Supplementary Information for

DPFunc: Accurately predicting protein function via deep learning with domain-guided structure information

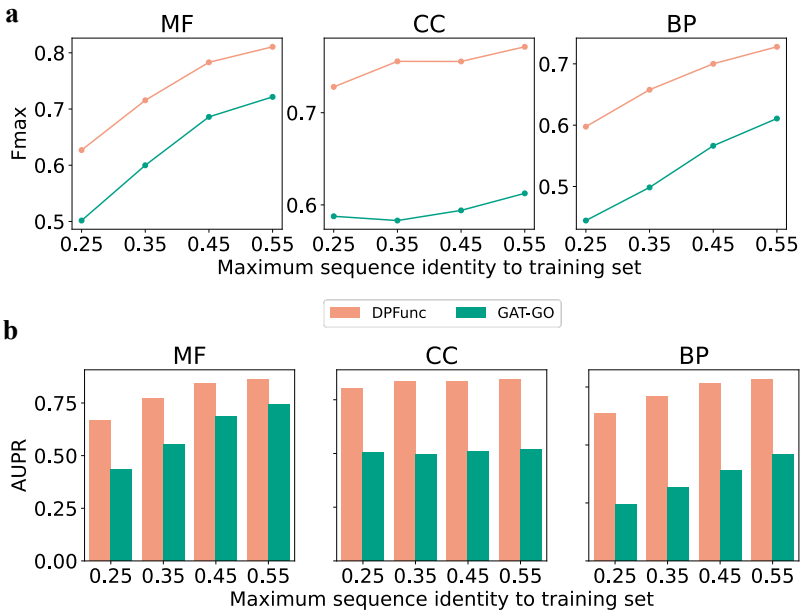
Wenkang Wang, Yunyan Shuai, Min Zeng, Wei Fan, Min Li

Correspondence to: limin@mail.csu.edu.cn

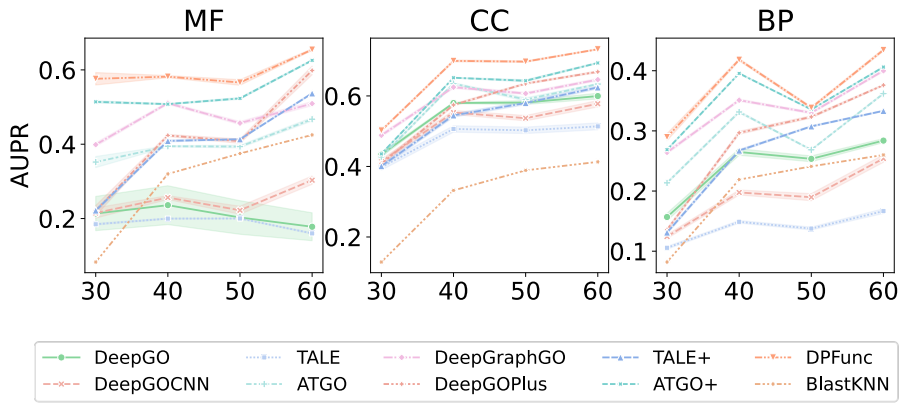
This PDF file includes:

Supplementary Figure S1 to S7.

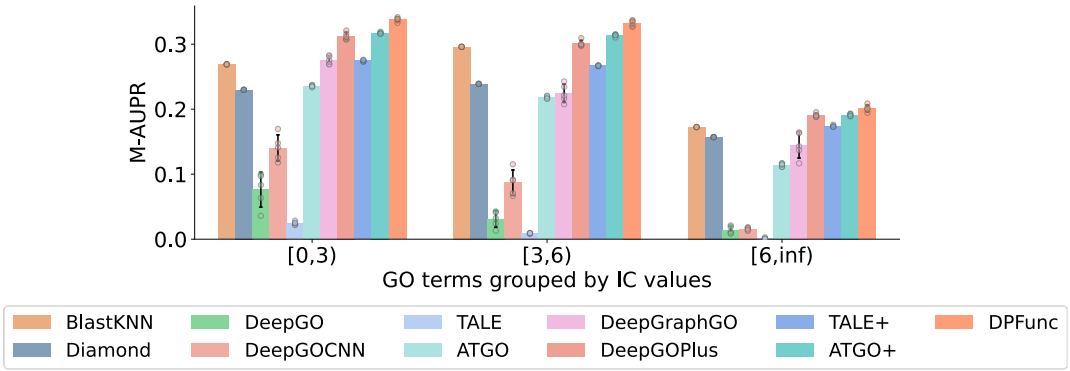
Supplementary Table S1 to S8.



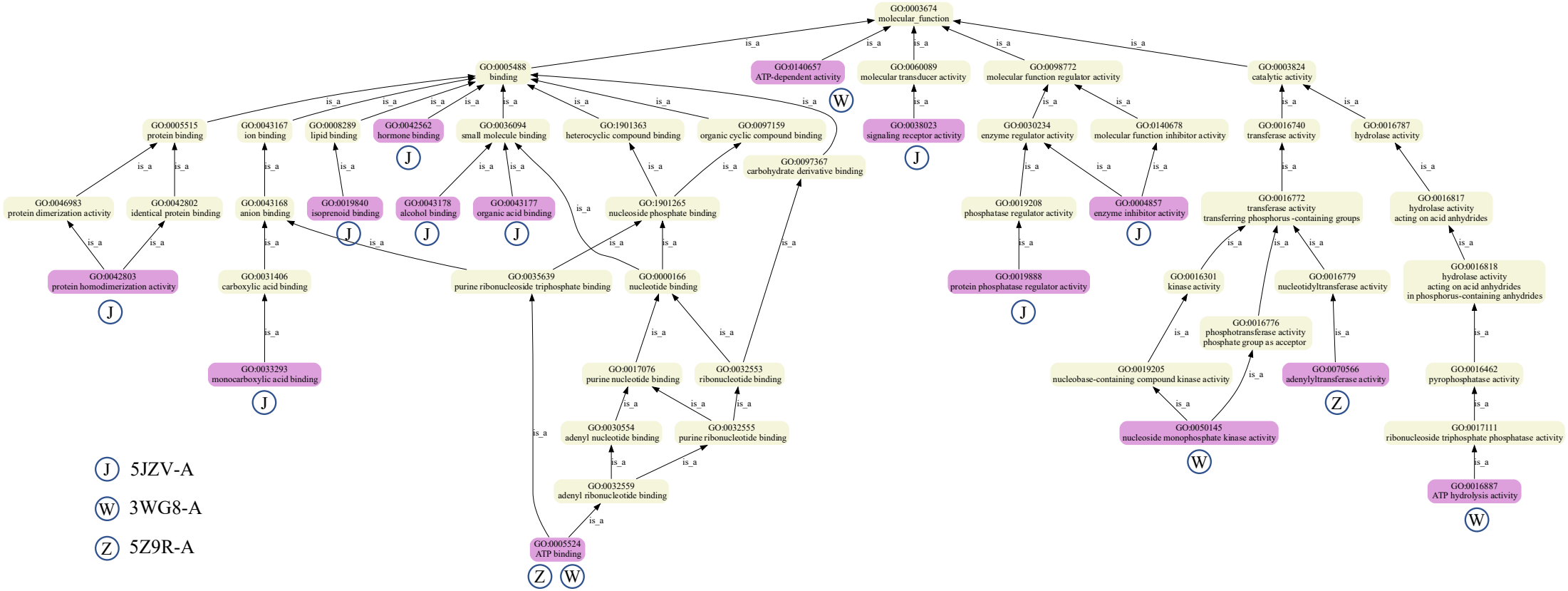
Supplementary Figure S1. The performance comparison of structure-based methods and DPFunc across different sequence identity cut-offs.



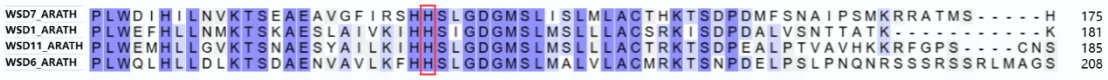
Supplementary Figure S2. The performance comparison of all SOTA methods and DPFunc on difficult protein sets with different sequence similarities to training proteins, where the data from five repeated experiments are presented as mean value +/- standard errors.



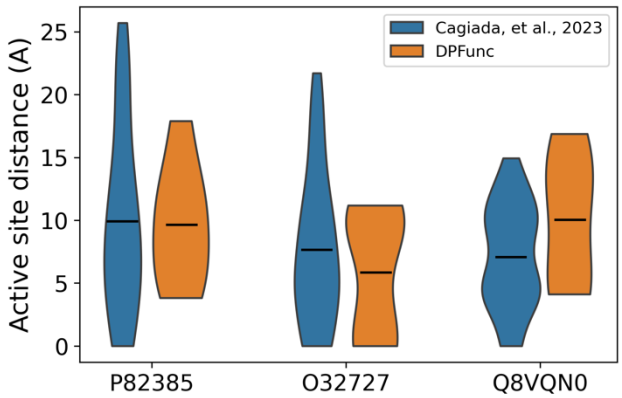
Supplementary Figure S3. The performance evaluation of all SOTA methods and DPFunc on rare GO terms with different IC values, where GO terms with higher IC values are more informative and valuable. The experiment is repeated five times for each method on the test data, reducing the effects from the random factor. The data are presented as mean value +/- standard deviation.



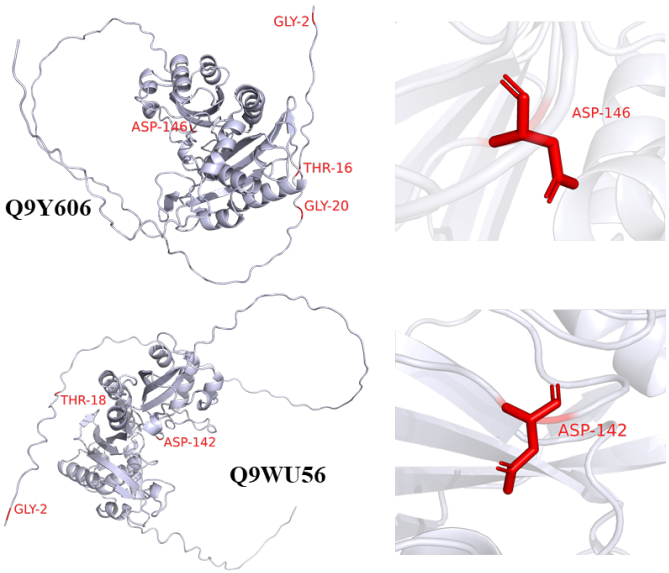
Supplementary Figure S4. The GO terms of three proteins (PDB ID: 5JZV-A, 3WG8-A, 5Z9R-A). The purple blocks are the deepest child GO terms, where "J", "W" and "Z" indicates 5JZV-A, 3WG8 and 5Z9R-A, respectively, and all of these GO terms are predicted by DPFunc accurately.



Supplementary Figure S5. The sequence alignment results of WSD1, WSD6, WSD7 and WSD11.



Supplementary Figure S6. The distribution of distances between predicted residues and the nearest known active sites. The central mark indicates the median.



Supplementary Figure S7. The detected residues by DPFunc and the validated active sites of two pseudouridylate synthases. The red positions shown in the structures are the key residues detected by DPFunc. The residues in the detailed graphs are the active sites that have been validated (ASP-142 for Q9Y606, ASP-146 for Q9WU56).

Supplementary Table S1. Predictive performance of DPFunc compared with web-servers on the large-scale dataset

	MF		CC		BP	
	Fmax	AUPR	Fmax	AUPR	Fmax	AUPR
COFACTOR $_{AF2}$	0.3767	0.2602	0.4401	0.2456	0.3009	0.1509
NetGO 3.0 $_{server}$	0.6308	0.5837	0.6363	0.6024	**0.4875**	0.4188
DPFunc $_{retrain}$	**0.6346**	**0.6537**	**0.6547**	**0.6887**	0.4534	**0.4290**

*$_{AF2}$ indicates that COFACTOR predict protein functions based on the structure predicted by AlphaFold2.

*$_{server}$ indicates that the results are from the web server of NetGO 3.0

*$_{retrain}$ indicates that our model is retrained on the same dataset with NetGO 3.0.

Best performance among all methods for each metric is shown in bold.

Supplementary Table S2. Predictive performance comparison on the large-scale dataset in terms of IC weighted AUPR

Methods	MF	CC	BP
Diamond	0.362	0.231	0.167
BlastKNN	0.459	0.309	0.219
DeepGO	0.216	0.386	0.187
DeepGOCNN	0.249	0.361	0.175
TALE	0.111	0.288	0.090
ATGO	0.391	0.424	0.267
DeepGraphGO	0.460	0.482	0.319
DeepGOPlus	0.512	0.468	0.305
TALE+	0.499	0.441	0.263
ATGO+	0.559	0.479	0.330
DPFunc	0.606	0.535	0.367

Supplementary Table S3. Coverage ratio of predicted functions by different methods to the known number of functions.

Methods	MF	CC	BP	Overall
Diamond	66.75%	60.18%	53.39%	56.17%
BlastKNN	75.95%	69.82%	61.25%	64.43%
DeepGO	50.20%	66.18%	29.99%	37.03%
DeepGOCNN	50.20%	66.55%	67.06%	64.45%
TALE	14.45%	27.27%	3.78%	7.98%
ATGO	69.25%	62.55%	54.93%	57.94%
DeepGraphGO	98.03%	99.27%	99.03%	98.90%
DeepGOPlus	83.05%	81.64%	78.98%	79.89%
TALE+	74.24%	66.00%	41.23%	48.96%
ATGO+	80.42%	73.09	62.36%	66.28%
DPFunc	100%	100%	100%	100%

Supplementary Table S4. Species associated with Bacillus subtilis (BACSU).

Target Species	Related Species
Bacillus subtilis (Taxon ID: 224308)	Salmonella choleraesuis (Taxon ID: 321314) Pseudomonas syringae pv. Tomato (Taxon ID: 223283) Salmonella typhimurium (Taxon ID: 99287) Escherichia coli (Taxon ID: 83333)

Supplementary Table S5. The process of detecting significant sites

Step	Description
	# First step – generate candidate residue set
1.	$Candidate\_set = set()$
2.	For W_{att-i} in $\{W_{att-1}, W_{att-2}, \dots, W_{att-n}\}$:
3.	Sort attention scores in descending order: $S = [(r_2, s_1), (r_{10}, s_2), \dots, (r_6, s_l)]$ # $s_1 > s_2 > \cdots > s_l$ *are the sorted attention scores and* $r_2, r_{10}, \dots, r_6$ *are the corresponding residues*
4.	Calculate gaps between neighbors: $G = [g_0, g_1, \dots, g_l]$ # $g_i = s_i - s_{i+1}$
5.	Calculate cut-off value: $cutoff = mean(G)$
	# Select residues from a high score to a low score
6.	For $index$ from 1 to l:
7.	If $g_{index} < cutoff$: $Candidate\_set.add(S_{index}[0])$
8.	Else: $Candidate\_set.add(S_{index}[0])$ and Break
	# Second step – filter residues with low confidence
9.	$Significant\_set = set()$
10.	For $residue$ in $Candidate\_set$:
11.	If $pLDDT(residue) \geq 50$: $Significant\_set.add(residue)$
12.	Return $Significant\_set$

Supplementary Table S6. The sequence identities between proteins calculated by BLAST.

	5JZV-A	3WG8-A	5Z9R-A
5JZV-A	100%	87.80%	89.74%
3WG8-A	87.80%	100%	90.24%
5Z9R-A	89.74%	90.24%	100%

Supplementary Table S7. The number of common functions between proteins.

	5JZV-A	3WG8-A	5Z9R-A
5JZV-A	31/31	5/31	20/31
3WG8-A	5/24	24/24	5/24
5Z9R-A	20/22	5/22	22/22

Supplementary Table S8. The statistical information of protein active sites.

Protein	PDB	Chain	Active Sites
P82385	1DO6	A	7
O32727	1L0O	A	16
Q8VQN0	1JC5	A	5