GCL-GO: A novel sequence-based hierarchy-aware method for protein function prediction

摘要

通过实验标注大量通过高通量技术快速拓展出的蛋白质是不切实际的。本研究为了解决这一问题，提出了一种基于序列的方法，即GCL-GO。该方法使用蛋白质语言模型表示序列，并通过图对比学习表示GO术语，从而结合这两个特征来预测蛋白质功能。GCL-GO通过对比GO图的结构特征和GO术语的语义特征学习GO术语之间的关系，提高了对未见过的GO术语的泛化能力，并减少了对训练数据的依赖，具有通用性和可拓展性。此外，引入了GCL-GO+，将基于序列相似性的方法与GCL-GO相结合，以提高性能。与传统的图对比学习方法不同，GCL-GO在功能泛化和可伸缩性方面展现了潜力，尤其在新的GO术语或在训练数据集中很少注释的GO术语上表现最佳。总体而言，该研究旨在改进蛋白质功能预测的基于序列的方法，提高对未见过的蛋白质序列和GO术语的泛化能力，并在CAFA3和TALE数据集上取得了优越的性能。

关键词：protein function prediction；gene ontology；graph constructive learning；protein language model

1 引言

本篇文章主要探讨了蛋白质功能的预测问题。蛋白质是生命活动中不可或缺的组成部分，它们的功能研究对于理解生物学机制和治疗各种疾病至关重要。尽管现代高通量技术能够迅速扩增大量蛋白质，但要确定这些庞大数量的蛋白质功能却非常具有挑战性。传统的湿实验方法，即在实验室中直接使用生物样本（如液体或组织）进行的实验，不仅耗时而且成本高昂，因此并不适合用于大规模的蛋白质功能确定。为了提升蛋白质功能预测的效率，研究人员开发了基于计算的方法。这些方法从蛋白质序列、结构和蛋白质相互作用网络等多种生物数据中推断蛋白质的功能。尤其是基于序列的方法，可以根据蛋白质序列进行功能预测。但这些方法通常难以从未见过的蛋白质序列中泛化，导致数据缺失的问题。为了解决这一问题，研究人员提出了一些自监督的图对比学习模型，应用于图表征学习任务。这些模型能够包含未标记的数据，并在学习图表征时实现泛化。同时，一些蛋白质语言模型也在蛋白质序列表征提取方面取得了成功。鉴于此，本研究提出了一种新颖的基于序列的层次感知方法，命名为GCL-GO。GCL-GO结合了蛋白质语言模型和图对比学习，分别用于获取序列表征和表征基因本体（Gene Ontology）的特征。通过整合这两种特征，GCL-GO能够有效预测蛋白质的功能。

2 相关工作

2.1 GO

GO（Gene Ontology，基因本体） [1]是一个广泛用于生物信息学领域的重要工具，它提供了一个统一的词汇来描述基因和蛋白质在不同生物体中的作用。GO 分为三个主要类别：生物过程 (Biological Process)、细胞组分 (Cellular Component) 和分子功能 (Molecular Function)。每个类别都是对基因和蛋白质功能的不同方面的描述。生物过程涉及到生物体内的过程和功能序列，细胞组分则关注蛋白质在细胞中的位置，而分子功能则描述了蛋白质的生化活动。GO 不仅提供了一个标准化的词汇，还包括了这些术语之间的关系，使得研究人员能够更有效地交流和共享关于基因和蛋白质功能的数据。通过这种方式，GO 对于基因组注释、基因表达分析、疾病模型研究等领域都至关重要。

2.2 传统计算方法

统计算法是一种常用的蛋白质功能预测方法。这些算法通常基于蛋白质序列或结构的统计特征，以及这些特征与已知功能之间的相关性。例如，通过比较新蛋白质序列与已知功能蛋白质序列的相似性，可以推测其潜在功能。其他常见的方法包括基于蛋

在蛋白质功能预测中，统计方法通常涉及利用已知数据的统计特性来预测蛋白质的功能。这些方法可能包括比较蛋白质序列的相似性、分析蛋白质结构的特定模体（如活性位点），或者使用计算溶剂映射技术来发现活性位点。例如，结构比对或基因组上下文分析方法可用于发现功能上有联系的蛋白质。这些统计方法能够揭示蛋白质可能的生物学功能，尤其在缺乏直接实验数据的情况下尤为有用。通过这种方式，研究者可以对大量未知功能的蛋白质进行高效的功能注释。

2.3 基于序列的平面预测方法

在蛋白质功能预测领域，基于序列的平面预测方法是一种重要的技术，它利用蛋白质的氨基酸序列来预测蛋白质的功能。这种方法的基本原理是，蛋白质的功能与其氨基酸序列密切相关，通过分析这些序列，可以推断出蛋白质可能的功能。

基于序列的平面预测方法首先利用已知功能的蛋白质序列作为参考，将待预测功能的蛋白质序列与之进行比较。通过寻找序列之间的相似性，可以推测未知蛋白质可能的功能。然后在蛋白质序列中寻找特定的模体（motif）或模式，这些模体是与特定功能相关的序列模式。通过识别这些模体，可以对蛋白质的功能进行预测。使用生物信息学数据库中的信息，如UniProt或Pfam，对蛋白质序列进行注释。这些数据库提供了大量关于已知蛋白质功能的信息，有助于预测新蛋白质的功能。随着生物信息学领域的发展，越来越多的机器学习技术被应用于蛋白质功能预测。这些方法可以从蛋白质序列中提取复杂特征，并进行有效的功能分类。

基于序列的平面预测方法是蛋白质功能预测的一个重要工具，它是一种基于氨基酸序列信息来理解和预测其生物学功能的有效途径。这种方法的核心优势在于它能够处理那些结构未知，但序列已知的蛋白质，通过分析其序列特征来推测功能。特别是在生物学研究和药物

开发领域，基于序列的平面预测方法为识别新蛋白质的潜在功能提供了一个重要工具，有助于加深我们对生命科学的理解。

2.4 基于序列的层次感知预测方法

该类方法利用蛋白质的氨基酸序列来推断其可能的功能。这种方法的核心思想是，蛋白质的功能与其氨基酸序列紧密相关，因为序列决定了蛋白质的结构，而结构又决定了功能。

在层次感知预测方法中，通常会采用机器学习或深度学习技术来分析蛋白质序列。这些技术能够从大量已知功能的蛋白质序列中学习模式，然后将这些模式应用于未知功能的蛋白质序列，从而预测其功能。这种方法的优势在于能够处理大量数据，并且随着数据库的不断丰富，预测准确性也在不断提高。

此外，层次感知预测方法还考虑到了蛋白质功能的层次结构。在生物学中，蛋白质的功能可以分为不同的层次，从广泛的类别（如酶、转运蛋白）到更具体的功能（如特定类型的生化反应）。基于序列的层次感知预测方法能够在这些不同层次上进行功能预测，提供更为细致和详尽的功能描述（例如特定的代谢途径或分子交互作用）。通过对这种层次结构的理解，研究人员可以更准确地预测蛋白质的多种潜在功能，从而为生物学研究和医学应用提供重要信息。

2.5 基于序列相似性的DIAMOND方法

DIAMOND [2]是一种生物信息学中用于蛋白质比对的高效工具。它主要用于在大规模基因组学和数据密集型进化项目中对测序读数与蛋白质参考数据库进行比对。DIAMOND的一个显著特点是它的运行速度远远超过传统的比对工具，如BLASTX。据报道，DIAMOND在处理短读数时的速度比BLASTX快大约20,000倍，同时保持了相似的敏感度。

DIAMOND之所以能够实现这样的高效率，部分原因是它采用了双重索引（double-indexing）方法和多重间隔种子（spaced seeds）技术。这种方法特别适合处理大量的查询和参考数据库，使得DIAMOND在处理庞大的数据集时仍能保持高效的运算性能。这对于需要快速分析庞大数据集的大规模基因组学研究尤为重要。

在敏感度方面，当使用大型基准数据集与其他工具进行比较时，DIAMOND显示出与BLASTP相似的敏感度水平，但计算性能却大大提高。随着时间的推移，DIAMOND的更新和改进不断提升其速度和敏感度，使其成为蛋白质比对领域的一个强大工具。DIAMOND在分布式计算环境（如超级计算机）中的可扩展性和效率也值得关注。它能够在极短的时间内完成生命各域之间的全面比对，这是使用其他工具（如BLAST）所无法比拟的。总的来说，DIAMOND因其快速、敏感的蛋白质比对能力而在生物信息学领域中脱颖而出。

3 本文方法

3.1 本文方法概述

本研究提出了一种名为GCL-GO的创新层次感知方法，用于蛋白质功能预测。该方法结合了蛋白质语言模型和图对比学习，以提取并预测GO术语特征，包括那些不常见的术语。GCL-GO的关键优势在于其能够准确嵌入GO术语的特征，同时对训练数据的依赖较少。此

外，研究还提出了GCL-GO+，它将DIAMOND的序列相似性方法与GCL-GO相结合，以增强性能。在CAFA3和TALE数据集的测试中，GCL-GO+表现优于其他基于序列的方法，并展示了在识别训练集中罕见的新GO术语方面的功能泛化和可扩展性。

实验总体的示意图如图1所示：

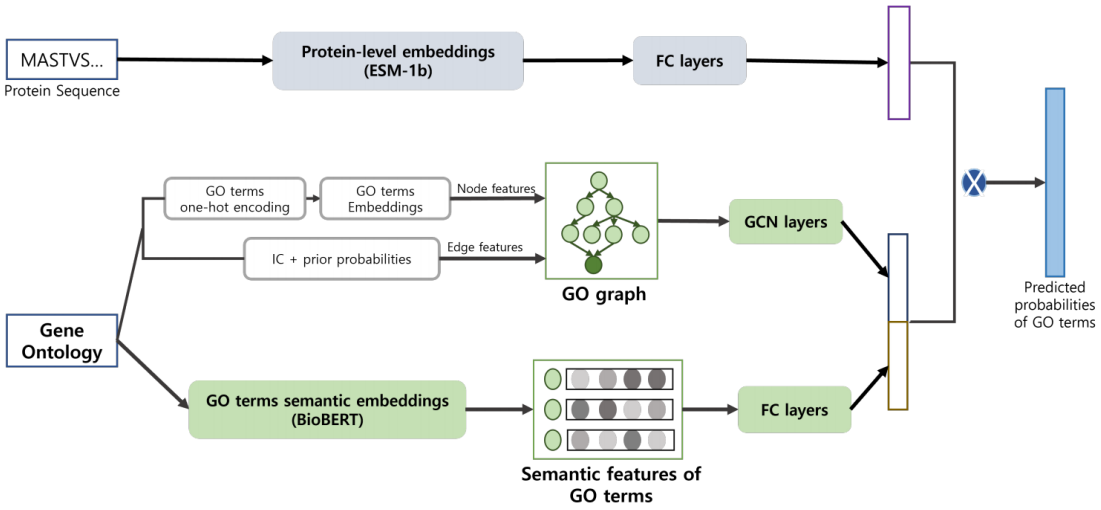

图 1. 方法示意图

3.2 蛋白质序列编码器

这部分主要在讲如何获得的蛋白质embedding，通过使用ESM-1b（一种蛋白质模型语言）来表示蛋白质序列，这是一种专门用于蛋白质序列表征学习的大型Transformer模型，拥有十亿参数。ESM（Evolutionary Scale Modeling）能够学习包括二级结构预测和接触图预测在内的蛋白质序列特征。这个模型通过训练超过2.5亿个UniRef50数据库中的蛋白质序列，能够捕捉蛋白质序列中的复杂模式，从而提高蛋白质结构预测任务的性能。ESM-1b产生的氨基酸级别嵌入每个嵌入含有1280个维度。通过对所有氨基酸的嵌入进行平均池化处理，从而获得蛋白质级别的嵌入表示，这有助于提取与蛋白质功能相关的序列特征，因为蛋白质的功能与整个序列相关，而非单个氨基酸。

3.3 GO编码器

受到SUGRL [4]启发，本研究使用图对比学习来表示GO项（包含它们之间的深度相关性）。

3.3.1 GO图构建

本研究使用GO数据集中构造图结构信息，使用GO术语的名称和定义构造语义特征。使用一个独热编码矩阵$H^0 \in \mathbb{R}^{N\times N}$构建初始节点特征，其中N是GO术语（节点）的数量，N是GO术语（节点）的数量。然后将稀疏的独热编码矩阵转换为稠密的矩阵$H^0 \in \mathbb{R}^{N\times d_0}$来构建节点特征嵌入，$d_0$是目标嵌入维度，这样做是为了降低维度并防止过拟合，降低计算复杂性，提高训练效率。

对于邻接矩阵的构建，邻接矩阵$A = P(U_s|U_t) + \frac{IC(s)}{\sum_{i \in child(t)} IC(i)}$，其中$IC(t) = -\log p(t)$、$p(t) = \frac{freq(t)}{freq(root)}$、$freq(t) = U_t + \sum_{i \in child(t)} freq(i)$，邻接矩阵$A$通过组合先验概率（prior probability，$P(U_s|U_t)$）和信息内容IC来构建，U_t是训练数据集中注释的GO术语的数量、P表示父节点t和子节点s在相同蛋白质中注释的条件概率，$p(t)$是GO数据集中每一个GO术语的概率，freq(t)是t的频率，它结合了每个子节点的频率和U_t。

3.3.2 构造GO术语的语义特征

每个GO术语都有自己的名称和定义，而语义相似的GO术语有相似的名称和定义，BioBERT是通过将BERT应用于生物医学文档来学习特定领域语言表示的大规模生物医学语料库预训练模型。本研究将名称和定义连接成一个句子，并对该句子使用BioBERT，BioBERT的输出被作为用作语义特征。

3.3.3 学习GO术语的表征

本研究使用两个多层的GCN来获得GO图中GO术语的结构信息，GCN层的定义为：$H^{l+1} = ReLU(\hat{A}H^l W^l)$，其中$H^l$是第$l^{th}$层特征，$\hat{A}$是$\tilde{D}^{-1/2}\tilde{A}\hat{D}^{-1/2} \in \mathbb{R}^{N \times N}$，$\hat{D}$是$\hat{A}$的度矩阵,这一个对角矩阵，其对角线元素为每个节点的度（邻居的数量），$\tilde{D}^{-1/2}$表示度矩阵的逆平方根，$\tilde{A} = A + I_N$，I_N是单位矩阵，这里是为了通过在对角线上添加自环（Self-Loop），可以确保每个节点的自身特征在聚合过程中也得到考虑，通过对邻接矩阵进行归一化操作，通过归一化操作，每个节点的邻居特征对中心节点的影响被缩放，使得在聚合过程中每个邻居的贡献更加平衡，这一步的目的是在聚合邻居节点的特征时考虑到每个邻居节点对中心节点的贡献，并防止在图卷积的过程中出现数值爆炸或消失的问题。

最后，使用全连接层来获取语义特征中的语义信息，为了将结构与语义信息结合起来，本研究将两条信息连接起来，以创建GO术语的最终表示

3.4 预测层定义

使用输入蛋白质的序列特征和GO术语特征的点积来预测输入蛋白质序列的单个GO术语概率，$Y = sigmoid(H^T S)$，S是蛋白质序列特征，H是GO术语特征。

3.5 损失函数定义

使用多重损失来探索SUGRL中的结构信息和语义信息之间的互补信息，将语义信息定义为锚定嵌入的内容，通过对语义信息进行打乱生成负嵌入，分别基于结构信息和语义信息设定了两种正嵌入类型，有两种多重损失函数，在结构部分，正嵌入是由GCN学到的结构信息，$L_{St} = \frac{1}{N}\sum_{i \in N}\left\{d(h, h^+)^2 - d(h, h_i^-)^2 + \alpha\right\}_+$，$L_{St}$是由结构正嵌入$h^+$得来的结构损失，在语义部分，正嵌入是通过整合邻居信息获得的，邻居信息是通过计算语义特征之间的余弦相似度生成的，本研究预先将每个GO term的前5个GO term指定为邻居术语，并通过整合邻居的语义信息生成正嵌入，$L_{Se} = \frac{1}{N}\sum_{j \in N}\left\{d(h, \tilde{h}^+)^2 - d(h, h_j^-)^2 + \alpha\right\}_+$，$L_{Se}$是由邻居正嵌入$\tilde{h}^+$得来的语义损失。

最后，将多重损失与二元交叉熵整合，最终的损失函数公式为：

$$L = \alpha(L_{St} + L_{Se}) + (1 - \alpha)L_{BCE}$$

4 复现细节

4.1 与已有开源代码对比

本文在论文中附有开源代码的网址，我自己参考着作者给出的开源代码自己重新复现了一遍代码。

4.2 实验环境

本研究在进行实验时，采用了一台配备Ubuntu操作系统的服务器硬件。为了充分发挥深度学习模型的性能，本研究使用了NVIDIA RTX3090 GPU，并借助CUDA和cuDNN等GPU加速库来加速模型的训练和推理过程。

本研究的深度学习框架选择了Pytorch，。该框架提供了丰富的工具和库，有效支持本研究所使用的模型结构和训练方法。本研究利用该框架的自动微分功能进行梯度下降优化。

在进行实验之前，本研究对数据集进行了预处理，以确保模型在训练过程中能够充分学习特征。本研究使用了CAFA3（Critical Assessment of Function Annotations3）数据集和TALE数据集 [3]。CAFA3数据集是一个生物信息学领域的挑战，旨在评估和预测蛋白质功能注释。

对比的方法有：DIAMONDScore，这是一个基于序列相似性的方法；DeepGoCNN、DeepGoPlus、TALE、TALE+是基于深度学习的方法，DeepGoCNN是一种使用卷积神经网络(CNN)作为序列编码器的平面方法。TALE是一种分层感知方法，利用一个transformer作为序列编码器，结合CNN学习联合嵌入序列的关系和GO术语特征了，DeepGoPlus和TALE+是他们所提议的模型(DeepGOCNN和TALE)按所需比例与DIAMODNScore的组合

4.3 创新点

本文介绍了一种名为GCL-GO的新型层次感知方法，该方法结合了图对比学习和蛋白质语言模型嵌入来表示GO术语，包括稀有的GO术语。不同于传统的图对比学习模型，GCL-GO专注于比较GO图的结构特征和BioBERT模型学习到的GO术语的语义特征。由于GO图是一个反映基因和蛋白质功能及其关系的本体论，破坏它可能会丧失其内在含义。因此，本研究通过对比结构和语义特征，学习GO术语之间的相关性。此外，还提出了GCL-GO+，它将GCL-GO与基于序列相似性的DIAMOND方法结合，实现了两种方法的协同。

5 实验结果分析

5.1 测试CAFA3数据集上的表现

首先模型在测试CAFA3数据集上的表现，因为本研究使用的GO术语比他们更多，所以本研究使用在GitHub上发布的代码重新训练了DeepGoCNN和TALE作为本研究的CAFA3训练数据集。在所有三个GO领域，GCL-GO+表现最佳。结果在表1中。

Method	$Fmax$			AUPR		
	MFO	BPO	CCO	MFO	BPO	CCO
DIAMONDScore	0.534	0.384	0.550	0.412	0.245	0.451
DeepGoCNN	0.532	0.456	0.640	0.446	0.328	0.613
DeepGoPlus	0.587	0.504	0.656	0.514	0.376	0.604
TALE	0.467	0.418	0.637	0.360	0.283	0.610
TALE+	0.532	0.498	0.643	0.396	0.369	0.581
GCL-GO	0.613	0.516	0.677	0.561	0.395	0.670
GCL-GO+	**0.637**	**0.540**	**0.681**	**0.570**	**0.413**	**0.660**

表1. 测试本文方法和对比方法在CAFA3数据集上的表现

5.2 测试模型在TALE数据集上的表现

测试模型在TALE数据集上的表现的结果在表2中，GCL-GO+在MFO和CCO中表现最好，在BPO中表现次之。

Method	$Fmax$			AUPR		
	MFO	BPO	CCO	MFO	BPO	CCO
DIAMONDScore	0.582	0.359	0.548	0.505	0.207	0.448
DeepGoCNN	0.476	0.266	0.616	0.419	0.162	0.563
DeepGoPlus	0.634	0.384	0.632	0.587	0.235	0.578
TALE	0.578	0.336	0.658	0.514	0.247	0.635
TALE+	0.667	**0.459**	0.677	0.604	**0.326**	0.643
GCL-GO	0.636	0.384	0.682	0.612	0.288	0.671
GCL-GO+	**0.686**	0.418	**0.686**	**0.679**	0.322	**0.686**

表2. 测试本文方法和对比方法在TALE数据集上的表现

5.3 测试GO术语下的通用性和可拓展性

最后一个实验测试GO术语下的通用性和可拓展性，通过计算每个测试序列GO术语的平均频率AF（average frequency）来了解GO术语的频率对性能的影响。AF的计算公式如下，其中$f(i) = \frac{1}{|N_{train}|}\sum_{j\in|N_{train}|} y_{i,j}$，$f(i)$是训练数据集中第i个GO术语的频率，$|N_{train}|$表示训

练数列的个数，$y_{i,j}$代表第j个训练序列的第i个GO术语，N_{yk}表示每个测试序列的功能注释数量,$y_{k,i}$代表第k个测试序列的第i个GO术语。

$$AF(k) = \frac{1}{|N_{y_k}|} \sum_{j \in N_{y_k}} f(i) \cdot y_{k.i}$$

根据[0,0.2)，[0.2,0.3)，[0.3,0.4)，[0.4,1]将CAFA3测试数据集划分为4个范围，实验最终结果在图2，图2显示了功能注释在4个频率范围内的性能。较低的范围意味着蛋白质具有训练数据中很少提及的功能注释。(a)为Fmax性能，(b)为竞争和本研究方法的AUPR性能。

在所有GO域中， GCL-GO和GCLGO+在[0,0.2)和[0.2,0.3)范围内都有出色的表现。 [0, 0.2)和[0.2,0.3)范围表示蛋白质具有训练数据集中很少提及的功能注释。这些结果证明了GCL-GO在深度GO术语下的优势。

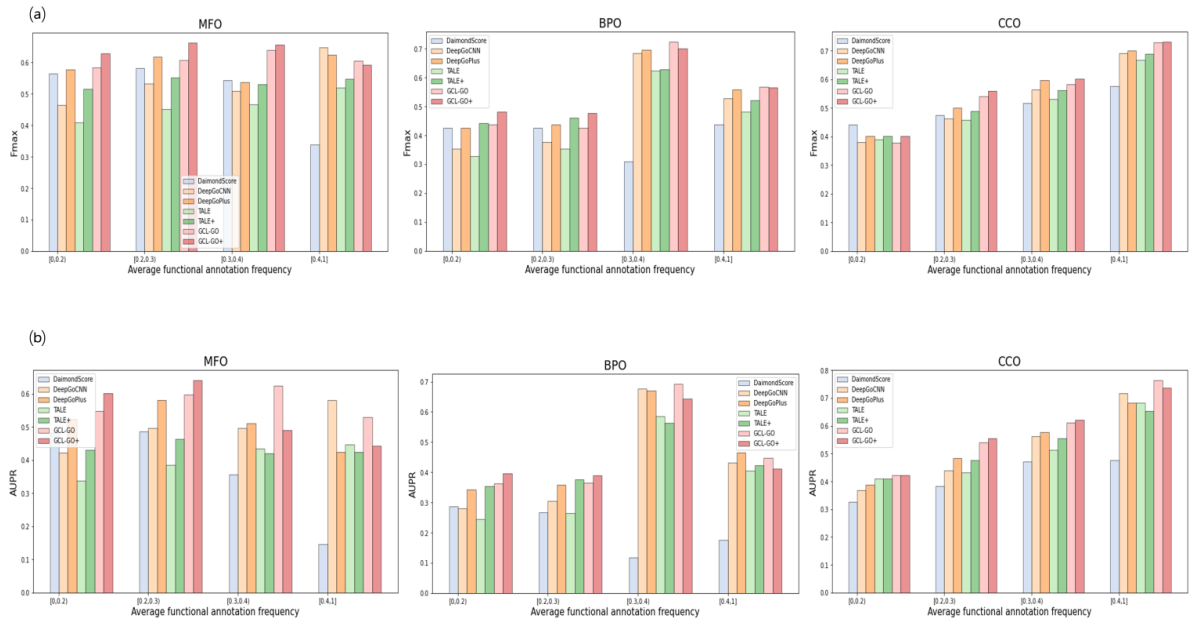

图 2 . 测试GO术语下的通用性和可拓展性的实验结果

6 总结与展望

本研究提出了GCL-GO，一种新型的基于序列的层次感知方法。这种方法使用ESM-1b模型来嵌入蛋白质序列， 并运用图对比学习技术来表示各种GO术语， 包括那些罕见或未知的GO术语。GCL-GO在CAFA3和TALE测试数据集上表现出色，其性能优于其他基于序列的方法。此外，结合了GCL-GO和DIAMOND技术的GCL-GO+版本，在这两个数据集上也展示了更好的性能。研究团队进一步测试了GCL-GO和GCL-GO+在处理罕见或全新GO术语方面的通用性和扩展性，希望这些模型能更有效地预测蛋白质功能及其相关的GO术语。

参考文献

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool

for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[2] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

[3] Yue Cao and Yang Shen. Tale: Transformer-based protein function annotation with joint sequence–label embedding. *Bioinformatics*, 37(18):2825–2833, 2021.

[4] Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7797–7805, 2022.