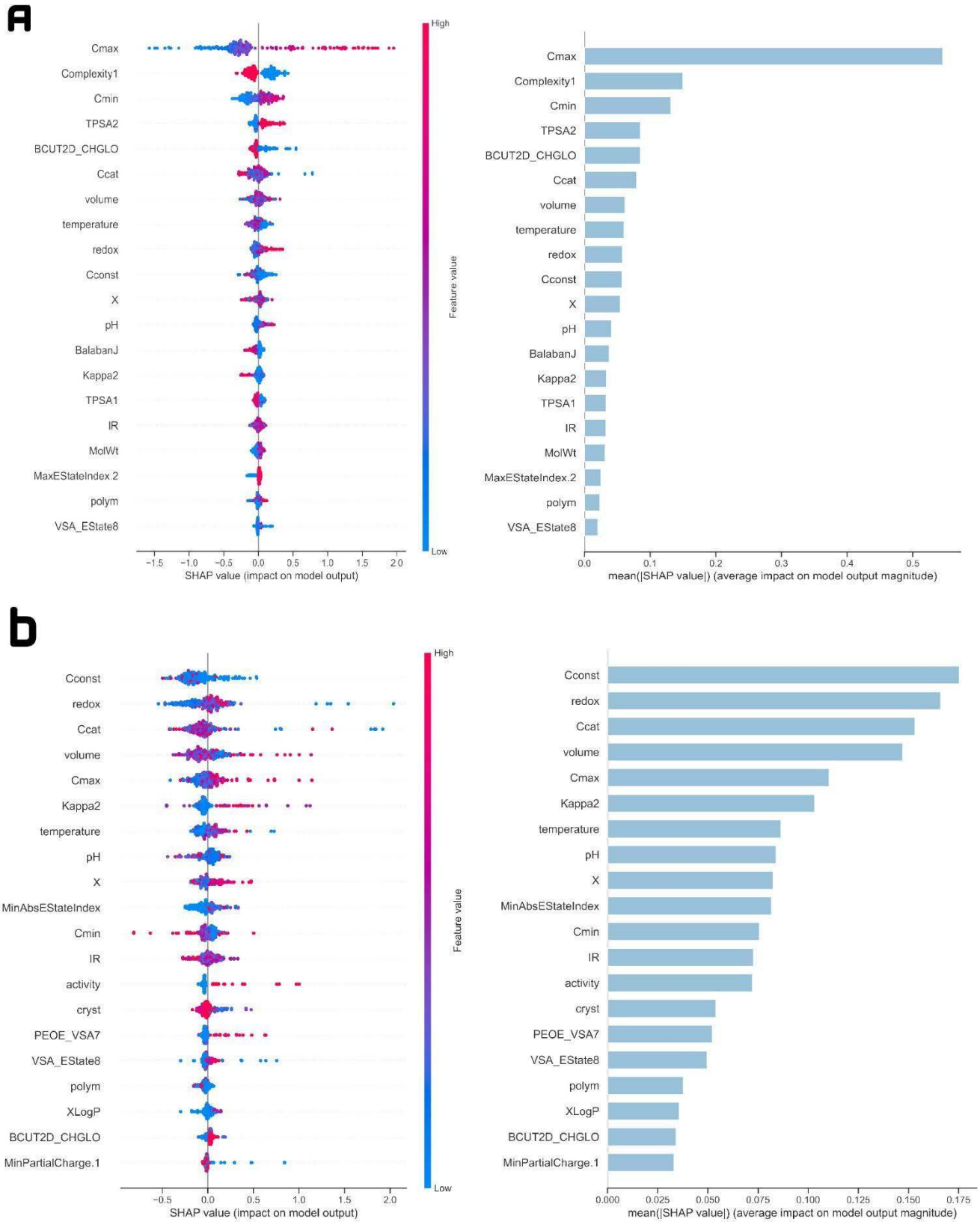

Figure S3. Feature importance SHAP-diagrams for histogram-based gradient boosting regression interpretation of prediction: a) lgKm, b) lgVmax.

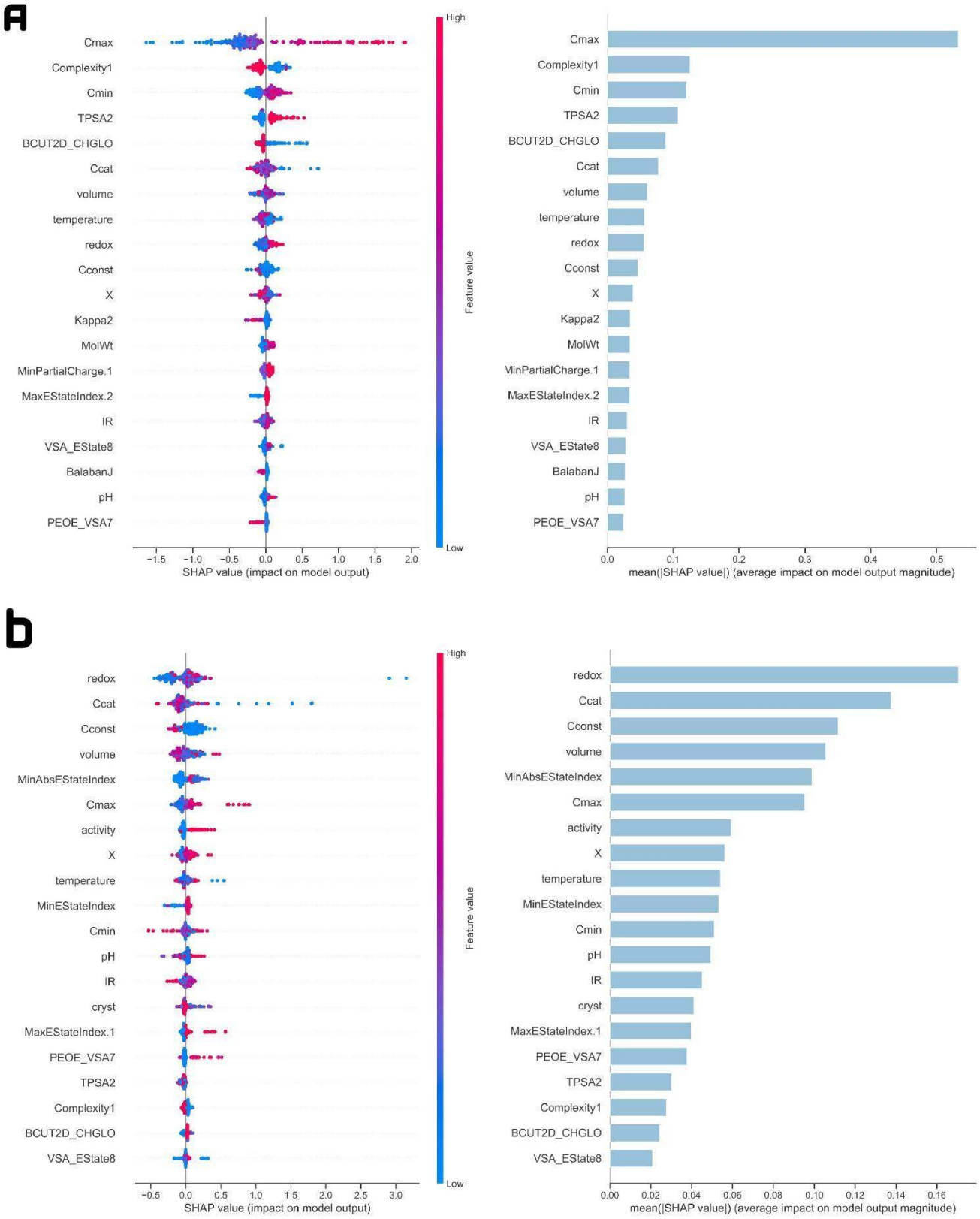

Figure S4. Feature importance SHAP-diagrams for light gradient boosting regression interpretation of prediction: a) lgKm, b) lgVmax.

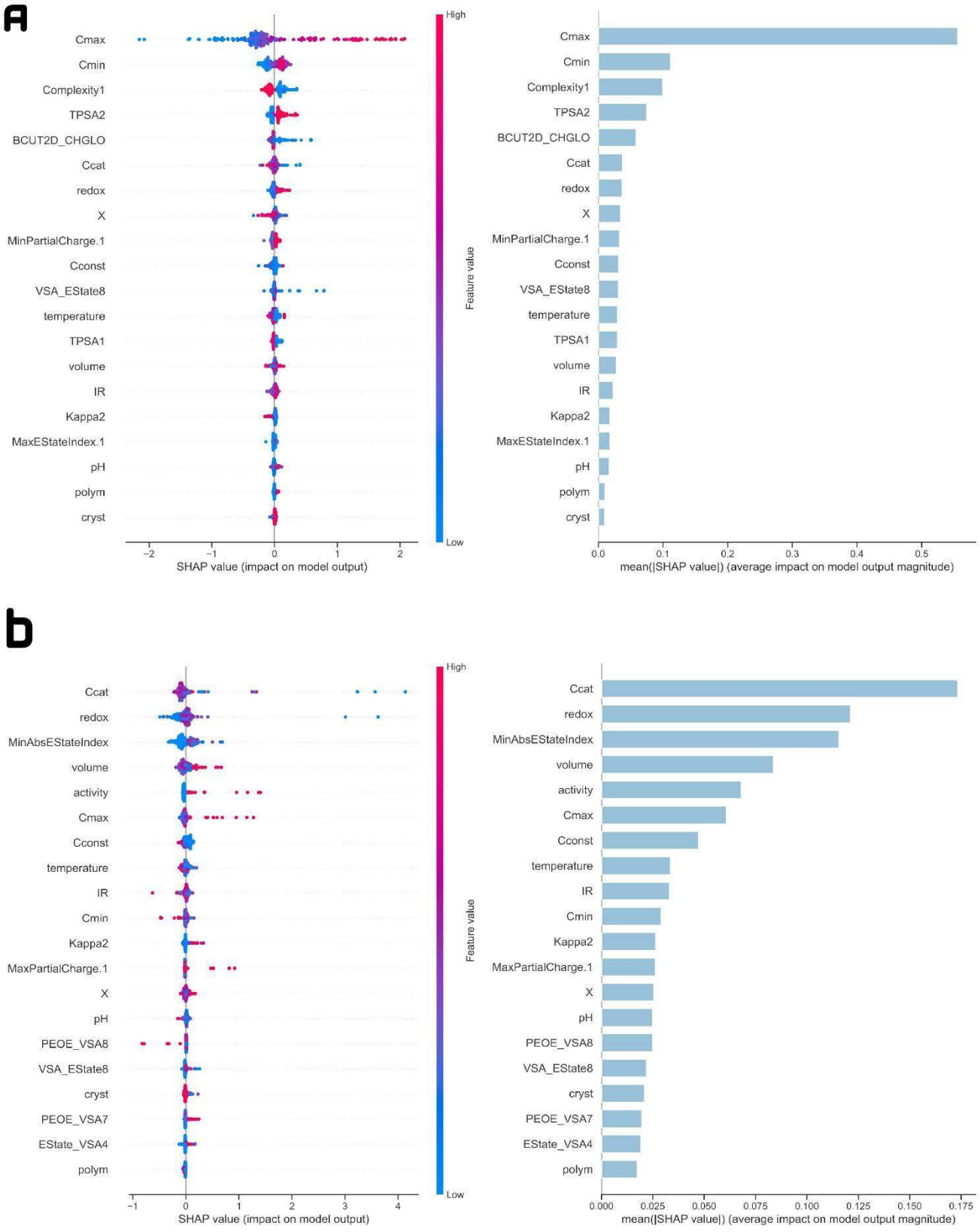

Figure S5. Feature importance SHAP-diagrams for random forest regression interpretation of prediction: a) lgKm, b) lgVmax.

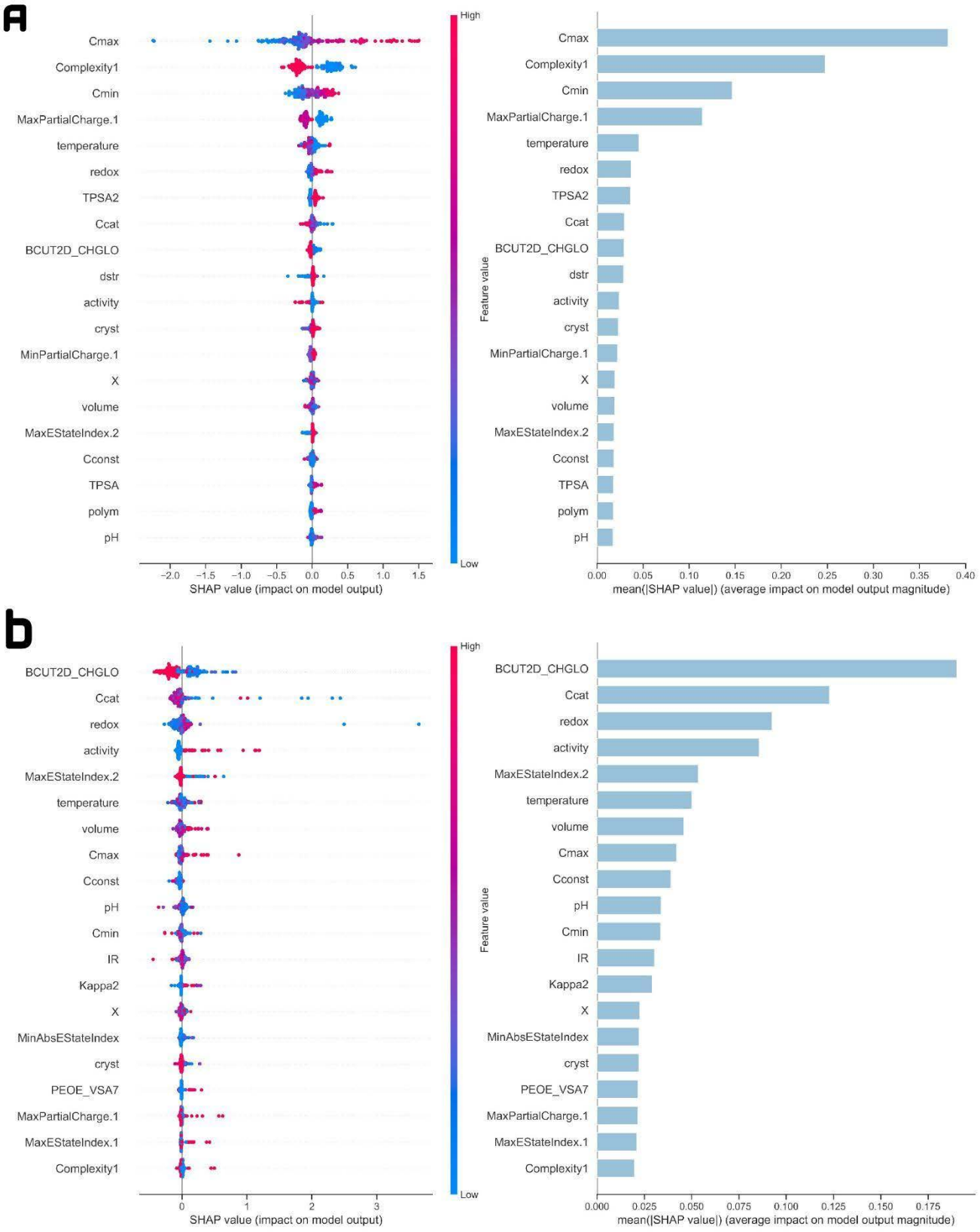

Figure S6. Feature importance SHAP-diagrams for extremely randomized trees regression interpretation of prediction: a) lgKm, b) lgVmax.

#	Chemical formula	Real lgKm	Predicted lgKm	$\left\lvert \underline{\Delta lgK_m} \right\rvert$	Real Vmax	Predicted Vmax	$\left\lvert \underline{\Delta lgV_{max}} \right\rvert$
1	$MnFe_2O_4$	-1.588	-1.605	0.017	0.190	-0.078	0.268
2	Fe_3O_4-MnO_2	-1.387	-0.009	1.378	-4.532	-4.183	0.349
3	MoO_2	0.947	-0.445	1.392	-4.284	-4.345	0.061
4	Fe_2O_3	-0.590	-0.355	0.235	-4.783	-4.836	0.054
5	VO_2	-0.836	-0.798	0.038	-5.883	-4.688	1.195
6	Ir	0.515	-0.101	0.616	-3.646	-4.006	0.360
7	Fe_3O_4	-0.018	-0.494	0.477	-4.284	-3.466	0.818
8	CeO_2	1.426	-0.463	1.888	-4.222	-3.964	0.258
9	CeO_2	0.255	-0.128	0.383	-3.301	-3.323	0.022
10	Co_9S_8	0.869	0.384	0.484	-3.456	-4.452	0.996
11	Mn_3O_4	0.100	-0.379	0.479	-3.626	-3.114	0.512
12	$Pt_{0.5}Ag_{0.5}$	-0.420	-0.438	0.018	-3.481	-3.859	0.378
13	FeS	-1.097	-1.041	0.056	-4.137	-4.269	0.132
14	Au@C	2.322	1.857	0.465	-3.481	-4.036	0.555
15	$Cu_{1.8}S$	0.236	0.053	0.182	-	-	-
16	$LaFeO_3$	-0.680	-0.019	0.661	-2.763	-4.516	1.752
17	$Fe_4(Fe(CN)_6)_3$	-0.119	-0.529	0.409	-	-	-
18	Ti_3C_2	-2.155	-0.373	1.782	-4.695	-4.198	0.497
19	Eu_2O_2S	-0.854	-0.756	0.097	-0.166	-3.638	3.473
20	Mo_5N_6	-1.086	-1.277	0.191	-5.046	-4.431	0.614
21	$Fe_{0.6}Ni_{0.4}S_2$	-0.099	-0.471	0.372	-3.764	-5.089	1.324

22	$Bi_2Fe_4O_9$	-1.155	-1.080	0.075	-4.664	-5.271	0.608
23	WTe_2	-0.658	-1.261	0.603	-3.813	-4.026	0.213
24	WTe_2	-0.509	0.691	1.199	-3.810	-4.276	0.466
25	MnOOH	-1.432	-1.479	0.047	-4.060	-3.334	0.726
26	$SiW_{12}@Co_3O_4$	-1.638	-1.004	0.634	-4.276	-3.630	0.645
27	Co	0.707	-0.117	0.824	-4.001	-3.760	0.241
28	Rh	-0.578	-0.801	0.222	-3.901	-3.361	0.540
29	$(FeS_2)_{0.3}$-SiO_2	-1.900	-0.461	1.438	-3.742	-4.293	0.551
30	MnO_2	-2.301	-0.980	1.321	-0.810	-4.080	3.270
31	$Pt_{0.5}$-MoO_3	-0.975	-0.550	0.425	-4.368	-5.208	0.841
32	PtFe	2.338	1.717	0.621	-4.087	-3.897	0.190
33	NiO	-1.943	-1.903	0.040	-3.317	-4.666	1.349
34	AuPt	0.946	1.163	0.216	-3.796	-2.916	0.880
35	VO_2	0.591	0.858	0.267	-3.948	-3.569	0.379
36	Pd@Pt-Ru_3	-0.201	-0.331	0.130	-2.699	-3.796	1.097
37	$Ag_{0.26}$-$Au_{0.3}@CeO_2$	0.477	-0.118	0.596	-4.000	-4.294	0.294
38	$CoMoO_4$	-1.369	-0.526	0.843	-4.059	-4.396	0.337
39	$Fe_2O_3/Fe_4(Fe(CN)_6)_4$	-2.000	-1.123	0.877	-3.910	-3.892	0.018
40	$NiCo_2O_4$-$Au_{0.44}$	-3.699	-3.773	0.074	-	-	-
41	Au-$Pt_{0.25}$	-0.788	-1.749	0.961	-3.916	-2.882	1.034
42	$NHMoO_3$	0.704	-0.204	0.908	-2.780	-2.442	0.338
43	Pt	-0.377	-0.554	0.177	-3.959	-4.380	0.421

44	RuTe	-1.726	1.523	3.249	-1.230	-3.170	1.940
45	Ag_3PO_4	0.320	-0.772	1.092	-1.616	-3.701	2.085
46	MoS_2@$MgFe_2O_4$	-0.623	-0.230	0.393	-3.423	-3.394	0.029
47	$Pd_{3.5}$-Ir	-0.398	-0.467	0.069	-3.051	-2.842	0.208
48	Au@Pt	-0.572	0.167	0.739	-4.815	-3.455	1.360
49	$Co_{1.5}Mn_{1.5}O_4$	-2.538	0.290	2.828	-4.724	-0.533	4.191
50	Pd	2.131	2.164	0.033	-2.229	-3.706	1.477

Table S3. Table of ensemble model performance on validation dataset.

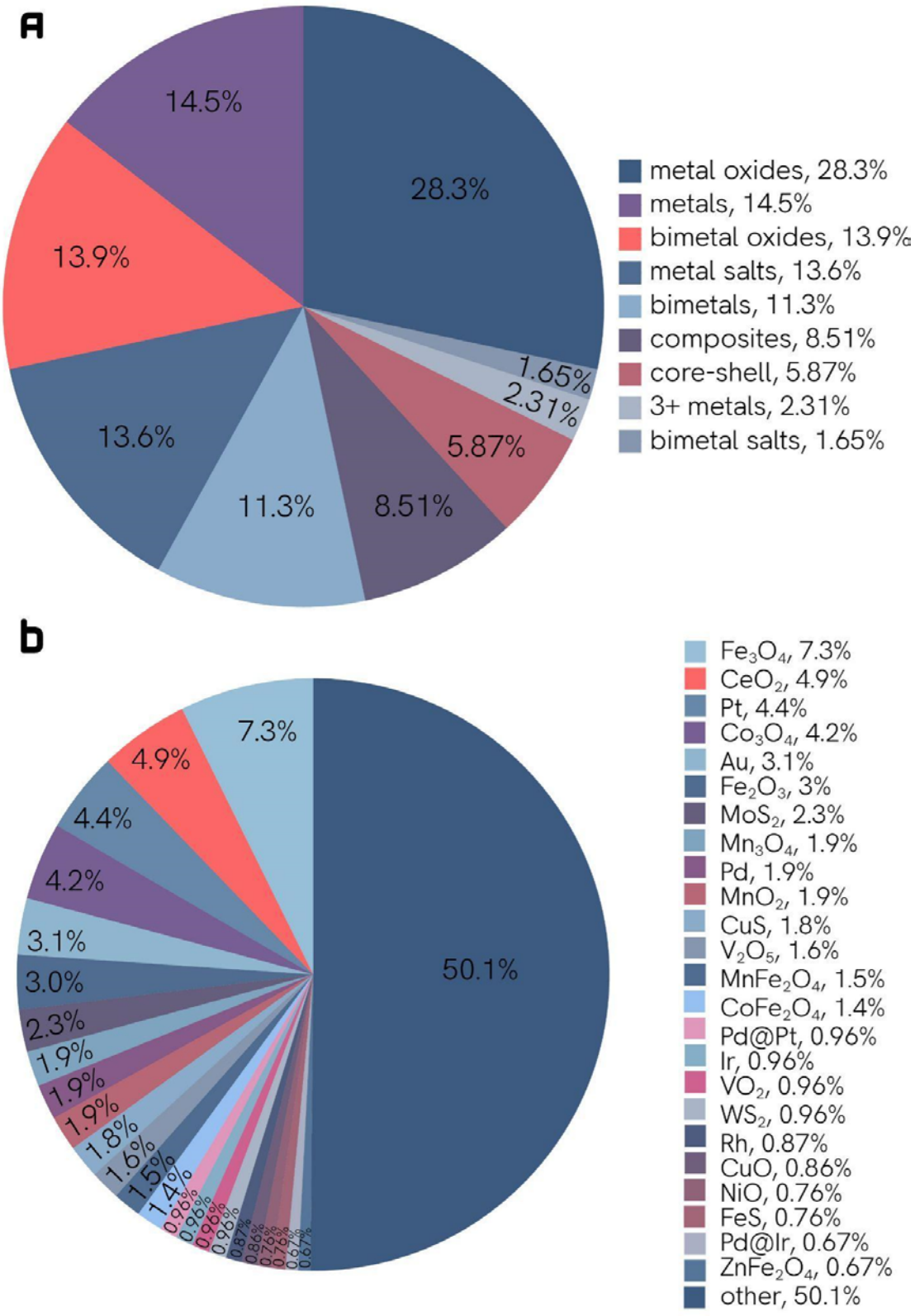

Figure S7. Pie chart of the distribution of compositions nanozymes in the database. a) most represented types of material compositions; b) most represented material compositions (amount more than 0.6%);

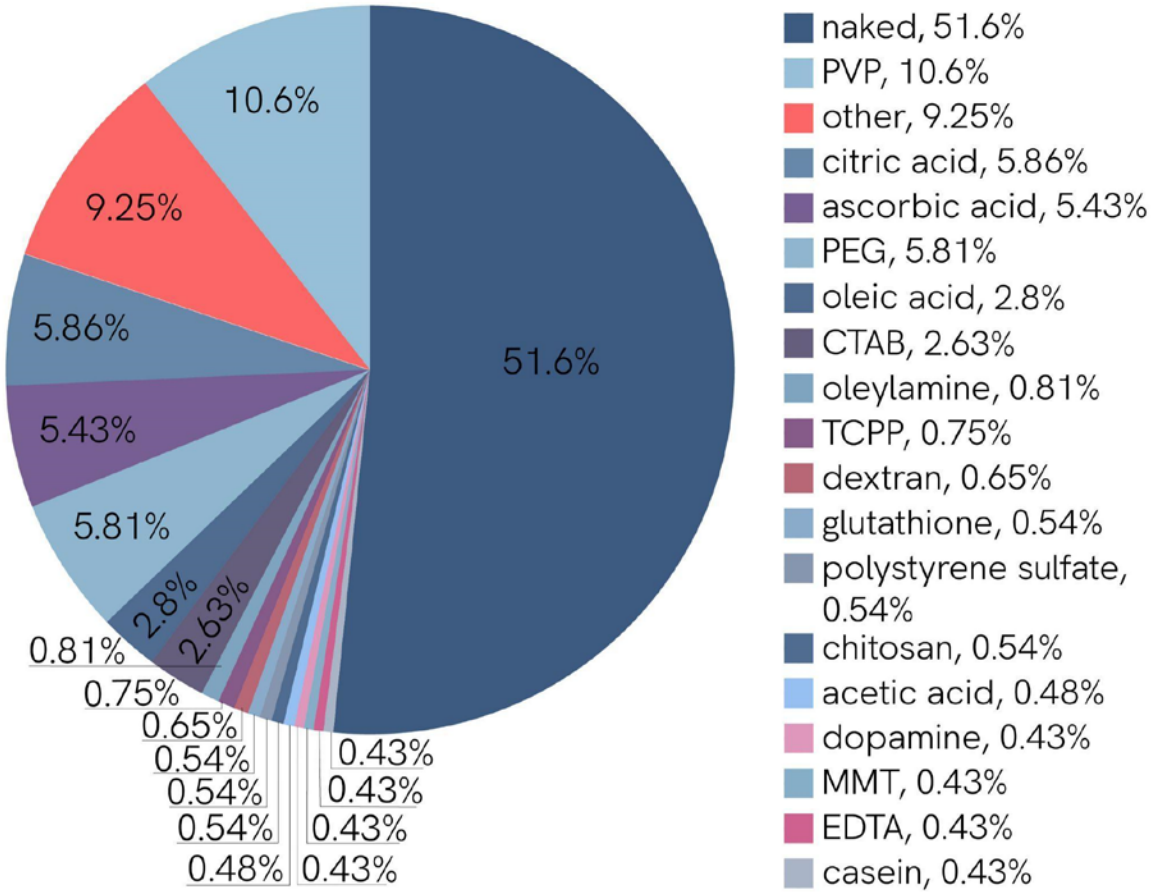

Figure S8. Pie chart of the distribution of most represented surface coatings molecules (amount more than 0.3%) in the database.

The DiZyme Assistant

The DiZyme Assistant, an LLM model based on ChatGPT, was recently evaluated on a nanozyme database question answering task. The assistant was evaluated on 150 query-response pairs where the queries covered different types of questions about nanozymes. The responses were compared against expert annotations as to whether the assistant's response contained the information needed to answer the query based on the referenced article.

The assistant demonstrated strong performance across several accuracy metrics:

- F1 score (micro-averaged): 0.847
- F1 score (macro-averaged): 0.842
- Precision: 0.974
- Recall: 0.784
- Accuracy: 0.847

The F1 score, precision, and recall metrics were calculated based on a binary classification of the model's responses into "I don't know" (label 0) and all other responses (label 1). The F1 score balances precision, the fraction of relevant responses out of all positive responses, and recall, the fraction of relevant responses retrieved. The micro-averaged F1 score calculates metrics globally, while the macro-averaged F1 calculates metrics for each class and takes the average leading to more weight for under-represented classes.

For this nanozyme task, the DiZyme Assistant achieved strong precision, returning relevant responses for 97.4% of its positive "answer exists" predictions. Its recall indicates there is still room for improvement in retrieving all existing answers. But overall, with micro and macro F1 scores above 0.84, the assistant demonstrates promising accuracy and balance of precision and recall for this challenging scientific domain. Enhancements to the model training and architecture hold promise for boosting recall further, while maintaining high precision.