Nucleic Acids Research, 2024, 1–8
https://doi.org/10.1093/nar/gkae381
Web Server Issue

GPSFun: geometry-aware protein sequence function predictions with language models

Qianmu Yuan [1], Chong Tian [1], Yidong Song [1], Peihua Ou [1], Mingming Zhu [1], Huiying Zhao [2,*] and Yuedong Yang [1,*]

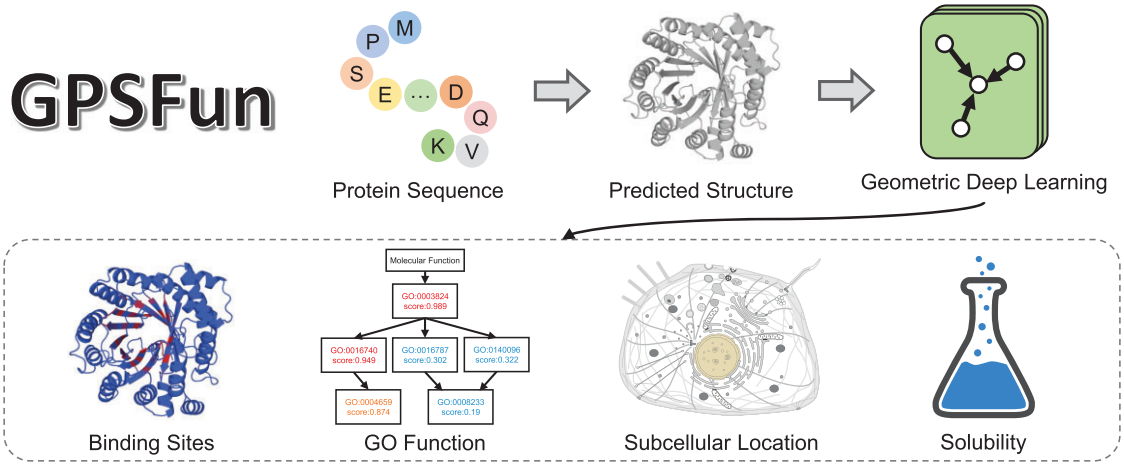
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510000, China
[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510000, China

*To whom correspondence should be addressed. Tel: +86 20 37106046; Fax: +86 20 37106020; Email: yangyd25@mail.sysu.edu.cn
Correspondence may also be addressed to Huiying Zhao. Email: zhaohy8@mail.sysu.edu.cn

Abstract

Knowledge of protein function is essential for elucidating disease mechanisms and discovering new drug targets. However, there is a widening gap between the exponential growth of protein sequences and their limited function annotations. In our prior studies, we have developed a series of methods including GraphPPIS, GraphSite, LMetalSite and SPROF-GO for protein function annotations at residue or protein level. To further enhance their applicability and performance, we now present GPSFun, a versatile web server for Geometry-aware Protein Sequence Function annotations, which equips our previous tools with language models and geometric deep learning. Specifically, GPSFun employs large language models to efficiently predict 3D conformations of the input protein sequences and extract informative sequence embeddings. Subsequently, geometric graph neural networks are utilized to capture the sequence and structure patterns in the protein graphs, facilitating various downstream predictions including protein–ligand binding sites, gene ontologies, subcellular locations and protein solubility. Notably, GPSFun achieves superior performance to state-of-the-art methods across diverse tasks without requiring multiple sequence alignments or experimental protein structures. GPSFun is freely available to all users at https://bio-web1.nscc-gz.cn/app/GPSFun with user-friendly interfaces and rich visualizations.

Graphical abstract



Introduction

Knowledge of protein function is crucial for comprehending metagenome functions, unraveling disease mechanisms and discovering new drug targets (1). Since biochemical experiments for protein function determination are expensive, time-consuming, and of low throughput (2), there is currently a widening gap between the rapid expansion of protein sequences and their limited function annotations (3). To this end, numerous computational tools have been developed for protein function predictions at residue and protein levels, such as protein–ligand binding sites (4–9), gene ontologies (GO) (10–14), subcellular locations (15–17) and protein solubility (18–20).

Despite the abundance of protein function predictors designed for various tasks, a one-stop comprehensive platform that offers high-quality predictions covering a wide range of functions is lacking. Furthermore, many existing sequence-based methods, such as TargetS (21), heavily rely on multiple sequence alignments (MSA), which are computationally expensive and futile for orphan proteins that lack close homologs. While our previous studies, LMetalSite (9) and SPROF-GO (12), have overcome this issue by substituting

Received: March 7, 2024. **Revised:** April 22, 2024. **Editorial Decision:** April 25, 2024. **Accepted:** April 26, 2024

MSA with language model representations, the absence of structural information still presents an opportunity for enhancing accuracy. By comparison, experimental structure-based approaches encoding protein structures via graph neural networks (GNN) (4–7,22–25) are often more effective. Nevertheless, most of these methods have not yet fully explored the geometry within the structure. More importantly, structure-based methods are not applicable to novel proteins with unsolved structures. Although our previously developed GraphSite (8) has shown the feasibility of leveraging AlphaFold2-predicted structures (26) for DNA-binding site prediction, the computationally intensive structure prediction pipeline hinders its application to sequences absent from the AlphaFold Protein Structure Database (27).

Based on the recent prosperity of protein language models (28,29), ESMFold (30) has emerged as a promising alternative to AlphaFold2, which replaces MSA with a large-scale pre-trained protein language model to significantly accelerate the prediction speed while maintaining comparable accuracy. To facilitate protein structure modeling, geometric deep learning has recently flourished in protein structure pre-training (31), protein design (32,33), protein docking (34,35), and binding site prediction (4–6,22). Building upon these recent advancements, it is promising to further enhance the applicability and performance of our previously well-validated methods for protein function annotations (7–9,12,18).

Here, we present GPSFun, a versatile web server for Geometry-aware Protein Sequence Function annotations, including protein binding sites for various ligands (i.e. DNA, RNA, peptide, protein, ATP, HEM, Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}), gene ontologies, subcellular locations and protein solubility. Specifically, starting from the protein sequences in FASTA format, GPSFun employs pre-trained language models to efficiently predict the 3D conformations of the proteins and extract informative sequence embeddings. Subsequently, geometric GNNs are utilized to synergistically capture the sequence and structure patterns in the protein graphs for diverse downstream tasks. Notably, GPSFun is independent of MSA and experimental protein structures, enabling fast and accurate predictions from sequences. Experiments demonstrate that GPSFun substantially outperforms state-of-the-art methods across various tasks. By providing user-friendly interfaces and rich interactive visualizations, GPSFun serves as a reliable and efficient tool for biologists and chemists. The GPSFun web server is freely available to all users at https://bio-web1.nscc-gz.cn/app/GPSFun.

Materials and methods

Benchmark datasets

The benchmark datasets for assessing binding site predictions of DNA, RNA, peptide, ATP and HEM are compiled from BioLiP (36). For each ligand, we collected the corresponding binding proteins with resolutions ≤3.0 Å and lengths ranging from 50 to 1500 released on 29 March 2023. We combined the binding site annotations of identical sequences and then removed redundant sequences sharing identity >25% over 30% alignment coverage using CD-HIT (37). Subsequently, each benchmark dataset was split into a training set with proteins released before 1 January 2021, and an independent test set with proteins released between 1 January 2021 and 29 March 2023. The datasets of protein-protein and protein-metal-ion (Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}) binding sites are directly obtained from our previous studies (7,9). To evaluate GO, subcellular localization, and solubility predictions, we adopted the datasets from (11), (17) and (20), respectively. More details of these benchmark datasets are provided in Supplementary Note S1 and Supplementary Tables S1–S4.

The workflow of GPSFun

The workflow of GPSFun is shown in Figure 1. For an input sequence, GPSFun first adopts the language model-based folding algorithm ESMFold (30) to predict the 3D conformation of the protein. Then, another pre-trained protein language model ProtTrans (version: ProtT5-XL-U50) (29) is used to extract sequence embedding, which is further normalized via min-max normalization as in (9,12). Subsequently, a geometric featurizer is employed to capture the residual and relational geometric contexts in the predicted structure. We also calculate the relative solvent accessibility and secondary structure profile from the predicted structure using DSSP (38) as done in our previous works (7,8). The resulting geometric-aware protein attributed graph is input to a set of GNNs to discover high-level patterns for various downstream tasks, including protein–ligand binding site, GO function, subcellular localization and solubility predictions.

The geometric featurizer

GPSFun represents a protein as a radius graph where residues constitute the nodes and adjacent nodes (distance between C_α < 15 Å) are connected by edges. An end-to-end featurizer is utilized to extract geometric features similar to (33), except that we additionally encode the sidechain conformations of the residues. Specifically, a local coordinate system is first defined at each residue based on the relative positions of the backbone C_α, N and C atoms. Then, several SE(3)-invariant geometric features are derived to capture the arrangements of backbone and sidechain atoms in or between residues. The geometric node features consist of intra-residue distances between any two atoms, relative directions of other inner atoms to C_α, as well as bond and torsion angles. The geometric edge features consist of inter-residue distances between any two atoms from the adjacent residues respectively, relative directions of all atoms in the neighboring residue to C_α of the central residue, as well as rotation angles between the two reference frames of the neighboring nodes. To encode the sidechain conformations, the centroids of the heavy sidechain atoms are calculated, which participate in the above feature calculations as regular atoms. The detailed definitions of the geometric features are given in Supplementary Note S2.

The deep graph neural networks

Given a protein attributed graph containing ProtTrans, DSSP and geometric node features, as well as geometric edge features, several GNN layers are adopted to learn the high-level residue representations. Specifically, we denote the hidden feature vectors of node i and edge $j \to i$ in layer l as h_i^l and e_{ji}^l, respectively. To update node i, the message passing in layer l is performed as follows:

$$\hat{h}_i^{l+1} = h_i^l + \sum_{j \in N(i) \cup i} \alpha_{ji}^l \left(W_V^l h_j^l + W_E^l e_{ji}^l \right) \quad (1)$$

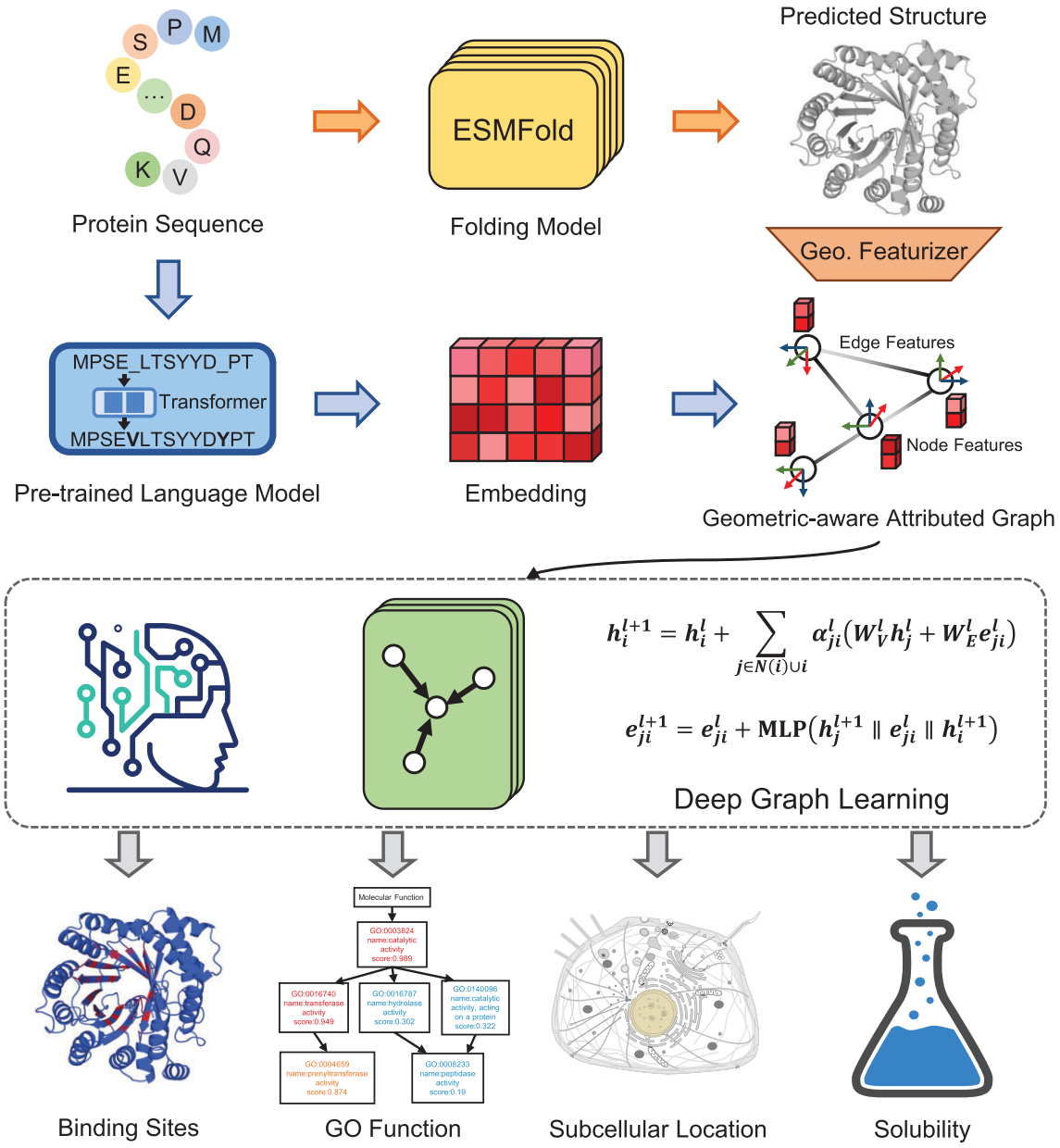


Figure 1. The workflow of GPSFun. For an input sequence, GPSFun first adopts the language model-based folding algorithm ESMFold to efficiently predict the 3D conformation of the protein. Then, another pre-trained protein language model is used to extract informative sequence embedding, and a geometric featurizer is employed to capture the residual and relational geometric contexts in the predicted structure. The resulting geometric-aware protein attributed graph is fed into a set of deep graph neural networks to discover high-level patterns for various downstream tasks, including protein–ligand binding site, GO function, subcellular localization and solubility predictions.

where the attention coefficient α_{ji}^l from node j to i is calculated by:

$$\begin{cases} w_{ji}^l = \dfrac{\left(W_Q^l h_i^l\right)^T \left(W_K^l h_j^l + W_E^l e_{ji}^l\right)}{\sqrt{d}} \\ \alpha_{ji}^l = \dfrac{\exp w_{ji}^l}{\sum_{k \in N(i) \cup i} \exp w_{ki}^l} \end{cases} \tag{2}$$

W_Q^l, W_K^l, W_V^l and W_E^l are learnable weight matrices, and $N(i)$ denotes the neighbours of node i. Then we update the features of an edge using its connecting nodes:

$$e_{ji}^{l+1} = e_{ji}^l + \mathrm{MLP}\left(\hat{h}_j^{l+1} \| e_{ji}^l \| \hat{h}_i^{l+1}\right) \tag{3}$$

where $\|$ denotes vector concatenation and MLP denotes multi-layer perceptron. We also exploit the global node update module in (33) to capture the global information.

Training and evaluation

To train the models for protein–ligand binding site, subcellular localization, and solubility predictions, we conducted five-fold cross-validation on the training sets. For GO prediction, the models were trained on the training sets using five different random seeds and evaluated on the pre-defined validation sets. All hyperparameters were optimized via grid search based on the performance of the validation sets. In the test phase, all five trained models (from cross-validation or different seeds) were

used to make predictions, which were averaged as the final prediction of GPSFun. Multi-task learning was employed to train the binding site data for different ligands concurrently as in LMetalSite (9), and we integrated the native and predicted structures to augment the training process. The homology-based label-diffusion in SPROF-GO (12) is also incorporated into GPSFun for GO and subcellular localization predictions. We adopted Pytorch (39) to implement GPSFun, and Adam optimizer (40) for model optimization with binary cross entropy loss. More details of the architecture and training strategy of GPSFun are provided in Supplementary Table S5. Besides, the implementations of the baseline methods are detailed in Supplementary Note S3. Consistent with previous studies, we use recall (Rec), precision (Pre), accuracy (Acc), Jaccard, F1-score (F1), maximum protein-centric F-measure (F_{max}), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPR) to evaluate the prediction performance, whose detailed definitions are given in Supplementary Note S4.

Web server implementation

GPSFun is run on a nginx (https://nginx.org/) server, with a backend based on Go (https://go.dev/) and a frontend based on Vue 3 (https://vuejs.org/). The combination of MySQL (https://www.mysql.com/) and MongoDB (https://www.mongodb.com/) is employed as the database solution. The interactive user interface components are provided by Element Plus (https://element-plus.gitee.io/en-US/). Protein structures are visualized using Mol* (41) (https://molstar.org/), and the GO function predictions are visualized by the directed acyclic graphs (DAG) based on Graphviz (https://graphviz.org/). The users' submitted jobs are queued and then run on a cluster of NVIDIA Tesla V100 GPUs (16 GB).

Results

The GPSFun web server

The GPSFun website (https://bio-web1.nscc-gz.cn/app/GPSFun) is free and open to all users and there is no login requirement. GPSFun neither utilizes cookies nor collects any personal information. GPSFun is compatible with most web browsers including Microsoft Edge, Google Chrome, Apple Safari and Mozilla Firefox across major operating systems including Windows, MacOS and Linux.

Inputs

The home page of GPSFun is shown in Figure 2A, where users can use the navigation bar to submit data, browse the brief introduction of GPSFun, read the detailed tutorial of the server, and download the datasets for training and evaluating GPSFun. To start, users can either paste the protein sequences of interest into the text box or upload a file in FASTA format. Batch predictions are supported for up to 20 proteins. An example input can be automatically loaded with a simple click. After submitting the example input or clicking the 'Example output' button, the prediction results for the example sequences will be displayed for demonstration.

Outputs

Once the sequences are submitted, users will be directed to a task page similar to Figure 2B. Users can bookmark this page to retrieve their results within 2 months. Typically, the initialization of the environment and loading of all pre-trained models require less than 5 min, while the annotation for a protein with 500 residues takes about 2 min. Upon task completion, a log file is available. Importantly, an overview of the submitted proteins is presented, including protein ID, length, predicted local distance difference test (pLDDT) and predicted TM-score (pTM) estimated by ESMFold. Higher scores of pLDDT and pTM indicate greater confidence in the predicted structures. Users can select a protein for detailed predictions and visualizations, or click the download button below to obtain the predictions for all proteins. Here, we discuss the outputs of GPSFun using the prediction results for the echPT1 gene of Aspergillus ruber (UniProt (42) ID: A0A2D1VNJ8) as an example.

The result page of GPSFun is divided into five sections. It starts with the basic information of the target protein including ID, length, pLDDT, pTM and the predicted solubility by GPSFun (Figure 2C). The secondary structure and relative solvent accessibility calculated from the ESMFold-predicted structure using DSSP are also visualized, where users can further hover the mouse over a sequence position to explore the detailed properties of a residue. The second section (Figure 2D) displays the protein binding site annotations covering ten available ligand types including DNA, RNA, peptide, protein, ATP, HEM, Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}. To explore the ligand-binding hotspots, users can examine the top n residues with the highest predicted scores in an interactive form. Notably, a structure view panel exhibits the predicted structure along with the GPSFun-predicted ligand-binding propensities. The confidence of the predictions is represented with a gradient of color from blue for non-binding to red for binding. The third section (Figure 2E) illustrates the GO function annotations regarding molecular function (MF), biological process (BP), and cellular component (CC). The hierarchy of the predictions is visualized by a DAG with nodes displayed in different colors based on the predictive scores of the GO terms. Users can click a node in the DAG to explore detailed information on the AmiGO website (https://amigo.geneontology.org/amigo). The fourth section presents the subcellular localization annotations of the protein (Figure 2F). Since similar sequences tend to share similar functions, the last section (Figure 2F) provides cross-links to other similar proteins in Swiss-Prot (42) for reference based on DIAMOND (43).

Validation

For protein–ligand binding site predictions, we compared GPSFun with state-of-the-art sequence-based methods including GraphSite (8), PepBind (44), PepBCL (45), TargetS (21), and LMetalSite (9), as well as experimental structure-based methods including GraphBind (23), GeoBind (22), aaRNA (46), PepNN (47), MaSIF-site (4), GraphPPIS (7), ScanNet (5), DELIA (48) and IonCom (49). As shown in Table 1 and Supplementary Figures S1 and S2, GPSFun surpasses all competing methods in AUPR by over 17.6%, 14.2%, 55.0%, 1.9%, 29.3%, 12.0%, 6.8%, 17.5%, 16.8% and 15.0% in the independent test sets of DNA, RNA, peptide, protein, ATP, HEM, Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}, respectively. To further illustrate the effectiveness of sequence embeddings and predicted structures from language models, we conducted ablation studies as shown in Supplementary Table S6. By employing ProtTrans embeddings as sequence features instead of the

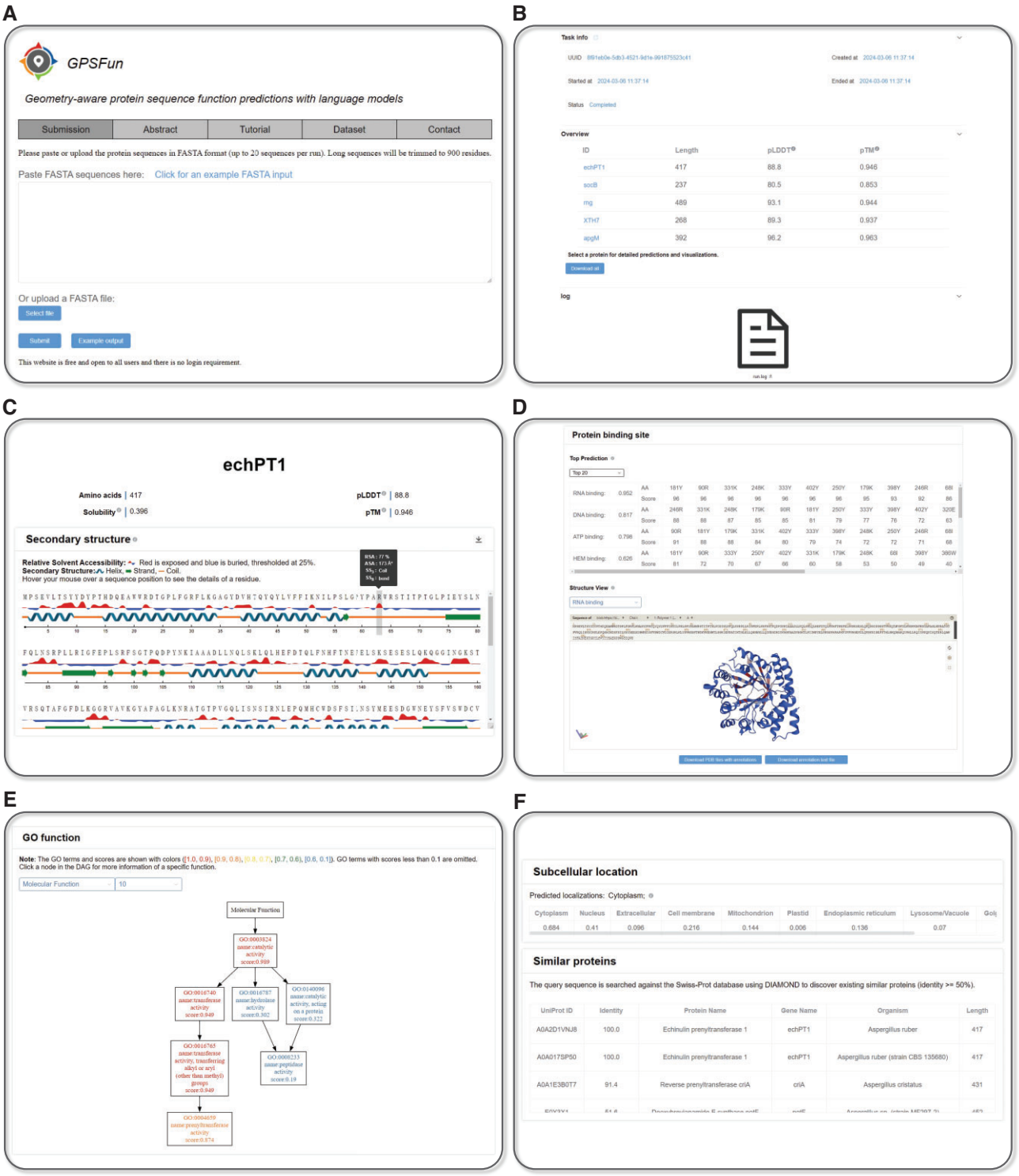


Figure 2. The GPSFun web server. (**A**) The home page of GPSFun. (**B**) The task page of GPSFun with an overview of the submitted proteins and a running log. (C–F) The outputs of GPSFun for the example input (echPT1 gene). (**C**) The confidence metrics and the secondary structure visualizations of the ESMFold-predicted structure. The solubility prediction is also provided. (**D**) Visualizations of the protein–ligand binding site predictions. (**E**) Visualizations of the GO function predictions. (**F**) The subcellular localization predictions, as well as cross-links to other similar proteins in Swiss-Prot.

Table 1. Performance comparison of GPSFun with state-of-the-art methods on the ligand-binding site test sets

Test set	Method	Rec	Pre	Acc	F1	MCC	AUC	AUPR
DNA	GraphBind	0.607	0.355	0.914	0.448	0.422	0.884	0.424
	GeoBind	0.520	0.442	0.935	0.478	0.445	0.896	0.443
	GraphSite	0.493	0.450	0.936	0.470	0.437	0.910	0.455
	GPSFun	0.477	0.552	0.948	0.512	0.486	**0.926**	**0.535**
RNA	aaRNA	0.422	0.360	0.870	0.389	0.318	0.803	0.359
	GeoBind	0.562	0.455	0.891	0.503	0.446	0.804	0.459
	GraphBind	0.633	0.400	0.871	0.491	0.436	0.861	0.506
	GPSFun	0.552	0.552	0.912	0.552	0.504	**0.901**	**0.578**
Peptide	PepBind	0.062	0.576	0.956	0.112	0.178	0.655	0.148
	PepNN	0.337	0.210	0.913	0.259	0.222	0.783	0.187
	PepBCL	0.168	0.389	0.951	0.234	0.233	0.758	0.222
	GPSFun	0.195	0.591	0.958	0.294	0.324	**0.846**	**0.344**
Protein	MaSIF-site	0.584	0.330	0.767	0.421	0.308	0.777	0.384
	GraphPPIS	0.670	0.320	0.745	0.434	0.328	0.794	0.422
	ScanNet	0.568	0.442	0.832	0.497	0.403	0.832	0.476
	GPSFun	0.613	0.419	0.820	0.498	0.403	**0.834**	**0.485**
ATP	GraphBind	0.529	0.473	0.967	0.499	0.483	0.901	0.503
	GeoBind	0.614	0.479	0.967	0.538	0.526	0.927	0.534
	DELIA	0.453	0.689	0.977	0.547	0.548	0.918	0.559
	GPSFun	0.720	0.678	0.981	0.698	0.688	**0.978**	**0.723**
HEM	GraphBind	0.733	0.505	0.939	0.598	0.578	0.926	0.638
	DELIA	0.604	0.670	0.957	0.636	0.614	0.928	0.664
	GeoBind	0.707	0.710	0.964	0.709	0.689	0.932	0.724
	GPSFun	0.707	0.787	0.970	0.745	0.730	**0.973**	**0.811**
Zn^{2+}	TargetS	0.454	0.749	0.987	0.566	0.578	0.874	0.593
	IonCom	0.852	0.137	0.898	0.236	0.317	0.937	0.671
	LMetalSite	0.681	0.859	0.992	0.760	0.761	0.976	0.803
	GPSFun	0.710	0.910	0.993	0.798	0.801	**0.982**	**0.858**
Ca^{2+}	GeoBind	0.279	0.515	0.985	0.362	0.372	0.895	0.348
	GraphBind	0.371	0.623	0.987	0.465	0.475	0.888	0.430
	LMetalSite	0.413	0.724	0.988	0.526	0.542	0.905	0.492
	GPSFun	0.398	0.848	0.990	0.542	0.577	**0.927**	**0.578**
Mg^{2+}	GeoBind	0.181	0.475	0.990	0.263	0.289	0.840	0.227
	GraphBind	0.273	0.414	0.989	0.329	0.331	0.776	0.231
	LMetalSite	0.245	0.728	0.991	0.367	0.419	0.865	0.316
	GPSFun	0.263	0.732	0.992	0.387	0.436	**0.895**	**0.369**
Mn^{2+}	GeoBind	0.569	0.479	0.988	0.520	0.516	0.938	0.454
	GraphBind	0.427	0.706	0.992	0.532	0.545	0.930	0.555
	LMetalSite	0.613	0.719	0.993	0.662	0.661	0.966	0.625
	GPSFun	0.662	0.730	0.994	0.695	0.692	**0.981**	**0.719**

Note: The best/second-best AUC and AUPR values are indicated by bold/underlined fonts.

Table 2. Performance comparison of GPSFun with state-of-the-art methods on the subcellular localization test set

	Micro			Macro				
Method	AUC	AUPR	F1	AUC	AUPR	F1	Acc	Jaccard
DeepLoc	0.812	0.599	0.487	0.726	0.458	0.368	0.360	0.404
DeepLoc 2.0	0.840	0.644	0.595	0.776	0.486	0.425	0.391	0.522
GPSFun	**0.876**	**0.700**	**0.629**	**0.802**	**0.538**	**0.483**	**0.416**	**0.551**

Note: Bold and underlined fonts indicate the best and second-best results, respectively.

Table 3. Performance comparison of GPSFun with state-of-the-art methods on the solubility test set

Method	Acc	MCC	AUC	AUPR
GraphSol	0.628	0.181	0.606	0.723
SoluProt	0.624	0.187	0.634	0.748
SWI	0.680	0.269	0.690	0.784
NetSolP	0.728	0.402	0.760	0.835
GPSFun	**0.734**	**0.435**	**0.792**	**0.859**

Note: Bold and underlined fonts indicate the best and second-best results, respectively.

MSA profiles we previously used (7,8), an increase of 4.2% in the average AUPR across the ten ligands is obtained. On the other hand, removing the structure information causes a substantial performance drop of 19.3% in the average AUPR. In addition, removal of the geometric featurizer within GPSFun also results in a considerable decline (11.5%) in the average AUPR, underscoring the significance of GPSFun's perception of protein geometry.

For GO predictions, GPSFun achieves superior performance to sequence-based methods BLAST-KNN, DeepGOPlus (10) and GOLabeler (50), predicted structure-based method Foldseek-KNN, as well as protein-protein interac-

tion network-based methods DeepGraphGO (11) and NetGO (13), by more than 11.6%, 25.3% and 5.8% in AUPR on the test sets of MF, BP and CC, respectively (Supplementary Table S7). Besides, GPSFun performs comparably to our previous SPROF-GO tool (12). GPSFun also generalizes well to non-homologous proteins as shown in Supplementary Table S8. Regarding subcellular localization prediction, GPSFun outperforms sequence-based predictors DeepLoc (16) and DeepLoc 2.0 (17) by more than 8.7% and 10.7% in micro and macro AUPR, respectively (see Table 2, Supplementary Table S9 and Supplementary Figure S3). GPSFun also exhibits improved performance compared to BLAST-KNN, Foldseek-KNN and the baseline model without structure information (Supplementary Table S10). For solubility prediction, GPSFun surpasses sequence-based predictors including GraphSol (18), SoluProt (19), SWI (51) and NetSolP (20) by more than 4.2% in AUC and 2.9% in AUPR (see Table 3 and Supplementary Figure S4). Similarly, GPSFun also outperforms BLAST-KNN, Foldseek-KNN and the baseline model without structures (Supplementary Table S11). Supplementary Tables S12–S15 also attest to the robustness of GPSFun according to the standard deviations of the models, as well as the benefits of the model ensemble technique.

Conclusions

Despite the availability of numerous protein function predictors tailored for diverse tasks, there is still a lack of a convenient computational platform for high-quality predictions that comprehensively cover a broad range of functions. Moreover, most existing sequence-based predictors are computationally intensive due to their reliance on MSA, and limited in accuracy owing to the absence of structural information. On the other hand, existing experimental structure-based approaches are hampered in genome-scale applications for novel proteins with unsolved structures.

Building upon our prior well-validated methods for protein function annotations at residue and protein levels, we present GPSFun, a versatile web server designed to annotate various functions for protein sequences, including protein–ligand binding sites, gene ontologies, subcellular locations and solubility. GPSFun is equipped with sequence embeddings and predicted structures from large language models, along with an advanced geometric protein encoder. Consequently, GPSFun achieves superior performance to state-of-the-art methods, while eliminating the need for MSA and experimental protein structures. The user-friendly interfaces and rich interactive visualizations offered by GPSFun enable biologists and chemists without programming backgrounds to readily understand the results. Serving as a reliable and efficient tool, GPSFun could facilitate the exploration of the intricate landscape of protein functions, thereby bridging the gap between genome and phenome.

Data availability

The source code of GPSFun and the data underlying this article are available in figshare, at https://doi.org/10.6084/m9.figshare.25324903.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

National Key Research and Development Program of China [2022YFF1203100]; Research and Development Project of Pazhou Lab (Huangpu) [2023K0606]; Shenzhen Science and Technology Plan Project [CJGJZD20220517142201004]. Funding for open access charge: Shenzhen Science and Technology Plan Project [CJGJZD20220517142201004].

Conflict of interest statement

None declared.

References

1. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
2. Costanzo,M., VanderSluis,B., Koch,E.N., Baryshnikova,A., Pons,C., Tan,G., Wang,W., Usaj,M., Hanchard,J., Lee,S.D., *et al.* (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**, aaf1420.
3. Cruz,L.M., Trefflich,S., Weiss,V.A. and Castro,M.A.A. (2017) Protein function prediction. *Methods Mol. Biol.*, **1654**, 55–75.
4. Gainza,P., Sverrisson,F., Monti,F., Rodolà,E., Boscaini,D., Bronstein,M.M. and Correia,B.E. (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods*, **17**, 184–192.
5. Tubiana,J., Schneidman-Duhovny,D. and Wolfson,H.J. (2022) ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods*, **19**, 730–739.
6. Krapp,L.F., Abriata,L.A., Cortés Rodriguez,F. and Dal Peraro,M. (2023) PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat. Commun.*, **14**, 2175.
7. Yuan,Q., Chen,J., Zhao,H., Zhou,Y. and Yang,Y. (2021) Structure-aware protein-protein interaction site prediction using deep graph convolutional network. *Bioinformatics*, **38**, 125–132.
8. Yuan,Q., Chen,S., Rao,J., Zheng,S., Zhao,H. and Yang,Y. (2022) AlphaFold2-aware protein-DNA binding site prediction using graph transformer. *Briefings Bioinf.*, **23**, bbab564.
9. Yuan,Q., Chen,S., Wang,Y., Zhao,H. and Yang,Y. (2022) Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings Bioinf.*, **23**, bbac444.
10. Kulmanov,M. and Hoehndorf,R. (2020) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, **36**, 422–429.
11. You,R., Yao,S., Mamitsuka,H. and Zhu,S. (2021) DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics*, **37**, i262–i271.
12. Yuan,Q., Xie,J., Xie,J., Zhao,H. and Yang,Y. (2023) Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings Bioinf.*, **24**, bbad117.
13. You,R., Yao,S., Xiong,Y., Huang,X., Sun,F., Mamitsuka,H. and Zhu,S. (2019) NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.*, **47**, W379–W387.
14. Yao,S., You,R., Wang,S., Xiong,Y., Huang,X. and Zhu,S. (2021) NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res.*, **49**, W469–W475.
15. Wan,S., Mak,M.W. and Kung,S.Y. (2017) FUEL-mLoc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. *Bioinformatics*, **33**, 749–750.
16. Almagro Armenteros,J.J., Sønderby,C.K., Sønderby,S.K., Nielsen,H. and Winther,O. (2017) DeepLoc: prediction of protein

subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.
17. Thumuluri,V., Almagro Armenteros,J.J., Johansen,A.R., Nielsen,H. and Winther,O. (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.*, **50**, W228–W234.
18. Chen,J., Zheng,S., Zhao,H. and Yang,Y. (2021) Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J. Cheminformatics*, **13**, 7.
19. Hon,J., Marusiak,M., Martinek,T., Kunka,A., Zendulka,J., Bednar,D. and Damborsky,J. (2021) SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics*, **37**, 23–28.
20. Thumuluri,V., Martiny,H.M., Almagro Armenteros,J.J., Salomon,J., Nielsen,H. and Johansen,A.R. (2022) NetSolP: predicting protein solubility in Escherichia coli using language models. *Bioinformatics*, **38**, 941–946.
21. Yu,D.J., Hu,J., Yang,J., Shen,H.B., Tang,J. and Yang,J.Y. (2013) Designing template-free predictor for targeting protein–ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **10**, 994–1008.
22. Li,P. and Liu,Z.P. (2023) GeoBind: segmentation of nucleic acid binding interface on protein surface with geometric deep learning. *Nucleic Acids Res.*, **51**, e60.
23. Xia,Y., Xia,C.Q., Pan,X. and Shen,H.B. (2021) GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res.*, **49**, e51.
24. Gligorijević,V., Renfrew,P.D., Kosciolek,T., Leman,J.K., Berenberg,D., Vatanen,T., Chandler,C., Taylor,B.C., Fisk,I.M., Vlamakis,H., *et al.* (2021) Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.*, **12**, 3168.
25. Lai,B. and Xu,J. (2022) Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings Bioinf.*, **23**, bbab502.
26. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
27. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
28. Rives,A., Meier,J., Sercu,T., Goyal,S., Lin,Z., Liu,J., Guo,D., Ott,M., Zitnick,C.L., Ma,J., *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Nat. Acad. Sci. U.S.A.*, **118**, e2016239118.
29. Elnaggar,A., Heinzinger,M., Dallago,C., Rehawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Steinegger,M., *et al.* (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.
30. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
31. Zhang,Z., Xu,M., Jamasb,A.R., Chenthamarakshan,V., Lozano,A., Das,P. and Tang,J. (2022) Protein representation learning by geometric structure pretraining. In: *The Eleventh International Conference on Learning Representations*.
32. Dauparas,J., Anishchenko,I., Bennett,N., Bai,H., Ragotte,R.J., Milles,L.F., Wicky,B.I.M., Courbet,A., de Haas,R.J., Bethel,N., *et al.* (2022) Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, **378**, 49–56.
33. Gao,Z., Tan,C. and Li,S.Z. (2022) PiFold: toward effective and efficient protein inverse folding. In: *The Eleventh International Conference on Learning Representations*.
34. Stärk,H., Ganea,O., Pattanaik,L., Barzilay,R. and Jaakkola,T. (2022) Equibind: geometric deep learning for drug binding structure prediction. In: *International conference on machine learning*. pp.20503–20521.
35. Lu,W., Wu,Q., Zhang,J., Rao,J., Li,C. and Zheng,S. (2022) Tankbind: trigonometry-aware neural networks for drug-protein binding structure prediction. *Adv. Neural Inform. Process. Syst.*, **35**, 7236–7249.
36. Zhang,C., Zhang,X., Freddolino,P.L. and Zhang,Y. (2024) BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **52**, D404–D412.
37. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
38. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
39. Paszke,A., Gross,S., Massa,F., Lerer,A., Bradbury,J., Chanan,G., Killeen,T., Lin,Z., Gimelshein,N. and Antiga,L. (2019) Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.*, **32**, 8026–8037.
40. Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. arXiv doi: https://arxiv.org/abs/1412.6980, 22 December 2014, preprint: not peer reviewed.
41. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
42. UniProt Consortium. (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
43. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
44. Zhao,Z., Peng,Z. and Yang,J. (2018) Improving sequence-based prediction of protein-peptide binding residues by introducing intrinsic disorder and a consensus method. *J. Chem. Inf. Model.*, **58**, 1459–1468.
45. Wang,R., Jin,J., Zou,Q., Nakai,K. and Wei,L. (2022) Predicting protein-peptide binding residues via interpretable deep learning. *Bioinformatics*, **38**, 3351–3360.
46. Li,S., Yamashita,K., Amada,K.M. and Standley,D.M. (2014) Quantifying sequence and structural features of protein-RNA interactions. *Nucleic Acids Res.*, **42**, 10086–10098.
47. Abdin,O., Nim,S., Wen,H. and Kim,P.M. (2022) PepNN: a deep attention model for the identification of peptide binding sites. *Commun. Biol.*, **5**, 503.
48. Xia,C.Q., Pan,X. and Shen,H.B. (2020) Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics*, **36**, 3018–3027.
49. Hu,X., Dong,Q., Yang,J. and Zhang,Y. (2016) Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics*, **32**, 3260–3269.
50. You,R., Zhang,Z., Xiong,Y., Sun,F., Mamitsuka,H. and Zhu,S. (2018) GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, **34**, 2465–2473.
51. Bhandari,B.K., Gardner,P.P. and Lim,C.S. (2020) Solubility-weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*, **36**, 4691–4698.

Received: March 7, 2024. **Revised:** April 22, 2024. **Editorial Decision:** April 25, 2024. **Accepted:** April 26, 2024