摘 要

多功能酶是具有可以催化多种基本化学反应的一种特殊类型的酶。研究表明，多功能酶可以以不同的形式来催化不同的化学反应，这使得多功能酶的研究价值和使用价值都要高于普通的单功能酶。传统的酶功能研究方法大都采用酶分析这一类的生物实验技术，对于多功能酶的分类以及数量快速增长的新酶的功能测定来说，基于生物技术的方法就显得十分耗时和代价昂贵。为应对上述挑战，近年来人们开始尝试借助机器学习等计算方法来处理酶功能分类问题。多功能酶预测问题本质上是一个多标签分类问题，目前已提出一些基于机器学习的多功能酶分类预测方法，且显示出了一定的有效性，但仍存在如下两个方面的重要挑战：一方面，已有方法几乎都只主要考虑了酶的序列特征，对于类别的标签特征并没有充分利用；另一方面，已有的多功能酶预测方法大多只考虑了多功能酶的主要类别的预测，而未能实现多功能酶完整 EC 编码的预测。针对上述挑战，本文进行了深入研究，提出了两种新的多功能酶分类预测模型。具体地，本文主要工作可概述如下：

（1）为充分利用酶类别的标签语义特征以及充分预测多功能酶的完整 EC 编码，提出了一种融合序列和多标签嵌入信息的多视角深度学习多功能酶预测方法 mlDGCnet (Multi-label Deep learning GCN-CNN net)。该方法引入了多视角学习、多标签分类机制和图卷积深度学习网络结构，通过提取酶序列的序列相关性特征和序列无关特征构建多视角特征集；同时使用图卷积网络对酶分类标签信息进行深度特征提取，并用于指导多视角学习过程；最终通过多标签分类器对多功能酶进行分类预测。相比于大部分现有的基于深度学习的多功能酶分类预测方法，本文提出的方法在多功能酶的各层 EC 码预测性能上均得到了一定提升。

（2）上述的多功能酶分类预测方法 mlDGCnet 虽然取得了较好的预测性能，但其在对酶序列进行提取特征时，学习的都是酶序列的局部特征。由于蛋白质的功能通常与其整体结构密切相关，因而，仅使用序列的局部特征很难获取有效的酶的整体结构和功能信息。另外，酶序列的局部特征的分布可能受到酶结构的变化影响，即酶在结构上的变化可能导致不同的局部特征分布，使得仅仅依赖局部特征难以实现酶功能的准确分类。针对此，在本文提出的 mlDGCnet 方法的基础上，对 mlDGCnet 中用于序列深度特征提取的 CNN-BiLSTM 模块进行改进，提出了一种融合局部和全局序列特征的多视角深度学习多功能酶分类预测方法 mlCBiGCnet。在该方法的序列特征提取部分，使用了带多头注意力机制的 CNN-BiGRUs 混合网络来对序列的深度局部特征和深度全局特征进行提取，以更好地捕捉多功能酶序列的整体结构信息，从而在一定程度上进一步提升模型的预测性能。实验结果表明，相比于上一个工作中的 mlDGCnet 方法，mlCBiGCnet 在 EC 编码每一层的预测性能又有了一定提升。

关键词：多功能酶分类，多视角深度特征学习，多标记信息辅助的特征学习，多标签分类，全局和局部特征

Abstract

A multifunctional enzyme is a special type of enzyme that catalyzes a variety of basic chemical reactions. Studies have shown that multifunctional enzymes can catalyze different chemical reactions in different forms, which makes the research value and application value of multifunctional enzymes higher than that of ordinary single-function enzymes. Traditional methods of enzyme function research mostly use biological experimental techniques such as enzyme analysis. For the classification of multifunctional enzymes and the function determination of the rapidly increasing number of new enzymes, biotechnology-based methods are time-consuming and expensive. In order to cope with the above challenges, computational methods such as machine learning have been attempted to deal with the problem of enzyme function classification in recent years. The multifunctional enzyme prediction problem is essentially a multi-label learning problem. At present, some classification and prediction methods of multifunctional enzymes based on machine learning have been proposed. Although the existing classification and prediction methods of multifunctional enzymes have shown certain effectiveness, there are still two important challenges as follows. On the one hand, almost all of the existing methods only consider the sequence features of the enzyme, and do not make full use of the label features of the class. On the other hand, most of the existing methods for predicting multifunctional enzymes only consider the prediction of the major classes of multifunctional enzymes, but fail to achieve the prediction of the complete EC coding of the multifunctional enzymes. In response to the above challenges, this paper conducts an in-depth study and proposes two novel multifunctional enzyme classification prediction models. Specifically, the main work of this paper can be summarized as follows:

(1) In order to make full use of the label semantic features of enzyme categories and fully predict the complete EC code of multifunctional enzymes, a multi-view deep learning multifunctional enzyme prediction method mlDGCnet (Multi-label Deep learning GCN-CNN net) combining sequence and multi-label embedding information was proposed. The method introduces multi-view learning, multi-label learning mechanism and graph convolutional deep learning network structure, and constructs multi-view feature set by extracting sequence correlation features and sequence independent features of enzyme sequences. At the same time, the graph convolutional network is used to extract the deep features of the enzyme classification label information, and it is used to guide the multi-view learning process. Finally, the multi-label classifier was used to classify and predict the multifunctional enzymes. Compared with most of the existing classification and prediction methods of multifunctional enzymes based on deep learning, the proposed method has a certain improvement in the

prediction performance of EC codes of each layer of multifunctional enzymes.

(2) Although the above multifunctional enzyme classification prediction method mlDGCnet has achieved good prediction performance, it only learns the local features of the enzyme sequence when extracting features of the enzyme sequence. Since the function of a protein is usually closely related to its global structure, it is difficult to obtain effective global structural and functional information of an enzyme by only using local sequence features. In addition, the distribution of local features of enzyme sequences may be affected by the changes in enzyme structure, that is, the changes in enzyme structure may lead to different distribution of local features, making it difficult to accurately classify enzyme functions only relying on local features. To this end, based on the mlDGCnet method proposed in this paper, the CNN-BiLSTM module used for sequence deep feature extraction in mlDGCnet is improved, and a multi-view deep learning multifunctional enzyme classification and prediction method mlCBiGCnet is proposed by fusing local and global sequence features. In the sequence feature extraction part of the method, a CNN-BiGRUs hybrid network with multi-head attention mechanism is used to extract the deep local features and deep global features of the sequence, so as to better capture the overall structural information of multifunctional enzyme sequences. Thus, the prediction performance of the model is further improved to a certain extent. The experimental results show that compared with the mlDGCnet method in the last work, the prediction performance of mlCBiGCnet at each layer of EC encoding is greatly improved.

Keywords： Multifunctional enzyme classification, multi-view deep feature learning, multi-label information assisted feature learning, multi-label classification, global and local features

目　录

第一章 绪论

1.1 研究背景及意义

酶[1]是所有生物体以及工业过程的关键组成部分，对加速基本的化学反应以及对抗疾病有积极作用。多功能酶（Multifunctional Enzyme）[2]是一类具有多个催化活性的酶，也被称为多催化酶（Multicatalytic Enzyme）或多反应酶（Multireaction Enzyme）。它们能够同时催化多个反应，包括不同的催化机制和化学反应类型，如氧化、还原、水解、缩合等。多功能酶在细胞代谢途径中扮演着重要的角色，能够将不同的代谢途径有机地结合在一起，促进代谢过程的高效进行。这些酶能够通过共享反应物和中间体的途径，完成多个生物合成和代谢通路中的多个步骤，从而起到协同调节和优化代谢途径的作用。与单功能酶相比，多功能酶在代谢通路中起到更加高效和紧密的调节作用，能够减少代谢物的浪费和能量消耗，同时提高代谢物的利用效率。多功能酶广泛存在于生物体中，包括细菌、真菌、植物和动物等各种生物。在细胞代谢途径中，多功能酶可以与其他酶共同组成代谢通路，从而实现代谢物的转化和利用。在人类体内，多功能酶的重要性也得到了广泛关注，例如人类的 Fatty Acid Synthase（FAS）[3]和 Polyketide Synthase（PKS）[4]等多功能酶被认为是肿瘤发生和发展的潜在靶点。多功能酶具有较高的结构复杂性和功能多样性，其中一些酶在同一多功能酶分子中的不同功能区域之间形成了协同作用，这些区域之间的相互作用对于多功能酶的催化活性和选择性起到了至关重要的作用。此外，多功能酶还能够将代谢过程中不同的反应物或中间产物直接转移，从而避免代谢通路中的堵塞。

多功能酶能以不同的功能和形式对生物体的生存、进化产生积极作用，还有助于提升利用生物体资源的效率，因此了解相关酶的功能就显得至关重要。下面将从三个方面来介绍多功能酶分类研究在不同应用领域的研究意义。

在酶工程方面，多功能酶分类具有重要的研究意义，有助于获得更多、更优质的多功能酶，并将其应用于酶催化反应、合成化学、生物转化等方面。首先，多功能酶分类研究能够帮助改进酶催化性能：多功能酶拥有多种催化活性，但其催化效率可能并不高。通过对多功能酶分类的研究，可以发现具有更高催化效率的多功能酶，并进一步优化其催化性能，以满足特定的应用需求。其次，多功能酶分类研究能够帮助获得具有多种催化活性的多功能酶：多功能酶可以同时催化多种反应，具有重要的应用价值。通过对多功能酶分类的研究，可以获得更多的多功能酶，并探索其在酶工程中的潜在应用，例如，将多功能酶用于制备具有高附加值的化合物。除此以外，多功能酶分类研究还能帮助获得具有新的催化功能的多功能酶：通过多功能酶分类的研究，可以获得具有新的催化功能的多功能酶，这些多功能酶可以用于开发新的酶催化反应或改进已有的反应。最后，有助于提高酶的稳定性和储存性：多功能酶分类的研究可以发现具有更好稳定性和储存

性的多功能酶，这对于酶工程中的大规模生产和应用具有重要意义。

在药物研发方面，多功能酶分类具有重要的研究意义，可以帮助研究多功能酶在药物研发中的应用。许多药物的作用机制涉及到多种酶的协同作用，多功能酶分类可以帮助识别多种酶，并提供药物研发的新靶点和思路。在药物研发中，多功能酶通常被认为是潜在的药物靶点。由于多功能酶可以催化多种化学反应，因此通过调节或抑制它们的活性，可以治疗多种疾病。例如，已有研究表明，多功能酶的活性与一些疾病的发展密切相关，如癌症、糖尿病、肝病等。其次，多功能酶分类和预测可以帮助研究人员确定潜在的靶点，加速新药的开发过程。通过对多功能酶的分类和预测，可以更好地理解其结构、功能和调节机制，从而设计出更加针对性的药物，并提高药物的效果和安全性。此外，多功能酶分类和预测还可以帮助预测药物的不良反应和副作用，从而加快药物的临床试验和上市进程。因此，多功能酶分类和预测在药物研发中具有重要的研究意义，对于促进医学科学的发展和推动药物研发进程具有积极的意义。

在生物进化方面，多功能酶分类具有重要的研究意义，多功能酶在生物进化中也起着重要的作用，多功能酶分类可以帮助研究不同物种之间的酶序列和功能的相似性和差异性，从而深入了解生物进化的过程和机制。一方面，能够揭示多功能酶的进化机制。多功能酶能够在同一催化中心上同时催化多个反应，为生物体提供了更多的适应性和生存优势。因此，研究多功能酶的进化机制对于理解生物进化的规律和机制具有重要意义。通过对多功能酶的分类、比较分析，可以揭示不同多功能酶之间的进化关系，推断它们的起源和演化路径，从而进一步了解生物进化的规律和机制。另一方面，能够探究多功能酶在生物适应性中的作用。多功能酶在生物体内具有多种催化活性，能够在复杂的代谢途径中协同作用，从而发挥重要的生物学功能。因此，研究多功能酶在生物适应性中的作用对于理解生物体对环境的适应性和生存策略具有重要意义。通过对多功能酶的分类、比较分析，可以揭示不同多功能酶在不同生物体中的分布情况和功能表现，进而推断它们在不同环境中的适应性和生存策略。

利用传统生物技术进行酶功能研究，费时费力且难以应对新酶数量的快速增长，对于多功能酶的研究更是如此。因此，有必要借助计算方法来研究解决方法[5]。

1.2 国内外研究现状

本节将从单功能酶和多功能酶分类预测的角度对酶功能智能预测的国内外进展进行回顾和分析。

1.2.1 单功能酶智能预测

随着机器学习[6-8]的发展以及机器学习应用场景的不断扩展，越来越多的机器学习模型，如支持向量机（SVM）[9-12]、K-近邻（KNN）[13, 14]、贝叶斯分类器[15-17]等，都被应用到了酶功能分类预测中。现已有许多基于机器学习的模型来预测酶的功能，代表性的方法有：EzyPred[18]、ECPred[19]、DEEPre[20]和 HECNet[21]。

EzyPred 方法采用自上而下的策略，构建了一个三层分类器，先预测酶的主要功能类别（如氧化还原酶、水解酶等），再进一步细分为更具体的亚类别（如醇脱氢酶、乙酰化酶等）。在第一层对输入的蛋白质序列进行酶与非酶的识别，第二层对酶的 6 大主要功能类进行识别，而第三层则进行 6 大功能类的子类的识别。该方法主要分为两个步骤：特征提取和分类器构建。特征提取步骤通过组合多种不同的特征（如氨基酸组成、位置特异性矩阵等）来描述酶序列，并使用主成分分析和线性判别分析等方法进行特征降维和选择。分类器构建步骤则使用随机森林算法[22-24]来训练模型，并采用交叉验证来评估模型性能。通过该模型方法，可以对酶序列进行高效准确的功能分类预测，有助于加速酶功能的研究和应用。EzyPred 融合了与蛋白质功能密切相关的功能域（FunD）信息和与蛋白质进化信息相关的伪位置特定评分矩阵（Pse-PSSM）信息，使用优化的证据理论 k 近邻（OET-KNN）分类器对酶功能进行识别，一定程度上提高了酶功能的预测成功率。

ECPred 是一种基于 EC 命名法的蛋白质酶功能预测工具，采用机器学习方法构建了多个二元分类器，其中每个二元分类器用于预测一个特定的酶 EC 码（酶功能）。该方法首先在 UniProt[25-27]数据库中搜索已知的酶序列来生成一个参考数据库。然后，利用已知的酶序列和其对应的 EC 号来构建一个多标签分类器，用于将未知序列分类到相应的 EC 号。具体来说，该分类器采用了基于随机森林（Random Forest）的多标签分类器，并通过提取蛋白质序列的各种特征，如 AAC（Amino Acid Composition）[28]、DPC（Dipeptide Composition）[29]、PSSM（Pposition Specific Scoring Matrix）[30, 31]等，将蛋白质序列表示成一个特征向量。最终，将该特征向量作为输入，经过多标签分类器输出一个 EC 号的概率向量，根据概率向量来预测蛋白质的酶功能。ECPred 结合了 SPMap[32]、BLAST-kNN 和 Pepstats[33] SVM 这三个独立的预测器，并使用集成预测的方法构建了分类器。其中 SPMap 基于子序列、BLAST-kNN 基于序列相似性、Pepstats SVM 基于氨基酸的理化特征进行预测。ECPred 分类器首先将输入的蛋白质序列分类为酶或非酶，紧接着对预测为酶的蛋白质序列进行主 EC 类的预测，确定 EC 主类后，ECPred 将进一步预测该酶的子类、子类的子类以及其底物类。ECPred 的预测精度经过多次实验验证，具有较高的准确性和可靠性。

DEEPre 分类器根据 EC 编码的树状结构，在其结构的每一个节点都构建了一个深度学习模型。对于第 0 级，采用一个模型用于预测是否为酶；在第 1 级，有 6 个模型，用于预测酶的 6 个主类。在第 2 级，同样有 6 个模型，分别对应于第 1 级的 6 大类，用于预测每个大类的子类。不同于以往的预测模型，DEEPre 在酶的特征选择上进行了创新，它使用了基于两种不同类型的蛋白质原始序列特征：序列长度相关特征和序列长度无关特征。利用这两种类型的输入特征，构造基于卷积神经网络(CNN)和递归神经网络(RNN)的深度学习分类器。DEEPre 使用的序列长度相关特征包含了蛋白质序列独热编码、溶剂可及性、二级结构，使用的序列长度无关特征包含了蛋白质功能域（FunD）。最终只需将蛋白质序列原始编码输入构建好的深度学习模型，DEEPre 便可对输入的序列进行特征提取和功能分类。

HECNet 提出了一种三连体网络框架来进行酶功能预测，其中包括了一种与 DEEPre 模型相似的网络用以处理具有大量酶数据的酶类，同时包含了一个暹罗三重网络（STNet）用以处理低层次水平上酶数据稀缺酶类。其中，STNet 由三个相同的子网络组成并且共享相同的权重。HECNet 使用 STNet 通过创建三元组来扩展数据，最终 STNet 将酶分类直到 EC 码的第四位。在 HECNet 模型中，对于其中几个 EC 类别，网络仅对第四级进行预测，并将第 2 级和第 3 级使用一个修改的三重损失函数合并成 EC 数的层次结构，以得到最终预测结果，这样会大幅度减少需要训练的模型数量。与 DEEPre 类似，HECNet 同样使用了序列长度相关特征和序列长度无关特征进行训练。

以上的这些代表性方法均通过提取酶的相关特征进行学习，进一步训练出分类器进行酶功能预测。但这些方法都有一个共同的局限性，这些方法都只能对单功能的酶进行分类预测，是一个单一标签的问题。研究表明，多功能酶可以以不同的形式来催化不同的化学反应，这使得多功能酶的研究价值和使用价值都要高于普通的单功能酶而对于多功能酶而言，其功能预测是一个多标签问题，以上的方法便不再适用。

1.2.2 多功能酶智能预测

近些年，针对于多功能酶分类[34-36]也已得到了初步的研究，代表性的方法有 EnzML[37]和 mlDEEPre[38]。其中，国内方面刘干提出了一种基于改进 PSSM 矩阵及二维 Gabor 变换局部特征提取的多功能酶预测方法。其方法提出了多重进化信息 PSSM 矩阵，并结合二维 Gabor 变换对其进行局部特征提取。通过 Gabor 变换可以将 PSSM 矩阵进行多尺度、多方向的分解，以获得 PSSM 更多的信息。最后使用随机 K 标签集成分类算法进行多功能酶分类预测。EnzML 是一种基于机器学习的多标签分类模型，用于预测酶的分类。该模型使用 InterPro 签名作为特征输入、使用多标签 k-最近邻（KNN）算法，通过多层感知器 MLP 模型对多功能酶进行预测。该模型首先通过 InterProScan[39, 40]将输入蛋白质序列的特征提取出来，例如域，反应物结合位点，功能注释等。然后将这些特征转换为二进制编码，用于输入 MLP 模型。MLP 模型包括多个隐藏层，每个隐藏层由多个神经元组成。在每个隐藏层之后，使用 dropout[41]技术以防止过拟合。最后一层是一个 Sigmoid[42]激活函数，用于预测多个酶类标签的存在或不存在。InterPro 签名是一个紧凑和强大的属性空间预测酶的功能。这种表示使得多标签机器学习在合理的时间内使用 Mulan[43]二元关联最近邻算法实现(BR-kNN[44])是可行的。

mlDEEPre 是单功能酶分类模型 DEEPre 的扩展，类似于 DEEPre 的结构，mlDEEPre 模型采用了层次化多标签分类的方法，先将输入序列分类为酶或非酶，并进一步将分类为酶的序列分类为单功能酶或者多功能酶。对于预测为单功能酶的序列，将其放入 DEEPre 分类器中进行功能预测；对于被预测为多功能酶的序列，mlDEEPre 将其分类为主要类别。mlDEEPre 模型采用了一个嵌套的深度学习架构，包括两个主要部分：1）基于自编码器的特征学习层，2）基于多标签分类器的层次分类层。模型利用基于自编码器的方法对酶的特征进行学习，以提高对酶功能的预测准确性。在特征学习层中，每个酶的氨基酸序列被编码成一个固定长度的向量，这个向量在后续的层次分类层中被用来

进行多标签分类。在层次分类层中，模型首先预测酶的大类功能，然后根据预测结果将酶分配到相应的子类中。层次分类器采用了一种新颖的损失函数，以在子类预测中保持一定的标签相关性，提高预测准确性。

1.2.3 问题与挑战

上述单功能和多功能酶分类预测的研究虽然显示出了一定的有效性，但仍存诸多挑战。特别地，对于本论文重点关注的多功能酶智能预测，在如下方面还有待深入研究：

（1）已有的酶功能预测方法都只主要考虑了酶的序列特征，对于类别的标签特征并没有充分利用。

（2）已有的多功能酶预测方法大多只预测了多功能酶的主要类别，即 EC 编号的第一位数字，且预测准确度都还有待提升。

（3）现有的多功能酶预测方法还存在没有充分利用酶序列全局特征的情况，可以通过全局特征和局部特征的使用来进一步提高预测精准度。

1.3 论文主要研究内容

针对上述现有多功能酶预测面临的挑战，本轮进行了深入的研究。具体地，本文的主要研究内容包含如下两个方面。

首先，针对 1.2.3 中的前两个问题，我们提出了一种融合序列和多标签嵌入信息的多视角深度学习多功能酶预测方法。在该方法模型中，我们首先对多功能酶序列进行初始系列特征抽取，包括酶序列长度相关性特征：氨基酸 One-hot 编码、蛋白质位置特异性矩阵 PSSM，以及长度无关性特征：蛋白质功能域 FuncD 两类特征。除此以外，我们使用词嵌入（Word Embedding）方法来提取酶 EC 类标签名之间的相关性特征。我们将 EC 类标签名的文字信息在 GloVe[45]模型上进行训练，生成标签相关矩阵，用以表示图节点信息。紧接着通过由 CNN 和 BiLSTM 组成的混合网络对上述三个视角的酶序列初始特征进行深度特征提取，与此同时，使用图卷积网络（GCN）[46]对由 EC 类标签名的标签相关矩阵进行深度特征学习，用以指导上述酶序列初始特征的深度特征提取过程。最后，利用训练好的多标签分类器对多功能酶进行分类预测。实验结果显示，加入 GCN 网络提取 EC 类名之间的相似度深度特征后，模型充分利用了酶标签名之间的相关性特征，在多功能酶分类预测方面性能提升较为明显。该模型对酶序列特征以及标签特征进行了充分的学习和利用，在相同数据集上的实验测试数据表明，相较于现有模型，我们的方法有了较为明显的性能提高。

进一步地，本文提出了一种融合局部和全局特征基于多视角深度学习结构的多功能酶分类预测方法。该方法是在上述第一种模型的基础上进行的改进。在第一个方法中的酶序列深度特征提取部分，我们仅使用了三个视角的局部特征，存在一定的局限性，例如：局部特征可能对酶的功能分类产生噪声，局部特征可能包含与酶的功能分类无关的信息，例如一些普遍存在于不同酶中的结构特征，这些特征可能会对分类产生误导，降

低分类的准确性。而全局特征相比于局部特征，可以涵盖整个酶分子的结构信息，包括其整体拓扑结构、构象和相互作用等，且全局特征不受蛋白质序列的局部变化影响，具有更好的鲁棒性和可靠性。因此我们考虑使用 CNN 和 BiGRU 组成的混合网络，对现有三个视角进行全局特征的抽取，同时结合局部特征进行深度特征提取，使模型具有更好的可解释性和更高的预测准确率。通过实验表明，使用新的由 CNN 和 BiGRU 组成的混合网络对酶序列进行全局特征和局部特征的深度特征抽取，使得相比于第一个方法提出的模型具有了更好的鲁棒性和可解释性，同时在分类预测指标上也有了进一步的提升。

1.4 论文组织结构

本文组织结构如下：

第一章为本文的绪论部分，主要阐述了多功能酶分类预测的研究背景和意义，同时介绍了当前在该研究方向上，国内外的研究进展和研究内容以及分析现有研究成果的可以改进的方向。紧接着概述了本篇论文所研究的主要内容，在本章最后，是本篇论文的组织结构。

第二章为本文的理论基础部分，主要介绍了基于深度学习的目标跟踪算法的基本原理。结合本文的研究内容，重点介绍了基于深度学习的目标跟踪算法的相关技术原理，这些技术包括：特征学习，数据关联，运动状态估计。

第三章为本文所提出的融合序列和多标签嵌入信息的多视角深度学习多功能酶预测方法，主要包含了对该方法以下几个方面的详细描述：1）算法框架，2）多视角初始特征抽取模块，3）多视角深度特征抽取网络模块，4）实验研究，5）结果分析。最后是对本章内容的小结。

第四章为本文所提出的融合局部和全局特征基于多视角深度学习结构的多功能酶分类预测方法，主要包含了对该方法以下几个方面的详细描述：1）算法框架，2）多视角局部特征抽取模块，3）多视角全局特征抽取网络模块，4）实验研究，5）结果分析。最后是对本章内容的小结。

第五章为本文的结论和展望部分，对全文研究的工作内容进行了总结，并指出本文所作研究的成果以及存在的局限性，最后对未来的研究方向和后续研究内容进行了展望。

第二章 相关生物信息学和机器学习理论基础

2.1 引言

相较于传统的酶分析[47]法进行酶功能的测定，将机器学习方法应用到酶功能分类中将更有利于提高多功能酶功能分类工作的效率。因此，本章将对在运用深度学习进行多功能酶分类任务中依据的相关生物信息学和机器学习基础理论进行阐述。2.2 节介绍了多功能酶功能分类的相关生物信息学理论基础，其中包括了酶功能分类的 EC 编码、多功能酶序列的线性氨基酸序列特征以及多功能酶序列的结构特征；2.3 节介绍了相关机器学习技术方法，其中包括了深度学习技术、多视角深度学习技术和多标签分类技术。

2.2 相关生物信息学理论基础

2.2.1 酶功能分类 EC 编码

EC 编码是一种用于分类酶的四级分类法，全称为“酶委员会编码（Enzyme Commission Number）[48]”。该编码由国际酶学委员会（International Union of Biochemistry and Molecular Biology）设立，并由其管理。

EC 编码由四个数字组成，用于描述酶催化反应的类型和催化剂。四个数字分别代表了酶在分类体系中的级别和酶催化反应的不同方面，如表 2-1 所示，第一个数字表示催化反应类型，包括了 7 个大类：

表 2-1 EC 编码第一位数字表示含义

层级	类型	英文名称	EC 编码
第一级别	氧化还原酶类	Oxidoreductases	1.x.x.x
第二级别	转移酶类	Transferases	2.x.x.x
第三级别	水解酶类	Hydrolases	3.x.x.x
第四级别	裂解酶类	Lyases	4.x.x.x
第五级别	异构酶类	Isomerases	5.x.x.x
第六级别	合成酶类	Ligases	6.x.x.x
第七级别	转位酶类	Translocases	7.x.x.x

第二个数字表示酶在特定催化类型中的子类别，用于描述酶所催化的具体反应类型，一般从 1 开始连续编号，例如：EC 1.1.x.x：氧化还原酶中，催化醛或酮物质和 NAD 或 NADP 的还原反应；EC 2.1.x.x：转移酶中，催化从一个基团向另一个基团的转移反应。第三个数字表示酶在特定催化子类别中的亚类别，同样从 1 开始连续编号，例如：第一类氧化还原酶中的 NAD(P)H 脱氢酶属于第二个亚类。第四个数字则表示亚型，通常是

一个整数或字母。对于某些类型的酶，第四个数字也可以表示其底物或产物的名称或结构。

EC 编码是酶功能分类的标准化体系，可用于描述酶催化反应的种类和特征，方便研究人员进行酶的分类和研究，通过 EC 编码系统，可以对酶的功能进行高效、准确的分类和命名，为研究酶的结构、功能和应用提供了重要的基础。

2.2.2 多功能酶序列的线性氨基酸序列特征

酶序列本质上是蛋白质序列，在蛋白质序列的一级结构特征中，氨基酸序列特征是最为基本和重要的一类[49, 50]。氨基酸是构成蛋白质的基本单元，每个氨基酸都有一个独特的侧链，这些侧链不同的化学性质决定了氨基酸的生化性质，从而决定了蛋白质的结构和功能。在酶序列中，氨基酸的类型、数量、顺序等特征会直接影响到酶的结构和功能。如表 2-2 所示，氨基酸有 20 种不同的常见类型，氨基酸序列特征可以通过多种方式表示，最常用的是单字母缩写表示法，例如"A"代表丙氨酸等。除此之外，还可以使用三字母缩写或全名等方式表示氨基酸。除了氨基酸的类型，它们在蛋白质序列中的排列顺序也是很重要的。一些序列中的氨基酸会组成一些具有特定生物功能的序列模式，例如信号肽、保守区域和结构域等。此外，氨基酸还与密码子相关联，三个不同的核苷酸（A、C、G、T/U）组成的一组被称为密码子，它们编码了不同的氨基酸。氨基酸密码子的不同组合方式决定了蛋白质的氨基酸序列和结构，从而决定了蛋白质的功能。

表 2-2 二十种氨基酸对应缩写遗传密码子

氨基酸	三字母	单字母	对应密码子
丙氨酸（Alanine）	Ala	A	GCU、GCC、GCA、GCG
精氨酸（Arginine）	Arg	R	CGU、CGC、CGA、CGG、AGA、AGG
天冬酰胺（Asparagine）	Asn	N	AAU、AAC
天冬氨酸（Asparticacid）	Asp	D	GAU、GAC
半胱氨酸（Cysteine）	Cys	C	UGU、UGC
谷氨酰胺（Glutamine）	Gln	Q	CAA、CAG
谷氨酸（Glutamicacid）	Glu	E	GAA、GAG
甘氨酸（Glycine）	Gly	G	GGU、GGC、GGA、GGG
组氨酸（Histidine）	His	H	CAU、CAC
异亮氨酸（Isoleucine）	Ile	I	AUU、AUC、AUA
亮氨酸（Leucine）	Leu	L	UUA、UUG、CUU、CUC、CUA、CUG
赖氨酸（Lysine）	Lys	K	AAA、AAG
甲硫氨酸（Methionine）	Met	M	AUG
苯丙氨酸(Phenylalanine)	Phe	F	UUU、UUC
脯氨酸（Proline）	Pro	P	CCU、CCC、CCA、CCG
丝氨酸（Serine）	Ser	S	UCU、UCC、UCA、UCG、AUG、AUC

苏氨酸（Threonine）	Thr	T	ACU、ACC、ACA、ACG
色氨酸（Tryptophan）	Trp	W	UGG
酪氨酸（Tyrosine）	Tyr	Y	UAU、UAC
缬氨酸（Valine）	Val	V	GUG、GUC、GUA、GUG

2.2.3 多功能酶序列的结构特征

PSSM（Position-Specific Scoring Matrix，位置特异性评分矩阵）[51]是用于描述蛋白质序列结构特征的一种方法。PSSM 特征[52-57]是从多序列比对中推导得到的，其本质是通过统计不同氨基酸在特定位置出现的频率，然后将这些频率转化成一个数值矩阵。该矩阵的每个元素代表了在一个给定的位置上，一个特定的氨基酸出现的概率与在整个数据集中出现该氨基酸的概率的比值，即表示该位置上这个氨基酸的特异性。

在生物学中，PSSM 特征常用于描述蛋白质序列的保守性和多态性。对于保守性而言，PSSM 特征可以用来预测蛋白质序列的保守域和功能域，以及酶催化位点等重要区域。对于多态性而言，PSSM 特征可以用来预测蛋白质序列的变异区域和疏水区域，从而揭示蛋白质序列的结构和功能特性。PSSM 特征能够提供关于蛋白质序列的结构和功能信息，对于许多蛋白质序列分析的任务（如蛋白质结构预测、蛋白质功能注释和分类等）都具有重要的生物学意义。

蛋白质功能域是指具有特定功能的蛋白质结构域，通常是由相对保守的氨基酸序列组成的。蛋白质的功能通常不是由整个蛋白质分子实现的，而是由其中特定的结构域完成的。这些功能域在不同的蛋白质中可能具有相似的序列和结构，表明它们在进化过程中具有一定的保守性。因此，识别和分析蛋白质的功能域可以为揭示蛋白质的结构和功能提供重要的线索。蛋白质功能域可以帮助我们了解蛋白质的结构和功能。通过对蛋白质序列进行功能域的识别和分析，可以了解到蛋白质的结构和功能信息，从而预测蛋白质的结构和功能。此外，功能域也是蛋白质的进化基本单元，对于研究蛋白质的进化关系和系统发育具有重要意义。功能域通常使用蛋白质序列比对和基于隐马尔可夫模型（HMM）的方法进行预测和注释。其中，Pfam 数据库是一个广泛使用的功能域数据库，它包含了大量已知的蛋白质功能域的 HMM 模型，可以用于预测蛋白质中的功能域。

2.3 相关机器学习技术理论基础

2.3.1 深度学习技术

深度学习是一种机器学习方法，其核心思想是通过构建多层神经网络来对输入数据进行学习和建模。与传统的机器学习方法相比，深度学习可以自动地从原始数据中提取和学习更高层次、更抽象的特征，从而实现更加准确和高效的模型训练和推理。可以用来解决复杂的模式识别和预测问题。深度学习模型通常包含多个神经网络层，这些层通过组合输入数据来学习复杂的特征表示，并可以在分类、回归、聚类等任务上进行预测。

深度学习主要包括以下几个方面的技术：神经网络模型设计、参数优化算法、模型评估和调优等。在神经网络模型设计方面，深度学习可以通过增加神经网络的层数、增加神经元的数量、改进激活函数等方式来提高模型的表达能力和泛化能力。在模型评估和调优方面，深度学习可以通过交叉验证、正则化、Dropout 等技术来防止过拟合和提高模型的泛化能力。

深度学习经典模型包括了感知机（Perceptron）[58, 59]、多层感知机（Multi-layer Perceptron，MLP）[60]、卷积神经网络（Convolutional Neural Network，CNN）[61]、递归神经网络（Recurrent Neural Network，RNN）[62]和生成对抗网络（Generative Adversarial Network，GAN）[63]。其中，感知机是最早的神经网络模型之一，用来解决二元分类问题。MLP 由多个感知机层组成的神经网络模型，可以解决更加复杂的分类和回归问题。CNN 通过卷积层、池化层等操作，可以有效地处理图像、视频等二维数据。RNN 可以处理序列数据，比如文本、语音等，通过记忆单元来保留过去的信息。GAN 由生成器和判别器两部分组成，可以用来生成逼真的图像、音频、文本等数据。

2.3.2.1 长短记忆网络 LSTM

对于 LSTM，其结构如图 2-1 所示，LSTM 采用门控机制来实现调节信息流，可以决定在何时对储存在神经元中的相关信息进行保留、更新或删除，在处理长序列数据时能够更好地保持梯度信息并避免梯度消失问题。具体来说，与普通的 RNN 相比，LSTM 引入了三个门（输入门、遗忘门和输出门）来控制信息的流动。

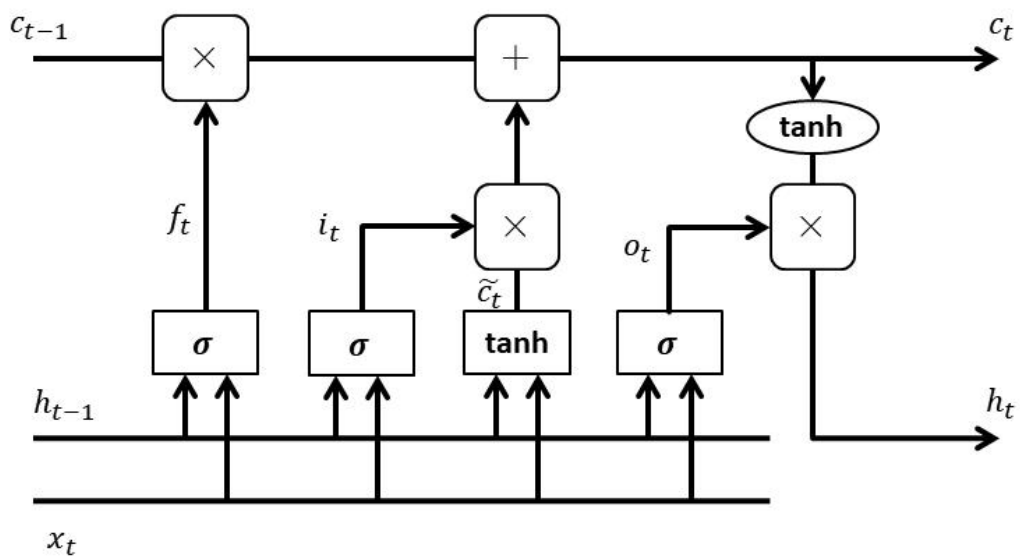

图 2-1 长短期记忆网络结构.

计算输入门的值 i_t：

$$i_t = \sigma\left(W_{xi} \cdot x_t + W_{hi} h_{t-1} + b_i\right) \tag{2.1}$$

其中，W_{xi} 和 W_{hi} 分别是输入 x_t、前一个时间步的输出 h_{t-1} 的权重矩阵，b_i 为偏置项；σ 是 Sigmoid 函数，用于将输入值映射到 0 和 1 之间的概率值。

计算遗忘门的值 f_t：

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{2.2}$$

其中，W_{xf} 和 W_{hf} 分别是输入 x_t、前一个时间步的输出 h_{t-1} 的权重矩阵，b_f 为偏置项；

更新记忆单元的值 c_t：

$$c_t = f_t c_{t-1} + i_t \widetilde{c_t} \tag{2.3}$$

$$\widetilde{c_t} = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{2.4}$$

其中，W_{xc} 和 W_{hc} 分别是输入 x_t、前一个时间步的输出 h_{t-1} 的权重矩阵，b_c 为偏置项，tanh 为双曲正切函数；

计算输出门的值 o_t：

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{2.5}$$

其中，W_{xo} 和 W_{ho} 分别是输入 x_t、前一个时间步的输出 h_{t-1} 的权重矩阵，b_o 为偏置项；

输出当前时间步的隐藏状态 h_t：

$$h_t = o_t \tanh\left(c_t\right) \tag{2.6}$$

LSTM 工作过程可以描述为：输入门 i_t 将接收的输入 x_t 和隐藏状态 h_{t-1} 通过公式(2.2)方式进行计算作为输入，紧接着通过计算公式(2.3)得到一个用于计算当前状态的候选值。f_t 表示遗忘门，控制从之前的单元 c_{t-1} 中遗忘多少信息，并通过 $f_t c_{t-1}$ 来构成 c_t 中的一部分。若经公式(3.3)计算 $f_t = 0$，则表示 c_{t-1} 没有信息需要传递给 c_t；若 $f_t = 1$，意味着 c_{t-1} 中的所有信息将被全部保留并传递给 c_t。c_t 是当前单元存储器，通过 $f_t c_{t-1}$ 计算来自 c_{t-1} 单元的信息保留量，同时通过 $i_t \widetilde{c_t}$ 来计算更新到 c_t 单元的信息量。o_t 为输出门，h_t 为 LSTM 单元的最终状态，前一个单元的隐藏状态、当前的输入以及当前单元状态共同决定了最终隐藏状态的输出。

2.3.2.2 注意力机制

注意力机制（Attention Mechanism）[64]是一种在深度学习领域广泛应用的技术，旨在让模型更加聚焦于输入数据中的重要部分，其结构如图 2-2 所示。相对于传统的全局池化方法，注意力机制更加灵活和精细，能够自动地学习到输入数据中的关键信息。注意力机制最初被应用于自然语言处理领域，以帮助机器翻译模型学习源语言和目标语言之间的对应关系。后来，注意力机制逐渐被引入到其他领域中，如计算机视觉、语音识别、推荐系统等，取得了很好的效果。注意力机制的主要思想是，对于每个输入，给予其不同的权重，以便模型更好地集中精力处理有用的信息。具体来说，模型通过计算每个输入与查询（Query）之间的相似度得到一个分数向量，然后将其输入到 softmax 函数中，得到归一化的注意力分布向量。最后，模型将输入和注意力分布向量加权求和，得

到注意力汇聚（Attention Pooling）表示。

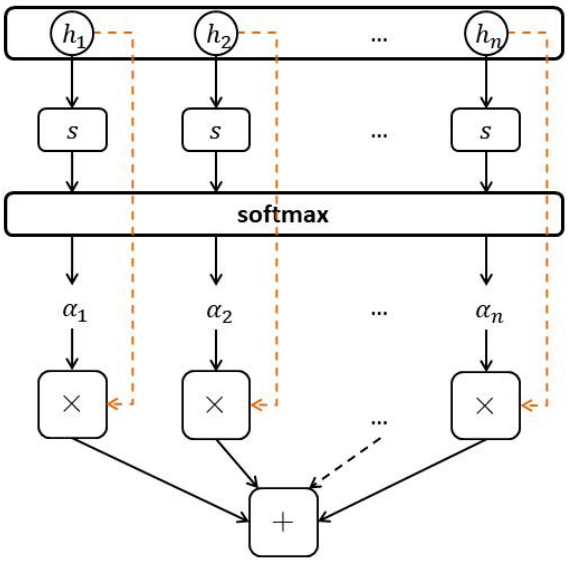

图 2-2 注意力机制结构.

常见的注意力机制包括单侧注意力、双侧注意力和多头注意力等。其中，单侧注意力只考虑查询与输入的相似度，而双侧注意力则同时考虑查询和输入之间的相互作用。多头注意力则是将注意力机制拓展到多个查询和多个输入之间，以更好地捕捉输入数据中的局部特征。注意力机制的应用范围非常广泛，常见的包括机器翻译、问答系统、语音识别、图像分类、推荐系统等。例如，在图像分类中，可以使用注意力机制让模型关注图片中的重要部分，而不是简单地对整个图片进行池化。在推荐系统中，可以使用注意力机制让模型注意用户历史记录中的重要项目，以更好地预测用户的兴趣。

注意力机制通常分为三个步骤：计算注意力权重，计算加权输入向量和最终预测。在第一步中，通过计算输入序列中每个位置的重要性得到注意力权重。这可以通过计算输入序列中每个位置的相似度得到，然后将相似度传递到 Softmax 函数中，得到每个位置的注意力权重。第二步是计算加权输入向量，这可以通过将输入序列的每个位置的向量乘以对应位置的注意力权重，然后将所有乘积相加得到。最后一步是通过将加权输入向量传递到最终预测层来得出最终的预测结果。对于注意力机制的三大主要部分：特征信息输入并计算相似度、注意力分布 α 、计算加权平均，Attention 层的计算公式如下：

$$O_i = s\left(h_i, b\right) \tag{2.7}$$

$$\alpha_i = \text{softmax}\left(O_i\right) = \frac{\exp\left(s\left(h_i, b\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(h_j, b\right)\right)} \tag{2.8}$$

$$Attention = \sum_{i=1}^{n} \alpha_i h_i \tag{2.9}$$

其中 O_t 为注意力得分，$s(\cdot)$ 为注意力打分函数，然后对 O_t 使用 Softmax 函数进行数值转换，以获得到注意力分布α，同时还可以进行归一化。最终使用注意力分布α对 h_i 进

行加权求和。分别对序列长度相关的两个特征经上述模型提取处理后，再将其输入到一个 RELU 激活的全连接层中进行维变换。

2.3.2 多视角深度学习技术

多视角深度学习[65-70]是一种利用不同视角的信息进行学习和决策的深度学习方法。在传统的深度学习中，通常使用一种或几种特征作为输入进行训练和预测。而多视角深度学习则利用多种特征视角，如图像的 RGB 通道、声音的频域和时域等不同视角，进行训练和预测，以提高模型的性能和泛化能力。在生物信息学中，多视角深度学习也被广泛应用于多种任务，如蛋白质结构预测、蛋白质功能预测和基因表达量预测等。例如，在蛋白质结构预测任务中，多视角深度学习可以同时利用蛋白质序列、进化信息和二级结构信息等多种特征视角，以提高结构预测的准确性和鲁棒性。

在深度学习中，多视角深度学习主要包含两个方面的技术：一是多层次特征融合（Multilevel Feature Fusion），通过不同的层次对多个视角进行特征融合，可以捕捉不同的特征层次信息，提高模型的鲁棒性和泛化能力；二是多任务学习（Multi-Task Learning[71, 72]），通过同时学习多个相关任务来共享底层特征，提高模型的效率和准确性。协同训练（Co-Training）[73, 74]、多核学习（Multi-Kernel Learning）[75]以及子空间学习（Subspace Learning）[76]等。

在多视角深度学习技术中，每个视角可以看做是一种特征表示，可以使用不同的方法来提取，比如 CNN、LSTM[77, 78]等模型。然后通过融合这些不同的特征表示来获得更全面的信息。多视角深度学习技术通常有以下几种实现方式：

1）串联模型：将不同视角的特征表示连接在一起，形成一个更大的特征向量，然后通过全连接层进行分类或预测。

2）并联模型：将不同视角的特征表示分别输入到不同的神经网络中，每个神经网络分别提取特征并进行分类或预测，最后将各自的结果进行组合得到最终结果。

3）集成模型：将不同的神经网络模型分别训练得到各自的预测结果，然后将这些结果进行加权或投票得到最终的预测结果。

2.3.3 多标签分类技术

多标签分类[79, 80]是一种机器学习技术，用于处理具有多个标签或类别的数据。与传统的单标签分类任务不同，多标签分类任务旨在将每个数据点与多个标签相关联。在多标签分类中，通常使用二进制分类方法。每个标签都被视为一个二元分类问题，因此分类器需要预测每个标签的存在或缺失。由于每个数据点可以有多个标签，因此多个分类器需要同时训练。这些分类器可以独立训练，也可以同时训练。常用的多标签分类算法包括基于决策树的方法、基于朴素贝叶斯的方法、基于支持向量机的方法、基于神经网络的方法等。其中，基于神经网络的方法，如多层感知机（MLP）和卷积神经网络（CNN），已经在多标签分类任务中取得了很好的效果。

ML-kNN 是一种基于 k 近邻算法的多标签分类算法，它的主要思想是利用相似度度量来预测每个标签的分类结果。ML-kNN 是一种基于实例的学习方法，通过利用训练集

中的实例来对新的实例进行分类。ML-kNN 模型的优点是简单、易于实现，并且可以有效地处理大规模的多标签分类问题。但是，由于它是一种基于实例的方法，对于噪声数据和高维稀疏数据集，其性能可能会受到影响。ML-kNN 算法模型的工作原理如下：

1）数据预处理：对于多标签数据集，首先需要对数据进行预处理，将每个实例的标签转换为二进制向量表示。例如，如果有 5 个标签，则每个实例的标签向量将有 5 个元素，其中每个元素表示该实例是否属于对应标签。

2）计算相似度：对于每个待预测的实例，使用距离度量（例如欧几里得距离、曼哈顿距离等）计算该实例与训练集中所有实例的相似度得分。

3）确定邻居数：根据预先设定的邻居数量 k，选取与待预测实例最相似的 k 个实例作为邻居。

4）标签预测：对于每个待预测实例，通过 k 个邻居的标签向量计算出每个标签的概率。

5）阈值设定：通过调整阈值，可以控制每个标签的预测结果的严格程度。例如：如果将阈值设为 0.5，则如果某个标签的预测概率大于等于 0.5，则将该标签分配给待预测实例。

Binary Relevance (BR)是一种常见的多标签分类方法，它将多标签分类问题转化为多个独立的二分类问题，每个二分类器用于判断一个标签是否存在。BR 的基本思想是将每个标签作为独立的二元分类任务，因此可以使用单标签分类器来解决问题。在 BR 中，对于每个标签，训练一个二元分类器，该分类器根据输入的特征，将样本分配给正类或负类，这里正类指该样本属于当前标签，负类则表示该样本不属于当前标签。在测试阶段，将输入的样本输入到所有二元分类器中，每个分类器都会返回一个二元分类结果，表示该样本是否属于当前标签。最终，将所有分类器的结果合并，得到每个标签的预测结果。具体地，假设训练集有 N 个样本，M 个标签，X_i 为第 i 个样本的特征，y_i 为该样本的 L 个标签的二元分类结果。对于每个标签 l，训练一个分类器 $f_l(x)$，其中 x 为输入的特征向量，$f_l(x)$ 的输出为二元分类结果 y_l，表示该样本是否属于标签 l。分类器分类器可以使用任何单标签分类器，如支持向量机（SVM）、逻辑回归（LR）、朴素贝叶斯（NB）[81-83]等。在测试阶段，对于输入的样本 x，通过所有二元分类器 $f_l(x)$，得到所有标签的预测结果 $\hat{y}$，最终的预测结果为 $\hat{y}=(\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_L)$。

BR 的优点在于简单易懂、易于实现，可以使用任何单标签分类器，适用于大多数多标签分类问题。但是，BR 忽略了不同标签之间的相关性，因此可能无法充分利用标签之间的相关性信息。此外，BR 可能存在类别不平衡的问题，即某些标签的正负样本比例非常不平衡，这可能导致分类器在这些标签上的表现不佳。

Classifier Chains (CC)是一种基于二元关系的多标签分类方法。它的主要思想是将标签之间的相关性建模为一个链式结构，并且利用这个链式结构来处理多标签分类问题。具体来说，CC 算法将每个标签看作是一个二元分类问题，并将标签链起来形成一个链式结构。假设有 M 个标签，那么就会形成一个长度为 M 的链式结构。在这个链式结构

中，第i个标签的分类器会把前$i-1$个标签的预测结果当做输入。CC 算法具体流程如表 3-1、表 3-2 所示。在训练过程中，CC 算法会训练M 个分类器，其中每个分类器都会负责预测标签链上的一个标签。在预测过程中，对于一个新的样本，CC 算法会先对第一个标签进行预测，然后将预测结果作为第二个标签的输入，继续预测下一个标签，直到预测完所有标签。

表 2-3 Classifier Chains 算法的训练流程

算法 1：针对训练集D和标签集L中的标签l进行训练
训练$(D=\{(x_1,y_1),\ldots,(x_N,y_N)\})$
算法过程：
1 for $j=1,\ldots,L$
2 do ▷第 j 个二分类和训练
3 $D'_j \leftarrow \{\}$
4 for $(x,y)\in D$
5 do $x' \leftarrow [x_1,\ldots,x_d,y_1,\ldots,y_{j-1}]$
6 $D'_j \leftarrow D'_j \bigcup (x',y_j)$
7 ▷训练h_j以预测y_j的二元相关性
8 $h_j : D'_j \rightarrow \{0,1\}$

表 2-4 Classifier Chains 算法的预测流程

算法 2：预测实例 x
CLASSIFY(x)
算法过程：
1 ▷global $h=(h_1,\ldots,h_L)$
2 $y \leftarrow [\hat{y}_1,\ldots,\hat{y}_L]$
3 for $j=1,\ldots,L$
4 do $x' \leftarrow [x_1,\ldots,x_d,\hat{y}_1,\ldots,\hat{y}_{j-1}]$
5 $\hat{y}_j \leftarrow h_j(x')$
6 return $\hat{y}$

由于标签之间的相关性被建模为一个链式结构，CC 算法能够捕捉到标签之间的依

赖关系，从而在多标签分类问题中表现出良好的性能。此外，CC 算法还具有良好的可扩展性，可以轻松地添加或删除标签。CC 算法的主要缺点是它无法处理标签之间的非线性相关性，因为链式结构无法捕捉到复杂的非线性相关性。此外，由于需要训练 M 个分类器，CC 算法的计算成本较高。

Label Powerset（LP）是一种基于转换的多标签分类方法，它将每个唯一的标签组合作为一个类别进行处理，并将多标签问题转换为单标签问题。具体而言，对于 L 个标签，LP 方法将它们的所有可能组合看作一个类，这样就得到了一个有 2^L 个类的单标签分类问题。在 LP 方法中，标签集合被视为有限状态自动机的状态，而状态转换是由输入特征和它们对应的标签触发的。因此，LP 方法允许将现有的单标签分类算法应用于多标签分类问题中。在 LP 中，训练集中的每个样本都被转化为一个向量，向量的每个元素对应于标签组合中的一个类别，如果该样本属于该类别，则该元素取值为 1，否则为 0。这些向量作为输入数据，可以用传统的单标签分类算法进行训练，例如朴素贝叶斯、决策树等。在预测时，将测试样本的预测结果向量与训练样本的向量进行比较，选择与测试样本向量最相似的训练样本向量对应的标签组合作为预测结果。LP 方法的主要优点是它可以使用现有的单标签分类算法，因此具有广泛的适用性和灵活性。此外，由于 LP 方法将多标签问题转换为单标签问题，因此它可以利用现有的单标签分类算法的优点，例如高效性和准确性。然而，由于标签组合的数量随着标签数目呈指数级增长，LP 方法在面对大规模标签集合时可能会面临计算和存储方面的挑战。

2.3.4 词向量模型 GloVe

GloVe（Global Vectors for Word Representation）是一个基于全局词频统计的词表征工具，它将一个单词表达成由实数组成的向量，该向量表征了单词之间一些语义特性，如相似性等。通过向量运算，如欧几里得距离，可以计算出两个单词之间的语义相似性。GloVe 算法的核心思想是将词的共现信息嵌入到词向量的学习中。具体来说，GloVe 算法在训练过程中通过计算全局词语的共现矩阵来捕捉单词之间的语义关系。该共现矩阵中的每个元素表示两个单词在同一个上下文中共同出现的次数。然后，GloVe 算法使用 SVD（奇异值分解）对该共现矩阵进行降维，从而得到每个单词的低维词向量表示。

相比于传统的基于词频的统计方法（如 LDA[84]、LSA[85]、Word2Vec）受到停用词、同义词和多义词等因素的影响，难以有效地捕捉单词的语义信息。而 GloVe 通过同时考虑共现词对之间的关系和全局词汇统计信息，能够更好地描述单词之间的语义相似度。GloVe 采用了一个加权平均的损失函数来优化模型，这种方法比起基于 Softmax 的损失函数更加稳定和快速。此外，GloVe 采用了基于线性缩放的学习率调整策略，可以使得模型在训练过程中更加平稳和快速。

GloVe 模型通过构造共现矩阵来捕捉词语之间的关系，同时将其转化为点积的形式，通过比较点积和词向量之间的关系来优化词向量的表示。具体来说，设 ω_i 和 ω_j 分别表示词汇表中的两个词，X 表示一个对称的共现矩阵，X_{ij} 表示 ω_i 和 ω_j 同时出现的次数，GloVe 模型将 ω_i 和 ω_j 的词向量分别表示为 υ_i 和 υ_j，然后通过如下公式计算它们的内积：

$$\omega_i^T \widetilde{\omega}_j = \sum_k \upsilon_i, k\upsilon_j, k \tag{2.10}$$

上述公式表明，ω_i和ω_j的内积等于它们词向量中对应维度上的值相乘之和。GloVe所构造的词向量与共现矩阵间的近似关系为：

$$\omega_i^T \widetilde{\omega}_j + b_i + \tilde{b}_j = \log X_{ij} \tag{2.11}$$

GloVe 模型的核心思想是：在这个内积中，两个词的线性关系，应该可以通过其他词对它们之间的共现频率进行解释。因此，GloVe 模型的优化目标是最小化以下损失函数：

$$J = \sum_{i,j=1}^{V} f(X_{ij})(\omega_i^T \widetilde{\omega}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{2.12}$$

$$f\left(X_{ij}\right) = \begin{cases} (\frac{X_{ij}}{x_{\max}})^{\alpha} & X_{ij} < x_{\max} \\ 1 & otherwise \end{cases} \tag{2.13}$$

其中，V 表示词表大小，b_i和b_j分别表示词汇ω_i和ω_j的偏置项，f 是一个权重函数，它的作用是对词频进行加权。

2.3.5 基于图卷积网络的标签特征学习

图卷积网络[86]（Graph Convolutional Network, GCN）是一种基于图结构数据的深度学习方法，适用于图节点的分类、聚类、链接预测、图表征等任务。GCN 的主要思想是在图结构中对每个节点和它的邻居进行卷积操作，从而提取节点在图中的特征表示。与传统的卷积神经网络不同，GCN 可以处理任意图结构，不需要事先设定固定的网络拓扑结构。GCN 基于谱图理论，利用图拉普拉斯矩阵对节点的邻居关系进行建模，将节点的邻居作为卷积核对节点特征进行卷积。相比于传统卷积神经网络，图卷积网络可以对节点和图进行表征学习，将图结构和节点特征结合起来表示，有利于图分析和挖掘。同时，相比于传统的图方法，GCN 可以使用深度学习方法进行端到端的学习，避免手动特征提取和复杂的预处理过程。

GCN 的工作原理如下：设 A 为邻接矩阵， W 为可训练的权重矩阵，H 为节点特征矩阵，那么 GCN 的基本公式为：

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{2.14}$$

其中，$\widetilde{A} = A + I_N$ 表示邻接矩阵加上自连接后的矩阵，I_N 为 $N \times N$ 的单位矩阵，$\widetilde{D}$ 为度矩阵，其对角线元素为每个节点的度数之和，$H^{(l)}$ 表示第 l 层的节点特征矩阵，$W^{(l)}$ 表示第 l 层的权重矩阵 σ 为激活函数，通常采用 ReLU 或 Sigmoid。GCN 的基本思想是将一个节点的特征表示为其周围节点的加权和，其中权重由邻接矩阵 A 和度矩阵 $\widetilde{D}$ 决定。

$\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ 可以看作是对邻接矩阵进行归一化，即使不同节点的度数相差不大，以避免高度连接的节点影响其他节点的特征。

图卷积网络（Graph Convolutional Network，GCN）最初是用于图像领域的深度学习模型，但随着研究的深入，人们发现它也可以用于处理图结构数据，例如文本分类。在传统的文本分类模型中，一般使用词袋模型将文本表示为向量，而忽略了单词之间的关系。然而，在某些任务中，例如情感分析和文本分类，单词之间的关系可能会提供重要信息。这就需要一种能够考虑单词之间关系的模型。GCN 正是一种能够处理图结构数据的模型，它通过学习节点之间的相对位置关系来获取节点的表示向量。在文本分类中，可以将每个单词看作一个节点，将单词之间的关系看作边，构建一张图来表示文本。然后，通过 GCN 模型学习每个单词的表示向量，并将它们汇总成整个文本的表示向量，最后将其输入分类器中进行分类。相比于传统的文本分类模型，使用 GCN 的好处在于它能够考虑单词之间的关系，因此能够更好地捕捉文本的语义信息。此外，GCN 还可以处理不同长度的文本，并且不需要预定义的词汇表，因此具有更广泛的适用性。因此，我们借用将图卷积网络用于文本分类的思想，将酶 EC 类标签的文本信息作为特征传入图卷积网络进行标签相关性的深度特征学习。

2.3.6 多标签分类

多标签分类（Multi-label Classification）是指一个样本可以属于多个标签中的一个或多个的学习任务。相较于传统的单标签分类（Single-label Classification），多标签分类更加贴近实际应用场景，比如图像标注、文本分类、药物作用预测等领域。多标签分类问题的难点在于每个样本可能属于多个标签中的一个或多个，因此需要考虑标签之间的相关性和依赖关系，同时需要选择合适的评价指标来衡量模型的性能。

在处理多标签分类问题时，常用的模型包括基于模型的方法和基于特征的方法。1）基于特征的方法主要包括特征提取和特征选择两个方面，其中，特征提取可以使用传统的特征提取方法，也可以使用深度学习技术，比如卷积神经网络 (CNN)、循环神经网络 (RNN) 等。特征选择则是在已有特征的基础上，选择最具代表性和区分度的特征进行分类，常用的特征选择方法包括基于过滤的方法、基于包装的方法和基于嵌入的方法等。2）基于模型的方法主要包括 Multi-Label k-Nearest Neighbors (ML-kNN)[87]、Binary Relevance (BR)、Classifier Chains (CC)[80]、Label Powerset (LP)等。其中，ML-kNN 是一种基于 KNN 算法的多标签分类方法，BR 是将每个标签独立看作一个二分类问题进行处理，CC 则是将标签之间的依赖关系建模为一个链式结构，LP 则是将所有标签的组合作为一个整体进行处理。

2.3.7 双向长短记忆网络 BiLSTM

BiLSTM 是一种双向循环神经网络，是 LSTM 的扩展形式。它通过增加反向循环层，处理输入序列的正反两个方向的信息。由于 BiLSTM 可以同时考虑过去和未来的上下文信息，因此在自然语言处理等序列建模任务中表现出色。与标准的 LSTM 相比，BiLSTM

包含两个方向的 LSTM 单元，一个按照时间顺序处理输入，另一个按照时间的逆序处理输入。这两个方向的输出在每个时间步骤被级联在一起，生成最终的输出。

BiLSTM 结构如图 2-3 所示，在双向 LSTM 中，正向 LSTM 按照时间顺序处理输入序列，而反向 LSTM 按照时间逆序处理输入序列。这样，每个时间步骤可以使用当前时间步的输入和前后时间步的上下文信息来计算隐藏状态。BiLSTM 的每个时间步骤包含以下步骤：

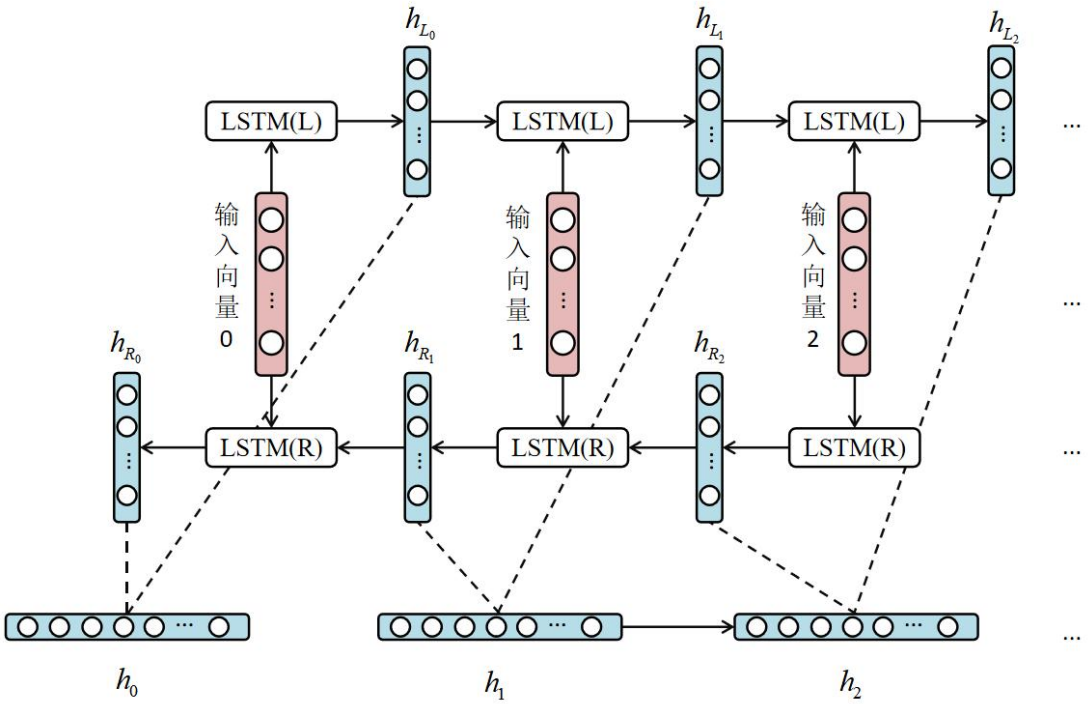

图 2-3 双向长短记忆网络 BiLSTM 结构.

1）输入门（Input Gate）：决定要更新当前时间步的单元状态，它将输入和前一个时间步的输出结合在一起，并传递到遗忘门和输出门。

2）遗忘门（Forget Gate）：决定保留或遗忘单元状态中的信息，由当前时间步的输入和前一个时间步的输出决定。

3）单元状态更新（Cell State Update）：使用输入门和遗忘门的输出，以及前一个时间步的单元状态来更新当前时间步的单元状态。

4）输出门（Output Gate）：根据当前时间步的输入和前一个时间步的输出决定要输出的信息。

5）隐藏状态（Hidden State）：根据当前时间步的单元状态和输出门的输出计算。

BiLSTM 可以用以下公式表示如下：首先定义输入序列为 $x=(x_1,x_2,\ldots,x_t)$，其中，$x_t \in R^{d_x}$ 为时间步 t 的输入向量，定义前向隐藏状态为 $\overrightarrow{h_t}$，后向隐藏状态为 $\overleftarrow{h_t}$，则有：

$$\overrightarrow{h_t} = \mathrm{LSTM}(\overrightarrow{h_{t-1}}, x_t) \tag{2.15}$$

$$\overleftarrow{h_t} = \mathrm{LSTM}(\overleftarrow{h_{t-1}}, x_t) \tag{2.16}$$

其中，LSTM 为标准的长短时记忆网络，根据上一时刻的前向/后向隐藏状态和当前时刻

的输入向量计算出当前时刻的前向/后向隐藏状态，具体表示为公式(3.2—3.7)。通过将前向和后向隐藏状态拼接得到最终的隐藏状态：

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{2.17}$$

其中，[·,·]表示拼接操作，最后，将得到的隐藏状态输入到一个全连接层进行分类或回归等任务。

2.4 本章小结

本章是整篇论文的理论基础部分，主要从两个方面简单介绍了所用到的理论知识。第一部分描述了多功能酶分类相关的生物信息学知识：一方面，对酶分类 EC 编码规则进行了详细介绍，另一方面，阐述了从序列顺序和结构两种类型的特征中提取的三个视角特征作为后续进行特征学习的内容，分别是序列氨基酸特征、序列位置特异性矩阵和蛋白质功能域。在第二部分中，对多功能酶分类所涉及到的机器学习方法理论进行了简单介绍。在我们的多功能酶分类预测模型中使用到的多视角特征学习技术以及多标签分类技术都在这一部分中进行了简要阐述。

第三章 融合序列和多标签嵌入信息的多视角深度学习多功能酶预测

3.1 引言

正如绪论 1.2.3 节所述，多功能酶智能预测还面临诸多挑战，其中第一、二个重要的挑战可进一步描述如下：（1）现有方法没有充分利用到酶 EC 标签的相关性特征；（2）现有方法在预测时往往只预测到了 EC 编码的前两位，并没有进行完整预测。针对上述挑战，本章提出了一种基于多视角分层深度学习和图卷积网络的多功能酶分类预测新方法。该方法的主要思路如下：

1）首先构建多功能酶序列的多视角初始特征，包括两种序列长度相关的视角的特征，即：酶序列 One-Hot 编码和蛋白质位置特异性矩阵 PSSM，以及一个序列长度无关的视角的特征功能域 FuncD，通过这两类酶序列特征进行了三个初始特征构建，此外，我们还进行了基于自然语言处理的酶 EC 类标签相关性特征构建，以用于指导酶序列特征的学习过程，在实验中，我们证实了酶 EC 类标签相关性特征在多功能酶分类预测工作上的有效性。

2）提出了多功能酶序列各视角的深度特征抽取方法：基于上述四个视角的初始特征，我们构建了一个包含 CNN-BiLSTM 网络和 GCN 网络两大核心模块的混合深度神经网进行深度特征抽取。其中 CNN-BiLSTM 网络用于抽取三个酶序列视角的深度特征，GCN 网络则用于抽取酶 EC 类标签相关性深度特征，以获取更具信息量的特征表达来进行进一步的分类预测任务。

3）提出了基于 GCN 网络的酶 EC 类标签之间的关联性特征抽取方法：多标签算法中，考虑标签之间关联性的算法是具有最优性能的，而作为一个多标签分类问题，现有的多功能酶预测方法中，几乎没有考虑到酶 EC 类标签之间相关性特征，因此在研究的模型中，我们利用自然语言处理技术构建了酶 EC 类标签之间的关联性矩阵，并通过 GCN 网络对酶 EC 类标签关联性的深度特征进行了抽取，用以指导上述三个视角的酶序列深度特征学习过程。

4）构建了适宜于多功能酶分类预测的多标签分类器：基于多标签学习技术，利用上一级网络提取到的多功能酶序列深度特征作为输入在多标签学习任务中进行分类学习，构建基于考虑标签之间关联性多标签分类算法的多功能酶分类器。

本章所做工作的主要贡献可归纳如下：

1）针对多功能酶分类预测问题，提出了一种新的深度学习模型。该模型对酶预测常用的三个特征：酶序列 One-Hot 编码、PSSM、功能域进行多视角特征提取。同时，使用图卷积网络提取 EC 类标签的相关性特征，并将其用于指导多视角特征学习过程，最终通过多标签分类器对多功能酶进行分类预测。

2）现有的酶功能预测模型大多只考虑使用了酶的序列特征，对于多功能酶来说，同个酶的不同功能之间同样存在一定联系，因此，我们在模型中使用图卷积网络对酶类标签进行相关性特征提取，用于指导多视角学习过程。

3）本文方法在 EC 编码第四层的子集精度能够达到 75.75%，其宏观 F1 参数能达到 90.41%，与现有方法相比，本文提出的方法在多功能酶的各层 EC 码预测性能上均得到了一定提升。

本章其余部分组织如下：第二部分、第三部分对所提方法的相关技术将进行了详细描述，在第四部分中，进行了详细的实验研究，以评估所提方法的有效性。最后一部分是对本章内容的总结。

3.2 融合序列和多标签嵌入信息的多视角深度学习多功能酶预测

3.2.1 模型框架

由于多功能酶分类预测是一个分层的多标签分类问题，本章提出了一个基于 EC 码树状分类结构的分层模型。其中，每一层的多标签分类模型整体框架如图 3-1 所示。每一层的 EC 类多标签分类器总共包含有六个部分，分别为：（1）多功能酶的多视角初始特征数据构建模块；（2）基于混合深度学习网络的酶序列深度特征提取模块；（3）酶 EC 类标签关系图构建模块；（4）基于图卷积网络的 EC 类相关性深度特征提取模块；（5）基于端到端的酶序列特征协同学习模块；（6）多标签分类器模块。模块（1）主要构造了多功能酶三个视角的初始特征，分别是酶序列 One-Hot 编码、蛋白质位置特异性矩阵（PSSM）以及功能域。在模块（2）中，我们是使用带注意力机制的 CNN-BiLSTM 网络来提取上述三个视角的深层特征。在模块（3）中，我们运用词嵌入方法，利用 GloVe 模型来构造 EC 类标签相关性特征，在后续 GCN 网络训练中，图节点分别表示每个不同的酶 EC 类标签，图节点之间的连线表示由 GloVe 模型提取到的标签相关性特征。模块（4）中对于每一层 EC 类的分类器，我们会构建一个与之对应的基于 l 个 EC 类标签的有向图 G，并对其应用图卷积网络进行 EC 类的标签嵌入，从而用以指导复合表征的学习过程。模块（5）为基于端到端的酶序列特征协同学习模块，在这个模块中，我们将训练好 GCN 网络输出权重与三个视角的深度特征进行融合，以构建新的复合特征，同时将新构建的复合特征与原有三个视角的深度特征进行拼接输入到后续分类器中。在模块（6）中，基于上述的多视角酶特征以及 EC 类标签相关图的标签嵌入学习，训练得到最终的多标签分类器，并对输入的测试数据进行分类预测。

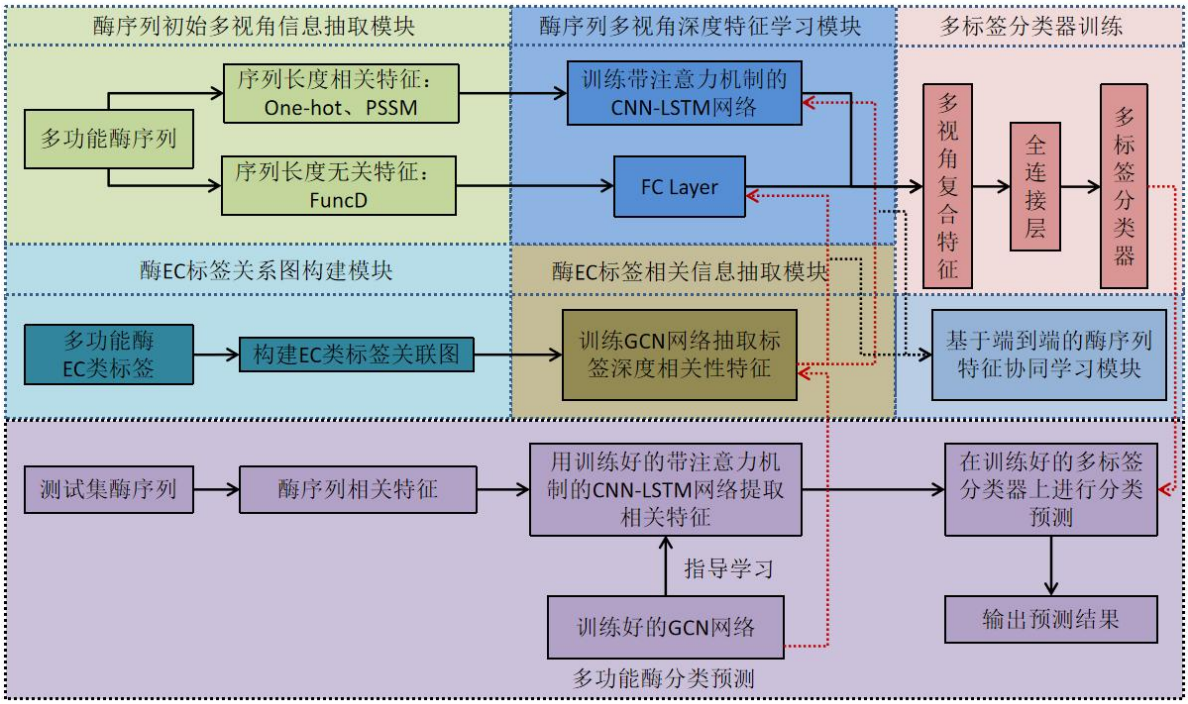

图 3-1 多功能酶分类模型框架.

3.2.2 酶序列初始多视角特征抽取模块

3.2.2.1 酶序列长度相关的视角特征抽取

在我们所研究的模型中，使用了酶序列长度相关性特征中的两个视角，分别是酶序列氨基酸 One-Hot 编码、酶序列蛋白质位置特异性矩阵 PSSM。

（1）酶序列氨基酸 One-Hot 编码 （视角 1 特征）

蛋白质是由许多氨基酸按一定顺序组成的多肽链经缠绕折叠形成的，酶作为蛋白质中的一种亦是如此。为了体现酶序列的氨基酸特性，我们使用 One-Hot 编码对序列进行特征表述。在酶序列中，One-Hot 编码可以用于将氨基酸序列转换为一个固定维度的向量表示，以便进行深度学习模型的训练和预测。这种编码方式可以保留酶序列中的信息，例如相邻的氨基酸之间的关系、序列中存在的结构域等，这些信息对于多功能酶分类预测任务非常重要，可以为深度学习模型提供足够的特征。此外每个氨基酸都有其独立的编码，没有重复的信息，不会影响模型的学习效果。在这种编码中，序列中的每个氨基酸被编码为一个长度为 20 的二进制向量，我们用 19 个 0，一个 1 共 20 位编码表示，其中只有对应的氨基酸位置上的二进制位为 1，其余为 0。这样，整个氨基酸序列就可以被表示为一个 $L\times 20$ 的矩阵，其中 L 表示序列长度。例如：丙氨酸的 One-Hot 编码可表示为 $(1,0,0,\ldots,0)$，谷氨酸的 One-Hot 编码可表示为 $(0,0,1,\ldots,0)$。

（2）酶序列蛋白质位置特异性矩阵 PSSM（视角 2 特征）

位置特异性矩阵（Position-Specific Scoring Matrix，PSSM）是在多序列比对的基础上，对蛋白质序列中每个残基进行分析的一种方法。在蛋白质分类中，PSSM 常用于提取蛋白质序列的位置相关信息，通过考虑不同位置上的保守性和变异性，更好地描述蛋白质序列的特征，进而提高分类的准确性。PSSM 的构建过程通常是先对一系列同源蛋

白序列进行比对，然后通过比对结果计算每个残基在不同位置出现的频率，再将这些频率转化为得分值。具体来说，对于每个残基，在每个位置上都会有一个得分矩阵，每个得分矩阵的大小为$L*20$（20 表示 20 种氨基酸类型，L表示蛋白质序列的长度），其中第i个残基在第j个位置上的得分为第i行第j列的值。这样，一个蛋白质序列就可以转化为一个大小为$L\times 20$的矩阵。PSSM 矩阵中的值可以代表在该位置上某个氨基酸出现的频率，以及该位置上某个氨基酸被替换的可能性，同时还能反映该位置的保守性，这些信息可以帮助分类器更好地识别酶序列的结构和功能，从而提高分类的准确性。此外，在进行蛋白质分类使时，PSSM 可以作为特征矩阵的一部分，与其他特征如 One-hot 编码等结合使用。

位置特异性矩阵 PSSM 能反应酶序列的进化信息，其可以通过比对算法来获取。本文通过 PSI-BLAST[88]算法，以数据集中的 2583 条多功能酶序列作为查询序列，以 Swissprot 作为数据库文件进行序列比对，经由多次迭代比对最终生成 2583 条酶序列的 PSSM 数据信息，以表征酶序列的进化特征。最终，每条酶序列通过执行 PSI-BLAST 程序生成的 PSSM 可编码为长度为$L*20$的特征向量，其形式如下：

$$PSSM=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} & \cdots & p_{i,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,j} & \cdots & p_{L,20} \end{bmatrix} \tag{3.1}$$

矩阵每一列表示 20 个氨基酸字母表中一项的对应残基类型，分别表示为{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}。$p_{i,j}$是一个分数，表示原氨基酸在进化过程中突变为j的几率; $p_{i,j}$的高分数通常表明突变频繁发生，该位置的相应残基可能是功能性的。

3.2.2.2 酶序列长度无关视角的特征抽取（视角 3 特征）

除了上述两个视角的酶序列相关性特征之外，我们还选择了一个酶序列长度无关性特征：蛋白质功能域 FuncD。蛋白质功能域是蛋白质序列中具有特定结构和功能的部分，是蛋白质功能的基本单位。在深度学习中，蛋白质功能域可以被看作是一种重要的特征，有助于提高蛋白质分类和预测的准确性。利用蛋白质功能域作为特征可以有效地提高蛋白质序列的表征能力。这是因为蛋白质功能域在不同的序列中具有一定的保守性和变异性，这种保守性和变异性可以通过深度学习模型进行学习和利用。此外，由于蛋白质功能域本身就具有特定的结构和功能，因此利用蛋白质功能域作为特征可以帮助深度学习模型更好地理解蛋白质序列的生物学意义，从而提高模型的预测准确性。在现有蛋白质功能分类方法中，通常会利用各种方法对酶序列进行功能域识别和注释，得到酶序列中的功能域信息。这些功能域信息可以作为输入特征送入深度学习模型，例如使用卷积神经网络或循环神经网络对蛋白质序列进行分类或预测。同时，还可以将蛋白质序列的功

能域信息与其他特征（如蛋白质序列 One-Hot 编码、位置特异性矩阵等）进行融合，以提高模型的表现。

在构建蛋白质功能域特征时，常常使用 HMMER[89]搜索 Pfam[90]数据库。Pfam 数据库是一个大规模的蛋白质家族和域的数据库，其中包含了许多已知的蛋白质域的信息，是根据多序列比对结果和隐马尔可夫模型表示的一系列蛋白质家族的集合。在构建功能域特征时，首先使用使用 HMMER（Hidden Markov Model-based profile）软件，将蛋白质序列与 Pfam 数据库中的 HMM 模型进行匹配，得到每个蛋白质序列匹配到的 Pfam 域以及相应的 E-value 和位于序列上的位置。然后根据匹配到的 Pfam 域和位置信息，可以将蛋白质序列转化为一个稀疏矩阵，其中每个 Pfam 域对应一个二元特征（存在或不存在），对应位置上的值为 1，其余位置上的值为 0，而这个稀疏矩阵就可以作为蛋白质的功能域特征，输入到深度学习模型中进行训练和预测。这种方式可以将蛋白质序列的每个位置对应到具体的功能域，具有较好的可解释性。使用 HMMER 搜索 Pfam 数据库构建蛋白质功能域特征，可以将蛋白质的功能信息加入到模型中，提高模型的分类性能，特别是对于多功能酶分类任务，更能发挥作用。

一个酶分子可能含有多个功能域，分别展示了不同的功能和进化信息。为表征酶分子的功能域特性，本文使用 HMMER 搜索 Pfam 数据库，对本文数据集中的 2583 个酶进行序列比对以寻找酶分子的功能域。PFam 数据库（Pfam 35.0，更新日期截止到 2021 年 11 月）中含有 19632 个条目，2583 个酶序列经过 HMMER 搜索后比对上的功能域共有 390 种。为此，本文使用 390 维矢量对搜索结果进行编码。如果某个序列的搜索结果报告对 390 种功能域的第 i 个条目比对成功，则 390 维矢量的相应位置置为 1，否则为 0。最终一条酶序列的功能域编码为如下形式的特征向量：

$$\text{Enzyme}_{\text{FuncD}} = \left[I_1, I_2, ..., I_i, ..., I_{390}\right] \tag{3.2}$$

其中，$I_i = \begin{cases} 1, & \text{当第}i\text{个条目搜索结果报告为比对成功} \\ 0, & \text{当第}i\text{个条目搜索结果报告为比对失败} \end{cases}$

3.2.3 酶 EC 类标签关系图构建

酶 EC 类标签名词嵌入是将酶序列中的酶名称转化为向量表示的过程，使用的是自然语言处理中的词嵌入技术。在词嵌入中，每个单词都被表示成一个向量，这个向量能够表达该单词在语言空间中的语义信息。类似地，酶类名也可以被看做是一组具有语义信息的单词，因此也可以通过词嵌入技术将其转化为向量表示。将酶类名转化为向量后，可以将其作为特征用于酶序列分类、酶功能预测等任务中，从而提高模型的性能。为充分利用酶 EC 类标签之间的相关性特征，以提高模型的预测性能。

在我们的方法中，在图卷积网络模块抽取标签相似度深度特征之前，对于每一层的分类模型，我们使用词嵌入（Word Embedding）方法，在 Wikipedia 数据集上训练了 GloVe-300d 来表示图节点（该层 EC 类的标签）的信息。当 EC 类标签名称中包含多个单词时，我们使用所有单词的嵌入平均值作为标签表示。最终我们得到一个大小为

$L\times300$ 的标签相关矩阵，其中 L 表示该层 EC 类标签数量（例如：EC 码第一层可分为 1-7 共七类，此时 $L=7$ ）。对于标签相关矩阵的构建，我们使用了经典的词向量模型 GloVe，其利用共现矩阵，结合词局部信息和整体的信息来训练词向量。

在公式（2.13）所示的权重函数中，我们将 α 设置为 0.75，以提高共现次数小的两个词的权重，进而提高低频词的词向量准确度， $x_{\max}$ 是共现次数的上限，设为 100。损失函数中的第一项 $\omega_i^T\widetilde{\omega}_j$ 表示 ω_i 和 ω_j 的内积，$\log X_{ij}$ 是对共现矩阵进行取对数的结果，该损失函数的目标是最小化模型预测的词向量内积与实际的内积之间的均方差，从而得到逼近共现矩阵的词向量表示。

3.2.4 基于 GCN 网络的深度酶 EC 类标签相关性特征抽取

对于酶 EC 类标签的特征提取部分，如图 3-2 所示，我们使用两个 GCN 层来完成。对所有 EC 类，其标签信息矩阵中的映射参数都是共享的，因此标签相关性可用其进行表征。GCN 网络部分将从 EC 标签的词嵌入中学习标签信息的相关性，并用于指导酶序列的特征学习过程。

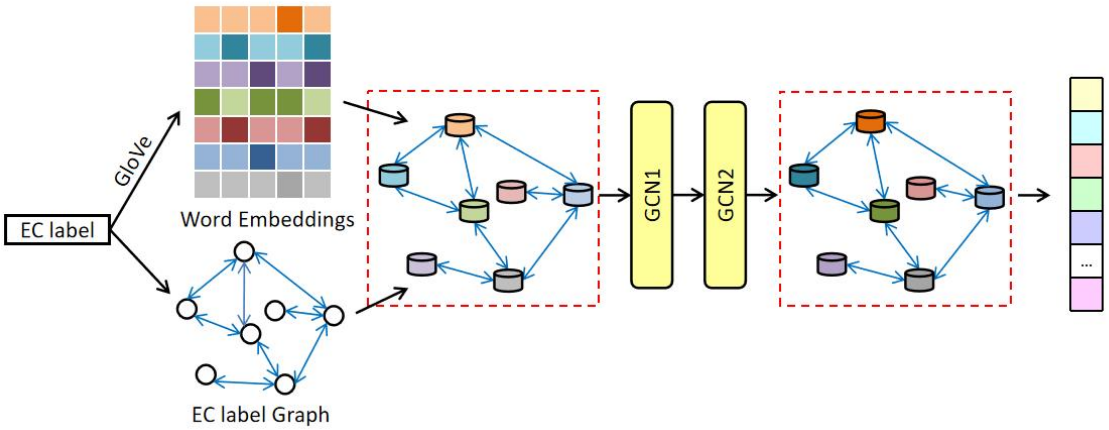

图 3-2 GCN 网络深度特征提取模块.

图卷积网络（GCN）是基于谱域图卷积的一阶局部近似。其神经网络结构是多层的。在传统 GCN 中，每层卷积层仅处理一阶邻域信息，多阶邻域的信息传递依赖若干卷积层的叠加实现。在训练时，GCN 将同时使用节点信息和结构信息。对于每一层卷积的定义，可表示为：

$$H^{l+1}=\sigma\left(\widetilde{P}\times H^{l}\times W^{l}\right) \tag{3.3}$$

在本文所构建的 GCN 网络中， H^{l+1} 表示所有节点在第 $l+1$ 层嵌入的标签， $\sigma(\cdot)$ 为激活函数，在我们的方法中采用的是 Leaky ReLU[91]。W^l 表示节点在第 $l+1$ 层的变换权重参数。P 是在做 EC 类标签词嵌入训练 GloVe 模型中的共现矩阵， $\widetilde{P}$ 则是 P 的邻接矩阵。在本文中， H^0 为经过 GloVe-300d 模型训练后得到的 EC 类标签相关矩阵，其维度为 $L\times300$，L 表示该层 EC 类标签数量。

对于 $\widetilde{P}$，由于没有对其进行归一化处理，所以可能会导致一些度数高的节点具有更

大的特征值，致使度数高低不同的节点在特征分布上显现出明显的差异，对特征提取结果产生影响，同时导致模型不正常的快速收敛。对于上述问题，可对原矩阵作傅里叶变换来解决，在$\widetilde{P}$经拉普拉斯矩阵变换得到的归一化矩阵后，GCN 的每一层卷积可重新表示为：

$$H^{l+1}=\sigma\left(\widetilde{D}^{-\frac{1}{2}}\times\left(\widetilde{D}-\widetilde{P}\right)\times\widetilde{D}^{-\frac{1}{2}}\times H^{l}\times W^{l}\right) \tag{3.4}$$

其中$\widetilde{D}$为图中节点的度矩阵，是一个对角矩阵，引入度矩阵的目的是提取节点自身的信息。

3.2.5 基于多视角深度学习的酶序列深度特征抽取

基于上述抽取到的多功能酶序列三个初始视角的特征矩阵，我门设计构建了一种混合神经网络用以提取相应的深度特征。该混合神经网络由 CNN 和 BiLSTM 组成，CNN 网络负责深度特征的抽取工作，得到更高层次、更抽象的特征表示。BiLSTM 循环神经网络可以保留多功能酶序列中的长期依赖信息，提取多功能酶序列中的语义特征，从而提高酶功能分类的准确性。混合网络 CNN-BiLSTM 则结合了上述两种网络的优势，有助于在后续学习中捕捉不同视角之间的相关性，同时有益于提升后续分类预测工作的准确度。此外，我们还构建了一个图卷积网络 GCN，用以酶 EC 类标签名相似度特征矩阵进行深度特征提取，该特征矩阵是由词嵌入模型 GloVe 训练得到的。上述两种深度神经网络对深度特征的抽取，旨在获得多功能酶相关特征有更高级别的抽象性和可泛化性的特征表示。

在我们的方法中，共提取了酶序列的两种特征，一种是序列长度相关性特征：One-Hot 编码、PSSM，其次是序列长度无关性特征：功能域。对于每一个序列长度相关性特征，我们使用 CNN 以及带有注意机制的双向长短期记忆网络（BiLSTM）网络提取相应深度特征。而对于序列长度无关性特征，我们将其直接传递给前馈神经网络的全连接层（FC layer）。

如图 3-3 所示为带有注意力机制的 CNN-BiLSTM 混合网络模型，在该模型中，我们重复使用了两组卷积神经网络（CNN）连接池化层（Max Pool）的结构，每组 CNN 由长度为 10 的 32 个卷积核组成，并使用 RELU 函数进行激活。对于 CNN 网络后连接的池化层，其窗口大小以及步长都设置为 3，用以聚合酶序列局部信息，并在将主要特征传递给 BiLSTM 层之前有效地向下采样(Downsamping)。在两组 CNN 连接 Pooling 的结构之后，我们将其输出传递给带有注意力机制的 LSTM 网络中。