Supporting Information

UniESA: a Unified Data-Driven Framework for Enzyme Stereoselectivity and Activity Prediction

Chun-Yue Weng [1, 2, 3] Jun Li [1, 2, 3], Qi-Lin Chen[1, 2, 3], Jia-Yi Han[1, 2, 3], Zhi-Tao Dong[1, 2, 3], Zhi-Qiang Liu[1, 2, 3, *], Yu-Guo Zheng[1, 2, 3]

[1]Key Laboratory of Bioorganic Synthesis of Zhejiang Province, College of Biotechnology and Bioengineering, Zhejiang University of Technology, Hangzhou 310014, China.

[2]Engineering Research Center of Bioconversion and Biopurification of Ministry of Education, Zhejiang University of Technology, Hangzhou 310014, China.

[3]National and Local Joint Engineering Research Center for Biomanufacturing of Chiral Chemicals, Zhejiang University of Technology, Hangzhou 310014, P. R. China.

***Corresponding author:**

Prof. Zhi-Qiang Liu

College of Biotechnology and Bioengineering

Zhejiang University of Technology

No. 18, Chaowang Road, Hangzhou 310014, China

E-mail: microliu@zjut.edu.cn

Tel: +86-571-88320379

Fax: +86-571-88320379

AAindex encoding and FFT

We encoded amino acid sequences into numerical matrices using entries from the AAindex database, replacing each amino acid's properties with the relevant numeric values from their respective AAindex descriptors. Such an encoding is performed for all entries of the AAindex database, which provide data for all 20 naturally amino acids (a list of all 566 indices can be obtained from **https://www.genome.jp/aaindex/AAindex/list_of_indices**).

For further featurization, AAindex encoded sequence vectors are transferred to the sequence domain using fast Fourier transform (FFT). The algorithm used for FFT constrains the shape of $F \in R^{1 \times N_{Dim}}$ to

$$N_{Dim} = 2^k,\ k \in N\#(S1)$$

If N_{Dim} does not hold for the upper equation, **F** is reshaped by adding columns of zeros (zero padding) until the condition is fulfilled. This normalized (and zero padded) matrix is in a next step FFT-ed following

$$\left|\hat{F}_j\right| = \left| \sum_{n=0}^{N_{Dim}-1} \tilde{x}_n exp\left(-\frac{2\pi i j n}{N_{Dim}}\right) \right|,\ \hat{F}_j \in C\#(S2)$$

with the spectrum frequencies $j \in [0,1,2,\cdots,N_{Dim}-1]$ and the imaginary number $i^2 = -1$.

Table S1. The initial dataset for machine learning

mutants	de_p (%)	Relative activity (%)	mutants	de_p (%)	Relative activity (%)
WT	66.6	100.00	F162M	74.7	103.00
V85Y	74.7	39.11	F162L	95.5	114.50
V85T	99.5	9.70	F162I	99.5	65.30
V85S	85.7	36.40	F162H	99.5	30.00
V85N	34.5	34.80	F162G	99.5	66.50
V85L	99.5	11.69	F162E	82.3	60.85
V85I	25.8	7.03	F162D	99.5	67.80
V85F/F216S	88	3.50	F162C	91.6	67.79
V85F/F216D	24.8	16.20	E87Y	96.1	111.52
V85F	-19.4	14.90	E87V	66.2	71.32
V85D	99.5	17.48	E87S	-11	7.90
V85C	55.1	32.18	E87R	-5.2	7.30
V85A	99.5	20.22	E87N	96.3	65.85
V239Y	74.3	18.90	E87F	99.5	69.97
V239S	92.7	66.90	E87D/V239N/A129Y	99.5	1.20
V239R	66.8	98.20	E87D/V239N/A129V	-78	43.00
V239N	40.2	105.10	E87D/V239N/A129T	-76.5	26.30
V239L	93.4	73.90	E87D/V239N/A129Q	99.5	2.60
V239I	95.7	9.42	E87D/V239N/A129N	93	9.70
V239H	71.6	17.10	E87D/V239N/A129M	-49	5.80
V239G	79.3	122.00	E87D/V239N/A129L	-99.5	6.80
V239F	84.2	27.20	E87D/V239N/A129K	99.5	3.00
V239D	61.9	53.10	E87D/V239N/A129I	-99.5	3.40
V239C	56.8	86.70	E87D/V239N/A129G	88.1	90.20
T217Y	83.6	6.50	E87D/V239N/A129D	-99.5	1.10
T217V	80.3	67.90	E87D/V239N	8.5	64.70

T217S	38.7	89.10	E87D/L96C	26.9	15.40
T217P	29.6	86.50	E87D/F216S	30.8	25.70
T217N	88.4	9.20	E87D/A129C/V239N	-99.5	7.20
T217L	99.5	10.30	E87D	-6.5	5.60
T217I	99.5	122.00	E87C	99.5	0.89
T217G	32.8	27.80	E134Y	98.2	123.20
T217F	35.8	16.40	E134V	97.9	114.20
T217C	99.5	23.00	E134Q	94.7	82.90
L96Y	85	13.15	E134P	99.5	126.70
L96V	90.8	135.60	E134N/V239N/A129V	-43.7	25.70
L96R	99.5	10.30	E134N/V239N	34.9	91.90
L96Q	99.5	37.08	E134N/F216S	96.8	31.70
L96N	-8.4	13.20	E134N/E87R	99.5	101.20
L96I	83	106.70	E134N/E87D	55.6	43.60
L96H	99.5	34.26	E134N	38.6	61.10
L96F	-28.5	55.00	E134L	98.7	126.90
L96E	99.5	20.73	E134K	98.9	78.40
L96C/F216D	43.3	7.30	E134I	99.1	116.60
L96C	-17.4	86.50	E134G	99.5	66.80
L96A	99.5	58.01	E134F	98.2	121.40
F216Y	70.7	86.30	E134D	98.6	134.30
F216V	54.8	89.00	E134C	98.2	131.70
F216S	8.8	68.00	E134A	99.5	7.16
F216L	45.7	40.80	C198T	85.1	91.00
F216G	27.4	76.70	C198S	96.8	77.00
F216D	9.2	22.30	C198P	98.3	72.60
F216C	63.3	24.20	C198L	97.2	70.30
F199Y	99.5	70.50	C198F	97.9	72.51
F199V	99.5	127.30	C198A	53.4	88.80

F199S	96.7	87.30	A129V/F216S	80	89.13
F199R	85.7	28.20	A129V/E87D	24.9	40.70
F199N	80.4	67.40	A129V	-6.8	18.90
F199L	96.5	110.50	A129S	99.5	10.25
F199I	99.5	49.70	A129P	99.5	63.92
F199H	99.5	35.20	A129N	99.5	18.73
F199G	99.5	70.20	A129M	-34.8	1.95
F199E	80.4	58.10	A129L	99.5	1.76
F199D	99.5	70.21	A129I	99.5	1.81
F199C	97.6	54.40	A129H	-31	10.80
F162Y	99.5	76.52	A129G	99.5	103.80
F162V	99.5	129.80	A129F	99.5	0.85
F162S	95.7	120.30	A129C/E87D	-39.8	3.10
F162R	83.7	25.70	A129C	-38.3	10.80
F162N	94	80.60			

Table S2. The Protein Language Models for sequences encoding

Protein Language Model	Size	Embedding Dim	Dataset
ProteinBERT	16M	347	UR90 2022_02
esm2_t12_35M_UR50D	35M	480	UR50/D 2021_04
esm1v_t33_650M_UR90S_[1-5]	650M	1280	UR90/S 2020_03

Table S3. The regression algorithms for model training

Machine learning regression algorithm	Module
KNeighborsRegressor	sklearn
SVR	sklearn
Ridge	sklearn
Lasso	sklearn
MLPRegressor	sklearn
DecisionTreeRegressor	sklearn
ExtraTreeRegressor	sklearn
RandomForestRegressor	sklearn
AdaBoostRegressor	sklearn
GradientBoostingRegressor	sklearn
BaggingRegressor	sklearn
PLSRegression	sklearn
BayesianRidge	sklearn
ElasticNet	sklearn

Table S4. The evaluation metrics for model validation

Evaluation metrics	Module
mean_absolute_error	sklearn.metrics
r2_score	sklearn.metrics
mean_squared_error	sklearn.metrics
pearsonr	scipy.stats
compare_nrmse	skimage.metrics

Table S5. Performance of UniESA on NOD and ANEH datasets

dataset	dataset split	encoding	R^2	RMSE	NRMSE	PCC
ANEH	Manual	WOLR810101	0.74	13.86	0.22	0.88
NOD	Manual	ProteinBERT	0.65	0.25	0.57	0.81
NOD	Automatic	ProteinBERT	0.80	0.16	0.40	0.89

Table S6. Performance of UniESA on LEH, IRED and ATAs datasets

dataset	Algorithm	encoding	R^2	RMSE	NRMSE	PCC
LEH	GradientBoosting	AAindex_ NAKH920105	0.98	6.85	0.09	0.99
IRED	BayesianRidge	AAindex_TANS770109	0.83	0.19	0.32	0.91
ATAs1	GradientBoosting	ProteinBERT	0.78	89.24	0.37	0.92
ATAs2	GradientBoosting	AAindex_NAKH900102	0.87	684.05	0.31	0.94

Table S7. Performance of AAindex (FFT) encoded sequence stereoselectivity prediction model on validation set

	Max_R²	
	AAindex	AAindex_FFT
KNeighborsRegressor	0.0000	0.0000
SVR	0.6700	0.7200
Ridge	0.4900	0.5900
Lasso	0.4900	0.6000
MLPRegressor	0.0000	0.0039
DecisionTreeRegressor	0.5552	0.7141
ExtraTreeRegressor	0.4161	0.6524
RandomForestRegressor	0.6077	0.6421
AdaBoostRegressor	0.6000	0.6600
GradientBoostingRegressor	0.6500	0.7000
BaggingRegressor	0.5598	0.6889
PLSRegression	0.4900	0.5900

BayesianRidge	0.1713	0.5366
ElasticNet	0.4900	0.5900

Table S8. Performance of PLMs encoded sequence stereoselectivity prediction model on validation set

	ProteinBERT				esm2_t12_35M_UR50D				esm1v_t33_650M_UR90S_[1-5]			
	MAE	RMSE	R^2	PCC	MAE	RMSE	R^2	PCC	MAE	RMSE	R^2	PCC
KNeighborsRegressor	0.2312	0.1633	0.6691	0.8215	0.2343	0.1825	0.5768	0.8205	0.2225	0.1894	0.6182	0.8212
SVR	0.3215	0.2518	0.0004	0.3680	0.3221	0.2519	-0.0028	0.0161	0.5095	0.3178	0.0000	0.7743
Ridge	0.2613	0.2049	0.7182	0.9047	0.3917	0.3530	0.0502	0.7756	0.3830	0.2912	0.1924	0.6903
Lasso	0.2449	0.2133	0.5778	0.9218	0.4246	0.3433	0.0000	-	0.3445	0.3140	0.0000	-
MLPRegressor	0.4834	0.3707	0.0116	0.8895	0.4151	0.3714	0.0025	0.2768	0.4009	0.3528	0.0081	0.6528
DecisionTreeRegressor	0.3747	0.2260	0.4203	0.6810	0.3692	0.2618	0.3804	0.6591	0.3249	0.2017	0.5060	0.7644
ExtraTreeRegressor	0.2830	0.1650	0.4316	0.7498	0.3497	0.2282	0.4764	0.7204	0.3804	0.2790	0.4564	0.7074
RandomForestRegressor	0.3386	0.2457	0.5266	0.7514	0.3503	0.2858	0.4257	0.7087	0.3237	0.2349	0.5126	0.7365
AdaBoostRegressor	0.3054	0.2037	0.6149	0.7901	0.3468	0.2845	0.5126	0.7271	0.2262	0.1828	0.6055	0.8089
GradientBoostingRegressor	0.2655	0.1899	0.6722	0.8333	0.3321	0.2560	0.4838	0.7261	0.3788	0.2896	0.4185	0.6740
BaggingRegressor	0.3507	0.2786	0.4927	0.7424	0.3907	0.2981	0.4266	0.6755	0.3523	0.2743	0.4969	0.7097
PLSRegression	0.4102	0.3041	0.4064	-	0.2604	0.1992	0.5800	-	0.2543	0.1838	0.5996	-
BayesianRidge	0.2433	0.1790	0.6336	0.7998	0.3545	0.2996	0.0091	0.4048	0.2570	0.1856	0.5909	0.7743
ElasticNet	0.3166	0.2804	0.3795	0.7840	0.3385	0.3060	-0.0265	-	0.3385	0.3062	0.0000	-

Table S9. Performance of AAindex (FFT) encoded sequence relative-activity prediction model on validation set

	Max_R²	
	AAindex	AAindex_FFT
KNeighborsRegressor	0.0000	0.0000
SVR	0.5900	0.5900
Ridge	0.4300	0.4600

Lasso	0.4500	0.4700
MLPRegressor	0.0000	0.2000
DecisionTreeRegressor	0.1398	0.2469
ExtraTreeRegressor	0.2726	0.5040
RandomForestRegressor	0.4935	0.5864
AdaBoostRegressor	0.5400	0.5200
GradientBoostingRegressor	0.5300	0.5700
BaggingRegressor	0.3543	0.6530
PLSRegression	0.4700	0.4800
BayesianRidge	0.1713	0.2485
ElasticNet	0.4000	0.3900

Table S10. Performance of PLMs encoded sequence relative-activity prediction model on validation set

	ProteinBERT				esm2_t12_35M_UR50D				esm1v_t33_650M_UR90S_[1-5]			
	MAE	RMSE	R^2	PCC	MAE	RMSE	R^2	PCC	MAE	RMSE	R^2	PCC
KNeighborsRegressor	0.3257	0.2518	0.4153	0.6670	0.3511	0.2910	0.3182	0.5843	0.3835	0.3284	0.2708	0.5337
SVR	0.3531	0.3100	0.0004	0.2769	0.3532	0.3101	0.0000	0.1654	0.3989	0.3563	0.0000	0.5120
Ridge	0.3650	0.2933	0.2634	0.5669	0.3618	0.3055	0.0076	0.2390	0.3668	0.3260	0.0201	0.3561
Lasso	0.3847	0.3374	0.1815	0.6304	0.4028	0.3614	0.0000	-	0.3532	0.3101	0.0000	-
MLPRegressor	0.3758	0.3193	0.0007	0.2478	0.4229	0.3741	0.0027	0.2294	0.3548	0.3221	0.0029	0.3432
DecisionTreeRegressor	0.3579	0.2691	0.2344	0.5810	0.3901	0.2919	0.1060	0.5790	0.4100	0.3339	0.0000	0.3783
ExtraTreeRegressor	0.3196	0.2256	0.3896	0.6497	0.3976	0.3144	0.0190	0.5464	0.4500	0.3655	0.0000	0.4314
RandomForestRegressor	0.3718	0.3330	0.1898	0.4683	0.3559	0.3018	0.2145	0.5947	0.3454	0.2837	0.1558	0.4531
AdaBoostRegressor	0.3881	0.3306	0.2533	0.5292	0.2938	0.2453	0.3893	0.6352	0.3761	0.3181	0.1691	0.4113
GradientBoostingRegressor	0.3567	0.2983	0.2853	0.5519	0.3008	0.2318	0.2761	0.5879	0.4005	0.3292	0.0705	0.2922
BaggingRegressor	0.3195	0.2692	0.1819	0.4286	0.3767	0.3045	0.2156	0.4757	0.3946	0.3326	0.1703	0.4511
PLSRegression	0.3118	0.2460	0.1887	-	0.3780	0.3210	0.1050	-	0.3577	0.3082	0.1086	-
BayesianRidge	0.3698	0.3215	0.1214	0.4643	0.4189	0.3696	0.0215	0.2393	0.4219	0.3783	0.0046	0.0730
ElasticNet	0.3634	0.3246	0.0376	-	0.3477	0.2991	-0.0083	-	0.3477	0.2991	0.0000	-

Prediction of *KmCR*

6,787 unique single-point *KmCR* mutant sequences (excluding sequences already present in the dataset) were generated for following virtual screening by UniESA-generated models.

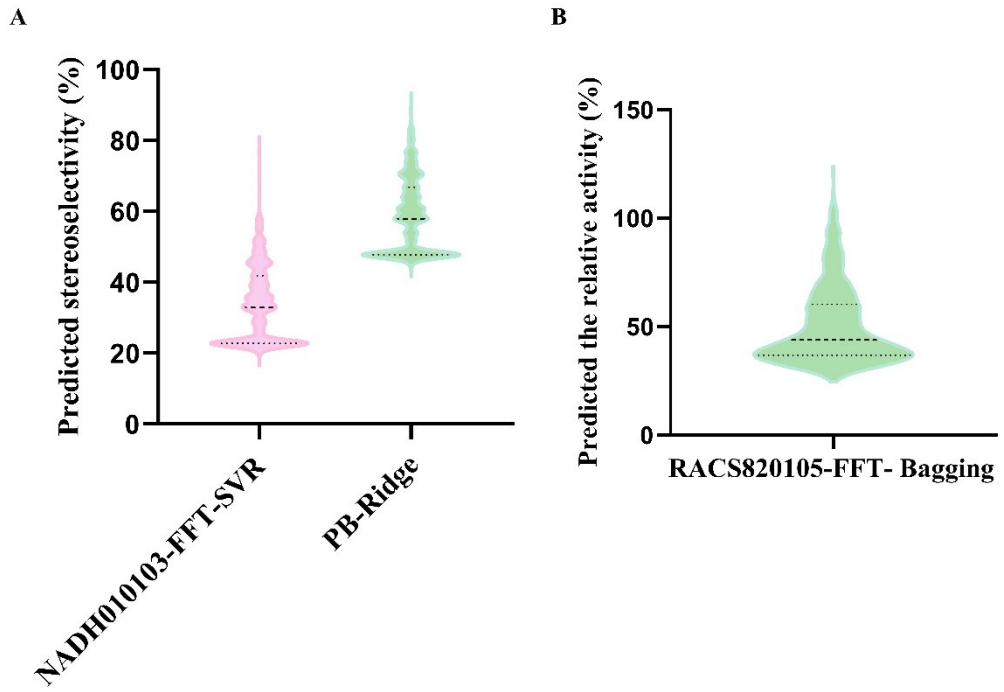

Figure S1 Prediction of single-point *KmCR* mutant sequences.

Experimental confirmation

Primer design and creation of *Km*CR mutants: Site-directed Mutagenesis of *Km*CR was constructed using whole-plasmid mutagenesis protocol with pET-28a-kmcr as the template. The PCR reaction system (50 μL) consisted of 1 μL forward primer, 1 μL reverse primer, 25 μL 2 × Phanta Buffer, 1 μL dNTP mixture, 1 μL plasmid template, 1 μL DNA Polymerase, and 20 μL ultra-pure water. The PCR conditions: 95 ℃ for 5 min, and then 30 cycles (95 ℃ for 30 s, 55 ℃ for 15 s, 72 ℃ for 4 min), 72 ℃ extension for 10 min. The PCR products were analyzed by electrophoresis on an agarose gel and digested at 37 °C for 30 minutes using *Dpn* I enzyme. After digestion, the resulting

recombinant plasmids were transformed directly into *E. coli* BL21 (DE3) receptor cells to produce mutants. Clones were screened by incubation at 37 °C for 12 h.

Protein expression: Each colony was transferred to 10 mL of LB medium supplemented with 50 μg/mL kanamycin, and then grown at 37 °C and 180 rpm for 12 hours. Subsequently, 1 mL of the culture was mixed with 1 mL of 30% (*v*/*v*) glycerol solution and stored at -80 °C. Next, 2 mL of the culture was transferred to 100 mL of LB medium containing 50 μg/mL kanamycin, and shaken at 37 °C and 180 rpm for 2 hours until the OD_{600} reached 0.6~0.8, followed by induction of protein expression with 0.2 mM IPTG. The cells were then grown at 28 °C and 180 rpm for 12 hours, harvested, and centrifuged at 8000 rpm for 10 minutes at 4 °C.

Explanation of the contribution of input features

We analyzed the feature contribution values of the prediction results of three models using "SHAP". We selected three mutants (C198G, E134M, F162P) that were successfully validated through experiments. We found that the "195" feature had the highest contribution value in the C198G result predicted by Model 1 (NADH010103-FFT-SVR), the "2212" feature had the highest contribution value in the E134M result predicted by Model 2 (PB-Ridge), and the "510" feature had the highest contribution value in the F162P result predicted by Model 3 (RACS820105-FFT- Bagging). We believe that different encoding methods result in different features that the model focuses on.

Model 1:NADH010103-FFT-SVR（C198G）

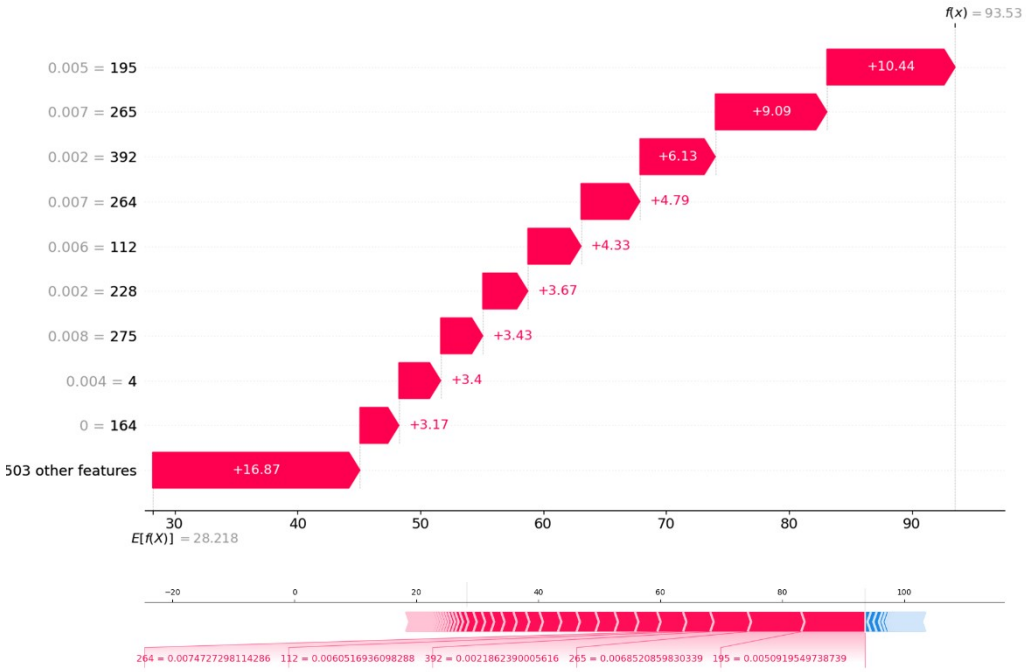

Figure S2 The marginal contribution of features to the model output (C198G).

Model 2: PB-Ridge(E134M)

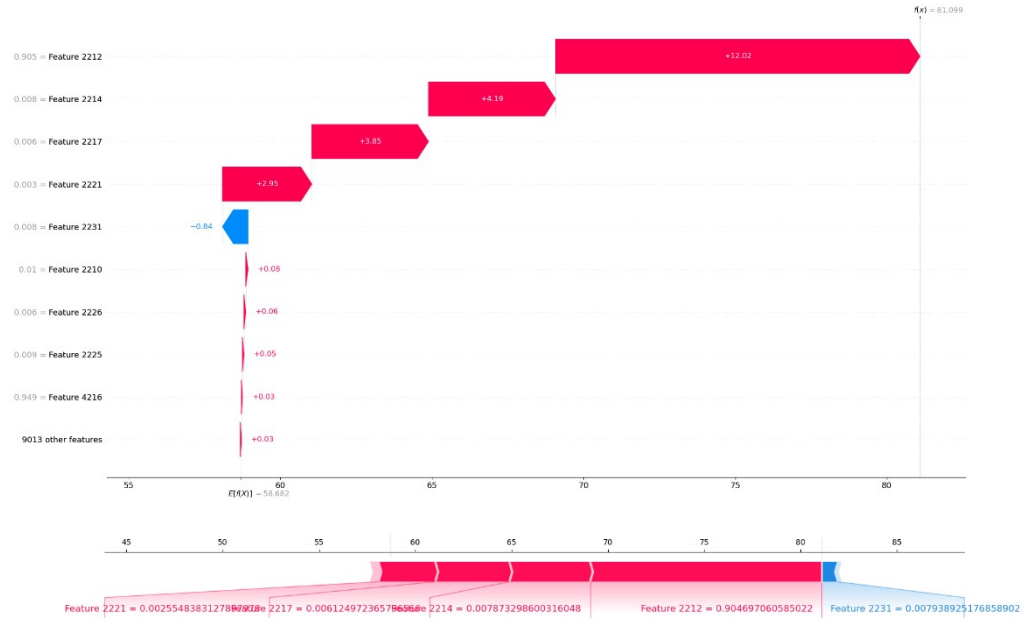

Figure S3 The marginal contribution of features to the model output (E134M).

Model 3: RACS820105-FFT- Bagging(F162P)

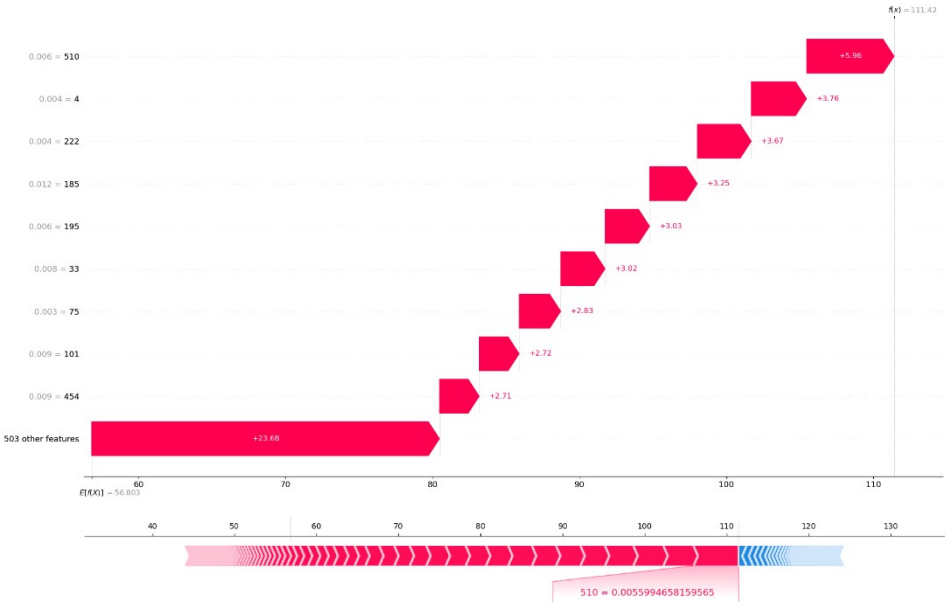

Figure S4 The marginal contribution of features to the model output (F162P).