特约评述

DOI: 10.12211/2096-8280.2025-002

基于文本数据挖掘的蛋白功能预测：机遇与挑战

张成辛[1, 2]
([1] 中国科学院深圳先进技术研究院，深圳合成生物学创新研究院，中国科学院定量工程生物学重点实验室，广东 深圳 518055；[2] 密歇根大学计算医学与生物信息学院，美国 密歇根州 安娜堡 48109)

摘要：理解蛋白质的生物学功能是定量合成生物学成功的前提。然而，除了少数模式生物外，大多数生物中有许多蛋白质的功能尚未通过实验进行解析。因此，开发自动、准确的蛋白质功能预测算法尤为重要。近年来，以深度学习为代表的人工智能算法成为蛋白质生物信息学发展的主流。在蛋白质功能预测领域，深度学习尤为显著。例如，在最近几届国际蛋白质功能预测大赛（Critical Assessment of Function Annotation，CAFA）中，排名靠前的算法使用深度学习模型（主要是大语言模型）实现基于文本数据挖掘的蛋白质功能预测。具体而言，这些方法或直接利用从科学文献中提取的文本特征来预测基因本体（Gene Ontology，GO），或通过具有相似文献的模板蛋白质来预测 GO。尽管在开发更强大的深度学习模型用于基于文本挖掘的蛋白质功能注释方面已有大量研究，基于文本挖掘的蛋白质功能预测算法在处理科学文献数据时仍存在一些长期被忽视的问题。本文首先回顾了蛋白质功能注释中现有的方法和挑战：第一，大多数基于文本挖掘的蛋白质功能预测器仅使用由 UniProt 数据库管理员为目标蛋白手工收集的 PubMed 摘要，忽略了尚未被 UniProt 收录的文献；第二，几乎所有方法都只处理摘要，而忽略了 PubMed Central 和 Europe PMC 等数据库中可获得的更详尽的全文文献；第三，鲜有研究工作能自动区分低通量实验、高通量研究和计算预测等不同类别的科研文献，这大大增加了基于文本进行功能注释的难度。此外，本文还提出了利用人工智能最新发展的有前景的方法，以改进基于文本挖掘的蛋白质功能注释。这有助于开发下一代文本挖掘工具，针对性攻克文本数据处理的现有困难，以实现更准确的功能注释。

关键词：蛋白质；生物学功能；基因本体；文本数据挖掘；深度学习

中图分类号：Q816　　**文献标志码：**A

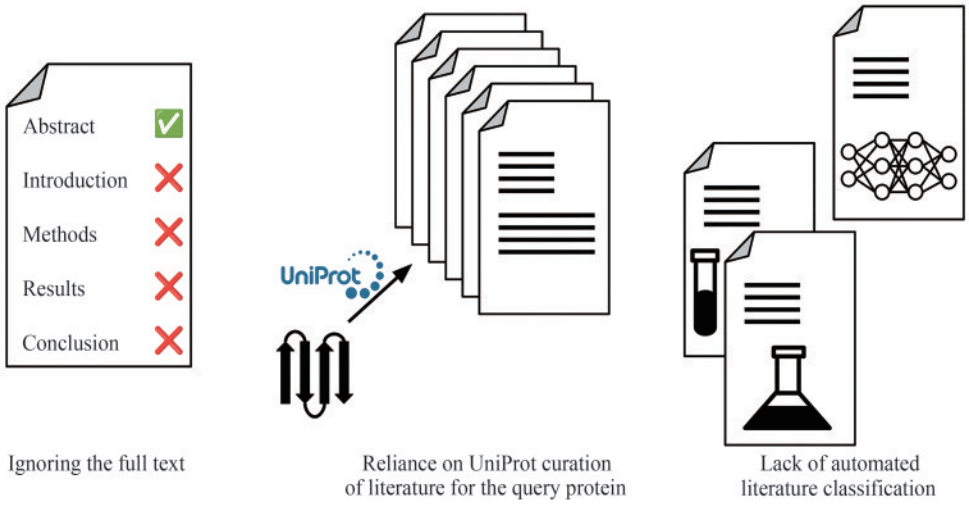
Challenges and opportunities in text mining-based protein function annotation

ZHANG Chengxin[1, 2]
([1]*CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China*; [2]*Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor 48109, Michigan, USA*)

Abstract: Understanding the biological function of proteins is crucial for advancing quantitative synthetic biology.

收稿日期：2025-01-02 修回日期：2025-03-04
引用本文：张成辛. 基于文本数据挖掘的蛋白功能预测:机遇与挑战[J]. 合成生物学, 2025, 6(3): 603-616
Citation: ZHANG Chengxin. Challenges and opportunities in text mining-based protein function annotation[J]. Synthetic Biology Journal, 2025, 6(3): 603-616

Except for a small number of model organisms, most species contain many proteins whose functions have not been experimentally verified, necessitating the development of accurate, automated protein function annotation methods. Recent progress in protein bioinformatics, particularly in predicting protein structures and functions, has been driven significantly by the application of artificial intelligence (AI) algorithms, with a notable emphasis on deep learning models. For instance, the top-ranked methods in recent Critical Assessment of Function Annotation (CAFA) challenge have used deep learning models, primarily large language models, to perform text mining-based protein function annotation. These methods either predict Gene Ontology (GO) terms directly from text features extracted from scientific literatures or from template proteins with databases. Despite the extensive work in developing increasingly powerful deep learning models for text mining-based protein function annotation, several major challenges have been overlooked when parsing scientific literature data. This manuscript reviews existing methods and challenges in protein function annotation. First, many text mining-based protein function predictors rely exclusively on PubMed abstracts collected by UniProt curators for the query protein, ignoring literatures that have not been reviewed by biocurators. Consequently, protein functions predicted by text mining might overlap with those from manual curation of the UniProt Gene Ontology Annotation. Second, nearly all methods only parse PubMed abstracts, ignoring the more informative full-text documents often available in the PubMed Central and Europe PMC repositories. Third, few studies have been proposed to automatically differentiate between different categories of literatures, such as low and high throughput experiments, and computational predictions. This manuscript also proposes promising approaches to enhance text mining-based protein function annotation using the latest development in AI, which is expected to contribute to the development of next-generation text mining tools for more accurate function annotation.



Keywords: proteins; biological functions; Gene Ontology (GO) terms; text mining; deep learning

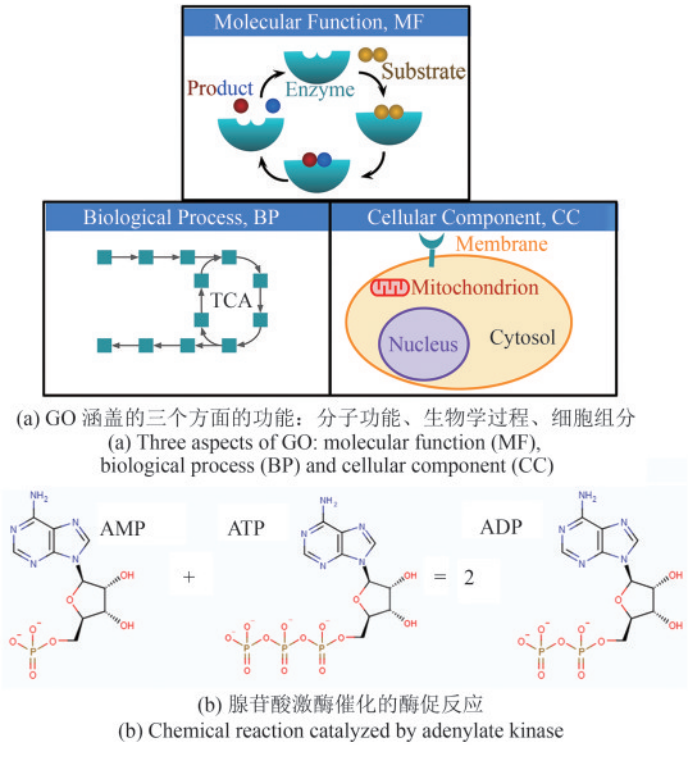
1 蛋白功能与基因本体（GO）注释

作为生物功能的直接执行者，承担催化、调控、转运和识别等功能。理解蛋白质功能是定量合成生物学成功的关键。为规范蛋白质功能的描述，多种分类体系被提出，包括基因本体（Gene Ontology，GO）[1]、酶学委员会（Enzyme Commission，EC）编号[2]、人类表型本体（Human Phenotype Ontology，HPO）[3]等。其中，GO是最常见与全面的体系［图1(a)］，涵盖生物过程

（biological process，BP）、细胞组成（cellular component，CC）和分子功能（molecular function，MF）三个方面。BP 描述生物通路或发育事件；CC 描述亚细胞定位或复合物形成；MF 描述蛋白质参与的分子事件，如酶促反应，因而常见的 EC 编号都有对应的 MF 项。GO 项以有向无环图（directed acyclic graph，DAG）组织，节点代表 GO 项，边表示从属关系。例如，腺苷酸激酶活性（GO：0004017 adenylate kinase activity，EC 编号 2.7.4.4）描述了催化 AMP 和 ATP 分子转化为 ADP 分子的酶促反应［图 1（b）］，该 GO 项在有向无环图中有 8 个父节点［图 1（c）］，包括单磷酸核苷激酶活性（GO：0050145 nucleoside monophosphate kinase activity），这是因为腺苷酸是单磷酸核苷的一种。

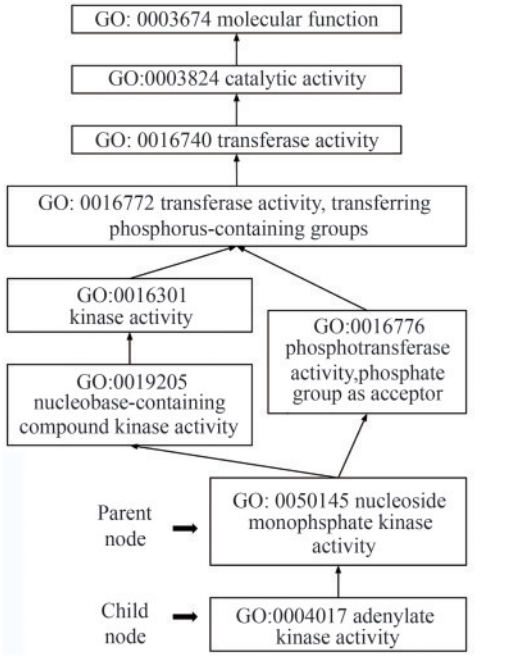
UniProt 数据库 [4] 对其蛋白质进行功能和文献的注释。首先，数据库管理员从 PubMed 等数据库中提取功能信息，以自由文本（free text）形式记录，并赋予 GO 注释 [5]。由于人工注释耗时耗力，许多文献的功能信息未能及时更新。例如，1999 年发表的 RasGap 蛋白质调控 Ras 蛋白质的酶活的文献 [6] 至今未在 UniProt 中有相应的功能注释。截止到 2024 年底，UniProt 收录 444 251 篇论文，仅 40.8%（181 306 篇）有相应的自由文本注释，40.5%（180 156 篇）有 GO 注释。同时，UniProt 已收录约 2.5 亿条蛋白质序列，仅约 57 万条有人工注释（图 2）。

为解决人工注释滞后问题，UniProt 引入功能预测算法作为补充。不同的注释方式对应不同的 GO 注释证据代码（Evidence code，https：//geneontology.org/docs/guide-go-evidence-codes/，表 1），这些代码分为低通量实验相关（EXP、IDA、IPI、IMP、IGI、IEP）、高通量实验相关（HTP、HDA、HMP、HGI、HEP）、计算预测（ISS、ISO、ISA、ISM、IGC、RCA）、系统进化树推断（IBA、IBD、IKR、IRD）[7] 等。其中 IEA 代码对应全自动预测，占 99.5%，错误率约 5%；基于系统发生树的注释占 0.3%，错误率高达 31%，实验相关注释错误率趋近于 0，但仅占 0.2% [8]，其余类型的注释占 0.2%。



(a) GO 涵盖的三个方面的功能：分子功能、生物学过程、细胞组分
(a) Three aspects of GO: molecular function (MF), biological process (BP) and cellular component (CC)

(b) 腺苷酸激酶催化的酶促反应
(b) Chemical reaction catalyzed by adenylate kinase



(c) 腺苷酸激酶活性对应的 GO 项及其父节点 GO 项构成的有向无环图
(c) Directed acyclic graph (DAG) consisting of the GO term for adenylate kinase activity and its parent GO terms

图 1　基因本体（GO）示意图

Fig. 1　Illustration of Gene Ontology (GO)

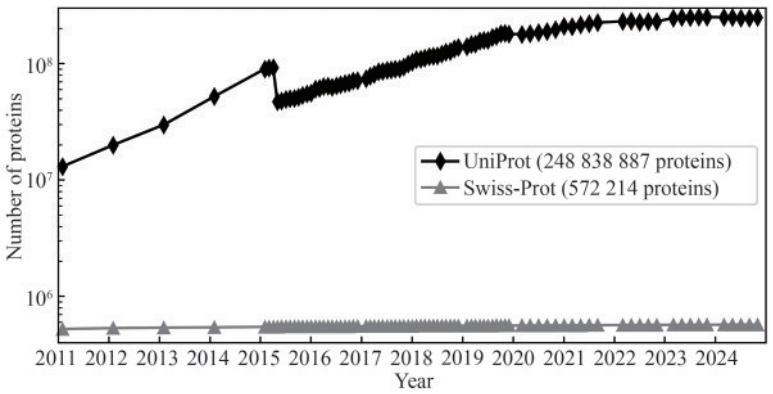


图2 UniProt与Swiss-Prot数据库收录的蛋白数目在过去14年间的增长情况

（在2015年，UniProt收录的蛋白数目有所下降，这是因为当时UniProt引入了一个主要针对微生物的冗余蛋白去除算法，具体而言，如果将两条序列高度相似的蛋白来自相同物种的不同菌株，则仅保留其中一条序列。）

Fig. 2 Accumulation of protein entries in the UniProt and Swiss-Prot databases within the past 14 years

(The drop in the number of UniProt proteins in 2015 was caused by the removal of redundant microbial proteins, *i.e.*, if two proteins are from different strains or isolates of the same species are almost identical, only one protein is kept.)

表1 GO注释证据代码

Table 1 Evidence codes used for Gene Ontology annotation

证据编码	详细解释
Inferred from Experiment（EXP）	实验验证的生物功能
Inferred from Direct Assay（IDA）	生物化学或细胞生物学实验验证的生物功能
Inferred from Physical Interaction（IPI）	实验验证的蛋白-蛋白、蛋白-核酸或蛋白-小分子配体相互作用
Inferred from Mutant Phenotype（IMP）	根据同一个基因的两个等位基因的功能差异推测的生物功能
Inferred from Genetic Interaction（IGI）	涉及两个或以上的基因的序列改变或者表达量改变的实验验证的生物功能
Inferred from Expression Pattern（IEP）	根据基因表达的位置或者基因表达时间推测的生物过程
Inferred from High Throughput Experiment（HTP）	高通量实验验证的生物功能
Inferred from High Throughput Direct Assay（HDA）	高通量生物化学实验或高通量细胞生物学实验验证的生物功能
Inferred from High Throughput Mutant Phenotype（HMP）	根据高通量实验中一个基因的两个等位基因的功能差异推测的生物功能
Inferred from Hight Throughput Genetic Interaction（HGI）	涉及两个或以上的基因的序列改变或者表达量改变的高通量实验验证的生物功能
Inferred from High Throughput Expression Pattern（HEP）	根据高通量实验中基因表达的位置或者基因表达时间推测的生物过程
Inferred from Sequence or Structural Similarity（ISS）	根据序列分析或者结构相似性预测并经过人工审核的生物功能
Inferred from Sequence Orthology（ISO）	根据直系同源关系预测并经过人工审核的生物功能
Inferred from Sequence Alignment（ISA）	根据序列比对预测的生物功能；功能预测与序列比对本身都经过人工审核
Inferred from Sequence Model（ISM）	基于隐马尔科夫模型（如Pfam）等蛋白家族的统计模型预测并经过人工审核的生物功能
Inferred from Genomic Context（IGC）	根据目标基因在基因组上邻近的其他基因元件预测并经过人工审核的生物功能
Inferred from Reviewed Computational Analysis（RCA）	根据大规模实验数据（如酵母双杂交、质谱、基因芯片）预测或者结合多种类型的数据预测并经过人工审核的生物功能
Inferred from Biological Aspect of Ancestor（IBA）	根据系统发生树中的先祖基因的功能推测的后代基因的生物功能
Inferred from Biological Aspect of Descendant（IBD）	根据系统发生树中的后代基因的功能推测的先祖基因的生物功能
Inferred from Key Residues（IKR）	根据关键氨基酸残基缺失推测的生物功能缺失
Inferred from Rapid Divergence（IRD）	根据后代基因与先祖基因在进化上的快速分歧推断的生物功能缺失
Traceable Author Statement（TAS）	根据综述文献或者实验文献的介绍或讨论章节中的引用文献总结的生物功能
Non-traceable Author Statement（NAS）	根据文献中没有明确实验依据或引用支持的文字描述总结的生物功能
Inferred by Curator（IC）	根据蛋白的已有功能注释推测的相关生物功能；例如，根据一个真核蛋白的已知功能“RNA聚合酶Ⅱ活性”推测该蛋白应具有功能注释“细胞核”
Inferred from Electronic Annotation（IEA）	无人工审核的计算预测得到的生物功能

2 蛋白功能预测与国际蛋白功能预测大赛（CAFA）

2.1 现有的蛋白功能预测算法

鉴于蛋白质功能注释，特别是GO注释对于生命科学研究的重要性以及蛋白质功能人工注释的稀缺性，准确高效的蛋白质功能预测算法显得尤为必要。于是，许多功能预测算法应运而生（表2）。这些算法的主流思路主要有两种：一是通过在数据库中搜索与目标蛋白质相似的模板蛋白质进行基于模板的功能预测；二是提取蛋白质的序列和/或结构特征，从头进行基于机器学习的功能预测，从而摆脱对模板的依赖。此外，还有一些混合算

表2 现有的蛋白功能预测方法

Table 2 Existing methods for protein function prediction

方法	功能预测的信息来源（特征）	机器学习模型
GOtcha、Blast2GO、BAR+	BLASTp搜索得到的同源序列	无
ConFunc、PFP、GoFDR	PSI-BLAST搜索得到的同源序列	无
HFSP	MMseqs2搜索得到的同源序列	无
ProFunc	BLASTp搜索得到的同源序列、SSM与Jess结构搜索得到的相似结构	无
COFACTOR	BLASTp与PSI-BLAST搜索得到的同源序列、TM-align结构搜索得到的相似结构、蛋白-蛋白互作	无
MetaGO	BLASTp与PSI-BLAST搜索得到的同源序列、TM-align结构搜索得到的相似结构、蛋白-蛋白互作	逻辑回归
StarFunc	BLASTp搜索得到的同源序列、Foldseek与TM-align结构搜索得到的相似结构、Pfam蛋白结构域家族、蛋白-蛋白互作、目标蛋白序列（ESM蛋白语言模型提取的特征）	逻辑回归、全连接神经网络、随机森林
DeepFRI、Struct2Go	三维结构提取的残基接触图、目标蛋白序列（独热编码）	图卷积神经网络
TALE-cmap	三维结构提取的残基接触图、多序列比对（ESM-MSA蛋白语言模型提取的特征）	Transformer
CLEAN-Contact	三维结构提取的残基接触图、目标蛋白序列（ESM蛋白语言模型提取的特征）	卷积神经网络
MS-kNN	同源序列、基因表达谱、蛋白-蛋白互作	*k*-最近邻
INGA	BLASTp搜索得到的同源序列、蛋白-蛋白互作、Pfam蛋白结构域家族	无
GOLabeler	BLASTp搜索得到的同源序列、InterPro蛋白结构域家族、目标蛋白序列（连续三个氨基酸残基序列片段的频率、ProFET程序提取的序列特征）	逻辑回归、梯度增强树
NetGO	BLASTp搜索得到的同源序列、InterPro蛋白结构域家族、蛋白-蛋白互作、目标蛋白序列（连续三个氨基酸残基序列片段的频率、ProFET程序提取的序列特征）	逻辑回归、梯度增强树
NetGO2.0	BLASTp搜索得到的同源序列、InterPro蛋白结构域家族、蛋白-蛋白互作、目标蛋白序列（连续三个氨基酸残基序列片段的频率、独热编码）、PubMed摘要	逻辑回归、双向长短期记忆神经网络、梯度增强树
DeepGO、DeepGOplus、ProteInfer、DeepEC、ECPICK	目标蛋白序列（独热编码）	卷积神经网络
ATGO+	BLASTp搜索得到的同源序列、目标蛋白序列（ESM蛋白语言模型提取的特征）	全连接神经网络
InterLabelGO+	DIAMOND搜索得到的同源序列、目标蛋白序列（ESM蛋白语言模型提取的特征）	全连接神经网络
DeepGO-SE	目标蛋白序列（ESM蛋白语言模型提取的特征）、蛋白-蛋白互作	全连接神经网络、图注意力网络
DeepECtransformer	DIAMOND搜索得到的同源序列、目标蛋白序列（ESM蛋白语言模型提取的特征）	注意力网络
CLEAN	目标蛋白序列（ESM蛋白语言模型提取的特征）	全连接神经网络

法将基于模板和从头预测的两种思路融合。

早期的蛋白质功能预测算法大多基于目标蛋白质与模板蛋白质的序列相似性。例如，GOtcha[9]、Blast2GO[10]和BAR-PLUS[11]等算法使用BLASTp[12]进行模板数据库的序列搜索，并将模板序列的功能注释转移到目标蛋白质上。ConFunc[13]、PFP[14]、GoFDR[15]等算法原理类似，只是将BLASTp替换为更加敏感的PSI-BLAST[12]。HFSP[16]则使用速度更快的MMseqs2[17]进行序列搜索。此外，还有一些算法[18-20]使用速度更快的DIAMOND[21]替代BLASTp和PSI-BLAST。最近的基准测试研究表明[22]，如果搜索参数设置恰当，使用BLASTp、DIAMOND或MMseqs2对基于序列相似性的功能预测算法的精度基本相当。

除了基于序列相似性的模板搜索，也使用基于结构相似性的模板搜索。例如，COFACTOR算法[23]及其后续版本MetaGO[24]利用TM-align结构比对程序[25]，对目标蛋白质的三维结构进行BioLiP蛋白质结构数据库[26]的模板搜索，从结构模板获取的功能信息再与BLASTp/PSI-BLAST搜索到的相似序列模板以及与目标蛋白质互作的蛋白质的功能信息结合，从而进行最终的功能预测。ProFunc[27]则利用SSM[28]和Jess[29]程序分别进行全局与局部的模板结构比对。而StarFunc算法[30]先使用Foldseek算法[31]对BioLiP结构数据库[26]与AlphaFold数据库[32]进行快速相似模板结构筛选，再使用TM-align进行更精细的结构比对，将得到的结构模板功能信息与序列模板、Pfam蛋白质结构域家族[33]、蛋白质互作网络以及InterLabelGO深度学习模型[34]相结合，通过随机森林（random forest）算法得到最终的功能预测。

目标蛋白质的三维结构不仅可用于结构模板的搜索，还可以用于提取氨基酸残基之间的相互接触图（contact map），并将其输入深度学习模型进行不依赖模板的从头功能预测。例如，早期的基于结构的GO预测深度学习模型DeepFRI[35]以及后来提出的Struct2Go[36]和TALE-cmap[37]都采用这一思路。此外，近期开发的CLEAN-Contact算法[38]也应用相似的思路进行EC编号的预测。

除了序列和结构信息，蛋白-蛋白互作、基因表达谱、蛋白质结构域家族等信息也被用于蛋白质功能预测。例如，MS-kNN算法[39]将相似序列、基因表达谱和蛋白-蛋白互作信息在一个*k*-最近邻（*k*-nearest neighbor）框架中进行处理。INGA算法[40]囊括了相似序列、蛋白-蛋白互作和Pfam数据库[33]的结构域家族信息。GOLabeler算法[41]将InterPro数据库[42]的蛋白质结构域家族、连续三个氨基酸残基序列片段的频率、同源序列等信息通过梯度增强决策树（gradient boosted decision tree，GBDT）[43]融合得到最终的GO预测。其后续版本NetGO[44]和NetGO2.0[45]分别增加了蛋白-蛋白互作和文本数据挖掘信息。

除了上述的基于随机森林、*k*-最近邻和梯度增强树等传统机器学习算法外，单纯从目标蛋白质序列出发的深度学习模型在蛋白质功能预测领域的应用日益受到重视。早期的用于GO预测的深度学习模型，包括DeepGO[46]、DeepGOPlus[18]和ProteInfer[47]，使用独热编码（one-hot encoding）将目标蛋白质的氨基酸序列（或三个氨基酸的片段）转化为数字特征，再将其输入到卷积神经网络（convolutional neural network，CNN）处理，并最终通过全连接神经网络（fully connected neural network）进行最终的功能预测。早期的用于预测EC的深度学习模型，如ProteInfer[47]、DeepEC[48]和ECPICK[49]，也使用相似的独热编码与CNN结合的方式。而近年来开发的新一代深度学习模型，包括ATGO+[50]、DeepGO-SE[51]和InterLabelGO[34]等GO预测模型以及DeepECtransformer[52]和CLEAN[53]等EC预测模型，不再使用独热编码生成的序列特征，而是采用ESM[54]和ProtT5[55]等基于Transformer架构[56]的蛋白质大语言模型（protein language model，PLM）进行蛋白质序列特征提取。PLM特征经过平均池化（mean pooling）和全连接网络，得到最终的功能预测。

近年来，无论是基于模板蛋白搜索的功能预测算法，还是基于机器学习的从头预测算法，其发展都得益于以AlphaFold和ChatGPT为代表的人工智能领域的重大突破。例如，StarFunc算法使用AlphaFold2预测的蛋白质结构作为功能预测的模板；而Struct2Go和CLEAN-Contact算法则直接将AlphaFold2预测的蛋白质结构作为训练数据，用于构建基于结构的功能预测模型。此外，以ChatGPT为代表的大语言模型（Large Language Model）的

开发，也推动了生物信息学领域蛋白质语言模型的进步，使得 ESM 等蛋白质语言模型成为蛋白质序列特征提取的主流工具。

2.2　国际蛋白功能预测大赛（Critical Assessment of Function Annotation，CAFA）

由于大量蛋白质功能预测算法的开发，对其预测精度进行客观公正的评估显得尤为重要。为此，美国爱荷华州立大学的 Iddo Friedberg 团队和印第安纳大学的 Predrag Radivojac 团队牵头组织了国际蛋白功能预测大赛，简称 CAFA [57]。从 2010—2011 年的 CAFA1 到 2023—2024 年的 CAFA5，CAFA 平均每三年组织一次，参与功能预测的团队逐年增加，从最初 CAFA1 的 30 个参赛团队到 CAFA5 的 1625 个团队，CAFA 已经成为蛋白质功能预测领域的顶级赛事，其地位相当于蛋白质结构预测领域的国际蛋白质结构预测大赛（Critical Assessment of Structure Prediction，CASP）。

CAFA 的开展利用了 UniProt 数据库中功能注释的自然增长现象，以下以 2016—2017 年开展的 CAFA3 [58] 为例讲解其具体大赛机制（图 3）。首先，CAFA 组织者在筹备阶段预先选定了 23 个模式生物和常见微生物的蛋白质组中的 130 827 个 Swiss-Prot 蛋白作为 CAFA3 目标蛋白，并于大赛开始日（2016 年 9 月）将其公布。68 个参赛组被要求对所有 130 827 个目标蛋白进行 GO 预测，并在截止日期（2017 年 2 月）之前将结果提交给 CAFA 组织者。截止日期之后直到 2017 年 11 月 UniProt GOA 对这 130 827 个蛋白中的 3328 个蛋白进行了新的基于实验文献的 GO 功能注释（证据代码：EXP、IDA、IPI、IMP、IGI、IEP、HTP、HDA、HMP、HGI、HEP、TAS、IC），而这 3328 个蛋白成为最终用于评估预测精度的基线测试集。

前文提到的一系列蛋白质功能预测算法中，BAR+、MS-kNN 是 CAFA1 [57] 中排名前列的算法；INGA 和 GoFDR 是 CAFA2 [59] 中排名前列的算法；GOLabeler 和 COFACTOR 是 CAFA3 [58] 中排名前列的算法；NetGO2.0 在 CAFA4 中排名第一；而 StarFunc、InterLabelGO 以及下文提到的 GOCurator [60] 和 PROTGOAT [61] 则是在 CAFA5 中排名前列的算法。

2.3　基于文本数据挖掘（text mining）的蛋白功能预算法

随着 CAFA 大赛的不断开展，文本数据挖掘算法的重要性日益明显，这些算法也随着机器学习领域对自然语言处理问题的深入研究而变得越发复杂与精细。比如说，CAFA1 中排名第一的算法 Jones-UCL 仅有简单的基于文本数据挖掘算法 [62]：通过统计具有不同功能的蛋白在 Swiss-Prot 中的描述文本中的单词频率，得到朴素贝叶斯（Naïve Bayes）模型，以进行 GO 预测。这种早期的模型并没有使用引用文献的文本数据，并且仅考虑单词的频率而忽略其上下文，因而有较大的局限性。

相较而言，CAFA4 中文本数据挖掘算法则更显成熟。比如，CAFA4 排名第一的算法 NetGO2.0 [45] 相较于其之前版本 GOLabeler [41] 和 NetGO [44]，引入了一个新的基于文本挖掘的功能预测模块 LR-Text［图 4（a）］，该模块借鉴了之前的工作 DeepText2GO [63]，具体原理如下：通过输入蛋白的 UniProt 编号找到该蛋白在 UniProt 数据库中相应的引用文献的 PubMed 编号，并通过 PubMed 数据库找到文献的标题和摘要，并从标题与摘要中用两种方法提取文本特征。第一种方法是词频-逆文档频（term frequency-inverse document frequency，TF-IDF），在 TF-IDF 中，单词 t 在文献 d 中的 TF-IDF 值见式（1）：

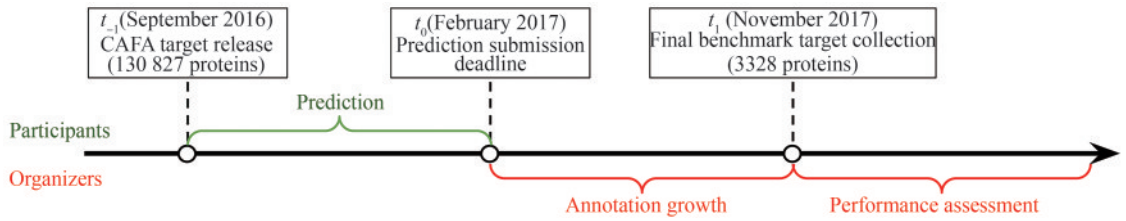
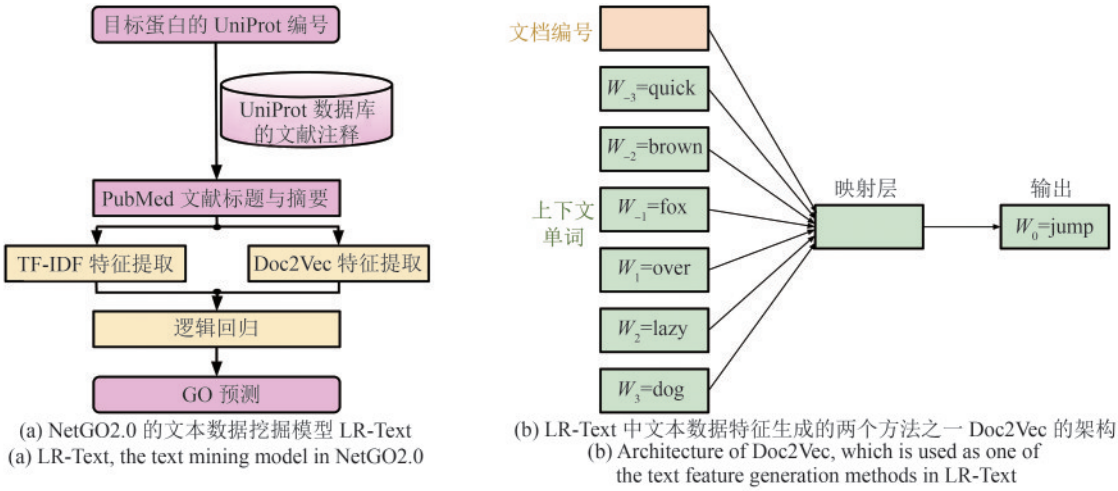


图 3　第三届国际蛋白功能预测大赛的时间节点

Fig. 3　Timeline of CAFA3



(a) NetGO2.0 的文本数据挖掘模型 LR-Text
(a) LR-Text, the text mining model in NetGO2.0

(b) LR-Text 中文本数据特征生成的两个方法之一 Doc2Vec 的架构
(b) Architecture of Doc2Vec, which is used as one of the text feature generation methods in LR-Text

图4 NetGO2.0中基于文本数据挖掘的GO预测

(在本例子中，Doc2Vec神经网络被训练预测在上下文为“The quick brown fox ___ over the lazy dog”中缺失的单词“jump”。原句中无意义的词“the”不包含在输入语句中)

Fig. 4 Text mining-based protein GO term prediction in NetGO2.0

(In this example, the Doc2Vec neural network model is trained to predict the masked word “jump” given its context in the sentence “The quick brown fox ___ over the lazy dog.” The word “the” is excluded from the input sentence as it does not have meaningful information.)

$$\mathrm{TFI\text{-}DF}(t,d)=\frac{f_{t,d}}{\sum_{t'\in d}f_{t',d}}\log\left(\frac{N}{N_t}\right) \quad (1)$$

式中，$f_{t,d}$是单词t在文献d中的频率；N是整个文献数据库的总文献数；N_t是整个文献数据库中含有单词t的文献数。

除了TF-IDF外，NetGO2.0还使用Doc2Vec[64]提取文本特征［图4(b)］。Doc2Vec是一个神经网络模型，训练这个网络的时候，输入文献的编号以及文献中一个单词的上下文，并通过神经网络的映射层，在输出层预测该单词；在应用该模型时候，只需要将PubMed文献的文本信息提供给输入层，就能将任意长度的文本变成映射层固定长度的特征。NetGO2.0把TF-IDF和Doc2Vec特征合并，用合并后的文本特征训练逻辑回归（logistic regression）模型，用以预测GO［图4(a)］，并将这些通过文本数据挖掘预测得到GO与通过BLASTp搜索到的相似序列、连续三个氨基酸残基片段的频率、InterPro蛋白结构域家族、蛋白-蛋白互作、深度神经网络共五种方法分别得到的GO相结合，使用梯度增强树得到最终的GO预测。

CAFA5中，排名第一和第四的方法GOCurator[60]和PROTGOAT[61]也使用了文本数据挖掘。其中，PROTGOAT对文本数据挖掘的使用相对简单，只使用了PubMed文献的TF-IDF特征。而GOCurator则是文本数据挖掘的集大成者，不但包含了其前身NetGO2.0中的LR-Text模型，还另外增加了三个文本数据挖掘模型：GOXML、LR-MEM和GORetriever。其中，GOXML只使用目标蛋白的PubMed文献标题与摘要作为输入数据，使用PubMedBERT大语言模型[65]进行文本特征提取，并使用AttentionXML多类别分类架构同时进行多个GO项的预测［图5(a)］。LR-MEM则是同时使用三种特征：PubMed文献的标题与摘要、目标蛋白在UniProt数据库中的文本描述以及氨基酸序列。其中，PubMed文献的特征使用文档水平的transformer模型SPECTER[66]进行特征提取，UniProt文本描述使用PubMedBERT进行特征提取，氨基酸序列特征使用ESM-1b蛋白语言模型[54]进行特征提取，三类特征合并后使用逻辑回归模型进行最终功能预测［图5(b)］。GOCurator中LR-Text、GOXML和LR-MEM三个文本挖掘模型都是不依赖模板蛋白的从头功能预测算法，而最后一个模型GORetriever则是一个利用模板蛋白信息的文本挖掘模型［图5(c)］。GORetriever首先获取目标蛋白引用的PubMed文献标题与摘要，并从中筛选出与蛋白功能相关的语句。同时，根据目标蛋白的UniProt文本描述，通过BM25算法搜索训练数据中具有相似文本描述的模板蛋白，

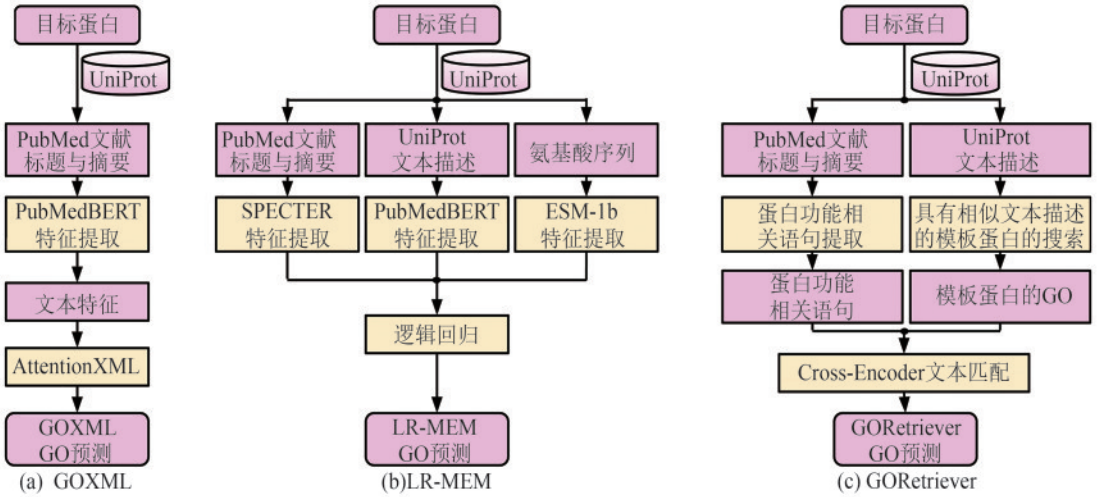


图5 GOCurator中三个基于文本数据挖掘的GO预测

Fig. 5 Three text mining-based models for protein GO term prediction in GOCurator

从而获取这些模板蛋白的GO注释以及相应的GO描述。目标蛋白PubMed文献中提取的功能相关语句和模板蛋白的GO描述进行基于Cross-Encoder[67]文本匹配，得到最终的GORetriever功能预测。

除了GO预测[68]外，文本数据挖掘也被用于预测其他类型的蛋白功能，比如蛋白-蛋白互作位点[69]、人类表型本体（HPO）[70]、代谢反应[71]、功能位点[72]等。

3 基于文本挖掘的功能预测的挑战

尽管文本数据挖掘在蛋白质功能预测，尤其是GO预测方面的作用已经在CAFA中得到充分展示，但仍有不少值得改进之处。在文献数据处理方面，这类方法目前至少存在以下不足：仅处理标题和摘要而忽略全文；完全依赖UniProt数据库管理员的文献注释；缺乏自动化的文献类型分类。

3.1 忽略文献全文

一篇研究一个蛋白或一组蛋白生物学功能的科研文献，通常包含标题（title）、摘要（abstract）、引言（introduction）、方法（methods）、结果（results）、讨论和总结（discussion and conclusion）部分。其中，引言部分往往包含对目标蛋白功能的过往研究认知，结果部分则是详细的实验或计算结果，讨论和总结部分可能包含对蛋白功能的概述及其可能具备的其他功能的推论。尽管这些章节提供了大量的功能信息，当前基于文本数据挖掘的蛋白质功能预测算法几乎都只处理标题和摘要。

实际上，获取PubMed摘要对应的全文的难度正在不断降低。随着开放存取（open access，OA）运动的推进，越来越多的文献作者在资助机构的要求下或主动选择将全文上传到PubMed Central和Europe PMC等全文数据库，或者预印本平台如bioRxiv和arXiv。例如，PubMed Central目前已经收录了约1 050万篇全文，占PubMed数据库所有文献的约28%。在UniProt于2024年收录的8935篇PubMed论文中，有约65%（5772篇）在PubMed Central拥有相应的全文，这意味着未来大部分PubMed论文都能找到相应的全文。此外，自2004年开始，PubMed Central也通过光学字符识别（optical character recognition，OCR）技术，陆续将146万篇仅有图片格式的影印版文献原文转化为纯文本格式，进一步方便了文本挖掘工具对这些文献的处理。因此，处理科研文献全文将成为未来基于文本挖掘的功能预测算法的一个突破方向。

相较于文献摘要的文本挖掘，文献全文的文本挖掘面临的一个主要挑战是有效信息的稀疏性。具体来说，文献摘要通常需要在数百单词的有限篇幅内对研究工作进行高度凝练的概括，因

此文本数据挖掘工具能够相对容易地定位摘要中关于目标蛋白功能的描述。然而，文献全文往往长达数千甚至上万单词，其中对蛋白功能的直接描述往往被大量的干扰信息所掩盖。在提取文献有效信息方面，基于大语言模型的聊天机器人（如ChatGPT和DeepSeek）可以提供一定程度的帮助。例如，可以通过以下指令："Summarize the biological functions of the protein discussed in the following text and propose Gene Ontology terms for its functions"，并结合需要处理的文献原文，引导大语言模型输出更为简洁的蛋白功能概述，并尝试生成相应的GO功能注释。

3.2 依赖UniProt的文献注释

目前几乎所有的基于文本挖掘的GO预测算法，其文本数据来源都是UniProt数据库对目标蛋白注释的PubMed文献。而UniProt的文献注释需要大量的人力资源，往往无法在文章发表后第一时间完成注释。例如，截至2024年底，UniProt数据库共收录了444 251篇文献，占整个PubMed数据库的约1.2%，其中8 935篇为2024年新收录的文献。这些新收录的文章中，并非最近三年发表的文献占60%，其中过半是发表十年以上的文献（图6）。这些数据说明UniProt数据库的文献收录具有明显的滞后性。因此，文本数据挖掘工具要充分利用已发表的文献，需要能够根据蛋白名、基因名、物种等关键词信息进行精确的PubMed和PubMed Central文献索引，而不能完全依赖UniProt的文献收录。

对一篇摘要或文献全文中所涉及的蛋白质与物种进行识别，本质上属于文本数据挖掘中的命名实体识别（named entity recognition， NER）问题，即从文本中提取具有特定意义的实体（如蛋白质名、基因名、物种名）。例如，对以下文献原文[73]进行基因名的命名实体识别"As an endogenous inhibitor of neutrophil adhesion， EDIL3 plays a crucial role in inflammatory regulation"，其正确识别结果为"As an endogenous inhibitor of neutrophil adhesion，［EDIL3］蛋白名 plays a crucial role in inflammatory regulation"。在蛋白质名与基因名的实体识别方面，目前已有一系列算法可供使用，例如EXTRACT[74]、PubTator[75]、HunFlair[76]、Saber[77]和OGER[78]等。这些算法可应用于基于文本挖掘的蛋白质功能注释研究中。

3.3 缺乏自动化的文献分类

一篇目标蛋白相关文献所报道的可能是低通

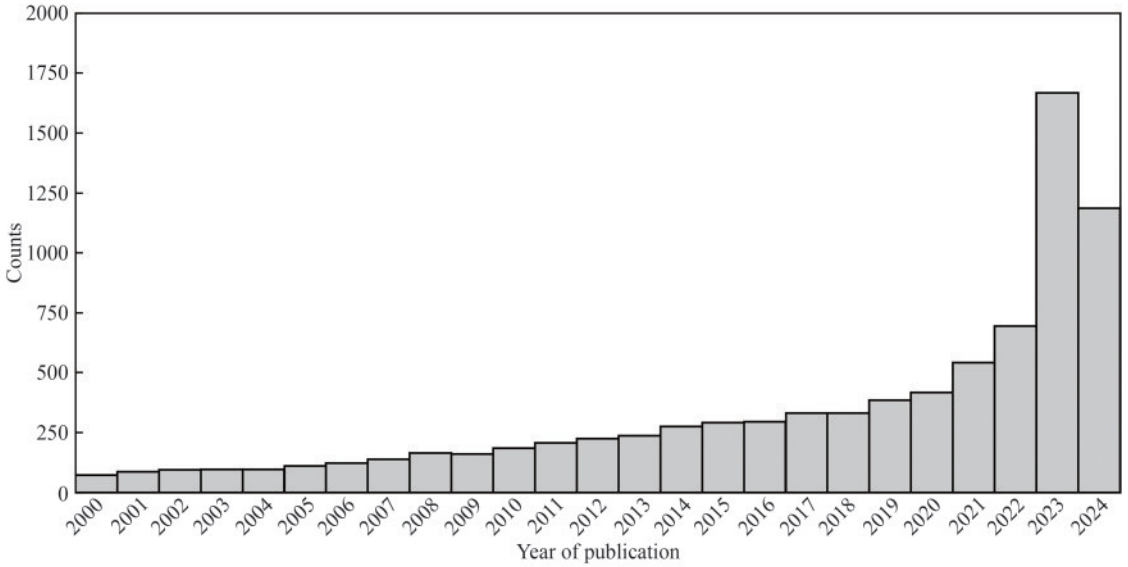


图6 UniProt数据库2024年收录的8935篇PubMed文献的发表年份

（1999年或之前发表的549篇文献没有在图中显示）

Fig. 6 Publication year for 8935 PubMed citations added to UniProt in 2024

(549 papers published in or before year 1999 are not shown)

量功能实验、结构解析工作、计算机预测、高通量实验、综述或方法学开发等不同类别的内容，而不同类型文献对蛋白功能预测的贡献显然不同。例如，UniProt数据库曾经为所有证据代码为IBA的GO注释引用了IBA算法对应的文献（PubMed：21873635）[7]，导致多达1 174 413个UniProt蛋白引用了该文献，直到2024年UniProt修正了这一引用。这类方法学开发相关的文献对基于文本挖掘的功能预测几乎没有帮助。因此，一个能够对科学文献进行基本类型分类，并对不同类别的文献按照其重要性自动加权的算法，将对后续的文本数据挖掘大有裨益。

4 结语与展望

ChatGPT和AlphaFold2的发布标志着生命科学研究已经进入人工智能时代，蛋白功能预测领域也不例外，而文本数据挖掘则是人工智能模型在蛋白功能预测中的一个重要应用。从早期基于朴素贝叶斯的统计模型，到基于Doc2Vec的经典神经网络，再到最近基于BERT架构的深度学习模型，基于文本数据挖掘的蛋白功能预测算法见证了从简单到复杂的发展历程，并充分利用了自然语言处理算法的突破，最近的CAFA5大赛更是凸显了文本数据挖掘在蛋白功能预测中的支配性地位。未来用于功能预测的文本挖掘算法将能够应用基于大语言模型的聊天机器人等最前沿的人工智能算法，更好地利用全文信息，实现更加自动和全面的文献检索，并能更好地对不同类型的文献进行自动分类和加权，从而实现更加准确的功能预测；同时这些算法的开发也能同步促进PubMed等文献数据库的文献分类，以及UniProt等蛋白数据库中将科研文献与相应蛋白相关联的工作，保障UniProt的人工功能注释过程能够高效及时地完成。

随着技术的进一步发展，未来的文本挖掘算法将不仅仅局限于对文献的检索、分类与功能注释，还将能够深入理解文献中的复杂语义信息。例如，通过结合知识图谱技术，算法可以自动构建蛋白与疾病、表型、药物代谢等多维关系的网络，从而为蛋白功能预测提供更加丰富的上下文信息。此外，基于多模态学习的方法也将成为未来的一个重要方向，通过整合文本数据与实验数据（如二代测序与质谱数据等），算法可以更全面地理解一个蛋白的功能机制及其与其他蛋白的协同关系。

在应用层面，这些先进的文本挖掘算法将极大地加速新药研发和个性化医疗的进程。例如，通过自动分析大量文献中蕴含的生物功能描述以及小分子-蛋白互作信息，算法可以为研究人员提供针对性药物的筛选建议并预测药物的药理、副作用和疗效。同时，这些算法还可以用于构建动态更新的蛋白功能数据库，为科研人员提供实时、精准的蛋白功能注释服务。

此外，随着人工智能技术的普及，未来的文本挖掘算法将更加注重用户友好性和可解释性。例如，通过开发交互式的可视化工具，研究人员可以直观地查看算法预测结果的直接依据，从而更好地理解蛋白功能的潜在机制以及实现该功能所需的生化环境。

总之，随着人工智能技术的不断进步，基于文本数据挖掘的蛋白功能预测算法将在未来发挥越来越重要的作用，不仅推动生命科学研究的深入发展，还将为人类健康和疾病治疗带来革命性的突破。

参 考 文 献

[1] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium[J]. Nature Genetics, 2000, 25(1): 25-29.

[2] International Union of Biochemistry, Nomenclature Committee. Enzyme nomenclature, 1978: recommendations of the Nomenclature Committee of the International Union of Biochemistry on the nomenclature and classification of enzymes [M]. New York: Academic Press, 1979.

[3] GARGANO M A, MATENTZOGLU N, COLEMAN B, et al. The human phenotype ontology in 2024: phenotypes around the world [J]. Nucleic Acids Research, 2024, 52(D1): D1333-D1346.

[4] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025[J]. Nucleic Acids Research, 2025, 53 (D1): D609–D617.

[5] HUNTLEY R P, SAWFORD T, MUTOWO-MEULLENET P, et al. The GOA database: gene Ontology annotation updates for 2015[J]. Nucleic Acids Research, 2015, 43(Database issue): D1057-D1063.

[6] FELDMANN P, EICHER E N, LEEVERS S J, et al. Control of

growth and differentiation by *Drosophila* RasGAP, a homolog of p120 ras-GTPase-activating protein[J]. Molecular and Cellular Biology, 1999, 19(3): 1928-1937.

[7] GAUDET P, LIVSTONE M S, LEWIS S E, et al. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium[J]. Briefings in Bioinformatics, 2011, 12(5): 449-462.

[8] WEI X Q, ZHANG C X, FREDDOLINO P L, et al. Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons[J]. Bioinformatics, 2020, 36(16): 4383-4388.

[9] MARTIN D M A, BERRIMAN M, BARTON G J. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes[J]. BMC Bioinformatics, 2004, 5: 178.

[10] CONESA A, GÖTZ S. Blast2GO: a comprehensive suite for functional analysis in plant genomics[J]. International Journal of Plant Genomics, 2008, 2008(1): 619832.

[11] PIOVESAN D, MARTELLI P L, FARISELLI P, et al. BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences[J]. Nucleic Acids Research, 2011, 39(Web Server issue): W197-W202.

[12] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25 (17): 3389-3402.

[13] WASS M N, STERNBERG M J E. ConFunc: functional annotation in the twilight zone[J]. Bioinformatics, 2008, 24(6): 798-806.

[14] HAWKINS T, CHITALE M, LUBAN S, et al. PFP Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data[J]. Proteins: Structure, Function, and Bioinformatics, 2009, 74(3): 566-582.

[15] GONG Q T, NING W, TIAN W D. GoFDR: a sequence alignment based method for predicting protein functions[J]. Methods, 2016, 93: 3-14.

[16] MAHLICH Y, STEINEGGER M, ROST B, et al. HFSP: high speed homology-driven function annotation of proteins[J]. Bioinformatics, 2018, 34(13): i304-i312.

[17] STEINEGGER M, SÖDING J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets[J]. Nature Biotechnology, 2017, 35(11): 1026-1028.

[18] KULMANOV M, HOEHNDORF R. DeepGOPlus: improved protein function prediction from sequence[J]. Bioinformatics, 2020, 36(2): 422-429.

[19] KULMANOV M, HOEHNDORF R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms[J]. Bioinformatics, 2022, 38(S1): i238-i245.

[20] YUAN Q M, XIE J J, XIE J C, et al. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion[J]. Briefings in Bioinformatics, 2023, 24(3): bbad117.

[21] BUCHFINK B, REUTER K, DROST H G. Sensitive protein alignments at tree-of-life scale using DIAMOND[J]. Nature Methods, 2021, 18(4): 366-368.

[22] ZHANG C X, LYDIA FREDDOLINO P. A large-scale assessment of sequence database search tools for homology-based protein function prediction[EB/OL]. bioRxiv, 2023: 2023.11.14.567021. (2023-11-16)[2024-12-01]. https://doi.org/10.1101/2023.11.14.567021.

[23] ZHANG C X, FREDDOLINO L, ZHANG Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information[J]. Nucleic Acids Research, 2017, 45(W1): W291-W299.

[24] ZHANG C X, ZHENG W, FREDDOLINO P L, et al. MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping[J]. Journal of Molecular Biology, 2018, 430 (15): 2256-2265.

[25] ZHANG Y, SKOLNICK J. TM-align: a protein structure alignment algorithm based on the TM-score[J]. Nucleic Acids Research, 2005, 33(7): 2302-2309.

[26] ZHANG C X, ZHANG X, FREDDOLINO L, et al. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions[J]. Nucleic Acids Research, 2024, 52(D1): D404-D412.

[27] LASKOWSKI R A. The ProFunc function prediction server[J]. Methods in Molecular Biology, 2017, 1611: 75-95.

[28] KRISSINEL E, HENRICK K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions[J]. Acta Crystallographica Section D, Biological Crystallography, 2004, 60(Pt 12 Pt 1): 2256-2268.

[29] BARKER J A, THORNTON J M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis[J]. Bioinformatics, 2003, 19(13): 1644-1649.

[30] ZHANG C X, LIU Q C, FREDDOLINO L. StarFunc: fusing template-based and deep learning approaches for accurate protein function prediction[EB/OL]. bioRxiv, 2024: 2024.05.15. 594113. (2024-05-18) [2024-12-01]. https://doi. org/10.1101/2024.05.15.594113.

[31] VAN KEMPEN M, KIM S S, TUMESCHEIT C, et al. Fast and accurate protein structure search with Foldseek[J]. Nature Biotechnology, 2024, 42(2): 243-246.

[32] VARADI M, ANYANGO S, DESHPANDE M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models [J]. Nucleic Acids Research, 2022, 50(D1): D439-D444.

[33] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: the protein families database in 2021[J]. Nucleic Acids Research, 2021, 49(D1): D412-D419.

[34] LIU Q C, ZHANG C X, FREDDOLINO L. InterLabelGO+: unraveling label correlations in protein function prediction[J]. Bioinformatics, 2024, 40(11): btae655.

[35] GLIGORIJEVIĆ V, RENFREW P D, KOSCIOLEK T, et al. Structure-based protein function prediction using graph convolutional networks[J]. Nature Communications, 2021, 12(1): 3168.

[36] MA W J, ZHANG S G, LI Z, et al. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures[J]. Journal of Chemical Information and Modeling, 2022, 62(17): 4008-4017.

[37] QIU X Y, WU H, SHAO J Y. TALE-cmap: protein function prediction based on a TALE-based architecture and the structure information from contact map[J]. Computers in Biology and Medicine, 2022, 149: 105938.

[38] YANG Y X, JERGER A, FENG S, et al. Improved enzyme functional annotation prediction using contrastive learning with structural inference[J]. Communications Biology, 2024, 7(1): 1690.

[39] LAN L, DJURIC N, GUO Y H, et al. MS-kNN: protein function prediction by integrating multiple data sources[J]. BMC Bioinformatics, 2013, 14(Suppl 3): S8.

[40] PIOVESAN D, TOSATTO S C E. INGA 2.0: improving protein function prediction for the dark proteome[J]. Nucleic Acids Research, 2019, 47(W1): W373-W378.

[41] YOU R H, ZHANG Z H, XIONG Y, et al. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank[J]. Bioinformatics, 2018, 34(14): 2465-2473.

[42] BLUM M, CHANG H Y, CHUGURANSKY S, et al. The InterPro protein families and domains database: 20 years on[J]. Nucleic Acids Research, 2021, 49(D1): D344-D354.

[43] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C/OL]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM, 2016: 785-794. (2016-08-13)[2024-12-01]. https://doi.org/10.1145/2939672.2939785.

[44] YOU R H, YAO S W, XIONG Y, et al. NetGO: improving large-scale protein function prediction with massive network information [J]. Nucleic Acids Research, 2019, 47(W1): W379-W387.

[45] YAO S W, YOU R H, WANG S J, et al. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information[J]. Nucleic Acids Research, 2021, 49(W1): W469-W475.

[46] KULMANOV M, KHAN M A, HOEHNDORF R, et al. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier[J]. Bioinformatics, 2018, 34(4): 660-668.

[47] SANDERSON T, BILESCHI M L, BELANGER D, et al. ProteInfer, deep neural networks for protein functional inference[J]. eLife, 2023, 12: e80942.

[48] RYU J Y, KIM H U, LEE S Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 13996-14001.

[49] HAN S R, PARK M, KOSARAJU S, et al. Evidential deep learning for trustworthy prediction of enzyme commission number [J]. Briefings in Bioinformatics, 2023, 25(1): bbad401.

[50] ZHU Y H, ZHANG C X, YU D J, et al. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction[J]. PLoS Computational Biology, 2022, 18(12): e1010793.

[51] KULMANOV M, GUZMÁN-VEGA F J, ROGGLI P D, et al. DeepGO-SE: protein function prediction as Approximate Semantic Entailment[EB/OL]. bioRxiv, 2023: 2023.09.26.559473. (2023-09-28)[2024-12-01]. https://doi.org/10.1101/2023.09.26.559473.

[52] KIM G B, KIM J Y, LEE J A, et al. Functional annotation of enzyme-encoding genes using deep learning with transformer layers[J]. Nature Communications, 2023, 14(1): 7370.

[53] YU T H, CUI H Y, LI J C, et al. Enzyme function prediction using contrastive learning[J]. Science, 2023, 379(6639): 1358-1363.

[54] LIN Z M, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[55] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: toward understanding the language of life through self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7112-7127.

[56] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C/OL]//Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017[2024-12-01]. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[57] RADIVOJAC P, CLARK W T, ORON T R, et al. A large-scale evaluation of computational protein function prediction[J]. Nature Methods, 2013, 10(3): 221-227.

[58] ZHOU N H, JIANG Y X, BERGQUIST T R, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens[J]. Genome Biology, 2019, 20(1): 244.

[59] JIANG Y X, ORON T R, CLARK W T, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy[J]. Genome Biology, 2016, 17(1): 184.

[60] YAN H Y, WANG S J, LIU H C, et al. GORetriever: reranking protein-description-based GO candidates by literature-driven

deep information retrieval for protein function annotation[J]. Bioinformatics, 2024, 40(S2): ii53-ii61.

[61] CHUA Z M, RAJESH A, SINHA S, et al. PROTGOAT: improved automated protein function predictions using Protein Language Models[EB/OL]. bioRxiv, 2024: 2024.04. 01.587572 .(2024-04-02) [2024-12-01]. https://doi.org/10.1101/2024.04.01.587572.

[62] COZZETTO D, BUCHAN D W, BRYSON K, et al. Protein function prediction by massive integration of evolutionary analyses and multiple data sources[J]. BMC Bioinformatics, 2013, 14(3): S1.

[63] YOU R H, HUANG X D, ZHU S F. DeepText2GO: improving large-scale protein function prediction with deep semantic text representation[J]. Methods, 2018, 145: 82-90.

[64] LE Q, MIKOLOV T. Distributed representations of sentences and documents; proceedings of the international conference on machine learning[C/OL]//Proceedings of the 31st International Conference on Machine Learning, PMLR, 2014, 32(2): 1188-1196[2024-12-04]. https://proceedings.mlr.press/v32/le14.html.

[65] GU Y, TINN R, CHENG H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare, 2021, 3(1): 1-23.

[66] COHAN A, FELDMAN S, BELTAGY I, et al. SPECTER: document-level representation learning using citation-informed transformers[EB/OL]. arXiv, 2020: 200407180. (2020-05-20) [2024-12-01]. https://doi.org/10.48550/arXiv.2004.07180.

[67] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[EB/OLJ]. arXiv, 2019: 190810084. (2019-08-27) [2024-12-01]. https://doi. org/10.48550/arXiv.1908.10084.

[68] WU J S, YIN Q, ZHANG C X, et al. Function prediction for G protein-coupled receptors through text mining and induction matrix completion[J]. ACS Omega, 2019, 4(2): 3045-3054.

[69] BADAL V D, KUNDROTAS P J, VAKSER I A. Text mining for protein docking[J]. PLoS Computational Biology, 2015, 11(12): e1004630.

[70] KAFKAS Ş, HOEHNDORF R. Ontology based text mining of gene-phenotype associations: application to candidate gene prediction[J]. Database, 2019, 2019: baz019.

[71] CZARNECKI J, NOBELI I, SMITH A M, et al. A text-mining system for extracting metabolic reactions from full-text articles [J]. BMC Bioinformatics, 2012, 13: 172.

[72] VERSPOOR K M, COHN J D, RAVIKUMAR K E, et al. Text mining improves prediction of protein functional sites[J]. PLoS One, 2012, 7(2): e32171.

[73] WEI X Q, ZOU S, XIE Z H, et al. EDIL3 deficiency ameliorates adverse cardiac remodelling by neutrophil extracellular traps (NET) -mediated macrophage polarization [J]. Cardiovascular Research, 2022, 118(9): 2179-2195.

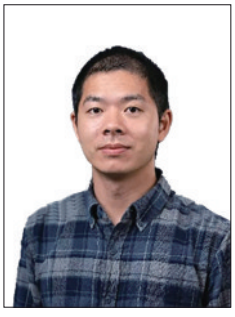
[74] PAFILIS E, BUTTIGIEG P L, FERRELL B, et al. EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation[J]. Database, 2016, 2016: baw005.

[75] WEI C H, KAO H Y, LU Z Y. PubTator: a web-based text mining tool for assisting biocuration[J]. Nucleic Acids Research, 2013, 41 (Web Server issue): W518-W522.

[76] WEBER L, SÄNGER M, MÜNCHMEYER J, et al. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition[J]. Bioinformatics, 2021, 37(17): 2792-2794.

[77] GIORGI J M, BADER G D. Towards reliable named entity recognition in the biomedical domain[J]. Bioinformatics, 2020, 36(1): 280-286.

[78] FURRER L, JANCSO A, COLIC N, et al. OGER++: hybrid multi-type entity recognition[J]. Journal of Cheminformatics, 2019, 11(1): 7.



第一作者及通讯作者：张成辛(1991—)，男，博士，研究员。研究方向为蛋白质与RNA的结构与功能预测。

E-mail：cx.zhang2@siat.ac.cn