Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures

Yidong Song[1,6], Qianmu Yuan[1,2,6], Sheng Chen[1], Yuansong Zeng[3], Huiying Zhao[4], Yuedong Yang[1,5*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, 510000, China.

[2]High Performance Computing Department, National Supercomputing Center in Shenzhen, Shenzhen, Guangdong, 518000, China.

[3]School of Big Data & Software Engineering, Chongqing University, Chongqing 400000, China.

[4]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, Guangdong, 510000, China.

[5]Key Laboratory of Machine Intelligence and Advanced Computing (MOE), Guangzhou 510000, China.

[6]These two authors contributed equally.

[*]Corresponding author.

E-mail addresses: yangyd25@mail.sysu.edu.cn (Yuedong Yang)

The Supplementary file includes:

Evaluation metrics

Figs. S1 to S13

Tables S1 to S8

Evaluation metrics

Consistent with prior research, we utilized the following metrics to evaluate the predictive performance: area under the precision-recall curve (AUPR), area under the receiver operating characteristic curve (AUC), Matthews correlation coefficient (MCC), precision, recall, and F1-score (F1).

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

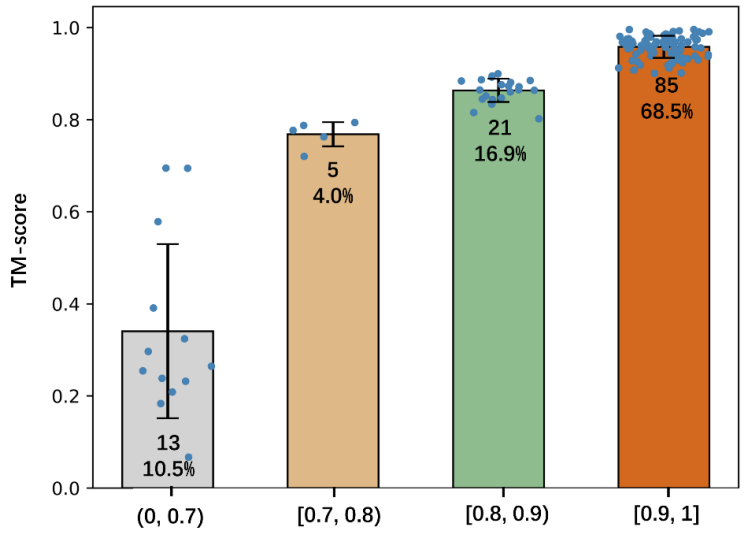

Fig. S1. The quality analysis of the predicted structures using TM-score on TS124. More than 85% of enzymes achieve TM-scores greater than 0.8, indicating the high quality of the ESMFold-predicted structures. Source data are provided as a Source Data file.

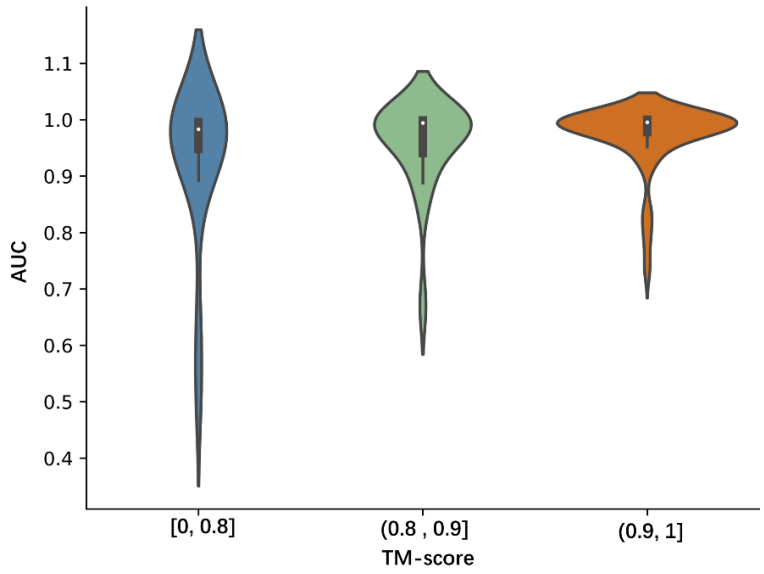

Fig. S2. The AUC of GraphEC-AS was changed by the TM-score.

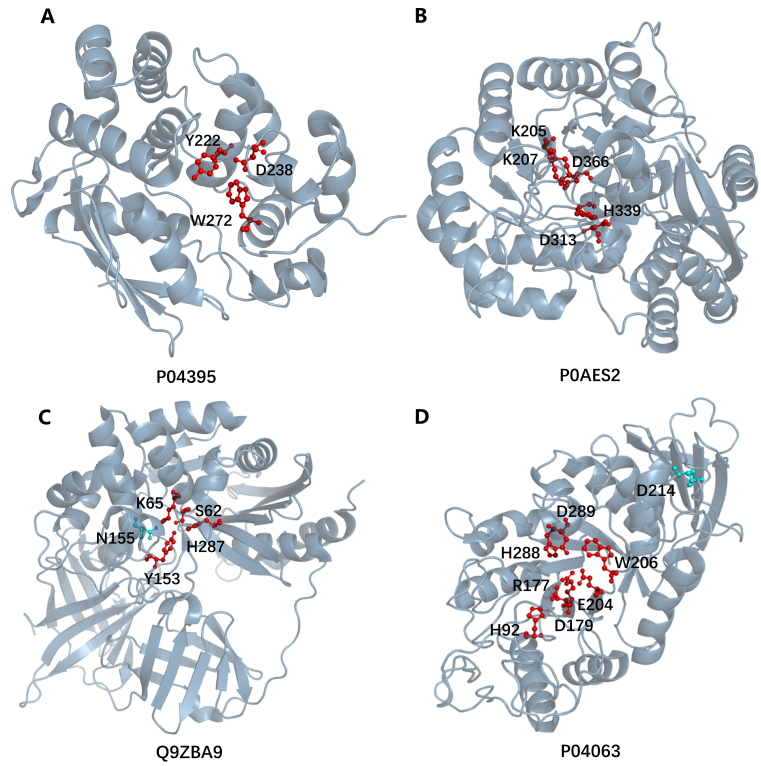

Fig. S3. The three-dimensional structures of several cases. Four enzymes (P04395, P0AES2, Q9ZBA9, and P04063) were visualized in their three-dimensional structures. True positives and false negatives are represented by red and cyan, respectively.

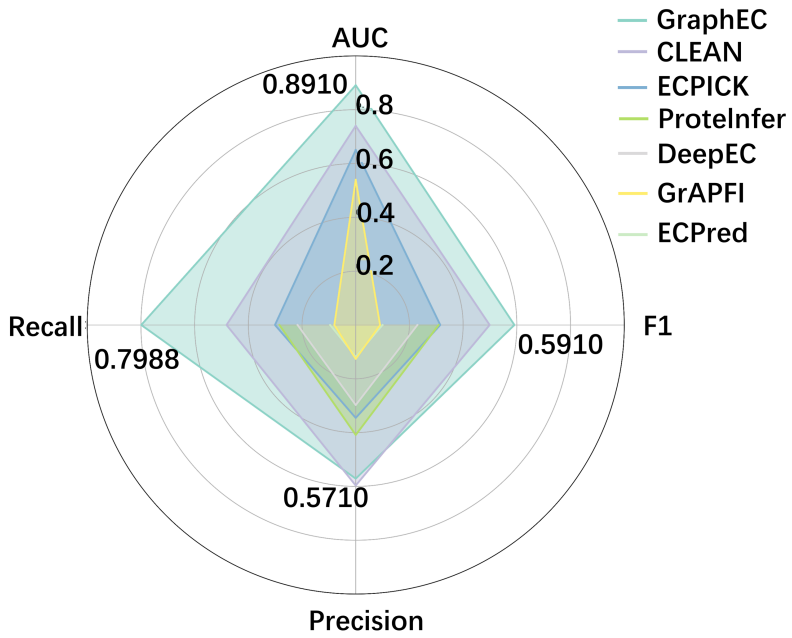

Fig. S4. The comparison between GraphEC and several state-of-the-art methods using AUC, recall, precision, and F1 on New-392. Source data are provided as a Source Data file.

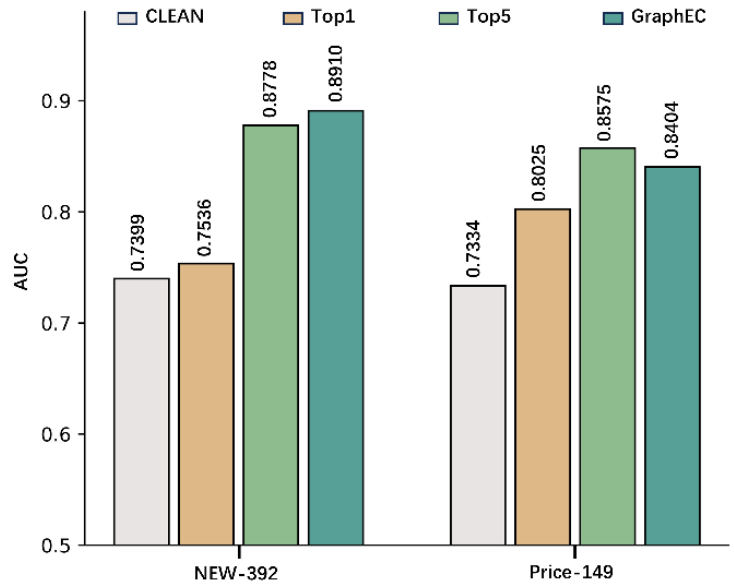

Fig. S5. Comparing the AUC of GraphEC with different methods, where Top1 and Top5 represent methods utilizing the top 1 and top 5 predictions from GraphEC as annotations, respectively. Source data are provided as a Source Data file.

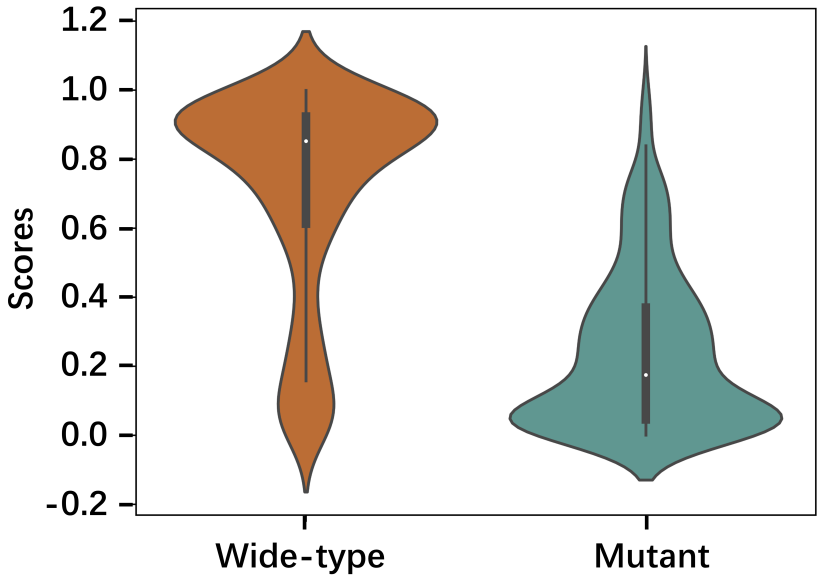

Fig. S6. The scores predicted by our model for the true EC numbers before (left) and after (right) the mutation.

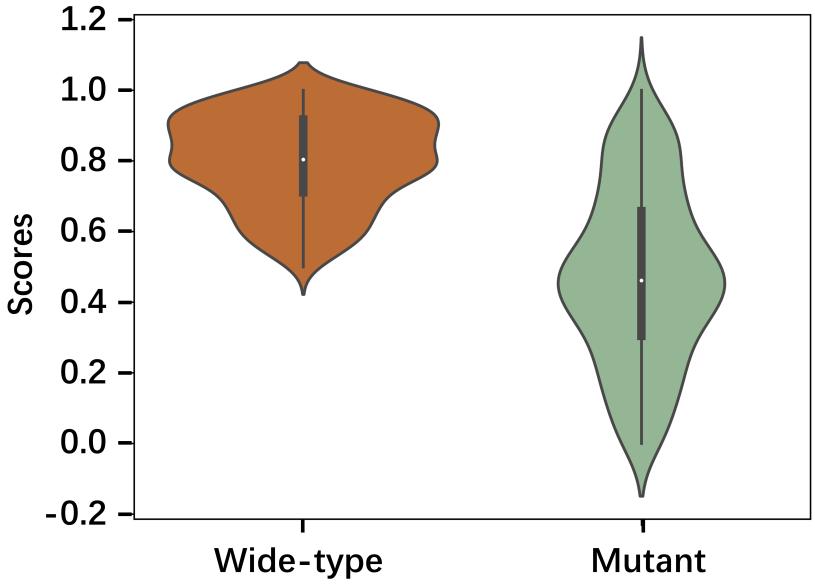

Fig. S7. The scores predicted by the model at active sites before (left) and after (right) mutation.

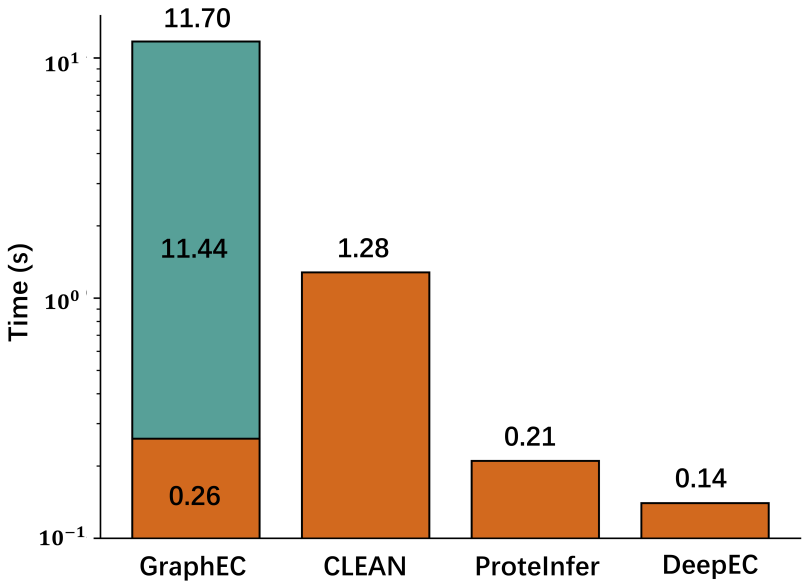

Fig. S8. Comparison of computational time of different methods. Green denotes the time taken by ESMFold to compute protein structures, while orange represents the inference time of models. Source data are provided as a Source Data file.

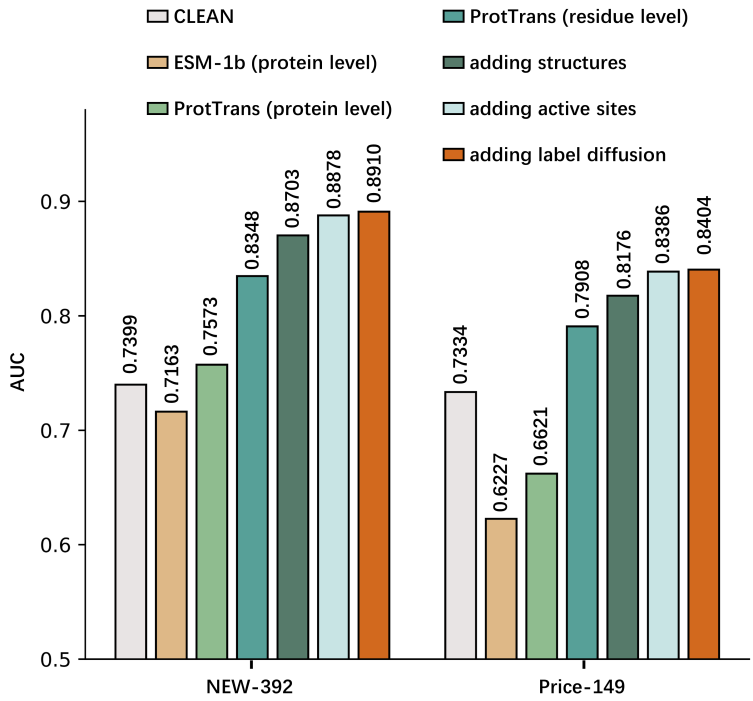

Fig. S9. The comparison of model performance using various representations and modules. ESM-1b (protein level), ProtTrans (protein level), and ProtTrans (residue level) represent the MLP models using different levels of representations, where the protein level representations indicate the mean value of residue level vectors. Adding structures represents the models utilizing geometric graph learning based on ESMFold-predicted structures. Adding active sites and Adding label diffusion refers to the sequential integration of active site guidance and label diffusion. Source data are provided as a Source Data file.

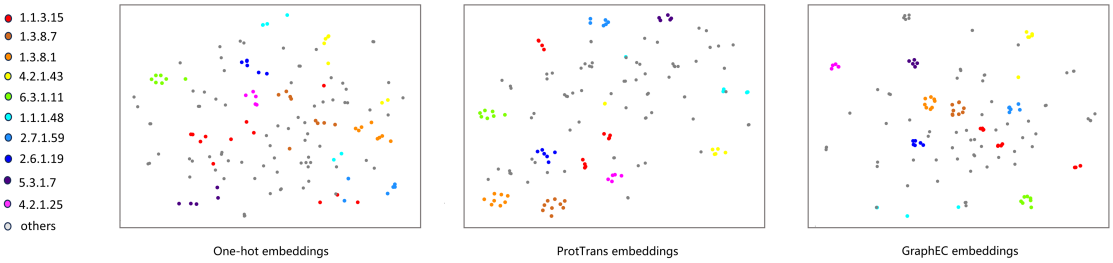

Fig. S10. Three embeddings were visualized on Price-149, with GraphEC embeddings representing the geometric embeddings learned by GraphEC, and one-hot embeddings and ProtTrans embeddings representing the one-hot vector and ProtTrans vector, respectively.

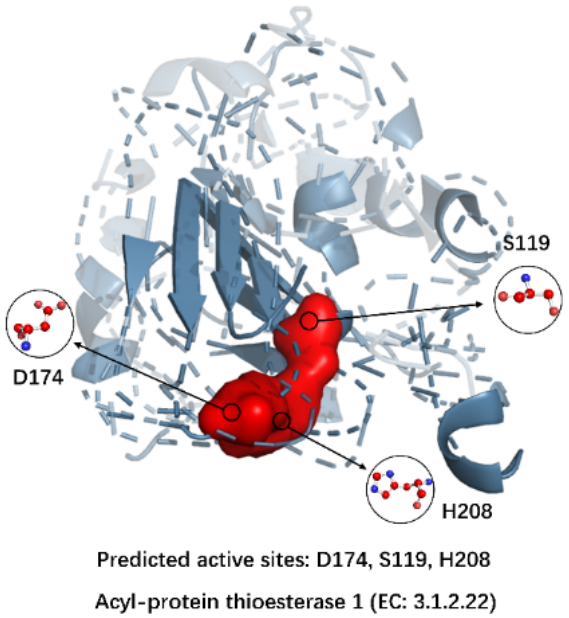

Fig. S11. The three-dimensional structure of O75608 was visualized, with the red portion representing the true active sites and the darker blue portion indicating higher attention scores.

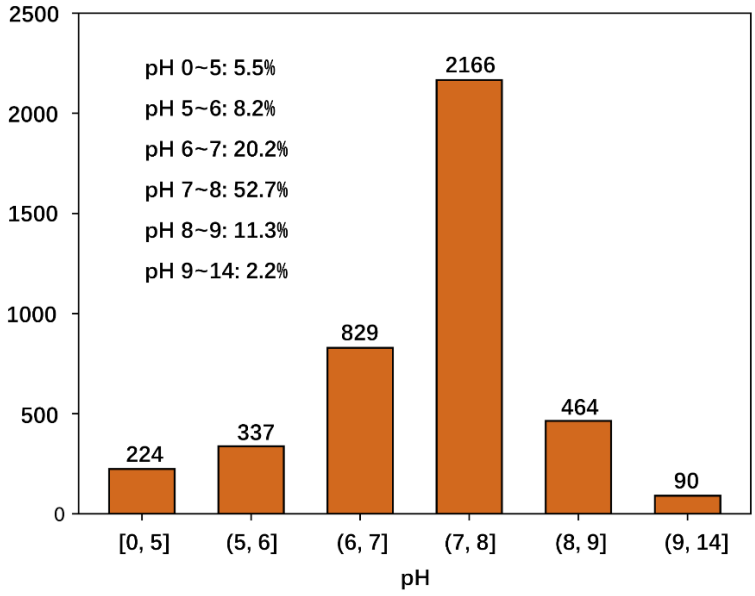

Fig. S12. The distribution of optimum pH on the dataset derived from BRENDA. 28.4% and 64.0% of enzymes exhibit an optimum pH range of 5 to 7 and 7 to 9, respectively.

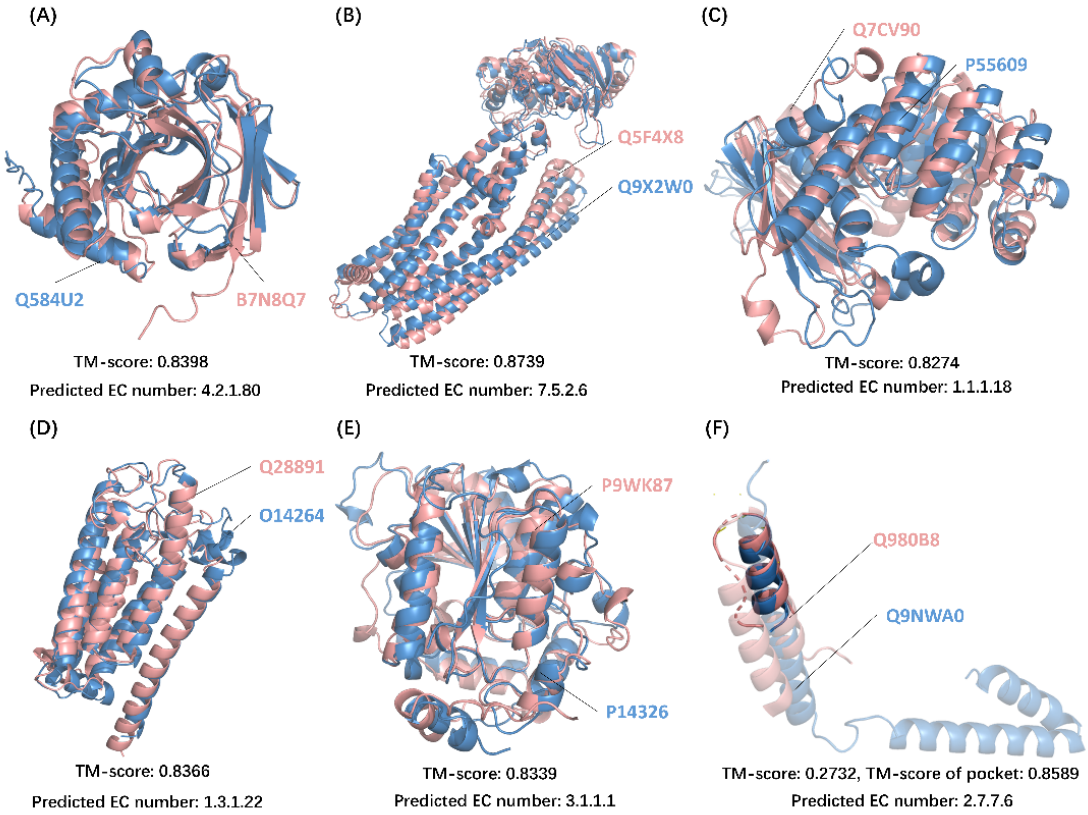

Fig. S13. The alignment of ESMFold-predicted structures with low sequence similarity, with the pink-orange portion representing enzymes in the training set and the blue portion representing the proteins from Swiss-Prot with less than 25% identity to the training set. The highlighted area in F represents the enzyme pocket.

Table S1. The Recall, Precision, MCC, and F1 of GraphEC-AS and comparison methods on TS124.

Methods	Recall	Precision	MCC	F1
HA (combination filter)	0.2401	0.1413	0.1757	0.1779
HA (residue identity filter)	0.2770	0.1606	0.2026	0.2033
Crpred (all residues)	0.5013	0.1438	0.2586	0.2235
CRpred (residues with coordinates)	0.5013	0.1469	0.2613	0.2272
PREvaIL_LR	0.4973	0.1406	0.2539	0.2192
PREvaIL_RF	0.6223	0.1487	0.2939	0.2400
GraphEC-AS	**0.7282**	**0.3467**	**0.4973**	**0.4698**

Table S2. The performance comparison of GraphEC, CLEAN, ECPICK, ProteInfer, DeepEC, GrAPFI, and ECPred on NEW-392 and Price-149.

Datasets	Methods	AUC	F1	Precision	Recall
NEW-392	ECPred	—	0.1000	0.1178	0.0954
	GrAPFI	0.5407	0.0927	0.1316	0.0817

	DeepEC	—	0.2297	0.2976	0.2167
	ProteInfer	—	0.3086	0.4088	0.2843
	ECPICK	0.6502	0.3148	0.3453	0.3008
	CLEAN	0.7399	0.4988	**0.5965**	0.4811
	GraphEC	**0.8910**	**0.5910**	0.5710	**0.7988**
Price-149	ECPred	—	0.0197	0.0197	0.0197
	GrAPFI	0.5095	0.0197	0.0197	0.0197
	DeepEC	—	0.0846	0.1184	0.0724
	ProteInfer	—	0.1662	0.2434	0.1382
	ECPICK	0.5888	0.1971	0.2763	0.1776
	CLEAN	0.7334	0.4947	0.5844	0.4671
	GraphEC	**0.8404**	**0.6131**	**0.6132**	**0.6908**

Table S3. The number of predictable EC numbers for different methods.

Methods	Number of predictable EC numbers	Release time
GraphEC, CLEAN	5106	2023
DeepEC	4669	2019
ECPred	858	2018

Table S4. The top 10 enzymes selected based on the degree of score decrease after mutation.

Uniprot ID	Predicted EC number (wide-type)	Predicted EC number (Mutant)	True EC number
Q59VG6	1.14.19.20	non-enzymes	1.14.19.20
A0A011QK89	1.1.99.2	non-enzymes	1.1.99.2
B4YA15	2.5.1.10, 2.5.1.1, 2.5.1.29	non-enzymes	2.5.1.10, 2.5.1.1
B1VB71	2.3.1.222	non-enzymes	2.3.1.222
Q0E2F9	3.4.19.12	non-enzymes	3.4.19.12
Q53FA7	1.6.5.5	non-enzymes	1.6.5.5
B1XJV9	2.5.1.29, 2.5.1.1, 2.5.1.10, 2.5.1.81	non-enzymes	2.5.1.29, 2.5.1.1, 2.5.1.10
WP_011717048	5.1.3.3	non-enzymes	5.1.3.3
WP_014879336	4.1.2.29, 4.1.2.13, 4.1.2.40	non-enzymes	4.1.2.29
WP_014235627	5.1.99.1, 4.4.1.5	non-enzymes	5.1.99.1

Table S5. The analysis of model performance in relation to the inclusion of physical properties.

	Methods	Recall	Precision	F1	AUC
Price-149	GraphEC	**0.6908**	0.6132	**0.6131**	**0.8404**

	GraphEC (add physical properties)	0.6776	**0.6228**	0.6084	0.8343
NEW-392	GraphEC	0.7988	**0.5710**	0.5910	**0.8910**
	GraphEC (add physical properties)	**0.8008**	0.5678	**0.6038**	0.8899

Table S6. Comparison of model performance when using AlphaFold2-predicted and ESMFold-predicted structures.

Datasets	Methods	Recall	Precision	F1	AUC
Price-149	GraphEC (AlphaFold2)	**0.6908**	0.6129	0.6127	**0.8406**
	GraphEC (ESMFold)	0.6908	**0.6132**	**0.6131**	0.8404
NEW-392	GraphEC (AlphaFold2)	**0.8267**	**0.5745**	**0.6044**	**0.9004**
	GraphEC (ESMFold)	0.7988	0.5710	0.5910	0.8910

Table S7. Comparison of model performance using graphs with different radii.

	Methods	Recall	Precision	F1	AUC
Price-149	GraphEC (radius = 8 Å)	0.6382	0.6123	0.5808	0.8150
	GraphEC (radius = 10 Å)	**0.6908**	**0.6132**	**0.6131**	**0.8404**
	GraphEC (radius = 12 Å)	0.6776	0.6119	0.6076	0.8344
	GraphEC (radius = 14 Å)	0.6776	0.6012	0.5962	0.8346
NEW-392	GraphEC (radius = 8 Å)	0.7729	0.5577	0.5459	0.8761
	GraphEC (radius = 10 Å)	**0.7988**	**0.5710**	**0.5910**	**0.8910**
	GraphEC (radius = 12 Å)	0.7928	0.5604	0.5481	0.8876
	GraphEC (radius = 14 Å)	0.7689	0.5637	0.5529	0.8753

Table S8. The homologous enzymes (acidophilic and non-acidophilic enzymes) correctly identified by GraphEC-pH. The E-value is calculated through DIAMOND.

Acidophilic enzymes	**Non-acidophilic enzymes**	**E-value**
A0A286JZ72	A5A699	5.89e-48
A0A286JZ72	K7ZLW6	1.43e-39
A0A286JZ72	Q7PC53	2.51e-39
A0A286JZ72	E5FGC5	2.97e-28
I0B0T0	A9ZND1	5.01e-07
D0QF43	A0A0U5B5Z4	1.77e-19
D0QF43	A0A075EEX4	3.15e-18
D0QF43	A1XM14	7.56e-08
D0QF43	A4XIF7	9.64e-33

A0A0C5H8Q7	C7NVN0	1.43e-08
I0B0T0	A3DD67	5.87e-72
D0QF43	A0A0F6MV31	1.5e-45
D0QF43	B9VSZ3	7.95e-41
E2GDF4	Q1KLC8	4.21e-17

108
109