分类号_____ 密级_____

UDC 注1 _____

南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

硕 士 学 位 论 文

基于蛋白质接触图的结晶倾向性预测

（题名和副题名）

王鹏浩

（作者姓名）

指导教师姓名 於东军 教 授

学 位 类 别 工学硕士

学 科 名 称 模式识别与智能系统

研 究 方 向 生物信息学

论文提交时间 2023 年 3 月

注 1：注明《国际十进分类法 UDC》的类号。

摘　要

X 射线晶体衍射是蛋白质三维结构测定的主要方法，在该方法中蛋白质结晶是最重要的步骤之一。由于不是所有的蛋白质都能轻易结晶，所以准确预测蛋白质的结晶倾向性能够指导结构测定的实验设计，还能够提高 X 射线晶体衍射实验的成功率。在本文中，我们基于蛋白质接触图对结晶倾向性预测方法进行了研究。蛋白质接触图描述了空间构象中残基之间的接触关系，可以看作一种相互作用网络，为了充分利用它携带的空间结构信息，我们在这项工作中使用了三种不同的方式来从接触图中提取与结晶高度相关的残基间相互作用特征，并利用该特征进行多阶段的结晶倾向性预测。

（1）设计了一种空间结构下的 K 间隔氨基酸对频率特征 CCmap-KAAP，并基于该特征和 XGboost 分类器构建了结晶倾向性预测模型 CCmapCrys。该特征描述了间隔为 K 的两个氨基酸构成的二肽在整个接触图中出现的频率，属于一种空间结构下的氨基酸组成成分特征。通过与其他基于机器学习的预测模型进行实验对比，CCmapCrys 取得了最优的预测结果；

（2）利用图注意力神经网络来自动提取接触图中的空间结构特征，并设计了新的预测模型 GCmapCrys。相比于手工特征，图注意力神经网络有更强大的特征提取能力。我们还使用了多种互补的蛋白质序列特征来进一步提高结晶预测的效果。在我们测试集上的实验结果表明，与其他最先进的结晶倾向性预测器相比，GCmapCrys 取得了最为显著的效果；

（3）使用图 Transformer 架构来进一步加强对全局氨基酸相互作用信息的提取，并设计了新的预测模型 GCmapTCrys。由于图卷积模型难以堆叠过深的层数，导致 GCmapCrys 模型无法提取更全面的结构信息，所以我们利用 Transformer 架构中的自注意力机制来提取全局的氨基酸相互作用信息。除此之外，我们还加入了残差连接和序列位置编码来增强 GCmapTCrys 模型的预测能力。通过与其他最先进的结晶预测模型进行对比，我们的 GCmapTCrys 模型取得了最先进的预测结果。而且我们还发现高精度的二级结构和相对溶剂可及性特征能够大幅提升模型的结晶倾向性预测能力。

关键词：蛋白质结晶倾向性预测，蛋白质接触图，K 间隔氨基酸对频率特征，图注意力神经网络，图 Transformer

Abstract

X-ray crystallography is the major approach for protein structure determination, in which crystallization is one of the most important steps. Since not all proteins can be easily crystallized, accurate prediction of protein crystallization propensity is critical to guiding the experimental design and improving the success rate of X-ray crystallography experiments. In this paper, we investigated the crystallization propensity prediction method based on predicted protein contact map. The protein contact map describes the contact relationships between residues and can be regarded as a kind of interaction network. In order to make full use of the spatial structure information, we used three approaches to extract residue interaction features that are highly relevant to crystallization from the contact map, and then used these features to predict crystallization propensity.

(1) We designed a K interval amino acid pair frequency feature CCmap-KAAP under spatial structure, which belongs to the feature of amino acid composition under a spatial structure and describes the frequency of dipeptides composed of two amino acids separated by K in the entire contact map. We constructed a prediction model CCmapCrys based on this feature and XGboost model. By comparing experiments with other machine learning-based prediction models, CCmapCrys achieved the best prediction results.

(2) We used graph attention network to automatically extract spatial structure features from contact map and designed a new model GCmapCrys, which has a more powerful feature extraction capability. We also used multiple protein sequence-based features to further enhance crystallization prediction. Experimental results on our test datasets showed that GCmapCrys achieves the most significant results compared to other state-of-the-art crystallization propensity predictors.

(3) We used the Graph Transformer architecture to further enhance the extraction of global residues interaction information and designed a new prediction model GCmapTCrys. It is difficult for the graph convolution model to stack too deep layers, which will cause the GCmapCrys model to fail to extract more comprehensive structural information. So, we use the self-attention mechanism in the Transformer architecture to extract global amino acid interaction information. We also added skip connection and sequence positional encoding to enhance the prediction of GCmapTCrys. In the end, our GCmapTCrys model achieved the most advanced prediction results. And we found that the high-precision secondary structure and relative solvent accessibility features can significantly improve the crystallization prediction.

Keywords: Protein crystallization propensity prediction, Protein contact map, K interval amino acid pair frequency feature, Graph attention network, Graph Transformer.

目 录

图表目录

1 绪论

1.1 研究背景

蛋白质有机大分子是生命活动的主要承担者，从其结构特征入手分析其功能作用是非常重要的。生命体内的基因规定了蛋白质的一级氨基酸序列，但是一级序列只有折叠成特定的三维空间构象才能具有相应的生物活性和功能。从结构的角度入手研究蛋白质不仅能够帮助人们认识到氨基酸序列是按照何种规则折叠成三级结构的，也有利于认识蛋白质是如何执行其功能的，这对现实生物、医疗行业有重要的指导意义。

测定蛋白质三维结构常用的生物实验手段包括 X 射线晶体衍射[1]（X-ray crystallography）、核磁共振（NMR）和冷冻电镜等。目前蛋白质结构数据库中由 X 射线晶体衍射技术测定的样本接近 90%[2]，这表明了 X 射线晶体衍射技术在蛋白质结构测定中拥有非常重要的地位。X 射线晶体衍射技术是首先将一定波长的 X 射线照射到高质量的蛋白质晶体上，之后 X 射线会通过规则排列的蛋白质晶体从而产生与晶体结构相对应的特有衍射图像[3]，最后人们就可以通过观察衍射图中的信息来推断蛋白质大分子的三维结构。因此要想通过 X 射线晶体衍射技术测定蛋白质结果，蛋白质结晶是必不可少的步骤之一。

在目前的生物实验中，蛋白质结晶是一项非常耗时且复杂的任务。蛋白质晶体一般是蛋白质在饱和溶液中逐步析出的结构化、有序化的晶体，研究人员需要选择适当的溶剂，并通过各种实验手段和试剂来严格控制溶液的饱和度以及系统的稳定性[4]。因此想要选定的蛋白质能够形成稳定的高质量晶体就需要对各种实验变量进行调节，以达到能够形成稳定晶体的严苛实验条件。这种不断试错的实验方法就导致了蛋白质结晶成为 X 射线衍射结构测定流程中最耗时的一个步骤，严重影响了整个结构测定流程的速度与质量。随着测序数据的增多，大量的序列的结构需要被解析，必须尽可能使用一些方法来减少结晶过程中不必要的时间和成本消耗。

在这样的问题背景下，从结晶过程数据中探寻蛋白质结晶的内在规律是提高结晶试验效率和成功率的最佳选择[5]。结晶倾向性预测是从蛋白质序列本身的物理、化学等内在特征属性判断其结晶倾向的问题，一般是分为可结晶与不可结晶两类。研究蛋白质内在结晶倾向性虽然不能指导结晶实验条件的选择，但是可以把候选蛋白质限制在那些具有可操作性的未知结构蛋白质上，并根据预期的兴趣和结晶倾向性确定优先次序，从而增加成功率、减少成本消耗以及缩短结构测定的时间[6]。对结晶倾向性的研究也能帮助人们更深层次地理解结晶与蛋白质内在属性的联系，加深对蛋白质的理解。在具体的结晶实验过程中，还可以划分为材料生产、纯化和晶体产生三个大致步骤，除了从头预测蛋白质结晶成功的概率之外，我们还预测了材料生产失败、纯化失败和晶体产生失败的

概率，这有助于更清晰地了解到蛋白质结晶失败的可能原因。在预测蛋白质结晶倾向时，已知的是蛋白质的一级序列结构，我们首先提取与结晶相关的特征因素，然后再利用这些特征预测候选蛋白质的结晶倾向性，如图 1.1 所示。

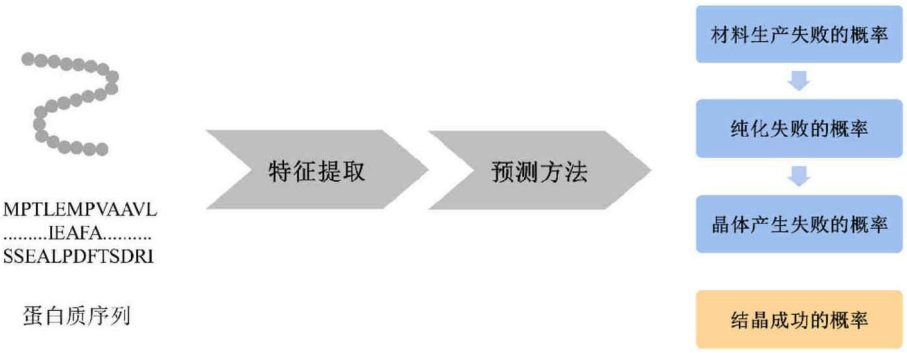


图 1.1 预测蛋白质结晶倾向性

1.2 影响蛋白质结晶倾向性的因素

蛋白质结晶困难的根源在于结晶的复杂物理化学性质，导致实验设计难以标准化，一个蛋白质在实验上能够形成稳定的晶体不仅严重依赖于蛋白质试剂浓度、温度、pH 值以及各种其他外部实验因素，还依赖于蛋白质本身的固有属性，因此在没有完整的结晶理论指导下，在实验上根据各种因素从头预测蛋白质是否结晶几乎是不可行的[5]。而且从数据的角度来看，结晶实验步骤中的各种实验信息也没有统一的标准，难以从统计学角度对完整的实验数据进行挖掘。因此目前更多的结晶预测问题是找到蛋白质固定的已知特性与成功结晶的相关性或分布关系，然后再利用这种分布进行结晶筛选，从而提高结晶效率。

Canaves 等人[7]对 Thermotoga maritima 蛋白质组上的结晶数据进行分析，以确定与结晶成功相关的蛋白质特性有哪些。他们对多个蛋白质物理化学参数进行筛选并最终过滤出了七个对蛋白质结晶有直接影响的序列衍生参数，包括蛋白质长度、计算的等电离点、带电残基百分比、平均疏水性、SEG 残基数量、预测的跨膜螺旋数量和预测的信号肽的数量等。Goh 等人[8]使用决策树进行分析发现了与结构基因组学管道中多个阶段（克隆、表达、纯化和结构确定）相关的蛋白质特性，包含序列保守性、带电残基百分比、疏水性、蛋白质结合伴侣的数量和蛋白质长度等。随着数据量的积累和技术的不断发展，Price 等人[9]通过 Northeast 结构基因组学联盟的大规模实验结果来研究控制蛋白质结晶的生物物理特性，结果表明结晶倾向性主要受控制蛋白质相互作用的有序表面特征的影响，具体来说主要包括暴露侧链的平均熵、预测的主链无序性、五种氨基酸（Ala，gly，phe，glu，lys）频率以及平均疏水性等。而且该文献还进一步表明了尽管结晶倾向性可

以受到溶液试剂影响，但它仍是一种内在的蛋白质属性。综上所述，由于受到不同的实验因素影响，这些文献得出的影响蛋白质结晶的内在因素可能不是完全正确的，但是仍然可以指导后续蛋白质结晶倾向性预测的研究方向，目前大部分蛋白质结晶预测方法也的确是从各种蛋白质内在因素入手分析它们和结晶倾向之间的相关性。

1.3 国内外研究现状

在蛋白质结晶倾向性预测方向中，研究者在过去的一段时间里提出了很多不同的预测模型，也都取得了较好的预测性能[10, 11]。这些预测模型大部分都是基于统计或机器学习算法进行开发的，着重于研究不同蛋白质物理化学性质与结晶之间的相关性。一般步骤是先提取各种不同的蛋白质特性，并将其整合成固定长度的特征向量，再将该特征向量作为输入送入到机器学习分类器中预测结晶倾向性。中间也会加入不同的特征选择方法来剔除一些无关特征并筛选出与结晶高度相关的蛋白质特性。这种基于蛋白质先验特征和机器学习的方法一直是研究蛋白质结晶倾向的主流方法，也取得了非常显著的预测性能。不过早期也有一些方法通过统计分析的方式在结晶数据集中寻找蛋白质特性与成功结晶之间的分布情况，以此来确定哪些蛋白质特性与结晶是高度相关的[12]。随着深度学习技术的发展和结晶实验数据的积累，一些研究者想要使用深度学习的方式直接从蛋白质序列本身中自动学习到与结晶相关的高层次特征[13-15]，但是从目前的结果来看，这些仅从蛋白质序列入手使用深度学习的端到端预测方法并没有取得过于优越的结果，再加上深度学习本身具有难以解释的性质，所以目前预测结晶倾向性的主流方式仍是优先将多种蛋白质特性作为模型输入，再使用机器学习或深度学习的方式进行结晶预测，不仅能够提高模型的预测性能，还能探寻各种蛋白质先验特性和结晶倾向性之间的关系。

1.3.1 基于统计分析的预测方法

如果想要了解某些蛋白质特性与结晶倾向性之间的关系，最简单的方法之一就是直接从结晶数据集中挖掘这些蛋白质特性的分布情况。

比如，Overton 和 Barton 等人[12]使用统计分析的方式分析了蛋白质结晶和等电离点（pI）、平均疏水性（Gravy）之间的关系。他们从 TargetDB 数据库[16]中分别提取了一部分可结晶的、不可结晶的和可溶的蛋白质并组成了三个数据集，还单独从 PDB 数据库[17]中抽取了一些已知结构分辨率小于 3Å 的蛋白质组成另外一个数据集 Dbrack_PDB。然后作者分别在这四个数据集上统计 pI、Gravy 的平均值和分布范围，以此来判断这两个蛋白质特性在可结晶与不可结晶的蛋白质数据集上是否有显著的差异。作者通过多次采样分析各个数据集上 pI-Gravy 联合特征的分布计算出了对应的 Z-Score 矩阵，并将 Dbrack_PDB 数据集上计算出的 Z-Score 矩阵称为 OB-Score，较高的 OB-Score 分数代表该蛋白质更易于结晶，低的 OB-Score 分数则相反。该项研究主要表明了 pI-Gravy 联合

特征与蛋白质结晶之间拥有较高的相关性。

ParCrys[18]是 Overton 另一篇从统计分析角度研究蛋白质结晶倾向性的方法，与之前的 OB-Score 方法相比主要有两方面的改进，一是考虑了更多的蛋白质特征，包括低复杂性区域的数量、平均疏水性、等电离点和标准氨基酸频率；二是使用 Parzen Window 概率密度函数[19]从不同的数据集中描述这四种蛋白质特征与蛋白质结晶之间的分布关系。

Slabinski 等人提出了 XtalPred 方法[20]，这种方法的思路比较简单直接，先根据经验与之前的文献选择出一些与结晶高度相关的蛋白质特征，再单独计算每种特征与成功结晶之间的概率分布，最后使用对数意见池（logarithmic opinion pool）方法[21]将多种特征的概率分布结合起来，多种概率的结合可以看作在不同权重影响下多种独立的特征概率相乘，最终得出多种特征影响下成功结晶的概率。这种方法的一大特点是尽可能模仿结构生物学家的工作，努力考虑每个可能对结晶造成影响的因素，并将这些因素结合起来判断结晶的概率，该模型最终输出的结晶概率被分为五个等级：最优、次优、一般、困难和极其困难。XtalPred 方法虽然没有使用复杂的预测模型，但是提供了一个比较好的预测基准，而且也取得了一定的预测性能。

上述三种基于统计分析的方法都在早期取得了不错的预测性能，而且从构建的分布函数中的确可以非常清晰地观察到结晶与某些蛋白质特性的相关性，但是这种方法也有较大的缺陷，首先是需要手动构建合适的分布函数，不合适的分布函数也许并不能将影响较大的蛋白质特性分离出来；其次，蛋白质结晶可能是受多种特性结合在一起控制的，不一定是各种特性独立地影响蛋白质结晶，所以这种基于统计分析的方法在预测性能上存在比较大的限制。

1.3.2 基于蛋白质特性和复杂网络模型的预测方法

相比于基于统计分析的方法，机器学习和深度学习的强大表征能力在语音识别和图像识别等领域都有所体现，因此目前更多的预测模型是使用机器学习或深度学习的方式来学习蛋白质特性与结晶之间的关系。

早期 Smialowski 等人[11]基于 SVM 和朴素贝叶斯构建了一个二级分类器 SECRET 来预测蛋白质结晶倾向性。该文献首先指出了预测蛋白质结晶倾向性的一个阻碍，那就是蛋白质结晶负样本的选择。不同于其它领域中对负样本的划分是理所当然的，在蛋白质结晶数据库中，许多没有成功结晶的蛋白质并没有被标记失败的原因（比如测序失败、克隆失败、表达失败等等），这就意味着很多蛋白质也许能够结晶但是由于其他外部因素从而导致实验停止。因此如果想要取得精准、有意义的预测结果就一定要处理好结晶负样本的选择。Smialowski 在该文献中首先将 PDB 数据库的可溶解蛋白质分为两组，由 X 射线衍射得来的蛋白质作为正样本（XRAY），而由核磁共振技术（NMR）技术得

来的蛋白质作为非正样本；然后将 NMR 对应的数据再细分为两部分，与 XRAY 数据中具有强序列同源性（>=75%）的蛋白质构成一个数据集（XRAY_NMR），基本没有同源性（<=10%）的蛋白质构成另外一个数据集（NMR_ONLY），其中 NMR_ONLY 数据集就作为结晶的负样本。这种划分的假设是仅由 NMR 技术确定的一组结构将富含不可结晶的蛋白质，因为结构基因组学联盟将核磁共振光谱作为一种处理困难结晶或难以结晶目标的方法[22]。不过目前 NMR 技术的一个主要限制是它通常只适用于相对较小的蛋白质，也就是蛋白质序列普遍较短，而 X 射线衍射技术则没有这个限制，为了消除正负样本数据长度分布不同的影响，SECRET 方法按照数据长度再次将数据划分为了 BIG 和 SMALL 两个部分，对于 XRAY 和 NMR_ONLY 数据集作者只划分了 SMALL 数据集。在蛋白质输入特征方面，作者将氨基酸聚类成具有类似物理化学或结构特性的组，原因是考虑到氨基酸代码中存在结构冗余，将 20 个字母的氨基酸表折叠成一个合适的浓缩版本不会导致信息的严重损失。而且在表示氨基酸频率特征时会大大减少维度。然后是特征选择方面，作者使用了 SVM 模型[23]和基于包裹式[24]的特征选择方法。最后是模型方面，作者将具有高斯核的 SVM 作为一级分类器对输入的特征向量进行分类，每一组蛋白质都对应一个 SVM，并输出可结晶和不可结晶两类。为了汇总多组一级 SVM 分类器的信息，作者将各个 SVM 的输出送入到朴素贝叶斯分类器[25]中构成一个二级分类器以获得最终的测试结果。虽然该方法在当时取得了不错的预测效果，但是仍然有很大的缺陷，一方面是该模型只适用于可溶性蛋白，而且由于将 NMR 技术产生的蛋白质结构作为负样本，只能适用于序列长度较小的蛋白质；另一方面是该模型考虑的特征因素较少，限制了模型预测精度的提高。

Chen 和 Kurgan 等人为了克服 SECRET 模型的一些缺陷，提出了 CRYSTALP 模型[26]。该模型仍然使用了和 SECRET 方法相同的数据集，也就是在数据上的缺陷并没有改动。作者提出了一种新的并置氨基酸对频率特征（p-collocated），相比于 SECRET 模型使用的单肽、二肽、三肽频率特征，这里的 p 代表间隔的意思，以前的二元氨基酸频率组成成分用来表示相邻的氨基酸对，现在可以加上间隔 p 可以表示相隔为 p 的二元氨基酸对，反映了更多的局部信息，三元氨基酸对频率同理。对于 $p = 0, 1, \dots, 4$ 的二元并置氨基酸对，一共有 $400 \times 5 = 2000$ 种特征，因为作者没有使用三元并置氨基酸对，所以在特征数量上比 SECRET 还要少，而且在使用 CFSS 方法[27]进行特征选择之后，只保留了 45 个二元并置氨基酸对频率特征，比 SECRET 模型在特征选择后的 103 个特征要少。在模型上作者只使用了朴素贝叶斯分类器，但是最终取得的预测效果要比 SECRET 高 10%，这也就表明了合适的特征对蛋白质结晶倾向性预测是非常重要的。

但是 CRYSTALP 模型仍旧没有摆脱只适用于短蛋白质链的限制，所以 Kurgan 等人基于改模型进行了改进，提出了 CRYSTALP2 模型[28]。该模型有三个改动，一是使用了新的训练数据集 FEAT，这个数据集来自于前面介绍的 ParCrys 模型，其中蛋白质的

长度没有限制；二是使用了更多的蛋白质特征，包括 CRYSTALP 没有使用的三元并置氨基酸对频率、等电离点和平均疏水性等；三是 CRYSTALP2 模型使用了归一化高斯径向基函数（RBF）网络，它是一种基于非线性高斯核函数的具有隐藏层的神经网络[29]。总体而言，在预测性能上比 SECRET、OB-Score、CRYSTALP 和 ParCrys 模型都要高。

随着神经网络模型在图像、语音等多个领域取得突破性的进展，这种学习模型也开始应用到蛋白质结晶倾向性预测领域中。Overton 等人基于人工神经网络提出了 XANNpred 结晶预测模型[30]。在输入的蛋白质特征方面，作者一共使用了 428 个特征，包括单肽频率、二肽频率、等电离点、平均疏水性、链和螺旋残基部分、RONN 无序性部分、序列长度、TMHMM2 跨膜区域以及分子量等。在模型方面采用了 SNNS 包[31]中的人工神经网络模型，包含 428 个输入节点、100 个隐藏层节点和一个输出节点。作者为了识别蛋白质序列中哪些区域最可能影响蛋白质结晶，额外采用了滑动窗口的方法，对于一个蛋白质序列而言，作者按照滑动窗口的方式在该序列上滑动，窗口大小为 61，然后会获得一定数量的窗口序列片段，这些片段会被当作子氨基酸序列并计算出 428 个特征，然后送入到人工神经网络模型中，最终得到每一个窗口片段成功结晶的概率。

2011 年，PPCpred[32]方法被提出，该方法整合了较为完备的蛋白质序列特征，而且还对 TargetDB 数据进行了整理，将蛋白质结晶预测问题划分为了多个阶段的预测问题。首先按照数据库中结晶实验的注释将结晶阶段分为蛋白质材料生产、纯化、晶体产生三个子步骤，只有通过了晶体产生步骤才算结晶成功。再针对这三部分生成三个对应的子数据集，分别预测蛋白质材料生产失败（MF）、纯化失败（PF）以及晶体生产失败（CF）的概率。同时和单阶段的方法相同，将所有可结晶与不可结晶的蛋白质再重新划分为一个只包含两个类别的数据集（CRYS），并基于该数据集预测整个结晶阶段结晶成功的概率。之后的许多模型也都开始将蛋白质结晶倾向性预测看作多阶段预测的问题，这有助于更清晰地了解到蛋白质结晶失败的可能原因。更具体的数据划分方式见本文第 2.2 章节对结晶实验数据的介绍。PPCpred 方法在特征筛选之前一共使用了 828 个特征，主要包括单肽频率、等电离点、AAindex 数据库中的 64 个基于疏水性和能量的特征指数、二级结构、无序性、溶剂可及性等。作者并没有直接使用这些原始的特征，而是对这些特征进行了人工的组合，比如计算 AAindex 在整个氨基酸序列上不同滑动窗口的最小、最大和平均值、二级结构中预测的暴露/掩埋残基数占全部暴露/掩埋残基数的比例等。在特征选择时作者对四个数据集都单独进行选择，因为不同阶段影响成功率的特征也可能不相同。最终，他们分别为 MF、PF、CF、CRYS 四个阶段对应的数据集选择了 86、100、115 和 95 个特征。在模型方面，作者使用了 SVM 分类器，每个阶段单独对应一个预测器，最终输出蛋白质在各个阶段成功或失败的分数。相比于其它的单阶段结晶倾向性预测模型，PPCpred 模型取得了最优越的预测性能。作者还根据模型结果分析了与预测结晶、纯化、材料生产倾向相关的因素，结果表明成功结晶取决于多种因素的组合，

疏水性和 Cys 残基对成功结晶有很重要的影响，而埋藏的 Cys 残基对于纯化步骤则很重要。PPCpred 模型能够取得成功最重要的原因有三个方面：一是使用了更新、更全面的数据集；二是使用了二级结构、溶剂可及性等结构衍生特征，并将序列衍生特征和结构衍生特征组合起来作为 SVM 预测模型的输入；三是从单阶段预测变为多阶段预测，还预测了结晶过程中多个阶段失败的原因。

相比于 PPCpred 模型，Jahandideh 和 Mahdavi 等人提出了 RFCRYS 方法[33]并没有使用结构衍生特征，而是只用序列衍生特征作为模型输入，作者引入了伪氨基酸组成成分特征（PseAAC）[34]来代替氨基酸组成成分特征（AAC），不过在文献的结果比对中并没有单独对该特征的有效性进行说明。在模型方面，作者使用了随机深林模型（RF）[35]预测结晶倾向性，而且他们使用 t 检验来减少输入特征的数量，因为当真正具有信息量的特征百分比很小而总特征的数量很大时，RF 的预测性能往往会急剧下降，原因是 RF 算法对每个节点都以相等的权重随机重采样从而选择特征。最终的预测结果表明 RFCRYS 方法对比于其他现有的预测方法，取得了最好的预测性能。

Charoenkwan 等人提出的 SCMCRYS[36]模型有一个很大的特点，他们的目的不是在追求预测精度的同时增加预测方法的复杂性和特征类型的数量，而是提供一种简单且高度可解释的方法，这种方法从生物学家的角度来看会具有很高的精度。考虑到二肽特征对蛋白质结晶的重要影响，作者只使用了 p-collocated 氨基酸对频率特征，该特征在 CRYSTALP 算法中首次引入并使用。在模型方面使用 SCM 方法[37]和集成方法预测结晶，首先对于每一种 p 的取值，都单独训练一个 SCM 模型，然后再对所有的 SCM 模型进行基于投票的集成，得到最终的预测结果。与其它模型的预测性能相比，要优于大部分模型，并与当时最好的两个模型性能相近，SCMCRYS 准确率为 76.1%，当时最好的两个模型是 PPCpred（准确率 76.8%）和 RFCRYS（80.0%）。SCMCRYS 模型的优势在于只使用了 p-collocated 氨基酸对频率特征也取得了很高的预测性能，侧面证明了二肽组成成分特征对蛋白质结晶的重要性。而且使用 SCM 算法还能分析每一种二肽相对于结晶的倾向性得分，这有助于生物学家设计表面残基的突变以增强蛋白质结晶倾向。

在 PPCpred 模型提出之后，很多研究者也开始倾向于预测结晶过程中多个阶段成功或失败的概率。Wang 等人提出的 Crysalis[38]也是一个多阶段预测模型，相比于 PPCpred 方法中的四种分组（三个结晶中间阶段加上一个预测从头结晶的部分），该文献增加了克隆失败阶段，一共预测五组数据。在特征方面，作者使用了四类特征：氨基酸组成（AAC），氨基酸指数（AAindex），K 间隔氨基酸对（KAAP）和分组的 K 间隔氨基酸对（GKAAP），可以发现使用的特征也都是基于序列衍生的特征，包括了多种形式的氨基酸组成成分。其中 K 间隔氨基酸对等价于 SCMCRYS 模型中使用的 p-collocated 氨基酸对，分组的 K 间隔氨基酸对主要是将 20 种氨基酸按照可及表面积、侧链取向、电荷、氢键供体、疏水性、和范德华潜力等物理化学性质分成三大类氨基酸，再对这三大类氨

基酸计算 K 间隔氨基酸对特征。这么做的原因是因为作者发现大多数氨基酸对几乎没有统计学意义，而且许多氨基酸对表现出相似的物理化学性质，比如氨基酸对 KD 和 RD 在离子对方面具有相似的性质，所以将氨基酸进行分组不仅合并了具有相似特性的对，而且还包含了它们多方面的理化特性。在特征选择方面作者先对 AAindex、KSAAP、GKSAAP 单独过滤，再将过滤后的特征集合在一起进行二级筛选，最终保留了前 100 个特征。然后将该特征向量送入到二级 SVM 模型中进行预测，预测结果与其他模型相比取得了最优的结果。Crysalis 模型的另外一个优势是没有使用结构衍生的特征从而大大提高了计算效率。

Wang 等人介绍了一种新的结晶预测模型 Crysf[39]。之前的所有模型都是使用蛋白质序列衍生的物理化学特征以及结构特征，而 Crysf 独特地使用了蛋白质的功能信息作为特征输入。作者从 UniProt[40]数据库中 Swiss-Prot 和 TrEMBL[41]部分分别提取蛋白质功能，因为 TrEMBL 数据库中的功能数据是未经过审查的，相比于 Swiss-Prot 数据库功能注释质量很差。Crysf 使用 LIBSVM 包中的 SVR 模型[42]来进行结晶预测，单纯从功能方面入手就已经取得了非常好的预测性能，而且运行效率也很高。但是 Crysf 预测模型也有非常大的缺陷，它只能用于 UniProt 中具有功能注释的蛋白质，而且还依赖于高质量的功能注释信息。

虽然基于蛋白质序列推导的结构信息能够帮助进行结晶倾向性预测，但是预测结构特征经常需要使用 PSI-BLAST 程序生成多序列比对[43]，就会导致很长的运行时间。所以 Meng 等人提出了高效的 fDETECT[44]方法来进行结晶倾向性预测，而且与 PPCpred 方法相同，也能全面预测结晶管道的四个步骤。在特征方面，原始特征包括了 1276 个特征：氨基酸组成、按物理化学性质对氨基酸分组、氨基酸物理化学性质、蛋白质物理化学性质、蛋白质链的序列复杂性和内在无序性。相对于 PPCpred 模型而言删除了一些耗时的结构特征，并添加了计算高效的序列衍生特征。在特征筛选时对每个阶段单独筛选，最终只保留了 9、8、4 和 11 个特征分别预测 MF、PF、CF 和 CRYS 步骤。在模型上选用了简单高效的逻辑回归模型，相比较于 SVM 和高斯径向基函数网络等，不仅能够显著降低运行时间，还具有良好的预测性能。从总体的特征输入和模型结构来看，fDETECT 方法尽可能在提高计算效率的同时保持了一定的预测精度。

Zhu 等人为了进一步提高结晶倾向性预测的性能，提出了 DCFCrystal[45]，从两个方面进行了改进。首先，以前的一些模型使用的数据集有一些过时，随着时间的推移，以前错误注释的数据被纠正，大量新的结晶注释数据正在逐渐积累，因此有必要构建一个新的高质量数据集，所以作者构建了两个新的高质量基准数据集 BD_CRYS 和 BD_MCRYS，其中 BD_CRYS 是通用数据集，BD_MCRYS 是由膜蛋白组成的特殊数据集。因为膜蛋白质在各种生物过程中起着至关重要的作用，占人类蛋白质组的四分之一以上[46]，但是由于膜蛋白的特殊属性导致其结晶倾向要比非膜蛋白质要困难的多，因此

作者在 DCFCrystal 基础上衍生出了 MDCFCrystal 模型，专门预测膜蛋白的结晶倾向，不过 MDCFCrystal 和 DCFCrystal 模型在模型结构上是一致的，只是在数据处理上稍有不同，所以这里只介绍 DCFCrystal 模型；最后是提出了一种新的伪预测混合溶剂可及性（PsePHSA）特征并使用深度级联深林（DCF）[47]模型进行结晶预测。深度级联模型属于一种深度学习模型，不同于常见的深度神经网络将单元节点进级联，DCF 是将多个随机深林（RF）和多个完全随机深林（CRTF）[48]组合在一起构成一个网络层，每一级 DCF 网络层接收前一级处理的特征信息，并将其处理结果发送给下一级。在特征方面，溶剂可及性一直被认为是影响结晶的重要因素，作者对预测的溶剂可及性信息重新设计并生成了一种新的特征信息 PsePHSA。最终的实验比对结果表明了 DCFCrystal 方法的优越性，在预测结果上取得了最优的结果，特征比对实验的结果也表明了 PsePHSA 特征对模型性能的贡献占比非常高。

综上所述，这些模型在进行结晶倾向性预测时存在不可或缺的两个部分，基于蛋白质推导出的特征以及机器学习或深度学习模型，特征大致可以分为四类：氨基酸组成成分类型、单个氨基酸或整个蛋白质的物理化学性质、预测的结构特征（二级结构、溶剂可及性、无序性区域等）、序列进化信息等。许多研究都不同程度的表明了这些特征对蛋白质结晶的成功率是有影响的，由于所有这些特征最终构成的特征空间是很大的，如何处理、整合并筛选这些特征成为了每一种方法都不可或缺的步骤。这种基于蛋白质先验特征和复杂网络模型的方法不仅能够取得很高的预测精准度，而且还能挖掘不同蛋白质特征和蛋白质结晶的内在相关性，因此是目前的主流研究方法。但是这种研究方式带来的弊端也是很明显的，一是这些特征都是人工提取并设计的手工特征，比较依赖于先验的结晶生物知识；二是想要处理这些蛋白质特征，一般需要将它们整合成固定长度的特征向量，这样才方便后续的模型预测，但是这也严重限制了特征和模型之间的匹配程度，比如蛋白质的二级结构、氨基酸编码、氨基酸物理化学性质等特征长度都是与蛋白质序列长度相关的，对于不同长度的蛋白质，就需要计算这些特征相对于整个蛋白质的特征信息，比较常用的就是组成成分（频率）、最大值、最小值、平均值等信息，因此也就一定程度限制了这些特征的使用。

1.3.3 基于深度学习的端到端预测方法

随着目前深度学习的发展，尤其是在自然语言处理领域取得的卓越成果，也有一些研究开始只基于蛋白质序列进行端到端的蛋白质结晶倾向性预测。下面就对这些端到端的方法进行简述。

Elbasir 等人[15]最早利用卷积神经网络模型来进行端到端的结晶倾向性预测，为了摆脱手工提取的蛋白质特征携带的一些弊端，他们设计了一种深度学习框架 DeepCrystal，无需从序列中手动设计额外的特征，只依据蛋白质序列本身从头预测蛋白质结晶的倾向

性。CNN 是一种特殊类型的深度神经网络（DNN）[49]，被广泛应用于计算机视觉和自然语言处理任务，并取得了优异的成绩。CNN 之所以能够应用到蛋白质结晶领域作者给出了两个原因，一是深度神经网络已经被应用于蛋白质结构预测[50]和蛋白质功能预测问题[51]等，可以将蛋白质序列与自然语言处理中的句子和文本序列之间进行类比[52]，20 个氨基酸中的每一个是基本的词语，他们一起构成了一个字典；二是根据之前的蛋白质结晶倾向性分析可知蛋白质序列中氨基酸的组成成分特征对结晶的成功率是有很大影响的，而深度卷积模型就可以提取蛋白质序列中局部的氨基酸组成特征。连续的 k 个氨基酸形成的序列一般称为 k-mers，当 $k = 1,2,3$ 时可以分别对应单肽、二肽、三肽特征，所以通过 CNN 模型就可以以 k-mers 的形式捕捉蛋白质序列中的局部特征，比如频繁出现的 k-mers，这种学习到的局部 k-mers 信息有助于预测蛋白质结晶倾向性。在数据输入方面，DeepCrystal 将蛋白质序列编码为 One-hot 形式的向量，一共 21 种类别，包括 20 种标准的氨基酸类别和一种非标准的氨基酸类别。在具体的模型架构方面，DeepCrystal 包含三个卷积模块，第一层卷积通过不同滤波器大小获取本地上下文信息；第二层识别有效的 k-mers 集，比如频繁出现的 k-mers；第三层卷积捕获这些有效的 k-mers 集合之间的交互。最终将 CNN 学习到的特征输入到全连接层和 sigmoid[53]输出层，预测蛋白质序列是否可以产生衍射质量的晶体。模型对比的结果显示 DeepCrystal 获得了当时最优的预测性能。在作者看来，传统方法在计算有关 k-mers 的特征时受限于 k 的大小，而 CNN 可以简单的通过设置卷积核的大小来自由地计算这种局部特征。而且深度学习模型相比较于传统的机器学习分类器有更强的拟合能力，虽然会出现过拟合，但是可以使用一些常见的调试手段比如 dropout[54]来取得比较好的泛化性能。

DeepCrystal 模型的产生促进了很多人开始研究如何更好地利用深度学习技术从头预测蛋白质结晶倾向性。虽然 DeepCrystal 可以使用 CNN 模型有效的提取 k-mers 氨基酸局部信息，但是难以学习蛋白质序列中远距离氨基酸的相互作用信息。所以 Xuan 等人[14]提出了利用 LSTM 网络[55]来提取这种远距离作用信息。首先是 k-mers 远距离的相互作用对于蛋白质形成稳定的空间结构是非常重要的，也被认为对蛋白质结晶有很大的影响，而 LSTM 模型能够学习数百个时间步长，也就能捕捉到这种全局的交互信息。Xuan 等人提出的模型架构为 CLPred，先利用 CNN 层提取不同长度的 k-mers 氨基酸片段信息，再将提取的特征送入到多层双向 LSTM（BLSTM）[56]网络模块中。在模型对比方面，取得了比 DeepCrystal 更好的测试结果。

Jin 等人[13]提出的 ATTCry 模型与 CLPred 模型的目的相同，首先使用多尺度卷积神经网络提取蛋白质序列的局部 k-mers 特征，然后再使用多头自注意力机制[57]联合卷积层提取得到的空间局部信息，以此来获取更复杂的长距离空间依赖信息，这两种模型结合在一起可以更有效的捕获蛋白质序列的局部特征和全局特征，从而增强蛋白质结晶倾向性的预测能力。而且该模型也是端到端的预测模型，相比于基于手工蛋白质序列特

征的模型，该模型只需要依赖蛋白质序列本身，在保持较高模型精度的同时也拥有非常快的预测速度。

上述这些模型都可以称为端到端的结晶倾向性预测模型，它们旨在不借助任何先验蛋白质特征，而仅通过蛋白质一级序列来预测蛋白质结晶倾向性。这样做的假设是蛋白质序列能够决定蛋白质的结构和功能，同时也能够决定一个蛋白质是否能够结晶或者结晶的难度。这些端到端的结晶倾向性预测模型利用深度学习技术来从蛋白质一级序列中提取与结晶相关的氨基酸相互作用信息，优点是拥有非常快的预测速度，但是由于缺乏了与结晶紧密相关的蛋白质特征对它的先验指导，在预测精准度上较差。

1.4 本文的工作

虽然已经有许多研究工作在预测蛋白质结晶倾向性方面做出了卓越的贡献，但是仍然有很大的提升空间。首先是在输入特征方面，大部分的结晶倾向性预测模型仍然只采用传统手工特征的方式，比如 TargetCrys[58]模型将氨基酸组成成分、位置特异性矩阵特征直接送入到双层 SVM 模型中进行结晶倾向性预测，BCrystal[59]模型采用多种基于序列的特征并使用 XGboost[60]模型进行预测。虽然有一些基于深度学习的模型在一定程度上克服了这个障碍，比如 DeepCrystal[15]使用卷积神经网络自动地从蛋白质序列中提取更有效的特征来预测结晶，DCFCrystal[45]模型在多种蛋白质序列特征辅助下使用深度级联深林来提取与结晶倾向性相关的特征，CLPred[14]模型和 ATTCry[13]模型分别使用双向 LSTM 和自注意力机制自动提取与结晶相关的远距离氨基酸相互作用信息。但是他们都只能从蛋白质一级序列的层面上来提取与结晶相关的氨基酸相互作用信息，而且在预测性能上仍有很大的不足。

为了有效的解决这些问题，我们引入蛋白质接触图来辅助进行结晶倾向性预测。蛋白质接触图描述了序列中氨基酸之间的空间接触关系，可以看作蛋白质三维空间结构的一种简化，相比于蛋白质一级序列和基于蛋白质序列的特征，蛋白质接触图的优势在于提供了先验的空间结构信息，使得我们能够在这个空间结构中挖掘出与结晶相关的更有效的特征。我们先后使用了三种方法来更加充分地利用蛋白质接触图来预测蛋白质结晶倾向性，下面我们对这三种方法进行简要概述。

（1）为了验证蛋白质接触图中是否包含有效的空间结构信息，我们设计了一种 CCmap-KAAP 手工特征，该特征是从接触图中提取出来的 K 间隔氨基酸对频率特征，相比于基于序列提取的 K 间隔氨基酸对频率特征（KAAP），它体现了氨基酸在空间中的近距离相互作用信息，还包含了一部分在序列上远距离接触的氨基酸相互作用信息。通过使用 XGboost 模型对这两种特征在结晶倾向性上的预测结果进行对比分析，我们发现 CCmap-KAAP 特征的确包含着有效的空间结构信息，而且与 KAAP 特征是互补的，将基于序列的特征和基于空间结构的特征融合在一起能够大大提高结晶倾向性的预测

能力，这也验证了蛋白质接触图在结晶倾向性预测方面的有效性。我们还提出了一个基于机器学习的新模型 CCmapCrys，它将 CCmap-KAAP 特征和更多的基于蛋白质序列推导的特征融合在一起，并使用 XGboost 模型进行多阶段的结晶倾向性预测。通过与其他基于机器学习的方法进行对比，我们发现在 CCmap-KAAP 特征的帮助下，我们的模型 CCmapCrys 取得了最好的预测性能。

（2）CCmap-KAAP 特征虽然有效，但是它仍属于一种手工特征，并没有完全解决我们之前所发现的那些问题，所以我们在第二种方法中使用了基于深度学习的图注意力网络模型（GAT）[61]来自动提取接触图中的空间结构信息。将蛋白质中的氨基酸看作节点，接触关系看作边，那么接触图就等价于接触网络，对于这种非结构化的数据，有很多研究表明使用图卷积类别的模型可以对其进行更好的表征[62-64]，所以我们基于一种特殊的图卷积模型 GAT 设计了一个新模型 GCmapCrys 来进行多阶段的结晶倾向性预测。GCmapCrys 模型利用图注意力机制迭代地更新氨基酸节点特征和对应的接触边特征，从而更有效地提取接触图局部结构信息。为了提高预测的精度，我们仍然融合了与第一种方法相同的多种互补的蛋白质特征来辅助结晶倾向性预测。通过与其他最先进的结晶倾向性预测模型进行对比，我们的 GCmapCrys 模型在多种指标上都获得了最高的预测结果。

（3）在第三种方法中我们使用了图 Transformer（Graph Transformer）[65]方法来提取蛋白质接触图中的信息。因为图卷积模型难以堆叠过深的卷积层来提取更高层次的特征，所以我们使用 Transformer[57]架构来进一步弥补这个缺陷。Transformer 模型的优势在于能够利用自注意力机制学习上下文中所有氨基酸的相互作用关系，能够进一步加强对长距离和全局氨基酸相互作用的学习。但是由于 Transformer 只能处理序列信息，无法直接应用到图数据结构上，所以我们先利用图注意力模型对接触图进行更新，使得每个节点都包含一定范围的局部结构信息，然后再将所有更新后的节点铺平为一个序列作为 Transformer 的输入，这样就使得氨基酸节点序列保留了部分接触图的原始拓扑结构信息。我们还在图注意力网络的基础上添加了残差连接层[66]来尽可能保留足够丰富的结构信息。因为蛋白质序列信息和结构信息是互补的，所以我们将蛋白质序列的位置信息进行编码并输入到 Transformer 中，以增强它对蛋白质序列位置信息的敏感度。最后我们将这个多阶段的结晶倾向性预测模型命名为 GCmapTCrys，通过与 GCmapCrys 和现阶段的其他先进的预测模型进行对比，我们发现 GCmapTCrys 拥有最好的预测性能。

1.5 本文结构

本文一共 6 章，章节安排如下：

第一章：绪论。我们首先介绍了结晶倾向性预测的概念、研究背景和意义；然后分析了影响蛋白质结晶的内在因素和外在因素有哪些，还介绍了关于本课题的国内外研究

现状；最后对本文的工作与结构进行简要概述。

第二章：蛋白质结晶预测相关基础知识。我们介绍了本文中涉及到的一些生物实验数据和相关基础概念，包括蛋白质接触图的基本概念、生成过程以及结晶数据集的构造；最后介绍了本文中使用的模型评估方法。

第三章：基于 CCmap-KAAP 特征的蛋白质结晶倾向性预测。我们首先介绍了将蛋白质接触矩阵转化为图结构数据的方法；然后设计了一种新的氨基酸组成成分特征——CCmap-KAAP，并通过实验验证了该特征的有效性；最后我们基于 CCmap-KAAP 特征提出了一个新的多阶段预测模型 CCmapCrys 并介绍了构造模型的具体过程；最后在实验部分与多种基于机器学习的预测模型进行了对比分析。

第四章：基于图注意力网络的蛋白质结晶倾向性预测。我们在本章节设计了一种新的多阶段结晶倾向性预测模型 GCmapCrys，首先介绍了蛋白质的图结构表征，详细讲述了如何将多种序列特征与接触图融合在一起构成模型的输入；然后分析了图卷积模型的概念以及 GCmapCrys 模型的整体流程；最后在实验部分介绍了本模型在不同评价指标下与其他最先进的模型的对比结果，还分析了不同蛋白质先验特征对于 GCmapCrys 模型的贡献。

第五章：基于 Graph Transformer 的蛋白质结晶倾向性预测。设计了一个新的多阶段结晶倾向性预测模型 GCmapTCrys，首先介绍了 Transformer 模型的基本概念以及将图与 Transformer 进行结合的三类方法；然后描述了 GCmapTCrys 模型的整体流程，包括图注意力网络层、残差连接、序列位置信息编码以及 Transformer 编码层等；最后在实验部分将我们的模型与其他预测模型进行对比分析，还对模型的不同部分进行了消融实验，验证了我们模型的有效性。

第六章：总结与展望。对本文提出的基于蛋白质接触图的结晶倾向性预测方法进行了总结，并对今后的工作进行了展望。

2 蛋白质结晶预测相关基础知识

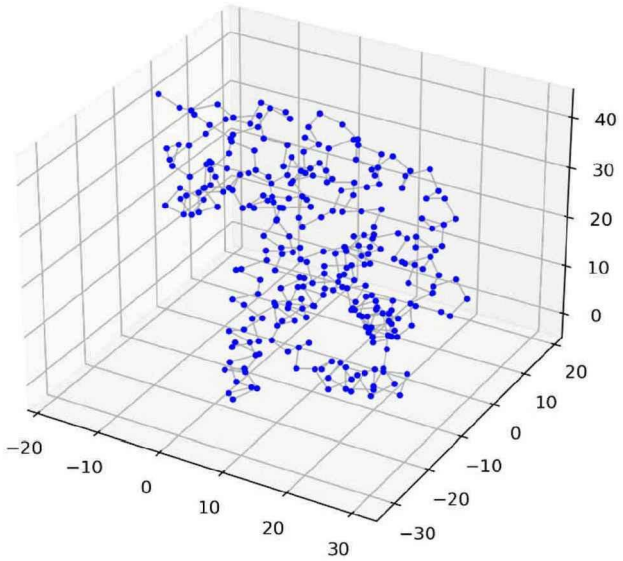
2.1 蛋白质接触图

蛋白质的形状通常可以分为四个层次：一级、二级、三级和四级。从蛋白质三级结构的角度来看，几百甚至几千个残基之间相互作用形成一个稳定的三维结构从而实现特定的功能，如果将残基看作一个顶点，残基节点之间的相互作用看作连接的边，那么网络建模的方法是非常适合来表征和分析蛋白质的结构与功能的。蛋白质三维结构包含很多复杂的信息，为了简化和模拟蛋白质结构，一种常用的方法就是将残基中原子之间的接触视为相互作用网络，忽略对应的二级结构和折叠类型，而蛋白质接触图就是由蛋白质三维结构简化而来的一种相互作用网络。

当给定一个蛋白质三维结构的 3D 坐标，我们可以根据这个坐标来构建蛋白质接触图。首先将蛋白质残基中的 C_β 原子（如果是 Glycine 残基则是 C_α 原子）看作网络结构中的残基顶点，忽略残基中其它原子的影响，可以形成如图 2.1（a）所示的简化版三维结构。然后再计算每对残基顶点之间欧氏距离，从而构建出一个距离矩阵，其中行和列表示对应的残基顶点，如图 2.1（b）所示。距离矩阵中的对角线元素始终为 0，因为相同残基自身到自身的距离为 0。 如果要根据这个距离矩阵判断任意两个残基之间是否相连，我们需要设置一个距离阈值，这个距离阈值一般取决于残基之间的非共价键作用范围，在不同的研究中有 5Å[67]、7Å[68]和 8.5Å[69]等多种截断值。确定了距离阈值之后，我们可以将距离矩阵中大于该阈值的元素设置为 1，表示接触，相反则设置为 0，表示不接触，从而形成相应的接触矩阵。我们进一步将接触部分标识为黑色，非接触部分标识为白色，可以形成如图 2.1（c）所示的蛋白质接触图。

虽然蛋白质接触图可以从真实的三维结构中推导出来，但是它更多的是通过预测得到并被用来辅助研究蛋白质折叠问题和蛋白质结构测定问题[70, 71]，在本文中我们也将使用预测的蛋白质接触图来辅助预测蛋白质结晶倾向性问题。目前比较常用的蛋白质接触图预测算法是基于序列进行从头预测，该方法试图通过分析多序列比对（MSAs）中目标残基对的进化相关性来预测残基之间的接触，如果一条序列中两个位置上的氨基酸存在某种协同进化相关性，那么这两个点很有可能就是相互接触的[72]。在本文中，为了获取更高的结晶倾向性预测能力，我们尽可能的从现有的接触图预测模型中选取预测精度较高的工具从而获取比较准确的蛋白质接触图，下面我们具体介绍生成蛋白质接触图的过程。

目前精度较高的蛋白质接触图预测工具都需要多序列比对作为直接或间接输入，所以在预测接触图之前我们需要先生成蛋白质样本对应的多序列比对文件。多序列比对的主要过程就是通过 BLAST[73]、HHblits[74]等比对工具在公共生物数据库中寻找待查询序



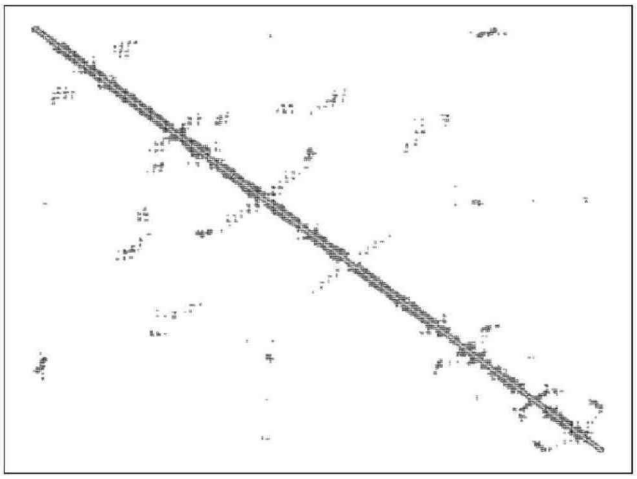
（a）基于 C_β 原子构建的蛋白质三维结构图

	M	I	S	N	S	S	A
M	0	3.8	6.53	9.36	12.82	16.3	18.12
I	3.8	0	3.82	6.53	9.74	13.28	14.75
S	6.53	3.82	0	3.79	6.5	10.2	11.74
N	9.36	6.53	3.79	0	3.79	6.96	9.23
S	12.82	9.74	6.5	3.79	0	3.8	5.57
S	16.3	13.28	10.2	6.96	3.8	0	3.81
A	18.12	14.75	11.74	9.23	5.57	3.81	0

大于8.5Å赋值为0
小于8.5Å赋值为1

	M	I	S	N	S	S	A
M	0	1	1	0	0	0	0
I	1	0	1	1	0	0	0
S	1	1	0	1	1	0	0
N	0	1	1	0	1	1	0
S	0	0	1	1	0	1	1
S	0	0	0	1	1	0	1
A	0	0	0	0	1	1	0

（b）蛋白质残基距离矩阵和接触矩阵局部示意图



（c）蛋白质接触图可视化

图 2.1 蛋白质接触示意图

列的多个同源序列。在本文中，我们使用了 HHblits 工具在 UniClust30[75]数据库中搜索待查询序列的多个同源序列。这些生成的多序列比对文件会被送入到接触图预测工具中进行预测。在预测工具方面，我们有四种待选择的方法，分别是 DeepCov[76]，PconsC4[77]，TripletRes[78]和 trRosetta[79]，下面我们分别从接触图的预测精度、运行时间和对硬件的依

赖三个角度来选择最合适的接触图预测工具。

表 2.1 四种方法在不同接触距离下的 top-L 预测精度

方法	平均距离	短距离	中距离	长距离
DeepCov	0.5686	0.2468	0.2667	0.4076
PconsC4	0.6016	0.2468	0.2787	0.4526
TripletRes	0.7266	0.2848	0.3357	0.5796
trRosetta	0.7516	0.2908	0.3397	0.6036

首先在预测精度方面，我们从 Zhang 等人[80]的实验中摘录出对这些方法的预测精度对比，如表 2.1 所示，它显示了这四种模型在 610 个非冗余蛋白质测试集上的对比结果，从中我们可以发现 PconsC4 方法排名第三。虽然 TripletRes 和 trRosetta 都具有很高的预测精度，但是这两种方法的可用性较差，分别受到运行时间和硬件设备的限制。以 018304_NYSGRC 蛋白质为例，PconsC4、TripletRes 和 trRosetta 三种方法在我们设备上的平均运行时间分别为 46.2、5503.3 和 3.4 秒，我们的硬件设备平台为 Intel(R) Xeon(R) E5-1620 CPU (256 G) 和 NVIDIA TITAN X (Pascal) GPU (12G)。其中 PconsC4 方法只依赖 CPU，而 TripletRes 和 trRosetta 同时依赖 CPU 和 GPU。从运行时间上我们可以发现 TripletRes 模型的运行时间要远远超过剩余两个模型，因为 018304_NYSGRC 蛋白质的长度接近我们数据集的平均长度，所以使用 TripletRes 模型预测我们数据集中超过 17000 个样本的全部时间近似为 1083（$5503.3 \times 17000 \div 3600 \div 24 \approx 1083$）天，因为运行时间太长，所以我们并不使用该工具进行预测。而对于 trRosetta 模型而言，虽然其运行速度很快，但是它会消耗大量的 GPU 资源，经过我们的验证，如果一个蛋白质序列的长度超过 400，同时其对应的同源序列数量超过 35000 时，就会发生显存超出限制的问题。如果用该工具预测全部的数据集样本，大约会有 16.5%的蛋白质序列无法预测成功，因此综合考虑之下，我们使用 PconsC4 工具来对我们的数据集预测蛋白质接触图。

值得注意的是，如果在测试过程中一条蛋白质序列在多序列比对时没有找到任何同源序列，此时 PconsC4 会预测失败，导致整个结晶倾向性预测程序无法运行，所以为了保证我们程序的鲁棒性，在测试阶段当蛋白质序列没有匹配到任何同源序列时，我们会为其随机生成一条序列一致性程度为 99%的同源序列，从而确保 PconsC4 工具能够正常预测出对应的蛋白质接触图。不过由于是随机生成的同源序列，所以蛋白质接触图的预测精度以及整个模型的结晶倾向性预测精度都会比较差。PconsC4 工具的最终输出的蛋白质接触图是一个 $L \times L$ 的概率矩阵，L 代表对应蛋白质序列的长度。其中矩阵中的每一个元素代表对应的两个残基之间接触概率的大小。

2.2 蛋白质结晶数据库

随着蛋白质结晶实验数据的缓慢积累，一些数据库开始逐步建立起来，2001 年，TargetDB[16]（http://targetdb.pdb.org/）数据库开始运行，其包括了目标选择、蛋白质序列、克隆、表达、纯化和结构测定阶段的实验数据，基本能够包含一个目标在结构测定实验流程中的各种实验数据。PepcDB[81]（蛋白质表达纯化和结晶数据库）于 2004 年左右建立，作为 TargetDB 的扩展，用于收集蛋白质结构生产管道中每个步骤的更详细的状态信息和实验细节。这些数据库中都包含了较为完整的蛋白质结晶实验数据。

在 PPCpred[32]方法出现之前，基于序列的蛋白质结晶倾向性预测器将标签分为可结晶与不可结晶两类，或者类似于 XtalPred[20]将“易于结晶”和“难以结晶”之间划分多个等级类别，本质等同于预测蛋白质可结晶的概率。但是 TargetDB 数据库并没有直接标明哪些目标是可结晶或者不可结晶的，需要根据其实验状态信息确定其标签。如果只划分可结晶与不可结晶两类，那么对于可结晶的目标，其 TargetDB 目标实验状态一般为 Crystal structure、in PDB，不可结晶的实验状态为已停止状态，包括测序失败（Sequencing failed）、克隆失败（Cloning failed）、表达失败（Expression failed）、纯化失败（Purification failed）、结晶失败（Crystallization failed）、衍射差（Poor diffraction）。对于大多数实验而言，其停止状态为空，这使得无法确定实验的最终结果以及其失败的原因（可能是能够结晶却由于其它原因导致实验停止，因此不能计入负样本中）。

表 2.2 多阶段数据划分标准

蛋白质类别	注释状态
蛋白质材料生产失败 (MF)	Selected
	Cloned
	Expressed
纯化失败(PF)	Soluble
	Purified
晶体生产失败 (CF)	Crystallized
	Diffraction
可结晶 (CRYS)	Crystal structure
	In PDB

对于 PPCpred 而言，它为了提供更精准和更全面的结晶倾向注释，将所有蛋白质分为四类，蛋白质材料生产失败（MF）、纯化失败（PF）、晶体生产失败（CF）与可结晶（CRYS），表 2.2 显示了详细的划分标准。然后作者将这四类蛋白质构成四种阶段的数据集，（1）MF_DS：MF 作为负样本，PF、CF、CRYS 合并作为正样本，代表蛋白质材料生产的失败和成功；（2）PF_DS：PF 作为负样本，CF、CRYS 合并作为正样本，

代表蛋白质纯化的失败和成功；（3）CF_DS：CF 作为负样本，CRYS 作为正样本，代表蛋白质结晶的失败和成功；（4）CRYS_DS：MF、PF、CF 作为负样本，CRYS 作为正样本，代表蛋白质是否可结晶。CF_DS 与 CRYS_DS 的区别在于不可结晶的样本是否通过了生产和纯化步骤。根据数据集的划分，我们将能够预测蛋白质材料生产失败、纯化失败、晶体生产失败、可结晶的模型称为多阶段预测模型，而只能预测蛋白质是否可结晶的模型称为单阶段预测模型，单阶段的预测模型只会用到 CRYS_DS 数据集。

在本文中，我们使用了 Zhu 等人[45]构建的多阶段数据集 BD_CRYS，它包含四个子数据集 MF_DS、PF_DS、CF_DS 和 CRYS_DS，这四种数据集的划分标准与 PPCpred 相同。我们在这四种数据集上分别预测蛋白质材料生产失败、纯化失败、晶体生产失败和可结晶的概率。每一种子数据集我们都使用 CD-HIT[82]工具进行过滤，使数据集中的序列保持 40%以下的序列一致性。

因为蛋白质接触图的预测性能依赖于多序列比对 MSA 的质量[83]，为了消除低质量的 MSA 对蛋白质接触图的预测性能造成影响，我们对原始的数据集进行了过滤。我们使用 Nf 指标来评估蛋白质 MSA 的质量：

$$Nf = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} w_n \tag{2.1}$$

其中 L 是蛋白质序列的长度，N 是 MSA 中蛋白质的同源序列数量，w_n 代表第 n 个同源序列的权重，该权重反映了这个同源序列包含进化信息的比重：

$$w_n = \frac{1}{1 + \sum_{m=1, m \neq n}^{N} I\left[S_{m,n} \geq 0.8\right]} \tag{2.2}$$

其中 $S_{m,n}$ 是同源序列 m 和 n 之间的序列一致性，$I[S_{m,n} \geq 0.8]$ 代表当 $S_{m,n} \geq 0.8$ 时取值为 1，否则为 0。当同源序列 n 与剩余所有的序列一致性都小于 0.8 时，此时 $w_n = 1$,代表当前同源序列能够包含较多的进化信息。最终我们根据 Zhang 等人[83]的理论设置 Nf 的阈值为 128，如果蛋白质 MSA 的 Nf 指标小于 128，我们将其从我们的数据集中剔除。同时为了防止生成的 MSA 包含的同源序列数量过多，我们去除高度相似且冗余的同源序列，从而尽可能使 MSA 中包含的同源序列数量小于 50000。最终我们四个子数据集 MF_DS、PF_DS、CF_DS 和 CRYS_DS 的样本总数量分别为 15476, 6389, 1994 和 15476。对于每个子数据集，我们随机筛选 90%的样本进行训练和验证，10%的样本进行测试，表 2.3 展示了数据集的具体数量信息。因为在模型对比实验时一些模型限制蛋白质序列的长度不超过 800，所以我们还从这四个数据集中过滤掉了长度超过 800 的蛋白质样本，从而又形成了 MF_DS800、PF_DS800、CF_DS800 和 CRYS_DS800 四种子数据集。

表 2.3 MF_DS、PF_DS、CF_DS 和 CRYS_DS 数据集的样本数量

	训练集		测试集	
	NP[a]	NG[b]	NP[a]	NG[b]
CRYS_DS	998	12930	111	1437
MF_DS	4366	9561	486	1063
PF_DS	1483	4266	165	475
CF_DS	1301	493	145	55

[a] NP 代表正样本的数量

[b] NG 代表负样本的数量

2.3 评估指标

结晶倾向性预测是一个二分类问题，根据以前常用的结晶倾向性预测评价指标，我们使用 Matthew's 相关系数（Matthew's correlation coefficient，MCC）、灵敏度（sensitivity，Sen）、特异性（specificity，Spe）、准确率（accuracy，Acc）和 ROC 曲线下面积（Area Under Curve，AUC）来对结晶倾向性预测模型进行评估，前四种指标的计算公式如下：

$$Sen = TP/(TP + FN) \tag{2.3}$$

$$Spe = TN/(TN + FP) \tag{2.4}$$

$$Acc = (TP + TN)/(TP + FP + TN + FN) \tag{2.5}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{2.6}$$

其中 TP、FP、TN、FN 代表的分别是被模型预测为正类的正样本、被模型预测为正类的负样本、被模型预测为负类的负样本和被模型预测为负类的正样本。因为对于大部分模型而言其输出是一个概率值，当概率大于 T 时表示预测为正类，小于 T 表示预测为负类，所以 TP、FP、TN 和 FN 都依赖于概率阈值 T 的选择。灵敏度 Sen 反应了我们的模型对正类样本的预测准确程度，特异性 Spe 反应了我们的模型对负样本的预测准确程度，这两个指标因为是针对于单一类别而言的，对整体样本是否均衡并不敏感，所以我们继续使用 Matthew's 相关系数来平衡表示整体的预测质量。准确率 Acc 反应了预测正确的结果占总样本数量的百分比。由于概率阈值的选取是不确定的，所以我们通过在训练集上最大化 MCC 来获取最优的阈值 T，并基于这个阈值计算测试集上的 Sen、Spe、Acc 和 MCC 指标。除了这四种指标之外，我们还使用了 ROC 曲线下的面积（AUC）作为评估指标，ROC 曲线是通过不同概率阈值下的真阳性率（True Positive Rate，TPR）和假阳性率（False Positive Rate，FPR）绘制得到的，不受阈值选取和类别不平衡的影响，TPR 和 FPR 两坐标轴围成的曲线下面积 AUC 能够很好的反应模型的总体性能，AUC

的值越接近于 1 表明模型的预测性能越高。

2.4 本章小结

因为蛋白质接触图作为本文研究的核心，所以在本章我们先详细介绍了蛋白质接触图的基本概念以及生成过程，为下文的三个方向进行了一些准备工作。我们最终从准确度、运行时间、硬件成本三个条件下选择了 PconsC4 工具进行蛋白质接触图预测。我们还介绍了蛋白质结晶相关实验数据信息，包括结晶实验数据的不同阶段以及多种标注状态。最后我们介绍了本文中 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四个数据集的详细构建过程。

3 基于 CCmap-KAAP 特征的蛋白质结晶倾向性预测

不论是蛋白质结晶倾向性预测领域还是其他生物信息学领域，氨基酸组成成分都是最常用的特征之一。该特征是根据蛋白质一维序列计算出的单肽或多肽频率，比如基本的氨基酸频率、二肽（Dipeptide）频率、三肽（Tripeptide）频率等组成成分特征。这些特征可以看作一个大小分别为 1、2、3 的窗口在连续的一维蛋白质序列上滑动得到的，属于一种局部性特征，能够表现出一些重要的氨基酸或多肽在蛋白质中出现的频率。随着后续结晶倾向性预测领域的发展，我们发现越来越多的模型开始探索远距离作用的氨基酸相互作用信息，比如 CLPred[14]方法利用 LSTM[55]网络来学习蛋白质序列中远距离氨基酸作用的信息，ATTCry[13]方法利用多头注意力机制[57]来学习这种远距离作用信息。从这个角度来看，传统的单肽、二肽、三肽组成成分特征就属于近距离氨基酸作用信息。我们在本文中引入的蛋白质接触图虽然只是三维空间构象的一种简化，但是相比于蛋白质一级序列，它仍然包含着丰富的结构信息，由于序列上距离很远的两个氨基酸在空间上可能相互接触，所以这种结构信息就可以看作一种远距离氨基酸作用特征。为了利用蛋白质接触图所带来的结构信息，在本章节中我们有两个重要的贡献，首先是基于蛋白质接触图设计出了一种新的氨基酸组成成分特征 CCmap-KAAP，并在 CRYS_DS 数据集上使用 XGboost[60]机器学习模型验证了该特征的有效性；其次我们将 CCmap-KAAP 特征和多种其他蛋白质序列特征整合在了一起，并使用 XGboost 模型来进行多阶段的蛋白质结晶倾向性预测，新的预测模型被命名为 CCmapCrys。通过与其他基于机器学习的方法进行对比，我们发现 CCmapCrys 在这些方法中取得了最优的预测结果。

3.1 CCmap-KAAP特征

3.1.1 氨基酸组成成分特征

氨基酸组成成分（Amino acid composition，AAC）特征是生物信息学中广泛使用的特征之一。构成蛋白质的标准天然氨基酸类别有 20 种，一般常用的 AAC 特征就是指这 20 种氨基酸在蛋白质序列中出现的频率。对于一个蛋白质序列 $P = A_1A_2A_3\ldots A_L$，AAC 的计算公式如下：

$$AAC_j = \frac{1}{L}\sum_{i=1}^{L} I(A_i = C_j), 1 \le j \le 20 \tag{3.1}$$

其中 A_i 是蛋白质序列 P 中第 i 位置的氨基酸，L 代表序列长度，一共 20 种标准氨基酸类别，用 C_j 来表示第 j 种氨基酸类别，函数 I 是指示函数，当 A_i 与 C_j 相等时输出为 1，否则输出 0。得到的 AAC 特征是一个长度为 20 的特征向量，AAC_j 是计算得到的第 j 类

氨基酸在蛋白质序列种出现的频率。除了可以计算单个氨基酸出现的频率，我们也可以计算多肽在蛋白质序列中出现的频率。相比与单个氨基酸，多肽能够反应局部氨基酸之间的相互作用信息，所以在分析蛋白质的性质时，也会分析对应蛋白质序列中二肽和三肽的出现频率，这两种特征一般简写为 Dip-AAC 和 Tri-AAC，都可以归类为氨基酸组成成分特征。按照 20 种氨基酸类别计算，一共有 $20 \times 20 = 400$ 种二肽类别，$20 \times 20 \times 20 = 8000$ 种三肽类别，我们用 Dip 表示所有二肽类别构成的集合，Tri 表示所有三肽类别构成的集合，则 Dip-AAC 和 Tri-AAC 的计算公式如下：

$$Dip\text{-}AAC_m = \frac{1}{L-1}\sum_{i=1}^{L-1} I(A_iA_{i+1} = Dip_m), 1 \le m \le 400 \tag{3.2}$$

$$Tri\text{-}AAC_n = \frac{1}{L-2}\sum_{i=1}^{L-2} I(A_iA_{i+1}A_{i+2} = Tri_n), 1 \le n \le 8000 \tag{3.3}$$

其中 Dip_m 代表第 m 种二肽类别，Tri_n 代表第 n 种三肽类别。一般需要计算的肽链长度不会超过 3，因为多肽特征的维度会随着肽链长度的增大而呈指数形式增长，当肽链长度等于 4 时，对应的特征维度已经达到了 $20^4 = 160000$，这种指数增长的趋势使得研究人员只能计算短肽链对应的频率特征。

为了获取更长范围的氨基酸组成信息，Chen 等人在 CRYSTALP[26]模型中引入了 K 间隔氨基酸对（KAAP）特征，通过在连续的氨基酸序列之间增加间隔的方式扩大长度且不会引起特征长度的增长。A_i-A_j，A_i--A_j，A_i---A_j，A_i----A_j分别是当间隔 $K = 1, 2, 3, 4$ 时的有效氨基酸对，“-”代表间隔位置，因此连续的二肽 A_iA_j 可以看作 $K = 0$ 时的特殊间隔氨基酸对。KAAP 的优势在于随着 K 的增大能够捕捉到更远距离的氨基酸相互作用信息，因此在结晶倾向性预测方面，CRYSTALP、CRYSTALP2[28]、SCMCRYS[36]、Crysalis[38]模型都使用了 KAAP 特征来提高结晶预测的准确率。但是这些特征都是基于蛋白质序列提取得到的，缺少了一定的结构信息，如果能够从蛋白质的三维结构上获取对应的氨基酸组成成分特征，就能够捕捉到更真实的氨基酸作用情况。比如在空间结构上相邻的氨基酸对，也许在蛋白质一级序列上相隔的距离很远，传统的多肽频率特征和 KAAP 特征都难以捕捉这种信息。蛋白质接触图作为三维空间构象的一种简化，仍然携带有丰富的结构信息，因此我们尝试从预测的蛋白质接触图中提取氨基酸组成成分特征来预测蛋白质结晶倾向性。

3.1.2 基于蛋白质接触图提取 CCmap-KAAP 特征

从 PcosnsC4 工具预测出来的蛋白质接触图是一个 $L \times L$ 的二维矩阵 P，L 代表蛋白质序列的长度，行和列分别对应蛋白质序列相应位置的氨基酸，如果将蛋白质接触图看作一个图结构的数据，这个二维矩阵就等价于图的邻接矩阵，矩阵中的元素代表残基 i 与残基 j 之间的接触概率 $P(i, j)$，也代表两个节点之间的边权重。对于一般的邻接矩

阵而言，$P(i,j)=0$ 代表 i 和 j 两个节点之间没有边，$P(i,j)>0$ 才代表存在边。但是预测的蛋白质接触图矩阵基本不存在为 0 的接触概率，如果两个节点之间在实际情况下不接触，在预测的接触图矩阵中对应的接触概率 $P(i,j)$ 一般比较小，但是不会为 0。所以不能使用 $P(i,j)>0$ 的标准来判断是否接触，需要设置一个阈值 d，只有当两个残基之间的接触概率大于 d 的时候才能定义两个残基之间存在边。d 属于一个超参数，我们在第 3.1.4 章节中分别对 $d=0.3,0.4,0.5,0.6,0.7$ 进行了参数搜索，最终选取了$d=0.3$。我们将蛋白质接触图用 G 来表示，图 3.1 对蛋白质接触图的概率矩阵形式和图形式进行了可视化。构建完蛋白质接触图 G 之后，为了更好的利用它携带的结构信息，我们从接触图 G 中提取了 KAAP 特征，简称为 CCmap-KAAP 特征。

0.000	1.000	1.000	0.466	0.484	0.261	0.150	0.037	0.003	0.001
1.000	0.000	1.000	0.536	0.500	0.445	0.160	0.007	0.000	0.000
1.000	1.000	0.000	1.000	0.511	0.499	0.248	0.007	0.002	0.000
0.466	0.536	1.000	0.000	1.000	0.500	0.450	0.080	0.007	0.000
0.484	0.500	0.511	1.000	0.000	1.000	0.500	0.328	0.035	0.001
0.261	0.445	0.499	0.500	1.000	0.000	1.000	0.972	0.214	0.008
0.150	0.160	0.248	0.450	0.500	1.000	0.000	0.999	0.798	0.066
0.037	0.007	0.007	0.080	0.328	0.972	0.999	0.000	0.982	0.505
0.003	0.000	0.002	0.007	0.035	0.214	0.798	0.982	0.000	0.933
0.001	0.000	0.000	0.000	0.001	0.008	0.066	0.505	0.933	0.000

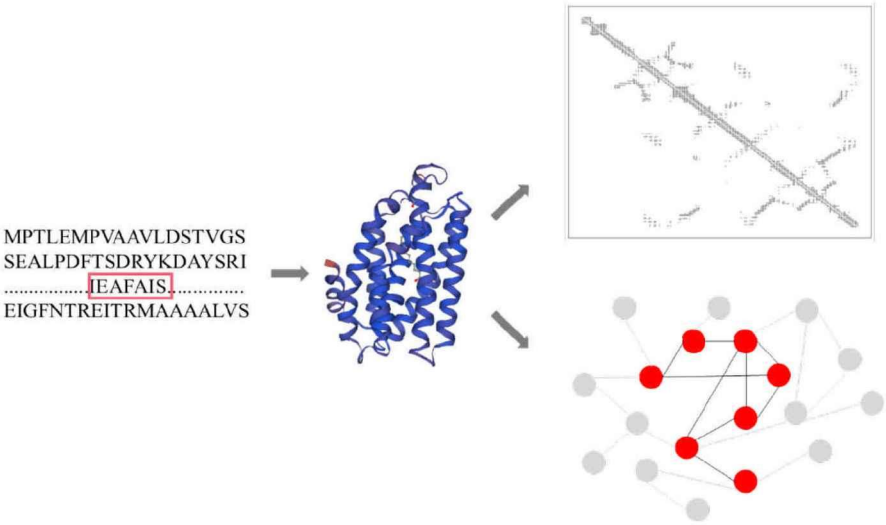


图 3.1 蛋白质接触图概率矩阵以及图结构的可视化

从序列中提取 KAAP 的手段是寻找两个间隔为 K 的氨基酸对，对于接触图来说，间隔等价于路径，间隔为 K 代表两个氨基酸顶点之间的路径长度为 $K+1$，所以提取 CCmap-KAAP 特征的过程就可以转化为求接触图 G 中每一个节点的 $K+1$ 阶邻居。我们用 N_i^K 表示接触图中顶点 A_i 对应的 $K+1$ 阶邻居集合，则 CCmap-KAAP 的计算公式如下：

$$CCmap\text{-}KAAP_j^K = \frac{\sum_{i=1}^{L}\sum_{V\in N_i^K} I(A_iV = Dip_j)}{\sum_{j=1}^{400}\sum_{i=1}^{L}\sum_{V\in N_i^K} I(A_iV = Dip_j)}, 1 \le j \le 400 \quad (3.4)$$

其中 V 是 A_i 的 $K+1$ 阶邻居之一，A_iV 是组合在一起的氨基酸对，也即是二肽类别，一共 400 种，所以 $CCmap\text{-}KAAP^K \in R^{400}$，公式中的分母用来进行归一化操作。

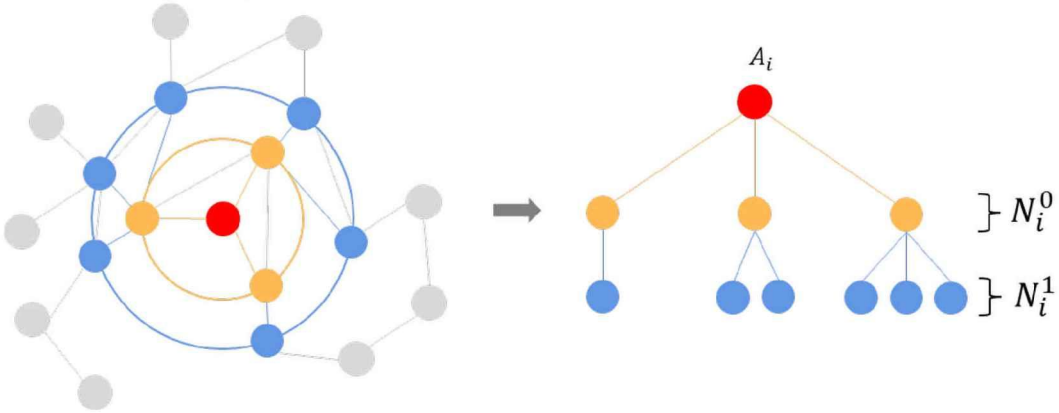


图 3.2 构建接触图的 K+1 阶邻居

图 3.2 展示了接触图的 $K+1$ 阶邻居示意图。因为接触图是一个无向带环图，所以一个节点即可能是源节点的 1 阶邻居，也是源节点的 2 阶邻居或其他阶邻居，比如 $A \leftrightarrow B \leftrightarrow C \leftrightarrow A$，连接的边是无方向的，所以 B、C 两个节点即是 A 的一阶邻居（$A \to B$，$A \to C$），也是节点 A 的二阶邻居（$A \to B \to C$，$A \to C \to B$），对于这种会出现冲突的情况，我们只选择最小的阶数。比如在上面的例子中，B，C 两个节点只是 A 的一阶邻居，而不能作为 A 的二阶邻居。然后我们就可以通过广度优先搜索的方式确定源节点任意且唯一的 $K+1$ 阶邻居集合，从而提取 CCmap-KAAP 特征。

3.1.3 XGboost 模型

为了测试 CCmap-KAAP 特征的有效性，我们使用 XGboost（eXtreme Gradient Boosting）[60]机器学习模型来拟合 CCmap-KAAP 和 KAAP 这两种特征与蛋白质结晶倾向性之间的关联性。XGboost 是一种优化的梯度提升决策树（Gradient Boosting Decision Tree，GBDT）[84]，主要是利用 boosting 算法[85]集成多个弱分类器来更好的评估蛋白质结晶倾向性。GBDT 在训练过程中每次迭代地学习一棵 CART 弱决策树[86]来拟合之前的残差，这里的残差是指前 $t-1$ 轮预测的输出值与真实值之间的误差，通过不断地拟合残差，最后将所有弱分类器串联起来得到一个强分类器。

对于每一条蛋白质序列，我们根据其对应蛋白质接触图来提取 CCmap-KAAP 特征。因为单个氨基酸的组成成分特征 AAC 是最基本的氨基酸组成成分特征，而且对于蛋白质序列和蛋白质接触图而言 AAC 是相同的，所以我们将 CCmap-KAAP 特征和 AAC 特征结合在一起作为 XGboost 模型的输入。具体来说，我们用 x 表示模型输入对应的特征向量，$x \in R^{20+400}$，特征向量的维度是 420。输入 x 所对应的标签 $y \in \{0,1\}$，0 代表负

样本，表示蛋白质无法成功结晶，1 代表正样本，表示蛋白质能够成功结晶。因为结晶倾向性预测属于二分类问题，我们使用了二分类逻辑回归作为 XGboost 模型的目标函数：

$$Loss(y,\hat{y}) = -[y \cdot log(\hat{y}) + (1-y) \cdot log(1-\hat{y})] \tag{3.5}$$

其中 $\hat{y}$ 代表模型的预测输出，是一个实数值，代表蛋白质能够成功结晶的概率。

在模型的训练阶段，我们基于普通 KAAP 特征在 CRYS_DS 数据集上对模型的多个超参数进行了网格搜索。对于 XGboost 模型来说，比较重要的参数包括树的最大深度（M），学习率（v），叶子节点的最小子权重（w）等等。我们先固定学习率，然后对 $M \in \{2,3,5,7,9\}$, $w \in \{3,4,5,6\}$ 进行了网格搜索，一共有 $5 \times 4 = 20$ 种情况，对每一组搜索参数，我们都在 CRYS 训练数据集上使用 5 折交叉验证的方式来评估最优的 AUC 指标。当我们确定了 M 和 w 之后，再对学习率进行了微调，最终确定的参数为 $M = 3$, $w = 6$, $v = 0.1$。除了这三个参数之外，剩余的特征采样率、样本采样率、正则化参数等，我们都使用了 sklearn 工具包中 XGboost 模型的默认值。

3.1.4 实验结果与评估

因为构建蛋白质接触图 G 需要一个超参数阈值 d，所以我们首先对超参数 d 的选取进行了实验验证。然后再验证 CCmap-KAAP 特征对蛋白质结晶倾向性的影响。两个实验用的数据集都是 CRYS_DS 数据集，该数据集的详细信息见第 2.2 章节。

（1）阈值 d 的选取

正如第 3.1.2 章节所述，我们构造蛋白质接触图 G 需要基于一个接触概率阈值 d。为了选择一个合适的阈值 d，我们首先对不同阈值下的蛋白质接触图 G 进行了可视化，如图 3.3 所示。在预测的蛋白质接触图矩阵中，只有当两个残基的接触概率大于 d 才会被看作接触，也即两个残基顶点之间存在边，随着 d 的增大，被认为接触的残基对也会相应减少。在图 3.3 中，黑色的点代表存在边，白色则代表不存在，我们可以看到当 $d = 0$ 时 G 是一个完全稠密的图，随着 d 的增大图变的越来越稀疏。

我们之所以面临这样的问题就在于我们的蛋白质接触图是通过工具 PconsC4 预测得到的，并不是从真实的三维结构中构建得到的，所以预测的蛋白质接触图矩阵中虽然两个残基的接触概率很小，但在真实情况下可能是接触的，或者预测的两个残基的接触概率很大，但在真实情况下可能完全不接触。这种由预测蛋白质接触图所带来的误差是不可避免的，想要减少这种误差，只能使用预测精度更高的工具来生成蛋白质接触图。但是目前的蛋白质接触图预测工具很难平衡计算效率和计算精度，一个预测精度很高的工具往往要花费非常多的计算时间才能得出计算结果。所以我们在对计算效率和计算精度之间进行平衡后，选择了 PconsC4 工具来预测蛋白质接触图。

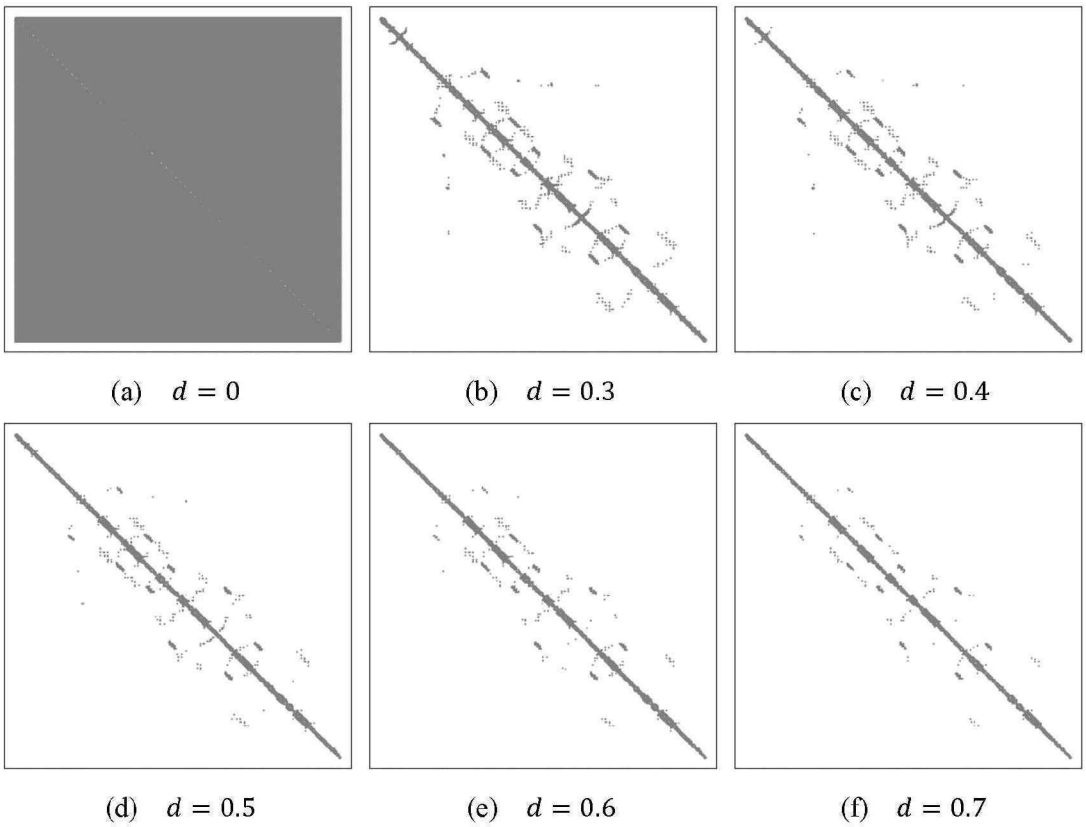


图 3.3 六种概率阈值下的接触图示意图

为了进一步对比不同的阈值 d 对 CCmap-KAAP 特征的影响，我们对不同阈值 d 生成的蛋白质接触图 G 都提取 $K = 0$ 对应的 CCmap-KAAP 特征（CCmap-KAAP^{0,d}），然后使用相同参数的 XGboost 模型在 CRYS_DS 验证数据集上对 CCmap-KAAP^{0,d} 特征进行验证。我们选取了 $d = 0.3, 0.4, 0.5, 0.6, 0.7$ 进行实验，图 3.4 显示了具体的实验结果。值得注意的是，在没有特殊说明的情况下， CCmap-KAAP 特征或者普通的 KAAP 特征默认都包含单个氨基酸的组成成分特征 AAC，也是即输入特征维度一直是 420。

图 3.4 是不同阈值下的 AUC 指标的对比结果，总体而言五种阈值下的 CCmap-KAAP^{0,d} 特征对结晶倾向性影响的差距并不是很大，当 $d = 0.3$ 时取得了最优的 AUC（0.769）。而且我们发现随着 d 的增大，对应的 AUC 结果有一直下降的趋势，这可能是由于 d 的增大导致提取到的有效氨基酸间隔对也一直在减少，从而导致预测结晶倾向性的精度下降。但是我们也不能取过于小的阈值，当 d 较小时，接触图中边的数量会急剧增加，一方面导致计算量的增加，另一方面也会带来冗余且错误的接触边，因此在权衡下我们选择了 $d = 0.3$ 作为最终的接触概率阈值。

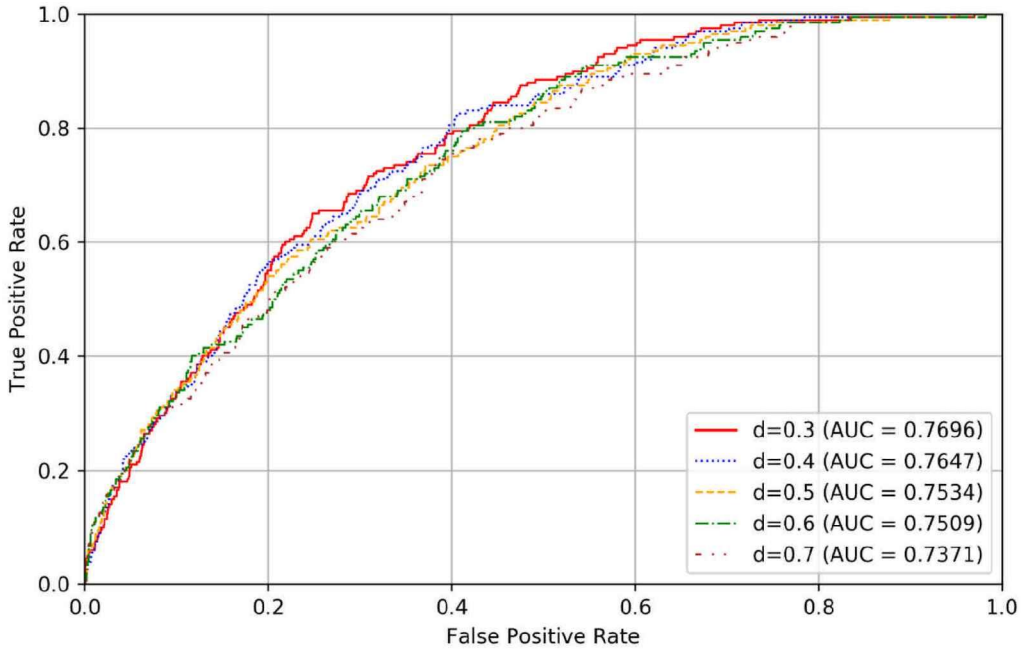


图 3.4 不同接触概率阈值对预测结晶倾向性的影响

（2）验证 CCmap-KAAP 特征的有效性

在选取了最优的阈值 d 之后，为了验证当前阈值下 CCmap-KAAP 特征的有效性，我们首先将其和普通的 KAAP 特征进行对比来判断这两种特征之间的关系。我们选取了 $K=0,1,2,3,4$ 对应的五种间隔，一共提取了 $KAAP^0$、$KAAP^1$、$KAAP^2$、$KAAP^3$、$KAAP^4$、$CCmap\text{-}KAAP^0$、$CCmap\text{-}KAAP^1$、$CCmap\text{-}KAAP^2$、$CCmap\text{-}KAAP^3$、$CCmap\text{-}KAAP^4$ 等 10 种特征，每种特征都默认添加了 AAC 特征，特征维度为 420。为了对这 10 种特征进行分析，我们分别将其作为 XGboost 模型的输入在 CRYS_DS 训练集上进行训练，并在 CRYS_DS 测试集上对这 10 个模型分别进行测试。

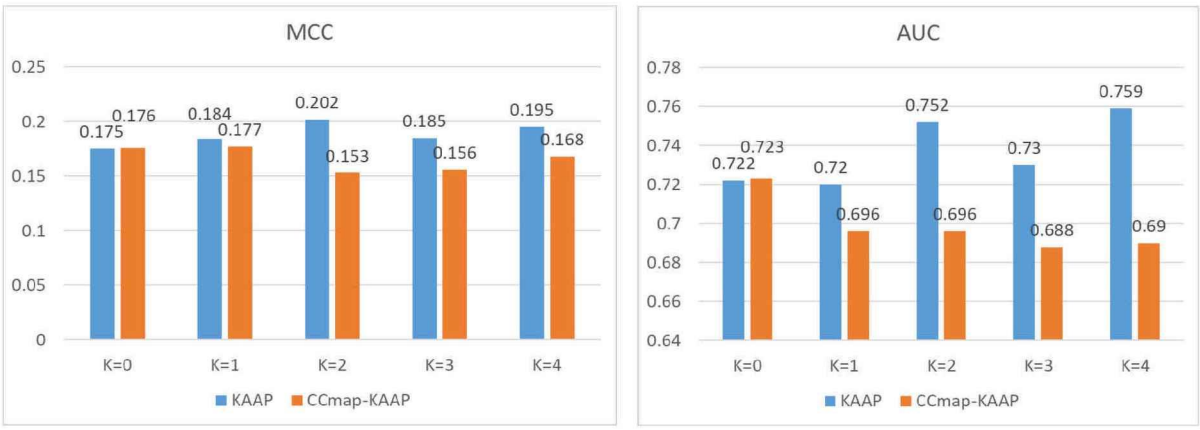
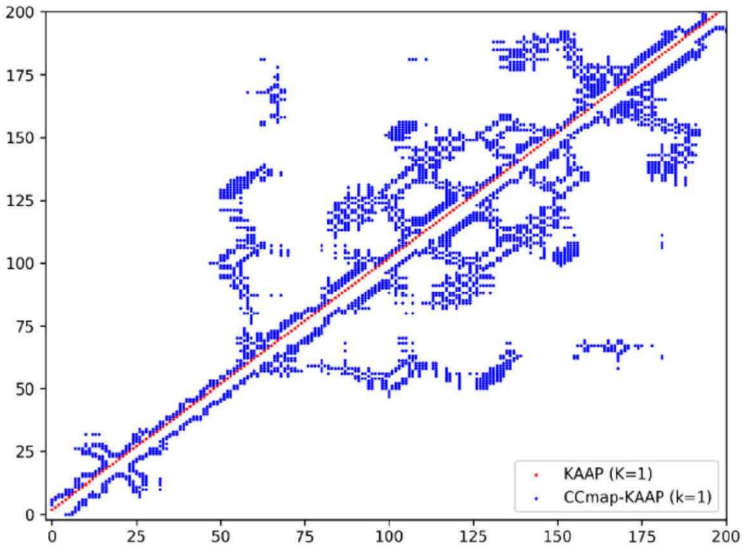


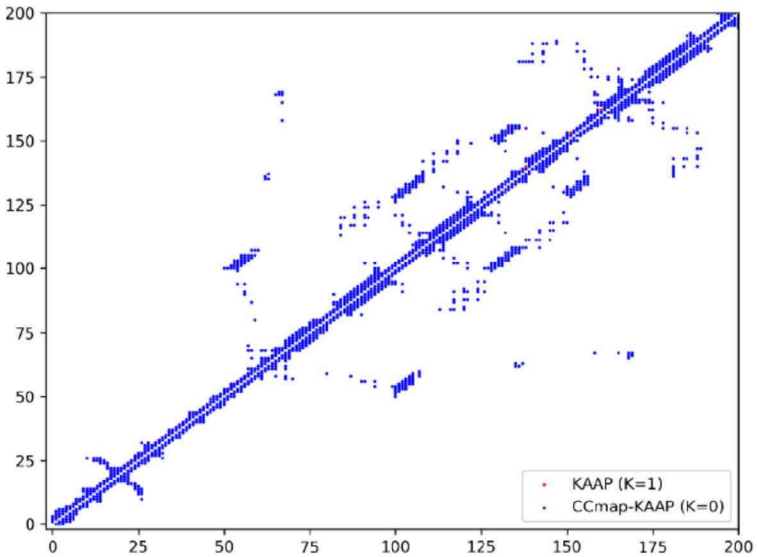
图 3.5 CCmap-KAAP 和 KAAP 特征在 CRYS_DS 测试集上的预测结果

图 3.5 展示了 10 种特征对应的 AUC 和 MCC 指标测试结果，我们发现不论是在 MCC 指标上还是 AUC 指标上，CCmap-KAAP 特征在结晶倾向性预测上的表现似乎都

不如普通的KAAP特征，除了在 $K=0$ 时两种特征的表现基本持平之外，剩下的四种间隔中普通的KAAP特征都取得了更好的结果，而且结果的差距较大，与预期的结果并不相符。从理论上来说，对于一个蛋白质 P，其一级序列上两个相隔为 K（$K \leq 4$）的两个氨基酸 i 和 j 在空间上的间隔也一定等于或小于 K，之所与会小于 K 是因为接触图 G 是一个有环图，氨基酸 i 在空间结构中可能有更短的路径到达 j。因此CCmap-KAAP特征应该包含了普通的KAAP特征，在结果上应该等于或优于KAAP特征对应的预测性能。为了进一步分析CCmap-KAAP特征和KAAP特征之间的关系，我们对 $K=0,1$ 时的特征进行了可视化。



(a) 蓝色代表 CCmap-KAAP^1，红色代表 KAAP^1



(b) 蓝色代表 CCmap-KAAP^0，红色代表 KAAP^1

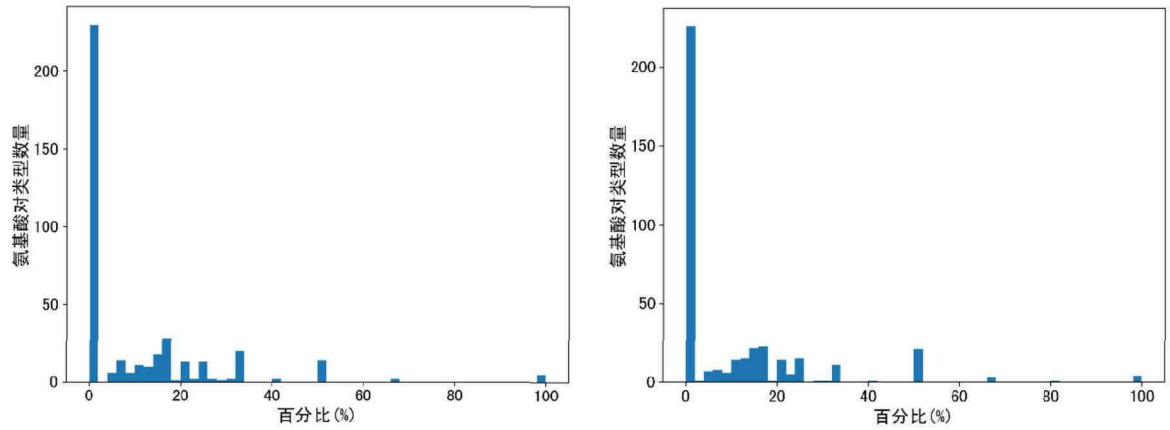
图 3.6 CCmap-KAAP 和 KAAP 特征的可视化

如图3.6所示，横纵坐标代表氨基酸在蛋白质序列中的位置索引，如果两个氨基酸 i 和 j 在序列上间隔为1，我们将对应的坐标点（i, j）标记为红色，如果在接触图 G 上

的空间间隔为 1，则标记为蓝色。通过观察图 3.6 (a)，我们发现 CCmap-KAAP^1 特征基本没有包含 KAAP^1 特征，也即是氨基酸 i 和 j 在序列上间隔为 1，但在空间上间隔为 0。这是由于序列上连续的三个氨基酸在蛋白质接触图中大概率会形成一个环，导致三个氨基酸相互之间间隔都为 0，所以理论上 CCmap-KAAP^0 特征包含了 KAAP^1 特征，正如图 3.6 (b)所示，蓝色的 CCmap-KAAP^0 完全覆盖了红色的 KAAP^1 特征。但是我们通过图 3.5 发现 CCmap-KAAP^0 在预测性能上的表现也没有优于 KAAP^1 特征。通过进一步分析图 3.6，我们发现计算 CCmap-KAAP 特征时对应的氨基酸对数要远远超过普通 KAAP 特征对应的氨基酸对数，那么在计算最终 400 种氨基酸对频率特征时，CCmap-KAAP^0 特征中 KAAP^1 特征的占比会非常低，我们通过以下方式计算 KAAP 特征在 CCmap-KAAP 特征中的占比：

$$Prob_j(K_1,K_2)=\frac{N\_KAAP_j^{K_1}}{N\_CCmap\text{-}KAAP_j^{K_2}},\ 1\le j\le 400 \tag{3.6}$$

其中 $N\_CCmap\text{-}KAAP_j^{K_2}$ 代表在 K_2 间隔下计算 CCmap-KAAP 特征时第 j 类氨基酸对的总数量，$N\_KAAP_j^{K_1}$ 代表在 K_1 间隔下计算 KAAP 特征时第 j 类氨基酸对的总数量。因为一级序列中 $K=0$ 间隔下的氨基酸对在空间中的间隔也一定为 0，一级序列中 $K=1$ 间隔下的氨基酸对在空间中的间隔基本也都为 0，所以 KAAP^0 特征和 KAAP^1 特征基本都包含在 CCmap-KAAP^0 中。



(a) KAAP^0 特征在 CCmap-KAAP^0 中的占比　　(b) KAAP^1 特征在 CCmap-KAAP^0 中的占比

图 3.7 CCmap-KAAP^0 特征中 KAAP^0 和 KAAP^1 特征的占比直方图

图 3.7 展示了 CCmap-KAAP0 特征中 KAAP^1 特征和 KAAP^0 特征占比的直方图，横轴代表占比数值 $Prob_j(K_1,K_2)$，纵轴代表等于该比例的类别数。我们可以发现 400 种氨基酸对类型中，超过 200 种氨基酸对类型的占比处在 0.1%附近，剩余的大部分类别也都只占比 20%左右，意味着 CCmap-KAAP 特征中的大部分氨基酸对是空间意义上的相邻，随着间隔的继续增大，KAAP 特征在 CCmap-KAAP 特征中的占比急剧减小，这也就导致在使用 CCmap-KAAP 特征预测结晶倾向性时，基本只有空间上远距离的氨基酸

对在起主导作用，忽略了序列上相邻的氨基酸对信息。

基于上面的分析，我们可以知道 CCmap-KAAP 特征能够提取蛋白质空间上远距离接触的氨基酸对频率信息，KAAP 代表蛋白质一级序列上近距离接触的氨基酸对频率信息，两者是互补而不是包含的关系。CCmap-KAAP 特征没有优于 KAAP 特征的另一个原因可能是蛋白质接触图的预测精度较差，导致在构建接触图 G 的时候出现了很多冗余且错误的边，从而降低了 CCmap-KAAP 特征的质量。

因为 CCmap-KAAP 特征与 KAAP 特征是互补的关系，所以想要取得好的预测结果，我们可以将不同间隔下的 5 种 CCmap-KAAP 特征和 5 种普通 KAAP 特征叠加起来进行结晶倾向性预测。

表 3.1 不同特征组合在 CRYS_DS 测试集上的预测结果

特征组合	Sen	Spe	Acc	MCC	AUC
KAAP	0.225	0.965	0.911	0.227	0.752
CCmap-KAAP	0.189	**0.969**	0.913	0.202	0.736
KAAP+ CCmap-KAAP	**0.252**	0.965	**0.913**	**0.254**	**0.771**

表 3.1 显示了在 CRYS_DS 测试数据集上的实验结果，为了进行对比，我们也将 5 种 CCmap-KAAP 特征和 5 种普通 KAAP 特征各自整合在一起进行测试。根据实验结果，我们看出 10 种特征整合在一起后取得了最优的预测结果，MCC 和 AUC 分别为 0.254、0.771，比单独使用 KAAP 特征提高了 2.7%（0.254-0.227）和 1.9%，比单独使用 CCmap-KAAP 特征提高了 5.2%和 3.5%。实验结果最终表明了 CCmap-KAAP 特征在结晶倾向性上的有效性，我们相信如果预测接触图能够有更高的精度，那么对于预测结晶倾向性的帮助会更大。

3.2 融合多源蛋白质特征进行结晶倾向性预测

验证了蛋白质接触图以及对应 CCmap-KAAP 特征的有效性之后，我们再次利用该特征进行蛋白质结晶倾向性预测，从之前基于机器学习的模型中可以看出，如果想要获得更好的预测精度，就需要获取与结晶倾向性关联性较高的特征，这也是早期结晶倾向性预测模型的研究重点。比如 OB-Score[12]模型表明 pI-Gravy 两个特征组合在一起对蛋白质是否能够成功结晶影响很大，PPCpred[32]模型通过整合二级结构、AAindex 数据库属性、无序性、相对溶剂可及性等特征提高了结晶倾向性的预测能力，TargetCrys[58]和 DCFCrystal[45]模型都使用位置特异性权重矩阵（PSSM）来辅助预测结晶倾向性。我们通过结合多种对结晶倾向性有较大影响的特征，再融合我们提出的 CCmap-KAAP 特征和普通 KAAP 特征来进行蛋白质结晶倾向性预测，在模型方面使用了 XGboost 机器学习模型，并分别在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 数据集上进行训练来预测多阶段的结晶倾向性，我们的整体模型简称为 CCmapCrys。

3.2.1 多源蛋白质特征

在本章节中我们一共使用了十种特征，分别是蛋白质的等电离点（isoelectric point，pI）、平均疏水性（grand average of hydrophobicity，Gravy）、AAindex 数据库中的氨基酸属性、二级结构特征（secondary structure，SS）、相对溶剂可及性（relative solvent accessibility，RSA）、位置特异性矩阵（Position-Specific Scoring Matrix，PSSM）、基于序列的 K 间隔氨基酸对频率特征（KAAP）、基于蛋白质接触图的 K 间隔氨基酸对频率特征（CCmap-KAAP）、单肽氨基酸组成成分（AAC）以及蛋白质序列长度（Length），这些特征都已经被许多方法证明有助于结晶倾向性预测[7-9]。有关 *K* 间隔氨基酸对频率特征以及单肽氨基酸组成成分特征的信息已经在前面详细介绍过，下面我们详细介绍其他类型的特征，并阐述每种特征的获取和处理方式。表 3.2 对所有的特征进行了总结。

（1）等电离点

蛋白质含有既能解离成带正电荷的氨基，又含有能解离成带负电荷的羧基，可以进行两性电离，受溶液 pH 值的影响，蛋白质在某一溶液中游离成阳离子、阴离子的程度是不一样的，会影响蛋白质在溶液中的带电类型。当蛋白质溶液处于某一 pH 值时，蛋白质游离成阳离子、阴离子的趋势相同，呈电中性，此时我们称该溶液 pH 值为对应蛋白质的等电离点 pI[87]。蛋白质溶液的 pH 值大于对应的等电离点时，该蛋白质带负电荷，反之则带正电荷，不同蛋白质都有各自不同的等电离点。在蛋白质纯化、结晶时，当溶液 pH 等于对应等电离点时，溶液的溶解性最低，蛋白质溶质容易析出，所以等电离点特征在结晶实验中占有很重要的地位。为了计算蛋白质的等电离点特征，我们使用了 IPC[88]工具，该工具有较高的计算精度和运算速度。对于蛋白质的等电离点特征，其特征维度为 1。

（2）平均疏水性

在天然蛋白质中，疏水键是疏水侧链为了避开水分子而聚集在一起的一种相互作用，疏水作用对于蛋白质的稳定性、构象和蛋白质功能具有重要作用[89]。蛋白质分子的疏水性会极大的影响蛋白质在溶液中的溶解性，一般平均疏水性越小，蛋白质溶解度也就越大。疏水性总程度可以用蛋白质各残基疏水值平均和（Gravy）来衡量，常用的计算方法是 Kyte-Doolittle[90]计算法，在本文中我们也依据该算法计算相应蛋白质的 Gravy 特征，该特征的特征维度为 1。

（3）AAindex 数据库

蛋白质中常见的 20 种氨基酸有多方面的特性，这些特性决定了蛋白质结构和功能的特异性和多样性，目前已经有大量的实验和理论来研究单个氨基酸的不同种类特性，并用具体的数值来表示这些特性。AAindex 数据库[91]就存储了其中大部分的氨基酸属性，该数据库主要由两部分组成：AAindex1 收集了氨基酸的多种物理化学以及生物属性，AAindex2 收集了氨基酸突变矩阵信息。这里我们只使用了 AAindex1 部分的氨基酸属性

信息，目前这个部分一共存在566个条目，主要包括了疏水性和能量相关的氨基酸属性。因为AAindex1中包含的特征都是针对于单个氨基酸的，所以我们需要计算整个蛋白质对应的特征信息。对于这566种物理化学属性，我们根据氨基酸对应的数值来分别计算其在整个蛋白质序列上的最小值（AAindex_Min）、最大值（AAindex_Max）和平均值（AAindex_Avg），最终构建的特征大小是 $566 \times 3 = 1698$ 。

表 3.2 多源蛋白质特征预处理

原始特征	预处理后特征	维度	描述
Length	Log_Length	1	对蛋白质序列长度进行Log归一化
AAC	-	20	单个氨基酸在序列中出现的频率
KAAP	-	2000	五种间隔下（$K = 0,1,2,3,4$）连续两个氨基酸对类型在序列中出现的频率
CCmap-KAAP	-	2000	五种间隔下（$K = 0,1,2,3,4$）连续两个氨基酸对类型在蛋白质接触图 G 中出现的频率
pI	-	1	蛋白质呈电中性时对应的溶液PH值
Gravy	-	1	序列中所有氨基酸的平均疏水性
AAindex	AAindex_Min, AAindex_Max, AAindex_Avg	1698	从AAindex中抽取566种物理化学属性，再根据氨基酸对应的数值来分别计算其在整个序列上的最小值、最大值和平均值
SS	Freq_SS8, Dip_Freq_SS8, Tri_Freq_SS8	584	单个二级结构类型、两个连续的二级结构类型、三个连续的二级结构类型在SS8序列中出现的频率
RSA	RSA20	20	在 20 种不同阈值下分别计算对应暴露残基数量占总残基数量的百分比
PSSM	PsePSSM	180	计算蛋白质序列中所有氨基酸突变为氨基酸类型 j 的总平均距离和 t 间隔平均距离

（4）二级结构

蛋白质的结构主要分为四种层次类型：一级结构、二级结构、三级结构和四级结构。二级结构是多肽链中各原子在局部的空间排布，即多肽链主链构象称为蛋白质的二级结构。对于蛋白质结晶倾向性预测而言，如果我们想要使用二级结构来辅助预测，需要使用预测工具来预测二级结构，而不能使用真实的二级结构。因为需要预测结晶倾向性的蛋白质样本一般都是没有完成三维结构测定的，不能使用真实的三维结构和二级结构来指导结晶倾向性预测。我们在本文中使用SCRATCH[92]工具来从头预测蛋白质的二级结构，并以两种形式输出预测的蛋白质二级结构，一种是分为三类的二级结构（SS3），包括螺旋（Helix，H）、卷曲（Coil，C）和折叠（Strand，E）；另一种是按照DSSP[93]

定义分为八类的二级结构（SS8），包括 3 转角螺旋（G）、α螺旋（H）、β螺旋（I）、氢键转角（T）、β折叠（E）、独立β桥内的残基（B）和卷曲（S）。这两种二级结构的输出都是长度为 L 的序列，L 代表蛋白质序列的长度，其中每一个位置是三种或八种类别之一，代表当前位置的氨基酸在二级结构中属于哪种结构类型。因为 8 种类别的二级结构包含的信息可能更丰富，所以我们使用了 SS8 作为输入。我们类比氨基酸组成成分特征来计算二级结构的组成成分特征，然后分别计算了单个二级结构（Freq_SS8）、两个连续的二级结构类型（Dip_Freq_SS8）、三个连续的二级结构类型（Tri_Freq_SS8）在 SS8 中的频率，最终构建的特征大小是 $8+8^2+8^3=584$ 。

（5）相对溶剂可及性

溶剂可及性一般也可以被描述为可及表面积或溶剂可及表面积，代表了溶剂可接触的生物分子表面积，能够极大的影响蛋白质的折叠和稳定性[94]，通过计算出的溶剂可及性数值，蛋白质的氨基酸残基可以分为掩埋和暴露两类。对于给定的蛋白质，我们也使用 SCRATCH 工具来预测蛋白质的溶剂可及性，该工具输出 acc 和 acc20 两种类型的相对溶剂可及性，这两种输出都是长度为 L 的序列，L 代表蛋白质序列的长度。对于 acc，当计算出来的氨基酸溶剂可及性小于 25%阈值时，定义当前位置的氨基酸为掩埋，相反则为暴露，所以 acc 是一个只包含 1 和 0 的序列。对于 acc20，则是定义了 20 种阈值，区间从 0 到 95%，间隔为 5%，acc20 序列中的某一个位置存放了其中一种阈值，代表当前位置的氨基酸只有在其对应的溶剂可及性大于该阈值时才能定义为暴露。在这里我们使用了 acc20 作为我们的特征输入，并在 20 种不同阈值下分别计算对应暴露残基数量占总残基数量的百分比，最终构建出了大小为 20 的特征（RSA20）。

3.2.1.6 位置特异性矩阵

位置特异性矩阵 PSSM 是一个维度为 $L\times 20$ 的特征矩阵，存储了蛋白质的进化信息。矩阵中某一元素 $PSSM(i,j)$ 代表了位置 i 上的氨基酸转变为氨基酸类型 j 的保守突变概率。PSSM 是通过 PSI-BLAST[43]程序多次迭代获取的，当一个残基在多序列比对的多次迭代中保持不变时，它可能具有很重要的生物学意义，这也就体现了 PSSM 矩阵的重要意义。由于 PSSM 矩阵不能直接送入到 XGboost 模型中，所以也需要对其进行预处理。我们使用 Chou 等人提出的方法依据 PSSM 矩阵提取伪特异性权重矩阵特征（PsePSSM）[95]。为了获取 PsePSSM 特征，我们首先对 PSSM 矩阵进行归一化：

$$N_{PSSM}(i,j)=\frac{1}{1+e^{-PSSM(i,j)}},1\le i\le L,1\le j\le 20 \tag{3.7}$$

其中 $N_{PSSM}(i,j)$ 是归一化后的结果。然后对归一化后的矩阵求解蛋白质序列中所有氨基酸突变为氨基酸类型 j 的平均概率 $v_{pssm}=(v_1,v_2,\ldots,v_{20})$：

$$v_j=\frac{1}{L}\sum_{i=1}^{L}N_{PSSM}(i,j) \tag{3.8}$$

接下来我们进一步计算蛋白质序列中所有氨基酸突变为氨基酸类型 j 的 t 间隔平均距离

$\mu_{PSSM}^{t}=(\mu_1^t,\mu_2^t,\ldots,\mu_{20}^t)$：

$$\mu_j^t=\frac{1}{L-t}\sum_{i=1}^{L-t}[N_{PSSM}(i,j)-N_{PSSM}(i+t,j)]^2 \tag{3.9}$$

我们选取 $t=1,2,3,4,5,6,7,8$ 得到 8 种距离下的平均距离 μ_{PSSM}。最后我们将 v_{pssm} 和 μ_{PSSM}^{t} 结合起来形成最终的 PsePSSM $=\left(v_{pssm},\mu_{PSSM}^{1},\mu_{PSSM}^{2},\ldots,\mu_{PSSM}^{8}\right)$。通过这种计算方式，我们可以得到维度为 $(1+8)\times 20=180$ 的 PsePSSM 特征。

最终我们将所有特征整合起来一共得到了维度大小为 6505 的特征输入。

3.2.2 实验结果与评估

我们将处理后的特征输入到 XGboost 模型中进行训练，因为是进行多阶段的结晶倾向性预测，所以我们分别在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 训练数据集上训练了四个模型来预测蛋白质材料生产失败、纯化失败、晶体生产失败和结晶成功的概率，并在对应测试集上进行评估。在模型对比方面，我们首先与 PPCpred、fDETECT、CrysalisI 和 CrysalisII等四个多阶段预测模型在四种测试集上进行对比，其次单独与 OB-Score、ParCrys 和 TargetCrys 三种模型在 CRYS_DS 测试数据集进行对比，因为这三种结晶倾向性预测模型都是单阶段预测器，只能预测蛋白质是否能够成功结晶，上述这些模型的详细信息见第 1.3 章节。值得注意的是，所有对比模型的预测结果是从对应的 Web 服务器上获取的。

表 3.3 是 CCmapCrys 模型与四种多阶段预测模型在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 测试集上的对比结果。我们发现 CCmapCrys 模型在四个测试集上都取得了最优的 Acc、MCC 和 AUC 值，在 Acc 指标上比其余四种模型平均高出了 9.3%、12.9%、14.2%、20.1%，在 MCC 指标上比其余四种模型平均高出了 10.6%、13.1%、18.2%、12.4%，以及在 AUC 指标上比其余四种模型平均高出了 8.7%、8.3%、11.9%、11.3%。对比结果显示了我们的模型要远远优于其他基于机器学习的多阶段结晶倾向性预测方法。主要原因是我们新设计的 CCmap-KAAP 特征与多种蛋白质序列特征整合在一起之后非常有助于结晶倾向性预测。

表 3.4 是 CCmapCrys 模型与三种单阶段预测模型在 CRYS 测试集上的对比结果，从中我们可以发现 CCmapCrys 模型在 Spe、Acc、MCC 和 AUC 四种指标上取得了最优结果，其中 MCC 指标的值为 0.304，比排名第二的 OB-Score 模型高了 15.1%，比排名最低的 TargetCrys 模型高了 19.7%，表明了我们的 CCmapCrys 在结晶倾向预测精度上取得了非常优异的表现。同时我们发现我们的模型相比于其它三种模型在 Sen 指标上比较低，但在 Spe 指标上非常高，这可能是由于 CRYS_DS 数据集中负样本的数量要比正样本的数量高出 12 倍左右，导致模型倾向于将负样本预测正确。

表 3.3 CCmapCrys 模型与四种多阶段预测模型对比结果

测试数据集	模型	Sen	Spe	Acc	MCC	AUC
MF_DS	PPCpred	**0.657**	0.537	0.619	0.184	0.628
	fDETECT	0.440	**0.819**	0.531	0.216	0.650
	CrysalisI	0.599	0.631	0.621	0.215	0.639
	CrysalisII	0.609	0.639	0.629	0.232	0.651
	CCmapCrys	0.591	0.739	**0.693**	**0.318**	**0.729**
PF_DS	PPCpred	**0.754**	0.491	0.686	0.231	0.667
	fDETECT	0.413	0.776	0.506	0.171	0.622
	CrysalisI	0.376	0.781	0.677	0.157	0.600
	CrysalisII	0.624	0.661	0.652	0.254	0.655
	CCmapCrys	0.442	**0.869**	**0.759**	**0.334**	**0.719**
CF_DS	PPCpred	0.296	0.917	0.749	0.273	0.654
	fDETECT	0.291	0.883	0.720	0.209	0.594
	CrysalisI	**0.979**	0.073	0.730	0.126	0.499
	CrysalisII	0.055	**1.000**	0.315	0.126	0.527
	CCmapCrys	0.910	0.400	**0.770**	**0.365**	**0.668**
CRYS_DS	PPCpred	0.324	0.876	0.836	0.150	0.669
	fDETECT	0.649	0.727	0.721	0.211	0.718
	CrysalisI	0.667	0.673	0.672	0.184	0.705
	CrysalisII	**0.685**	0.647	0.650	0.177	0.712
	CCmapCrys	0.279	**0.971**	**0.921**	**0.304**	**0.814**

表 3.4 CCmapCrys 模型与三种单阶段预测模型的对比结果

模型	Sen	Spe	Acc	MCC	AUC
OB-Score	**0.937**	0.321	0.365	0.153	0.656
ParCrys	0.712	0.516	0.530	0.118	0.615
TargetCrys	0.802	0.399	0.428	0.107	0.608
CCmapCrys	0.279	**0.971**	**0.921**	**0.304**	**0.814**

3.2.3 特征有效性分析

模型对比实验表明了我们的 CCmapCrys 模型拥有较高的预测性能，本章节我们继续对模型中使用的特征进行有效性分析，探寻哪些特征对于结晶倾向性有更大的影响。我们使用 sklearn 中的工具对 CRYS_DS 数据集上训练好的 XGboost 模型进行特征重要性分析，图 3.8 展示了对 XGboost 模型影响最大的前 20 个特征（共 6505 种特征），我

们发现等电离点、序列长度、溶剂可及性以及位置特异性矩阵等特征对于结晶倾向性预测的帮助是最大的，剩余影响较大的特征中就包括了我们新提出的CCmap-KAAP特征，这进一步表明了依据蛋白质接触图来提取 K 间隔的氨基酸对频率特征是有意义的。

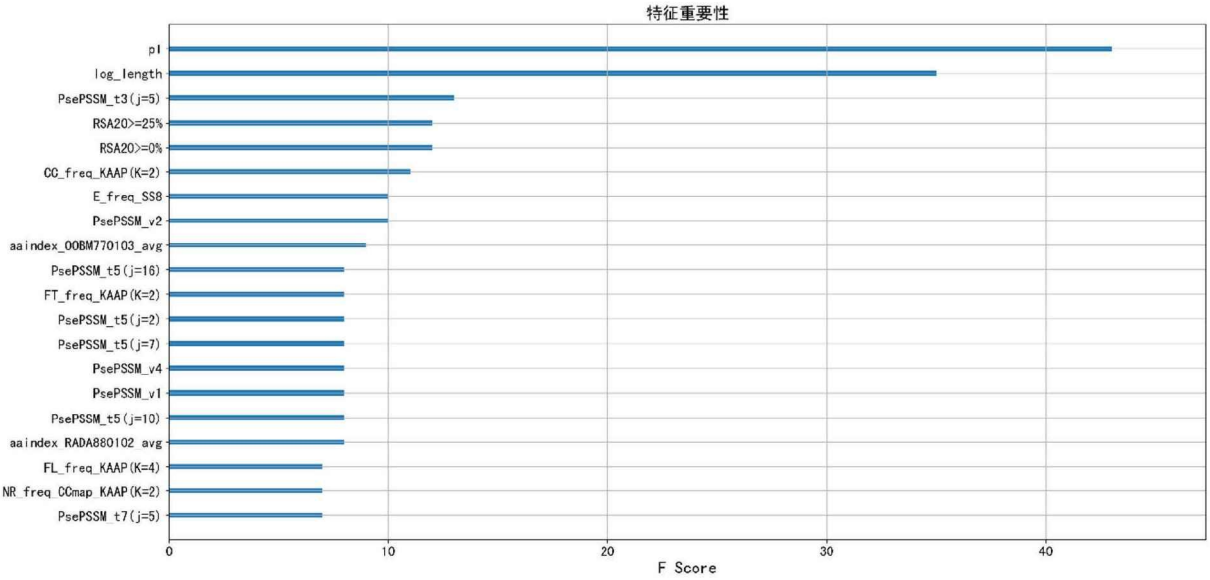


图 3.8 特征重要性

3.3 本章小结

本章我们根据蛋白质接触图设计了一种新的空间 K 间隔氨基酸对组成成分特征CCmap-KAAP，通过与普通的基于序列提取的 K 间隔氨基酸对组成成分特征 KAAP 进行实验对比，我们发现 CCmap-KAAP 包含的空间信息与 KAAP 特征包含的序列信息是互补的，两者结合在一起能够大大提高结晶倾向性的预测能力。然后我们继续将 CCmap-KAAP 特征与其它多种蛋白质序列特征整合在一起并使用 XGboost 模型进行多阶段的结晶倾向性预测，整体模型命名为 CCmapCrys。通过与其它基于机器学习模型的结晶倾向性预测方法进行对比，我们发现我们的 CCmapCrys 具有非常好预测性能。我们进一步对该模型进行特征分析，发现等电离点、序列长度、位置特异性矩阵和相对溶剂可及性特征在蛋白质结晶倾向性预测中都占有非常重要的比重。

4 基于图注意力网络的蛋白质结晶倾向性预测

第三章中我们利用蛋白质接触图设计了一种新特征 CCmap-KAAP，并用它来进行结晶倾向性预测，虽然取得了不错的预测性能，但是仍然有很多的提升空间。从特征的角度来看，常用于结晶倾向性预测的特征主要是基于序列和结构的特征。氨基酸组成成分、氨基酸物理化学属性、长度、位置特异性矩阵等常用特征都是基于序列衍生出来的，虽然位置特异性矩阵包含了一定程度的保守性突变信息，但是它仍然是对蛋白质序列进行多序列比对得到的，而二级结构、相对溶剂可及性等常用信息是属于结构上的特征。我们引入的蛋白质接触图相比于这些特征提供了更高层次的三维结构信息，这也是CCmap-KAAP 特征能帮助预测结晶倾向性的原因之一。

CCmap-KAAP 特征虽然提取了部分空间结构信息，但是它属于手工提取的特征，对预测性能的提升有很大的限制。随着深度学习技术快速的发展，其在图像、语音、文本都多个领域都取得了卓越的成果，相比于传统的机器学习技术，深度学习的一大优势就是能够自动地提取更有效的特征，从而获得相较于手工特征更好的结果。我们希望能够利用一种深度学习技术来充分地提取蛋白质接触图中隐含的空间特征，比如远距离氨基酸之间的相互作用信息。最终我们选取了图卷积模型来提取接触图的特征，主要是考虑到相比于普通的卷积模型，图卷积模型能更有效的提取蛋白质接触图中的空间信息。普通卷积模型中的卷积核是矩阵形式，比如对于 3x3 的卷积核，在提取局部特征时默认 9 个位置在空间上是紧密相连且等权重的。这个条件对于图像是成立的，但是对于非结构化的数据就是不成立的。因为在图结构数据中，一个中心节点的周围邻居节点在数量上是不固定的，而且由于边权重的不同，这些邻居节点相对于中心节点的权重也大概率不同，所以普通的卷积模型难以直接应用到图结构的数据中。相比之下，图卷积模型更适合来解决图结构数据对应的问题[96]。

早期的图神经网络一般利用拉普拉斯矩阵或其变体来构建卷积算子，从而刻画节点的局部特征，比如 ChebyNet[97]、GCN[98]等。现在更常用的图卷积神经网络是基于聚合函数的方法，通过定义聚合函数来聚合中心节点及其对应邻居节点的信息，比如GraphSAGE[99]模型提出图采样聚合网络，通过对邻居节点进行随机采样，再聚合采样的邻居节点信息来更新目标节点信息。通过多次迭代，每个节点都能聚合到一定范围内的邻居节点信息，并根据这种局部结构信息来学习到更有效的节点表征。对于蛋白质接触图而言，氨基酸代表顶点，接触代表顶点之间的边信息，则基于蛋白质接触图预测结晶倾向性的问题可以看作一个图分类任务。因为图卷积的过程主要是学习单个节点的有效特征，所以对整个图进行分类一般需要再次聚合所有节点信息得到图的表示信息，然后将该图的表示输入到分类器中以预测图的标签。由于图卷积模型在处理图结构数据时的

卓越表现，它已经被广泛地用于推荐系统、生物化学、计算机视觉等多个领域中[100-102]。

在本章节中，我们基于图注意力神经网络（GAT）[61]和蛋白质接触图来预测蛋白质结晶倾向性。GAT 模型是一种特殊的图神经网络模型，它使用自注意力机制来聚合节点之间的信息，在聚合过程中，每个节点的权重依赖于节点的特征信息，从而更灵活的自适应于不同的任务。预测的蛋白质接触图包含了两个氨基酸之间的接触概率，我们将这个接触概率作为两个顶点之间的边权重，为了利用边的权重信息，我们在 GAT 网络中加入了边特征的信息，使得每个节点的权重不仅依赖于节点的特征信息，还依赖于对应的边特征信息。为了提高蛋白质结晶倾向的预测精度，我们仍然使用了基于蛋白质序列推导的特征信息作为模型输入的一部分，使用的特征包括蛋白质的等电离点（isoelectric point，pI）、平均疏水性（grand average of hydrophobicity，Gravy）、AAindex 数据库中的氨基酸属性、预测的二级结构特征（predicted secondary structure，PSS）、预测的相对溶剂可及性（predicted relative solvent accessibility，PRSA）、位置特异性矩阵（Position-Specific Scoring Matrix，PSSM）、氨基酸序列编码（Amino Acid Coding, AACD）以及蛋白质序列长度（Length）。相比较于传统的机器学习模型，我们不需要将这些特征转换为固定长度的一维向量就可以直接整合到蛋白质接触图中作为模型的输入，这就减少了特征变换带来的信息损失。我们将构建的模型命名为 GCmapCrys，并在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四个数据集上训练来进行多阶段的结晶倾向性预测，最终经过与多个现有的模型进行对比，我们发现 GCmapCrys 取得了最优的预测性能。

4.1 蛋白质的图结构表征

我们在第 3.1.2 章节中已经介绍了如何根据蛋白质接触图生成对应的蛋白质图 G，在本章节中我们继续对图 G 添加对应的点特征 X 和边特征 E，从而将整个蛋白质图 G 作为图注意力模型的输入。图 4.1 展示了构造蛋白质图 G 的流程。

对于一条长度为 L 的蛋白质序列，我们先使用 HHblits[74]和 PconsC4[77]工具生成预测的蛋白质接触图矩阵 P，再根据接触阈值 d 来生成蛋白质图 G。与第一章节的接触阈值相同，我们取 $d = 0.3$，当 $P(i,j) > d$ 时认为氨基酸 i 和 j 之间相互接触，因为我们需要保留边的权重信息，也即接触的概率 $P(i,j)$，所以我们用以下的公式表示 E：

$$e_{i,j} = P(i,j) \tag{4.1}$$

$$E = \{e_{i,j} | e_{i,j} > d\}, E \in R^{M\times 1} \tag{4.2}$$

其中 $e_{i,j}$ 表示蛋白质图中单条边的信息，M 表示边的数量。

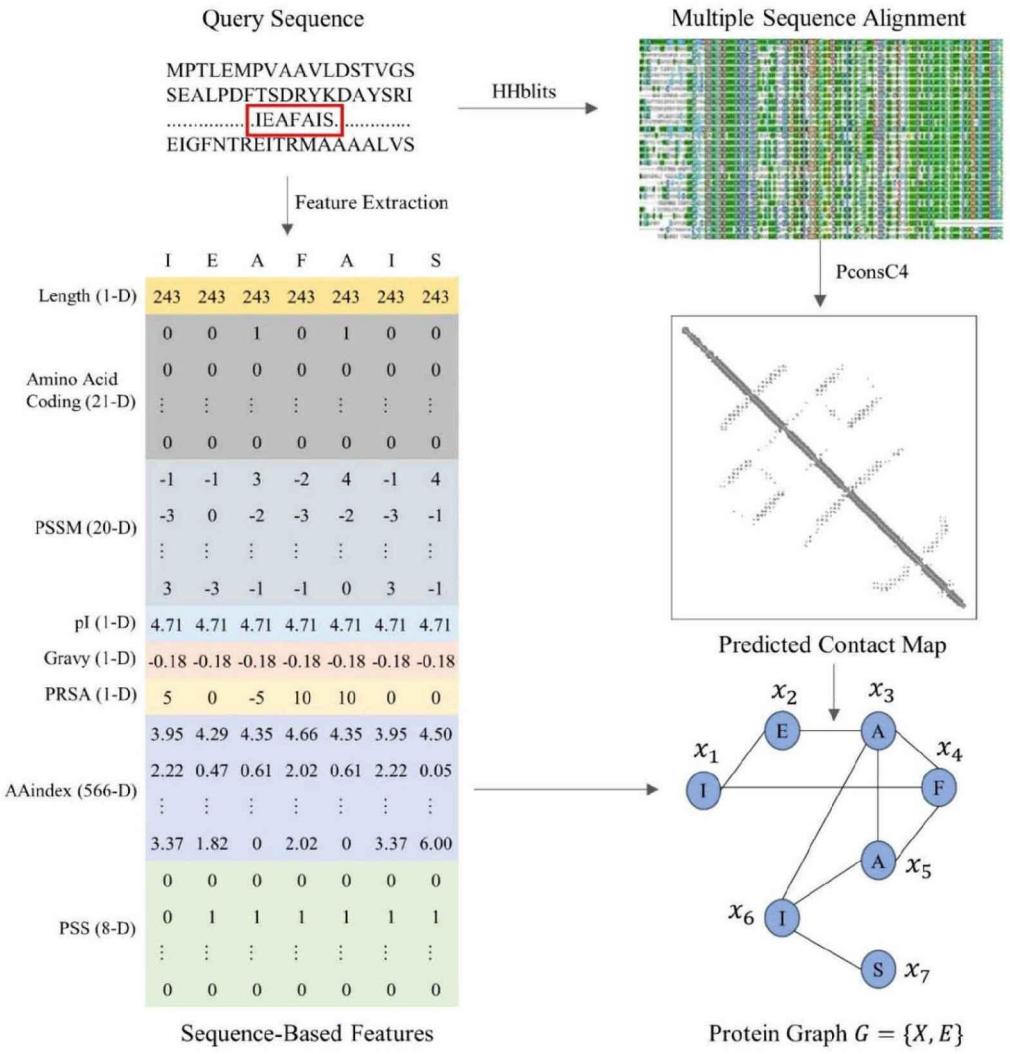


图 4.1 蛋白质图的构造过程

在节点特征方面，我们使用了与第一章节相同的基于序列推导得到的蛋白质特征信息，但是在处理方式上有所不同。因为对于第一章中的机器学习模型来说，我们需要将这些特征转换为固定长度的一维特征向量从而确保能够以正确的格式输入到模型中。但是这就带来了两个问题，首先是人工的特征变换如果没有较强的先验知识所指导，势必会带来一定的信息损失；其次是转换后的特征代表的是整个蛋白质的特征，无法再表示单个氨基酸的特征。所以为了更好的利用这些基于序列的蛋白质特征，我们将每种特征都映射到接触图中的氨基酸节点上。具体来说，对于氨基酸序列编码（AACD）特征，我们使用 one-hot 编码来离散化 21 种氨基酸种类，从而获得一个 $L \times 21$ 维的特征向量，每个氨基酸节点都对应一个类别编码，其中 21 包括了 20 种标准氨基酸类别和一种非标准氨基酸类别；对于位置特异性矩阵（PSSM）特征，我们按照公式 3.7 进行归一化之后获得一个 $L \times 20$ 维的特征向量，每个氨基酸节点对应 20 个特征，表示当前位置的氨基酸突变为 20 种氨基酸的保守性概率；对于预测的二级结构（PSS），我们使用了 8 种类

别的 SS8 作为输入，并使用 one-hot 编码来离散化 8 种二级结构类别，从而获得一个 $L \times 8$ 维的特征向量，每个氨基酸节点都对应一个二级结构的类别编码；对于预测的相对溶剂可及性特征（PRSA），我们使用 20 种不同阈值下的 RSA20 作为输入，将每个氨基酸位置对应的阈值除以 100 可以近似得到单个氨基酸暴露的概率，此时的特征维度是 $L \times 1$；而 AAindex 数据库中的 566 种氨基酸物理化学属性可以直接对应到单个氨基酸节点的特征上，所以我们将这 566 种氨基酸属性各自归一化之后就可以得到一个 $L \times 566$ 维的特征向量，每个氨基酸节点可以对应 566 种属性；剩余的等电离点（pI），平均疏水性（Gravy）和序列长度（Length）三种特征都属于蛋白质层级的特征，无法分解为单个氨基酸的特征，所以我们认为这三种特征是被所有氨基酸节点所共有的，从而在每个氨基酸节点都添加这三种特征，从而形成了一个维度为 $L \times 3$ 的特征向量。最终每一个氨基酸节点我们都得到 619（$21 + 20 + 8 + 1 + 566 + 3 = 619$）种特征，蛋白质图 G 的全部节点表征 $X \in R^{L \times 619}$ 。

4.2 GCmapCrys模型架构

我们提出了一个新的基于图注意力框架的结晶倾向性预测模型 GCmapCrys，该模型以蛋白质接触图作为信息来源，通过多次使用图注意力网络来捕捉蛋白质接触图的空间信息，从而帮助预测蛋白质的结晶倾向性。GCmapCrys 模型的整体框架如图 4.2 所示。

GCmapCrys 模型的输入是一条蛋白质查询序列，输出是预测的结晶倾向性概率。首先，长度为 L 的蛋白质输入序列会被转化为一个蛋白质图 G，它包含 L 个氨基酸节点和 M 条接触边，其中每个节点 x 被表征为一个 619 维的特征向量，每条边 e 代表残基间接触的概率，表示图 G 中两个氨基酸顶点的边权重信息。然后这个蛋白质图 G 被输入到一个图注意力网络结构中预测对应蛋白质的结晶倾向性概率。这个图注意力网络模型可以分为以下三个步骤：首先是三个连续的图注意力层（Graph Attention Layer），用来从蛋白质接触图中提取氨基酸之间的空间相互作用信息。每一个图注意力层都会对当前输入的所有节点 X 和边 E 进行更新，将输入的图 $G = \{X, E\}$ 更新为一个新图表征 $G' = \{X', E'\}$，从而提高整个蛋白质图的表征能力；然后经过三次更新后的蛋白质图 G' 会被送入到一个全局池化层（Global Pooling Layer）中，该池化层的目的是对所有节点 X 进行平均池化来得到整个蛋白质图的特征向量；最后这个代表整张图的全局平均特征向量会经过两个全连接层构建的分类器，其中第二个全连接层后使用一个 Sigmoid 函数来输出蛋白质结晶倾向性的预测分数。

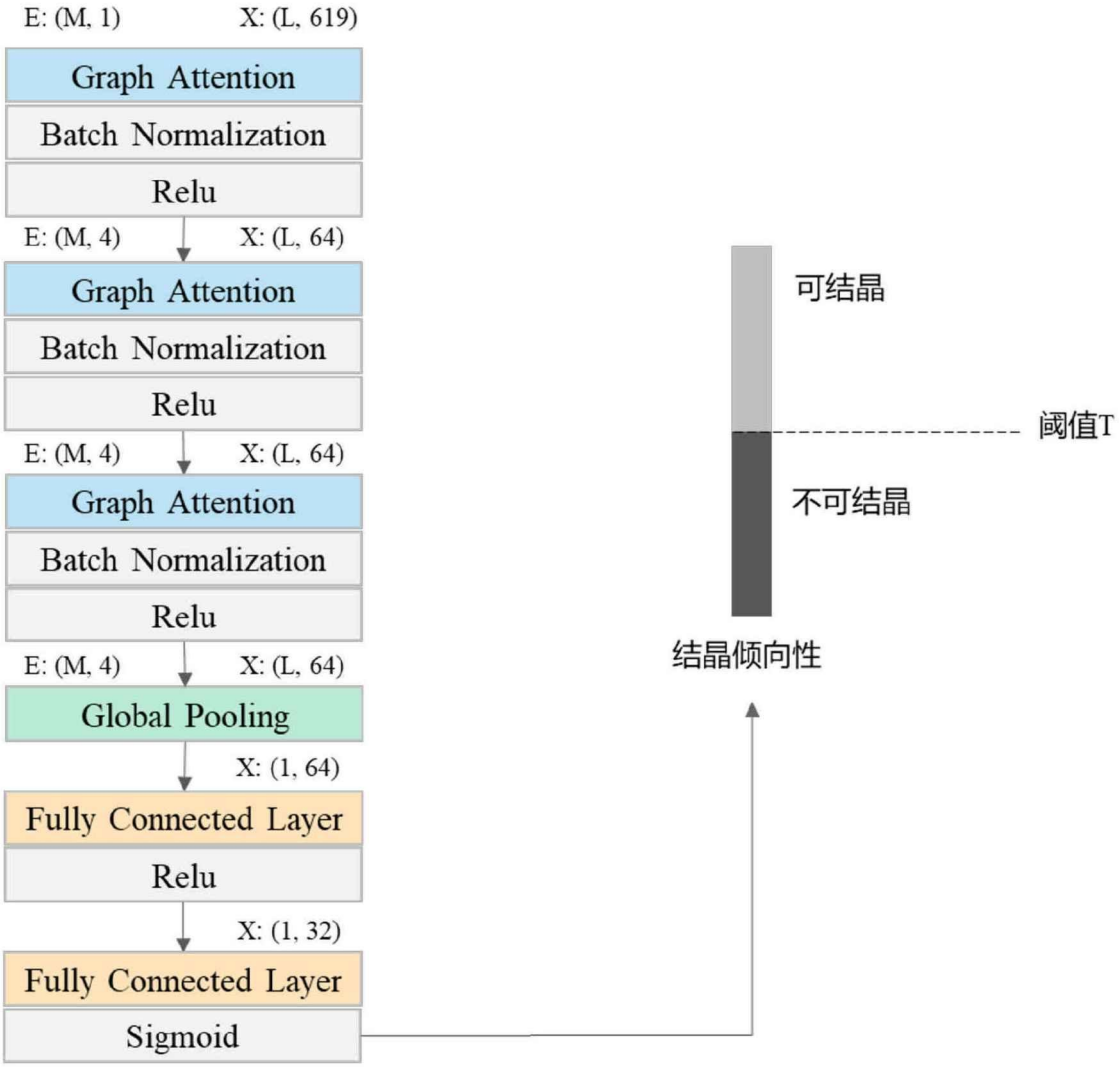


图 4.2 GCmapCrys 模型架构

4.2.1 图注意力层

为了便于描述图注意力层的前向传播过程，我们定义氨基酸节点 x 的特征维度为 D，$x \in R^D$，初始时 $D = 619$。所有的节点共同形成了一个特征矩阵 $X = [x_1, x_2, \ldots, x_L]^T$，$X \in R^{L \times D}$，其中 L 是蛋白质序列的长度，也即是氨基酸节点的数量，$\cdot^T$ 代表矩阵转置。考虑到氨基酸之间的接触是有概率的，我们将接触的边 $e_{i,j}$ 看作一个特征向量，特征向量的维度表示为 F，$e_{i,j} \in R^F$，初始时 $F = 1$，仅代表两个氨基酸之间接触的概率。所有的边形成一个特征矩阵E，$E \in R^{M \times F}$，其中 M 代表边的数量。

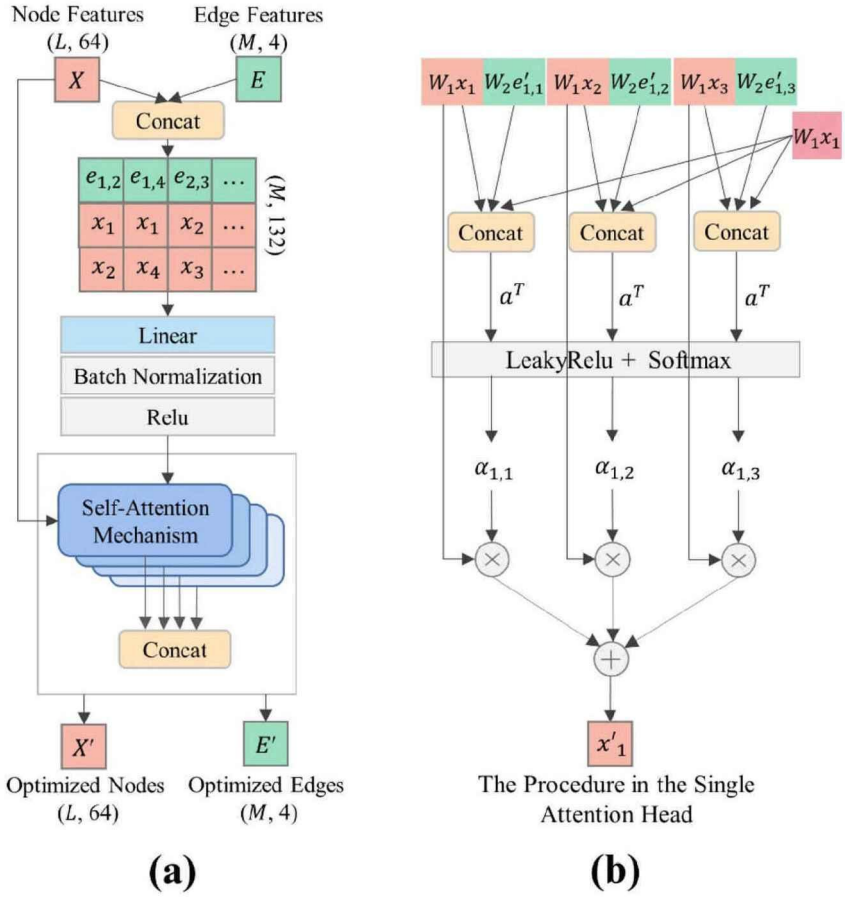


图 4.3 图注意力层

图注意力网络是一种特殊的图卷积模型，从图的消息传播角度来看，图卷积的核心在于如何定义节点之间的聚合函数，基于不同的聚合函数，每个节点会以不同的方式和它对应的邻居节点进行信息叠加，这种叠加的过程能够使得节点学习到它周围的局部结构信息。在经过多次迭代地聚合之后，每一个节点都会学习到更远距离的结构信息。这种信息的叠加也可以类比为图像卷积模型中的卷积过程，可以通过卷积核来提取图像某个范围内的局部特征信息，并不断叠加卷积层来扩大感受野，不同的是图卷积中聚合的邻居节点数目是不固定的。图注意力网络的聚合函数利用了自注意力机制，通过计算中心节点和所有邻居节点的相互作用关系来更新中心节点特征。但是这里的相互关系是只根据节点特征计算得到的，没有考虑带有权重的边信息，而边特征与自注意力机制计算得到的相互作用关系是相似的，都可以代表节点之间的一种权重信息，因此我们将边特征加入到图注意力模块中来帮助聚合节点的信息。具体来说，我们先结合氨基酸节点 X 的信息来更新边 E 的特征得到 E'，然后再利用自注意力机制和更新后的边特征 E' 来计算中心节点和邻居节点之间的权重系数 α。这个权重系数被用来帮助聚合这些节点的信息，从而得到具有更强表征能力的节点特征 X'，这个过程可以看作更新节点的过程。图 4.3（a）展示了 GCmapCrys 模型中第二个图注意力层的传播过程，下面我们具体阐述如

何更新边特征和节点特征。

（1）更新边特征

对于一条边 $e_{i,j}$ 而言，能够影响这条边对应特征信息的最大因素是节点 x_i 和 x_j，所以为了得到有效的边特征，我们需要结合 $e_{i,j}$ 对应的两个氨基酸节点特征信息来帮助更新边特征，这个过程可以表示为：

$$e'_{i,j} = \sigma\left(W\left(e_{i,j} \parallel x_i \parallel x_j\right)\right), \quad e'_{i,j} \in R^{F'} \tag{4.3}$$

其中 $\parallel$ 表示特征向量进行拼接，边 $e_{i,j}$、顶点 x_i 和顶点 x_j 进行拼接之后会形成一个维度为 $F + 2D$ 的特征向量。F' 代表更新后边特征 $e'_{i,j}$ 的维度。$W \in R^{F' \times (F+2D)}$ 是权重矩阵，用来对拼接后的特征进行线性变化，σ 代表 ReLU 非线性激活函数。

（2）更新节点特征

节点更新的步骤是通过聚合邻居节点的特征信息来更新自身特征，如图 4.3（b）所示。具体的聚合过程可以用下述公式所表示：

$$x'_i = \sigma\left(\sum_{j \in N(i)} \alpha_{i,j} W_1 x_j\right), x'_i \in R^{D'} \tag{4.4}$$

其中 $N(i)$ 中心节点 x_i 的一阶邻居节点集合，D' 代表更新后点特征 x'_i 的维度，$W_1 \in R^{D' \times D}$ 是一个权重矩阵，用来对节点 x 进行线性变化，对于图中的所有节点而言，线性变化的权重 W_1 是共享的。$\alpha_{i,j}$ 是由自注意力机制计算出来的中心节点 x_i 和邻居节点 x_j 之间的权重系数，是一个实数值。σ 代表 ReLU 非线性激活函数。总体来看，图注意力网络的聚合过程就是对邻居节点进行带权重求和的过程，权重系数 $\alpha_{i,j}$ 会根据节点特征值的不同而自动发生改变，从而在聚合邻居节点信息时加重对某些节点的信息的关注并忽略那些权重系数很低的节点信息。权重系数 $\alpha_{i,j}$ 的计算过程如下：

$$\alpha_{i,j} = \frac{exp(\sigma(a^T[W_1 x_i \parallel W_1 x_j \parallel W_2 e'_{i,j}]))}{\sum_{t \in N(i)} exp(\sigma(a^T[W_1 x_i \parallel W_1 x_t \parallel W_2 e'_{i,t}]))} \tag{4.5}$$

其中 $W_2 \in R^{D' \times F'}$ 是用来对更新后的边特征 $e'_{i,j}$ 进行线性变换的权重矩阵，我们将线性变换后的 x_i，x_j 和 $e'_{i,j}$ 拼接在一起并使用权重矩阵 $a \in R^{3D'}$ 来计算三者之间的相关性，为了将这个相关性变为概率值，我们先使用 σ 代表的 LeakyRelu 对这个相关系数进行非线性激活，然后再针对一阶范围内的所有邻居节点使用 SoftMax 函数，从计算出两个节点之间的权重系数 $\alpha_{i,j}$，其数值范围为 0~1。值得注意的是，上述我们描述的节点更新过程是针对单头的注意力机制模块，为了提高模型的特征提取能力，我们在模型中同时使用了数量为 4 的多头注意力机制模块，可以并行计算出四种不同的权重系数，从而关注不同特征空间中的更多信息。每一个单头注意力机制是由不同的 W_1, W_2, 和 a 三个参数所决定，最终计算出来的四种更新节点会在特征维度上拼接起来形成最终的结果 $x'_i \in R^{4D'}$。

4.2.2 模型训练

经过连续三层的图注意力网络模块之后，我们会得到更新的边特征和点特征，因为我们进行的整张图的分类任务而不是节点分类任务，所以我们进一步使用全局池化层来对所有的节点特征进行平均池化，并将池化后的特征向量作为整张蛋白质图的表征。之后我们使用两个全连接层构成一个分类器来学习全局平局特征向量和结晶倾向性之间的关系。考虑到结晶倾向性预测是一个二分类问题，我们使用 Sigmoid 函数来将全连接层输出的置信分数进行概率化，代表蛋白质能够成功结晶的概率。

在模型训练方面，我们使用二值交叉熵损失函数[103]来计算训练时的误差：

$$loss(y', y) = \frac{1}{N}\sum_{n=1}^{N} -[y_n \cdot log\, y_n' + (1 - y_n) \cdot log(1 - y_n')] \tag{4.6}$$

其中 N 代表训练时采用的 batch-size 大小，$y_n \in [0,1]$ 是蛋白质样本的真实标签，y_n' 是模型预测的蛋白质结晶倾向性的概率值，真实标签只包含两类，0 代表负样本，表示蛋白无法结晶，1 代表正样本，表示蛋白质能够成功结晶。

下面我们对模型的一些超参数进行简单的说明。首先，对于三层连续的图注意力层网络，更新后的边特征维度 F' 和点特征维度 D' 分别是 4, 4, 4 和 16, 16, 16。然后两个全连接层的神经网络单元数分别是 32 和 1，1 代表的是最终的输出是一个实数值。在模型训练阶段，对应的 batch-size, 学习率和训练迭代次数分别是 64，0.001 和 200。为了避免过拟合，我们使用了衰减因子为 0.001 的L_2正则化，并执行了及早停止的训练手段。

4.3 实验结果与评估

为了验证我们提出的 GCmapCrys 模型的有效性，我们分别在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四个测试数据集上对其进行评估，并与现有的其他结晶倾向性预测模型进行对比。我们还基于 CRYS_DS 数据集重新训练了 DeepCrystal 模型，以此来分析图注意力网络和蛋白质接触图的有效性。最后我们还进一步分析了不同的蛋白质序列特征对于 GCmapCrys 模型的贡献程度，从而得出哪些蛋白质序列特征与结晶倾向性的相关性最高。

4.3.1 模型对比

（1）与单阶段预测模型进行对比

我们首先在 CRYS_DS 测试数据集上与现有的四个单阶段结晶倾向性预测模型 TargetCrys[58]、ParCrys[18]、OB-Score[12]和 ATTCry[13]进行对比。其中 OB-Score 使用 pI 和 Gravy 特征来计算结晶倾向性的置信分数，ParCrys 模型基于氨基酸组成成分特征和 Parzen window 概率密度函数[19]来预测结晶倾向性，TargetCrys 模型则是使用了双层 SVM 模型来帮助预测结晶倾向性。这三种模型都属于机器学习模型，而 ATTCry 模型目前最

先进的基于多尺度卷积神经网络和自注意力机制的端到端结晶倾向性预测模型。因为ATTCry 模型限制蛋白质的最大的长度为 800，所以我们单独从 CRYS_DS 测试数据集上抽取长度小于 800 的蛋白质样本重新构成了一个新的子数据集 CRYS_DS800，并在 CRYS_DS800 测试数据集上对 GCmapCrys 模型和 ATTCry 模型进行对比。下面的表 4.1 和图 4.4 展示了 GCmapCrys 模型在 CRYS_DS 测试数据集上与 TargetCrys, ParCrys、OB-Score 模型的对比结果。

表 4.1 GCmapCrys 与单阶段预测模型的对比结果

模型	Sen	Spe	Acc	MCC	AUC
OB-Score	**0.937**	0.321	0.365	0.153	0.656
ParCrys	0.712	0.516	0.530	0.118	0.615
TargetCrys	0.802	0.399	0.428	0.107	0.608
GCmapCrys	0.550	**0.960**	**0.931**	**0.496**	**0.895**

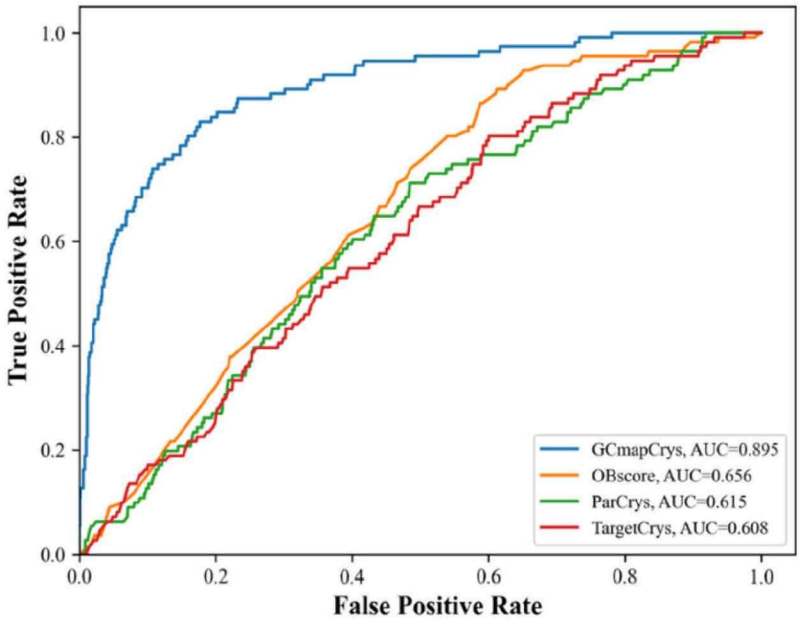


图 4.4 GCmapCrys 与单阶段预测模型在 AUC 指标上的对比结果

从表 4.1 中我们可以发现我们的 GCmapCrys 模型在 Spe、Acc、MCC 和 AUC 指标上都取得了最好的结果。相比于排名第二的 OB-Score 模型，我们的模型在 MCC 获得了 34.3%的提升。同时，如图 4.4 所展示的那样，GCmapCrys 模型在 AUC 上的取值为 0.895，比 TargetCrys、ParCrys、OB-Score 模型分别高了 28.7%、28.0%和 23.9%。而且 GCmapCrys 模型在 Spe 指标上的值为 0.960，Acc 指标上的值为 0.931，相比于其他三个模型平均高了 54.8%和 49.0%。值得注意的是 OB-Score 模型在 Sen 指标上最高的 0.937 但是在 Spe 指标上取得了最低的 0.321，这意味着 OB-Score 模型将很多的负样本预测为正例，从而拉低了整体的 MCC 指标。

下面我们继续在 CRYS_DS800 数据集上对 GCmapCrys 模型和 ATTCry 模型进行对比，表 4.2 展示了详细的对比结果。我们发现 GCmapCrys 模型仍然取得了最好的预测性能。GCmapCrys 模型在Acc、MCC 和 AUC 指标上的取值比 ATTCry 模型分别高了 32.7%、28.3%和 13.6%，这个提升效果仍然是非常显著的，表明了我们 GCmapCrys 模型在单阶段结晶倾向性预测中的拥有卓越的预测性能。

表 4.2 GCmapCrys 与 ATTCry 模型的对比结果

Model	Sen	Spe	Acc	MCC	AUC
ATTCry	**0.974**	0.587	0.602	0.203	0.765
GCmapCrys	0.541	**0.959**	**0.929**	**0.486**	**0.901**

（2）与多阶段预测模型进行对比

表 4.3 GCmapCrys 与多阶段预测模型的对比结果

测试数据集	模型	Sen	Spe	Acc	MCC	AUC
MF_DS	PPCpred	**0.657**	0.537	0.619	0.184	0.628
	fDETECT	0.440	**0.819**	0.531	0.216	0.650
	CrysalisI	0.599	0.631	0.621	0.215	0.639
	CrysalisII	0.609	0.639	0.629	0.232	0.651
	GCmapCrys	0.537	0.794	**0.713**	**0.332**	**0.755**
PF_DS	PPCpred	**0.754**	0.491	0.686	0.231	0.667
	fDETECT	0.413	0.776	0.506	0.171	0.622
	CrysalisI	0.376	0.781	0.677	0.157	0.600
	CrysalisII	0.624	0.661	0.652	0.254	0.655
	GCmapCrys	0.600	**0.840**	**0.778**	**0.432**	**0.817**
CF_DS	PPCpred	0.296	0.917	0.749	0.273	0.654
	fDETECT	0.291	0.883	0.720	0.209	0.594
	CrysalisI	**0.979**	0.073	0.730	0.126	0.499
	CrysalisII	0.055	**1.000**	0.315	0.126	0.527
	GCmapCrys	0.855	0.545	**0.770**	**0.410**	**0.766**
CRYS_DS	PPCpred	0.324	0.876	0.836	0.150	0.669
	fDETECT	0.649	0.727	0.721	0.211	0.718
	CrysalisI	0.667	0.673	0.672	0.184	0.705
	CrysalisII	**0.685**	0.647	0.650	0.177	0.712
	GCmapCrys	0.550	**0.960**	**0.931**	**0.496**	**0.895**

因为我们的 GCmapCrys 是一个多阶段的结晶倾向性预测器，还可以预测蛋白质材

料生产失败、纯化失败、晶体生产失败的概率，所以我们分别在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四个测试数据集上与 PPCpred[32]、fDETECT[44]、CrysalisI 和 CrysalisII 四个多阶段预测模型进行对比，其中 CrysalisI 和 CrysalisII 是 Crysalis[38]的两个不同版本。具体来说，我们在对应的四个训练数据集上分别训练 GCmapCrys 模型从而得到四个单独的 GCmapCrys 子模型，分别预测蛋白质材料生产失败、纯化失败、晶体生产失败和结晶成功的概率，再与其他四个多阶段预测模型在四个测试数据集上进行对比，这四种其他模型的对比结果都是从对应的 Web 服务器上获取得到的。表 4.3 显示了 GCmapCrys 模型与四种多阶段预测模型在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四个测试数据集上的对比结果。

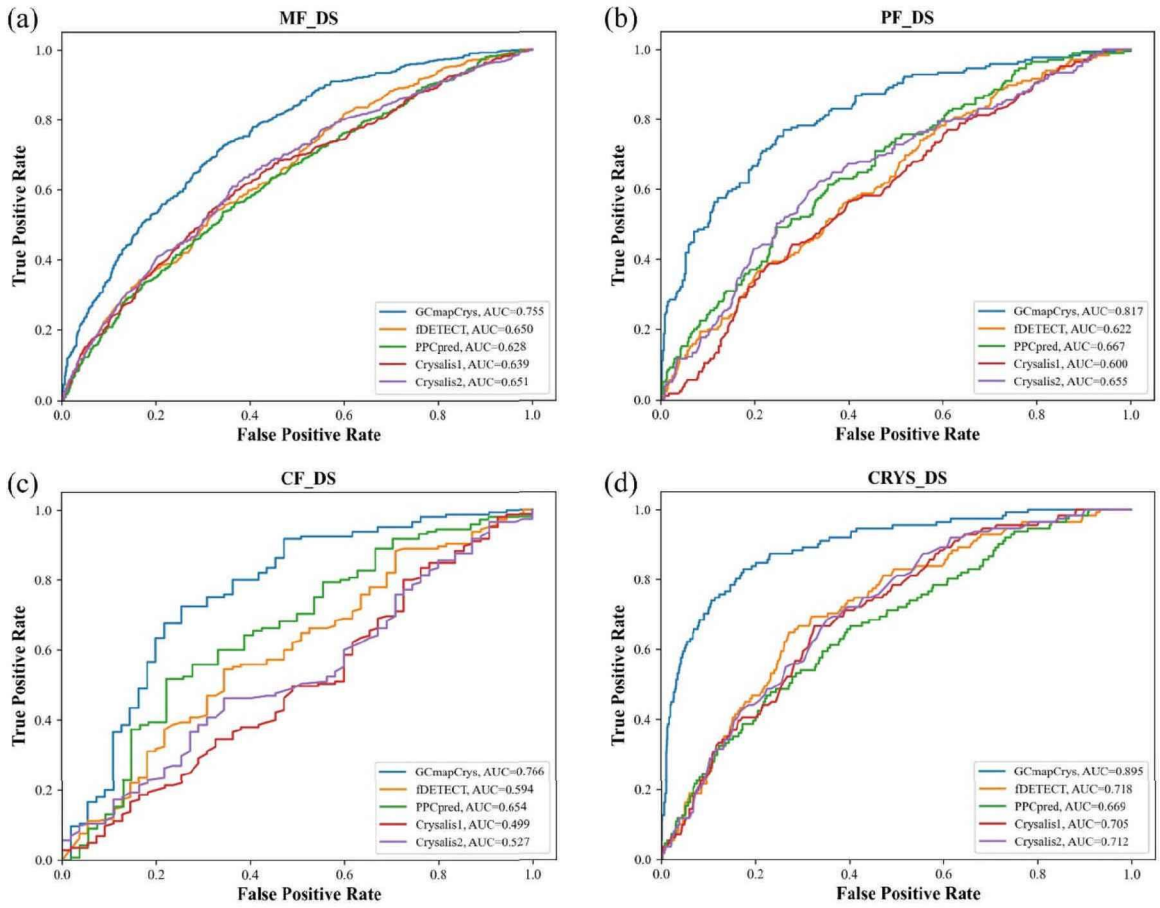


图 4.5 GCmapCrys 与其他多阶段方法在 AUC 指标上的对比结果

从表 4.3 中我们可以发现我们的 GCmapCrys 模型在四个测试数据集上都取得了最高的 Acc、MCC 和 AUC 指标值。具体来说，GCmapCrys 模型在四个测试数据集上的 MCC 值相比于其他四种模型分别平均提高了 12%、22.9%、22.7%和 31.6%。同时，通过观察图 4.5 显示的五种模型在四个测试数据集下的 ROC 曲线下面积，我们发现我们的 GCmapCrys 模型在四种测试数据集中都取得了最高的 AUC 值，而且提高的效果都很显著。这表明了我们模型在所有的结晶阶段中都有着最优的预测性能。而且我们观察到

GCmapCrys 在 PF_DS 和 CRYS_DS 测试数据集上取得了最高的 Spe 值，分别为 0.840 和 0.960，表示了我们的模型在预测纯化步骤和整个结晶流程时对负样本有较高的预测能力。值得注意的是，CrysalisI 模型在 CF_DS 测试数据集上获得最高的 Sen 值 0.979 和最低的 Spe 值 0.073，同时 CrysalisII 模型在 CF_DS 测试数据集上获得最低的 Sen 值 0.055 和最高的 Spe 值 1.000，这导致两个模型在 CF_DS 测试数据集上的 MCC 指标值都比较低。

（3）验证图注意力网络和蛋白质接触图的有效性

上述关于单阶段和多阶段的结晶倾向性预测实验表明了我们的 GCmapCrys 模型的确拥有最优的预测性能，这主要是由于以下两个因素，首先是我们采用图注意力网络来提取蛋白质接触图中的空间结构信息，相比于蛋白质一级序列信息，蛋白质接触图提供了一定程度的空间信息来辅助预测结晶倾向性，而且我们设计的图注意力网络相比于普通的 CNN 模型能够更有效的提取空间中氨基酸的相互作用信息；第二个因素是我们融合了多种互补的蛋白质序列特征，包括氨基酸序列编码、位置特异性矩阵、预测的二级结构和相对溶剂可及性、氨基酸的物理化学属性等。这些特征已经被很多研究证明是有助于蛋白质结晶倾向性预测的。为了进一步对这两种因素进行验证，我们将 GCmapCrys 模型和 DeepCrystal 模型在 CRYS_DS800 测试数据集上进行对比。DeepCrystal 模型是一个单阶段预测模型，也是首个使用卷积神经网络来预测蛋白质结晶倾向性的模型，它仅依赖于蛋白质序列对应的氨基酸编码作为模型输入，再使用多层的卷积神经网络来提取与结晶倾向性相关的特征，属于直接从氨基酸序列预测蛋白质结晶倾向性的端到端模型。为了与 DeepCrystal 模型进行对比，我们进行了两组实验对比，第一组实验中两种模型都只使用蛋白质序列氨基酸编码特征作为模型输入，不包含额外的特征，第二组实验中两种模型都使用包含 AACD、pI、Gravy、Length、PSSM、PSS、PRSA、AAindex 在内的全部八种特征作为模型输入，这两组实验对应的命名分别为 FR_I 和 FR_II。值得注意的是，DeepCrystal 原始模型是不包含额外的蛋白质序列特征的，我们按照第 4.1 章节中构造节点特征的方式为 DeepCrystal 模型构造同样的特征输入，并在 CRYS_DS800 训练数据集上重新训练后再在对应的测试数据集上获取预测结果。表 4.4 总结了两种模型在 FR_I 和 FR_II 两组实验上的对比结果。

表 4.4 GCmapCrys 与 DeepCrystal 模型在不同特征输入条件的对比结果

特征类型	Model	Sen	Spe	Acc	MCC	AUC
FR_I[a]	DeepCrystal	**0.523**	0.844	0.821	0.246	0.780
	GCmapCrys	0.468	**0.879**	**0.850**	**0.255**	**0.807**
FR_II[b]	DeepCrystal	0.369	**0.972**	0.928	0.395	0.870
	GCmapCrys	**0.505**	0.964	**0.931**	**0.477**	**0.875**

从表 4.4 中我们可以发现 GCmapCrys 模型在 Acc、MCC 和 AUC 指标上都取得了最好的结果。对于 FR_I 实验而言，GCmapCrys 模型取得了 0.255 的 MCC 和 0.807 的 AUC，

比 DeepCrystal 模型高了 0.9%和 2.7%。同时在 FR_II 实验上 GCmapCrys 模型在 MCC 和 AUC 指标上也比 DeepCrystal 模型高了 8.2%和 0.5%。这两个实验结果表明了在同样的条件下 GCmapCrys 模型要优于 DeepCrystal 模型，也就证明了使用图注意力网络和蛋白质接触图在结晶倾向性上的预测效果是要优于DeepCrystal模型所使用的普通卷积模型。而且这两个模型在添加了蛋白质序列特征后在预测精度上都取得了很大的提升，表明这些互补的蛋白质序列特征对于预测蛋白质结晶倾向性是非常有帮助的。

4.3.2 特征消融实验

为了进一步分析不同的蛋白质序列特征在 GCmapCrys 模型中对于结晶倾向性预测的贡献，我们对这些特征进行了消融实验。我们首先将我们使用的特征分为四组：氨基酸编码（AACD）、位置特异性矩阵（PSSM）、预测的结构信息编码（PSBC）和物理化学属性（PCP），其中 PSBC 包括预测的二级结构和预测的相对溶剂可及性，PCP 包括等电离点、平均疏水性、AAindex 和序列长度。而且我们将所有的这些特征组合在一起时命名为 APSC。为了验证这四组特征对模型的贡献，我们从 APSC 中分别去除 AACD、PSSM、PSBC 和 PCP 从而形成 PSC、ASC、APC 和 APS 四种类型的特征组合，我们将包含 APSC 在内的这五种特征组合作为 GCmapCrys 模型的输入并在 CRYS_DS 数据集上进行训练和测试，图 4.6 显示了这五种特征组合在 CRYS_DS 测试集上的实验结果。

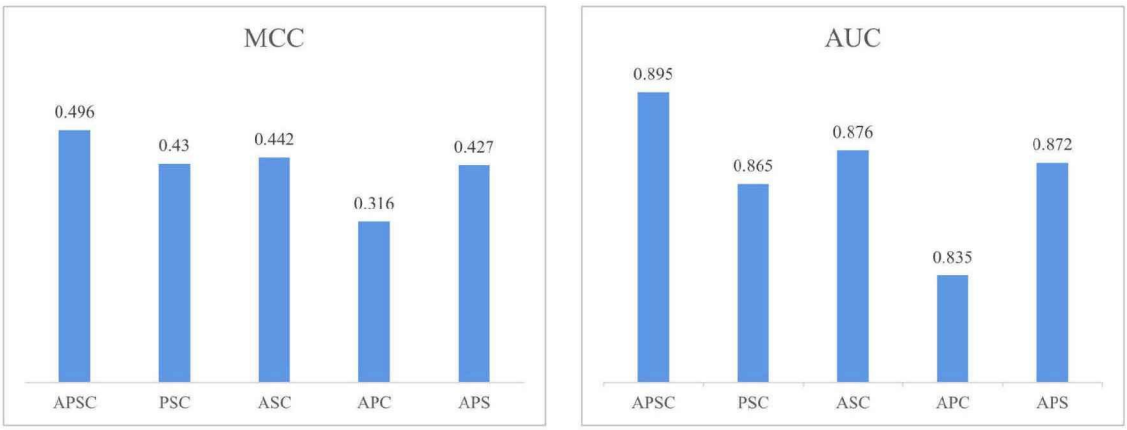


图 4.6 GCmapCrys 在五种特征组合下的预测结果

从图 4.6 中我们可以发现相比于 APSC 组合特征，PSC, ASC, APC 和 APS 对应的 MCC 指标分别下降了 6.6%, 5.4%, 18%和 6.9%，AUC 指标分别下降了 3%, 1.9%, 6%, 和 2.3%。从中我们可以发现分别去除 AACD、PSSM、PSBC 和 PCP 特征组合后，模型的整体性能都有所下降，表明了这四种特征组合都有助于蛋白质结晶倾向性预测。另一方面，去除 PSBC 特征组合后模型的预测性能下降最多，表明了预测的二级结构和相对溶剂可及性对于模型预测性能的贡献是最高的，其次分别是 AACD、PSSM 和 PCP。

4.4 本章小结

在本章我们提出了一种新的深度学习方法，GCmapCrys，通过整合图注意力神经网络和预测的蛋白质接触图来实现多阶段的蛋白质结晶倾向性预测。在 MF_DS、PF_DS、CF_DS 和 CRYS_DS 四种数据集上的实验结果表明，我们提出的 GCmapCrys 取得了最好的预测性能，优于最先进的单阶段和多阶段预测模型。我们对实验结果的分析表明 GCmapCrys 的优势主要归功于两个方面：首先，图注意力网络与预测的蛋白质接触图可以有效地捕捉与结晶相关的氨基酸相互作用信息。最后是多种互补的蛋白质序列特征可以进一步辅助结晶预测，从而提高结晶倾向性预测的准确率。

5 基于 Graph Transformer 的蛋白质结晶倾向性预测

相比之前的结晶倾向性预测方法，我们第 4 章提出的 GCmapCrys 模型基于蛋白质接触图和图注意力网络有效地提取了蛋白质序列中远距离的氨基酸相互作用信息。但是我们无法叠加过多的图卷积层来扩大卷积过程中的感受野，这就导致我们只能学习蛋白质接触图的小范围局部结构信息。我们在第 3 章设计 CCmap-KAAP 特征的目的是探索三维空间结构中 K 间隔氨基酸对组成成分特征对蛋白质结晶倾向性的影响，其中 K 间隔代表的就是蛋白质接触图中的 $K+1$ 阶范围邻居，从图 3.5 中我们可以看出 $K=3$（4 阶邻居）、4（5 阶邻居）时 CCmap-KAAP 特征仍然能够帮助预测蛋白质结晶倾向性，所以我们希望通过深度学习技术提取更大范围的空间结构信息。

ATTCry[13]是目前前沿的端到端结晶倾向性预测方法，在没有空间先验信息的辅助下，他使用多头自注意力网络[57]提取全局的远距离相互作用信息。自注意力机制最早应用在自然语言处理方面，能够处理序列中上下文的长距离依赖关系，由于其卓越性能表现，已经被广泛地应用在图像、语音多个领域[104, 105]。其中 Transformer[57]架构使用注意力机制在多个领域取得了进一步的成功，而且许多研究表明 Transformer 在处理图结构数据时有非常大的潜力[65, 106, 107]，所以在本章节中，我们结合图卷积模型和 Transformer 架构来处理蛋白质接触图中的氨基酸的长距离相互作用信息。

有许多方法已经尝试将 Transformer 和图卷积进行结合，比如 GROVER[107]、Graph-BERT[106]和 Graphormer[108]等。比较常用的方式是先使用图卷积神经网络提取局部结构信息，再将所有更新后的节点表征作为一个序列，并使用 Transformer 来计算这个序列中所有节点之间的相互作用信息，从而增强全局的表征能力[65]。在本章节我们使用同样的方式将图注意力层和 Transformer 层结合起来，首先使用第 4 章中构建的图注意力网络来提取接触图的局部结构特征，然后使用 Transformer 的编码层来提取局部结构信息之间的相互作用关系。为了更加充分地利用局部结构特征，我们还使用残差连接将图注意力网络的所有中间层输出合并在一起作为 Transformer 层的输入。Transformer 虽然具有强大的上下文表征能力，但是它抛弃了序列的位置信息，也即不论我们如何调整序列中元素的相对位置关系，都不影响 Transformer 层的输出。这也是我们无法直接将 Transformer 应用到原始的图结构数据中的原因，因为将最初的所有节点铺平成序列之后就丧失了图的拓扑结构信息。由于蛋白质序列和蛋白质接触图的关系是紧密相连的，接触图中的每一个氨基酸节点或局部结构信息都应该有相应的位置信息，所以为了弥补这个缺陷，我们将蛋白质序列的位置信息进行编码并与图注意力网络输出的节点进行结合，从而使得每一个氨基酸节点和其对应的结构信息都包含相应的位置信息。我们将最终构建的模型称为 GCmapTCrys，它是一个多阶段的结晶倾向性预测模型，在蛋白质接触图

作为输入的基础上充分利用了图卷积网络和自注意力机制来同时提取局部和全局的结构信息，从而更有效的进行结晶倾向性预测。在 MF_DS800、PF_DS800、CF_DS800 和 CRYS_DS800 四种测试集上的实验结果表明我们的 GCmapTCrys 多阶段结晶倾向性预测模型取得了最优异的结果。

5.1 Graph Transformer

2017 年 google 提出的 Transformer[57]架构在机器翻译 seq2seq 任务中取得了 SOTA 的成绩，它替代了自然语言处理任务中传统的 RNN[109]循环神经单元网络和 CNN[49]卷积神经网络，使用自注意力机制作为基础计算单元来处理词语之间的上下文依赖关系。传统的 RNN 循环神经网络在处理过长的序列信息时很容易丢失掉之前学习的信息，导致难以应对长文本或高分辨率图像的场景。而自注意力机制在计算两个词语的关系时考虑了上下文背景中的所有词语，从而避免了信息丢失的问题。Transformer 架构的另一个优势是对应的自注意力计算单元并行程度很高，能够大大加快训练和测试的时间。由于 Transformer 的优异表现，它已经是 NLP 领域最重要的成果之一，而且随着该框架的持续发展，它已经在图像、语音、生物信息等多个领域取得了突破性的进展。

图作为一种特殊的数据结构，大致可以分为节点分类、图聚类、图分类等任务。目前图数据与深度学习结合最紧密的领域是图卷积神经网络结构，利用信息在节点和边之间的传递对节点进行迭代地聚合，从而达到类似卷积的效果来提取图的局部结构特征。但是图卷积网络也有很大的缺陷，对于普通的图卷积网络 GCN，频繁的对节点特征进行聚合会导致整张图出现过平滑的现象[110]，过深的图卷积网络结构也会不断压缩远距离邻居节点的信息从而出现过渡挤压现象[111]，这两个问题导致我们不能简单地通过叠加深层网络来提取更好的特征。对于图注意力神经网络而言，自注意力机制虽然一定程度上缓解了过平滑问题，但是仍然不能堆叠过深的网络[112]。而 Transformer 的出现为解决图结构数据的问题提供了另外一种方向，但是 Transformer 只能直接作用到序列信息上，如果将图的所有节点铺平为一个序列，那么这个序列就丧失了图的拓扑结构信息，所以将 Transformer 和图进行结合的一个难点就是如何在保留图结构信息的同时对图进行序列化。

目前已经有多种将 Transformer 和图结合的模型能够解决这类问题，根据 Min 等人[65]的研究，我们可以将这些模型分为三类：使用 GNN 作为辅助模型；将图编码为具有位置信息的序列；将图的先验信息融合到自注意力机制中。第一种方法使用 GNN[63]模型来提取每个节点对应的局部子图结构信息，然后在将这些带有局部结构信息的节点铺平进行序列化，从而在保留了局部拓扑结构的同时将图数据进行了序列化，比如 GraphTrans[113]和 GROVER 模型[107]。第二种方法是为序列中的节点赋予位置信息，这个位置信息就是图拓扑结构的一种映射，比较常用的位置编码方式是根据邻接矩阵计算拉

普拉斯特征值或 SVD 向量。第三种方法并不是将图数据进行序列化，而是通过各种方式将图的先验结构信息融合到自注意力机制的计算过程中，可以看作对自注意力机制的一种变形，比如图注意力网络[61]中的局部自注意力机制，只对节点周围的邻居节点进行自注意力计算，而不是计算所有节点的相关系数。

在本章中我们主要利用了第一种方法即将图卷积网络作为 Transformer 的辅助模型，但是在整体模型架构中为了提高模型的预测能力，我们也对图节点赋予了位置信息，同时使用图注意力神经网络中的局部自注意力机制增强了提取局部结构信息的能力，因此从这个角度来看，我们的 GCmapTCrys 模型同时融合了这三类方法，从而提高了模型的整体性能。

5.2 GCmapTCrys模型架构

我们提出一个新的图 Transformer 架构 GCmapTCrys 来预测蛋白质结晶倾向性，该模型以蛋白质接触图转化而来的图 G 作为输入，然后使用 GCmapCrys 模型中的图注意力层来提取图 G 的局部子图结构，该注意力层包含边更新和点更新两个子模块，其中点更新的过程利用了局部自注意力机制和更新后的边信息来自适应地聚合中心节点和邻居节点的信息，提高了局部特征的表征能力，为了便于表示，我们将一个完整的图注意力层命名为 GAT Block，更详细的介绍见第 4.2 章节。经过 K 个连续的 GAT Block 层之后，我们使用一个残差连接将所有 GAT Block 层的输出节点结合起来送入到 Transformer 编码层中。为了使 Transformer 对蛋白质序列的氨基酸位置信息敏感，我们对蛋白质序列进行了位置编码并添加到 Transformer 的输入中。然后我们使用五个连续的 Transformer 编码层来提取蛋白质图中的远距离空间相互作用信息，再使用一个全局池化层和三个全连接层执行蛋白质结晶倾向性预测任务，整体的工作流程见图 5.1。

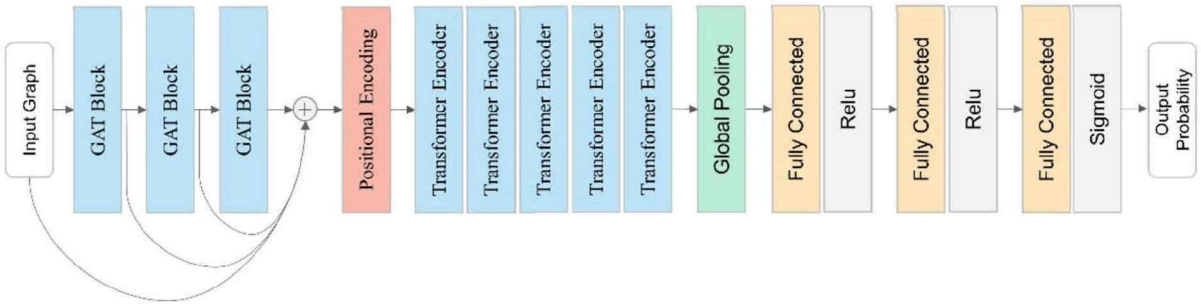


图 5.1 GCmapTCrys 整体模型架构

5.2.1 GAT Block

对于输入的蛋白质图 $G = \{X, E\}$，我们将其送入到 K 个连续的 GAT Block 层中，

K 是一个超参数，我们选取 $K=3$。经过 3 次图注意力层的迭代之后，我们可以得到更新后的蛋白质图，具体流程如下：

$$X^k, E^k = F^k_{GAT\text{-}Block}(X^{k-1}, E^{k-1}) \tag{5.1}$$

$$X' = LayerNormal\left(\sum_{k=1}^{K}\sigma(W^k X^k + b^k) + \sigma(WX+b)\right) \tag{5.2}$$

其中 $F^k_{GAT\text{-}Block}$ 代表第 k 层 GAT Block，对前一层的输入 X^{k-1}, E^{k-1} 进行更新并得到当前层的节点更新值 X^k 和边更新值 E^k。公式 5.2 表示一个残差连接层，其中 W^k 和 W 是权重矩阵，用来对输入进行线性变换，b^k 和 b 是偏置，σ 代表 ReLU 非线性激活函数，$LayerNormal$ 是层间归一化函数[114]，不同于 $BatchNormal$ 批归一化函数对不同样本的同一个通道做特征归一化，它是对同一个样本的不同通道做特征归一化。这个残差层将所有中间层更新的节点输出 X^k 和原始输入 X 结合起来形成 $X' \in R^{L\times d}$，其优势在于提取的局部结构信息叠加在一起之后会减少逐层损失和过渡平滑的现象，而且也能够使得 Transformer 输入层能够接收多种不同范围的局部结构信息，使得铺平之后的节点序列能够保留更多的拓扑结构信息。

5.2.2 序列位置信息编码

位置信息是一种特别的结构信息，它能够表示节点之间的先后顺序关系。对于蛋白质图 G 而言，每个节点没有绝对的空间位置信息，只有节点之间相对的位置关系，这种相对位置关系就体现了整张图的拓扑结构。但是当把图的节点铺平为一个序列时，节点之间的相对位置关系就会消失，所以使用图卷积神经网络提取每个节点对应的局部结构信息能够保留一定程度的相对位置关系。但是除了这种空间结构上的相对位置关系，蛋白质序列还拥有序列上的位置关系，按照蛋白质序列决定结构、结构决定功能的生物理论，这种序列上的顺序关系也是至关重要的。所以我们对蛋白质的序列位置信息进行编码，并将其整合到 Transformer 中来增强输入。

因为蛋白质序列与自然语言中的文本有非常相似的结构，文本中的词语对应蛋白质中的氨基酸，所以我们按照 Transformer 为文本赋予词序的方式对蛋白质序列的位置信息进行编码，对于长度为 L 的蛋白质序列，我们用 $P=[p_1, p_2, \ldots, p_L]$ 表示这个序列对应的位置向量，$p_l \in R^d$，d 是对应氨基酸节点的特征维度。

$$p_l^{(i)} = \begin{cases} sin\left(\dfrac{l}{10000^{2n/d}}\right), & i = 2n \\ cos\left(\dfrac{l}{10000^{2n/d}}\right), & i = 2n+1 \end{cases} \tag{5.3}$$

$$S = X' + P \tag{5.4}$$

计算出每个位置对应的编码向量之后，我们将其与 GAT Block 层残差连接之后的输

出 X' 相加得到 Transformer 的输入 S，两者的特征维度相同，$P \in R^{L\times d}$，$X' \in R^{L\times d}$。这样就能够将序列位置信息注入到模型的输入中，使得模型具备了学习氨基酸序列位置信息的能力。

5.2.3 Transformer 编码层

从模型架构上来说，Transformer 是一个基于注意力机制的编码器-解码器架构，编码器通过多层自注意力机制对序列中的词语进行编码，而解码器使用带有掩码的多层自注意力机制预测某一时刻的输出。因为我们是二分类的整图预测任务，不是 seq2seq 类型的任务，所以只需要使用编码层来提取氨基酸节点的全局相互作用信息。

在经过图注意力层 GAT Block 与序列位置编码层之后，我们得到了新的氨基酸节点表征 $S \in R^{L\times d}$，S 即包含了蛋白质接触图的局部子图结构信息，还包含了氨基酸序列的位置信息，然后我们使用 Transformer 编码层处理输入 S 来更好地捕捉氨基酸节点之间的相互作用信息，第一层 Transformer 编码器的输出 Z 可以用公式表示如下：

$$M = LayerNormal_1\left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V + S\right) \tag{5.5}$$

$$\mathrm{F} = FFN(M) = \sigma\left(MW_{f1} + b_{f1}\right)W_{f2} + b_{f2} \tag{5.6}$$

$$Z = LayerNormal_2(F + M) \tag{5.7}$$

其中Q、K、V 分别对应自注意力机制中的 query（$Q = SW_Q$）、key（$K = SW_K$）和 value（$V = SW_V$），$W_Q, W_K, W_V \in R^{d\times d_k}$ 是线性映射对应的权重矩阵，d_k是特征 K 的维度，主要作用是进行归一化。然后自注意力机制计算出来的结果会和输入 S 进行残差连接，并使用 $LayerNormal$进行层间归一化得到中间输出 M。FFN 代表对每一个位置进行前馈神经网络（MLP）计算，$W_{f1} \in R^{d_k\times d_f}, W_{f2} \in R^{d_f\times d_k}$是线性映射对应的权重矩阵，$b_{f1}, b_{f2}$ 是偏差，σ 代表 ReLU 非线性激活函数，d_f 是隐藏层的神经元个数。最后将 FFN 的输出 F 与自注意力的输出 M 进行残差连接，并使用 $LayerNormal$ 归一化得到最终的输出 Z。总体来看一层 Transformer 编码层包含两个子层，第一个子层使用多头自注意力机制计算上下文的相互作用，第二个子层使用前馈神经网络进行更新，每个子层都会使用残差连接和 $LayerNormal$ 更新输出，其中多头自注意力能够使得模型具有更强的表征能力，我们用 h 表示多头自注意力的数量，每个自注意力机制的输出被拼接在一起形成 $h \times d_k$ 大小的特征维度。

如图 5.1，我们使用了五个连续的 Transformer 编码层来进行特征提取，最终得到更新后的节点特征。因为每个蛋白质序列的长度不同，所以我们对所有的节点使用全局平均池化（Global Pooling）来得到整张蛋白质图的表征。然后我们将整张图的表征向量送入到三个全连接层（Fully Connected）中进行分类，其中最后一个全连接层是输出层，

用来计算蛋白质结晶倾向性的置信分数，最后再使用 Sigmoid 函数来将置信分数转换为输出的结晶倾向性概率。在模型训练方面，与第 4 章相同，我们使用二值交叉熵损失函数[103]来计算训练时的误差。在模型超参数方面，我们总结在了表 5.1 中，模型的总参数量约为 1.4MB。

表 5.1 GCmapTCrys 模型超参数

GAT Block 层数（K）	3
GAT Block 输出节点维度（d）	64
多头自注意力数量（h）	4
自注意力隐藏层维度（d_k）	16
前馈神经网络隐藏层维度（d_f）	100
三层全连接层维度	[64, 32, 1]
Batch-size	32
学习率（lr）	0.001
L2 正则化权重衰减系数	0.001

5.3 实验结果与评估

为了验证 GCmapTCrys 模型的有效性，我们首先与其他单阶段和多阶段的结晶倾向性预测模型进行对比，然后再通过消融实验验证 GAT Block 中残差连接和序列位置信息编码的有效性。因为我们的 GCmapTCrys 模型仍然使用了与第 4 章相同的多种蛋白质特征来辅助结晶倾向性预测，而且在第 4.3.2 章节中我们发现预测的二级结构和相对溶剂可及性特征对结晶倾向性的影响最大，所以我们又单独对这两种特征进行了分析。

5.3.1 模型对比

（1）单阶段预测模型对比

我们与三种单阶段预测模型进行了对比，分别是 DeepCrystal[15]，ATTCry[13]和 BCrystal[59]。其中 DeepCrystal 和 ATTCry 是端到端的单阶段结晶倾向性预测模型，分别利用卷积神经网络和自注意力机制来提取蛋白质序列中氨基酸的全局相互作用信息。而 BCrystal 模型将多种蛋白质序列特征作为输入，并使用机器学习模型 XGboost 来进行单阶段预测。由于我们自己的模型 GCmapTCrys 在处理较长序列的时候容易超出设备的显存限制，而且 DeepCrystal 和 ATTCry 模型也都限制蛋白质序列的长度不超过 800，所以我们在 CRYS_DS800 测试集上对这四种模型进行了对比。具体对比结果见表 5.2。

从表 5.2 中我们可以发现 BCrystal 模型取得了最好的预测结果，而我们的 GCmapTCrys 模型取得了排名第二的结果，相比较于 DeepCrystal 和 ATTCry 模型，GCmapTCrys 在 Acc、MCC、AUC 指标上平均提升了 22.5%、28.1%和 11.3%。但是

BCrystal 模型的测试结果要远远高于 GCmapTCrys，通过分析 BCrystal 模型所使用的蛋白质序列特征，我们发现对 BCrystal 模型影响最大的两种特征是预测的二级结构和相对溶剂可及性特征，与我们在第 4.3.2 章节进行特征消融实验得出的结果一致。因此我们进一步研究了这两种特征与结晶倾向性之间的相关性。

表 5.2 GCmapTCrys 与三种单阶段模型在 CRYS_DS800 测试集上的对比结果

模型	Sen	Spe	Acc	MCC	AUC
DeepCrystal	0.523	0.844	0.821	0.246	0.780
ATTCry	**0.974**	0.587	0.602	0.203	0.765
BCrystal	0.901	0.954	**0.951**	**0.715**	**0.976**
GCmapTCrys	0.523	**0.968**	0.936	0.505	0.885

（2）预测的二级结构、相对溶剂可及性特征与结晶倾向性的相关性分析

与 BCrystal 相同，我们同样使用了 SCRATCH 工具来预测蛋白质的二级结构和相对溶剂可及性。SCRATCH 输出的二级结构和相对溶剂可及性包含两种结果，一种是基于同源分析方式得到的结果，另一种是基于序列从头预测得到的结果。基于同源分析的方法是将待预测的序列与数据库中已知二级结构和相对溶剂性的序列进行相似性比较，利用打分矩阵计算出相似性得分，再根据相似性得分构建出待预测片段的二级结构和相对溶剂可及性。因为我们的数据集中的大部分蛋白质序列已经被收录在数据库中，就导致使用同源分析方法计算出来的二级结构和相对溶剂可及性非常接近于真实的数据。相反使用从头预测的方法预测出来的二级结构和相对溶剂可及性的准确率就会比较差。所以高准确率的二级结构和相对溶剂可及性特征可能是 BCrystal 模型在结晶倾向性上取得高准确率的关键原因，为了验证这个猜想，我们使用基于同源分析得到的二级结构和相对溶剂可及性特征来替换之前使用的这两种特征，得到新的模型 GCmapTCrys_hom，然后我们在 CRYS_DS800 测试数据集上对新的模型进行验证。

表 5.3 GCmapTCrys_hom 模型在 CRYS_DS800 测试集上的测试结果

模型	Sen	Spe	Acc	MCC	AUC
BCrystal	**0.901**	0.954	0.951	0.715	0.976
GCmapTCrys	0.523	0.968	0.936	0.505	0.885
GCmapTCrys_hom	0.829	**0.978**	**0.978**	**0.770**	**0.978**

表 5.3 显示了 GCmapTCrys_hom 模型在 CRYS_DS800 测试集上的测试结果，从表中我们可以发现当使用了更高准确率的二级结构和相对溶剂可及性特征时，我们的 GCmapTCrys_hom 模型获得了非常高的预测结果，相比于 GCmapTCrys 模型在 MCC 指标上提高了 26.5%，相比于 BCrystal 模型提高了 5.5%。所以综合表 5.2 与 5.3，我们可以得出两个结论；一是我们构建的基于 Graph Transformer 和蛋白质接触图的模型框架对蛋白质结晶倾向性预测是非常有效的；二是高准确率的二级结构和相对溶剂可及性特

征能非常显著的提升结晶预测的准确率，表明了二级结构和相对溶剂可及性与蛋白质结晶拥有非常紧密的关系。

虽然 GCmapTCrys_hom 模型拥有非常高的结晶倾向性预测结果，但是我们还是需要选取 GCmapTCrys 模型作为最后的预测模型。因为在真实的结晶预测场景中，许多待查询蛋白质无法在数据库中匹配到高度同源的蛋白质序列，从而导致预测的二级结构和相对溶剂可及性特征非常差，所以我们仍然采用基于从头预测的方式获取这两种特征，从而保证训练数据和测试数据的一致性。

（3）多阶段预测模型对比

我们的 GCmapTCrys 模型也是一个多阶段的结晶倾向性预测模型，除了预测蛋白质能否通过整个结晶过程，还能预测蛋白质材料生产、纯化和结晶产生三个步骤各自成功的概率。为了验证该模型的预测性能，我们在 CRYS_DS800、MF_DS800、PF_DS800、CF_DS800 四种数据集上与第 4 章节的 GCmapCrys 模型进行对比，表 5.4 是具体的对比结果。

表 5.4 GCmapTCrys 与 GCmapCrys 模型的对比结果

测试数据集	模型	Sen	Spe	Acc	MCC	AUC
MF_DS800	GCmapCrys	0.520	**0.789**	**0.704**	0.310	0.715
	GCmapTCrys	**0.607**	0.731	0.692	**0.324**	**0.719**
PF_DS800	GCmapCrys	**0.650**	0.813	0.771	0.439	0.785
	GCmapTCrys	0.540	**0.883**	**0.795**	**0.442**	**0.792**
CF_DS800	GCmapCrys	0.841	**0.574**	0.769	0.415	**0.720**
	GCmapTCrys	**0.890**	0.500	**0.784**	**0.421**	0.706
CRYS_DS800	GCmapCrys	0.505	0.964	0.931	0.477	0.875
	GCmapTCrys	**0.523**	**0.968**	**0.936**	**0.505**	**0.885**

从表 5.4 中我们可以发现相比于 GCmapCrys 模型，GCmapTCrys 模型在四种测试数据集上都有更好的预测性能。在 CRYS_DS800 测试集上，GCmapTCrys 模型取得了 0.523（Sen）、0.968（Spe）、0.936（Acc）、0.505（MCC）和 0.885（AUC）的最优结果，在这五种指标上比 GCmapCrys 模型分别提高了 1.8%、0.4%、0.5%、2.8%和 1.0%。在 MF_DS800 和 PF_DS800 测试集上，GCmapTCrys 模型在 MCC 和 AUC 指标上也都取得了最好的预测结果，比 GCmapCrys 模型平均提升了 0.85%和 0.55%。在 CF_DS800 测试集上，GCmapTCrys 模型在 Acc 和 MCC 上取得了最好的结果。从这四种测试集上的对比结果来看，GCmapTCrys 在总体上要优于 GCmapCrys 模型，表明了使用 Graph Transformer 模型架构来捕捉蛋白质接触图的全局氨基酸作用信息是非常有效的。

5.3.2 消融实验

在 Graph Transformer 的模型架构下，我们使用残差连接（skip connect）和氨基酸序列位置编码（positional encoding）来提高模型的预测能力。下面我们在 CRYS_DS800 测试集上使用消融实验来验证两者的有效性。

表 5.5 GCmapTCrys 模型消融实验

Skip connect	Positional encoding	Sen	Spe	Acc	MCC	AUC
		0.477	0.968	0.932	0.469	0.879
√		0.505	0.970	0.936	0.500	0.884
	√	0.423	**0.978**	**0.938**	0.474	0.883
√	√	**0.523**	0.968	0.936	**0.505**	**0.885**

如表 5.5 所示，我们在 GCmapTCrys 模型的基础上分别去掉残差连接和序列位置编码两个子模块并测试其在 CRYS_DS800 测试集的预测结果。我们发现去掉残差连接之后，模型的 MCC 指标和 AUC 指标分别下降了 3.1%和 0.2%，去除序列位置编码后分别下降了 0.5%和 0.1，同时去除两者之后，分别下降了 3.6%和 0.6%。这个结果表明了残差连接和序列位置编码对我们的预测模型都有一定的帮助，而且残差连接能够带来的性能提升更大。最主要的原因是残差连接能够尽可能的保留图注意力网络提取的子图结构信息，从而让 Transformer 编码层更有效地提取蛋白质接触图中的氨基酸全局相互作用信息。

5.4 本章小结

在本章节中我们设计了一个新的多阶段结晶倾向性预测模型 GCmapTCrys。借助预测的蛋白质接触图，我们使用 Graph Transformer 的框架来提取全局的氨基酸相互作用信息，并利用该信息进行蛋白质结晶倾向性预测。我们首先使用第 4 章节中的图注意力网络层来提取蛋白质接触图的子图结构信息，并利用残差连接层来保留更全面的子图结构信息；然后我们将更新后的蛋白质图依照节点铺平为一个序列，并对其进行位置编码；最后使用 Transformer 编码层来提取全局的氨基酸节点相互作用信息，并预测相应的结晶倾向性。通过与其他单阶段和多阶段的预测模型进行对比，我们发现 GCmapTCrys 模型拥有最好的预测性能。而且我们发现当使用更高预测精度的二级结构和相对溶剂可及性特征时，能够极大地提升模型在结晶倾向性上的预测性能，表明了这两种特征与蛋白质结晶的关系是非常紧密的。最后我们对残差连接和序列位置编码进行消融实验，实验结果表明这两个子模块都能够有效地提升模型的预测性能。

6 总结与展望

在本文中我们主要利用预测的蛋白质接触图来预测蛋白质结晶倾向性。蛋白质接触图是三维结构的简化，但仍然包含着丰富的残基间接触信息，我们的主要研究方向就是使用合适的方法来提取接触图中的氨基酸相互作用信息，并利用该信息进行蛋白质结晶倾向性预测。

在第 3 章节中我们首先从蛋白质接触图中提取了一种新的氨基酸组成成分特征 CCmap-KAAP，该特征表示空间结构中 K 间隔下的氨基酸对频率。相比于基于序列提取的 K 间隔氨基酸对频率特征 KAAP，CCmap-KAAP 特征包含了丰富的结构信息。最终的实验结果表明基于空间结构的 CCmap-KAAP 特征和基于序列的 KAAP 是互补的，将这两者整合在一起之后能够极大的提升蛋白质结晶倾向性预测的能力。然后我们还将 CCmap-KAAP 和其他多种互补的蛋白质序列特征整合在一起，并基于 XGboost 设计了一个新的多阶段结晶倾向性预测模型 CCmapCrys，通过与其他基于机器学习的预测模型进行对比，我们发现 CCmapCrys 模型取得了最优的预测性能。

但是考虑到 CCmap-KAAP 特征仍然属于一种手工特征，所以在第 4 章我们使用图注意力神经网络模型来自动提取蛋白质接触图中的空间结构特征，并设计了一个新的多阶段预测模型 GCmapCrys。该模型利用图卷积和局部多头自注意力机制来更新接触图中的氨基酸节点特征，使其更有利于结晶预测，并融合了多种蛋白质序列特征来进一步增强氨基酸节点表征。最终的实验结果表明了利用图注意力神经网络和蛋白质接触图能够很好地帮助结晶倾向性预测。而且我们还进行了特征有效性分析，并发现预测的二级结构和相对溶剂可及性与蛋白质结晶有更高的相关性。

第 5 章我们使用了 Graph Transformer 框架来预测蛋白质结晶，因为较浅的图注意力神经网络只能提取一定范围内的局部结构信息，所以我们使用 Transformer 框架来进一步提取全局的氨基酸相互作用信息。在此基础上我们还使用了残差连接和序列位置编码两个模块来提高模型的预测能力，残差连接使得图注意力神经网络能够保留充分的局部结构信息，序列位置编码能够让 Transformer 对蛋白质序列的位置信息更加敏感。我们将设计的新模型 GCmapTCrys 模型与目前最先进的预测模型进行对比，发现 GCmapTCrys 取得了最优的预测性能。而且在与 BCrystal 模型进行对比时，我们发现高精度的二级结构和相对溶剂可及性能够大幅提升模型的预测性能。

我们提出的三个预测模型 CCmapCrys、GCmapCrys 和 GCmapTCrys 都表明了蛋白质接触图对于结晶倾向性预测是非常有效的。但是仍旧有许多待提升的空间。首先是蛋白质接触图的预测精度，因为我们将预测的接触图作为信息来源，所以接触图的预测精度很大程度决定了我们模型的预测精度，在未来我们可以使用更高效且更精准的接触图

预测工具来进一步提升结晶倾向性预测的准确率；其次是对蛋白质序列特征的选择，与结晶相关的蛋白质序列特征对于模型预测准确率的提升是非常巨大的，我们在第 4 章和第 5 章都发现高精度的二级结构和相对溶剂可及性特征能够有效地提升模型的预测性能，因此在未来我们可以通过研究更先进的二级结构和相对溶剂可及性预测工具来辅助结晶预测；最后是预测速度方面，不论是蛋白质接触图还是其他基于序列的蛋白质特征，在获取时都会花费大量的计算时间，所以有一些端到端的模型只利用蛋白质一级序列来预测结晶倾向性。但是我们的实验结果表明相比于基于深度学习和蛋白质序列特征的方法，端到端的结晶倾向性预测模型拥有过低的预测准确率，所以在未来我们可以发展新的端到端结晶倾向性预测模型来更好的平衡蛋白质结晶倾向性的预测速度和预测准确率。

参考文献

[1]熊强, 丁立新, 姜晓燕, et al. X 射线测定蛋白质结构的技术进展与研究现状 [J]. 癌变畸变 突变, 2019, 31(1): 82-5.

[2]Jang K, Kim H G, Hlaing S H S, et al. A Short Review on Cryoprotectants for 3D Protein Structure Analysis [J]. Crystals, 2022, 12(2): 138.

[3]Warren B E. X-ray Diffraction [M]. Courier Corporation, 1990.

[4]Durbin S, Feher G. Protein crystallization [J]. Annual review of physical chemistry, 1996, 47(1): 171-204.

[5]Rupp B, Wang J. Predictive models for protein crystallization [J]. Methods, 2004, 34(3): 390-407.

[6]Brenner S E. Target selection for structural genomics [J]. nature structural biology, 2000, 7(11): 967-9.

[7]Canaves J M, Page R, Wilson I A, et al. Protein Biophysical Properties that Correlate with Crystallization Success in Thermotoga Maritima: Maximum Clustering Strategy for Structural Genomics [J]. J Mol Biol, 2004, 344(4): 977-91.

[8]Goh C-S, Lan N, Douglas S M, et al. Mining the Structural Genomics Pipeline: Identification of Protein Properties that Affect High-Throughput Experimental Analysis [J]. J Mol Biol, 2004, 336(1): 115-30.

[9]W Nicholson Price I, Chen Y, Handelman S K, et al. Understanding the physical properties controlling protein crystallization based on analysis of large-scale experimental data [J]. Nature biotechnology, 2009, 27(1): 51.

[10]Bertone P, Kluger Y, Lan N, et al. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics [J]. Nucleic acids research, 2001, 29(13): 2884-98.

[11]Smialowski P, Schmidt T, Cox J, et al. Will my protein crystallize? A sequence - based predictor [J]. Proteins: Structure, Function, Bioinformatics, 2006, 62(2): 343-55.

[12]Overton I M, Barton G J. A Normalised Scale for Structural Genomics Target Ranking: the OB-Score [J]. FEBS Lett, 2006, 580(16): 4005-9.

[13]Jin C, Gao J, Shi Z, et al. ATTCry: Attention-based neural network model for protein crystallization prediction [J]. Neurocomputing, 2021, 463: 265-74.

[14]Xuan W, Liu N, Huang N, et al. CLPred: a sequence-based protein crystallization predictor using BLSTM neural network [J]. Bioinformatics, 2020, 36(Supplement_2): i709-i17.

[15]Elbasir A, Moovarkumudalvan B, Kunji K, et al. DeepCrystal: A Deep Learning Framework for Sequence-Based Protein Crystallization Prediction [J]. Bioinformatics, 2019, 35(13): 2216-25.

[16]Chen L, Oughtred R, Berman H M, et al. TargetDB: a target registration database for structural genomics projects [J]. Bioinformatics, 2004, 20(16): 2860-2.

[17]Berman H M, Westbrook J, Feng Z, et al. The Protein Data Bank [J]. Nucleic acids research, 2000, 28(1): 235-42.

[18]Overton I M, Padovani G, Girolami M A, et al. ParCrys: A Parzen Window Density Estimation Approach to Protein Crystallization Propensity Prediction [J]. Bioinformatics, 2008, 24(7): 901-7.

[19]Parzen E. On Estimation of a Probability Density Function and Mode [J]. Ann Stat, 1962, 33(3): 1065-76.

[20]Slabinski L, Jaroszewski L, Rychlewski L, et al. XtalPred: A Web Server for Prediction of Protein Crystallizability [J]. Bioinformatics, 2007, 23(24): 3403-5.

[21]Genest C, Weerahandi S, Zidek J V. Aggregating opinions through logarithmic pooling [J]. Theory decision, 1984, 17(1): 61.

[22]Abraham R J, Fisher J, Loftus P. Introduction to NMR spectroscopy [M]. Wiley New York, 1998.

[23]Suykens J A, Vandewalle J. Least Squares Support Vector Machine Classifiers [J]. Neural Process Lett, 1999, 9(3): 293-300.

[24]Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial intelligence, 1997, 97(1-2): 273-324.

[25]Heckerman D. A tutorial on learning with Bayesian networks [J]. Innovations in Bayesian networks, 2008: 33-82.

[26]Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs [J]. Biochemical biophysical research communications, 2007, 355(3): 764-9.

[27]Hall M A. Correlation-based feature selection for machine learning [D]; The University of Waikato, 1999.

[28]Kurgan L, Razib A A, Aghakhani S, et al. CRYSTALP2: sequence-based protein crystallization propensity prediction [J]. BMC structural biology, 2009, 9(1): 1-14.

[29]Moody J, Darken C J. Fast learning in networks of locally-tuned processing units [J]. Neural computation, 1989, 1(2): 281-94.

[30]Overton I M, Van Niekerk C J, Barton G. XANNpred: Neural nets that predict the propensity of a protein to yield diffraction - quality crystals [J]. Proteins: Structure, Function,

and Bioinformatics, 2011, 79(4): 1027-33.

[31]Simulator S N N. User Manual, Version 4.1 [J]. Institute for Parallel Distributed High Performance Systems, University of Stuttgart, 1995.

[32]Mizianty M J, Kurgan L. Sequence-Based Prediction of Protein Crystallization, Purification and Production Propensity [J]. Bioinformatics, 2011, 27(13): i24-i33.

[33]Jahandideh S, Mahdavi A. RFCRYS: Sequence-Based Protein Crystallization Propensity Prediction by Means of Random Forest [J]. J Theor Biol, 2012, 306: 115-9.

[34]Chou K C. Prediction of protein cellular attributes using pseudo - amino acid composition [J]. Proteins: Structure, Function, Bioinformatics, 2001, 43(3): 246-55.

[35]Breiman L. Random forests [J]. Machine learning, 2001, 45(1): 5-32.

[36]Charoenkwan P, Shoombuatong W, Lee H-C, et al. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs [J]. PloS one, 2013, 8(9): e72368.

[37]Larson P D, Halldorsson A. What is SCM? And, where is it? [J]. Journal of Supply Chain Management, 2002, 38(3): 36-44.

[38]Wang H, Feng L, Zhang Z, et al. Crysalis: An Integrated Server for Computational Analysis and Design of Protein Crystallization [J]. Sci Rep, 2016, 6(1): 1-14.

[39]Wang H, Feng L, Webb G I, et al. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity [J]. Brief Bioinformatics, 2018, 19(5): 838-52.

[40]Consortium U. UniProt: a hub for protein information [J]. Nucleic acids research, 2015, 43(D1): D204-D12.

[41]Bairoch A, Apweiler R. The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000 [J]. Nucleic Acids Res, 2000, 28(1): 45-8.

[42]Chang C-C, Lin C-J. LIBSVM: a library for support vector machines [J]. ACM transactions on intelligent systems technology, 2011, 2(3): 1-27.

[43]Schäffer A A, Aravind L, Madden T L, et al. Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-Based Statistics and Other Refinements [J]. Nucleic Acids Res, 2001, 29(14): 2994-3005.

[44]Meng F, Wang C, Kurgan L. fDETECT Webserver: Fast Predictor of Propensity for Protein Production, Purification, and Crystallization [J]. BMC Bioinform, 2017, 18(1): 1-11.

[45]Zhu Y, Hu J, Ge F, et al. Accurate Multistage Prediction of Protein Crystallization Propensity Using Deep-Cascade Forest with Sequence-Based Features [J]. Brief Bioinformatics, 2021, 22(3): bbaa076.

[46]Varga J K, Tusnády G E. TMCrys: predict propensity of success for transmembrane protein

crystallization [J]. Bioinformatics, 2018, 34(18): 3126-30.

[47]Zhou Z, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks; proceedings of the International Joint Conference on Artificial Intelligence, F, 2017 [C].

[48]Geurts P, Ernst D, Wehenkel L. Extremely randomized trees [J]. Machine learning, 2006, 63(1): 3-42.

[49]Lecun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition [J]. P IEEE, 1998, 86(11): 2278-324.

[50]Li Z, Yu Y. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks [J]. arXiv preprint arXiv:07176, 2016.

[51]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): 660-8.

[52]Asgari E, Mofrad M R. Continuous distributed representation of biological sequences for deep proteomics and genomics [J]. PloS one, 2015, 10(11): e0141287.

[53]Yin X, Goudriaan J, Lantinga E A, et al. A flexible sigmoid function of determinate growth [J]. Annals of botany, 2003, 91(3): 361-71.

[54]Baldi P, Sadowski P J. Understanding dropout [J]. Advances in neural information processing systems, 2013, 26.

[55]Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-80.

[56]Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition; proceedings of the International conference on artificial neural networks, F, 2005 [C]. Springer.

[57]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[58]Hu J, Han K, Li Y, et al. TargetCrys: Protein Crystallization Prediction by Fusing Multi-View Features with Two-Layered SVM [J]. Amino Acids, 2016, 48(11): 2533-47.

[59]Elbasir A, Mall R, Kunji K, et al. BCrystal: an interpretable sequence-based protein crystallization predictor [J]. Bioinformatics, 2020, 36(5): 1429-38.

[60]Chen T, Guestrin C. Xgboost: A scalable tree boosting system; proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, F, 2016 [C].

[61]Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks; proceedings of the International Conference on Learning Representations, F, 2018 [C].

[62]Wang Z, Chen J, Chen H. EGAT: Edge-Featured Graph Attention Network; proceedings of the International Conference on Artificial Neural Networks, F, 2021 [C]. Springer.
[63]Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model [J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.
[64]Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks [J]. arXiv preprint arXiv:01261, 2018.
[65]Min E, Chen R, Bian Y, et al. Transformer for Graphs: An Overview from Architecture Perspective [J]. arXiv preprint arXiv:08455, 2022.
[66]He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks; proceedings of the European conference on computer vision, F, 2016 [C]. Springer.
[67]Privett H K, Kiss G, Lee T M, et al. Iterative approach to computational enzyme design [J]. Proceedings of the National Academy of Sciences, 2012, 109(10): 3790-5.
[68]Emerson I A, Gothandam K, Applications I. Network analysis of transmembrane protein structures [J]. Physica A: Statistical Mechanics, 2012, 391(3): 905-16.
[69]Bagler G, Sinha S. Network properties of protein structures [J]. Physica A: Statistical Mechanics and its Applications, 2005, 346(1-2): 27-33.
[70]Bhavani S D, Sinha S. Mining of protein contact maps for protein fold prediction [J]. Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery, 2011, 1(4): 362-8.
[71]Zheng W, Li Y, Zhang C, et al. Deep‐learning contact‐map guided protein structure prediction in CASP13 [J]. Proteins: Structure, Function, Bioinformatics, 2019, 87(12): 1149-64.
[72]於东军, 李阳. 蛋白质残基接触图预测 [J]. 南京理工大学学报: 自然科学版, 2019, 43(1): 1-12.
[73]Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: a better web interface [J]. Nucleic acids research, 2008, 36(suppl_2): W5-W9.
[74]Remmert M, Biegert A, Hauser A, et al. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment [J]. Nat Methods, 2012, 9(2): 173-5.
[75]Mirdita M, Von Den Driesch L, Galiez C, et al. Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments [J]. Nucleic Acids Res, 2017, 45(D1): D170-D6.
[76]Jones D T, Kandathil S M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features [J]. Bioinformatics, 2018, 34(19): 3308-15.
[77]Michel M, Menéndez Hurtado D, Elofsson A. PconsC4: Fast, Accurate and Hassle-Free Contact Predictions [J]. Bioinformatics, 2019, 35(15): 2677-9.

[78]Li Y, Zhang C, Bell E W, et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks [J]. PLoS Comput Biol, 2021, 17(3): e1008865.

[79]Du Z, Su H, Wang W, et al. The trRosetta server for fast and accurate protein structure prediction [J]. Nat Protoc, 2021, 16(12): 5634-51.

[80]Zhang H, Huang Y, Bei Z, et al. Inter-Residue Distance Prediction From Duet Deep Learning Models [J]. Front Genet, 2022, 13.

[81]Kouranov A, Xie L, De La Cruz J, et al. The RCSB PDB information portal for structural genomics [J]. Nucleic acids research, 2006, 34(suppl_1): D302-D5.

[82]Li W, Godzik A. Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences [J]. Bioinformatics, 2006, 22(13): 1658-9.

[83]Zhang C, Zheng W, Mortuza S, et al. DeepMSA: Constructing Deep Multiple Sequence Alignment to Improve Contact Prediction and Fold-Recognition for Distant-Homology Proteins [J]. Bioinformatics, 2020, 36(7): 2105-12.

[84]Friedman J H. Greedy function approximation: a gradient boosting machine [J]. Annals of statistics, 2001: 1189-232.

[85]Freund Y, Schapire R E. Experiments with a new boosting algorithm; proceedings of the icml, F, 1996 [C]. Citeseer.

[86]Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees [M]. Routledge, 2017.

[87]黄福祥, 栗大超, 宋冰, et al. 基于 GGBP 蛋白绑定的表面等离子共振葡萄糖浓度测量 [J]. 纳米技术与精密工程, 2010, 8(2): 132-6.

[88]Kozlowski L P. IPC 2.0: Prediction of Isoelectric Point and pKa Dissociation Constants [J]. Nucleic Acids Res, 2021, 49(W1): W285-W92.

[89]黄曼, 卞科. 蛋白质疏水性测定方法研究进展 [J]. 粮油食品科技, 2004, 12(2): 31-2.

[90]Kyte J, Doolittle R F. A Simple Method for Displaying the Hydropathic Character of a Protein [J]. J Mol Biol, 1982, 157(1): 105-32.

[91]Kawashima S, Kanehisa M. AAindex: Amino Acid Index Database [J]. Nucleic Acids Res, 2000, 28(1): 374-.

[92]Cheng J, Randall A Z, Sweredoski M J, et al. SCRATCH: A Protein Structure and Structural Feature Prediction Server [J]. Nucleic Acids Res, 2005, 33(suppl_2): W72-W6.

[93]Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen‐Bonded and Geometrical Features [J]. Biopolymers, 1983, 22(12): 2577-637.

[94]Savojardo C, Manfredi M, Martelli P L, et al. Solvent accessibility of residues undergoing

pathogenic variations in humans: from protein structures to protein sequences [J]. Frontiers in molecular biosciences, 2021, 7: 626363.

[95]Chou K-C, Shen H-B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM [J]. Biochemical biophysical research communications, 2007, 360(2): 339-45.

[96]徐冰冰, 岑科廷, 黄俊杰, et al. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43(5): 755-80.

[97]Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering [J]. Advances in neural information processing systems, 2016, 29.

[98]Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:02907, 2016.

[99]Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs [J]. Advances in neural information processing systems, 2017, 30.

[100]Wang S, Hu L, Wang Y, et al. Graph learning approaches to recommender systems: A review [J]. arXiv preprint arXiv:11718, 2020.

[101]Baldassarre F, Menéndez Hurtado D, Elofsson A, et al. GraphQA: protein model quality assessment using graph convolutional networks [J]. Bioinformatics, 2021, 37(3): 360-6.

[102]Li X, Ying X, Chuah M C. Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving [J]. arXiv preprint arXiv:07792, 2019.

[103]De Boer P-T, Kroese D P, Mannor S, et al. A Tutorial on the Cross-Entropy Method [J]. Ann Oper Res, 2005, 134(1): 19-67.

[104]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:11929, 2020.

[105]Kim J, El-Khamy M, Lee J. T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement; proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), F, 2020 [C]. IEEE.

[106]Zhang J, Zhang H, Xia C, et al. Graph-bert: Only attention is needed for learning graph representations [J]. arXiv preprint arXiv:05140, 2020.

[107]Rong Y, Bian Y, Xu T, et al. Self-supervised graph transformer on large-scale molecular data [J]. Advances in Neural Information Processing Systems, 2020, 33: 12559-71.

[108]Ying C, Cai T, Luo S, et al. Do Transformers Really Perform Bad for Graph Representation? arXiv, 2021 [Z].

[109]Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN

encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv:14061078, 2014, 2014.

[110]Zhao L, Akoglu L. Pairnorm: Tackling oversmoothing in gnns [J]. arXiv preprint arXiv:12223, 2019.

[111]Rong Y, Huang W, Xu T, et al. Dropedge: Towards deep graph convolutional networks on node classification [J]. arXiv preprint arXiv:10903, 2019.

[112]Zhao W, Wang C, Han C, et al. Exploring Over-smoothing in Graph Attention Networks from the Markov Chain Perspective [J].

[113]Dwivedi V P, Bresson X. A generalization of transformer networks to graphs [J]. arXiv preprint arXiv:09699, 2020.

[114]Ba J L, Kiros J R, Hinton G E. Layer normalization [J]. arXiv preprint arXiv:06450, 2016.