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

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

Aslı Gök

LifeLU Reading Club – 16.01.2025

Motivation

- Till now, less than 1% of protein sequences are annotated by Gene Ontology (GO) terms, which can be divided into three ontologies:

1. molecular functions (MF),
2. cellular components (CC),
3. and biological process (BP).

- Consequently, developing computational methods for automated protein function prediction is crucial for bridging the widening gap between the number of known annotations and protein sequences.

Overview of DPFunc

- DPFunc is a deep learning-based method for protein function prediction using domain-guided structure information.

- It consists of three modules:

1. **a residue-level feature learning module** based on a pre-trained protein language model and graph neural networks for propagating features between residues through protein structures
2. **a protein-level feature learning module** for extracting the whole structure features from residue-level features guided by domain information from sequences.
3. **a protein function prediction module** for annotating functions to proteins based on protein-level features.

Dataset

- **PDB dataset:** non-redundant set by clustering all PDB chains at 95% sequence identity
- The structures of proteins are obtained from the Protein Data Bank (PDB)

- **CAFA dataset :** collected from the UniProt and Gene Ontology database.
- Predicted structures from the AlphaFold database are used.

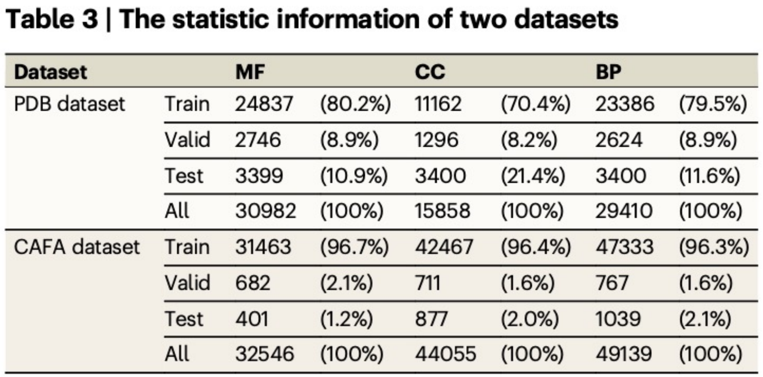

Table 3 | The statistic information of two datasets

Dataset		MF		CC		BP	
PDB dataset	Train	24837	(80.2%)	11162	(70.4%)	23386	(79.5%)
	Valid	2746	(8.9%)	1296	(8.2%)	2624	(8.9%)
	Test	3399	(10.9%)	3400	(21.4%)	3400	(11.6%)
	All	30982	(100%)	15858	(100%)	29410	(100%)
CAFA dataset	Train	31463	(96.7%)	42467	(96.4%)	47333	(96.3%)
	Valid	682	(2.1%)	711	(1.6%)	767	(1.6%)
	Test	401	(1.2%)	877	(2.0%)	1039	(2.1%)
	All	32546	(100%)	44055	(100%)	49139	(100%)

Methodology

- The core idea is to leverage domain information within protein sequences to guide the model toward learning the functional relevance of amino acids in their corresponding structures, highlighting structure regions that are closely associated with functions.

- DPFunc first extracts residue-level features from a pre-trained protein language model and then employs graph neural networks to propagate features between residues.

- Simultaneously, it scans the sequences and generates domains, converting them into dense representations through embedding layers. Inspired by the transformer architecture, DPFunc introduces an attention mechanism that learns whole structures and predicts functions under the guidance of corresponding domain information.

- To assess the importance of different residues, inspired by the transformer architecture, an attention mechanism is introduced to interweave the protein-level domain features and residue-level features, which detects the importance of each residue.

Residue-level feature learning module

- To learn residue-level features, DPFunc first constructs graphs based on protein structures. Specifically, for a target protein pi, its residues are considered as nodes, and two residues are connected if the distance of their corresponding Cα atoms is less than 10 Å.

Protein -level feature learning module

- DPFunc first scans the protein sequences and generates corresponding domain properties by InterProScan.

Function prediction, postprocessing procedure and significant residues detection

- Finally, DPFunc integrates these two modules and predicts protein functions

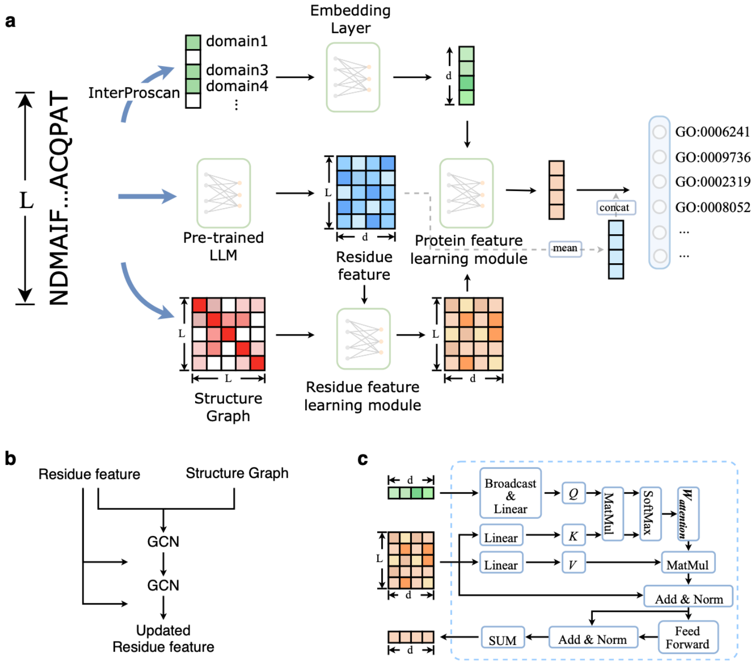

Fig. 1 | Model architectures of DPFunc. a The overview of DPFunc. It mainly consists of three aspects, including the domain information via scanning protein sequences, the residue features generated from the pre-trained protein language model, and the structure graphs constructed based on the predicted or native structures. Based on these features, a residue feature learning module and a protein feature learning module are designed to learn the residue representations and significance of residues in the structure, which are used to predict functions subsequently. **b** The details of the residue feature learning module. It utilizes GCN layers and residual operation to update residue features based on the pre-trained features and structure graphs. **c** The details of the protein feature learning module. Inspired by self-attention, it takes domain information and residue representations as input, and calculates the importance of different residues in structures to generate protein features.

Results

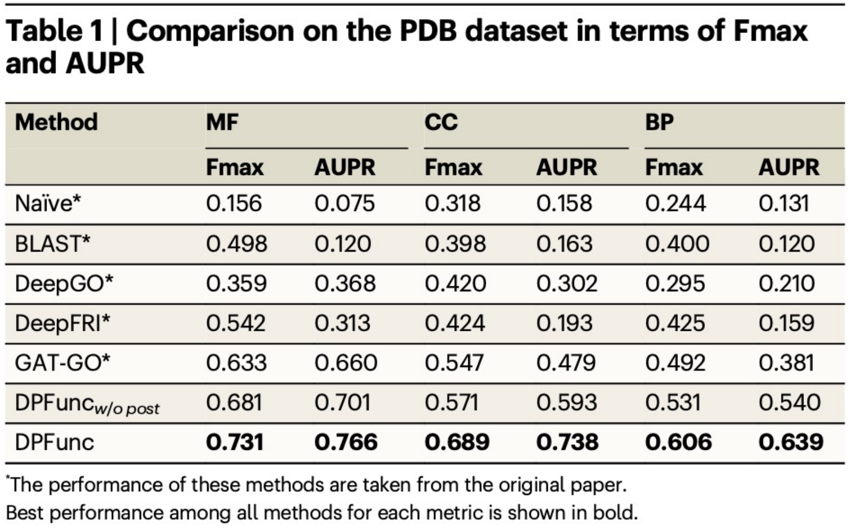

Table 1 | Comparison on the PDB dataset in terms of Fmax and AUPR

Method	MF		CC		BP	
	Fmax	AUPR	Fmax	AUPR	Fmax	AUPR
Naïve*	0.156	0.075	0.318	0.158	0.244	0.131
BLAST*	0.498	0.120	0.398	0.163	0.400	0.120
DeepGO*	0.359	0.368	0.420	0.302	0.295	0.210
DeepFRI*	0.542	0.313	0.424	0.193	0.425	0.159
GAT-GO*	0.633	0.660	0.547	0.479	0.492	0.381
$\text{DPFunc}_{w/o\ post}$	0.681	0.701	0.571	0.593	0.531	0.540
DPFunc	**0.731**	**0.766**	**0.689**	**0.738**	**0.606**	**0.639**

*The performance of these methods are taken from the original paper.
Best performance among all methods for each metric is shown in bold.

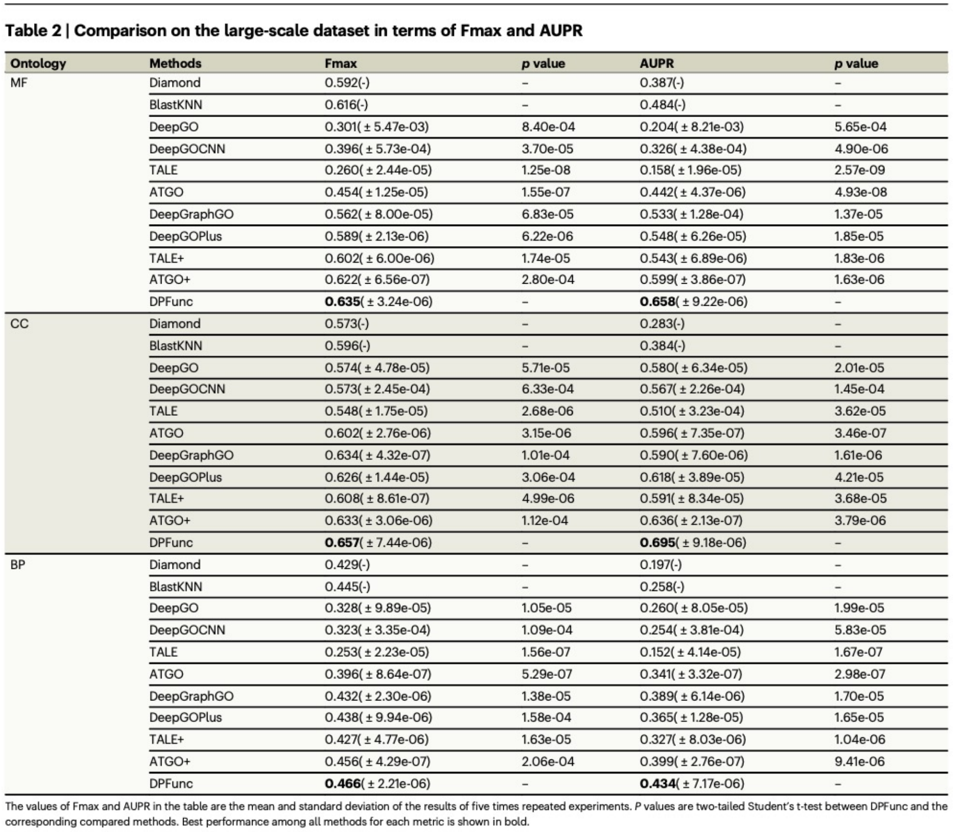

Table 2 | Comparison on the large-scale dataset in terms of Fmax and AUPR

Ontology	Methods	Fmax	*p* value	AUPR	*p* value
MF	Diamond	0.592(-)	–	0.387(-)	–
	BlastKNN	0.616(-)	–	0.484(-)	–
	DeepGO	0.301(± 5.47e-03)	8.40e-04	0.204(± 8.21e-03)	5.65e-04
	DeepGOCNN	0.396(± 5.73e-04)	3.70e-05	0.326(± 4.38e-04)	4.90e-06
	TALE	0.260(± 2.44e-05)	1.25e-08	0.158(± 1.96e-05)	2.57e-09
	ATGO	0.454(± 1.25e-05)	1.55e-07	0.442(± 4.37e-06)	4.93e-08
	DeepGraphGO	0.562(± 8.00e-05)	6.83e-05	0.533(± 1.28e-04)	1.37e-05
	DeepGOPlus	0.589(± 2.13e-06)	6.22e-06	0.548(± 6.26e-05)	1.85e-05
	TALE+	0.602(± 6.00e-06)	1.74e-05	0.543(± 6.89e-06)	1.83e-06
	ATGO+	0.622(± 6.56e-07)	2.80e-04	0.599(± 3.86e-07)	1.63e-06
	DPFunc	**0.635**(± 3.24e-06)	–	**0.658**(± 9.22e-06)	–
CC	Diamond	0.573(-)	–	0.283(-)	–
	BlastKNN	0.596(-)	–	0.384(-)	–
	DeepGO	0.574(± 4.78e-05)	5.71e-05	0.580(± 6.34e-05)	2.01e-05
	DeepGOCNN	0.573(± 2.45e-04)	6.33e-04	0.567(± 2.26e-04)	1.45e-04
	TALE	0.548(± 1.75e-05)	2.68e-06	0.510(± 3.23e-04)	3.62e-05
	ATGO	0.602(± 2.76e-06)	3.15e-06	0.596(± 7.35e-07)	3.46e-07
	DeepGraphGO	0.634(± 4.32e-07)	1.01e-04	0.590(± 7.60e-06)	1.61e-06
	DeepGOPlus	0.626(± 1.44e-05)	3.06e-04	0.618(± 3.89e-05)	4.21e-05
	TALE+	0.608(± 8.61e-07)	4.99e-06	0.591(± 8.34e-05)	3.68e-05
	ATGO+	0.633(± 3.06e-06)	1.12e-04	0.636(± 2.13e-07)	3.79e-06
	DPFunc	**0.657**(± 7.44e-06)	–	**0.695**(± 9.18e-06)	–
BP	Diamond	0.429(-)	–	0.197(-)	–
	BlastKNN	0.445(-)	–	0.258(-)	–
	DeepGO	0.328(± 9.89e-05)	1.05e-05	0.260(± 8.05e-05)	1.99e-05
	DeepGOCNN	0.323(± 3.35e-04)	1.09e-04	0.254(± 3.81e-04)	5.83e-05
	TALE	0.253(± 2.23e-05)	1.56e-07	0.152(± 4.14e-05)	1.67e-07
	ATGO	0.396(± 8.64e-07)	5.29e-07	0.341(± 3.32e-07)	2.98e-07
	DeepGraphGO	0.432(± 2.30e-06)	1.38e-05	0.389(± 6.14e-06)	1.70e-05
	DeepGOPlus	0.438(± 9.94e-06)	1.58e-04	0.365(± 1.28e-05)	1.65e-05
	TALE+	0.427(± 4.77e-06)	1.63e-05	0.327(± 8.03e-06)	1.04e-06
	ATGO+	0.456(± 4.29e-07)	2.06e-04	0.399(± 2.76e-07)	9.41e-06
	DPFunc	**0.466**(± 2.21e-06)	–	**0.434**(± 7.17e-06)	–

The values of Fmax and AUPR in the table are the mean and standard deviation of the results of five times repeated experiments. *P* values are two-tailed Student's t-test between DPFunc and the corresponding compared methods. Best performance among all methods for each metric is shown in bold.

Results

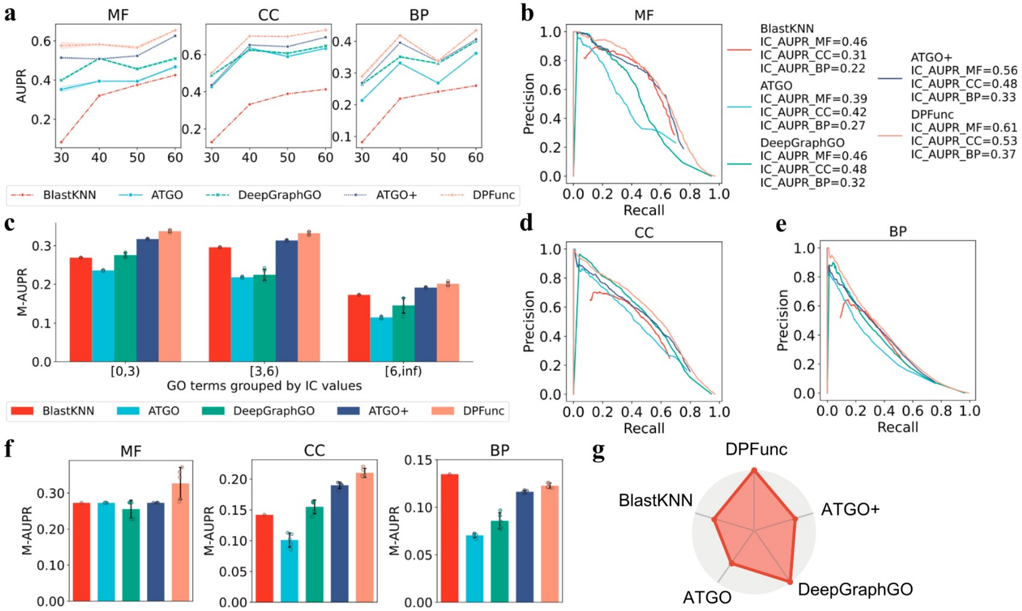

Fig. 2 | Detailed analyses of model performance. a The performance comparison of DPFunc and other representative methods on difficult protein sets with different sequence similarities to training proteins, where the data from five repeated experiments are presented as mean value +/- standard errors. **b**, **d**, **e** The IC weighted PR curve of DPFunc and other representative methods on MF, CC and BP, respectively. **c** The performance evaluation of DPFunc and other representative methods 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. **f** The performance of DPFunc and other representative methods on GO terms with deeper depths, where the distances between GO terms and root node (MF/CC/BP) are larger than 8, 6, and 8, respectively. 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. **g** The coverage of predicted functions from DPFunc and other representative methods. DPFunc can predict all known functions while others can only predict parts of functions.

Results

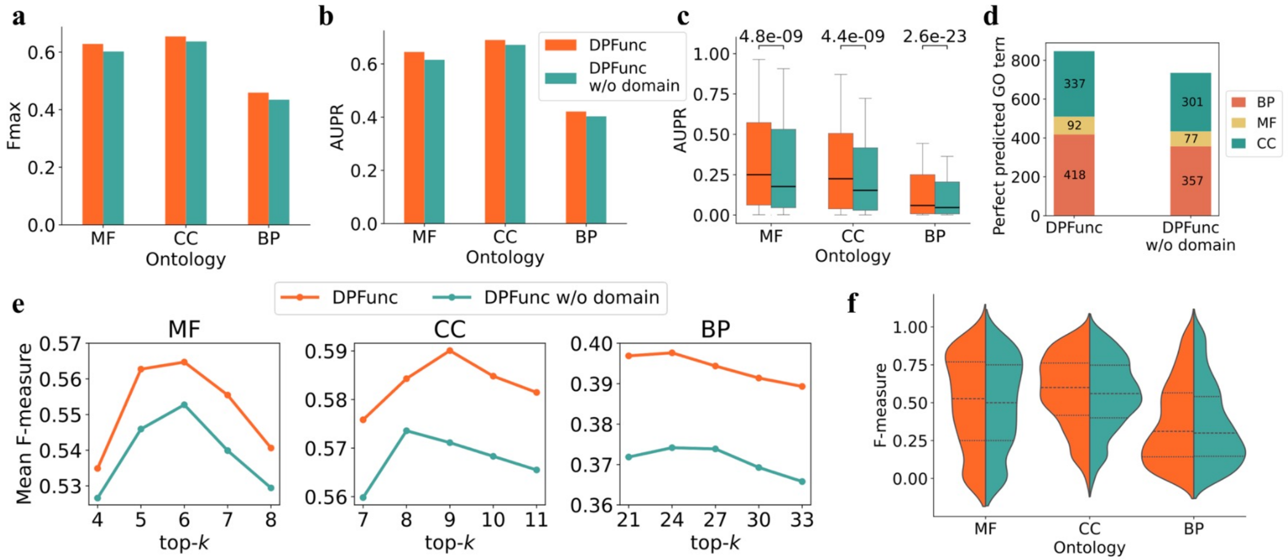

Fig. 3 | The analyses of the role of domain information. a,b The comprehensive comparison of DPFunc and DPFunc w/o domain in terms of Fmax and AUPR. **c** The performance on each function. AUPR values are calculated separately for each GO term (remove the perfect predicted GO terms which are shown in Fig. 3d). The median is represented by the centerline of the boxplot, while the first and third quartiles are indicated by the bounds of the box. The whiskers represent the 0.8 interquartile range (IQR). Specifically, there are 424 MF GO terms, 457 CC GO terms and 3283 BP GO terms for DPFunc. And there are 460 MF GO terms, 472 CC GO terms and 3343 BP GO terms for DPFunc w/o domain. Two-side paired t-tests are conducted on the overall performance of these two models and the resulting P values are annotated at the top of the boxes. **d** The number of perfect predicted GO terms. **e** The performance of top-*k* predicted functions of each protein. Since there are 8, 10, and 30 GO terms per protein on average in MF, BP, and CC, different ranges of *k* are selected (4-8 for MF, 7-11 for CC, and 21-33 for BP, respectively). **f** The performance of top-*k* predicted functions of each protein, where *k* is exactly set as 5, 9, 24 for MF, CC, and BP, respectively.

- To further illustrate the potential of DPFunc in detecting similar structural motifs, even in the absence of sequence similarity, they conducted two case studies: P0C617 and Q8NGY0, two pivotal plasma membrane proteins that separate the cell from its external environment.

Results

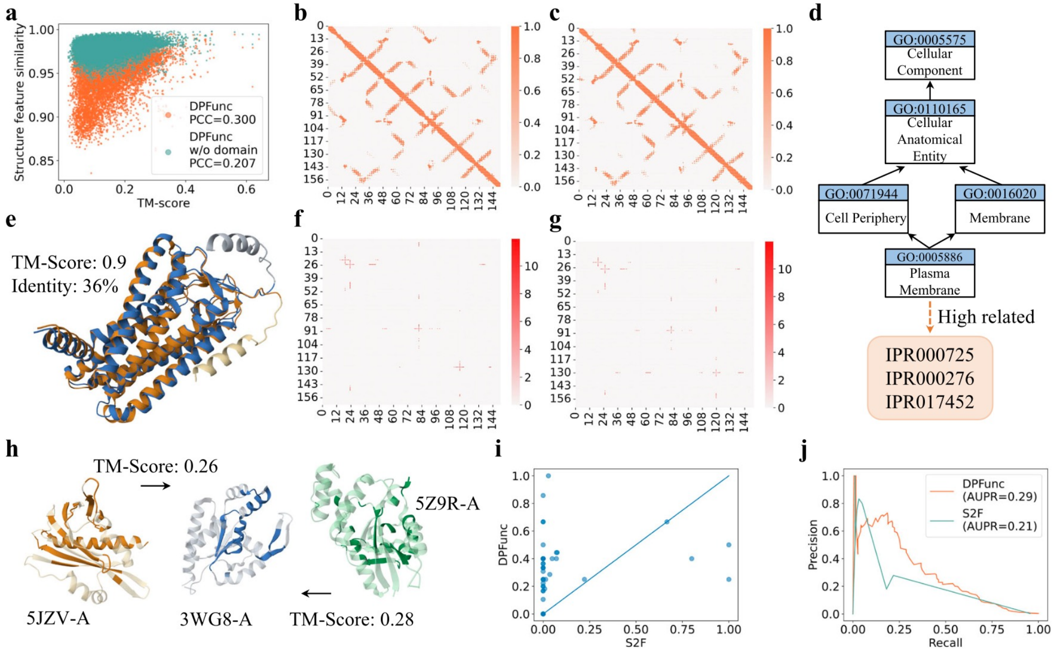

Fig. 4 | The performance of DeepDugest on structure motifs learning. a The correlations between the learned structure feature similarities and structure similarities on protein pairs with low sequence similarities. **b**, **c** The constructed structure graphs of two proteins, P0C617 and Q8NGY0, where orange points represent the edges between residues. **d** The functions of these two proteins (P0C617 and Q8NGY0) and corresponding related domains. **e** The structure alignment results of P0C617 and Q8NGY0. **f**, **g** The views of attention maps of P0C617 and Q8NGY0, where red points represent the key residues and regions detected by DPFunc. **h** The structure alignment results between 5JZV-A, 3WG8-A and 5Z9R-A. Dark colors in each protein represent residues that are aligned and light colors represent residues that are not aligned. **i** The performance of DPFunc and S2F on 47 proteins from bacteria (*Bacillus subtilis*, BACSU). The coordinates of each scatter indicate the F-measure values of these two methods on one protein. **j** The PR curve and AUPR values of DPFunc and S2F on BACSU proteins.

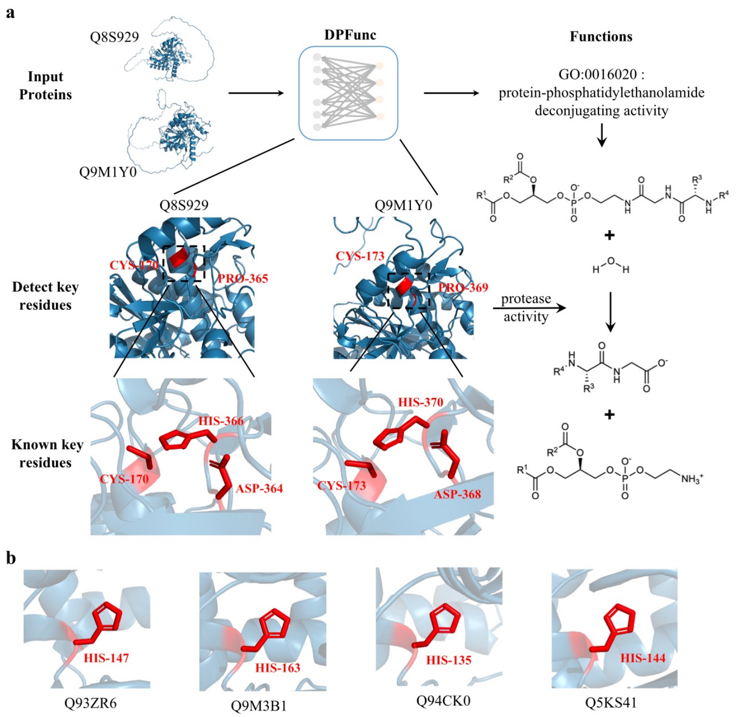

Fig. 5 | Key residues detected by DPFunc. a The details detected by DPFunc on two important cysteine proteases. The red positions shown in the structures are the key residues detected by DPFunc (CYS-170, PRO-305 for Q8S929, and CYS-173, PRO-369 for Q9M1Y0). The three red residues in the detailed graphs are the active sites that have been validated (CYS-170, ASP-364, HIS-366 for Q8S929, and CYC-173, ASP-368, HIS-370 for Q9M1Y0). These residues play significant roles in autophagy and perform the functions (mediating both proteolytic activation and delipidation of ATG8 family proteins). **b** The red positions shown in the structures are the validated active sites of four *Arabidopsis thaliana* proteins (HIS-147 for Q93ZR6/WSD1, HIS-163 for Q9M3B1/WSD6, HIS-135 for Q94CK0/WSD7 and HIS-144 for Q5KS41/WSD11), performing the same functions, involving in cuticular wax biosynthesis.