


HiFun: homology independent protein function prediction by a novel protein-language self-attention model

Jun Wu [†], Haipeng Qing[†], Jian Ouyang, Jijia Zhou, Zihao Gao, Christopher E Mason, Zhichao Liu and Tielu Shi

Corresponding authors. Zhichao Liu. E-mail: zhichao.liu@boehringer-ingenheim.com; Tielu Shi. E-mail: tielushi@yahoo.com

[†]Jun Wu and Haipeng Qing should be regarded as joint first authors.

Abstract

Protein function prediction based on amino acid sequence alone is an extremely challenging but important task, especially in metagenomics/metatranscriptomics field, in which novel proteins have been uncovered exponentially from new microorganisms. Many of them are extremely low homology to known proteins and cannot be annotated with homology-based or information integrative methods. To overcome this problem, we proposed a Homology Independent protein Function annotation method (HiFun) based on a unified deep-learning model by reassembling the sequence as protein language. The robustness of HiFun was evaluated using the benchmark datasets and metrics in the CAFA3 challenge. To navigate the utility of HiFun, we annotated 2 212 663 unknown proteins and discovered novel motifs in the UHGP-50 catalog. We proved that HiFun can extract latent function related structure features which empowers it ability to achieve function annotation for non-homology proteins. HiFun can substantially improve newly proteins annotation and expand our understanding of microorganisms' adaptation in various ecological niches. Moreover, we provided a free and accessible webservice at <http://www.unimd.org/HiFun>, requiring only protein sequences as input, offering researchers an efficient and practical platform for predicting protein functions.

Keywords: protein function prediction, deep-learning, self-attention, homology-independent, metagenome, protein structure

Jun Wu is an associated researcher at East China Normal University. He received his Bachelor and Master degree in Pattern Recognition and Intelligent Systems from Changchun University of Science and Technology, and Ph.D degree in Control Science and Engineering from Shanghai Jiao Tong University. He completed a post-doctoral fellowship in Biomedical Engineering at Shanghai Jiao Tong University. Currently his research is focused on developing novel deep-learning methods to uncover unknown species and functions in microbiome data, and integration of multi-omic data for disease diagnosis and therapy.

Qin Haipeng is a passionate individual with a love for mountain climbing and camping. He holds a B.S. degree in Bioinformatics from Chongqing University of Posts and Telecommunications in Chongqing City, China, as well as an M.S. degree in Biochemistry and Molecular Biology from the School of Life Sciences at East China Normal University in Shanghai City, China. Driven by curiosity and a thirst for exploration, Qin developed a keen interest in the intersection of deep learning and metagenomics, specifically focusing on large language models (LLMs). He is committed to contributing to advancements in this field by exploring the possibilities of integrating deep learning techniques into the realm of metagenomics analysis.

Jian Ouyang having earned his B.S. and M.S. degrees in Biomedical Engineering from the University of Shanghai for Science and Technology, Shanghai, China, in the years 2013 and 2016, respectively. He is currently working toward the Ph.D. degree in Biochemistry and Molecular Biology with the School of East China Normal University, Shanghai, China. His primary research focus revolves around the data mining and modeling based on clinical and multi omics data, constructing databases, developing omics analysis pipeline or tools and is deeply engaged in microbiomics research.

Jijia Zhou received the B.S. degree in Biotechnology from Henan Normal University, Xinxiang, China, in 2017 and the M.S. degree in Pathogen biology from Chinese Center for Disease Control and Prevention, Beijing, China, in 2020. She is currently working toward the Ph.D. degree in biochemistry & molecular biology with the School of East China Normal University, Shanghai, China. Her research interests include gut microbiota, metagenome and big data.

Zihao Gao is a doctoral candidate in Shi Tielu Research Group, School of Life Sciences, East China Normal University. Prior to coming to ECNU, Gao graduated with a bachelor's degree from the School of Life Sciences at Dalian University of Technology, majoring in bioinformatics. Gao's primary research interests are virology, metagenome/metatranscriptome and other sequencing data related fields.

Christopher E Mason is an assistant professor of Computational Genomics at Weill Cornell Medical College. He completed his B.S In Genetics and Biochemistry from University of Wisconsin-Madison and Ph.D in Genome Evolution and postdoctoral in Neuroscience from Yale University. His laboratory work utilizes computational and experimental methodologies to identify and characterize the essential genetic elements that guide the function of the human genome. He performs research in three principal areas: (1) the functional annotation of the human genome by mutational profiling in families with brain malformations and cancer patients, (2) the examination of the elements that orchestrate the development of the human brain and their evolutionary changes, and (3) the development of models for systems and synthetic biology. We use high-throughput methods to generate cell-specific molecular maps of genetic, epigenetic, and transcriptional activity and we use them to create multi-dimensional molecular portraits of development and disease. We develop algorithms to detect, catalog and functionally annotate variants in the genetic pathways that control developmental processes. He has more than 50 publications.

Zhichao Liu is head of Computational Toxicology at Nonclinical Drug Safety (NDS) of Boehringer Ingelheim Pharmaceuticals. At BI, Dr. Liu is leading efforts to grow AI-based solutions for next-generation drug safety evaluation and risk assessment. Dr. Liu comes to BI from US FDA, where he led Artificial intelligence (AI) Research Force (AIRForce). Dr. Liu's background spans the fields of chemistry, biology, and computer science. He led many cutting-edge projects over the past decade by designing, implementing, and deploying AI/ML solutions for advanced regulatory science and drug development. His accomplishments are reflected by >100 peer-reviewed publications and numerous scientific awards.

Tielu Shi is a full professor at East China Normal University. He received his Master degree in Plant Physiology from Shanghai Institute of Plant Physiology, Academy of Sciences in 1992; Master degree in Computer Science in 1999 and Ph.D degree in molecular biology in 2000 from the University of Louisville, USA. After obtained PhD degree, he pursued bioinformatics research and joined the Bioinformatics Center, Shanghai Institute of Biological Sciences, Chinese Academy of Sciences between 2002 and 2008. He moved to East China Normal University by the end of 2008. He has published over 150 papers. His current research interests include Clinical data standardization and integration analysis; Disease gene, disease mechanism and biomarker discovery based on multi-level data integration of omics and clinical information, Methodology developments and applications in the high through-put data (NGS data, proteomic data, etc.), Gene regulatory network analysis and protein-protein interaction network analysis, Drug target, drug efficacy and adverse analysis.

Received: March 13, 2023. **Revised:** July 31, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

INTRODUCTION

Proteins play an essential role in a large variety of biological processes, thus elucidating their function is always the major task in the post-genomic era, which can help us better understand the role of proteins in disease pathobiology, novel biomolecular tools discovering and even drug targets finding [1–6]. With the advent of next-generation sequencing techniques, an enormous amount of sequences have been produced, which exponentially expanded the protein sequences databases, but their functional annotation is lagging far behind and the gap between unannotated and annotated proteins is widening [7, 8]. In the Uniprot database [9], there are >1% of the available proteins are reliable annotated [7, 10]. Consequently, there is an urgent need to develop high-efficiency and accurate computational methods to predict protein function to shrink the gap.

The commonly used methods (e.g. BLAST/PSI-BLAST [11, 12] and Diamond [13]) to predict the protein functions are achieved by aligning the query protein to the proteins with known functions and assigning the functions based on the sequence similarity. However, there are numerous proteins with similar functions while their sequences are distinct. To address this problem partially, the domain- and motif-based methods [14–17] have been developed based on the similarity of conserved sequence domains or motifs, which are obtained by performing multiple sequence alignment of proteins belonging to the same protein family with known function. These methods fall into two major limitations, one of which is that high-quality sequence alignments are hard-won especially when the sequences with low homology, and the other one is that high quality of the functional annotation of domain/motifs is challenged. Besides the sequence features, additional functional experimental evidences (e.g. protein-protein interaction, gene expression, gene neighborhood and gene co-occurrence) can be also supplied to improve the ability the protein function prediction. The Critical Assessment of Functional Annotation Challenge (CAFA) has showed that combing multiple sources of information using integrative machine learning and statistical methods outperform traditional sequence alignment-based methods, such as deepGOplus [2, 18], deepFunc [7] and S2F [10]. However, these methods also suffered the weakness of low homology protein function prediction.

From another perspective, the 3D structure of a protein is believed to be more involved in its biological function, and can be considered as an important attribution for the protein function annotation, such as FFPred [19], COFACTOR [20] and DeepFRI [21]. Maranga *et al.* [22] developed a novel metagenome analysis pipeline which includes the deep learning-based functional annotation from DeepFRI, and performed the functional annotations for the >1000 infant metagenomes from the DIABIMMUNE cohort. However, very few experimentally verified 3D structures information is available for the proteins with known functions. Although several high-accuracy theoretical methods have been developed to predict the protein structures, such as AlphaFold2 [23, 24], RoseTTAFold [25] and ProFold [26], the accuracy still need to be improved, especially when no homologs structure is available [27]. Even so, the success of AlphaFold2 *et al.* revealed that the sequence of protein indeed implies its structure related information.

Most recently, inspired by approaches proposed for natural language processing, the proteins could be represented as protein languages. Several methods have been proposed that could resemble the protein or DNA sequences as protein language and declared that the complex sequence-structure-function relationships can be extracted [28, 29]. Inspired by these, in this study, we

introduced a novel deep-learning model HiFun to achieve the protein function prediction directly from the protein sequences. Our main idea is to extract function related latent features directly from the protein sequence using a pretrained embedding model as well as a unified deep neural network architecture. The evaluation results showed that our method outperformed other state-of-the-art general-purpose protein function prediction methods with respect to most of the evaluation metrics. Using the proposed method, we annotated the unknown proteins in the Unified Human Gastrointestinal Protein (UHGP) database and performed *de novo* motifs discovery and structure-related analysis.

MATERIAL AND METHODS

Sequence embedding based on BLOSUM62 matrix. BLOcKS SUBstitution Matrix (BLOSUM) matrix is one of the most common substitution matrices used for protein sequence alignment [30] and contains the evolutionary information of protein sequence. According to the identity of the sequences used for the BLOSUM matrix construction, several sets of BLOSUM matrices exist and named with numbers, such as BLOSUM45, BLOSUM62 and BLOSUM80. In this work, we used BLOSUM62 matrix, which was experimentally proven to be among the best for detecting most weak protein similarities, to encode protein sequences, and each amino acid was represented as the corresponding row of the BLOSUM62 matrix. For each protein sequence, we trimmed or padded the protein sequences into a fix-length of 1000 AAs (amino acids). Hence, for each protein sequence can be converted into a matrix with a dimension of 1000-length numeric vector \times 25 amino acids ('A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V', 'B', 'J', 'Z', 'X', '*').

Sequence embedding with FastText. To capture the nature of protein sequence beyond homology, we further assimilated protein sequence as a type of comprehensive yet straightforward language and employed a FastText word embedding model [30] to train all the publicly available bacterial protein sequence and then digitalize each protein sequence. Briefly, we first downloaded all 335 066 reviewed bacterial protein sequences from the UniProt database (<https://www.uniprot.org/>, release date: 12 July 2021), and then the sequences were trimmed or padded into a fix-length of 1200 AAs. After that, the FastText sequence embedding model was trained with an n-gram of character of 3 and the output embedding was set as 200. Consequently, the pretrained FastText sequence embedding model could translate the sequence into a matrix with a dimension of 200-length numeric vector \times 400-mers.

HiFun model architecture. Two types of embedding methods described above were used to transfer the protein sequences into numeric vectors/matrices, and we named the outputs for these two embedding layers as BLOSUM62-based vectors and FastText-based matrices for convenience, respectively. The BLOSUM62-based vectors were further fed into the convolutional layers to extract the evolutionary features of the input proteins. Simultaneously, the FastText-based matrices went through a sub-architecture constituted by connecting Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) with a self-attention mechanism in series. The convolutional layers were designed to extract n-gram features from vector embeddings of input sequences, while the BiLSTM layers access both the preceding and succeeding contextual features by combining a forward hidden layer and a backward hidden layer. The attention mechanism for the single amino acid (i.e. 3 k-mer combinations) representation pays more attention

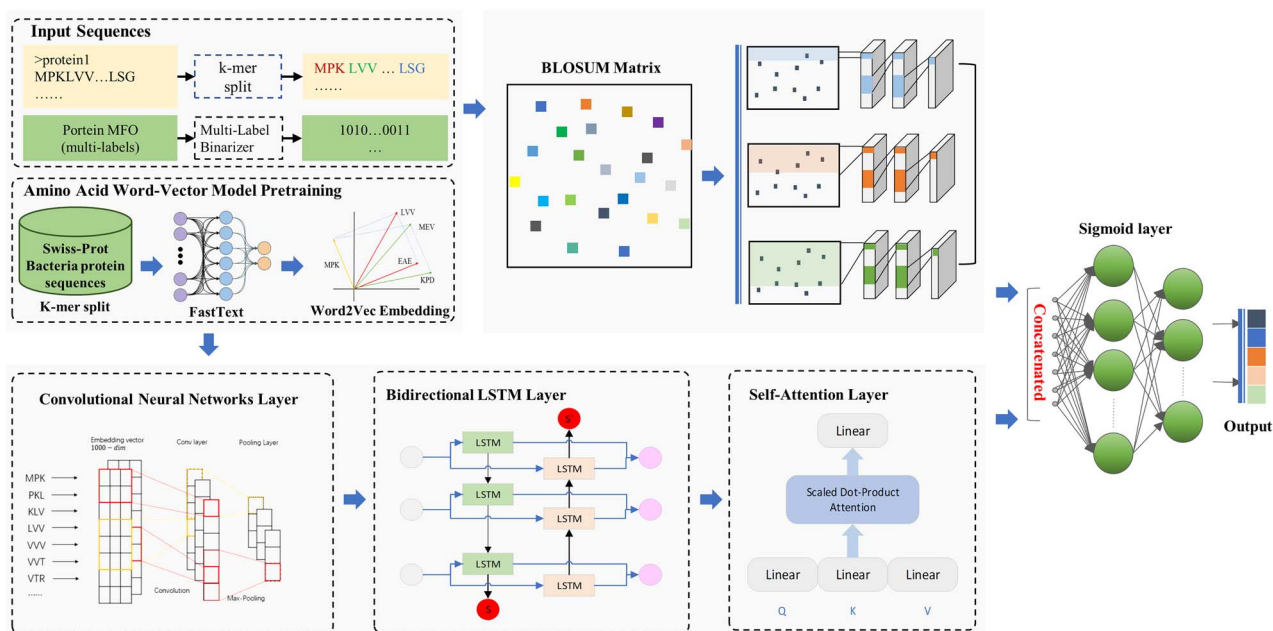


Figure 1. The architecture of HiFun. The protein sequences were embedding with both the BLOSUM matrix and the pretrained word2vec model. Two parallel sub-architectures were designed to extract the latent features and then go through a dense layer to generate the output.

to the amino acids related to the sentiment of the sequence. The outputs of the two parts were concatenated and then fed into a dense layer for protein function prediction (Figure 1).

The detail of the architecture was described as below

Input unit

Two input units (BLOSUM62-based and FastText-based) were used in our study, which were described as above.

CNN unit for BLOSUM62-based input

The BLOSUM62 embedding matrix of each sequence was fed into three parallel subnetworks and each subnetwork consisting of three convolutional layers with different sizes of filters to extract the latent features, which are processed by the subsequent max-pooling layer. Rectified Linear Units activation is used after each convolutional layer. The size of all the max-pooling layers is set 2.

CNN unit for FastText-based input

The CNN unit is used to extract both local and global features from the sequence represented by the embedding matrix. The embedding matrix of each sequence was fed into the three-layered 1-D convolutional layers with two max-pooling layers inspected between 1-D convolutional layers. Specifically, three 1-D convolutional layers consist of 32, 16 and 8 filters and window size 10, 10 and 5 to move across the embedding matrix for the feature extraction. The size of two max-pooling layers is set 2. Furthermore, the reshaping process is required before giving the output of CNN as an input to the BiLSTM, which accepts 1D in its input.

BiLSTM unit

The extracted feature vectors of sequences from the CNN unit are further passed to the two BiLSTM layers, which access the preceding and succeeding contextual features by different controlling gates (like input gate, output gate and forget gate) [31]. The 32, 16 units were used in the two BiLSTM layers. A dropout rate of 0.5 is applied to the Bi-LSTM layers. Throughout the BiLSTM unit, the

query matrix, key matrix and value matrix are generated as the input of the following self-attention unit.

Self-attention unit

To overcome the problem of long-distance dependency in the protein sequence and differentiate the association of individual amino acid to ARG and its category classification, we introduced the self-attention unit to estimate the importance of each amino acid. The self-attention mechanism is performed on the output of the BiLSTM unit to execute the importance estimation of each amino acid in the sequence. Technically, the attention process is implemented by calculating a context vector for a decoder containing the most useful information from all hidden states of the encoders, with an averaging of weights done on. Attention width of 15 and a kernel regularizer is used in the self-attention mechanism.

Output unit

The outputs of the CNN units for BLOSUM62-based input and Self-Attention unit were concatenated to 9032 feature maps, which were further processed by two fully connected layers. As our task is multi-label classification, the sigmoid layer is used to determine the GO terms of the proteins. Considering the imbalance of proteins respected to different GO terms, we applied focal loss function as the loss function [32] and the alpha and gamma of focal loss are set as default.

GO annotation propagation. We first separated each class by their sub-ontology. According to the hierarchical structure of the GO (release date: 10 September 2020), we computed the parent and child classes locally within the sub-ontology, and then the GO annotation was propagated. For example, if a protein P is annotated with a class C which has a 'part-of' relation to a class D, then we annotate P with the class D. This procedure is repeated until no further annotation can be propagated.

Performance metrics

In this study, we applied the same strategy used in the CAFA3 challenge [33] to evaluate the performance of our model. Briefly, the

individual classes are first separated by their sub-ontology (MFO, BPO and CCO) and then parent and child classes are computed locally within the sub-ontology. After that, we used the metrics F_{\max} based on the precision-recall (PR) curve and S_{\min} based on the remaining uncertainty-misinformation curve [34] to measure the performance. F_{\max} , which is a maximum F-measure computed over all prediction thresholds, is calculated using the following equations:

$$pr(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f I(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f I(f \in P_i(\tau))},$$

$$rc(\tau) = \frac{1}{n} \sum_{i=1}^n \frac{\sum_f I(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f I(f \in T_i)},$$

$$F_{\max} = \max_{\tau} \left\{ \frac{2 \times pr(\tau) \times rc(\tau)}{pr(\tau) + rc(\tau)} \right\},$$

where f is a GO term, $P_i(\tau)$ is a set of predicted annotation of a protein i with threshold τ , T_i is the corresponding ground-truth set of terms for the protein, $m(\tau)$ is the number of proteins with at least on predicted annotation with threshold τ , $I(\bullet)$ is an indicator function and n is the total number of proteins. F_{\max} was computed for prediction thresholds $\tau \in [0, 1]$ with step size of 0.01. The high F_{\max} indicated the higher performance.

S_{\min} is the minimum semantic distance between real and predicted annotations and is calculated as follow:

$$S_{\min} = \min_{\tau} \left\{ \sqrt{ru(\tau)^2 + mi(\tau)^2} \right\},$$

where $ru(\tau)$ and $mi(\tau)$ are the average remaining uncertainty and misinformation [34], respectively. The low S_{\min} indicated the higher performance.

Preprocessing of Uniprot bacterial proteins for HiFun building

In total, 335 066 bacterial proteins with MFO annotation were downloaded from the Swiss-Prot database which is a high quality manually annotated protein sequence database of UniProtKB (release date: 12 July 2021). The duplicated proteins were removed by clustering all their sequences with CD-Hit, discarding all except with 100% identity and the same length. The MFOs for each protein were propagated with the strategy mentioned above. Statistically, 3848 MFOs were retained (261 level 3 MFOs, 644 level 4 MFOs, 1806 level 5 MFOs, 734 level 6 MFOs and 403 level 7 MFOs), and >92.62% and 93.86% of the proteins harbored level 3 and level 4 MFO, respectively. To build the deep-learning model, we only considered the MFOs with >50 protein sequences. For each protein, only the level 3 and level 4 MFOs were used as the labels. Proteins without level 3 or level 4 MFOs were removed. MFOs with <50 protein sequences were also excluded for model training. Finally, 223 991 proteins along with 120 level 3 and 204 level 4 MFOs were retained for model construction.

Motif discovery and comparison

Mafft (version 7.487) was used to perform the multiple sequence alignment and the Fasttree (version 2.1.11) was applied to construct the phylogenetic tree. The de novo motif discovery is performed using MEME tool [35] with the classic mode and default

parameters. The similarity between motifs was achieved by using the TOMTOM tools in the MEME suite [36] with the default parameters.

Protein Structure alignment

The structures of UHGP proteins were predicted using RoseTTAFold [25] and structure of Swiss-Prot proteins were downloaded from the Swiss-Prot database. Structure alignment was achieved using TM-align [37] and the protein structure was visualized using Mol*Viewer [38] in the PDB webserver.

RESULTS

Evaluation and comparison

We first evaluated HiFun using the latest CAFA3 benchmarking dataset [39], which contain both the training sequences and experimental annotation, based on the metrics (F_{\max} , S_{\min} and AUPR) used in the CAFA3 challenge (see Methods and Materials). Using the GO ontology released on 10 September 2020, we propagated the GO annotation in the CAFA3 dataset (see Methods and Materials), and the final details about our updated CAFA3 annotation are shown in Table 1.

We compared our method against the top 10 methods that were evaluated in CAFA3 [33] as well as the state-of-the-art method deepGOplus. We also included two baseline methods, Naïve and sequence-based (BLAST) baseline methods [39, 40], which were used in the CAFA evaluation in our comparison. According to the CAFA3 evaluation metrics, HiFun achieves F_{\max} of 0.608, 0.462 and 0.644 for the molecular function (MFO), biological process (BPO) and cellular component (CCO) evaluations, respectively (Figure 2). For the F_{\max} metric, HiFun performed best in both the BP and CCO and it was the second-best performing method in the MFO evaluation. For the MFO, HiFun achieved the second-best performance respect to S_{\min} . While HiFun did not perform so well for the S_{\min} metric for BPO and CCO, it did considerably better than DeepGOplus.

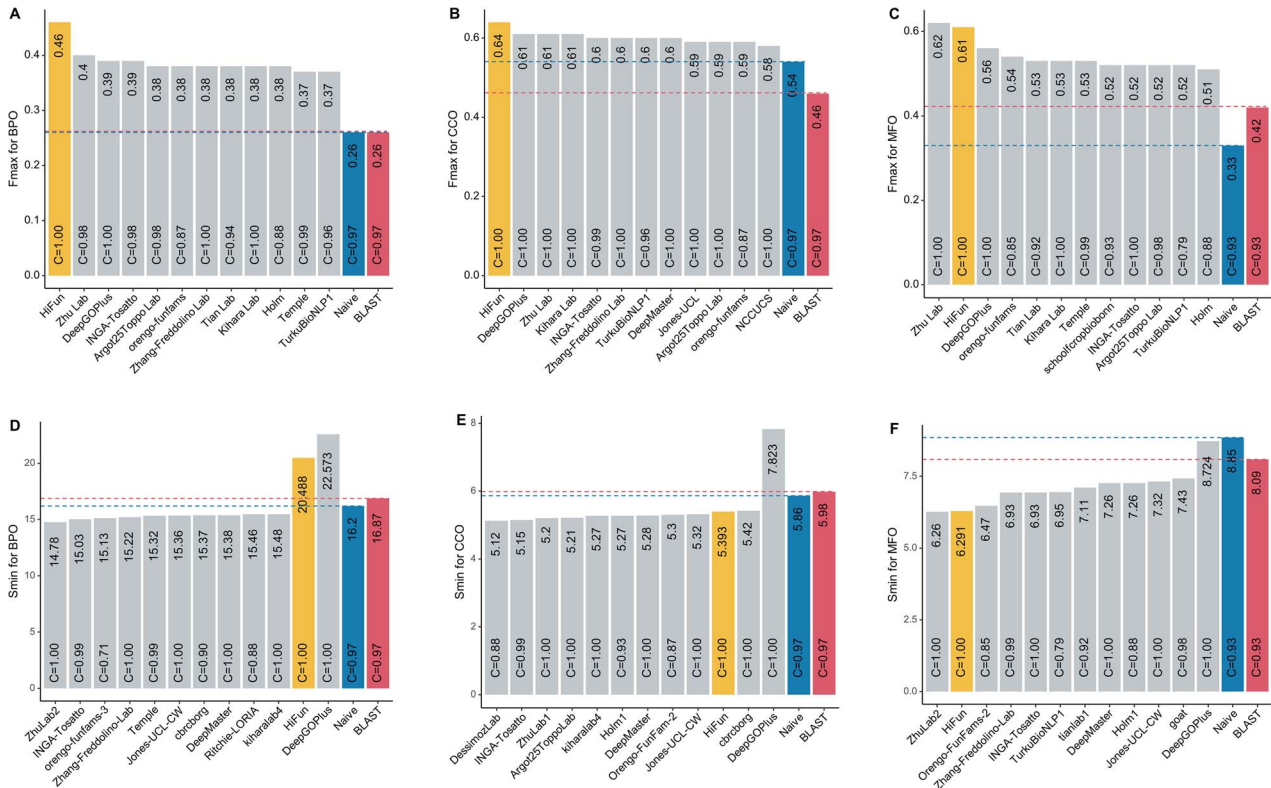
To further validate homology independence of HiFun, we divided the protein in the CAFA3 test dataset into seven groups according to the levels of sequence identity with the proteins in training dataset and inspected the performance of HiFun as well as the competing deepGOplus for each protein group. The results showed that the performance of HiFun was stable across different protein groups while deepGOplus gave a low performance for test proteins with low percentage identity with the proteins in the training dataset (Figure 3).

Establishing HiFun model for bacterial protein function prediction

The reliability and high performance of our method were proved using the CAFA datasets through comparing with the state-of-the-art methods. In this part, we aimed to build the ultimate model for the bacterial protein function prediction and only the MFO was considered. We first downloaded all the bacterial proteins with MFO annotation from the Swiss-Prot database. After removing the duplicated proteins and propagating the MFOs (see Materials and Methods), 223 991 proteins along with 334 MFO terms were remained for the model construction. These proteins were then divided into training set, validation set and independent test set with the percentage of 80, 18 and 2%, respectively (Figure 4A). According to the results for the test dataset, we can see that HiFun achieves S_{\min} of 1.34, AUPR of 0.65 and AUC of 0.98. While the threshold was set as 0.24, the F-score can reach its maximum $F_{\max} = 0.69$ (Figure 4B).

Table 1. The updated CAFA3 dataset for each sub-ontology

Dataset	Statistic	#MFO	#BPO	#CCO	#All
CAFA3	Training size	36 110	53 500	50 596	66 841
CAFA3	Testing size	1137	2392	1265	3328
CAFA3	Number of classes	556	3358	296	4210

**Figure 2.** Comparison of HiFun with CAFA3 top 10 methods as well as deepGOplus in three sub-ontologies based on the F_{\max} and S_{\min} metrics. (A-C) bar plots showing the F_{\max} of the 14 methods. Coverage of methods was labeled inside the bars. (D-F) bar plots showing the S_{\min} of the 14 methods. Coverage of methods was labeled inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods.

Considering a practical issue that massive unknown organisms and proteins have been uncovered with the rapid application of metagenomic/metatranscriptomic sequencing, the functional annotation for the proteins of newly uncovered organisms is in high demand. The network-based methods, which rely heavily on the interaction transferring of known proteins (e.g. deepFunc [7], deepGO [18] and S2F [10]), usually favor the well-characterized organisms. To investigate the performance bias of HiFun between the organisms with and without well-characterization, we applied HiFun on the proteins of three model bacteria (*Streptomyces coelicolor*, *Bacillus subtilis* and *Escherichia coli*) and three non-model bacteria (*Pseudomonas*, *Lactococcus* and *Klebsiella*) [41]. The results showed that HiFun achieves high and stable performance for both the model and non-model bacteria, which also indicated that HiFun is particularly suitable for the metagenome data which contain mounts of uncharted species and proteins (Figure 4C).

Annotating unknown proteins in the UHGG database

With the extensive application of metagenomic sequencing technology in the human gut studies [42–45], large amounts of unknown species and proteins have been uncovered, which also spur the annotation of novel species and proteins. To establish nonredundant dataset of human gut genomes, Almeida

et al. present the Gastrointestinal Genome (UHGG) collection, which consists of 204 938 non-redundant genomes encoding >170 million proteins. However, about 40% of these proteins lack functional annotations, which put off our comprehensively functional characterization of the human gut microbiota. In this part, we aimed to annotate these unknown proteins with our method.

We first retrieved the proteins without MF annotations (hereafter referred to as unknown proteins) from the UHGP-50 catalog, which was generated by clustering all coding sequences with >50% protein identity. Considering there were lots of partial proteins in the UHGP-50 catalog, we further removed the redundant protein fragments. We first performed pairwise alignment between the UHGP-50 proteins using DIMAOND [13], and then the proteins with E-value < 1E-4 were clustered. For each cluster, we inspected the overlap region between each protein pair and the proteins covered by any other proteins were removed. Finally, 2 218 032 non-redundant unknown proteins were remained for further analysis. With threshold leading the maximum F-score in the independent test dataset mentioned above, HiFun can annotate 99.76% (2 212 663) of these unknown proteins, involving 176 MFOs (including 80 level 3 MFOs and 96 level 4 MFOs). More than half of these proteins harbored more than four MFOs of level 3, and more than half harbored more than two MFOs of level 4 (Figure. 5A and B). The top three most common MFO at the

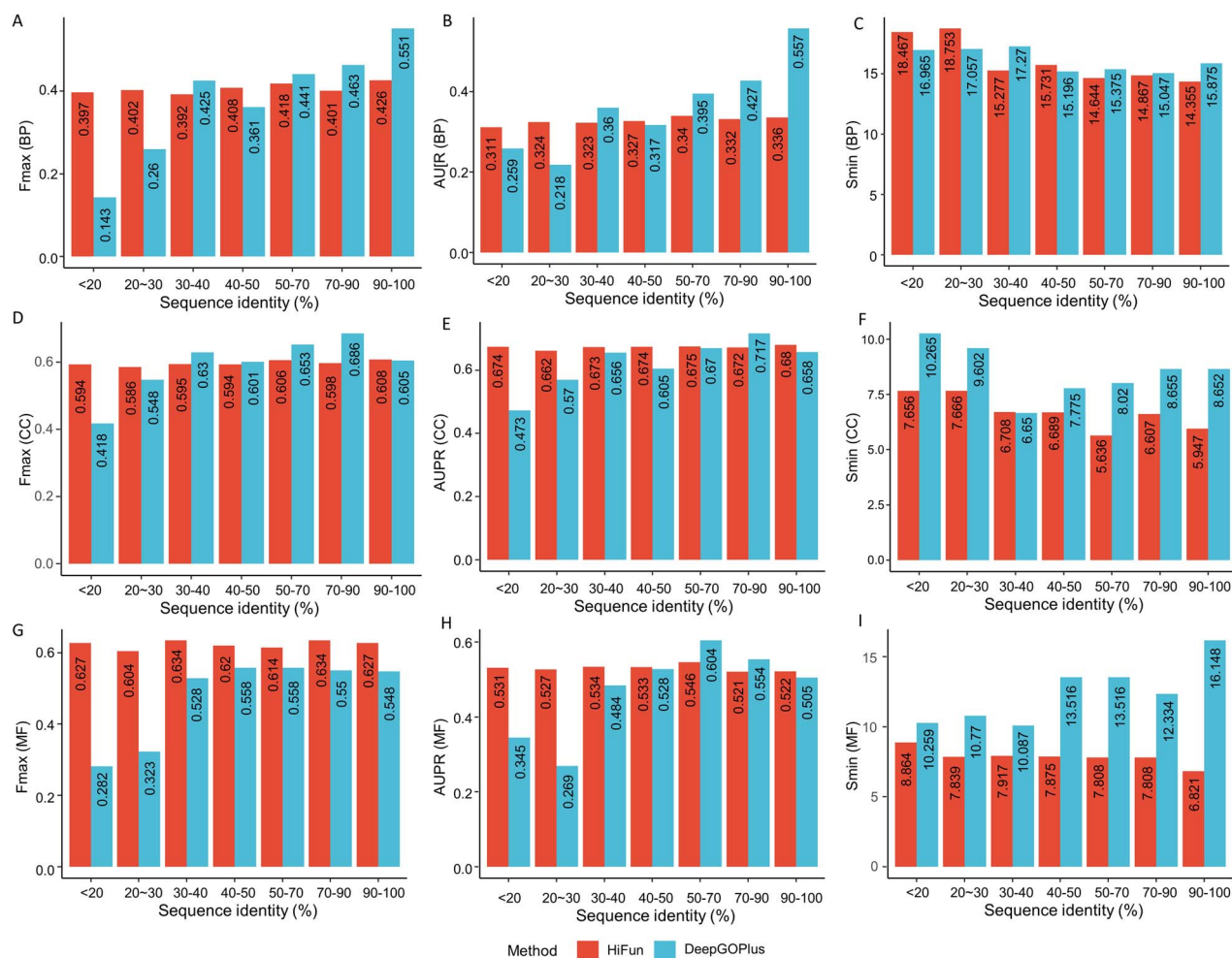


Figure 3. Performance of HiFun for the proteins with different sequence identity to the reference proteins. Bar plots showing the F_{max} , AUPR and S_{min} of HiFun and deepGOPlus respect to BPO (A-C), CCO (D-F) and MFO (G-I). The high values of F_{max} and AUPR indicated good performance, while low S_{min} means good performance.

third level are GO: 0043169 (cation binding), GO: 0003676 (nucleic acid binding) and GO: 0016788 (hydrolase activity, acting on ester bonds), while the top three most common MFOs at the fourth level are GO:0046872 (metal ion binding), GO:0003677 (DNA binding) and GO:0004518 (nuclease activity) (Figure 5C and D).

De novo motif discovery for novel protein families

Protein sequence motifs are one of the most important signatures of protein families and can often be used as tools for the protein function prediction. From the above, we have uncovered mounts of proteins that cannot be annotated with the HMM-based egnog-mapper tool [46, 47], which indicated that it could contain many novel motifs for the protein families. To address it, we took GO:0016628 (oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor, which was reported to be associated with the toxicity of antiviral drug sorivudine [48, 49]) for example and tended to uncover novel motifs associated with GO:0016628. We randomly selected one novel protein (*GUT_GENOME201382_01770*), which was annotated as GO:0016628 by HiFun, from the UHGP-50 dataset and search its analogous with sequence identity >62% (referring to the criterion used to construct the BLOSUM62 matrix) from the UHGP-

100 dataset, and finally 46 analogous proteins were retrieved. Additionally, we extracted 1524 the proteins annotated as GO:0016628 from Swiss-Prot database and grouped them into 166 blocks in which the proteins are at least 62% identical (referring to the criterion used to construct the BLOSUM62 matrix). After that, we selected one block with 49 Swiss proteins as well as the 47 UHGP proteins (*GUT_GENOME201382_01770* and 46 analogous proteins in the UHGP-100 dataset) to perform the phylogenetic analysis and de novo motif discovery.

The phylogenetic analysis results showed that the 46 novel UHGP proteins and 49 Swiss proteins are clearly divided into two families, indicating that these 46 novel UHGP proteins formed a novel protein family related to GO:0016628 (Figure 6A). Using MEME tool [35], 12 motifs were detected from the 46 novel UHGP proteins and 13 motifs were detected from the 49 Swiss proteins (Figure 6B, see Materials and Methods). Through comparison, we found that two motifs detected in the novel protein family were significantly similar to the motifs detected in the Swiss-Prot proteins (E-value <0.05, Figure 6C, see Materials and Methods). We also compared the motifs detected in the novel proteins against all the available motifs in the PROSITE database [50], and three motifs were observed to be significantly similar to the PROSITE motifs (E-value <0.05, Figure 6D) and the other seven motifs could be novel motifs.

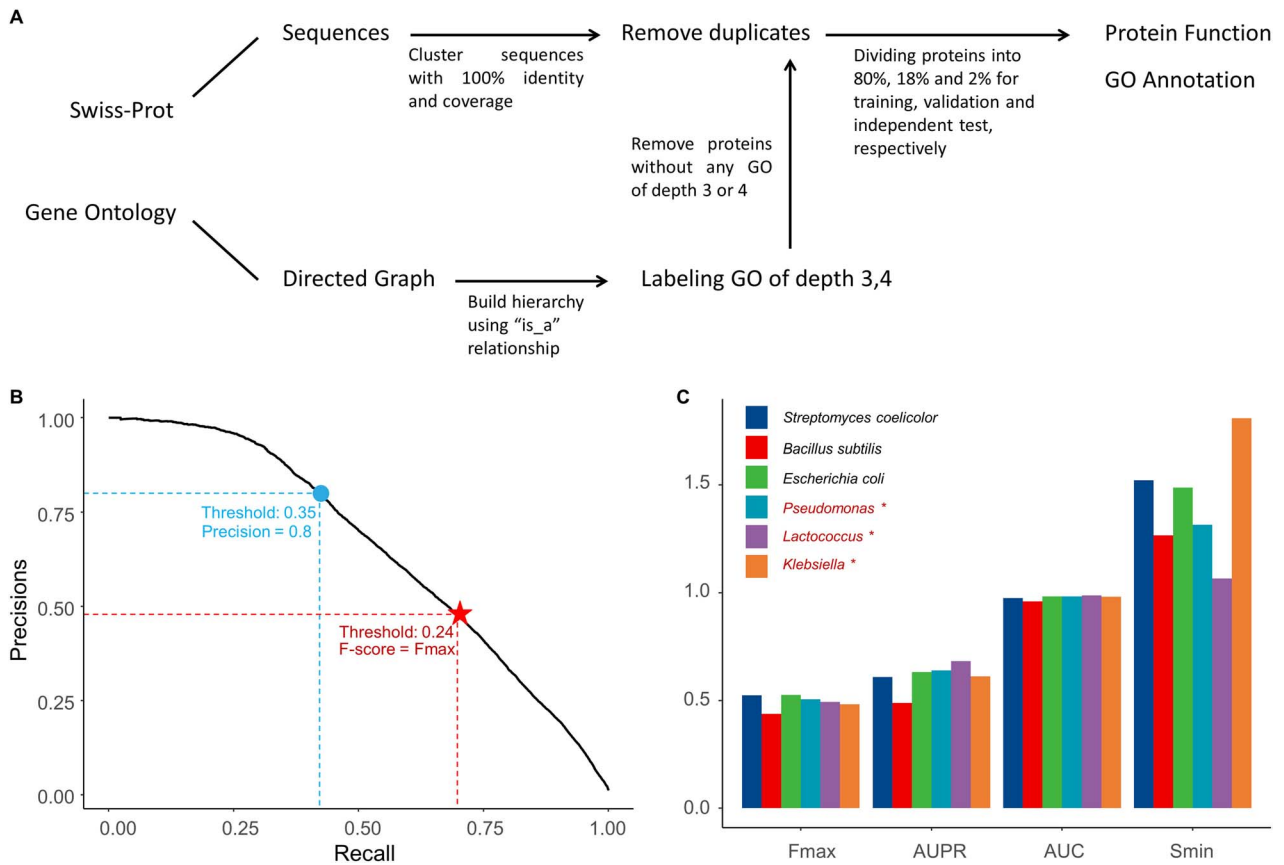


Figure 4. Build HiFun using Swiss-Prot bacterial proteins. **(A)** Removing duplicated proteins and GO propagating. **(B)** Precision-Recall curve of HiFun on the independent test data. The point marker indicated when set the threshold of HiFun as 0.35 can lead the HiFun with 80% precision. The pentagram marker indicated when set the threshold as 0.24 the F-score can reach the maximum. **(C)** Performance of HiFun between model and non-model bacteria respect to MFO. The bacteria with asterisk next to they names are non-model.

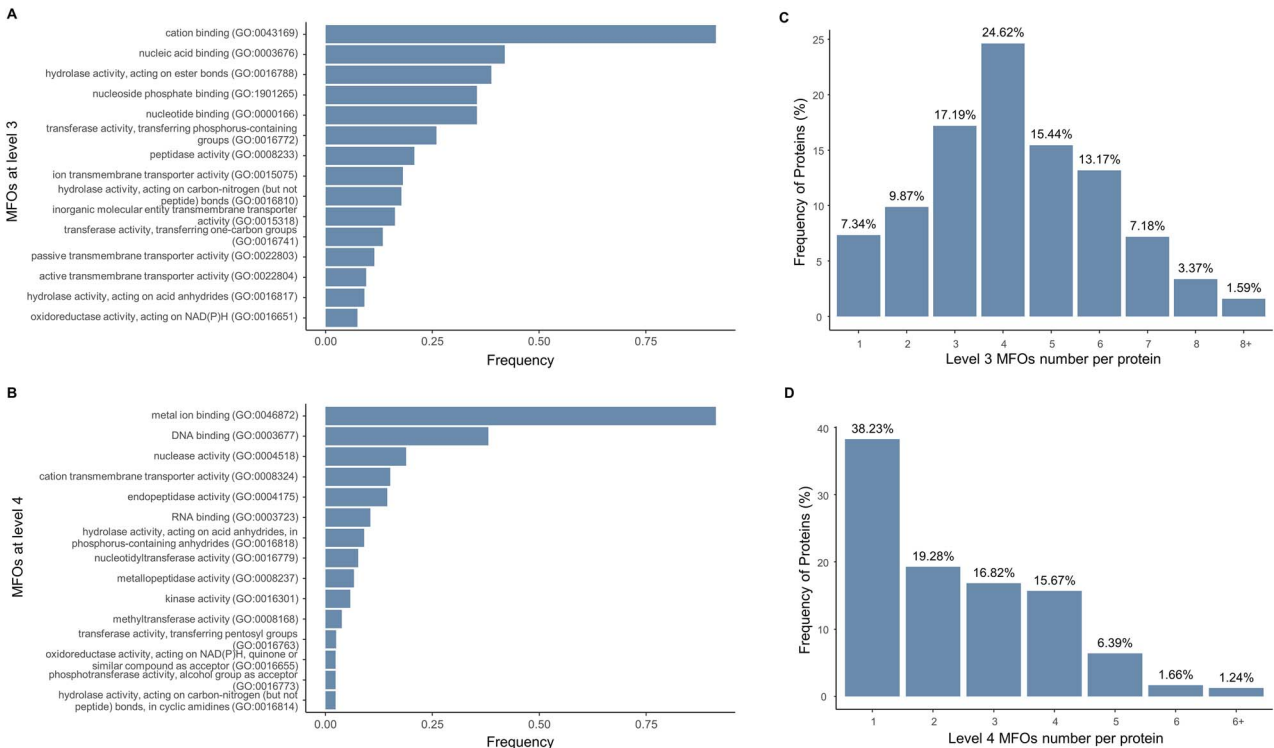


Figure 5. Functions of unknown proteins in the UHGP-50 catalog. **(A-B)** Top 15 MFOs of unknown proteins predicted by HiFun at level 3 and level 4, respectively. **(C-D)** The number of level 3 and level 4 MFOs harbored by each unknown protein.

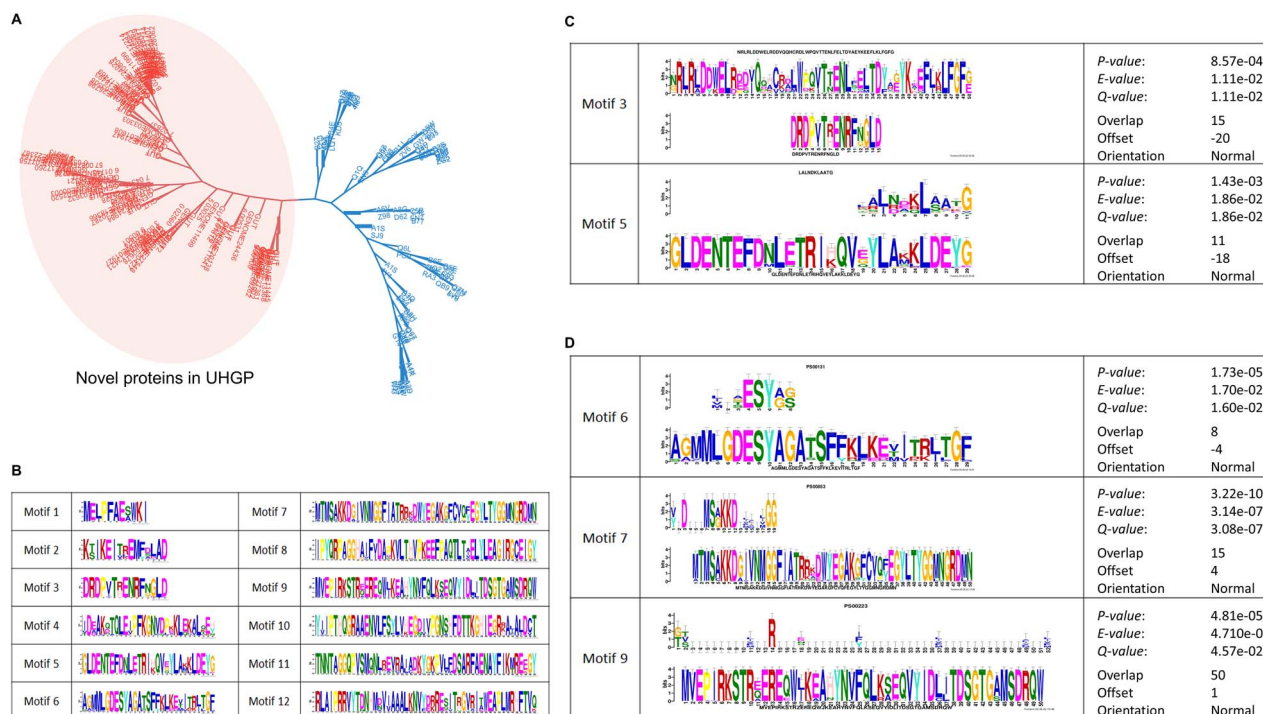


Figure 6. De novo motifs discovery for proteins annotated as GO:0016628. (A) Phylogenetic tree of 45 novel proteins (colored with red) as well as 49 Swiss-Prot proteins (colored with blue). (B) Twelve motifs detected from the 45 novel UHGP proteins. (C) Comparison of the motifs between novel protein family and Swiss-Prot protein family. (D) Comparison of the motifs between novel protein family and motifs in the PROSITE database.

Latent structure nature can be extracted by HiFun

Many proteins harbor the same function while their amino acid sequences were highly dissimilar, which limit the application of commonly used homology-based protein function prediction methods. Protein structure dictating biological function of protein is a highly reliable feature for protein function prediction [21]. Recently, deep learning-based approaches for protein structure prediction, exemplified by AlphaFold2 and RoseTTAFold, have generated shock waves in the structural biology community. However, these methods usually yield very low confidence for the proteins without available similar local patterns (in the absence of homologs) or well-defined structural packing [51, 52].

To further demonstrate the utility of HiFun, we took the unknown proteins annotated as GO:0046872 for example. Considering the median length of the Swiss-prot proteins annotated as GO:0046872 was 342 AAs, we randomly selected 200 unknown UHGP proteins ranged in length between 300 and 400 AAs and annotated as GO:0046872 by HiFun with probability larger than 0.2, 0.4, 0.6 and 0.8, respectively. The pair-wise sequence alignment with BLASTP showed that, only 2.43% of these protein pairs (483 in C_{200}^2) showed E-value <10, which means that these proteins were extremely dissimilar to each other respect to the sequence. It should be emphasized that these 200 proteins cannot be annotated with the traditional HMM-based method, indicating that these proteins do not have detectable motifs and homolog to known proteins, which may lead the low confidence of the structure prediction.

The structures of these 200 proteins were predicted by Robetta server with RoseTTAFold method, and the results showed that only 18% of these unknown UHGP proteins (36 in 200) can be given a structure with confidence >0.7 (Figure 7A). Even so, the pair-wise structure alignment among these proteins showed that TM-scores of 89.38% protein pairs were >0.2 (scores below 0.2

correspond to randomly chosen unrelated proteins [37]), which indicated that there may be similarities appearances among these unknown proteins to a certain extent.

To further explore the association between the structural similarity and sequence similarity, we focused on the protein structures with moderate confidence (≥ 0.5) and 2145 protein pairs were selected. We divided the protein-pairs into two groups and the protein-pairs with BLASTP E-value < 10 (default setting) were regarded as matched, otherwise were unmatched. From the matched protein-pairs, we observed that that structural similarity (TM-score) showing significantly positive correlation to the sequence similarity ($R=0.7$, P-value=6.13E-11, Figure 7B), and the TM-score of matched protein pairs were significantly higher than that of unmatched protein pairs (Wilcoxon test P-value=0.0025).

We further aligned these predicted structures to the X-ray structures of 1394 Swiss-Prot proteins annotated as GO:0046872 in the RCSB PDB database to validate its ability of function prediction. We only consider the RCSB structures with protein lengths larger 80 AAs as they typically have relatively simple topologies [53]. For each structure pair, the TM-score normalized by the smaller protein size are used. As the TM-score of two proteins > 0.5 assume generally the same fold in SCOP/CATH, we regarded two proteins with TM-score >0.5 as matched. The results showed that, without considering the confidence of Rosetta result, 51% of these unknown proteins were structurally related to at least one of the Swiss-prot protein annotated as GO:0046872, and two UHGP proteins were related with >75 Swiss-prot protein (Figure 7C). From the results, we observed two UHGP unknown proteins (GUT_GENOME189133_01419, GUT_GENOME283701_01607) with high-confidence (RoseTTAFold confidence ≥ 0.79) predicted structures harbored the most similar Swiss-prot proteins in term of structure, and the most structural similar proteins of these two unknown proteins were

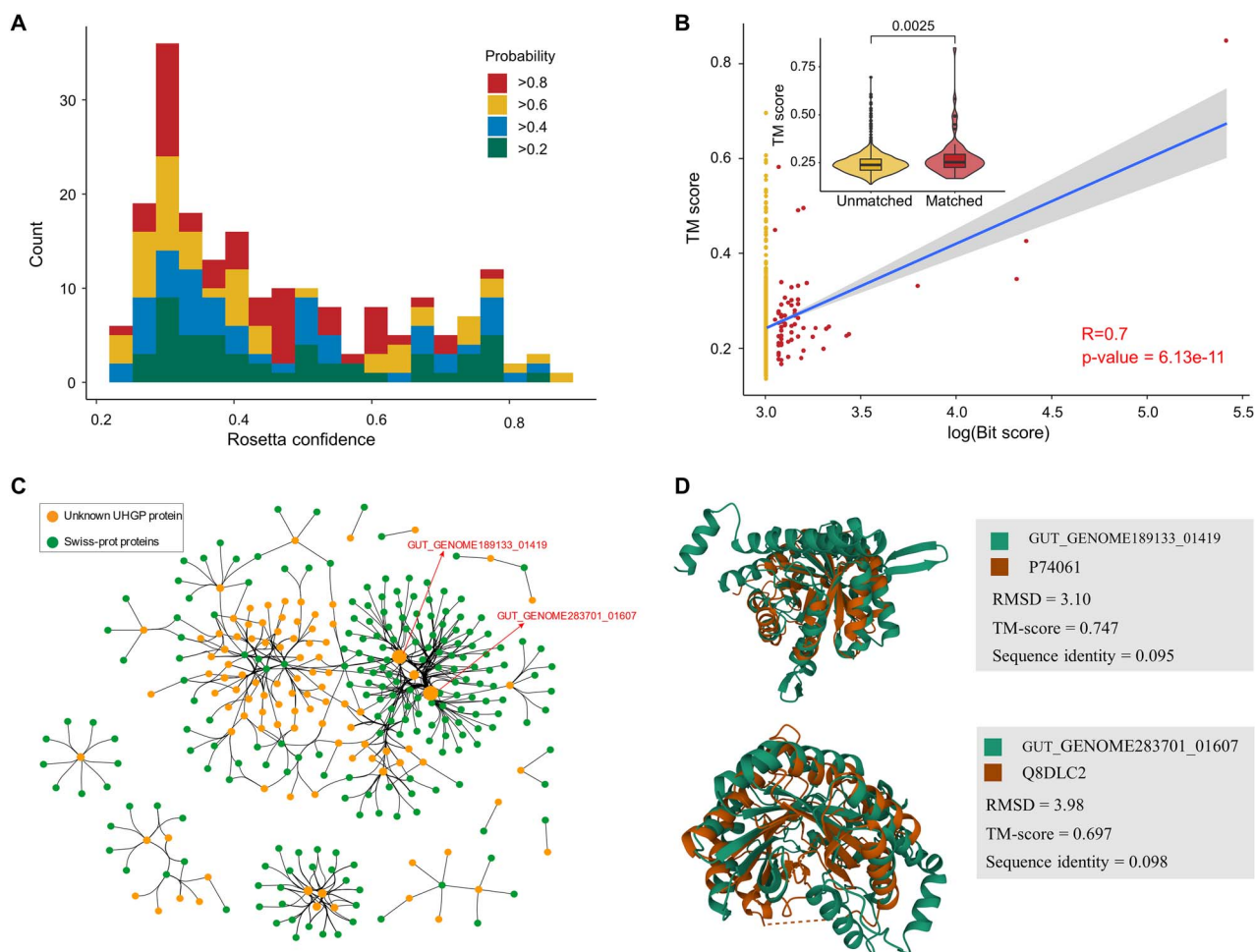


Figure 7. Structural alignment of UHGP unknown proteins. **(A)** Confidence distribution of structures predicted by Rosetta. The 200 proteins were selected according to the probability of HiFun. **(B)** Correlation between the TM-score and BLASTP bit score. Proteins with BLASTP *E*-value < 10 were regarded as matched and *R* is the Pearson correlation coefficient calculated according to the matched protein pairs. The Wilcoxon test method was used to measure the significance of difference between the TM-score of matched and unmatched protein pairs. **(C)** Structure alignment between the UHGP unknown proteins and the Swiss-prot proteins. The structure of UHGP unknown proteins were predicted with Rosetta and the Swiss-prot proteins' structures were downloaded from the RCSB PDB database. The orange nodes were the UHGP unknown proteins annotated as GO:0046872 by HiFun and the green nodes were the Swiss-prot proteins annotated as the same function. Proteins with TM-score > 0.5 were linked. **(D)** Structure alignment between two proteins with high confidence predicted structures and the matched Swiss-prot proteins with the highest TM-score.

Ribulose-phosphate 3-epimerase (Swiss-prot id: P74061, RCSB PDB id: 1TQJ, TM-score = 0.747) and Lipoyl synthase 2 (Swiss-prot id: Q8DLC2, RCSB PDB id: 4U00, TM-score = 0.697), respectively (Figure 7D).

DISCUSSION

With the rapid development of metagenomic/metatranscriptomic sequencing technologies, a massive amount of new microorganisms and proteins have been uncovered in various ecological environments which make them evolve specific features to adapt those unique niches. It was reported that up to 80% of these proteins in a microorganism show no similarity to proteins with known functions [45, 54, 55]. Although these unknown proteins vary significantly in their amino acid sequence, given that bacteria exist in a high variety of ecological niches, they naturally differ widely in their exact ecophysiology, but it can be anticipated that they would still carry out the similar general biological processes by using very similar molecular machinery. Therefore, deciphering the function of those novel proteins with computational approaches by simply inspecting their amino acid sequence

is one of the major challenges in the post-genomic era, especially when no homology information, genomic context or experimental resource available. Many of the classical protein function prediction methods have been developed to overcome the challenge but with limited success, because those methods normally extract the amino acid features or homology information based on proteins of known function, which cannot be applied to those unknown proteins with extremely low homology to known proteins.

Most recently, the success of several protein structure prediction models (e.g. AlphaFold2, RoseTTAFold and ProFold) indicates that the complex sequence-structure-function relationship can be extracted using deep learning model. Inspired by these, in this study, we developed a novel protein function prediction method HiFun, which is independent of homology information (sequence alignment free), by reassembling the protein sequences as protein language. The performance of HiFun was validated through comparing with the state-of-art methods using the CFA3 benchmark data. The structure alignment results also revealed that the proposed method can extract latent structure features from the protein sequence which empowers its ability to achieve

function annotation for non-homology proteins. Using HiFun, we further annotated 2 212 663 unknown proteins collected in the UHGP-50 database and showcased the utility of the HiFun framework in the expansion of novel protein family discovery, suggesting its great potential in real-world application. Through structure alignment analysis, we demonstrated that our method can acquire the relationship between protein function and informative internal representations extracted from protein sequence by learning the latent sequence features and the potential association between amino acids, which can overcome the limitations of homology-based approaches due to the lack of similar sequences with known function, or to misleading alignment results.

In our study, we observed that nearly half of the level 3 and level 4 MFOs had > 50 proteins, while only <30% MFOs of their child levels (5–7 levels) had > 50 proteins. In addition, as the task of protein function prediction is a multi-label classification, the number of labels to be predicted can highly affect the complexity and robustness of the model. Generally, the more classes to be predicted, the lower performance the model can achieve, which has also been observed in the results of CAFA3 challenge (e.g. the models always achieve lower performance for the BPOs prediction with greater number labels than predicting MFOs or CCOs) [33]. Therefore, taking the coverage of the proteins in training set, model complexity and the coverage of MFOs into account, in this study, we only focus on the molecular function of level 3 and level 4. Although it seems to be generic, our model can still provide important information for the experimental function validation. Moreover, our approach provided a complementary way to structure-based methods, such as DeepFRI, since the protein structures were usually hard to obtain (e.g. experimentally verified structure is limited and structure prediction models always take huge computational resource and is not applicable in normal laboratory). To facilitate the user's application, we also developed a user-friendly web server (<http://www.unimd.org/HiFun>).

Although HiFun could give us a sizable performance boost for the protein function prediction especially for the non-homology proteins, it is still worth considering additional investigations to further improve the model performance of HiFun and confirm the findings from this study: (i) In our study, we only constructed a model for Molecular Function prediction, the Biological Process and Cellular Component prediction can also be achieved in the future. (ii) The recent method mainly focused on the levels 3 and 4 of GO tree. In the future, we will expand the method to not focus on fixed level, but rather for each 'GO path' find specific and most informative depth. (iii) The number of proteins respected to different functions was highly imbalanced, and this could result in the degradation of performance for the functions with fewer protein sequences. Hence, a proper loss function or preprocessing strategy to balance the unbalance prevalence should be considered. (iv) In the current study, we showed a great potential to utilize AI-powered language models for protein function prediction. Besides the proposed model architecture in the HiFun framework (i.e. CNN + BiLSTM with self-attention mechanism), transformer-based language models such as Bidirectional Encoder Representations from Transformers and its derivatives may have a great potential for further improvement [56]. (v) Moreover, nearly half of the MFOs were excluded in our analysis as lacking sufficient protein sequences in those labels for training, these orphan MFOs could be involved in some environmental specific functions. Hence, uncovering proteins with these orphan MFOs could be one of the further important tasks.

Key Points

- The architecture of HiFun mainly consists of two parts. The first part is a BLUSOM62-based embedding unit followed by a convolutional neural network (CNN) module, which was used to learning the evolutionary features of the queried proteins. For the second part, we pretrained a protein language model to convert the protein sequences into numerical matrices to capture the nature of protein sequence beyond homology, and then the embedding results were fed into a sub-architecture constituted by connecting CNN and Bidirectional Long Short-Term Memory (BiLSTM) with a Self-Attention mechanism in series. The outputs of these two parts were concatenated and then passed through a softmax layer to generate the final prediction.
- The performance of HiFun was evaluated using the benchmark datasets and metrics in the CAFA3 challenge, and the results revealed great improvement over state-of-the-art methods. And most specially, for the proteins with extremely low identity to the proteins in reference dataset, HiFun can provide more robust performance than the homology-based method - deepGOplus.
- We established a HiFun model for the bacterial protein function prediction with 223 991 non-redundant Swiss-Prot bacterial proteins involving 334 MFOs and demonstrated the utility of HiFun by expanding the annotation of 2 218 032 non-redundant proteins that cannot be annotated with traditional methods in the UHGP-50 catalogue. About 99.76% of these unknown proteins can be annotated with HiFun.
- To demonstrate the ability of capturing latent structure nature, we applied RoseTTAFold to predict the 3D structure of 200 randomly selected unknown UHGP proteins. As these unknown proteins with no clear homolog to current annotated proteins, only 18% of these unknown UHGP proteins (36 in 200) can be built a structure with confidence >0.7. Through comparing these predicted structures with the X-ray structures of 1394 Swiss-Prot proteins with the same molecular function annotation in the RCSB PDB database, we found that 51% of these unknown proteins were structurally related to at least one of the Swiss-prot proteins.

ACKNOWLEDGEMENTS

T.S. and J.W. conceived the study. J.W., H.Q., Z.G., J.O., and J.Z. collected data and conducted data pre-processing. J.W. and H.Q. built the model and performed data analysis. J.O., H.Q., and J.W. developed the web server and software. T.S., J.W., Z.L., and H.Q. wrote the manuscript with. T.S., C.E.M, J.W., and Z.L. revised the manuscript. All authors read and approved the final manuscript.

FUNDING

Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), the Fundamental Research Funds for the Central Universities and the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU.

DATA AVAILABILITY STATEMENT

The HiFun model development source code is available at <https://github.com/Junwu302/HiFun>. A free and accessible webservice accessible at <http://www.unimd.org/HiFun>, offering an alternative platform for researchers to predict protein functions. An unrestricted and easily accessible webservice, located at <http://www.unimd.org/HiFun>, provides researchers with an alternative platform for predicting protein functions. The annotation results of the unidentified proteins within the UHGP-50 database are available for download through the designated 'download' web directory on our web service.

REFERENCES

- Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. *Nature* 2000;**405**:823–6.
- Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 2020;**36**:422–9.
- Zhang XX, Liu ZB, Xu W, et al. Genomic insights into versatile lifestyle of three new bacterial candidate phyla. *Science China-Life Sciences* 2022;**65**:1547–62.
- Thakur A, Sharma A, Alajangi HK, et al. In pursuit of next-generation therapeutics: antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications. *Int J Biol Macromol* 2022;**218**:135–56.
- Chamoli T, Khera A, Sharma A, et al. Peptide utility (PU) search server: a new tool for peptide sequence search from multiple databases. *Heliyon* 2022;**8**:e12283.
- Kim DI, Han SH, Park H, et al. Pseudo-isolated alpha-helix platform for the recognition of deep and narrow targets. *J Am Chem Soc* 2022;**144**:15519–28.
- Zhang F, Song H, Zeng M, et al. DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics* 2019;**19**:1900019.
- Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* 2018;**46**:D493–6.
- UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
- Torres M, Yang HX, Romero AE, Paccanaro A. Protein function prediction for newly sequenced organisms. *Nature Machine Intelligence* 2021;**3**:1050–60.
- Boratyn GM, Camacho C, Cooper PS, et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 2013;**41**:W29–33.
- Ding SY, Li Y, Shi ZX, Yan SJ. A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile. *Biochimie* 2014;**97**:60–5.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29–37.
- Engelhardt BE, Jordan MI, Srouji JR, Brenner SE. Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Res* 2011;**21**:1969–80.
- Finn RD, Attwood TK, Babbitt PC, et al. InterPro in 2017–beyond protein family and domain annotations. *Nucleic Acids Res* 2017;**45**:D190–9.
- Chen RZ, Deng YW, Ding YL, et al. Rice functional genomics: decades' efforts and roads ahead. *Science China-Life Sciences* 2022;**65**:33–92.
- Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 2018;**34**:660–8.
- Lobley AE, Nugent T, Orengo CA, Jones DT. FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res* 2008;**36**:W297–302.
- Zhang CX, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res* 2017;**45**:W291–9.
- Gligorijevic V, Renfrew PD, Kosciolk T, et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**(1):3168.
- Maranga M, Szczerbiak P, Bezshapkin V, et al. Comprehensive functional annotation of metagenomes and microbial genomes using a deep learning-based method. *mSystems* 2023;**8**(2):e0117822.
- Cramer P. AlphaFold2 and the future of structural biology. *Nat Struct Mol Biol* 2021;**28**:704–5.
- Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
- Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
- Chen D, Tian X, Zhou B, Gao J. ProFold: protein fold classification with additional structural features and a novel ensemble classifier. *Biomed Res Int* 2016;**2016**:6802832–10.
- Xu J, McPartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell* 2021;**3**:601–9.
- Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**(15):e2016239118.
- Unsal S, Atas H, Albayrak M, et al. Learning functional properties of proteins with language models. *Nature Machine Intelligence* 2022;**4**:227–45.
- Hess M, Keul F, Goesele M, Hamacher K. Addressing inaccuracies in BLOSUM computation improves homology search performance. *Bmc Bioinformatics* 2016;**17**:189.
- Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch W, Kacprzyk J, Oja E, Zadrozny S, (eds). *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*. ICANN, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg; 2005;**3697**.
- Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**(2):318–27.
- Zhou NH, Jiang YX, Bergquist TR, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* 2019;**20**:244.
- Clark WT, Radivojac P. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* 2013;**29**:53–61.
- Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;**2**:28–36.
- Gupta S, Stamatoyannopoulos JA, Bailey TL, et al. Quantifying similarity between motifs. *Genome Biol* 2007;**8**:R24.
- Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;**33**:2302–9.

38. Sehnal D, Bittrich S, Deshpande M, et al. Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;**49**:W431–7.
39. Radivojac P, Clark WT, Oron TR, et al. A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**: 221–7.
40. Jiang YX, Oron TR, Clark WT, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 2016;**17**:184.
41. Lammens EM, Nickel PI, Lavigne R. Exploring the synthetic biology potential of bacteriophages for engineering non-model bacteria. *Nat Commun* 2020;**11**:5294.
42. Sberro H, Fremin BJ, Zlitni S, et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* 2019;**178**:1245–1259.e14.
43. Qin JJ, Li YR, Cai ZM, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60.
44. Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from human microbiome metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**:649–662.e20.
45. Nayfach S, Paez-Espino D, Call L, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 2021;**6**:960–70.
46. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**:D309–14.
47. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast genome-wide functional annotation through Orthology assignment by eggNOG-mapper. *Mol Biol Evol* 2017;**34**:2115–22.
48. Clarke G, Sandhu KV, Griffin BT, et al. Gut reactions: breaking down xenobiotic-microbiome interactions. *Pharmacol Rev* 2019;**71**:198–224.
49. Chen M, Wang J, Yang Y, et al. Redox-dependent regulation of end-binding protein 1 activity by glutathionylation. *Science China-Life Sciences* 2021;**64**:575–83.
50. Sigrist CJ, de Castro E, Cerutti L, et al. New and continuing developments at PROSITE. *Nucleic Acids Res* 2013;**41**: D344–7.
51. Jones DT, Thornton JM. The impact of AlphaFold2 one year on. *Nat Methods* 2022;**19**:15–20.
52. Bondarenko V, Wells MM, Chen Q, et al. Structures of highly flexible intracellular domain of human alpha7 nicotinic acetylcholine receptor. *Nat Commun* 2022;**13**:793.
53. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* 2010;**26**:889–95.
54. Nayfach S, Shi ZJ, Seshadri R, et al. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 2019;**568**:505–10.
55. Danko D, Bezdan D, Afshin EE, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* 2021;**184**:3376–3393.e17.
56. Liu ZC, Roberts RA, Lal-Nag M, et al. AI-based language models powering drug discovery and development. *Drug Discov Today* 2021;**26**:2593–607.