

**Metabolic engineering and deep learning-driven protein
engineering for N-Acetylneuraminic acid biosynthesis in
*Escherichia coli***

Nan-Kai Wang^{1,2,3}, Song Yue^{1,2,3}, Jin-Ping Chen^{1,3}, Chang Su^{1,3}, Zhen-Ming Lu², Jin-Song Gong^{1,3,*}, Wei E. Huang⁴, Zheng-Hong Xu^{2,3,5}, Jin-Song Shi^{1,3,*}

¹Key Laboratory of Carbohydrate Chemistry and Biotechnology of Ministry of Education, School of Life Sciences and Health Engineering, Jiangnan University, Wuxi 214122, PR China;

²National Engineering Research Center for Cereal Fermentation and Food Biomanufacturing, School of Biotechnology, Jiangnan University, Wuxi 214122, PR China;

³Institute of Future Food Technology, JITRI, Yixing 214200, PR China;

⁴Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK.

⁵College of Biomass Science and Engineering, Sichuan University, Chengdu 610065, PR China;

*Corresponding author:

Lihu Avenue No. 1800, Jiangnan University, Wuxi 214122, PR China

E-mail: shijs@163.com (J.S. Shi)

jinsonggong.bio@hotmail.com (J.S. Gong)

26 **Supplementary Table 1.** Strains used in this study for optimizing RBS intensity.

Strains	RBS-Mgglms	RBS-Cegna1	RBS-AnaAGE	RBS-NmNeuB
G15	MgWT	CeWT	AnWT	NmWT
R1	RM1	CeWT	AnWT	NmWT
R2	RM2	CeWT	AnWT	NmWT
R3	RM3	CeWT	AnWT	NmWT
R4	RM4	CeWT	AnWT	NmWT
R5	RM5	CeWT	AnWT	NmWT
R6	RM3	RC1	AnWT	NmWT
R7	RM3	RC2	AnWT	NmWT
R8	RM3	RC3	AnWT	NmWT
R9	RM3	RC4	AnWT	NmWT
R10	RM3	RC5	AnWT	NmWT
R11	RM3	RC4	RA1	NmWT
R12	RM3	RC4	RA2	NmWT
R13	RM3	RC4	RA3	NmWT
R14	RM3	RC4	RA4	NmWT
R15	RM3	RC4	RA5	NmWT
R16	RM3	RC4	RA5	RN1
R17	RM3	RC4	RA5	RN2
R18	RM3	RC4	RA5	RN3
R19	RM3	RC4	RA5	RN4
R20	RM3	RC4	RA5	RN5

27

28

29 **Supplementary Table 2.** The predicted sites of AGE^{Bf} by PROSS.

V402L	T295I	S84A	S341A	C320H
N195P	E86Q	E53K	S296M	N318K
N301E	K372P	I80M	I251E	Y312F
Q263L	Q109H	K293V	K178R	F407Y
Y378F	L189Q	K40R	H324Q	G345A
F74Y	S91A	L196M	E70D	C106A
Y117H	T367S	N46D	G111A	D45N
T298W	S132Q	S128A	Q162E	N161D

30
31
32
33

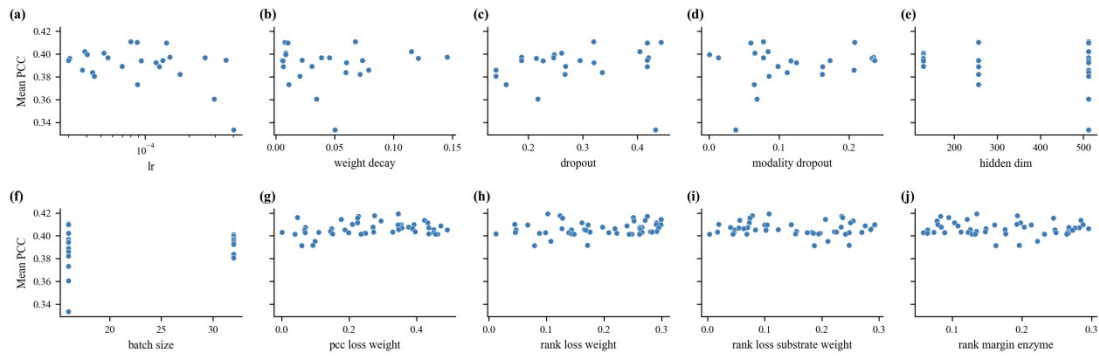
34 **Supplementary Table 3.** The predicted sites of NeuB^{Nm} by DLcatalysis.

Ranking	Mutant	Value
1	I343K	1699.82
2	H57K	714.23
3	I343R	609.63
3212	G281A	278.10
3213	D276T	278.10
3214	P284T	278.09
6611	H215L	44.89
6612	H215I	44.31
6613	H215Y	43.40

35

36

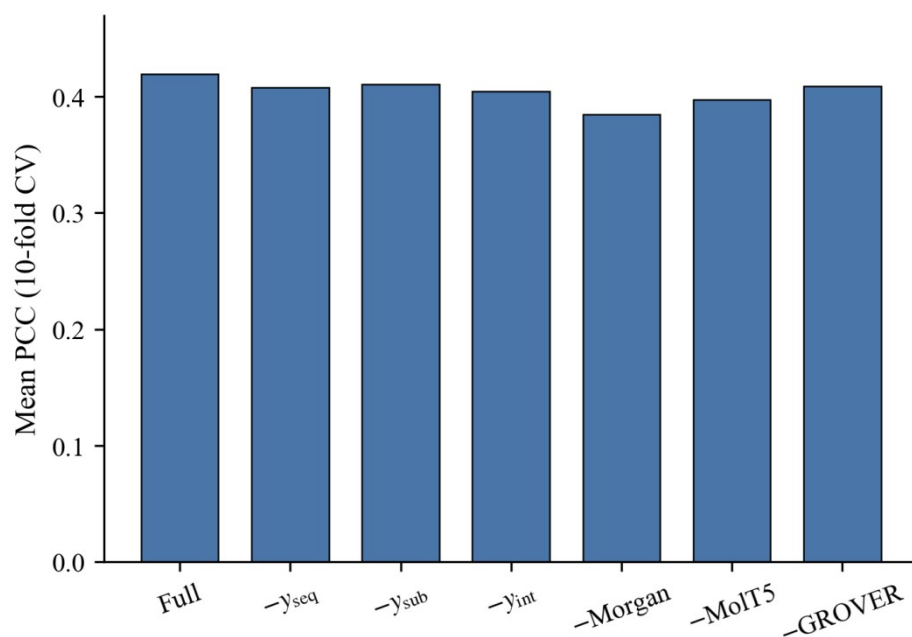
37



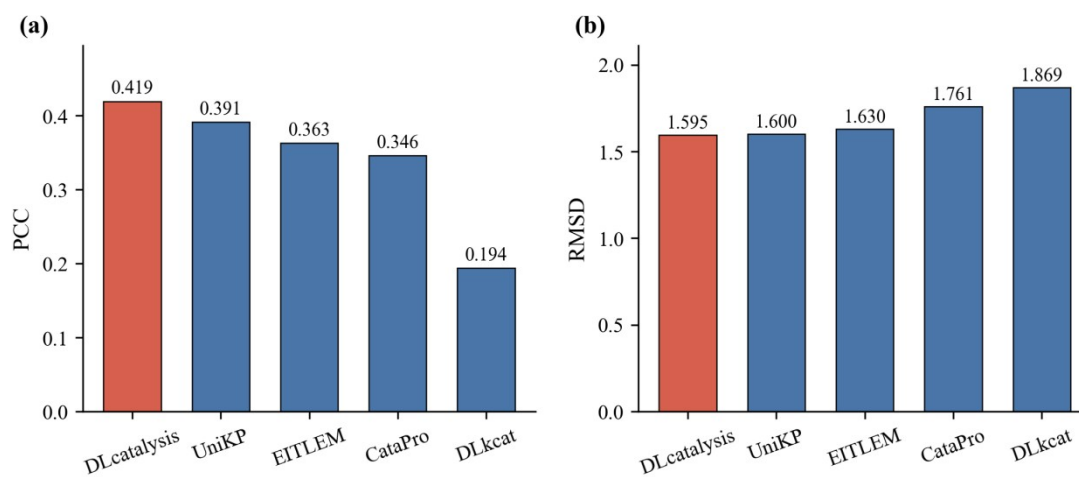
38

39 **Supplementary Fig. 1.** Scatter plots showing the relationship between individual
40 hyperparameters and mean Pearson correlation coefficient (PCC) during
41 hyperparameter optimization. Each point represents one tested configuration. The
42 evaluated hyperparameters include (a) learning rate, (b) weight decay, (c) dropout rate,
43 (d) modality dropout rate, (e) hidden dimension, (f) batch size, (g) PCC loss weight, (h)
44 ranking loss weight, (i) substrate-wise ranking loss weight, and (j) enzyme-wise
45 ranking loss weight. Source data are provided as a Source Data file.

46



Supplementary Fig. 2. Component ablation of DLcatalysis.(a) Mean PCC from ten-fold cross-validation for the full DLcatalysis model and ablated variants. Error bars indicate variability across folds. Source data are provided as a Source Data file.

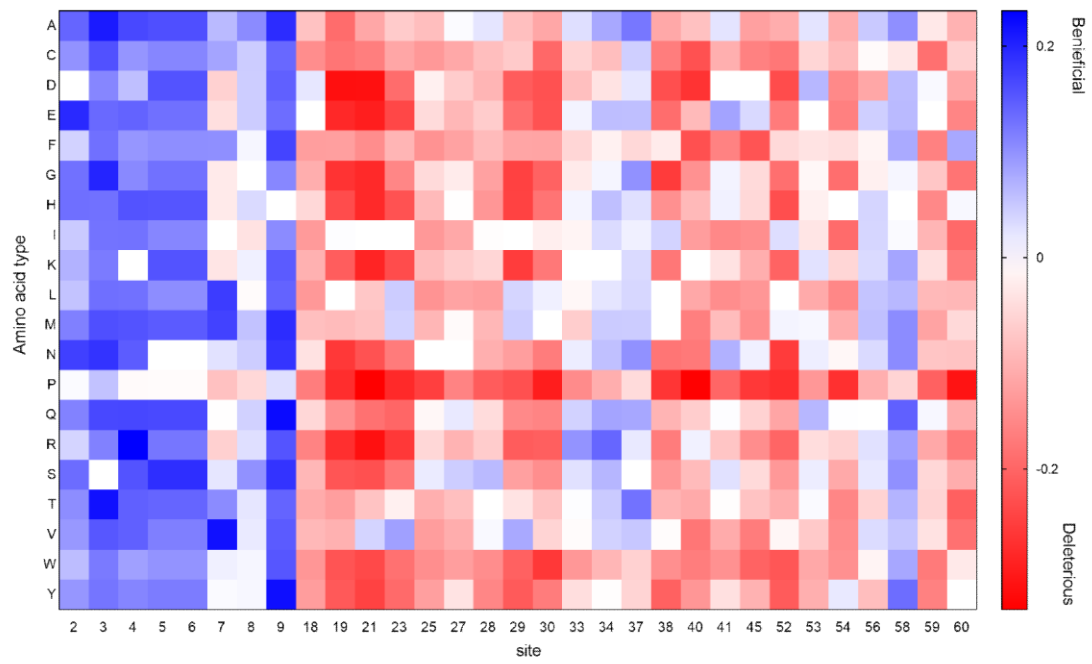


Supplementary Fig. 3. Comparison of DLcatalysis with existing k_{cat} K_{m}^{-1} predictors.

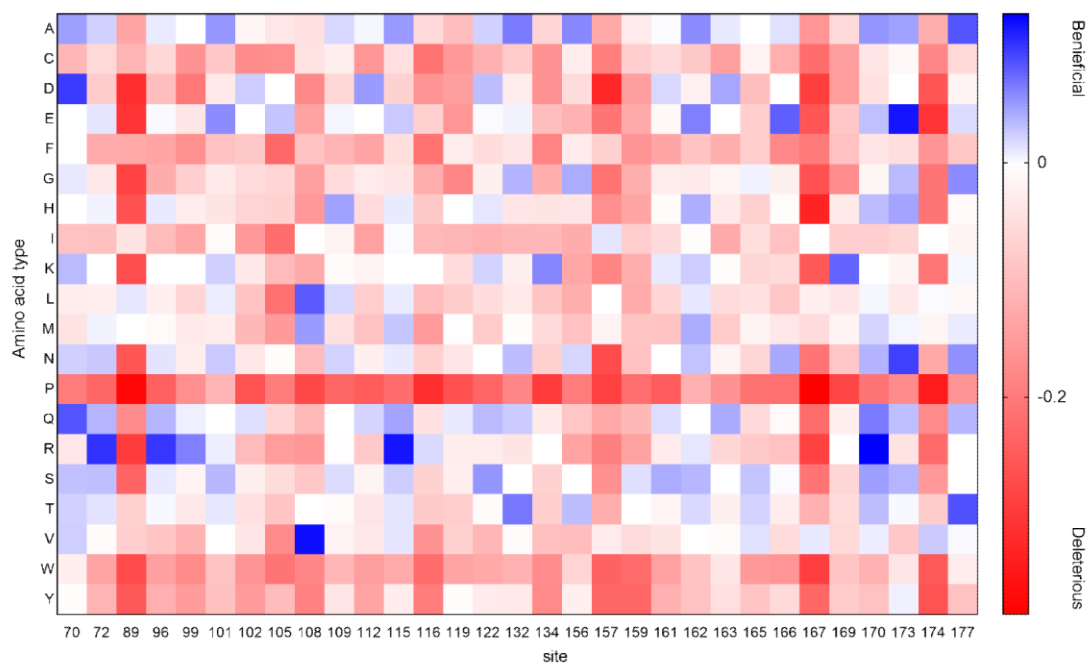
(a) PCC and (b) RMSD of DLcatalysis, UniKP, EITLEM, CataPro and DLkcat

evaluated under the same ten-fold cross-validation protocol. Source data are provided

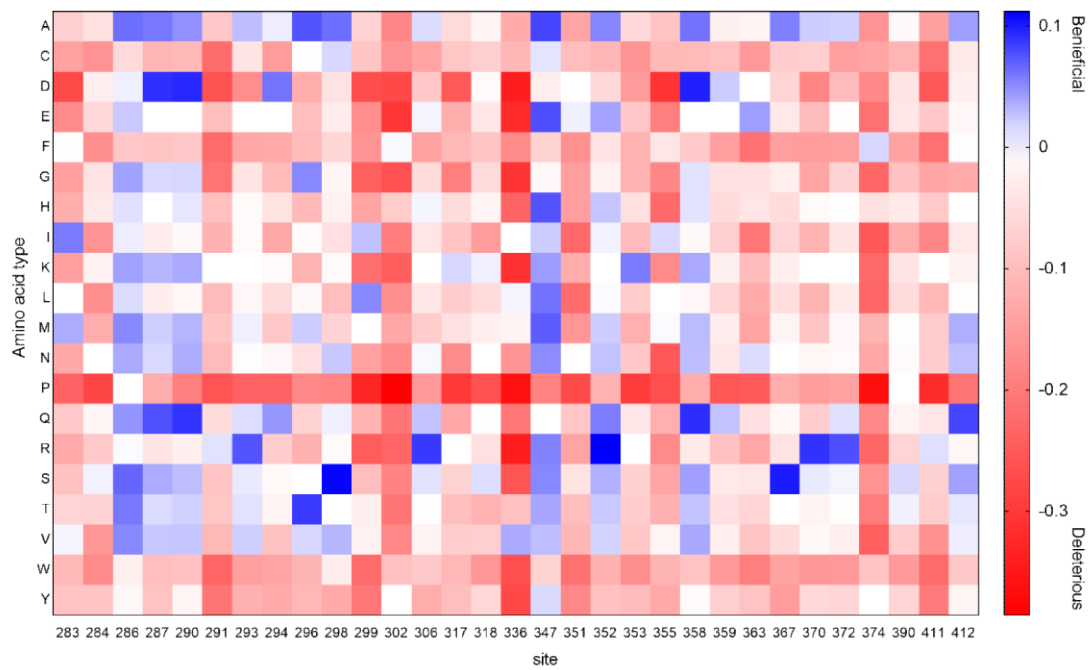
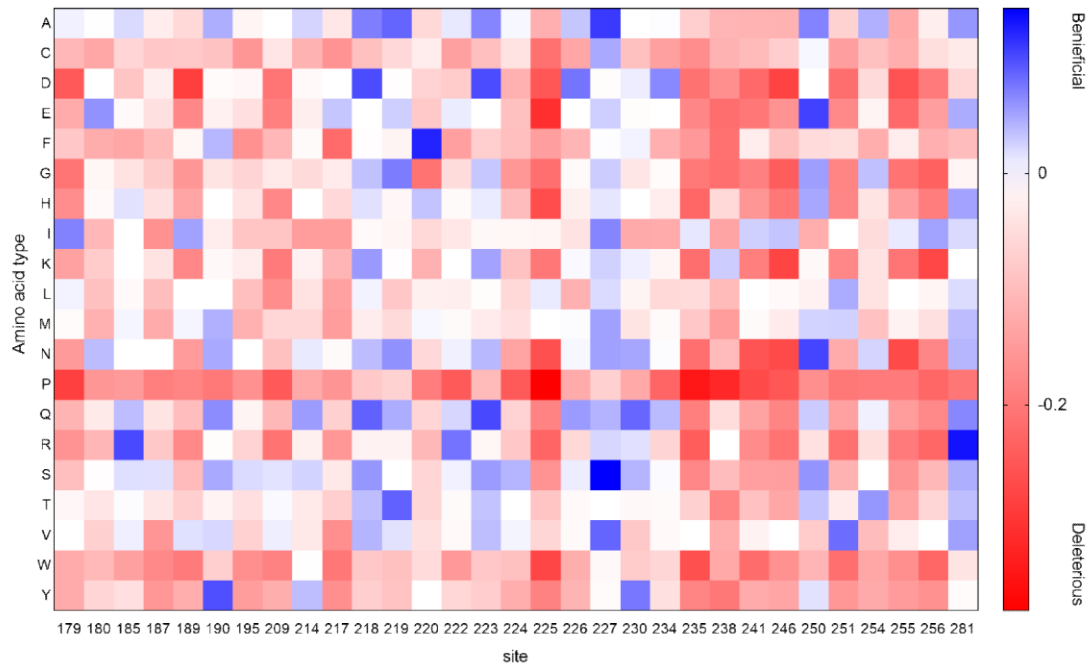
as a Source Data file.



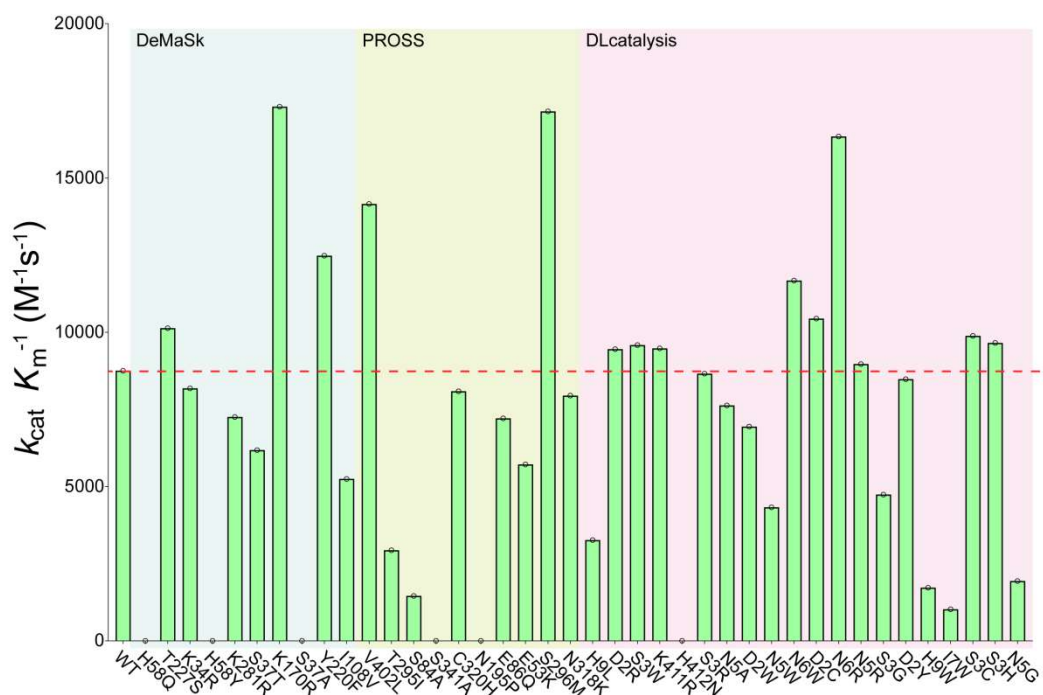
58



59

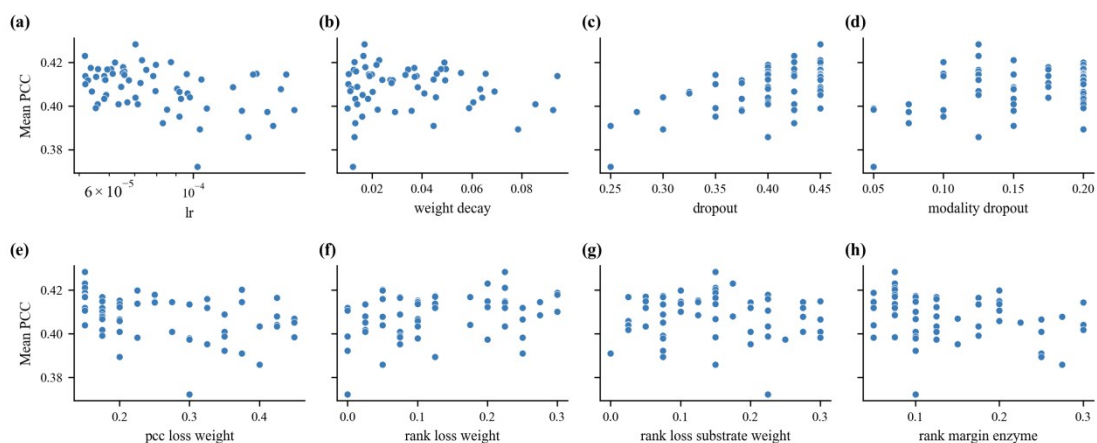


Supplementary Fig. 4. The scores of predicted mutants by DeMaSK. Source data are provided as a Source Data file.

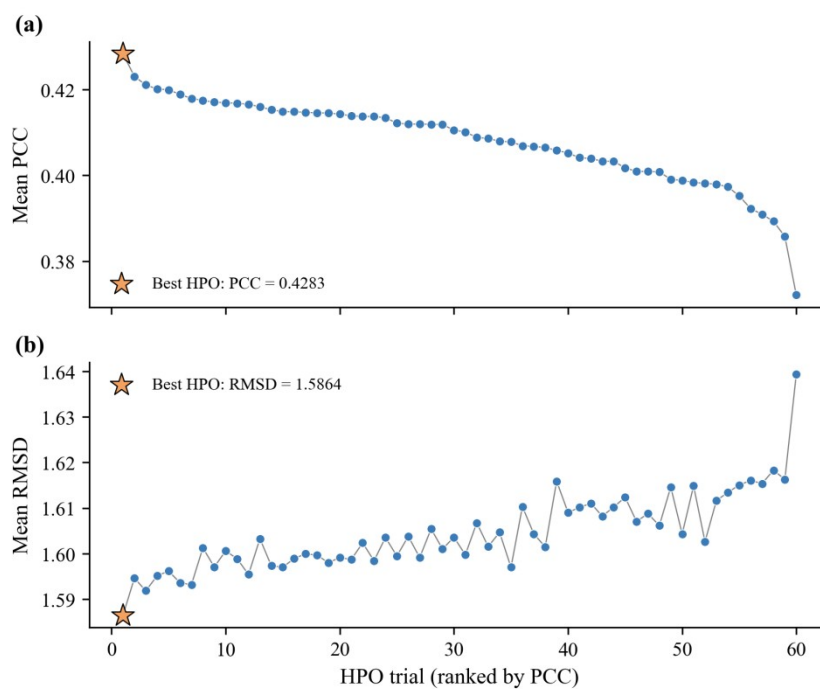


Supplementary Fig. 5. SEED-guided active learning for AGE^{Bf} variant prioritization.

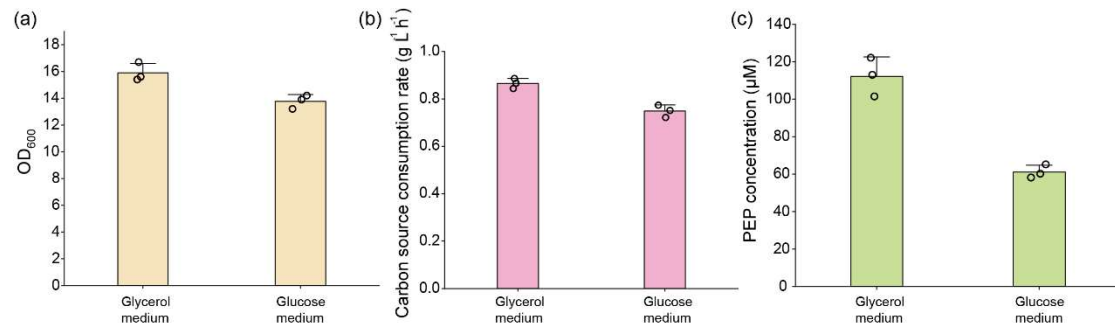
The catalytic efficiencies of AGE^{Bf} variants obtained through the SEED-based active-learning workflow. Variants in the blue-shaded region represent the top ten candidates predicted by DeMaSk, variants in the yellow-shaded region represent the top ten candidates predicted by PROSS, and variants in the pink-shaded region represent the top ten candidates predicted by DLcatalysis after active learning. The red dashed line indicates the $k_{cat} K_m^{-1}$ value of AGE^{Bf}. Source data are provided as a Source Data file.



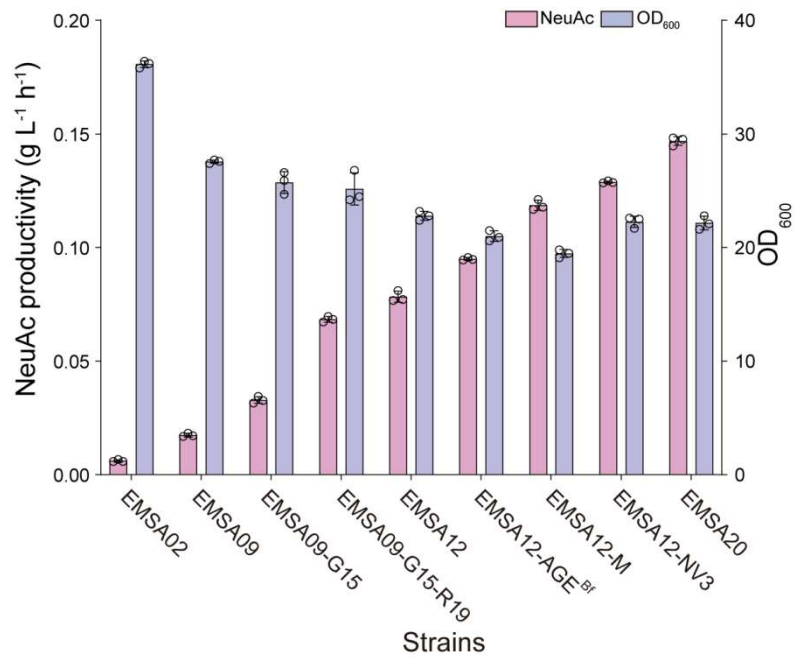
Supplementary Fig. 6. Hyperparameter optimization of DLcatalysis after active learning. Mean PCC values obtained under different hyperparameter settings during optimization. Panels show the effects of (a) learning rate, (b) weight decay, (c) dropout, (d) modality dropout, (e) PCC loss weight, (f) ranking loss weight, (g) substrate-wise ranking loss weight, and (h) enzyme-wise ranking margin. Each point corresponds to one hyperparameter trial. Source data are provided as a Source Data file.



Supplementary Fig. 7. Hyperparameter optimization performance of DLcatalysis. (a) Mean PCC and (b) mean RMSD across hyperparameter optimization trials ranked by PCC. Each point represents one tested configuration. The selected configuration, HPO1, is marked with a star and achieved the highest mean PCC of 0.428 with a corresponding RMSD of 1.586. Source data are provided as a Source Data file.



Supplementary Fig. 8. Cultivation results of strain EMSA20 in glycerol- and glucose-based media. (a) The OD₆₀₀ of strain EMSA20 in different culture medium. (b) The carbon source consumption rate of strain EMSA20 in different culture medium. (c) The PEP concentration of strain EMSA20 in different culture medium. Data are presented as mean $\pm$ SD of $n = 3$ independent biological replicates. Source data are provided as a Source Data file.



Supplementary Fig. 9. The comparison of the productivity of each improved strain. Data are presented as mean $\pm$ SD of $n = 3$ independent biological replicates. Source data are provided as a Source Data file.