特约评述

DOI: 10.12211/2096-8280.2022-078

深度学习在蛋白质功能预测中的应用

宋益东，袁乾沐，杨跃东
（中山大学计算机学院，广东　广州　510000）

摘要：蛋白质功能预测是生物信息学中的一项重要任务，在疾病机制的阐明和药物靶点发现等领域有着重要作用。因为传统的测定蛋白质功能的生化实验通常成本高、耗时长、通量低，所以开发出高效且准确的蛋白质功能预测计算方法十分重要。蛋白质功能预测可以分为残基水平的结合位点预测和蛋白水平的基因本体论（gene ontology，GO）预测。本文首先介绍该领域常用的数据库及蛋白质特征信息，接着对当下最新的蛋白质功能预测方法进行总结。在结合位点预测方面，根据配体类型分别介绍了最新的蛋白质-蛋白质、蛋白质-多肽、蛋白质-核酸和蛋白质-小分子或离子配体的结合位点预测方法；在GO预测方面，按照预测方法的类别分别介绍了最近的基于序列、基于结构和基于蛋白相互作用网络的方法。最后，对目前的蛋白质功能预测方法进行总结、分析优劣，并展望该领域未来的发展方向。

关键词：深度学习；蛋白质；功能预测；结合位点；基因本体论

中图分类号：Q71　　**文献标志码：**A

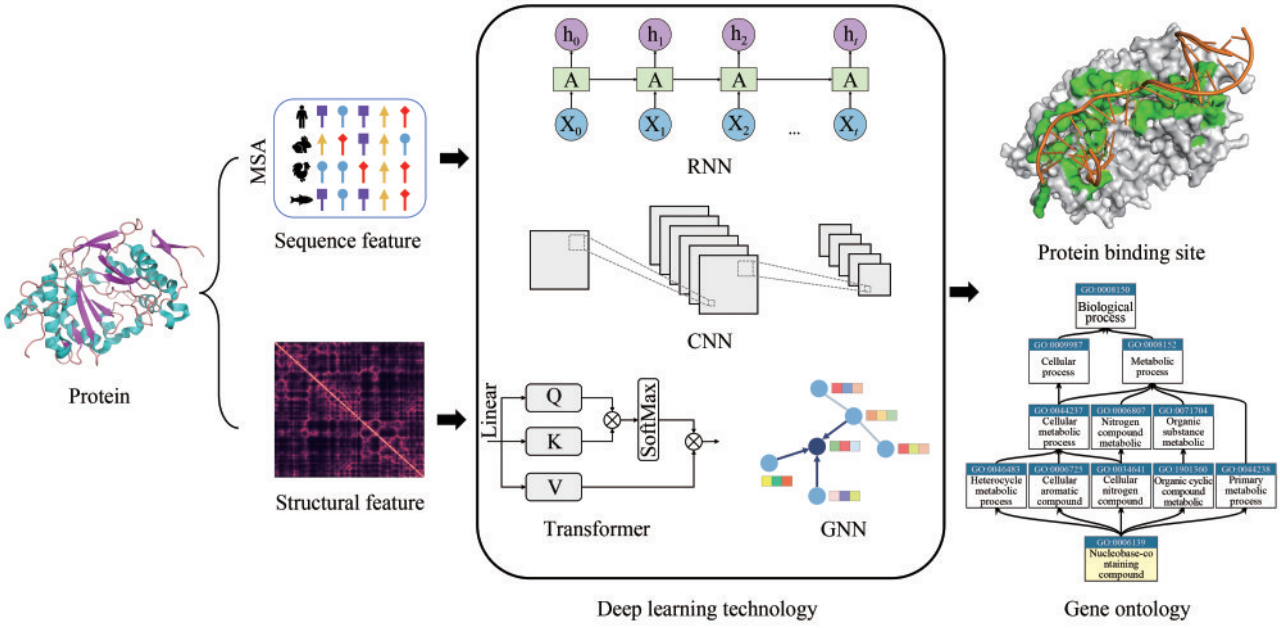
Application of deep learning in protein function prediction

SONG Yidong，YUAN Qianmu，YANG Yuedong
（*School of Computer Science and Engineering*，*Sun Yat-Sen University*，*Guangzhou 510000*，*Guangdong*，*China*）

Abstract: Protein function prediction is essential for bioinformatics analysis, which benefits a wide range of biological studies such as understanding the functions of metagenomes, uncovering mechanism underlying diseases, and finding new drug targets. With the rapid development of high-throughput sequencing technology, protein sequence data have been increased quickly, but functions of most proteins have not yet been identified. Since traditional biochemical experiments to determine protein functions are usually expensive, time-consuming, and less efficient, developing more efficient and effective computational methods for protein function prediction is of great significance. Deep learning technology has made breakthroughs in many fields, including image recognition, natural language processing, genomic analysis and drug discovery. In this review, we address applications of deep learning in protein function prediction, which can be divided into residue-level binding site prediction and protein-level gene ontology (GO) prediction. Protein binding sites are regions that bind to specific ligands, which play an important role in signal

收稿日期：2022-12-31 修回日期：2023-03-07
基金项目：国家重点研发计划（2022YF1203100）；国家自然科学基金（12126610）
引用本文：宋益东，袁乾沐，杨跃东. 深度学习在蛋白质功能预测中的应用[J]. 合成生物学，2023，4(3)：488-506
Citation：SONG Yidong，YUAN Qianmu，YANG Yuedong. Application of deep learning in protein function prediction[J]. Synthetic Biology Journal，2023，4(3)：488-506

transduction, metabolism, revealing molecular mechanisms underlying diseases, and designing new drugs. Gene ontology is a standard function classification system for genes, which provides a set of annotations to comprehensively describe the properties of genes and gene products. Firstly, we introduce commonly used large-scale protein structure and function databases. Secondly, discriminative protein sequence and structure features are described. Thirdly, we summarize the latest protein function prediction methods: in terms of the prediction of binding sites, we introduce the latest methods based on the ligand type, including protein, peptide, nucleic acid and small molecule as well as ion ligand, and in the aspect of GO prediction, we highlight the latest sequence-based, structure-based, and protein interaction network-based methods developed with protein information. Finally, we comment the advantages and disadvantages of the current protein function prediction methods, and discuss the future development in this field.



Keywords: deep learning; protein; function prediction; binding site; gene ontology

蛋白质在生物体内发挥着至关重要的作用，包括信号转导、催化代谢反应、维持细胞结构等，准确的蛋白质功能鉴定有助于疾病机制的阐明和药物新靶点的发现[1]。由于传统测定蛋白质功能的生化实验通常成本高、耗时长、通量低，开发高效且有效的蛋白质功能预测计算方法十分重要[2]。同时，传统的计算方法如分子动力学模拟、统计能量函数、分子对接等需要耗费大量资源且耗时较长，限制了这一领域的发展[3-5]。随着深度学习的蓬勃发展，通过深度学习进行蛋白质功能预测已经成为生物信息学的研究热点[6-8]。蛋白质功能预测可以分为残基水平的结合位点预测和蛋白水平的基因本体论（gene ontology， GO）预测，下面我们将从这两个方面逐一进行介绍。蛋白质的结合位点预测和GO预测是两个不同水平的预测，GO预测研究的是蛋白质具有的不同功能，而结合位点预测则是研究蛋白质在残基水平所具有的一些性质，两者是对蛋白质功能不同水平的刻画[6, 9]。

蛋白质结合位点是蛋白质上与特异性配体相结合的区域，蛋白质的结合位点预测在信号转导、运输和代谢[10]、揭示疾病的分子机制[11]和设计新药[12]等方面有着重要作用。目前蛋白质结合位点预测的方法可以分为基于序列和基于结构的方法。基于序列的方法如DELPHI[13]、PepNN[14]等，利

用序列提取的特征学习生物理化特征的局部模式，其优点是它们可以通过序列对任意蛋白进行预测。然而，由于结合残基的潜在模式并不能仅从它们的序列中显式地体现，而可能在空间结构[15]中是保守的，从蛋白质序列中捕获的特征可能不足以充分地表示残基。因此，基于序列的方法的性能可能受到限制。与基于序列的方法不同，以实验结构为输入的基于结构的方法往往更加准确，其一般可分为基于模板的方法、基于机器学习的方法和混合方法。基于模板的方法如MIB[16]使用比对算法来转移模板的结构信息并推断结合位点。然而，当缺少高质量的模板时，这些方法将受到严重的限制。基于结构的机器学习方法从蛋白质结构提取几何特征，然后再将其输送到神经网络，如DELIA[17]。另外，也可以考虑蛋白质结构的上下文拓扑信息，并使用端到端的方式进行训练，如GraphBind[7]。对于混合方法，如COACH[18]和IonCom[19]，则同时集成了基于模板和基于机器学习的方法。相对于基于序列的方法，基于结构的方法更加准确，但这种方法应用范围有限，只适用于存在实验三维结构的蛋白。

蛋白质功能可通过GO中的功能项描述[20]，其中GO涵盖了分子功能（molecular function， MF）、生物过程（biological process， BP）和细胞组分（cellular component， CC）三个生物学领域。通常一个蛋白质会与多个GO项相关，因此蛋白质功能预测可以看作是一个大规模、多类别、多标签的问题。此外，GO是一个有向无环图（directed acyclic graph， DAG），如果蛋白质被注释了GO项，那么它所有的祖先项也应该被注释。因此，蛋白质功能预测应该考虑GO的层次结构并产生合理的输出：一个GO项的预测概率必须等于或大于其所有子项[21]。为了促进蛋白质GO功能预测的发展，CAFA比赛（critical assessment of functional annotation）已成功举办了四次。具体来说，给定一个蛋白质，参加者需要在T_0之前提交预测结果，几个月后（T_1）组织者会收集具有最新实验注释的蛋白质作为测试集，对不同的方法进行评估。现有的蛋白质GO功能预测的方法根据所使用的信息大致可以分为三类：基于序列、基于结构和基于生物网络。大多数基于序列的方法利用序列相似性，搜索序列域，或者采用深度学习捕获判别性特征来进行预测。其中，由于相似的序列往往具有相似的功能，一种基本的方法就是直接从已知功能的同源序列中转移注释，如Blast2GO[22]。此外，另一种方法是寻找序列的结构域或蛋白家族进行预测。例如，GOLabeler[23]利用排序学习（learning to rank， LTR）[24]算法整合了序列同源性、蛋白质结构域和家族信息。随着深度学习技术的发展，通过设计复杂的神经网络，如DeepGOPlus[9]中的卷积神经网络和TALE[25]中的Transformer，也可以从序列中自动提取判别性嵌入信息。然而，目前基于序列的方法预测精度较低。相比于基于序列的方法，基于结构的方法具有更高的预测精度。基于结构的方法使用天然的蛋白质结构作为输入，通常使用图神经网络（graph neural networks， GNN）学习局部三级模式进行功能预测，如DeepFRI[26]。此外，基于网络的方法，利用生物网络（例如蛋白质-蛋白质相互作用或代谢网络）中连接的蛋白质可能具有相同功能的原理[27]，对蛋白质GO功能进行预测。例如，NetGO[28]在STRING[29]中集成了多个蛋白质网络，在网络中从最近的邻居转移注释至目标蛋白。NetGO 2.0[30]将文献和序列信息融入到NetGO中，进一步提高性能。尽管CAFA比赛表明结合多种信息的集成预测方法通常优于基于序列的方法，但这些额外的特征对于大多数蛋白质来说往往是不可用、不完整或难以获得的，这限制了它们的应用范围。单独从序列中预测蛋白质功能的方法则更具有普遍性，适用于大多数尚未被广泛研究的蛋白质。

通过与实验结合，使用计算方法对蛋白质功能进行准确预测具有重要意义。由于对蛋白质的全链筛选耗时且昂贵，预测方法可以帮助缩小潜在的结合位点范围。在我们之前的合作研究[31]中，通过计算预测方法并结合湿实验成功验证了JAK2激酶中的结合残基。同时，SPOT-Struc[32]使用结构比对和蛋白质结合亲和力预测对糖结合蛋白进行识别，并成功找到了糖结合蛋白。准确的蛋白质功能预测也可以为许多致病基因突变机制提出假设或结论，例如影响mRNA转运的*THOC2*突变[33]。在新的药物设计中，结合位点预测可用

于预测药物的可药用性[34]或作为从头分子设计的生成模型的条件[35]。综上所述，研究高效准确的蛋白质功能预测方法在生命科学领域具有重要作用，这也突出了这项研究的重要意义。

在本文中，我们将从残基水平的结合位点预测和蛋白水平的GO预测两方面对蛋白质功能预测进行详细的介绍。首先，我们将介绍该领域常用的数据库和蛋白特征。然后，在结合位点预测方面，我们按照配体的不同类型分别介绍了蛋白质-蛋白质、蛋白质-多肽、蛋白质-核酸和蛋白质-小分子或离子配体的结合位点预测方法，着重分析了每种方法的优缺点及不同方法之间的区别。同时，我们根据GO预测所使用的信息分别介绍了基于序列、基于结构和基于网络的GO预测方法，对这些方法进行详细的对比分析。最后，本文综合前面的介绍进行总结与展望，希望能推动该领域的进一步发展。

1 常用数据库介绍

首先，我们介绍蛋白质功能预测领域的常用数据库，如表1所示。

在蛋白质结构方面，PDB数据库是目前最重要的生物大分子结构数据库，包括蛋白质、核酸、多糖等的结构数据。在蛋白质序列方面，UniProt数据库包含世界上大部分公开可用的蛋白质序列，是研究蛋白质序列的重要资源。BioLiP数据库是重要的研究蛋白质与配体相互作用的数据库。GO数据库包含了不同生物体的基因功能的计算表示。基因本体注释（gene ontology annotation， GOA）数据库则在GO数据库的基础上对UniProt数据库进行注释，广泛地应用于GO预测研究中。

1.1 PDB数据库

蛋白质结构数据库（protein data bank， PDB）是美国Brookhaven国家实验室于1971年创建的，由结构生物信息学研究合作组织（Research Collaboratory for Structural Bioinformatics， RCSB）维护。该数据库是结构生物学研究中的重要资源，并且每周更新，截至2022年11月，PDB数据库已收集了约20万条实验测得的结构数据。PDB数据库是目前最主要的收集生物大分子（蛋白质、核酸、多糖和病毒）结构的数据库，其中的三维结构主要通过X射线单晶衍射、核磁共振、电子衍射等实验手段确定。PDB储存的内容包括生物大分子的原子坐标、参考文献、一级和二级结构信息，也包括了晶体结构因数以及NMR实验数据等。

1.2 BioLiP数据库

BioLiP是一个半自动半手动生成的生物相关的配体-蛋白质相互作用数据库。此前，大多数配体结合位点预测方法使用PDB中的蛋白质结构作为模板。然而，并非PDB中存在的所有配体都具有生物学相关性，因为小分子通常用作解析蛋白质结构的添加剂。为了促进基于模板的配体-蛋白质对接、配体虚拟筛选和蛋白质功能注释，BioLiP开发了一种分层程序来评估PDB结构中存在的配体的生物学相关性，包括四步的生物特征过滤以及仔细的人工验证。简单来说，判断配体与蛋白质受体具有生物相关性的要求是配体不在人工添加物列表中且同一个PDB文件中出现次数小于15次，与配体相互作用的受体结合位点残基不少于2个且结合位点残基不连续，如果配体在人工添加物列表中则检查其是否在PDB相关文献的PubMed摘要中被提及，如果被提及则是生物

表1 常用数据库介绍

Table 1 Commonly used databases

名称	内容	下载
PDB数据库	蛋白质结构	https://www.rcsb.org/
BioLiP数据库	配体-蛋白质相互作用数据	https://zhanggroup.org/BioLiP/
UniProt数据库	蛋白质序列和注释数据	https://www.uniprot.org/
GO数据库	细胞组分，分子功能，生物学过程	http://geneontology.org/
GOA数据库	基因本体注释数据	https://www.ebi.ac.uk/GOA/

相关的。BioLiP数据库一般一周更新一次，其中的每个条目都包含以下注释：配体结合残基、配体结合亲和力、催化位点、酶学委员会注释编号、基因本体论项以及连接到其他数据库的链接。

1.3 UniProt数据库

UniProt数据库是蛋白质序列和注释数据的综合资源，是欧洲生物信息学研究所（EMBL-EBI）、瑞士生物信息学研究所（SIB）、蛋白质信息资源（PIR）的合作项目。UniProt数据库包含三个部分，分别是UniProt Knowledgebase（UniProtKB）、UniProt Reference Clusters（UniRef）和UniProt Archive（UniParc）。UniProtKB是收集蛋白质功能信息的中心枢纽，具有准确、一致和丰富的注释。UniRef提供来自UniProt（包括异构体）和选定的UniParc记录的集群序列集，以便在多个分辨率下获得对序列空间的完整覆盖，同时从视图中隐藏冗余序列。UniParc是一个全面且非冗余的数据库，包含世界上大部分公开可用的蛋白质序列，蛋白质可能存在于不同的源数据库中，也可能存在于同一数据库中的多个副本中。UniParc通过仅存储每个唯一序列一次并为其提供稳定且唯一的标识符（UPI）来避免这种冗余，从而可以从不同的源数据库中识别相同的蛋白质。UniParc仅包含蛋白质序列，有关蛋白质的所有其他信息必须使用数据库交叉引用从源数据库中检索。

1.4 GO数据库

GO数据库提供了来自许多不同生物体的基因功能的计算表示，该数据库于2000年由GO组织（Gene Ontology Consortium）建立，拟在构建一个结构化的标准生物学模型，建立基因及其产物知识的标准词汇体系，包括细胞组分（cellular component）、分子功能（molecular function）、生物学过程（biological process）三个部分。

1.5 GOA数据库

GOA数据库旨在使用GO数据库的标准化词汇为UniProt数据库（Swiss-Prot、TrEMBL和PIR-PSD）提供高质量的电子和手动注释。作为GO注释的补充存档，GOA通过将UniProt注释转换为公认的计算格式来促进UniProt中表示的知识与其他数据库的高度集成。通过联合其他模型生物组的GO注释，GOA将专业知识和专家意见进行整合，以确保数据仍然是最新生物信息的关键参考。GOA已经越来越多地用于评估文本挖掘或蛋白质相互作用实验产生的GO预测，同时也用于增强特定模型生物体或基因表达数据集的注释。

2 常用特征介绍

本节首先介绍蛋白质序列的独热（one-hot）编码，基于20位的状态存储器对蛋白质序列进行编码；之后介绍位置特异性打分矩阵和隐马尔科夫矩阵；以及SPIDER3[36]软件，该软件在蛋白质序列及其他信息的基础上，进一步生成蛋白质的预测结构信息；此外还包括氨基酸物理化学性质和最新的语言模型特征。语言模型特征包括ESM和ProtTrans，这类模型通过在大规模数据上进行训练并学习相关生物学特性。同时还介绍了常用的结构特征，包括DSSP和蛋白距离图，该类特征用于基于结构的方法中，包含丰富的特征信息。

2.1 独热（one-hot）编码

由于蛋白质序列共包含20种氨基酸，属于离散特征，且取值之间无大小的意义，因此可以对氨基酸序列进行one-hot编码，即对序列中的每个氨基酸使用20位的状态寄存器表示，这20位状态寄存器中只有一位是有效的，将其记为1，其余寄存器记为0。最后我们可以得到大小为$L \times 20$的矩阵，其中L为序列长度。

2.2 位置特异性打分矩阵（position-specific scoring matrix，PSSM）

由进化保守氨基酸形成的蛋白基序（motif）常常与蛋白结合功能密切相关，如蛋白结合性质。

我们可以通过PSI-BLAST[37]程序进行多序列比对得到氨基酸序列的进化信息，在UniRef90[38]数据库上进行迭代搜索，为每个蛋白质生成PSSM特征，该特征将表示为$L \times 20$的矩阵，其中矩阵的每一行表示序列中特定位置氨基酸残基发生替代的对数似然分值，共L行。由于共有20种氨基酸，因此PSSM矩阵$\boldsymbol{M}$共包含20列。其中$\boldsymbol{M}_{i,j}$表示蛋白质序列在进化过程中第i个位置的氨基酸发生突变成为第j种氨基酸类型的分值，高度保守的位置将会获得较高的分值，而低度保守的位置会取得较低的得分。

2.3 隐马尔科夫矩阵

HHblits[39]是一种应用于蛋白质序列搜索和比对的开源工具包。相比于PSI-BLAST，HHblits能够更快且更准确地搜索出数据库中具有相似序列的蛋白质。HHblits将查询蛋白序列和数据库中的蛋白序列相互转换，接着通过隐马尔科夫模型（hidden Markov model，HMM）进行统计。HMM是一种在序列比对过程中统计出现突变可能性的模型，能够有效地提高子序列相似性搜索的准确率和灵敏度。通过在Uniclust30[40]数据库上运行HHblits以生成隐马尔科夫模型的序列谱，该特征将表示为$L \times 30$的矩阵。

2.4 SPIDER3

通过SPIDER3[36]软件可生成蛋白质的预测结构信息，SPIDER3的输入包括蛋白质序列以及通过PSI-BLAST和HHblits获得的PSSM和HMM特征，输出包括：①ASA（solvent accessible surface area），指的是溶剂可达（可以接触到溶剂）的生物分子表面积，简称溶剂可及性；②二面角（torsional angles），蛋白质主链的二面角通常包括5个，即θ、φ、ψ、ω和τ。由于蛋白质的平面性，ω通常是180°，所以一般不使用ω二面角，其余4个二面角分别取其正弦和余弦值，因此共8个特征；③CN，指的是在三维空间内，以残基为中心，给定的长度为半径的球体内包含了多少其他氨基酸，代表了这片区域内氨基酸的疏密度；④半球暴露（half-sphere exposure，HSE），这是CN特征的一个扩展，它将CN中的球体拆分成了上半部分和下半部分，对其分别计数，HSE以蛋白质二级结构中的C_α-C_α方向向量和C_α-C_β方向向量来确定两个半球的分界；⑤三个二级结构（即α螺旋、β折叠和无规卷曲）的预测概率值。

2.5 氨基酸物理化学性质

7维的氨基酸物理化学性质特征向量[41]，包括了空间参数、疏水性、体积、极化率、等电点、螺旋概率和片状概率，也常被用于蛋白表征学习当中。

2.6 语言模型特征ESM

ESM（evolutionary scale modeling）[42]是由Facebook提出，基于2.5亿个蛋白质序列训练大容量的Transformer语言模型，并使用该语言模型学习生物学特性。在人工智能领域，无监督学习所带来的数据规模和模型能力的结合，让表征学习和统计生成取得了重大进步。ESM团队使用无监督学习在跨越进化多样性的2.5亿个蛋白质序列上训练了一个包含860亿个氨基酸的深度上下文语言模型。得到的模型表示中包含了有关生物属性的重要信息，该信息在一系列应用中具有普适性，可以实现对突变效应和二级结构的监督预测，并改进用于远程接触预测的最新特征。

2.7 语言模型特征ProtTrans

ProtTrans[43]是一个蛋白质语言模型（protein language model，pLM），在包含3930亿个氨基酸的UniRef[38]和Big Fantastic Database数据集上进行训练，将自然语言处理（natural language processing，NLP）中的语言模型概念进行复制，将蛋白质序列中的氨基酸看作语言模型的词，将整个蛋白质视为语言模型中的句子。首先，将这些语言模型以自监督的方式进行训练，本质上是学习预测已知序列中隐藏的氨基酸。在训练完成后，需要确定语言模型捕获了相关信息。然后，通过提取嵌入信息来迁移语言模型学习到的内容，同时将其作为输入用于监督训练每个残基和每个蛋白质的预测任务。

2.8 结构特征DSSP

使用DSSP[44]软件可以对蛋白的PDB结构进行特征提取，计算出三个类别的结构特征：①8维的one-hot二级结构分类；②肽骨架扭转角PHI和PSI，取其正弦值和余弦值；③溶剂可及性表面积，随后根据对应氨基酸类型的最大ASA归一化为相对溶剂可及性（relative solvent accessibility，RSA）。

2.9 结构特征蛋白距离图（distance map）

根据蛋白质的PDB文件，可以得到每个氨基酸的C_α原子坐标，然后计算所有氨基酸对之间C_α原子坐标的欧氏距离，即可得到一个$L \times L$的蛋白距离图。一种常见的处理方法是设定一个距离阈值，距离图中大于此阈值的值转换为0，小于此阈值的转换为1，从而得到一个邻接矩阵，用于表示蛋白质氨基酸之间接触与否。此邻接矩阵可用于表示蛋白图，从而应用图卷积神经网络（graph convolutional network， GCN）等图模型进行学习。另一种处理方法是将蛋白距离图矩阵转换为热力图，从而运用卷积神经网络等图像学习模型进行学习。

在上面介绍的特征中，由于GO预测的数据规模较大，而PSSM、HMM、SPIDER3特征需要的计算时长较长，因此此类特征一般不适合进行GO预测。同时GO数据集没有结构，DSSP特征对这类问题也不适合。GO预测问题一般使用one-hot特征，或者使用当下最新提出的语言模型（ESM或ProtTrans）提取丰富的特征信息作为输入。对于结合位点预测问题，上面所介绍的各种特征被广泛用于多种预测方法，该类问题使用的特征范围更广。

3 最新方法介绍

3.1 结合位点类方法

在这里按照不同的配体类型选择部分结合位点预测方法进行介绍，方法总结于表2。

3.1.1 蛋白质-蛋白质结合位点预测方法

DELPHI[13]是一种基于序列的PPI位点预测框架，集成了卷积神经网络（CNN）和循环神经网络（RNN）进行结合位点预测。DELPHI使用的特征有GO词频、序列对比信息、氨基酸三联体（3 mer）、蛋白家族信息、结构域和基序、ProFET[58]序列特征，同时该方法具有开源代码和可供使用的服务器。DELPHI使用不同的模型去捕获不同的信息，模型主要由三部分组成，分别是卷积神经网络模块、循环神经网络模块以及集成模块。CNN和RNN组件的核心层分别为卷积和双向门控循环单元（GRU）层，而集合模型主要负责对前两个分量的输出进行解码。除了提出一种基于CNN和RNN的集成模型之外，DELPHI又一重要贡献是提出了三种全新的特征，并将这三种特征首次用到PPI位点预测中，具有重要意义。相比于基于序列的方法，基于结构的方法使用了蛋白质的结构信息，这类方法通常具有较高的准确度。

GraphPPIS[8]是一种基于结构的方法，使用深度图网络进行蛋白质结合位点的预测。该模型将蛋白质视为无向图，将PPI位点预测视为图节点分类问题，同时综合进化信息和结构信息构建节点特征，计算成对氨基酸之间的距离构建邻接矩阵。然后，使用初始残差和恒等映射实现深度图卷积框架，并用来捕获来自高阶氨基酸邻居的信息。GraphPPIS使用的特征有PSSM、HMM和DSSP，并且具有可下载的代码及可使用的web服务器。GraphPPIS通过初始残差连接以及恒等映射的方式使得GCN克服了堆叠高层数时出现的梯度消失以及过平滑现象，并能够很好地捕捉到蛋白质图的远程邻居消息。普通图卷积网络已经被证明会逐步将节点的低阶邻居信息聚合到自身，这在多数图相关的任务上可以取得很好的性能效果，但限制了其感知远程邻居的能力，且本身还存在过平滑现象。GraphPPIS通过初始残差连接以及恒等映射将普通GCN扩展为深层GCN，与普通GCN相比，深层GCN有两个优势：第一是在一定程度上能够保证层数堆叠起来之后仍然保留蛋白质的初始结构消息，从而能够减缓梯度消失以及过平滑现象；第二是为权重矩阵加入了恒等映射矩阵，它保证了深层GCN在仅堆叠少数基层的时候

表2 结合位点预测最新方法总结

Table 2 Summary of the latest binding site prediction methods

方法		数据来源	年份	特征[①]	算法	是否开源[②]
蛋白-蛋白	SPPIDER[45]	PDB	2007	物理化学性质，基于MSA的进化信息，DSSP结构信息，dSA（预测的和真实RSA的差值）	全连接神经网络	S
	SCRIBER[46]	BioLip	2019	相对溶剂可及性，进化保守性，相对氨基酸结合倾向性，物理化学性质，内部无序性，二级结构，残基位置	逻辑回归	S
	DELPHI	PDB, BioLip	2020	高分值片段对，ProtVec1D，PSSM，进化保守性，相对溶剂可及性，相对氨基酸结合倾向性，亲水性，内部无序性，物理化学性质，PKx，位置信息	CNN+GRU	S, C
	DeepPPISP[47]	PDB	2020	PSSM，二级结构，one-hot蛋白序列	CNN	S, C
	MaSIF[48]	—	2020	表面几何与物理化学特征，如局部曲率、Poisson-Boltzmann静电、氢键供体或受体以及亲水性	几何深度学习	C
	GraphPPIS	PDB	2021	PSSM，HMM，DSSP	GCN	S, C
蛋白-多肽	SPRINT[49]	PDB	2016	one-hot蛋白序列，PSSM，相对溶剂可及性，二级结构，物理化学性质	SVM	S
	PepBind[50]	BioLiP	2018	PSSM，HMM，二级结构，内部无序性	SVM+基于模板的方法	S
	Visual[51]	BioLiP	2020	PSSM，半球暴露，二级结构，溶剂可及性，扭转角，物理化学性质	CNN	C
	$BiteNet_{p_p}$	BioLip	2021	体素化的11种原子密度	3D CNN	S, C
	PepNN	PDB	2022	残基间距离，C_α的相对方向，局部坐标系间旋转矩阵，残基的相对位置，one-hot蛋白序列，扭转骨架角，语言模型特征	互注意力机制+GNN	C
蛋白-核酸	DNAPred[52]	PDB	2019	PSSM，预测的二级结构和溶剂可及性，结合与非结合氨基酸的频率差	SVM	S
	NucBind[53]	PDB	2019	PSSM，HMM，预测的二级结构，预测结构	SVM+COACH-D[54]	S
	NCBRPred[55]	—	2021	PSSM，HMM，预测的二级结构和溶剂可及性	GRU	S, C
	GraphBind	BioLiP	2021	残基的原子特征，DSSP，PSSM，HMM	GNN	S, C
	GraphSite	BioLiP	2022	AlphaFold2 single特征，PSSM，HMM，DSSP	Graph Transformer	S, C
蛋白-小分子或离子配体	TargetS[56]	PDB	2013	PSSM，预测的二级结构，相对氨基酸结合倾向性	AdaBoost	S
	IonCom[19]	BioLiP	2016	PSSM，预测的二级结构和溶剂可及性，保守性，氨基酸的离子结合频率，预测结构	AdaBoost+SVM+COFACTOR[57]+S-SITE[18]+TM-SITE[18]	S, C
	MIB[16]	PDB	2016	结构模板数据	Fragment Transformation	S
	DELIA	BioLip	2020	PSSM，HMM，二级结构，可溶性，S-SITE特征，基于结构的距离矩阵	CNN	S
	LMetalSite	BioLiP	2022	语言模型特征	Transformer+多任务学习	S, C
综合不同类型配体	MTDsite	BioLip	2021	PSSM，HMM，SPIDER3，溶剂可及性表面积，扭转角，分界线内的残基数，半球暴露	BiLSTM+多任务学习	C
	DeepDISOBind	DisProt	2022	one-hot蛋白序列，相对氨基酸亲和性，二级结构，内部无序性	CNN+多任务学习	S, C

[①]PKx表示解离常数负对数。

[②]S和C分别表示网页服务器和源代码可用。

仍然能够保持性能不下降。该方法是第一个使用深度图卷积网络进行蛋白质结合位点预测的工作，可以很容易地扩展到其他功能位点预测的任务中。

3.1.2 蛋白质-多肽结合位点预测方法

在蛋白质-多肽结合方面，最新的方法有$BiteNet_{p_p}$[59]、PepNN[14]，$BiteNet_{p_p}$和PepNN分别是基于3D卷积神经网络和图神经网络构建的模型，

两者均是当前比较突出的模型。$BiteNet_{P_p}$和PepNN均是基于结构的方法，其中$BiteNet_{P_p}$基于三维图像的目标检测进行蛋白质-多肽结合位点预测，PepNN则提出了一种相互注意力模块（reciprocal attention），增强了输入之间的信息流动。

$BiteNet_{P_p}$是一种基于结构的深度学习模型，通过将蛋白质结构视为目标检测的三维图像来识别蛋白质-多肽结合位点。$BiteNet_{P_p}$使用的特征有体素化的11种原子密度并且具有可下载的代码和web服务器。基于从PDB蛋白质数据库中检索到的蛋白质-多肽复合物的非冗余集合，模型训练了一个3D卷积神经网络进行蛋白质-配体结合位点预测模型，据悉，这是首次使用域自适应技术将蛋白质-小分子复合物的模型微调为蛋白质-多肽复合物的模型。$BiteNet_{P_p}$使用了一种基于张量的空间蛋白质结构表示，并将其输入到3D卷积神经网络，利用3D卷积神经网络对蛋白质结构进行体素化表示，即对蛋白进行3D单元表示，最终输出蛋白质-多肽结合位点的坐标及其概率得分。$BiteNet_{P_p}$使用了域自适应技术，即在蛋白质-多肽数据集上微调在蛋白质-小分子复合物上训练的原始$BiteNet_{P_p}$模型，通过这种域适应技术来提高模型性能。该方法可以对大规模的结合位点进行快速检测，只需要不到1 s的时间就可以分析单个蛋白质结构。

PepNN是一种基于结构和序列的蛋白质-多肽结合位点预测方法。预测蛋白质-多肽的结合位点的一个主要困难是多肽的柔性及其在结合时发生的构象变化，考虑到这些因素，PepNN提出了一种相互注意力模块（reciprocal attention），在增强对称性的同时同步更新多肽和蛋白质残基的编码，允许两个输入之间的信息流动。PepNN将该模块与图神经网络层集成，并在训练时使用迁移学习来弥补蛋白质-多肽复合物信息的稀缺性。在这项研究中，作者整合了语言模型、可用的蛋白质-蛋白质复合物数据和基于任务的注意力架构，分别开发了基于结构和基于序列的并行模型PepNN-Struct和PepNN-Seq。由于蛋白质-多肽复合物数据较为稀缺，PepNN-Struct和PepNN-Seq使用了集成了迁移学习的基于注意力的深度学习模块，来弥补这种数据限制。此外，PepNN的成功可以作为相互注意力机制有效性的证明，该模块可以有效地用于建模数据点对之间的双向关系，因此可以扩展到其他生物分子相互作用，包括蛋白质-蛋白质和蛋白质-DNA的相互作用。在这些情况下，序列或结构信息可以通过序列或图注意力模块进行传播，然后相互注意力模块可以有效地将受体蛋白与之联系起来，同时保持两者的对称性。

3.1.3 蛋白质-核酸结合位点预测方法

GraphBind[7]是一种基于结构的蛋白质-核酸结合位点预测器，基于端到端图神经网络，通过层次图神经网络（HGNN）学习蛋白质结构上下文嵌入规则，并用于识别与核酸结合的残基。GraphBind输入的特征包括残基的原子特征、DSSP、PSSM和HMM，由于结合位点在局部三级结构上往往表现出高度的保守模式，GraphBind首先根据目标残基的结构上下文及其空间邻域构建图。然后，使用层次图神经网络学习结构与理化特征的局部模式的隐含嵌入用于识别结合的残基。对于每个目标残基，首先基于目标残基的局部环境构建一个图。初始节点特征向量由进化保守性、二级结构信息、其他生物理化特征和位置嵌入组成，其中位置嵌入是通过定义结构上下文中残基空间关系的几何知识来计算的。之后再构建一个分层图神经网络来学习潜在的局部模式，并用于结合残基预测，其中设计了边更新模块、节点更新模块和图更新模块来学习目标残基的高级几何和生物理化特征。此外，GraphBind还利用门控循环单元[60]堆叠了多个GNN-blocks，充分利用了所有block的信息，避免了梯度消失问题。总的来说，GraphBind的优越性主要表现在两个方面：①基于结构上下文的图表示适合表示目标残基局部环境的几何和生物物理化学知识；②在预测结合残基方面，HGNN是一种高效的学习高级模式的算法。同时，GraphBind也有一定的局限性，当使用预测的结构作为GraphBind的输入时会降低GraphBind的性能，这表明结构质量与几何知识有关，而几何知识对HGNN非常重要。GraphBind需要找到一种新的构建异质图的方法，使得对结构信息具有更好的鲁棒性。

GraphSite[6]是一种基于序列的方法，通过使

用AlphaFold2预测的结构对DNA结合残基进行精确预测。GraphSite结合了图Transformer和AlphaFold2预测的蛋白质结构，并应用于DNA结合残基的预测。GraphSite将结合位点预测问题转化为图节点分类任务，并使用Transformer变体模型来考虑蛋白质的结构信息，通过利用预测的蛋白质结构和图转换器，GraphSite相较于最新的基于序列和基于结构的方法都有了很大的改进。具体来说，GraphSite在计算Transformer中的注意力分数时，融合了多序列比对（multi-sequence alignment，MSA）信息和结构信息来构建残差特征，并通过计算成对氨基酸距离来覆盖空间上距离较远的氨基酸。这是第一个利用AlphaFold2预测的结构和图转换器进行蛋白质-DNA结合位点预测的工作。总的来说，GraphSite的优越性主要体现在两个方面：①AlphaFold2可以预测出较高质量的蛋白质结构；②在结合残基的预测方面，结构感知的图转换器是学习模式的有效算法。同时，GraphSite模型仍然存在一些局限性，GraphSite的性能很大程度上受到AlphaFold2预测质量的影响。这可以通过添加其他信息丰富的序列衍生特征来提高模型对结构预测质量的鲁棒性来解决。在GraphSite[6]的文章中，GraphSite和其他众多方法在测试集Test_129上进行了比较。其中，GraphSite、GraphBind和NucBind均表现出较好的性能，其AUC分别为0.934、0.928和0.858。GraphSite借助于AlphaFold2预测的蛋白质结构，使用图Transformer对DNA结合残基进行预测，相较于目前的方法有了很大的改进。GraphBind则基于层次图神经网络（HGNN）对与核酸结合的残基进行识别。该方法的优势在于基于结构上下文的图表示包含了重要的特征信息，同时HGNN是一种高效的学习高级模式的算法，在结合位点预测中较为有效。NucBind则基于所输入的PSSM、HMM、预测的二级结构、预测结构等特征对结合位点进行了很好的预测。

3.1.4 蛋白质-小分子或离子配体结合位点预测方法

DELIA[17]是一种新的基于深度学习的蛋白质-配体结合残基的预测方法。该方法输入的特征有PSSM、HMM、二级结构、可溶性，S-SITE特征和基于结构的距离矩阵，同时该方法提供了一个可供使用的web服务器。DELIA设计了一种混合深度神经网络，将基于序列的一维特征与基于结构的二维氨基酸距离矩阵进行融合。同时为了克服结合残基和非结合残基之间严重的数据不平衡问题，DELIA设计了小批量过采样、随机欠采样和堆叠集成的策略来增强模型，并且在五个基准数据集上达到很好的效果。为了开发出更强大的蛋白质-配体结合残基预测的预测器，DELIA设计了一种融合卷积神经网络和双向长短时记忆网络（BiLSTM）的混合深度神经网络来处理异质蛋白质数据，包括一维序列特征向量和二维距离矩阵[61-62]。其中距离矩阵是蛋白质结构的有效表示，表达的是蛋白质结构中每一对残基之间的距离信息。为了从距离矩阵中挖掘出更多的信息，DELIA中使用CNN从距离矩阵中提取局部信息，并且设计深度架构来学习用于结合位点识别的高层表示。同时，与体素化表示相比，二维距离矩阵更加紧凑，对旋转和平移具有不变性，因此更适合此类问题。

LMetalSite[63]是一种无需序列比对的预测BioLiP中最常见的四种金属离子（Zn^{2+}，Ca^{2+}，Mg^{2+}和Mn^{2+}）结合位点的方法。LMetalSite利用预训练的语言模型快速生成信息丰富的序列表示，并使用Transformer捕获长程依赖关系。同时采用多任务学习来弥补训练数据的稀缺性，捕捉不同金属离子之间的内在相似性，并在多个基准数据集上取得较好效果。LMetalSite利用最近发布的预训练语言模型ProtTrans[43]以避免耗时的数据库搜索，在短时间内生成信息丰富的序列表示。其还利用多任务学习，通过弥补训练数据的稀缺性和更好地建模不同金属离子之间的内在相似性来进一步提高预测质量。具体来说，LMetalSite使用Transformer模型[64-65]作为共享网络来捕获蛋白质序列中的长程依赖等常见的结合机制，然后使用四个针对于不同离子的特异性多层感知器（MLP）来学习特定金属离子的结合模式。总的来说，LMetalSite仅使用蛋白质序列就取得了很好的性能（超越了最好的基于结构的方法），这有望同时解决当前基于结构和基于序列方法的局限性。同时LMetalSite所采用的多任务学习技术能够进一步提高预测质量，而其他方法则忽略了相似离子之间的潜在关系。

此外，LMetalSite仍然存在可以改进的空间，如元学习（meta-learning，指的是在多个学习阶段改进学习算法的过程）[66-67]在多任务问题中有重要的作用，LMetalSite可以结合元学习进行更深的探索。

3.1.5 多任务整合不同类型的配体

MTDsite[68]是一种新的结合位点预测器，采用多任务深度学习策略，基于序列来同时预测具有多个重要分子类型的结合位点。MTDsite输入的特征包括PSSM、HMM、SPIDER3、溶剂可及性表面积、扭转角、分界线内的残基数、半球暴露等，同时该方法提供了可下载的源代码。通过合并DNA、RNA、多肽和糖结合蛋白的4个训练集，MTDsite在各自的独立测试集上获得了准确和鲁棒的预测。而且据我们所知，这也是第一个使用多任务框架同时预测多个分子结合位点的方法。在MTDsite中，不同的任务之间共享一个网络，互相促进学习，从而获得相对较强的抽象能力，其中长短期记忆网络（LSTM）作为共享网络来收集蛋白质链中长距离残基的信息。同时，针对四种不同的个体类型MTDsite分别训练了四个小的特定子网络，用来提取个体属性。

DeepDISOBind[69]是一种创新的深度多任务架构，可以从蛋白质序列中准确预测与DNA、RNA和蛋白质结合的内在无序的区域（IDRs）。该方法通过输入one-hot蛋白序列、相对氨基酸亲和性、二级结构、内部无序性等特征进行结合位点预测，并且提供了可下载的源代码和web服务器。DeepDISOBind依赖于一个信息丰富的序列谱，该序列谱由一个创新的多任务深度神经网络处理，并且在后续层逐渐特异化，以预测特定模式的结合。其中普通输入层会链接到区分蛋白质和核酸结合的层，该层再进一步链接到区分DNA和RNA相互作用的层。实证检验表明，与单一任务设计相比，这种多任务设计在三种不同类型任务中提供了统计上显著的预测质量增益。多任务学习旨在通过使用共享表示来预测相关学习任务[70-71]并进一步提高预测性能，该方法可以推广到其他领域。

3.2 GO预测

我们根据使用信息的不同对蛋白质GO预测的方法进行了逐一介绍，并着重分析了部分最新的方法，表3将各种预测方法进行了总结。

表3 最新GO预测类方法总结

Table 3 Summary of the latest GO prediction methods

	方法	年份	特征	算法	是否开源①
基于序列	GOLabeler	2018	GO词频，序列对比信息，氨基酸三联体（3 mer），蛋白家族信息，结构域和基序，ProFET[58]序列特征	LTR	S, C
	DeepGOPlus	2020	基于序列和基序的功能信息	CNN	S, C
	TALE[25]	2021	one-hot蛋白序列，GO层次结构矩阵、序列相似性	Transformer+CNN	C
	GAT-GO	2022	one-hot蛋白序列，PSSM，HMM，ESM-1b嵌入信息	GAT	
	DeeProtGO[72]	2022	SeqVec序列嵌入、序列相似性、物种分类、InterPro蛋白结构域和蛋白家族信息、GO注释信息	层次化的全连接神经网络	C
基于结构	COFACTOR[73]	2017	蛋白序列、结构信息和PPI网络	序列比对+结构比对+基于网络邻居的功能聚合	S
	DeepFRI	2021	蛋白质接触图，语言模型特征	GCN	S, C
基于网络	DeepGO[74]	2018	蛋白序列，PPI网络	CNN+层次化的全连接神经网络	S, C
	NetGO	2019	GO词频，序列对比信息，氨基酸三联体（3 mer），蛋白家族信息，结构域和基序，ProFET[58]序列特征，蛋白质相互作用网络	LTR	S
	NetGO 2.0	2021	GO词频，基于序列信息，蛋白质相互作用网络，序列中的深层模式，文献信息	LTR	S
	S2F	2021	同源信息，HMMER特征，InterPro特征，进化信息，PPI网络	label diffusion	S, C
	DeepGraphGO	2021	InterPro特征，PPI网络	GCN	C

①S和C分别表示网页服务器和源代码可用。

3.2.1 基于序列的方法

GOLabeler[23]是一种用于预测未知蛋白质功能的新方法，它集成了5个组件分类器，并从不同的特征中进行训练，包括GO项频率、序列比对、氨基酸三联体（3 mer）和生物物理特性等，同时该方法提供了可供下载的代码并且具有web服务器。GOLabeler在基于排序学习（LTR）的框架中进行训练，其中排序学习是机器学习中的一种范式，对于多标签分类尤为有效。GOLabeler的基本思想是在排序学习的框架下整合不同类型的基于序列的信息。LTR的逻辑是，对于排名较低的正样本会受到更多的惩罚，而在常规分类中，它们会受到无区分平等的处理。LTR最初是为了使网页排序与网页和用户查询之间的相关性一致而开发的。如果关注二进制相关性，那么排序问题就变成了预测给定查询的相关网页的问题。这正是多标签分类，将网页视为标签，查询视为示例。LTR可以通过对标签进行排序并选择排名靠前的标签来解决这类问题。因此，以GO项为标签，以蛋白质为例，可以将LTR应用于相应的自动功能预测（automated function prediction，AFP）中。另外，LTR的另一个值得注意的优点是GOLabeler可以有效地集成多个基于序列的信息，这些信息是由不同类型的分类器（或组件）生成的，其中所有的信息都来自于序列。总的来说，基于序列的蛋白质大规模AFP（SAFP）是一个重要的问题，主要具有三方面的挑战：①结构化的本体；②每个蛋白质有许多标签；③每个蛋白质的GO条目数量变化大。针对上面的问题，GOLabeler进行了针对性设计，并解决了以下问题：①使用GO的DAG结构中所有对应的GO项；②通过排序学习，进行更有效的多标签分类；③通过LTR，允许不选择每个蛋白质的GO项数量。

DeepGOPlus[9]是一种新颖的单独从序列预测蛋白质功能的方法，将深度卷积神经网络模型与基于序列相似性的预测相结合，在多个基准数据集上达到了很好的效果。DeepGOPlus使用的特征有基于序列和基序的功能信息，并且该方法具有web服务器。DeepGOPlus在2017年提出的DeepGO[74]基础上进行了改进，克服了其在序列长度、缺失特征和预测类别数量方面的限制。DeepGOPlus模型将输入的长度增加到2000个氨基酸（覆盖了UniProt中99%以上的序列），同时将新模型的架构进行改进，使其能够分割更长的序列和扫描更小的模块来进行功能预测。在模型方面，DeepGOPlus将神经网络预测与基于序列相似性的方法相结合，以捕获直接和间接的相互作用信息。总的来说，DeepGOPlus是一种从蛋白质序列中预测蛋白质功能的快速而准确的工具。特别地，DeepGOPlus对氨基酸序列的长度没有限制，因此可以用于蛋白质功能的基因组尺度注释，这在新测序的生物体中尤为重要。DeepGOPlus也不对蛋白质所属的分类做任何假设，因此可以进行宏基因组学的功能预测。此外，DeepGOPlus速度较快，即使在单个CPU上也能在几分钟内注释数千个蛋白质，这使其能够进一步应用于宏基因组学或大量未知功能蛋白质的鉴定项目。

GAT-GO[75]是一种基于图注意网络（graph attention network，GAT）的方法，可以通过利用预测的结构信息和蛋白质序列的嵌入信息来大幅提高蛋白质功能的预测能力。GAT-GO使用的特征有one-hot蛋白序列、PSSM、HMM和ESM-1b嵌入信息。GAT-GO使用RaptorX[76]预测的蛋白质的结构信息，并使用Facebook的ESM-1b[42]生成其嵌入信息。即使在测试蛋白与训练蛋白的序列一致性较低的情况下，GAT-GO也优于传统的基于同源性的算法，如BLAST[77]和以前的深度学习方法[9]。最近的两项研究[26, 78]探索了GCN和蛋白质嵌入信息在蛋白质功能预测方面的作用，但与仅基于序列的方法相比，它们的改进有限。GAT-GO与GCN方法DeepFRI[26]的不同之处在于：GAT-GO使用了GAT[79]代替传统的GCN，GAT可以通过自注意力机制进行灵活的节点特征聚合来增强模型容量。此外，GAT-GO使用了拓扑池化[80]实现更高效的下采样，提高模型的泛化能力。通过结合序列特征、蛋白质嵌入信息和残基间接触图，GAT-GO可以从局部和全局信息中预测蛋白质功能。相反，基于序列的方法不能利用预测的结构信息，因此不善于处理与任何训练序列不相似的测试序列。同时，GAT-GO没有使用非常大的宏基因组数据库来生成用于残基间接触预测的多序列比对，从而节约了搜索这些数据库所需要的计算资源。

3.2.2 基于结构的方法

DeepFRI[26]是一种基于图卷积网络（GCN）的蛋白质功能注释和检测蛋白质中功能区域的方法，称为深度功能残基识别（deep functional residue identification，DeepFRI）。DeepFRI输入的特征包括蛋白质接触图和语言模型特征，并且具有可供使用的web服务器。DeepFRI通过利用从蛋白质语言模型和蛋白质结构中提取的序列特征来预测蛋白质的功能，具有显著的去噪能力，并且其类激活映射使其达到了较高分辨率的预测。DeepFRI具有一个两阶段的体系结构，将蛋白质结构和来自预先训练的、与任务无关的语言模型的序列表示作为输入，并表示为3D结构中氨基酸相互作用的图。尽管高质量的序列比对往往足以传递折叠或结构信息[53]，但由于不同功能需要不同的阈值、部分比对、蛋白质兼并和新功能化等原因，序列比对很难用于传递函数。因此，DeepFRI的一个重要优势是能够通过提取局部序列和全局结构特征进行超越同源比对的功能预测[2]。总之，DeepFRI描述了一种将计算生物学中的两个关键问题（蛋白质结构预测和蛋白质功能预测）联系起来的方法。DeepFRI将深度学习与越来越多的可用序列和结构数据联系起来，有可能满足不断增长的基因组序列数据带来的挑战，为我们解释蛋白质生物多样性提供了新的见解。

3.2.3 基于网络的方法

NetGO[28]是一个能够通过整合海量蛋白质-蛋白质网络信息来进一步提高大规模蛋白质自动功能预测（AFP）性能的Web服务器。该方法使用的特征包括GO词频、序列对比信息、氨基酸三联体（3 mer）、蛋白家族信息、结构域和基序、ProFET序列特征、蛋白质相互作用网络。NetGO的基本思想是将基于网络的信息整合到GOLabeler框架中[23]，从而提高大规模AFP的性能，其主要的优势有以下3个方面：①NetGO依靠机器学习强大的排序学习框架，有效整合了蛋白质的序列和网络信息；②NetGO利用了STRING数据库中所有物种（大于2000）的海量网络信息，而不仅仅是一些特定的物种；③即使某个蛋白质不包含在STRING中，NetGO仍然可以利用网络信息通过同源转移来注释蛋白质。NetGO将网络信息与其他类型的数据相结合，以进行更好的蛋白质功能预测，其将几个组件集成到一个有效的框架中，在大规模网络的综合实验中取得了最好的性能。同时，NetGO网络服务器运行速度快，具有可视化界面，适合大规模蛋白质功能预测，是一款高性能Web服务器。另外，在2021年该团队提出了更新版本NetGO 2.0[30]，其在NetGO的基础上，将通过逻辑回归得到的文献信息和循环神经网络提取的序列信息纳入框架。实验结果表明，NetGO 2.0在生物过程（BP）和细胞成分（CC）子本体上的表现明显优于NetGO。进一步分析，NetGO 2.0的优越性能表明：①额外信息的使用有助于AFP，NetGO 2.0进一步结合了SwissProt[81]中通过逻辑回归手动注释的每个蛋白质的文献信息和RNN的潜在序列信息，这些信息将有助于提供大规模AFP的性能；②神经网络可以进一步提取隐藏在序列中的高阶信息；③排序学习框架可以很好地集成新的信息和方法。在NetGO 2.0[30]文章中，NetGO 2.0和其他众多方法在测试集（testing data）上进行了比较，NetGO 2.0、NetGO和GOLabeler均达到了较好的性能。其中，NetGO 2.0的MFO（AUPR）、BPO（AUPR）和CCO（AUPR）分别是0.655、0.269和0.593；NetGO分别为0.653、0.239和0.583；GOLabeler的分别是0.647、0.193和0.193。NetGO 2.0是在NetGo的基础上，加入了文献信息和循环神经网络提取的序列信息，更进一步地提高了模型的性能。GOLabeler则是在排序学习的框架下整合不同类型的基于序列的信息，所使用的特征包括GO项频率、序列比对、氨基酸三联体（3 mer）和生物物理特性等，在蛋白质功能预测方面有很好的性能。

S2F[82]是一种新颖的基于网络传播的预测蛋白质功能的方法，其主要思想是系统地将功能相关的数据从模式生物转移到新测序的生物，从而可以使用标签传播方法。S2F引入了一种新颖的标签扩散算法，可以解释具有相关功能的蛋白质重叠在网络中的重叠（overlapping）效应。S2F将网络传播算法应用于只有序列信息可用的生物体，通过系统地传递模式生物的功能数据来创建网络，并利用这些网络来组合和增强通过同源性或可识别的蛋白质特征获得的一些初步的GO标签。使用

网络上的扩散过程是提高简单同源性预测的有效方法，S2F通过一个扩散过程，将同源信息和可识别的蛋白质特征以及同源映射图中包含的进化信息有效地融合在一起。同时，S2F允许通过学习到的系数对不同网络进行线性组合，其组合方法与GeneMANIA[83]中使用的方法类似，但它允许学习这些线性权重，而不依赖于初始的已知功能标签集。

DeepGraphGO[84]提出了一种基于端到端的多物种图神经网络AFP方法，该方法充分利用了蛋白质序列和高阶蛋白质网络的信息，其多物种策略允许对所有物种训练一个单一的模型，这使得DeepGraphGO比现有方法拥有更多的训练样本。DeepGraphGO是一种半监督的深度学习方法，通过图神经网络[85]同时利用蛋白质序列和网络信息，并且具有3个显著特点：①蛋白质表示是由InterPro数据库[86]生成，InterPro结合了Pfam[87]、SUPERFAMILY[88]、CATH-Gene3D[89]和CDD[90]等14个不同的数据库，提供了蛋白家族、结构域和基序等多种类型的功能信息。②DeepGraphGO包含多个图卷积神经网络（GCN）层。GNN已被开发用于各种任务，如节点嵌入、链接预测、节点分类和图分类[91]。GCN是一种典型的GNN，它可以通过一个GCN层获得每个节点的表示向量，该层聚合了相邻节点的表示。而在DeepGraphGO中使用了多层GCN，有助于捕获节点之间的高阶信息，提升模型性能。③DeepGraphGO具有多物种策略。DeepGraphGO使用所有物种的蛋白质只训练一个单一的模型，这种被称为多物种策略的方法与以往专注于单个物种的工作相比，它可以利用更多的数据来达到更好的性能，特别是对于那些缺少注释数据的物种尤为重要。

4 总结与展望

本篇文章首先介绍了与蛋白质功能预测有关的数据库（PDB、BioLiP、UniProt、GO和GOA数据库），然后介绍了常用的特征。之后根据配体类型分别介绍了最新的蛋白质结合位点预测方法，并根据使用的信息介绍了基于序列、基于结构和基于网络的蛋白质GO功能预测方法。

总的来说，蛋白质结合位点预测的方法可以分为基于序列和基于结构的方法。基于序列的方法只需从序列中对任意蛋白进行预测，但由于结合残基的潜在模式并不能仅从它们的序列中显式地体现，而在空间结构[15]中是保守的，基于序列的方法相对于基于结构的方法性能上受到一定限制。基于结构的方法可分为基于模板的方法、基于机器学习的方法和混合方法。基于模板的方法是该领域早期的研究主流，然而对于不存在高质量模板的输入蛋白，基于模板的方法准确率通常较低，这使得后来的主流方法主要基于机器学习，或结合机器学习与模板搜索。基于结构的机器学习方法是从蛋白质结构提取几何特征，然后再将其输送到神经网络，或者直接考虑蛋白质结构的上下文拓扑结构，并使用端到端的方式进行训练。基于结构的混合方法则同时集成了基于模板和基于机器学习的方法。相对于基于序列的方法，基于结构的方法更加准确，但这种方法受限于实验测得的蛋白质结构的数量，只适用于具有可用三级结构的蛋白质。蛋白质GO功能预测的方法按照使用信息的不同大致可以分为基于序列、基于结构和基于网络的方法。大多数基于序列的方法利用序列相似性，搜索序列域，或者采用深度学习捕获判别性特征来进行预测。目前基于序列的方法预测精度较低，相比于基于序列的方法，基于结构的方法使用天然的蛋白质结构进行GO功能预测，具有更高的准确度。此外，基于网络的方法利用生物网络中连接的蛋白质可能具有相同功能的原理[27]进行预测。

尽管当前蛋白质功能预测的方法已经达到了很好的效果，但是仍然存在一些可以改进的地方。首先，在对蛋白质功能进行预测时，不同配体之间存在潜在联系，如蛋白和多肽以及不同金属离子之间，因此可以使用多任务学习提高预测质量。然而最新的研究表明，元学习[66-67]也可以很好地应用在多任务问题中，并能够快速适应标签有限的未知任务，因此可以尝试使用元学习进一步提升模型性能。其次，基于语言模型的预测结构已经被证明对结合位点问题有用，如GraphSite[6]。而ESMfold[92]实验证明具有和AlphaFold2[93]相近的准确率，因此可以使用ESMFold快速生成高质

量三维结构，并通过更好的几何学习模型捕捉结构信息，如GVP[94]和Graph Transformer[95]等，以此来提高预测性能。同时，对于数据不均衡问题，可以使用先进的采样技术加以解决。对比学习[96]是一种自监督学习方法，用于在没有数据标注的情况下，让模型学习同类数据之间的相似和不同类数据之间的差异性，从而学习数据的一般特征，目前对比学习方法也被应用到了蛋白质GO预测领域[97]。在使用PPI网络预测GO时，可以将对比学习应用于PPI网络，以最大化网络邻居之间的功能相似性，进一步提高预测性能。另外，知识图谱技术[98]也可以引入到这一问题中，用以融合药物和疾病信息。可以探索蛋白质结合位点预测和GO预测之间的关系，如使用不同配体的结合位点的预测信息作为GO预测的特征，进一步丰富特征表示，提高性能。同时，还可以进一步增加新的特征信息来提高预测性能，包括生物进化树、宏基因组、基因表达信息等。通过对蛋白质进行更加丰富的表达，深入探索蛋白质功能的内在联系，更好地进行预测。

参 考 文 献

[1] EISENBERG D, MARCOTTE E M, XENARIOS I, et al. Protein function in the post-genomic era[J]. Nature, 2000, 405 (6788): 823-826.

[2] 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.

[3] ISRALEWITZ B, BAUDRY J, GULLINGSRUD J, et al. Steered molecular dynamics investigations of protein function[J]. Journal of Molecular Graphics & Modelling, 2001, 19(1): 13-25.

[4] KLEPEIS J L, LINDORFF-LARSEN K, DROR R O, et al. Long-timescale molecular dynamics simulations of protein structure and function[J]. Current Opinion in Structural Biology, 2009, 19(2): 120-127.

[5] PIERRI C L, PARISI G, PORCELLI V. Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening[J]. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2010, 1804(9): 1695-1712.

[6] YUAN Q M, CHEN S, RAO J H, et al. AlphaFold2-aware protein-DNA binding site prediction using graph transformer[J]. Briefings in Bioinformatics, 2022, 23(2): bbab564.

[7] XIA Y, XIA C Q, PAN X Y, et al. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues[J]. Nucleic Acids Research, 2021, 49(9): e51.

[8] YUAN Q M, CHEN J W, ZHAO H Y, et al. Structure-aware protein-protein interaction site prediction using deep graph convolutional network[J]. Bioinformatics, 2021, 38(1): 125-132.

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

[10] ZHANG J, KURGAN L. Review and comparative assessment of sequence-based predictors of protein-binding residues[J]. Briefings in Bioinformatics, 2018, 19(5): 821-837.

[11] KUZMANOV U, EMILI A. Protein-protein interaction networks: probing disease mechanisms using model systems[J]. Genome Medicine, 2013, 5(4): 37.

[12] WELLS J A, MCCLENDON C L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces[J]. Nature, 2007, 450(7172): 1001-1009.

[13] LI Y W, GOLDING G B, ILIE L. DELPHI: accurate deep ensemble model for protein interaction sites prediction[J]. Bioinformatics, 2021, 37(7): 896-904.

[14] ABDIN O, NIM S, WEN H, et al. PepNN: a deep attention model for the identification of peptide binding sites[J]. Communications Biology, 2022, 5: 503.

[15] CHEN J W, XIE Z R, WU Y H. Understand protein functions by comparing the similarity of local structural environments[J]. Biochimica et Biophysica Acta (BBA) -Proteins and Proteomics, 2017, 1865(2): 142-152.

[16] LIN Y F, CHENG C W, SHIH C S, et al. MIB: metal ion-binding site prediction and docking server[J]. Journal of Chemical Information and Modeling, 2016, 56(12): 2287-2291.

[17] XIA C Q, PAN X Y, SHEN H B. Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data[J]. Bioinformatics, 2020, 36(10): 3018-3027.

[18] YANG J Y, ROY A, ZHANG Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment[J]. Bioinformatics, 2013, 29(20): 2588-2595.

[19] HU X Z, DONG Q W, YANG J Y, et al. Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transferals[J]. Bioinformatics, 2016, 32 (21): 3260-3269.

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

[21] DAVIS J, GOADRICH M. The relationship between Precision-Recall and ROC curves[C]//Proceedings of the 23rd international conference on Machine learning. June 25-29, 2006, Pittsburgh, Pennsylvania, USA. New York: ACM, 2006: 233-240.

[22] CONESA A, GÖTZ S, GARCÍA-GÓMEZ J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. Bioinformatics, 2005, 21(18): 3674-3676.

[23] 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.

[24] LI H. A short introduction to learning to rank[J]. IEICE Transactions on Information and Systems, 2011, E94-D(10): 1854-1862.

[25] CAO Y, SHEN Y. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding[J]. Bioinformatics, 2021, 37(18): 2825-2833.

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

[27] OLIVER S. Guilt-by-association goes global[J]. Nature, 2000, 403(6770): 601-602.

[28] 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.

[29] SZKLARCZYK D, GABLE A L, NASTOU K C, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets[J]. Nucleic Acids Research, 2021, 49(D1): D605-D612.

[30] 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.

[31] WANG S Y, LIANG K, HU Q S, et al. JAK2-binding long noncoding RNA promotes breast cancer brain metastasis[J]. The Journal of Clinical Investigation, 2017, 127(12): 4498-4515.

[32] TIRALONGO J, COOPER O, LITFIN T, et al. YesU from *Bacillus subtilis* preferentially binds fucosylated glycans[J]. Scientific Reports, 2018, 8: 13139.

[33] KUMAR R, CORBETT M A, VAN BON B W M, et al. *THOC2* mutations implicate mRNA-export pathway in X-linked intellectual disability[J]. The American Journal of Human Genetics, 2015, 97(2): 302-310.

[34] SCHMIDTKE P, BARRIL X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites[J]. Journal of Medicinal Chemistry, 2010, 53(15): 5858-5867.

[35] XU M Y, RAN T, CHEN H M. *De novo* molecule design through the molecular generative model conditioned by 3D information of protein binding sites[J]. Journal of Chemical Information and Modeling, 2021, 61(7): 3240-3254.

[36] HEFFERNAN R, YANG Y D, PALIWAL K, et al. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility[J]. Bioinformatics, 2017, 33(18): 2842-2849.

[37] 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.

[38] SUZEK B E, HUANG H Z, MCGARVEY P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters[J]. Bioinformatics, 2007, 23(10): 1282-1288.

[39] REMMERT M, BIEGERT A, HAUSER A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment[J]. Nature Methods, 2012, 9(2): 173-175.

[40] MIRDITA M, VON DEN DRIESCH L, GALIEZ C, et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments[J]. Nucleic Acids Research, 2017, 45 (D1): D170-D176.

[41] MEILER J, MÜLLER M, ZEIDLER A, et al. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks[J]. Molecular Modeling Annual, 2001, 7(9): 360-369.

[42] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[43] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing [EB/OL]. arXiv, 2020: 2007.06225[2023-02-01]. https://arxiv.org/abs/2007.06225.

[44] KABSCH W, SANDER C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.

[45] POROLLO A, MELLER J. Prediction-based fingerprints of protein-protein interactions[J]. Proteins: Structure, Function, and Bioinformatics, 2007, 66(3): 630-645.

[46] ZHANG J, KURGAN L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences[J]. Bioinformatics, 2019, 35(14): i343-i353.

[47] ZENG M, ZHANG F H, WU F X, et al. Protein-protein interaction site prediction through combining local and global features with deep neural networks[J]. Bioinformatics, 2020, 36 (4): 1114-1120.

[48] GAINZA P, SVERRISSON F, MONTI F, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning[J]. Nature Methods, 2020, 17(2): 184-192.

[49] TAHERZADEH G, YANG Y D, ZHANG T, et al. Sequence-based prediction of protein-peptide binding sites using support vector machine[J]. Journal of Computational Chemistry, 2016, 37(13): 1223-1229.

[50] ZHAO Z J, PENG Z L, YANG J Y. Improving sequence-based prediction of protein-peptide binding residues by introducing intrinsic disorder and a consensus method[J]. Journal of Chemical Information and Modeling, 2018, 58(7): 1459-1468.

[51] WARDAH W, DEHZANGI A, TAHERZADEH G, et al. Predicting protein-peptide binding sites with a deep convolutional neural network[J]. Journal of Theoretical Biology, 2020, 496: 110278.

[52] ZHU Y H, HU J, SONG X N, et al. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines[J]. Journal of Chemical Information and Modeling, 2019, 59(6): 3057-3071.

[53] SU H, LIU M C, SUN S S, et al. Improving the prediction of protein-nucleic acids binding residues *via* multiple sequence profiles and the consensus of complementary methods[J]. Bioinformatics, 2019, 35(6): 930-936.

[54] WU Q, PENG Z L, ZHANG Y, et al. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking[J]. Nucleic Acids Research, 2018, 46(W1): W438-W442.

[55] ZHANG J, CHEN Q C, LIU B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning[J]. Briefings in Bioinformatics, 2021, 22(5): bbaa397.

[56] YU D J, HU J, YANG J, et al. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(4): 994-1008.

[57] ROY A, YANG J Y, ZHANG Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation[J]. Nucleic Acids Research, 2012, 40(W1): W471-W477.

[58] OFER D, LINIAL M. ProFET: feature engineering captures high-level protein functions[J]. Bioinformatics, 2015, 31(21): 3429-3436.

[59] KOZLOVSKII I, POPOV P. Protein-peptide binding site detection using 3D convolutional neural networks[J]. Journal of Chemical Information and Modeling, 2021, 61(8): 3814-3823.

[60] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. arXiv, 2014: 1406.1078[2023-02-01]. https://arxiv.org/abs/1406.1078.

[61] GRAVES A. Long short-term memory[M]//Studies in Computational Intelligence: Supervised sequence labelling with recurrent neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 37-45.

[62] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time series[M/OL]//The handbook of brain theory and neural networks. Cambridge, MA, USA: MIT Press, 1995 [2023-02-01]. https://citeseerx. ist. psu. edu/viewdoc/download; jsessionid=0B925DD52A8A879C47A4032DEC9CE5E4? doi= 10.1.1.32.9297&rep=rep1&type=pdf.

[63] YUAN Q M, CHEN S, WANG Y, et al. Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning[J]. Briefings in Bioinformatics, 2022, 23(6): bbac444.

[64] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in neural information processing systems[C/OL]//Advances in Neural Information Processing Systems 30-NeurIPS 2017[2023-02-01]. https://proceedings. neurips. cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[65] ZHENG S J, RAO J H, ZHANG Z Y, et al. Predicting retrosynthetic reactions using self-corrected transformer neural networks[J]. Journal of Chemical Information and Modeling, 2020, 60(1): 47-55.

[66] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. August 6-11, 2017, Sydney, NSW, Australia. New York:

ACM, 2017: 1126-1135.

[67] WANG J H, ZHENG S J, CHEN J W, et al. Meta learning for low-resource molecular optimization[J]. Journal of Chemical Information and Modeling, 2021, 61(4): 1627-1636.

[68] SUN Z, ZHENG S J, ZHAO H Y, et al. To improve prediction of binding residues with DNA, RNA, carbohydrate, and peptide *via* multi-task deep neural networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(6): 3735-3743.

[69] ZHANG F H, ZHAO B, SHI W B, et al. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning[J]. Briefings in Bioinformatics, 2022, 23(1): bbab521.

[70] ZHANG Y, YANG Q. An overview of multi-task learning[J]. National Science Review, 2018, 5(1): 30-43.

[71] CARUANA R. Multitask learning[J].Machine Learning, 1997, 28(1): 41-75.

[72] MERINO G A, SAIDI R, MILONE D H, et al. Hierarchical deep learning for predicting GO annotations by integrating protein knowledge[J]. Bioinformatics, 2022, 38(19): 4488-4496.

[73] ZHANG C X, FREDDOLINO P 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.

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

[75] LAI B Q, XU J B. Accurate protein function prediction *via* graph attention networks with predicted structure information[J]. Briefings in Bioinformatics, 2022, 23(1): bbab502.

[76] XU J B, MCPARTLON M, LI J. Improved protein structure prediction by deep learning irrespective of co-evolution information[J]. Nature Machine Intelligence, 2021, 3(7): 601-609.

[77] ALTSCHUL S F, GISH W, MILLER W, et al . Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[78] VILLEGAS-MORCILLO A, MAKRODIMITRIS S, VAN HAM R C H J, et al. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function[J]. Bioinformatics, 2021, 37(2): 162-170.

[79] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. arXiv, 2017[2023-02-01]. https://arxiv.org/pdf/1710.10903.pdf.

[80] LEE J Y, LEE I Y, KANG J W. Self-attention graph pooling[C/OL]// Proceedings of the 22nd international conference on Machine learning, 9-15 June 2019, Long Beach, California, USA, 97: 3734-3743[2023-02-01]. https://proceedings. mlr. press/v97/lee19c.html.

[81] BOUTET E, LIEBERHERR D, TOGNOLLI M, et al. UniProtKB/Swiss-prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view[M]//Plant Bioinformatics. New York: Springer New York, 2016: 23-54.

[82] TORRES M, YANG H X, ROMERO A E, et al. Protein function prediction for newly sequenced organisms[J]. Nature Machine Intelligence, 2021, 3(12): 1050-1060.

[83] MOSTAFAVI S, RAY D, WARDE-FARLEY D, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function[J]. Genome Biology, 2008, 9(Suppl 1): S4.

[84] YOU R H, YAO S W, MAMITSUKA H, et al. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction[J]. Bioinformatics, 2021, 37(Supplement_1): i262-i271.

[85] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. arXiv, 2016: 1609.02907 [2023-02-01]. https://arxiv.org/abs/1609.02907.

[86] MITCHELL A L, ATTWOOD T K, BABBITT P C, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations[J]. Nucleic Acids Research, 2019, 47(D1): D351-D360.

[87] FINN R D, COGGILL P, EBERHARDT R Y, et al. The Pfam protein families database: towards a more sustainable future[J]. Nucleic Acids Research, 2016, 44(D1): D279-D285.

[88] OATES M E, STAHLHACKE J, VAVOULIS D V, et al. The SUPERFAMILY 1.75 database in 2014: a doubling of data[J]. Nucleic Acids Research, 2015, 43(D1): D227-D233.

[89] LEWIS T E, SILLITOE I, DAWSON N, et al. Gene3D: extensive prediction of globular domains in proteins[J]. Nucleic Acids Research, 2018, 46(D1): D1282.

[90] MARCHLER-BAUER A, BO Y, HAN L Y, et al. CDD/SPARCLE: functional classification of proteins *via* subfamily domain architectures[J]. Nucleic Acids Research, 2017, 45(D1): D200-D203.

[91] ZHOU J, CUI G Q, HU S D, et al. Graph neural networks: a review of methods and applications[J]. AI Open, 2020, 1: 57-81.

[92] LIN Z M, AKIN H, RAO R S, et al., Language models of protein sequences at the scale of evolution enable accurate structure prediction [EB/OL]. bioRxiv, 2022[2023-02-01]. https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.

[93] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583-589.

[94] JING B W, EISMANN S, SURIANA P, et al. Learning from protein structure with geometric vector perceptrons[EB/OL]. arXiv, 2020: 2009.01411[2023-02-01]. https://arxiv.org/abs/2009.01411.

[95] YUN S J, JEONG M Y, KIM R Y, et al. Graph transformer networks[C/OL]//Advances in Neural Information Processing Systems 32-NeurIPS 2019[2023-02-01]. https://proceedings. neurips. cc/paper/2019/file/9d63484abb477c97640154d40595a3bb-Paper.pdf.

[96] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 1597-1607.

[97] 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.

[98] ZHENG S J, RAO J H, SONG Y, et al. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa344.

通讯作者：杨跃东（1980—），男，教授，博士生导师。研究方向为高性能生物信息计算，蛋白质结构与功能预测，智能药物设计，跨尺度多组学数据挖掘，及生物医药超算平台。

E-mail：yangyd25@mail.sysu.edu.cn

第一作者：宋益东（1998—），男，博士研究生。研究方向为蛋白质功能预测、蛋白质无序预测等。

E-mail：songyd6@mail2.sysu.edu.cn