蛋白质复合物是蛋白质相互结合完成某一项生物功能的集合。生物学上蛋白质复合物的识别与研究对细胞组成分析、药物预测等至关重要。生物实验的方法可以识别蛋白质复合物，但其成本较高、周期较长，无法满足大规模数据时代的研究需求。 现有的蛋白质复合物预测算法主要是基于计算的算法，将蛋白质之间广泛的相互作用抽象成图，蛋白质复合物抽象为图中的局部结构，此时蛋白质复合物预测问题转换为局部子图发现问题。复合物预测算法可以从互作网络中挖掘出大量的局部子图样本，但是由于预测算法本身的局限性，预测结果中会存在部分不符合复合物形成规律的样本。基于计算的复合物预测算法不具有对其预测结果的评价能力，无法识别并剔除这部分样本。

为了识别和剔除这部分对预测结果具有干扰性的样本，本文提出了蛋白质复合物筛选模型，构建了融合结点特征和邻边特征的蛋白质相互作用网络，并进一步构造了复合物特征子图样本集。本文提出了多种复合物筛选模型并针对每个模型进行了模型训练以及复合物筛选验证，复合物筛选模型包括基于图卷积的模型、基于邻边卷积的模型和基于点边消息传递网络的模型。

本文的研究工作及贡献：

1）融合结点特征和邻边特征的蛋白质相互作用网络（简称特征互作网络）。基于蛋白质互作数据构建了蛋白质相互作用网络，基于图自编码器、深度随机游走等网络嵌入方法为网络添加了结点特征，基于生物学上的多种相似性特征提取方法为网络添加了邻边特征，最终构建出特征互作网络。

构造了多种类别的蛋白质复合物特征子图样本集。原始蛋白质复合物数据集以蛋白质集合的形式存在，无法满足样本后筛选工作的要求。本文基于特征子图提取方法，在特征互作网络中构建对应蛋白质复合物数据集的特征子图样本集，包括基于邻居相似性融合多个标准集构建的正样本数据集；基于COACH算法构建的中间样本数据集；基于改进的随机算法构建的负样本数据集；基于核心附属结构的复合物预测算法构建的待筛选数据集。

2）基于图卷积神经网络的复合物筛选模型。从拓扑数据出发，本文提出了基于图卷积神经网络的复合物筛选模型。该模型基于图自编码器和深度随机游走嵌入得到蛋白质全局特征，并依据蛋白质复合物子图的拓扑结构，基于图神经网络对蛋白质全局特征进行深度融合。本文将该模型与无全局特征的模型、基于拓扑统计特征的模型进行了对比与分析，结果表明了融合全局特征的图卷积神经网络方法的有效性。

3）基于邻边卷积的复合物筛选模型。从生物数据出发，针对蛋白质复合物生物数据抽象程度较低的问题，本文提出了基于邻边卷积的复合物筛选模型。该模型在蛋白质互作网络中嵌入了表达蛋白质功能关联的多种相似性数据，作为互作网络邻边特征，并基于邻边卷积的方法将子图拓扑和邻边特征进行融合。实验对比了有无邻边特征情况下的邻边卷积模型，结果显示了带邻边特征的邻边卷积模型的有效性。

4）基于点边消息传递网络的复合物筛选模型。从特征融合角度出发，本文提出了基于点边消息传递网络的复合物筛选模型。该模型保留了表示全局拓扑特性的结点特征和表示生物功能特性的邻边特征，采用消息传递的更新方法在蛋白质复合物局部拓扑结构中同时更新结点特征和邻边特征，实现了全局拓扑特征与生物功能特征的动态更新和深度融合。结果表明基于点边消息传递网络地复合物筛选模型能更好地挖掘蛋白质复合物的形成规律，在多个实验中其评价指标的提升达到了最优。

A protein complex is a collection of proteins that bind to each other to accomplish a biological function. The identification and research of protein complexes in biology are very important for cell composition analysis, drug prediction and so on. The method of biological experiment can identify protein complex, but its cost is high and the cycle is long, which can not meet the research demand of large-scale data age. The existing protein complex prediction algorithm is mainly based on the calculation algorithm, abstracting the wide interaction between proteins into a diagram, protein complex abstraction into the local structure of the diagram, at which time the protein complex prediction problem is transformed into a local sub-graph to find the problem. The composite prediction algorithm can extract a large number of local subgraph samples from the mutual network, but due to the limitations of the prediction algorithm itself, there will be some samples in the prediction results that do not conform to the formation law of the compound. The calculation-based compound prediction algorithm does not have the ability to evaluate its prediction results, and cannot identify and reject this part of the sample.

In order to identify and reject the samples which are disturbing to the prediction results, this paper presents a protein complex screening model, constructs a protein interaction network that combines node features and adjacent features, and further constructs a sample set of composite feature sub-maps. In this paper, a variety of composite filtering models are proposed and model training and composite filtering validation are carried out for each model, including models based on tortogram remnants, models based on adjacent curly volume, and models based on point edge messaging networks.

The research work and contribution of this paper:

1) A network of protein interactions (referred to as a network of features) that combines node features with adjacent features. Based on the protein interoperability data, the protein interaction network is constructed, the node characteristics are added to the network based on the network embedding methods such as graph self-encoder and deep random walk, and the adjacent features are added to the network based on various similarity characteristic extraction methods in biology, and the characteristic interoperability network is finally constructed.

A sample set of sub-maps of protein complex features in various categories is constructed. The original protein complex dataset exists in the form of a protein collection and cannot meet the requirements of post-sample screening. Based on the feature sub-graph extraction method, this paper constructs the characteristic sub-map sample set of the corresponding protein complex data set in the feature interoperability network, including the positive sample data set built on the basis of multiple standard sets of neighbor similarity fusion, the intermediate sample data set based on THECH algorithm, the negative sample data set based on the improved random algorithm, and the pending filtering data set based on the composite prediction algorithm of the core satellite structure.

2) A composite screening model based on a refride neural network. Based on the topological data, this paper presents a composite screening model based on the refride neural network. The model obtains the global characteristics of proteins based on graph self-encoder and deep random walk embedding, and based on the topology of protein complex subpic, the global characteristics of proteins are deeply fused based on the graph neural network. In this paper, the model is compared and analyzed with the model without global characteristics and based on topological statistical features, and the results show the effectiveness of the topographical neural network method of fusing global features.

3) A composite filtering model based on neighboring coils. Based on the biological data, in view of the low abstraction of the biological data of protein complexes, this paper puts forward a composite screening model based on the neighboring reftric. The model embeds a variety of similarity data expressing protein function association in the protein interoperability network, as a mutual network neighbor feature, and fuses the subgraph topology and neighbor feature based on the method of neighbor converse. The experiment compares the neighborhood remnoms model with or without neighbor features, and the results show the validity of the neighboring reuter model with neighbor features.

4)A composite filtering model based on a point-edge messaging network. From the point of view of feature fusion, this paper proposes a composite filtering model based on point-edge messaging network. The model retains the node characteristics representing the global topological characteristics and the neighbor characteristics representing the biological functional characteristics, and uses the method of message delivery to update the node features and adjacent features in the local topology of the protein complex at the same time, and realizes the dynamic update and deep fusion of the global topological features and biological functional features. The results show that the formation law of protein complex can be better excavated based on the point-edge message transmission network composite screening model, and the improvement of its evaluation index has reached the optimal level in many experiments.

蛋白质相互作用

蛋白质相互作用网络

# 绪论

## 研究背景

生物信息学是研究生物数据的采集、处理、存储、传播和解释等各方面的学科。 生物信息学可以帮助人类从海量的生物数据中挖掘内部的生理过程规律，从而指导进一步的生物学研究。在规模化实验大力发展和海量数据爆发的如今，如何有效的利用生物数据变得越来越中了。 生物信息学分为三个主要的发展阶段，前基因组时代主要建立了各种生物数据库和序列比较算法、基因组时代进行了大规模的基因测序以及目前所处的后基因组时代。 后基因组时代研究重心以生物数据分析为主，并且挖掘的层次逐渐深入。已经从对基因组直接的结构的研究逐渐转向对基因功能的研究，其中的主要侧重点包括基因组学、转录组学以及蛋白质组学等。

蛋白质是生物细胞和组织的重要组成部分，是生命体的物质基础，也是遗传信息的直接表达手段，涉及生物体载体、免疫、激素等方方面面。蛋白质分子深度参与了组织的构成与修复、生理功能的调节和能量的供给。 蛋白质组学是在蛋白质表达层面研究生理生化功能为主的一门学科，目的是揭示蛋白质的基本生命活动规律，其中研究主要关注蛋白质结构、蛋白质丰度、蛋白质修饰以及蛋白质相互作用。

蛋自质在细胞活动中发挥着巨大的作用。但是在多数情况下单个蛋自质无法独立的执行生物功能，只有构成蛋白质复合物，才能有效的参与到细胞活动中。因此蛋白质复合物的结构、功能及形成方式的研究就显得尤为重要。图[[fig:swr1\_complex]](#fig:swr1_complex)是具有染色质重塑功能的酵母SWR1复合物。

酵母菌SWR1复合物

酵母菌SWR1复合物

[fig:swr1\_complex]

生物实验中检测蛋白质复合物主要通过串联亲和纯化与质谱分析、酵母双杂交两种技术对蛋白质复合物进行分离和鉴定。串联亲和纯化与质谱分析通过靶蛋白标定以及自然条件下亲和纯化获取可能的蛋白质复合物，再使用质谱分析进行鉴定。酵母双杂交技术利用转录调控因子中的组件特征研究蛋白质之间的相互作用关系。虽然基于实验测定的方法具有生物学上的可解释性，但是生物实验往往条件困难、实验步骤多且成本昂贵，无法满足快速增长的研究需求。

蛋白质与生理环境存在广泛的相互作用，蛋白质复杂功能的实现同蛋白质之间、DNA与蛋白质、RNA与蛋白质的相互作用密切相关，蛋白质复合物正是一组强相关的蛋白质组合共同作用的结果。随着生物信息学的发展以及高通量技术的发展，蛋白质相互作用关系（Protein-ProteinInteraction，，后简称为互作关系）得到了大量的补充，促成了大规模互作网络的构建，即蛋白质相互作用网络（Protein-ProteinInteractionNetwork，）。图[[fig:ppi]](#fig:ppi)为酵母菌蛋白质相互作用网络。

酵母菌蛋白质相互作用网络（部分）

酵母菌蛋白质相互作用网络（部分）

已知的蛋白质复合物可以视作互作网络里面的一系列子网络，因此利用图论和计算模型学习子网络的分布模式，即可在互作网络中发掘潜在的蛋白质复合物。利用图论的方法，蛋白质互作网络可以转换为无向图 。其中为图中结点的集合，表示所有的蛋白质，为图中邻边的集合，表示所有的蛋白质互作关系。转换为图结构之后，图论上的计算模型和深度学习方法就可以迁移到蛋白质复合物的研究中。2002年，Tong等人提出了密集子图的假说，蛋白质复合物在是密集链接的子图，而与其他的蛋白质连接相对稀疏。在中预测蛋白质复合物的问题就转换为了密集子图的挖掘问题，并在此基础上逐渐发展出了利用和计算方法识别蛋白质复合物的理论。

蛋白质复合物的识别与评价也具有重要的研究价值。对复合物局部子图的识别有助于研究蛋白质复合物形成的内部机理，同时可以使得预测结果更加精确，有利用后续生物学上的进一步研究。

## 国内外研究现状

现有的基于计算方法的蛋白质复合物预测方法主要分为五类：基于网络结构的图聚类方法、融合生物数据的图聚类方法、核心附属扩展方法、动态网络方法以及监督学习方法。本章在[1.2.1](#subsection:research:Prediction)部分分别对这五类方法的研究现状做简单的介绍。本章在[1.2.2](#subsection:research:Selection)部分简要介绍蛋白质复合物筛选算法的现有算法。

### 蛋白质复合物预测

下面介绍五类基于计算方法的蛋白质复合物预测算法。

###### 1) 基于无权网络局部密集子图的预测算法

现有大多数复合物预测方法为基于网络结构的图聚类方法，基本思路是在无权无向图中挖掘密集子图。这类方法较为简单明确，取得了一定的成果。

MCODE算法是最早提出基于网络结构构造密集子图的复合物预测算法。首先算法会计算所有结点的局部邻居密度，其中密度超过平均值的结点成为种子，视作初始子图。满足相应阈值条件的邻居结点不断扩充子图，直到阈值条件饱和，最终子图视为预测的复合物，算法最终会过滤掉结点数少的复合物。 Clique算法通过穷举法、超顺磁性聚类和蒙特卡洛模拟三种方法搜索完全图来检测蛋白质复合物。 RNSC算法以随机聚簇最为初始聚簇，按照代价函数逐渐削减聚簇，最终形成蛋白质复合物。 Pereira‐Leal提出将马尔科夫聚类算法应用于蛋白质复合物检测，通过转移矩阵的自乘来扩展连通区域，通过幂运算进行膨胀操作只保留生成概率高的区域。膨胀和扩展操作交替运行，收敛之后获得蛋白质复合物。 LCMA算法从中较小的完全图开始，通过不断合并重合率高的完全图来检测蛋白质复合物。CFinder算法进一步定义了搜索与合并的策略，算法首先寻找网络中的k阶完全图，如果两个子图之间有k-1个公共结点，则定义为两个k阶完全图相邻，将两个子图合并。算法通过不断合并k阶完全图预测蛋白质复合物。 SCAN算法认为一对蛋白质的公共邻居超过阈值时，这对蛋白质可被视为结构可达，可以作为种子蛋白质继续扩展其余结构可达蛋白质。 ClusterONE算法充分利用了蛋白质复合物内部连接密集，外部连接稀疏的假设，并且明确定义了子图紧密性。其主要思路是首先按照结点度排序获取种子结点，种子结点向外扩展操过程中可以添加或删除结点，以达到局部子图最佳紧密性。 Zheng等人进一步改进了图紧密性定义，提出根据子图中3阶完全图个数来定义局部子图连紧密性。

有部分研究将中的结点做预分类以达到更优的结果。 CPridict算法首先计算蛋白质之间的功能相似性，然后以功能相似性做谱聚类，将蛋白质分为多个分组，各个分组独立进行复合物预测。CPridict2.0算法进一步基于FunCat功能目录对蛋白质进行分组。 CODEC算法基于质谱实验将蛋白质分类为诱饵蛋白和靶标蛋白，从靶标蛋白及其邻居结点开始，通过增减结点获得最大子图得分。

###### 2) 基于加权网络局部密集子图的预测方法

蛋白质复合物的预测是一个复杂的生物学问题，而由于实验手段的限制，互作数据存在着高假阴性和高假阳性的缺陷。研究者开始尝试挖掘网络拓扑特征，对网络结构进行修复和连边加权处理，提高的可信度以获取更精确的聚类结果。 DPClus算法最早对做加权处理，根据一对结点公共邻居数量给这对结点之间的边加权，结点的权重所有邻边的权重之和。权重较大的结点作为种子结点，通过紧密型结点的连接预测蛋白质复合物。 PCP算法利用使用FS-weight计算邻边的可靠度，并移除网络中可靠度较低的边，之后利用完全图发现合并算法检测复合物。 CMC算法采用了一种迭代式的边加权方法，不断的以结点的加权公共邻居数更新邻边权重，完全图评价分数也和子图的权重有关。 Bootstrap算法基于boostrap采样法预测复合物，在网络中做有放回重复采样，多次采样确定蛋白质相互作用权重，再进行多次执行马尔可夫聚类算法，形成蛋白质聚簇。最终以蛋白质在聚簇网络中的贡献性重构权重。

除了基于拓扑结构的加权方法，还可以利用生物数据对相互作用加权。已有研究表明，形成功能团的蛋白质通常具有相同的基因表达。部分研究借助基因表达数据更改权重，重新定义密集子图的评价函数以达到更优的聚类效果。 MATISSE算法使用基因表达数据的相关性衡量蛋白质的互作强度，并以此定义的权重，再将聚类算法运用到加权网络预测蛋白质复合物。 GFA算法基于最密子图算法获取互作网络中的最密子图，将子图的密度定义为子图内基因表达量之和，在聚类结果中选取密度较大的子图为复合物。

蛋白质复合物中，蛋白质的功能表达往往趋向于相近或者相同，因此功能表达数据也有助于蛋白质相互作用加权。Lubovac等人在基因本体信息（GENE Ontology，简称GO）形成的有向无环图上，使用SWEMODE算法计算两个蛋白质的功能相似性。在中结点权重通过聚合邻边权重以及邻居数计算，后续借鉴MCODE在结点和邻边均加权的网络中预测蛋白质复合物。WCOACH算法同样使用蛋白质功能表达数据加权，在加权网络上运用改进的COACH算法。

有研究表明，蛋白质结构域互作信息也会对蛋白质复合物的形成产生影响，多对蛋白质无法作用于互作界面的重复区域。Jung等人利用这个特性，剔除了MCODE和LCMA生成的结果中具有结构域冲突的复合物，提高了结果的准确性。DACO算法将蛋白质相互作用和结构域相互作用数据结合起来，使用图聚类算法预测蛋白质复合物时，进行结构域的紧密优化，预测出来的复合物确保存在结构域互作的可能性。

###### 3) 基于核心附属结构的预测算法

SCAN算法已经提出了基于种子蛋白质扩展的复合物预测方法。然而种子蛋白质仅仅时拓扑结构上的聚类核心，并不具有明显的生物学意义。研究人员通过酵母菌中复合物的结构发现，部分蛋白质存在于多种复合物中，且这些蛋白质之间存在大量的相互作用关系，组成了复合物的功能单元，此类蛋白质被称为核心蛋白质。在一个复合物中其余蛋白质附着于核心蛋白质，被称为附属蛋白质。如图[[fig:core\_append]](#fig:core_append)所示，其中深色椭圆形底框部分为复合物中的核心模块。

COACH算法是基于蛋白质复合物核心附属形成机理提出的算法，通过计算两个蛋白质公共邻居之间互作情况发现核心蛋白质，通过核心蛋白质周围相互作用关系添加附属蛋白质。

核心附属结构示例——Core64部分为核心结构

核心附属结构示例——Core64部分为核心结构

[fig:core\_append]

###### 4) 基于监督学习的预测算法

为了进一步利用现有数据，多个研究小组提出了基于监督学习的复合物预测算法。Qi等人将复合物的拓扑结构分为多类，使用真实的复合物作为训练样本，在PPI网络中随机生成子图作为复合物候选样本，选取了其中的生物及拓扑结构作为特征，使用概率贝叶斯网络作为分类器，判断其是否为复合物，再使用模拟退火算法对候选子图进行简单修改，预测复合物。Shi等人使用了神经网络方法，采用与Qi等人类似的样本生成策略进行复合物的预测。

###### 5) 基于动态网络的预测算法

由于蛋白质会随着生理生化周期不断变化，因此蛋白质相互作用也具有动态特性，部分研究将基因表达时序数据和蛋白质相互作用结合，构建动态蛋白质相互作用网络。 Tang等人设定阈值，将蛋白质基因表达数据超过阈值的蛋白质视为活跃蛋白质，以活跃蛋白质构建动态网络。Ou-Yang等人使用基因表达数据区分瞬态和稳态蛋白质。

### 蛋白质复合物筛选

蛋白质复合物的筛选主要是将复合物中的蛋白质集合与蛋白质相互作用网络结合起来，构建蛋白质复合物的局部子图，以局部子图作为蛋白质复合物的代表进行研究。部分监督学习方法提出了提取子图的多项拓扑特征，利用机器学习方法如线性回归、支持向量机等等对拓扑特征进行分类从而对评价蛋白质复合物。Yu提出了SLPC算法，基于GO注释数据对蛋白质相互作用网络加权，采用回归模型分类蛋白质复合物。

Liu等人在其论文中提出了在蛋白质相互作用网络中使用深度随机游走获取蛋白质编码，用蛋白质复合物子图中的蛋白质编码最大值代表复合物的特征，进而使用机器学习分类器进行蛋白质复合物预测的方法。

### 方法总结

表[[tab:MethodSummary]](#tab:MethodSummary)给出了现有复合物预测方法的类别、思想以及相关特点。

L3cmL6cmL5cm **方法类别** & **研究思想** & **特点**  
基于无权网络局部密集子图的预测算法 & 挖掘互作网络中的密集子图 & 倾向于预测连接紧密的子图，易受到蛋白质互作假阳性、假阴性的影响  
基于加权网络局部密集子图的预测方法 & 使用多种生物数据对互作网络加权，在加权网络的基础上挖掘密集子图 & 提高了互作网络可信度，依旧倾向于预测密集子图  
基于核心附属结构的预测算法 & 提取核心蛋白质，再在核心蛋白质周围不断添加附属将蛋白质 & 仅仅针对具有核心附属结构的蛋白质复合物  
基于监督学习的预测算法 & 利用已知复合物辅助的生物、拓扑特征训练分类模型，基于分类模型预测复合物 & 数据融合和模型训练具有一定的难度  
基于动态网络的预测算法 & 基于生物数据建立动态蛋白质复合物，再基于动态网络某一时刻的数据预测蛋白质复合物 & 动态网络更贴近生命活动的真实过程，然而预测算法依旧基于密集子图发现算法

已有的蛋白质复合物的筛选方法通常是针对特定的复合物生成算法做补充，训练出来筛选模型用于对应算法中蛋白质复合物的评分以及过滤，而无法作为通用的蛋白质复合物识别模型。Liu等人提出了直接使用全局拓扑特征做过滤的筛选算法，然而该算法未考虑蛋白质复合物的拓扑结构，未融合生物数据，具有一定的局限性。

## 研究动机、思路和基础

本节介绍研究动机、思路和基础，分析现阶段蛋白质复合物预测方法存在的问题、可行的改进思路以及相关的研究基础。

### 研究动机

国内外学者对复合物的识别已经提出了诸多方法，总体趋势也是趋向于融合生物数据和网络数据并达到更精确的预测，但是目前现有的方法还存在以下的不足。

* 无法充分利用已有的先验知识。上述的研究方法主要是基于无监督方法，在复合物的研究问题上，无监督方法具有训练简单、结构明确的优点。但是随着实验技术更新、生物数据量增加以及图数据挖掘算法的井喷发展，无监督学习方法无法利用已有的先验知识，其预测准确率无法得到进一步提升。有监督学习方法可以通过学习已有复合物的分布特征，挖掘出特定网络中蛋白质复合物的分布规律。目前已有的监督方法如Shi提出的方法仅仅学习复合物的拓扑特征的映射函数，且映射函数只作用于复合物扩展过程中单个结点的取舍判断，算法尚存在大量的改进空间。
* 复合物准确度较低。部分模型方法如RNSC算法、Clique算法等等具有很强的随机性，此类方法为了达到较高的复合物召回率，往往倾向于产生过量的候选复合物，导致结果的准确率的降低。预测的复合物中存在大量错误的样本，对后续的生物研究有一定的影响。如何在维持复合物预测召回率的情况下，提高预测的准确度是一个值得研究的问题。
* 生物数据在中融合度低。本文在基于加权网络局部密集子图的预测算法中提到了融合生物数据的诸多方法，然而这些方法只停留在利用生物数据更改邻边权重的层面，不同的生物数据如GO注释、基因表达信息、蛋白质保守性等数据被统一编码到了相互作用的边权重中，这个编码转换过程会丢失原有丰富的生物数据。编码过程仅仅作为数据的预处理，生物数据无法动态地参与到复合物预测模型的框架设计中。同时，目前也没有方法将生物数据编码到蛋白质结点中，如何有效的利用结点编码增强复合物的预测质量也是一个亟待解决的问题。

针对以上蛋白质复合物预测问题中存在的不足，从蛋白质复合物的预测到对预测结果的评价之间存在改进的空间。不同复合物预测方法可以得到大量的复合物预测结果，由于预测方法存在的缺陷，预测结果中存在部分不符合复合物形成规律的样本，而预测方法无法精确的识别并剔除这部分复合物。因此本文基于预测后的评价与筛选模型研究（简称后筛选）动机设计可行且有效的模型框架。

### 研究思路

本文提出了一种基于监督学习和融合特征的复合物后筛选模型。 算法的核心思想是利用已有的生物数据，包括、生物注释信息、已知复合物等等，将蛋白质复合物结点数据转换为可学习的结构化数据，并为结构化的数据添加标签，通过监督学习的方式构建一个鲁棒性的复合物分类模型。

通过模型设计以及合适的训练过程，模型能够融合蛋白质复合物的多种数据，包括该复合物具有哪些蛋白质、具有哪些邻边关系以及特征等等，较为准确的辨别某一个复合物是否为真实复合物，学习复合物的形成规律。 在此模型训练成功的前提下，将一般性的复合物预测算法，比如DPClus算法、Clique算法等产生的复合物交由分类模型做进一步筛选，，剔除其中得分较低的不符合复合物形成规律的样本，保留得分高的样本代替为相应算法预测的最终结果，相较于未筛选前的样本，最终预测结果的多项评价指标可以得到提升。

### 研究基础

两阶段筛选方法在已有领域的应用。 基于两阶段筛选方法在目标检测领域已经得到了广泛的运用，比如RCNN算法。在一张图片中检测目标可以分为两个阶段，第一阶段为获取所有可能的框图，第二阶段为对候选框做筛选与调整。对应到蛋白质相互作用网络中预测蛋白质复合物，也可以做类似的处理，首先获取对应的候选复合物，再利用复合物筛选算法，在候选复合物基础上获得更加精确的预测样本，以达到更优的预测结果。在复合物预测领域，这种两阶段思想具有一定的可行性。

复合物筛选领域已有部分筛选方法。 在复合物预测领域，已有部分研究提出过类似的方法，比如RNSC算法提出了基于设定得分函数筛选预测复合物的思想，论文提出了使用复合物拓扑性质对复合物做分类的方法。但是目前已有的复合物筛选方法研究均较为有限，已提出的方法未偏计算统计的方法，尚未有算法将生物数据与复合物的拓扑结构结合起来，无法学习蛋白质复合物深层次特征，且此类方法通常只将筛选部分作为特定模型的补充，而非成体系的研究，无法将筛选模型当成一般性方法应用于复合物筛选领域。为了改进朴素的基于计算的预测方法，本文提出以GCN为基础的一系列复合物分类模型，将复合物的特征与结构结合起来。

可供筛选的复合物样本。 部分复合物筛选算法，如Clique算法、Dpclus算法等，是基于随机种子以及子图扩展的方法。这类方法由于能够广泛的挖掘整个网络的所有结点，因此其预测结果能够匹配出多数已知的蛋白质复合物。但是此类方法会过多样本，其中大量的具有随机性的复合物对总体的预测结果具有干扰性。使用复合物筛选模型剔除具有随机性的负样本，可以在保持该类方法高召回率的同时，提高预测结果的准确度。

## 研究工作和贡献

本节介绍本文的主要研究工作以及相关成果。在总体层面上将所有的研究内容构建成工作流程图，并对每一部分的工作做了详细的介绍。

本文的主要贡献如图[[fig:main-flow]](#fig:main-flow)所示，主要由四个阶段组成，分别是构建融合结点特征和邻边特征的蛋白质相互作用网络；利用特征互作网络构建复合物子图样本集；基于训练数据集中的子图样本训练图分类模型；基于已训练复合物评价模型筛选复合物及对比。其中重要的阶段为第二阶段和第三阶段。

该算法框架补充了已有的复合物筛选算法的部分不足。针对现有的复合物预测算法无法利用先验知识，算法框架提出了基于监督学习的复合物评价模型，构建了监督学习的特征子图训练集；针对复合物预测结果精度低，框架提出了基于已有复合物预测算法的二阶段筛选方法，并进行对比实验验证筛选前后的评价指标；针对现有预测方法无法融合拓扑结构与相关的特征数据，框架提出了多种特征融合方法。

实验结果表明，本文筛选模型能够挖掘复合物深层次的规律。在多个复合物预测方法上，经过筛选模型过滤，预测样本的多项评价指标得到了提升，更加精确的预测结果更具有生物学上的参考意义。

研究内容和贡献图示

研究内容和贡献图示

###### 第一阶段，构建融合结点特征和邻边特征的蛋白质相互作用网络

利用蛋白质相互作用数据集构建蛋白质相互作用网络，其中每一个结点代表一个蛋白质，每一条邻边代表蛋白质之间的相互作用关系。然后利用多种结点嵌入方法和邻边嵌入方法给蛋白质相互作用网络添加相应的特征数据。 最终形成特征相互作用网络。 该阶段总体过程如图中标号部分所示。

基于多种网络嵌入方法和多种生物数据嵌入方法，本文得到了兼具结点特征和邻边特征的网络结构，最终，中每一个蛋白质具有82维特征，每一对蛋白质互作具有12维特征，具体特征分布如下。

* 结点特征部分：包括Deeowalk特征、GAE特征、蛋白质长度特征；
* 邻边特征部分：包括蛋白质功能注释GO特征、结构域特征、亚细胞定位特征、局部拓扑相似性特征；

###### 第二阶段，利用特征互作网络构建复合物子图样本集

蛋白质复合物数据集中每一个蛋白质复合物数据都是以蛋白质集合的形式存在。将中的蛋白质对应到复合物相互作用网络中的结点集合，在特征互作网络中抽取结点集合对应的子图，得到特征子图数据。

依据复合物数据集来源的不同，特征子图样本可分为已知复合物得到的正样本子图、COACH算法得到的中间样本子图、随机数据得到的负样本子图以及待筛选复合物预测算法得到的待筛选样本子图。 依据已有的复合物评价指标，正样本得分为1、中间样本得分为01、负样本得分为0，这些携带标签的样本组合起来作为训练数据集。 该阶段总体过程如图中标号部分所示。

###### 第三阶段，基于训练数据集中的子图样本训练图分类模型

训练集中的样本均为图结构数据，每一个图结构数据代表一个蛋白质复合物，且具有结点和邻边特征，同时具有分类标签和评分。在此基础上，基于图分类模型可以训练从特征子图到分类和评分的映射函数，该映射函数作为复合物评价模型。

基于多种图分类和评价模型，本文进行了复合物评价模型的设计与训练。本文涉及了三种图分类和评价模型，分别是基于结点特征模型（对应章节[3](#chapter:NodeConv)）、基于邻边特征模型（对应章节[4](#chapter:EdgeConv)）和基于特征融合模型（对应章节[5](#chapter:MPNN)）。 该阶段具体过程如图中标号所示。模型具体细节如下。

* 基于结点特征模型：提出了基于结点的GCN方法，融合结点的全局拓扑特征进行复合物预测；
* 基于邻边特征模型：提出了基于邻边的GCN方法，将初始的邻边特征转换为结点特征，再基于edge-base的GCN方法对特征进行融合；
* 基于特征融合模型：提出了融合结点特征和邻边特征的MPNN方法，提出了结点特征和邻边特征交替更新的方法，在特征融合同时保留自身关键特征。子图特征分为MPNN得到的可学习特征和基于图论得到的计算统计特征，两部分特征结合具有更好的预测准确度；

###### 第四阶段，基于已训练复合物评价模型筛选复合物及对比

在蛋白质相互作用网络中预测初始复合物，作为待筛选复合物，复合物包括DPClus算法、clique算法等的预测结果。依据标号复合物评价模型的评价结果，有一部分评分较高，其余部分评分较低，保留其中评分搞得复合物样本作为筛选后的结果，筛选过程如图中标号部分所示。最后本文使用多种的复合物预测评价指标对筛选前后的总体样本进行综合评价，如图中对比部分所示。

本文进行了多项实验验证模型的有效性，在DIP网络和Biogrid网络上进行分别训练了各个预测模型，并在Dpclus、IPCA等方法上验证筛选结果，验证的指标包括f1值和复合物预测领域的融合指标。

## 论文组织

本论文分为五大章节，具体如下：

第一章，绪论。介绍了主流的蛋白质复合物预测算法，并根据存在的问题提出了可改进的空间。对研究背景、研究意义以及本文的主体研究流程做了相关介绍。 介绍了筛选模型的总体框架结构。

第二章，蛋白质复合物特征子图样本生成。阐述了具有结点特征和邻边特征的蛋白质相互作用网络的构建方法，包括GAE和Deepwalk的结点特征嵌入、以及多种基于蛋白质相似性计算的邻边嵌入方法；后续阐述了如何从特征相互作用网络中提取蛋白质复合物子图，以及训练子图样本、待筛选子图样本的生成方法。

第三章，基于图卷积神经网络的复合物筛选模型。从拓扑特征入手，提出了利用结点的拓扑特征并基于结点的图卷积神经网络的复合物筛选模型，挖掘网络中结点嵌入对复合物预测的作用。同时提出了本文算法框架中的参数细节，评价指标等等。后续进行了实验并验证模型的有效性。

第四章，基于邻边卷积网络的复合物筛选模型。从特征子图中的邻边相似性出发，探讨了使用生物特征转换为的邻边数据的基础上，复合物筛选模型可达到的效果提升。提出了使用基于邻边卷积的图神经网络方法以及相应的模型，阐述了邻边特征如何初始化结点特征，以及EdgeConv的信息流更新过程。后续进行了相关对比实验并验证模型的有效性。

第五章，基于点边消息传递网络的复合物筛选模型。融合全局特征以及生物相似性特征的情况下，提出了基于点边消息传递网络的复合物筛选模型。 特征子图具有GAE和Deepwalk特征等结点特征、多种蛋白质相似性嵌入邻居特征，提出了改进的MPNN更新方法，实现了结点与邻边交替融合更新的模型。最后对比了利用结点特征的基于图卷积的模型，利用相似性特征的基于邻边卷积的模型，并进行实验验证。

第六章，总结与展望。对所有研究工作的概括，以及未来可深入研究点的展望。

# 蛋白质复合物特征子图样本生成

本章主要介绍蛋白质复合物特征子图的生成，其中包括蛋白质相互作用网络构建、结点特征生成与嵌入、邻边特征生成与嵌入以及特征子图样本的提取。重点在于本章使用了深度学习方法、网络嵌入方法、多种邻边相似性特征提取方法构建具有丰富特征的蛋白质相互作用网络，并在此基础上构建了特征子图样本数据集，包括正样本数据集、中间样本数据集、负样本数据集、待筛选样本数据集。

## 引言

目前公开的已验证蛋白质复合物均为以蛋白质集合表示，每一个蛋白质集合代表一个复合物，表示该集合内若干个蛋白质之间具有一定的相互作用并形成复合物。但是集合的数据不具有可学习性，计算方法无法有效识别某个蛋白质集合能否形成复合物，因此需要将集合数据转换为复合物的结构化数据，结构化数据为从蛋白质复合物特征子图。特征子图中包含了复合物中的蛋白质集合、蛋白质相互作用数据、蛋白质特征以及蛋白质相互作用特征。在蛋白质复合物结构化数据的基础上，深度学习的方法可以有效的挖掘蛋白质复合物的内部规律，学习蛋白质复合物的形成条件。

蛋白质复合物特征子图样本的构造需要以下几步。首先利用已知的蛋白质相互作用数据集构建蛋白质相互作用网络，再在互作网络的基础上嵌入丰富的结点特征和邻边特征，形成带特征的蛋白质相互作用网络，最后根据蛋白质复合物中的蛋白质集合作为子图结点集合从带特征的蛋白质相互作用网络中提取特征子图。按照蛋白质复合物数据来源划分，特征子图可以区分为正样本特征子图、中间样本特征子图、负样本特征子图和待筛选特征子图，前三者作为训练集训练蛋白质复合物筛选模型，最后复合物筛选模型会在待筛选特征子图进行测试。

## 蛋白质相互作用特征网络构建

蛋白质相互作用特征网络构建目标是将蛋白质的相互作用数据转换为图结构，并为图结构添加结点特征和邻边特征。以下分别介绍图结构的形成、结点特征的添加以及邻边特征的添加。

### 基于互作关系构建蛋白质相互作用网络

酿酒酵母（Saccharomyces cerevisiae）的蛋白质相互作用中有较深入的研究，其标准复合物数据集，蛋白质相互作用数据较为完备，因此本文基于酿酒酵母的蛋白质互作网络展开研究。

本文分析了主流的酿酒酵母细菌的数据集，包括DIP、Krogan、Biogrid和Gavin。

DIP数据集是结合各种来源的信息创建的一致的数据集，可以自动从论文中提取关系数据并更新数据库，同时有专业人员管理。 Krogan数据集使用技术检测蛋白质相互作用并借助机器学习方法提高结果的可靠度，数据集分为核心和附属两个部分，本文将核心附属网络融合进行实验。 Biogrid数据集融合了多个数据集来源，同时排出了不符合多重验证标准的相互作用。 Gavin数据集基于亲和纯化与质谱技术检测蛋白质相互作用，并使用socio-affinity指数评估检测结果的可靠度。

每个数据集都由若干蛋白质相互作用对组成，表示两个蛋白质之间存在相互作用关系，依据关系数据，可以相应的构造。以的形式存在，为一个图结构（Graph），为图结构的所有结点，每一个结点代表一个蛋白质，为图结构中的所有邻边，每一条邻边代表两个蛋白质之间存在相互作用关系，如代表第i个蛋白质和第j个蛋白质之间存在互作关系。图结构以邻接矩阵的形式存在。

数据集具体的网络结构如图[[fig:ppi-datasets]](#fig:ppi-datasets)所示，具体统计学特征如表[[tab:PPIStatic]](#tab:PPIStatic)所示。

0.5cm

L4.2cmL2cmL2cmL2cmL2cm **网络特征** & **DIP** & **Krogan** & **Biogrid** & **Gavin**  
蛋白质数 & 4928 & 3672 & 5573 & 1855  
相互作用数 & 17201 & 14317 & 59636 & 7669  
平均度数 & 6.98 & 7.80 & 21.40 & 8.27  
密度 & 0.0014 & 0.0021 & 0.0038 & 0.0035  
平均聚类系数 & 0.0945 & 0.1203 & 0.2483 & 0.4674  
连通个数 & 28 & 14 & 1 & 43  
最大连通结点数 & 4873 & 3642 & 5573 & 1727  
传递性 & 0.0728 & 0.1001 & 0.0657 & 0.5676  
直径 & 11 & 10 & 6 & 13

从表[[tab:PPIStatic]](#tab:PPIStatic)中可以看出，DIP数据集和Biogrid数据集是结点数和邻边数最大的两个数据集，同时其度数差异性较大，DIP数据集密度0.0014最低，平均度数最低，代表稀疏连接的蛋白质相互作用网络，而Biogrid数据集密度0.0038为最高，平均度数最高，代表密集连接的蛋白质相互作用网络。这两个数据集具有一定的数据基础，同时其网络特征具有一定的代表性，本文的研究工作主要围绕这两个数据集展开。

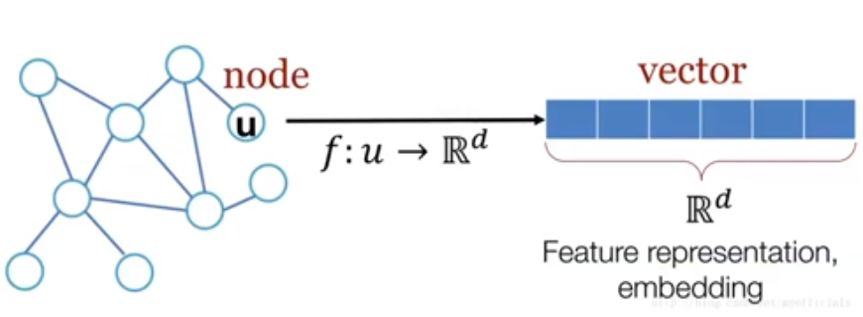
利用数据集构建相应的网络结构之后，可以将蛋白质特征以及蛋白质相互作用特征添加到网络中去，形成带有生物特征的蛋白质相互作用网络。在蛋白质相互作用网络结点上添加蛋白质自有特征和嵌入特征作为结点的初始特征。蛋白质自有特征包括蛋白质序列长度、分子重量特征，蛋白质嵌入特征包括图自编码器器（Graph Auto Encoding，简称GAE）和结点嵌入模型（Node2Vec）中的深度随机游走方法（Deepwalk）得到的图嵌入特征。在蛋白质相互作用网络邻边上添加通过生物计算的蛋白质相互作用特征，包括蛋白质功能注释特征、拓扑域相似性特征和亚细胞定位相似性特征，以及蛋白质拓扑域相似性特征作为图网络里面邻边的初始特征。

下面分别介绍结点特征的嵌入方法和邻边特征的嵌入方法。

### 基于图自编码器器和深度随机游走的结点特征嵌入

图结构是高维的复杂的数据，包含噪声和冗余信息，直接的图结构数据无法应用于机器学习或数据分析任务中。结点特征嵌入就是将网络数据映射到低维空间中，从而学习到网络节点的低维向量表示，低维向量的长度一般为固定值，远小于网络的结点数量。在低维表示下，可以应用丰富的数据分析方法进行进一步处理，比如结点分类、链路预测、图分类等等。

传统的结点表示方法是基于拓扑的，直接通过邻接矩阵的简单运算得到结点的特征，其中可能包含过量的噪声或冗余信息。



结点特征嵌入示意图

结点特征嵌入映射到二维平面示意图

结点特征嵌入映射到二维平面示意图

图[[fig:node-vec]](#fig:node-vec)展现了结点特征表示的过程，结点u经过特征嵌入后，可以转换为长度为d的向量表示，向量完全表达了结点信息。图[[fig:node-embedding]](#fig:node-embedding)展现了在结点分类任务上，网络嵌入的分类效果，经过结点嵌入的各个结点可以转换为一系列的向量，将该系列的向量映射到二维空间之后，可以依旧按照原始网络中的分布规律在平面内聚簇分布，如图中不同颜色的聚簇表示。

本文基于图自编码器和深度随机游走嵌入对蛋白质相互作用网络中的蛋白质结点特征进行嵌入，同时在蛋白质结点中保留了部分蛋白质自有的生物学信息嵌入。具体介绍如下所示。

#### 基于图自编码器的结点特征嵌入

图自编码器是一种基于深度学习的结点嵌入方法，其核心思想是结点的嵌入特征需要具备有效恢复整个网络结构的信息。在此前提基础上图自编码器主要有两个过程，编码过程和解码过程，编码到解码的中间状态的结点表示即为图中的结点特征嵌入。 图自编码器的编码过程基于图卷积神经网络实现，解码过程基于链路预测和网络还原实现。

下面分别对图卷积神经网络、自编码器以及图自编码器做简单介绍，后续介绍基于图自编码器的蛋白质互作网络结点嵌入模型。

###### 1) 图卷积神经网络

卷积神经网络快速发展，其具有高效的特征提取能力。相比传统方法，卷积神经网络给图像处理和自然语言处理等领域带来了空间的提升，如机器翻译、图像识别和语音识别等。但是传统的卷积神经网络只能处理欧氏空间数据，欧式空间数据具有平移不变性，其定义的全局共享的卷积核可以处理所有的数据点，如图像数据中，图片转换为像素点，任意像素点为中心都可以按同样的卷积核提取周围的信息。

图结构数据是非欧式空间数据，图数据中每个结点的局部结构各异，这使得图数据不具有平移不变性。图卷积神经网络（Graph Convolution Network，简称GCN）的提出解决了该缺陷，GCN基于邻居特征聚集以及将转换矩阵作为卷积核的思想巧妙的实现了在每一个结点做相同卷积过程。

单个的图卷积过程具体如公式[[equ:gcn]](#equ:gcn)所示。

其中为经过拉普拉斯平滑的邻接矩阵；代表第l层的结点特征矩阵，其中每一个结点特征维度为m；代表卷积核，代表每一个结点的特征转换；为激活函数；代表经过一轮卷积之后的输出结果。

类似于传统的卷积方式，GCN也可以多层卷积组合到一起进行，如图[[fig:gcn\_main]](#fig:gcn_main)所示。图卷积的输入为邻接矩阵代表的图结构和特征矩阵代表的结点特征，隐层之间采用Relu的激活函数，最后得到图的结点特征。

GCN示意图

GCN示意图

[fig:gcn\_main]

###### 2) 自编码器

自编码器（Auto Encoder）是无监督模型，由编码器和解码器两部分构成。编码器的任务是将原始的输入数据转换为编码数据表示，类似于数据压缩数据或者特征提取过程。解码器是将编码器得到的数据重新还原成原始的输入数据。自编码器的训练目标是使最终输出数据和原始输入数据之间的误差尽量小。基于该目标，自编码器的中间层编码数据具有复原输入的能力，可以认为这种编码数据表示有效地提取到了原始数据的特征。通常情况下，编码数据较原始数据维度低，具有一定的数据降维能力，同时也能去除数据中的噪声，有效提高数据的可用性。

自编码器示意图

自编码器示意图

[fig:gae\_autoencoder]

图[[fig:gae\_autoencoder]](#fig:gae_autoencoder)为经典的图片自编码器示意图，输入图像为，中间编码为，输出图像为，输入输出差距为$\varDelta X$，编码器的函数过程可以用表示，解码器过的函数过程可以用表示。总体过程可以如式[[equ:autoencoder]](#equ:autoencoder)所示。通过反向传播不断的减小$\varDelta X$，最终隐层编码表示即可学习到输入图像的特征。

$$\label{equ:autoencoder}
X^\prime =G(F(X)),
\varDelta X={\| X^\prime - X \Vert}$$

###### 3) 图自编码器

图自编码器是结合了图卷积神经网络和自编码器的一种用于处理图结构数据的深度神经网络。其具体过程如图[[fig:gae\_main]](#fig:gae_main)表示，其目标是对图结构进行编码，得到结点的特征表示。

相较于普通的自编码器，图自编码器的输入为图的邻接矩阵和图的特征矩阵，编码过程是多层的图卷积神经网络，解码过程是用编码之后的结点特征做基于相似性计算的链路预测，最终对比经过链路预测得到的邻接矩阵和输入的邻接矩阵的差异。

GAE示意图

GAE示意图

[fig:gae\_main]

###### 4) 基于图自编码器的蛋白质互作网络结点嵌入模型

本节介绍在蛋白质互作网络上基于图自编码器的结点嵌入模型。

本文的自编码器模型的初始输入特征为常数，本文设定为1，这种情况会削弱输入结点特征对模型的影响，使模型能够关注并学习到相互作用网络的全局的拓扑特征。

编码过程采用3层GCN网络，第一层GCN网络的特征转换矩阵为，第二层GCn网络的特征转换矩阵为，第三层GCn网络的特征转换矩阵为，即最终得到的结点隐层表示维度为16维，整个网络中结点的隐层特征为。 采用三层GCN网络可以使得结点可以获取3跳之外的邻居信息，通过GCN融合到一起就可以整个直径范围为7的信息，通过[[tab:PPIStatic]](#tab:PPIStatic)的统计可以得出，不同的的全局图直径范围在6至12，因此这个卷积方法可以融合至少半个图结构范围的信息，一定程度上可以体现结点的全局拓扑特征。 在每层GCN中，模型采用了LeakyRelu函数作为激活函数，同时模型采用了全局0.1的Dropout，一定程度上防止模型的过拟合。

基于图自编码器的蛋白质互作网络结点嵌入模型

基于图自编码器的蛋白质互作网络结点嵌入模型

[fig:gae\_mygae]

解码过程采用了基于神经网络的链路预测模型，对于图中的每一条可能的邻边，基于邻边两个端点的特征组合到一起，经过两层神经网络，再经过Softmax激活函数得到一个0至1的数值，代表该条邻边存在的可能性，具体过程如式[[equ:decoder]](#equ:decoder)所示。

式中和为经过编码器编码的结点i和j的特征，为邻边存在的概率。最终预测得到的图邻接矩阵为。

损失函数为预测的邻接矩阵和真实邻接矩阵的MSELoss（Mean Square Error Loss）。

其中G代表图结构，代表图结构中i结点和j结点连边，而为预测的边权重，为原始图中真实的边权重，如果原始图中邻边不存在，权重为0。

训练过程的其他训练细节如下所示。训练过程中采用了Adam优化器，学习率为0.01，模型训练轮数为80epoch。

最终基于图自编码器的蛋白质互作网络结点嵌入模型采用最后一轮的隐层作为最终的结点嵌入，最终的结点嵌入结果为的特征矩阵。

具体算法如下所示。

: PIN graph struct data; : adjacent matrix of ; : epoch index; : node num in PIN; : node feat after GAE Process; initial nodefeat=1, ;initial i=0;initial : Multi-layer perceptifier,: node-base graph convolution network; compute hidden feat 1, ,,; compute hidden feat 2, ,; compute hidden feat 3, ,; ,extract hidden node feats; ,reconstruct each edge; compute loss, ; optimization, ; i=i+1; extract node feats in final epoch to , =;

算法中输入包括蛋白质相互作用网络的图结构G以及其邻接矩阵A，GAE的运行轮数为80。第1行表示GAE的初始化，包括初始化结点特征为常数1，初始化轮数为0，以及其中编码器和解码器的参数初始化。 第3行至14行表示具体的GAE过程，其中3、4、5行为三层GCN组成的编码器，第6行表示取最后一层隐层特征作为结点编码，第7至11行表示解码得到G中每一条邻边的权重，第12行和13行为计算损失和进行GAE的优化，第16行表示最终的GAE结果为最后一个epoch的隐层特征。

#### 基于Deepwalk的结点特征嵌入

Deepwalk是基于随机游走的图结点嵌入模型，基本思想是通过在图中随机游走的方式获取若干结点的序列，这些序列被当作具有上下文语言的语句，而结点被当作语句中的一个单词。通过Word2Vec的模型，将语句中的单词转换为向量，作为结点的特征，即可得到结点特性嵌入。在图中游走获取的结点序列能够在一定程序上包含结点周围的拓扑信息，若干周围信息组合到一起则可以提取出整个图结构的拓扑信息。通过序列获取结点嵌入后使结点依旧保持图结构的拓扑信息。

RandomWalk是一种可重复访问已访问节点的深度优先遍历算法。给定当前访问起始节点，从其邻居中随机采样节点作为下一个访问节点，重复此过程，直到访问序列长度满足预设条件。该采样过程得到的序列如图[[fig:deepwalk\_main]](#fig:deepwalk_main)中的b部分所示。

Skip-gram是一种词向量嵌入模型，基本思路是利用序列中的单个词预测周围词的模型，在学习了所有序列之后，可以得到每一个结点的特征嵌入。利用Skip-gram模型获取结点的嵌入过程如图[[fig:deepwalk\_main]](#fig:deepwalk_main)中的c部分所示。

deepwalk嵌入示意图示意图

deepwalk嵌入示意图示意图

[fig:deepwalk\_main]

考虑到的规模，本文的Deepwalk嵌入模型的序列长度为100，基于每一个结点均随机游走采用10跳序列，最终的序列数据为，n为不同的的结点数。最终获取的Deepwalk嵌入特征维度为64维。

#### 基于蛋白质自有特征的结点特征嵌入

在细胞中，蛋白质可以由一系列氨基酸构成，基于氨基酸序列本文获取了部分蛋白质特征，分别为氨基酸序列的长度和蛋白质的重量。

其中氨基酸长度可以代表由氨基酸生成该蛋白质的困难程度，需要多少个氨基酸才能组合成该蛋白质。而蛋白质的重量一定程度上反映了生成该蛋白质消耗的元素量，体现了蛋白质重要程度。本文采用了蛋白质的序列长度和质量两项特征作为蛋白质的自有特征。

#### 结点特征嵌入汇总

在结点特征嵌入阶段，计算了中蛋白质的GAE嵌入特征、Deepwalk嵌入特征以及蛋白质的自有特征。中每个结点特征嵌入维度为82维，其计算公式如[[equ:feat:nodecat]](#equ:feat:nodecat)所示，具体分布如表[[tab:PINedgeFeatNums]](#tab:PINedgeFeatNums)所示：

其中表示第i个结点，为特征拼接函数，为第i个结点融合之后的82维特征，为基于GAE得到的16维结点特征，为基于Deepwalk得到的64维结点特征，为基于蛋白质自有数据得到的2维结点特征。

C3cmC5cmC3cm **GAE嵌入特征** & **Deepwalk嵌入特征** & **结点自有特征**  
16 & 64 & 2

### 基于多种相似性计算方法的邻边特征嵌入

除了蛋白质的结点特征，蛋白质之间也存在复杂的相互所有关系，不同的相互作用关系由于作用的蛋白质结构域不同、功能不同、发生作用的生理位置不同等，具有很大的差异性。在一般的方法中，这部分差异性仅仅映射在作用强度一个单一的数值中，在本文中，将这种丰富的作用差异作为邻边的特征完整的嵌入到图结构中，如图[[fig:edgefeat\_main]](#fig:edgefeat_main)所示。

edgefeat嵌入示意图

edgefeat嵌入示意图

[fig:edgefeat\_main]

本文为蛋白质相互作用网络的所有邻边嵌入相同维度的特征，特征中相同标号的数据具有同样的生物学意义。本文为邻边嵌入的特征包括基于蛋白质GO功能注释相似性的特征、基于蛋白质拓扑域相似性特征、基于蛋白质亚细胞定位相似性的特征和基于拓扑相似性的特征。下面分别介绍这四类相似性特征。

#### 蛋白质GO功能注释相似性的邻边特征嵌入

基因本体信息即GO注释信息是描述蛋白质功能表达的词汇表，将生物活动过程进行统一编码表示，其在生物信息学领域得到了广泛的应用。对应到蛋白质中，蛋白质分子参与的若干生物过程均可对应到GO注释表，形成对该蛋白质的综合功能描述。GO注释项总体分为三类，分别是生物过程（Biological Process）、分子功能（Molecular Function）及细胞组成（Cellular Component），生物过程描述分子功能有序组成的生命活动，分子功能描述分子在生物学上的活性，细胞组成描述了亚细胞结构及位置信息。GO注释项之间由于描述涵盖范围的不同，具有一定的从属关系，所有的GO注释项可以组成有向无环图（Directed Acyclic Graph，简称DAG）。图[[fig:go-dag]](#fig:go-dag)为QuickGO网站中获取的生殖细胞发育过程对应的GO注释DAG图。GO注释信息的从属关系主要分为两种，分别是和，其中表示包含关系，如图中黑色箭头所示，表示从属关系，如图中蓝色箭头所示。图中每一个方框展示了GO注释项及其生物功能。

生殖细胞发育对应的GO示意图

生殖细胞发育对应的GO示意图

通常情况下，蛋白质复合物中蛋白质倾向于相同的功能表达，同时功能表达相似性越高时，蛋白质产生相互作用的可能性越大。部分算法基于功能表达数据提出了连边权重更新的方法。 本文采用两种方法计算蛋白质功能相似度，将计算结果作为网络中邻边的蛋白质功能特征。

###### 基于DAG图的GO相似性计算

本方法参考了Wang等人提出的方法，给定一个GO注释项，从GO的总体DAG图中提取子图，其中是及所有的祖先结点的注释项，表示所有截取的注释项的连接关系。 定义中每一个结点对的语义值如下：

其中是和之间的权重，定义两种类型的边和权重各自为0.8和0.6。有公式可以直观得出，祖先结点点中离注释越远的注释项，其对的语义值越小，而对自身的语义值贡献为1。 定义了所有祖先结点对的语义值之后，可以直接得出的语义值得分，计算公式如下：

按照该计算方法，可以得到每一个注释项各自的语义值，在此基础上可以得到两个注释项，语义值的相似度，具体计算方法如下：

其中为和的所有公共祖先结点。由于蛋白质通常是由多个注释项组成，因此蛋白质之间的功能相似性需要考虑两个蛋白质各自注释项之间的综合结果，其具体计算方法如下：

其中，表示两个蛋白质，其中分别具有，的功能注释项。 这个方法的核心思想是将两个蛋白质的注释信息相互关系矩阵构造出来，只考虑源蛋白质的某一个注释项和目标蛋白质所有注释项最大的相关性，而蛋白质的相互作用关系是所有最大值的平均。

###### 基于频率统计的GO相似性计算

本方法参考了Lin等人提出的方法。首先计算某两个注释项之间的相关性，具体计算过程如下：

其中表示两个注释项的公共祖先，而表示某一个注释项在所有蛋白质中出现的概率。这个计算公式直观的认为，两个具有相互关联的功能注释，其关联结点越稀有，则表示两个注释项的相关性越强。计算蛋白质功能相似性的计算方法如下：

#### 蛋白质结构域相似性的邻边特征嵌入

蛋白质结构域时蛋白质具有独立功能和特异结构的区域，结构域互作信息影响蛋白质复合物的形成。结构域之间具有相互作用关系，这些相互作用关系可以从三维交互域（3did）数据库提取，并构建成结构域相互作用网络（Domain-Domain Interaction Network，简称DDI）。 单个蛋白质所具有的结构域可以映射到结构域互作网络的若干结点，成为一个结构域集群，而蛋白质域相似性可以转换为两个结构域集群的相互作用关系。 本文中两个蛋白质的结构域互作特征包括如下方面：两个蛋白质结构域交集、并集以及相互之间的差集；映射到结构域相互网络之后，两个结构域集群之间的相互作用关系数量，其具体的计算公式如下：

其中、分别时蛋白质、之间的结构域，代表结构域、在结构域相互作用网络中的作用强度，对于无权网络可以取值为1。

Huang等人在运用结构域相互作用网络预测蛋白质相互作用时，提出了使用一阶邻居和二阶邻居计算相似性的想法，其扩展方式如图[[fig:domain-second]](#fig:domain-second)所示。本文也同时补充了一阶和二阶扩展之后的蛋白质相似性特征。

DDI多阶邻居示意图

DDI多阶邻居示意图

#### 蛋白质亚细胞定位相似性的邻边特征嵌入

细胞是一个高度有序的结构，胞内根据空间分布和功能不同，可以分成不同细胞器或细胞区域，蛋白质只有转运到正确的部位才能参与细胞的各种生命活动，ZHANG等人认为蛋白质的亚细胞定位有助于研究蛋白质的生物学功能，同时对蛋白质的其他研究如相互作用、进化等也能提供必要的信息。Huang等人提出使用蛋白质复合物和亚细胞定位信息预测关键蛋白质，蛋白质复合物，蛋白质复合物的生成和功能实现和其所处的细胞位置相关，因此蛋白质的亚细胞定位能对蛋白质复合物的形成产生一定的影响。本文将两个蛋白质亚细胞定位数据的交集和并集作为蛋白质的亚细胞定位特征。图[[fig:subcell]](#fig:subcell)为苏氨酸蛋白激酶TOR1亚细胞定位图，其中黄色的部分表示其主要的注释区域，包括细胞膜和液膜。

苏氨酸蛋白激酶TOR1亚细胞定位图

苏氨酸蛋白激酶TOR1亚细胞定位图

#### 蛋白质拓扑相似性的邻边特征嵌入

蛋白质在中所处的拓扑环境也可以作为蛋白质相似性特征的计算方式。基于DPClus算法中提出的相似性计算方法，本文将结点的公共邻居数作为拓扑相似性特征。具体计算方法如下：

其中，为相互作用的蛋白质，和分别为两个蛋白质在中的邻居。

#### 邻边特征嵌入汇总

在邻边特征嵌入阶段，计算了中所有蛋白质相互作用之间的相似性特征，其中包括蛋白质功能注释特征、结构域特征、亚细胞定位特征以及网络拓扑特征。中每条相互作用边所具有的特征维度为12维，具体计算如公式[[equ:feat:edgecat]](#equ:feat:edgecat)所示，具体分布如表[[tab:PINedgeFeatNums]](#tab:PINedgeFeatNums)所示：

其中表示第i个结点和第j个结点之间的邻边，为特征拼接函数，为邻边ij融合之后的12维特征，为基于功能注释相似性得到的2维邻边特征，为基于结构域相似性得到的7维邻边特征，为基于亚细胞定位相似性得到的2维邻边特征，为一阶邻居相似性得到的1维邻边特征。

C3cmC3cmC3cmC3cm **功能注释特征** & **结构域特征** & **亚细胞定位特征** & **网络拓扑特征**  
2 & 7 & 2 & 1

### 融合结点特征和邻边特征的蛋白质相互作用网络构建

通过蛋白质相互作用数据集构造蛋白质相互作用网络，再在网络的基础上应用多种结点嵌入方法和邻边嵌入方法为网络添加结点特征和邻边特征，其工程如图[[fig:featedgraph]](#fig:featedgraph)所示，其中的小矩形代表某一个元素的特征向量，橘色为结点特征向量，蓝色的邻边特征向量。结点总特征维度为82维，邻边总特征维度为12维。

结点和邻边特征嵌入示意图

结点和邻边特征嵌入示意图

## 训练和待筛选复合物特征子图样本生成

在具有结点特征和邻边特征的蛋白质相互作用网络（）的基础上，按照复合物的对应结点抽取蛋白质复合物特征子图（），该局部图包含了复合物内部蛋白质之间的相互作用关系、相关性特征数据及部分蛋白质特征数据，因此可以将局部图作为后续复合物筛选模型的输入数据。该局部子图的示意图如下所示。

复合物子图样本的构造示意图

复合物子图样本的构造示意图

在图中，左侧为整个蛋白质相互作用网络，结点为某一个蛋白质复合物的蛋白质集合对应的结点，在获取局部子图的过程中，首先保留结点集合及其结点特征，然后保留结点集合之中的相互作用关系，如邻边等。最后形成蛋白质复合物对应的具有结点特征和邻边特征的子图。

本文正样本由酵母菌标准复合物数据集CYC2008和MIPS构成，负样本为上在一定限制条件下产生的随机子图。COACH算法作为在复合物预测方法中具有一定准确度和鲁棒性的方法，其产生的样本作为中间样本。待筛选样本为Dpclus算法、Clique算法和IPCA算法在蛋白质复合物网络中的预测结果对应的样本。

### 正样本数据：基于已知蛋白质复合物融合生成

蛋白质复合物数据集通常是由单独的研究机构或学校维护，会每隔一段时间更新生物学上已发现的蛋白质复合物。其中CYC2008数据集和MIPS数据集是针对酿酒酵母的蛋白质复合物数据集，且蛋白质复合物数量较多，更新较快，因此本文采用了CYC2008和MIPS数据集作为标准数据集。

为了训练统一的能够识别真正蛋白质复合物的模型，需要将已知的蛋白质复合物做融合，成为统一的正样本数据。由于不同的正样本数据集可能产生高度相似甚至相同的蛋白质复合物，为了保证样本的独立性，本文做了冗余数据的剔除处理。

本文以此对比了每一对蛋白质复合物并计算其邻居相似性，邻居相似性的计算方法如公式[[equ:compComplexSim:NA]](#equ:compComplexSim:NA)所示，邻居相似性代表了两个蛋白质复合物的重合程度。参考已有方法，当两个复合物邻居相似性高于0.8时，本文剔除其中结点数较低的复合物，从而保证最终的正样本集合中无相似样本，具有一定的独立性。最终形成了734个样本的蛋白质复合物数据集。

其中和是均表示蛋白质复合物，为复合物中的结点集合，为复合物中的相互作用集合。

在PPI网络中，由于标准复合物数据集和复合物相互作用数据集可能存在不一致性，部分结点标准复合物数据集存在而互作网络中不存在，为了保证标准复合物的准确性，在抽取子图的过程中，本文会舍弃其中映射到网络中出现结点缺失的复合物。

经过上述操作，为在Biogrid网络中剔除结点缺失的样本之后最终的正样本子图数据个数为621个，在DIP网络中剔除结点缺失的样本之后最终的正样本子图数据个数为371个。

### 中间样本数据：基于COACH算法生成

COACH算法是一种基于密集子图发现的蛋白质复合物预测算法，可以直接在无权网络中预测的蛋白质复合物样本。该方法算法性能优良，且在多个已知的相互作用网络数据集中预测之后，预测结果的召回率和准确率均有较高的数值，其预测的复合物具有一定的指导性，因此本文将COACH算法的预测结果作为独立的中间样本数据。中间样本数据对标准复合物和随机复合物具有一定的区分度，能对分类器识别复合物具有指导作用。COACH算法中的参数处理如下，其中DENSITY\_THRESHOLD为预测复合物密度的阈值，AFFINITY\_THRESHOLD和CLOSENESS\_THRESHOLD与子图的合并有关。

C5cmC2cm **参数名称** & **参数值**  
DENSITY\_THRESHOLD & 0.7  
AFFINITY\_THRESHOLD & 0.225  
CLOSENESS\_THRESHOLD & 0.5

一般情况下，通常复合物预测算法无法无法得到和标准集中的复合物完全重合的结果。当预测的某一个复合物其和标准复合物的最大邻居相似性大于0.25时就认为复合物预测成功。

依据基于邻居相似性的分类标准，COACH生成的样本中有部分复合物预测成功，剩余的复合物预测失败，本文对分类模型的要求时可以区分这两类样本。在可区分该两类样本的情况下，模型对其他算法产生的结果进行筛选时，才可以正确的定位到其中预测成功的样本。因此本文将COACH生成的样本分为两类。 由于邻居相似性处于0.25附近时，样本可能具有较大的迷惑性，反而影响模型的正常训练，因此剔除相似性在0.25附近的部分样本。

本文对中间样本具体做如下处理：

* 计算各个样本和标准复合物的最高邻居相似性，得到0$\thicksim$1的分数;
* 剔除其中相似性分数高于0.8的样本，防止样本与标准集重合;
* 将相似性分数在0.4$\thicksim$1的样本作为A样本;
* 将相似性分数在0$\thicksim$0.2的样本作为B样本;

经过上述处理，分别得到Biogrid网络和DIP网络运行COACH算法的中间样本数据集，其中Biogrid的A样本为192个、B样本为975个，DIP的A样本为179个、B样本为345个。

### 负样本数据：基于权重动态随机密集子图生成

随机符合物是在互作网络中随机游走产生的随机图，结合标准样本和中间样本具有紧密型子图的特点，本文对随机游走算法做了一定限制。

首先按照图结点的权重选取初始结点，视作子图。对子图周围所有邻居结点设置权重，邻居结点与子图连接数越高，则该邻居结点权重越高。按照邻居结点权重加权随机抽取下一个结点，并将该结点加入子图，如此反复进行直至子图结点个数达到阈值。按照该种随机子图生成方法，子图趋向于密集连接。由于负样本在子图结构上和正样本相近，在进行分类任务时，可以避免模型学习简单的拓扑差异，迫使模型学习复合物内在的形成规律，达到主动学习的目的。图[[fig:diffrent-random-garphs]](#fig:diffrent-random-garphs)在DIP网络中采用不同随机算法得到的随机子图差异。

0.5cm

本文在网络中按照真实样本的结点数分布抽取负样本，负样本个数为标准样本和中间样本之和。

### 待筛选样本数据：基于核心附属复合物预测算法生成

待筛选样本是用于测试模型有效性的测试样本，由已有的复合物预测算法产生。本文采用了基于核心附属的蛋白质复合物预测算法，包括Dpclus算法、Clique算法和IPCA算法，该类算法均具有较高的复合物样本产生量，其召回率的评价指标较高，具有较高的可优化空间。

分别在DIP和Biogrid蛋白质相互作用网络中使用运行复合物预测算法，其得到的预测结果统计如数据汇总与分析所示。

### 数据汇总与分析

最终产生的训练数据集如下表所示：

C3cmC2cmC3cmC3cmC2cm **网络** & **正样本数** & **中间样本数（A）** & **中间样本数（B）** & **负样本数**  
DIP数据集 & 371 & 179 & 345 & 764  
Krogan数据集 & 621 & 192 & 975 & 819

其中正样本数如[2.3.1](#subsection:allSample:positiveSampleData)所示，CYC2008数据集和MIPS数据集作为原始的正样本数据集，相似性融合之后正样本总数为734，映射到Krogan网络中完整正样本数剩余621，映射到DIP网络中完整的正样本数为371。中间样本数如[2.3.2](#subsection:allSample:middleSampleData)所示，其中对于Krogan数据集，采用无权重COACH算法预测蛋白质复合物，A类中间样本即评分高于0.4的样本数为192，B类中间样本即评分低于0.2的样本数为957。对于DIP数据集，采用 相似的方法，A类中间样本数为179，B类中间样本数为345。负样本数如[2.3.3](#subsection:allSample:negtiveSampleData)所示，Krogan数据集中随机产生的负样本数为819，DIP数据集中随机产生的负样本数为764。

最终产生的待筛选数据集如下表所示：

C3cmC2cmC3cmC3cmC2cm **网络** & **Dpclus算法** & **Clique算法** & **IPCA算法**  
DIP数据集 & 1178 & 3224 & 840  
Krogan数据集 & 1078 & 6250 & 438

其中Dpclus算法中D\_THRESHOLD参数设置为0.9，CP\_THRESHOLD参数设置为0.5；在Clique算法中使用了默认参数；在IPCA算法中T\_IN参数设置为0.5。在这些参数条件下，分别在Krogan网络和DIP网络中运行上述的复合物预测算法，可以得出对应的复合物预测结果如表中所示。

## 本章小结

本章主要提出了蛋白质复合物特征子图的生成算法，主要分为两个阶段完成了这个过程。第一阶段是特征蛋白质相互作用网络的构建，其中包括蛋白质相互作用网络构建、结点特征生成与嵌入、邻边特征生成与嵌入。提出了Deepwalk和GAE为基础的结点特征嵌入方法、提出了多种基于蛋白质相似性的邻边特征嵌入方法。后续阐述了如何在特征蛋白质相互作用网络基础上上构建了特征子图样本数据集。

# 基于图卷积神经网络的复合物筛选模型

本章介绍利用结点特征基于图卷积神经网络的复合物筛选模型，介绍了模型实现的细节和整体的模型实现流程。同时本章介绍了通用的相关的实验参数与指标。最后本章进行了相关实验并对实验结果进行了分析。

## 引言

已有的基于监督学习的复合物预测算法，基本都是基于挖掘复合物子图的拓扑特征，比如图的密度、图的结点个数、平均度数等等，将图的所有拓扑特征置于一个向量中，构成一个的向量，用该向量代替子图，最后将向量用于训练分类模型。具体过程如图[[fig:entropy-classification]](#fig:entropy-classification)所示。

基于图拓扑特征的蛋白质复合物分类模型

基于图拓扑特征的蛋白质复合物分类模型

但是蛋白质复合物中，有大部分复合物是小尺度复合物，其蛋白质个数小于5个。此类复合物的拓扑结构结构简单，提取的拓扑特征具有趋同性。随机的子图在结点数较少的情况下可能形成相同的拓扑特征，此时分类模型无法区分子图是否为真正的复合物。

蛋白质复合物网络是一个高维结构，可以通过挖掘蛋白质在网络中所处的位置和其周围相互作用关系挖掘蛋白质的潜在特征表示，将网络的高维数据转换为蛋白质的低维嵌入表示，成为每一个蛋白质初始的特征向量。

在不引入额外的生物学信息的情况下，本文提出了基于结点的全局拓扑特征的复合物局部子图分类模型。该模型仅使用网络的拓扑特征，利用图卷积神经网络在将拓扑信息与局部子图相互融合，最终得到局部子图的总体特征并做相应的分类与回归处理。

## 基于图卷积神经网络的复合物筛选模型

本节分别介绍基于结点卷积单层更新的具体实现，以及基于该单层更新方法构建的蛋白质复合物评价模型，最后介绍复合物筛选模型的总体实现流程。

### 基于结点卷积单层更新

本文在[2.2.2](#subsection:featPPINetwork:nodeFeatConstruct)部分简单介绍了图卷积神经网络。本节介绍在基于图卷积神经网络的复合物筛选模型中单一图卷积网络的计算方法。

在单一的卷积层中，结点特征的计算方法如下图所示。

结点更新过程中数据流程示意图

结点更新过程中数据流程示意图

[fig:gcn\_nodeupdate]

如图中所示，结点A的特征将有结点共同确定，图中左侧为未更新状态，右侧为更新结点特征过程中的数据流动。其具体过程可以表示为如下的公式。

其中表示结点i的所有邻居，表示结点i和结点j度乘积的平方根，其具体计算为，表示激活函数，表示卷积过程中的偏移值。

在模型中，偏移值设置为0，激活函数为LeakyRelu函数。而在卷积过程中，为了保持结点特征的连续性，每一个子图的结点均添加一条自环，从式中可以得出，添加自环之后，结点特征更新过程中会有一部分信息由自身提供，本模型中平衡系数设置为1。

上式为单个结点的更新过程，将子图中所有结点更新一轮之后即完成一层图卷积过程，图中结点特征从第l层更新到第l+1层。

### 特征子图评价模型

下面从数据流的层面阐述从特征子图数据到样本评价数据的实现流程。基于全局特征的复合物分类模型总体结构如图[[fig:gcn\_flow]](#fig:gcn_flow)所示。

基于图卷积神经网络的复合物筛选模型总体流程

基于图卷积神经网络的复合物筛选模型总体流程

[fig:gcn\_flow]

模型的初始输入特征结点特征，维度为80维，其中包括基于Deepwalk嵌入特征64维和GAE嵌入特征16维。 在初始特征之后加入了两层图卷积模型，图卷积模型如[3.2](#section:NodeConv:detail)所示，图卷积可以融合图中结点特征和拓扑结构对结点特征进行更新。图卷积过程中隐层维度设置为64维。两层图卷积之后，采用了对结点特征MaxPooling和MeanPooling的池化方法。两种池化方法的结果进行拼接之后，形成的128维的特征代表图整体特征，如图中1和2部分所示，1部分代表MaxPooling得到的64维特征，2部分代表通过MeanPooling得到的64维特征。

分类与评分部分，1和2两种Pool方法拼接作为代表子图的特征向量。 后续分别为两层感知器和softmax层得到图的分类预测，两层感知器和sigmoid层得到图的评分预测。而两项预测结果采用公式[[equ:loss]](#equ:loss)的损失函数统一起来。通过不断的优化损失函数，可以使得特征向量同时具备分类与评分特征。

分类指标和评分将作为蛋白质复合物的综合评价指标，具体计算方法如[3.3](#section:NodeConv:allExperienceDesign)中损失函数计算部分所示。

### 筛选模型算法

本节介绍基于图卷积神经网络的复合物筛选模型的具体算法，包括数据处理，模型训练，样本筛选与结果比对。 具体如算法[[alg:nodegcn-screen]](#alg:nodegcn-screen)所示。

: train protein complexes; : protein complexes before screen; : golden bench protein complexes; : protein-protein interaction network : predicted protein complexes F1/SPA metrix after and before screen; :PIN node number;; :compute node feat by GAE algorithm; :compute node feat by Deepwalk;algorithm; ; : feated subgraph extract algorithm, see in [2.3](#section:featSubNetworkConstruct:allSample); ; ; ; ; ;; ; optimization ; : feated subgraph extract algorithm, see in [2.3](#section:featSubNetworkConstruct:allSample); ;; ; ; ;

模型算法中输入数据包括本文[2.3](#section:featSubNetworkConstruct:allSample)节构造的训练数据集、待筛选数据集以及蛋白质相互作用网络图结构数据，其中训练数据集由蛋白质复合物的正样本数据集，中间样本数据集和负样本数据集构成，训练数据集中每一个蛋白质符合物样本均具有分类标注与评分标注。

模型的输出数据为待筛选数据集筛选之前的所有样本与筛选之后剩余样本的评价指标，评价指标需要参照真实复合物样本集计算，计算的指标包括F1值和综合评价指标SPA。最终的输出为筛选后F1评价指标、筛选前F1评价指标、筛选后SPA评价指标、筛选前SPA评价指标，如算法中输出部分所示。

算法中第2行代表采用图自编码器方法对蛋白质相互作用网络中每一个结点进行编码，得到编码结果为结点特征矩阵，其中n代表互作网络中结点的数量。第3行代表采用深度随机游走获取的结点特征矩阵。通过的方法将GAE特征和Deepwalk特征拼接到一起，得到80维的特征矩阵，如第4行所示。算法中第5行和第17行表示结合互作网络，特征数据和蛋白质复合物中的蛋白质集合提取取蛋白质复合物特征子图的方法，其中表示训练数据集提取的所有特征子图，为待筛选数据集提取的所有特征子图，其具体实现如[2.3](#section:featSubNetworkConstruct:allSample)所示。

第6行至第16行表示基于图卷积神经网络的蛋白质评价模型的训练过程，模型采用分批次的训练方法，同一个批次的若干子图训练损失加和作为整个批次的损失，并进行批次更新参数，如第13行和14行所示。算法中第9行和第10行表示两层基于结点的图卷积过程，该过程具体的更新方式如[3.2.1](#section:NodeConv:single)所示，模型实现细节已有阐述。第11行表示图读出过程，图读出采用所有结点隐层特征的最大池化和平均池化拼接作为子图的特征，其特征维度维，其中代表该特征子图中的结点数量。第12行表示以子图特征为基础，经过两层分类感知器得到子图的分类结果，分类结果为的矩阵，其和为1，每一个数据表示子图输出对应类别的概率；经过两层回归感知器得到子图的评分结果，评分数据为0至1，代表样本为真实复合物的概率值。第14行表示某一个复合物样本的训练损失，分为分类损失和回归损失，其中有损失平衡系数。

第18行至第22行为模型筛选以及评价阶段。第18行表示对于待筛选复合物中的每一个特征子图样本，由已经过训练的复合物评价模型得出对应待筛选样本的分类数据和评分数据。在筛选阶段，如果该样本被模型分类为真样本或者其评分高于0.25，则将该样本归类为高评分复合物，添加到筛选结果中，如算法中第19行所示。最终第21行表示筛选后与筛选前复合物样本分别与标准复合物做比对，并分别得出其F1值评价指标，第22行对应SPA值评价指标。评价指标的具体计算如[3.3](#section:NodeConv:allExperienceDesign)部分所示。

## 实验参数与评价指标

本节介绍实验相关环境、实验通用超参数设置、相关优化器以及评价指标，该部分内容适用于本文所有复合物筛选模型的实验，后续章节中不做重复阐述。

### 实验环境

本文的实验基于Linux系统Ubuntu16.04版本，处理器型号为Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz，内存大小为128GB。本文代码基于Python编写，其中深度学习框架基于Pytorch，图神经网络框架基于DGL库。

### 实验参数

由于训练数据的不平衡性，训练模型之前，数据进行了重采样处理，保持各分类训练样本相等。按照的比率分割各分类样本，形成训练集和测试集。在分类模型的训练过程中，观察训练集和测试集的损失函数，根据早停策略得到未过拟合最佳模型。 由于各项特征数据存在不平衡性，部分数据如蛋白质长度等数值过大，本文将所有的特征数据进行了统一的归一化和标准化处理。

本文采用0.001的学习率，batch大小设置为32，隐层维度为64，模型使用了两层的GCN网络以及两层的分类器。 采用LeakyRelu作为激活函数和AdamOptimizer作为模型优化器。LeakyRelu激活函数如图[[fig:leakyrelu]](#fig:leakyrelu)所示。

Leaky-Relu 激活函数示意图

Leaky-Relu 激活函数示意图

LeakyRelu激活函数会保留所有为正值的数据，相较于Relu激活函数将所有输入的负值以0输出，LeakyRelu会保留一定的负值输出。

每一个训练复合物样本都具有复合物类别以及相似性评分两项数据，针对这两项属性，本文实验中分别设置了两项损失函数，使用交叉熵损失（Cross Entropy Loss）计算复合物类别分类损失，使用BCE损失（Binary Cross Entropy Loss）计算复合物得分回归损失。由于两项损失在数值上的差异，设置了相应的损失平衡系数。具体损失计算如下所示。

其中，为预测类别，为真实类别，为预测评分，为真实评分，为损失平衡系数。

### 评价指标

模型测试阶段，本文将复合物生成算法A在网络中运行得到算法A预测的复合物，用复合物分类模型筛选所有预测结果，其中如果预测样本在筛选模型中被分类为真样本或者评分高于0.25时，预测样本即可通过筛选。最后所有通过筛选的复合物形成复合物集合。最后使用F1值和PPV指标评估和的复合物质量。 评价复合物预测结果的常用指标为精准率（precision）、召回率（recall）和F1值（F1-score）。预测复合物和标准复合物计算邻居相似性（NA-similarity）[[equ:compComplexSim:NA]](#equ:compComplexSim:NA)。

按照一般性标准，邻居相似性大于0.25时，预测复合物和标准复合物具有匹配性，表示复合物预测成功。 精准率计算如公式[[equ:precision]](#equ:precision)所示。 其中为预测复合物集合，为预测复合物中具有匹配性的所有复合物集合。精准率表示预测复合物中，预测成功的样本所占比重，衡量预测结果的“查准率”。 召回率段计算如公式[[equ:recall]](#equ:recall)所示。 其中为标准复合物集合，为标准复合物中被预测复合物匹配的集合。召回率表示标准复合物中，能被预测复合物匹配的样本所占的比重，衡量预测结果的“查全率”。 F1值综合考虑查准率和查全率，计算如公式[[equ:f1]](#equ:f1)所示。 从定义看出，当查准率和查全率中有一项指标偏小时，F1值就会偏小。

除此之外，研究者还提出了多种评价复合物预测质量的指标，包括Sn值、PPV值（Cluster-wise Positive Predictive Value）和Acc值等等。

Sn值具体计算方式图式[[equ:Sn]](#equ:Sn)所示，

其中表示所有的预测复合物，表示所有的标准复合物，表示两个复合物之间共同蛋白质的数量，表示真实复合物中蛋白质的个数。值可以反应真实复合物中蛋白质分子被匹配到个数的平均值。

PPV值具体计算方式如式[[equ:PPV]](#equ:PPV)所示，

可以看出PPV值反应了预测的复合物的匹配程度，PPV值越高，表示预测复合物和标准复合物的平均重合度越高。

Acc值综合考虑了Sn值和PPV值的结果，能更全面的反应预测复合物集合的质量，其具体计算方法如式[[equ:Acc]](#equ:Acc)所示。通常情况下，可以将Sn值、PPV值和Acc值的和作为综合的衡量指标[[equ:SPA]](#equ:SPA)。

## 实验设计及结果分析

本章提出了基于全局特征的复合物筛选模型，该模型使用蛋白质网络得到蛋白质编码，在复合物子图中使用GCN网络将蛋白质编码和子图拓扑结构进行融合。 该模型未引入额外的生物学信息，为了验证模型的有效性，本文进行了如下的对比试验。

第一个对比的模型是使用随机特征GCN图分类模型，子图GCN网络输入的结点特征为随机数据，通过该模型的对比验证全局特征的有效性。

第二个对比的模型是采用图论拓扑统计特征的分类模型[[fig:entropy-classification]](#fig:entropy-classification)。计算若干维度的图拓扑特征并进行标准化处理，作为子图特征的代表，用于后续多层感知器分类模型的输入。具体特征提取拓扑特征如表[[tab:datasets:statisticgraphfeat]](#tab:datasets:statisticgraphfeat)所示。

C3cmC2cmL8cm **拓扑特征类型** & **维数** & **具体描述**  
图总体特征 & 2 & 图结点数，图密度  
度分布特征 & 4 & 度数平均值、最大值、最小值、方差  
聚类系数特征 & 3 & 聚类系数最大值、平均值、方差  
其他特征 & 1 & 度同配型（degree assortativity）

###### 实验方案（一）

稀疏网络中基于图卷积神经网络筛选模型实验

在本文介绍的四个网络中，DIP网络的密度最低，为0.0014，其平均度数最低6.98，具有一定的稀疏性。且DIP网络的结点数量较多，为4928个，仅次于Biogrid网络，数据完整性较高。因此方案（一）选择DIP网络作为稀疏网络进行相关实验。

0.2cm

[fig:dip\_node]

C2cmC2cmC2cmC2cmC2cm **F1值** & **原始数据** & **随机GCN筛选** & **拓扑方法筛选** & **全局特征GCN筛选**  
Dpclus算法 & 0.414 & 0.352 & 0.328 & 0.434  
Clique算法 & 0.296 & 0.222 & 0.238 & 0.477  
IPCA算法 & 0.291 & 0.300 & 0.310 & 0.336

C2cmC2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **随机GCN筛选** & **拓扑方法筛选** & **全局特征GCN筛选**  
Dpclus算法 & 1.333 & 1.340 & 1.330 & 1.424  
Clique算法 & 0.658 & 0.596 & 0.656 & 0.815  
IPCA算法 & 0.794 & 0.729 & 0.744 & 0.922

[tab:result/DIP/node]

本文在DIP网络中提取了训练特征子图，其组成为正样本特征子图371个，A类中间样本179个，B类中间样本345个以及负样本764个，其具体的构造方法如[2.3](#section:featSubNetworkConstruct:allSample)所示。接下来在DIP网络中运行Dpclus算法、Clique算法和IPCA算法，分别得到了预测的蛋白质复合物结果，分别为1178、3224和840个样本，接下来分别构造了特征子图，最终形成了三个待筛选特征子图样本数据集。

实验中在训练特征子图数据集基础上训练了基于图卷积神经网络的评价模型。模型训练完成之后，分别对三个待筛选特征子图样本数据集中的样本进行了测试将其中符合要求的复合物保留下来，作为筛选后样本数据集。接下来分别对比了筛选前后样本数据集F1值和综合评价指标。

按照以上的实验方法本章对比了基于随机结点特征模型和基于图论拓扑特征的模型，随机结点特征的对比可以确定嵌入结点特征的有效性，基于图论拓扑特征的模型作为拓扑特征的Base模型，总体的实验结果如下所示。

图[[fig:dip\_node]](#fig:dip_node)为在DIP网络中，随机结点特征模型、图论拓扑特征模型以及全局特征模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/DIP/node]](#tab:result/DIP/node)为实验结果具体数据。

###### 实验方案（二）

稠密网络中基于图卷积神经网络筛选模型实验

在本文介绍的四个网络中，Biogrid网络的密度最高，为0.0038，其平均度数最低21.40，远高于其他网络，相较于其他网络Biogrid网络稠密性较高。且Biogrid网络的具有最高的结点数量，数据为5573。因此方案（二）选择Biogrid网络作为稠密网络进行相关实验。

0.2cm

[fig:biogrid\_node]

C2cmC2cmC2cmC2cmC2cm **F1值** & **原始数据** & **随机GCN筛选** & **拓扑方法筛选** & **全局特征GCN筛选**  
Dpclus算法 & 0.425 & 0.444 & 0.254 & 0.470  
Clique算法 & 0.202 & 0.216 & 0.196 & 0.372  
IPCA算法 & 0.293 & 0.123 & 0.331 & 0.365

C2cmC2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **随机GCN筛选** & **拓扑方法筛选** & **全局特征GCN筛选**  
Dpclus算法 & 1.796 & 1.187 & 1.382 & 1.871  
Clique算法 & 0.901 & 0.532 & 0.675 & 1.009  
IPCA算法 & 0.859 & 0.886 & 0.810 & 0.970

[tab:result/Biogrid/node]

本文在Biogrid网络中提取了训练特征子图，其组成为正样本特征子图621个，A类中间样本192个，B类中间样本975个以及负样本819个，其具体的构造方法如[2.3](#section:featSubNetworkConstruct:allSample)所示。接下来在DIP网络中运行Dpclus算法、Clique算法和IPCA算法，分别得到了预测的蛋白质复合物结果，分别为1078、6250和438个样本，接下来分别构造了特征子图，最终形成了三个待筛选特征子图样本数据集。 后续其他具体实验细节如方案（一）。

图[[fig:biogrid\_node]](#fig:biogrid_node)为在Biogrid网络中，随机结点特征模型、图论拓扑特征模型以及全局特征模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/Biogrid/node]](#tab:result/Biogrid/node)为实验结果具体数据。

###### 实验结论

实验方案（一）中图[[fig:dip\_node]](#fig:dip_node)显示在DIP网络中随机GCN筛选和基于拓扑GCN筛选可能导致F1值的降低和评价指标的降低，原因为简单模型在学习的过程中会直接学习直观的拓扑关系，导致分类的错误率较高，在筛选过程中会错误的剔除正样本，导致筛选之后样本质量反而降低。 实验方案（二）中图[[fig:biogrid\_node]](#fig:biogrid_node)显示全局特征的GCN筛选对Clique算法的提升十分显著，Clique算法是一种基于3-clique结构挖掘复合物的算法，该算法是一种基于密集连边发现复合物的算法，在Biogrid网络中，网络密度相较于DIP网络高了5.2，因此在Biogrid网络中，Clique算法会相应的产生更多的样本，从而为算法优化提供空间。

基于全局拓扑数据结点嵌入的复合物筛选模型，在稀疏网络和密集网络中对比的无特征方法、局部拓扑特征方法中均取得了最佳的F1值和SPA值，证明了结点嵌入的复合物筛选模型的有效性。

## 本章小结

本章基于拓扑结构出发，提出了基于全局拓扑数据嵌入的复合物筛选模型，通过GAE和Deepwalk嵌入结点特征，基于图卷积神经网络的蛋白质复合物子图评分模型将结点特征融合。本章详细阐述了单层卷积过程以及整体模型的结构与算法流程。同时本章介绍了通用性的实验环境、实验参数以及评价指标。最后本章在子图拓扑特征模型，无嵌入图卷积模型并进行了相关实验，实验结果表明了结点嵌入特征和筛选模型的有效性。

# 基于邻边卷积网络的复合物筛选模型

本章首先介绍了生物数据对复合物预测的重要性，介绍了生物数据在蛋白质相互作用网络中的邻边嵌入。同时介绍了基于邻边嵌入的相关图卷积算法，并提出了基于邻边卷积网络的复合物筛选模型。最后针对该模型进行了相关的对比实验并分析了实验结果。

## 引言

已有方法已经充分表明，在PPI网络中进行蛋白质复合物预测不仅仅是一个图论中的聚簇发现问题，更是一个信息融合问题。无论是加权网络方法、核心附属结构预测方法，都致力于最大程度的利用生物学信息进行辅助预测。

而已有的算法通常将生物学信息，包括GO功能注释、拓扑域、亚细胞定位等转换为蛋白质之间的相似性数据，即映射为邻边数据。已有的基于加权图的复合物预测算法是将邻边特征映射到固定的权重计算公式中，相较于基于无权图的算法，其预测复合物的质量有较好的提升，说明邻边数据在复合物预测领域具有一定的重要性。

然而，硬编码的过程只能将丰富的生物数据进行简单的映射，无法做到对生物数据的高维抽象。 目前该研究领域尚未出现对邻边特征进行高维非线性抽象的算法。 基于以上前提，本文提出了利用生物特征基于邻边卷积网络的复合物局部子图分类模型。

## 邻边卷积网络介绍

邻边卷积网络（EdgeBase Convolution）由点云学习领域提出，是一种基于邻边做特征转换以及特征聚合的方法。点云学习指的是使用计算机的方法学习与分析空间中独立数据点的集合，数据点的信息包括坐标、颜色、分类、强度值等等信息，数据点在空间中的集合状态即为点云。点云学习的任务是对点云数据做分割、物体识别等等。

在点云任务中，会与待更新结点i周围的K个最近邻居建立连边，基于连边做特征更新，然后所有的连边特征汇聚到一起作为结点的新特征。连边特征更新的过程中会考虑源点与汇点的信息流动，其具体的过程如图[[fig:edgeconv\_main]](#fig:edgeconv_main)所示。

基于邻边的图卷积神经网络更新示意图

基于邻边的图卷积神经网络更新示意图

[fig:edgeconv\_main]

对于结点i的更新，需要从结点i周围的5个邻居建立到结点i的邻边。对于其中某一条邻边，其两个端点分别具有各自的特征，将特征拼接到一起，然后经过邻边卷积的多层感知器计算，即可得到邻边特征更新后的结果如图中左侧所示。 更新结点i周围所有的邻边特征，然后计算其和值，就形成了结点i更新之后的特征，如图[[fig:edgeconv\_main]](#fig:edgeconv_main)中右侧所示。

## 基于邻边卷积网络的复合物筛选模型

本节分别介绍基于邻边卷积单层更新的具体实现，以及基于该单层更新方法构建的蛋白质复合物评价模型，最后介绍复合物筛选模型的总体实现流程。

### 基于邻边卷积单层更新

在通常的图卷积过程中，结点A的特征为周围所有结点的特征和自生特征决定，可以采取平均值的计算方法，如公式[[equ:NormalGCNNodeFlow]](#equ:NormalGCNNodeFlow)所示。

其中代表A结点的特征，代表A结点所有邻居的特征，为信息汇聚之后的特征更新函数。

在基于邻边卷积的复合物预测方法中，结点特征更新由其邻边共同决定。其具体计算过程如公式[[equ:EdgetoNodeFeat]](#equ:EdgetoNodeFeat)所示。

其中表示所有连接到A的邻边特征。其具体的汇聚过程如图[[fig:edge-feat-flow]](#fig:edge-feat-flow)所示，A结点汇聚周围所有邻边特征，包括、、以及。

边特征初始化结点特征示意图

边特征初始化结点特征示意图

蛋白质生物数据是基于邻边的数据，因此在对复合物子图做图卷积的时候，基于生物特征的模型采用了基于edge的GCN模型，其具体的邻边更新方式如式[[equ:EdgeConv]](#equ:EdgeConv)所示。

其中表示第层邻边更新之后的特征数据，表示从邻居结点到结点的数据流，为结点保留信息，和分别对应更新函数。为邻边特征激活函数。

结点更新的方式如式[[equ:EdgeConv-node]](#equ:EdgeConv-node)所示，

式中为第第层结点j更新之后的特征数据，表示结点的所有邻居，表示邻边数量，表示某一条邻边更新之后的特征。

每一轮基于edgeConv的卷积过程可以图[[fig:edgeconv\_edgeconv]](#fig:edgeconv_edgeconv)所示，结合公式和图片，可以得出基于edgeConv的卷积过程首先是由结点更新邻边特征，如图中上半部分所示，在子图中每一条邻边特征会获取其两个端点（源点和汇点）的特征，第一部分更新为源点至汇点的特征相减并做特征转换，代表邻边的数据流，第二部分为源点特征做特征转换，这两部分特征结合到一起经过激活形成邻边更新的特征。然后是用更新之后的邻边特征聚合物新的结点特征，如图中下半部分所示，结点会汇聚其所有邻边的特征并取平均值，同时依据最大值池化结点会保留一部分自身特征。

总体EdgeConv图卷积过程示意图

总体EdgeConv图卷积过程示意图

[fig:edgeconv\_edgeconv]

### 特征子图评价模型

本文处理了大量的生物学数据，包括蛋白质功能注释数据、结构域相互作用数据、亚细胞定位数据等，应用相关的生物领域的处理方法将生物学数据转换为可应用的特征数据，最终得到了11维度的特征数据。所有的特征都描述蛋白质相互作用，因此在图结构中，这些特征编码为图结构邻边的特征。 基于邻边卷积特征的复合物筛选模型总体流程如图[[fig:edgeconv\_flow]](#fig:edgeconv_flow)所示。

第一阶段如标号1所示，需要将初始的邻边特征以汇聚方法初始化结点特征，如图中上半部分所示。该汇聚方法为初始结点特征由其周围所有邻边特征的平均值替代。其具体的计算过程如公式[[equ:EdgetoNodeFeat]](#equ:EdgetoNodeFeat)所示，其中激活函数为直接映射。其过程可被视为聚合阶段的基于边卷积神经网络过程，如图中[[fig:edgeconv\_edgeconv]](#fig:edgeconv_edgeconv)下半部分所示。

第二阶段如标号2所示，需要基于邻边卷积将特征与拓扑结构融合，基于pool方法得到子图结构的特征表示，并基于子图特征做分类预测以及评分预测。其具体过程如图中下半部分所示。

模型使用了两层的基于邻边卷积的图神经网络，图卷积过程中隐层维度设置为64维。 两层图卷积之后，采用了对结点特征分别做最大值池化和平均值池化的方法，拼接到一起得到128维的子图特征。后续分别为两层感知器和softmax层得到图的分类预测，两层感知器和sigmoid层得到图的评分预测。

基于邻边卷积特征的复合物筛选模型总体流程

基于邻边卷积特征的复合物筛选模型总体流程

[fig:edgeconv\_flow]

### 筛选模型算法

本节介绍基于邻边卷积的复合物筛选模型的具体算法，包括数据处理，模型训练，样本筛选与结果比对。 具体如算法[[alg:edgegcn-screen]](#alg:edgegcn-screen)所示。

: train protein complexes; : protein complexes before screen; : golden bench protein complexes; : protein-protein interaction network : predicted protein complexes F1/SPA metrix after and before screen; :PIN node number;:PIN edge number;; ; ; : feated subgraph extract algorithm; ; ; ; ;; ; optimization ; : feated subgraph extract algorithm, see in [2.3](#section:featSubNetworkConstruct:allSample); ;; ; ; ;

模型算法中输入数据包括本文[2.3](#section:featSubNetworkConstruct:allSample)节构造的训练数据集、待筛选数据集以及蛋白质相互作用网络图结构数据，其中训练数据集由蛋白质复合物的正样本数据集，中间样本数据集和负样本数据集构成，训练数据集中每一个蛋白质符合物样本均具有分类标注与评分标注。

模型的输出数据为待筛选数据集筛选之前的所有样本与筛选之后剩余样本的评价指标，评价指标需要参照真实复合物样本集计算，计算的指标包括F1值和综合评价指标SPA。最终的输出为筛选后F1评价指标、筛选前F1评价指标、筛选后SPA评价指标、筛选前SPA评价指标，如算法中输出部分所示。

算法中第2行代表采用多种蛋白质相似性特征计算方法嵌入的邻边特征，其中包括基于GO相似性得到的2维特征，基于蛋白质拓扑域相似性得到的7维特征以及基于蛋白质亚细胞定位得到的2维特征，这些特征具体的计算方法如[2.2.3](#subsection:featPPINetwork:edgeFeatConstruct)所示。通过的方法将三种蛋白质相似性特征拼接到一起，得到11维的特征矩阵，如第3行所示，其中m表示的是蛋白质相互作用网络中的所有连边数量。算法中第4行和第17行表示结合互作网络，特征数据和蛋白质复合物中的蛋白质集合提取取蛋白质复合物特征子图的方法，其中表示训练数据集提取的所有特征子图，为待筛选数据集提取的所有特征子图，其具体实现如[2.3](#section:featSubNetworkConstruct:allSample)所示。

第5行至第16行表示基于邻边卷积的蛋白质评价模型的训练过程，模型采用分批次的训练方法，同一个批次的若干子图训练损失加和作为整个批次的损失，并进行批次更新参数，如第13行和14行所示。算法中第8行为基于邻边特征转换为结点特征的算法，其具体过程为遍历子图中的每一个结点，将结点特征初始化为其所有邻边的特征和，具体计算方法如[[equ:EdgetoNodeFeat]](#equ:EdgetoNodeFeat)所示，然后基于一层神经网络进行特征转换得到结点初始化特征，该神经网络的参数为。 第9行和第10行表示两层基于邻边卷积过程。对于单层邻边卷积神经网络，其更新过程为邻边特征更新与结点特征汇集，邻边特征更新为两个方面的和，分别为数据流更新和源点特征更新，结点特征汇集包括为两个方面的最大值，分别为自身特征和邻边特征平均值，其更新的具体过程如[4.3.1](#subsection:EdgeConv:single)所示。由于数据维度的不同，更新网络的参数规模为，更新网络的参数规模为。

第11行表示图读出过程，由于在该模型中邻边特征初始化为了结点特征，同时结点特征经过两层邻边卷积神经网络已经和子图拓扑结构进行了融合，此时结点特征具有较强的表达能力，隐层隐层图读出采用所有结点隐层特征的最大池化和平均池化拼接作为子图的特征。

第12行表示以基于子图特征的分类预测与评分预测。 第18行至第22行表示基于以训练模型筛选蛋白质复合物样本以及评价阶段。其具体细节如算法[[alg:nodegcn-screen]](#alg:nodegcn-screen)描述所述。

## 实验设计及结果分析

为了验证生物特征的添加对算法模型的提升，本文对比了在不添加生物数据，而使用相同的GCN结构的情况下，复合物分类模型对样本质量的提升程度。

###### 实验方案（一）

稀疏网络中基于邻边卷积筛选模型实验

在本文介绍的四个网络中，DIP网络的密度最低，为0.0014，其平均度数最低6.98，具有一定的稀疏性。且DIP网络的结点数量较多，为4928个，仅次于Biogrid网络，数据完整性较高。因此方案（一）选择DIP网络作为稀疏网络进行相关实验。

0.2cm

[fig:dip\_edge]

C2cmC2cmC2cmC2cm **F1值** & **原始数据** & **随机GCN筛选** &**邻边特征GCN筛选**  
Dpclus算法 & 0.414 & 0.352 & 0.484  
Clique算法 & 0.296 & 0.222 & 0.453  
IPCA算法 & 0.291 & 0.300 & 0.420

C2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **随机GCN筛选** & **邻边特征GCN筛选**  
Dpclus算法 & 1.333 & 1.340 & 1.490  
Clique算法 & 0.658 & 0.596 & 0.731  
IPCA算法 & 0.794 & 0.729 & 0.838

[tab:result/DIP/edge]

依据[3.4](#section:NodeConv:experience)部分中实验方案（一）中提到的方法，针对DIP网络构造了具有四类分类的训练特征子图数据集，基于三种方法生成了待筛选复合物数据集。实验中在训练特征子图数据集基础上训练了基于邻边卷积网络的评价模型。模型训练完成之后，分别对三个待筛选特征子图样本数据集中的样本进行了测试并将其中符合要求的复合物保留下来，作为筛选后样本数据集。接下来分别对比了筛选前后样本数据集F1值和综合评价指标。

按照以上的实验方法本章对比了基于随机邻边特征的邻边卷积模型，随机邻边特征模型的对比可以确定嵌入的邻边相似性特征的有效性，本章总体的实验结果如下所示。

图[[fig:dip\_edge]](#fig:dip_edge)为在DIP网络中，随机特征模型和基于生物特征的模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/DIP/edge]](#tab:result/DIP/edge)为实验结果具体数据。

###### 实验方案（二）

稠密网络中基于邻边卷积筛选模型实验

在本文介绍的四个网络中，Biogrid网络的密度最高，为0.0038，其平均度数最低21.40，远高于其他网络，相较于其他网络Biogrid网络稠密性较高。且Biogrid网络的具有最高的结点数量，数据为5573。因此方案（二）选择Biogrid网络作为稠密网络进行相关实验。

0.2cm

[fig:biogrid\_edge]

C2cmC2cmC2cmC2cm **F1值** & **原始数据** & **随机GCN筛选** &**邻边特征GCN筛选**  
Dpclus算法 & 0.425 & 0.445 & 0.529  
Clique算法 & 0.203 & 0.216 & 0.510  
IPCA算法 & 0.294 & 0.122 & 0.373

C2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **随机GCN筛选** & **邻边特征GCN筛选**  
Dpclus算法 & 1.796 & 1.119 & 1.950  
Clique算法 & 0.901 & 0.532 & 1.011  
IPCA算法 & 0.859 & 0.886 & 0.891

[tab:result/Biogrid/edge]

依据[3.4](#section:NodeConv:experience)部分中实验方案（二）中提到的方法，针对Bigrid网络构造了具有四类分类的训练特征子图数据集，基于三种方法生成了待筛选复合物数据集。实验中训练了基于邻边卷积网络的评价模型，并分别对三个待筛选特征子图样本数据集中的样本进行了测试，获取了筛选后样本数据集，最后对比了筛选前后样本数据集F1值和综合评价指标。

按照以上的实验方法本章对比了基于随机邻边特征的邻边卷积模型，随机邻边特征模型的对比可以确定嵌入的邻边相似性特征的有效性，本章总体的实验结果如下所示。 图[[fig:biogrid\_edge]](#fig:biogrid_edge)为在Biogrid网络中，随机特征模型和基于生物特征的模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/Biogrid/edge]](#tab:result/Biogrid/edge)为实验结果具体数据。

###### 实验结论

从实验方案（一）从图中可以看出，添加了生物特征后，在DIP网络中复合物的生成质量得到了相应的提升，同样的由于CLique算法的特性，该算法的前后对比提升幅度最为明显。 和实验方案（二）从图[[fig:biogrid\_edge]](#fig:biogrid_edge)中可以看出添加了生物特征后，Biogrid网络相应样本的F1值均有较为明显的提升，而Clique算法的样本提升最为明显。综合评价指标SPA值也具有一定的提升。

基于相似性数据邻边嵌入的复合物筛选模型，对比无相似性嵌入模型取得较高的F1值和SPA值提升，稠密网络具有更多的邻边信息，在稠密网络中其提升幅度更为明显，最高的F1值提升在Biogrid的Clique算法中，从0.203提升至0.510。

## 本章小结

本章基于蛋白质相似性特征出发，提出了使用基于邻边卷积的复合物特征子图评价模型以及相应的筛选算法。通过多种蛋白质相似性计算方法为蛋白质复合物子图嵌入11维度的邻边特征，提出了邻边特征初始化结点特征的方法以及蛋白质复合物中基于邻边卷积的信息流更新过程。最后本章对比了基础的无特征邻边卷积模型，实验结果表明了生物相似性特征和筛选模型的有效性以及邻边卷积对特征数据具有融合能力。

# 基于点边消息传递网络的复合物筛选模型

本章总结了基于结点和基于邻边卷积网络各自的特点，提出了融合特征的重要性。介绍了适合于融合特征的图卷积神经网络算法以及相关消息传递机制的实现。提出了基于点边消息传递网络的复合物筛选模型，并进行了相关的实验与分析。

## 引言

模型[3.1](#section:NodeConv:Put)和模型[4.2](#section:EdgeConv:intro)分别是结点嵌入和邻边嵌入模型，蛋白质复合物的形成和蛋白质特征以及蛋白质互作特征均具有相关性。 无论普通的GCN模型还是EdgeConv模型，均只能直接基于结点特征进行卷积运算，或者邻边特征转换为结点特征之后再进行图卷积运算，这个过程将邻边特征均匀的分配给了其相邻的结点，在模型训练过程邻边作为独立的计算元素依旧缺乏针对邻边本身的更新。这种情形下，邻边特征只能作为结点特征的补充，无法与复合物互作网络动态的结合起来。

为了将邻边特征与结点特征整合到模型的动态参数更新中，本节基于消息传递网络（Message Passing Neural Network，简称MPNN）提出蛋白质复合物特征融合模型，同时处理蛋白质复合物子图中的结点数据和邻边数据，将生物数据、全局拓扑特征以及局部拓扑特征动态的融合起来。

## 消息传递网络介绍

MPNN网络是一种基于消息传递的图结构学习框架。不同于图卷积模型必须将特征绑定到结点，MPNN将图结构的特征当成消息，特征在结构中的流动被视为消息的传递。在此框架下，图卷积神经网络可以被当作MPNN的特例，是只传递结点消息的MPNN网络。

MPNN网络的具体实现如式[[equ:MPNNPassing]](#equ:MPNNPassing)所示。

其中表示图中结点的邻居，为时间步。公式[[equ:MPNNPassing]](#equ:MPNNPassing)表示消息传递阶段（Message Passing），表示消息传递的更新函数，代表结点w向结点v传递消息的过程中，其传递的消息的更新方式。公式[[equ:MPNNReadout]](#equ:MPNNReadout)表示更新阶段，表示更新阶段的函数，代表结点v汇总所有周围消息后的数据更新方式。

## 基于点边消息传递网络的复合物筛选模型

本节分别介绍基于点边消息传递网络单层更新的具体实现，以及基于该单层更新方法构建的蛋白质复合物评价模型，最后介绍复合物筛选模型的总体实现流程。

### 基于点边消息传递网络单层更新

基于点边消息传递网络的复合物筛选模型是基于MPNN网络的改进。 为了使得结点特征和邻边特征具有融合的能力，蛋白质复合物特征融合模型中采用了结点更新邻边，邻边更新结点的方式。在每一层的MPNN过程中，结点t时刻特征更新为邻边t+1时刻特征、邻边t特征更新为结点t+1时刻特征。

结点特征的更新具体方法如下所示。

式中为一层感知器，表示邻边特征的自我更新，为激活函数，本文采用了LeakyRelu函数。式[[equ:MineMPNNne]](#equ:MineMPNNne)表明了从邻边均值到结点特征的更新过程，为从邻边更新过来的邻边特征，式[[equ:MineMPNNnup]](#equ:MineMPNNnup)表示结点特征结合邻边特征的更新。 可以看出，结点的特征更新来源分为两部分，其一为汇聚所有邻边的特征，同时为了保留一部分结点中具有关键性的原始特征，其更新方式采用了maxPooling。

邻边特征的具体更新方法如下所示。

式中和均为感知器，为对源点到汇点特征流的更新，为源点特征更新，表示依据端点更新的邻边特征。式[[equ:MineMPNNeup]](#equ:MineMPNNeup)表示邻边特征结合结点特征的更新。 从式中得出，类似于结点更新模型，邻边的特征更新也可以分为两部分，分别为邻边自有特征和经过结点特征转换更新的邻边特征，最后同样采用了maxPooling的方式保留关键邻边特征。经过结点特征转换更新的方式综合考虑了源点到汇点的特征流动部分以及汇点的特征保留部分。

### 特征子图评价模型

子图数据保留GO注释特征、拓扑域特征以及亚细胞定位特征等邻边特征，保留Deepwalk特征、GAE特征等结点特征。形成兼具结点和邻边特征的复合物特征子图数据，其中结点特征维度为82维，邻边特征维度维12维。

基于特征融合的分类模型总体流程

基于特征融合的分类模型总体流程

[fig:mpnn\_flow]

在模型中，添加了两层结点和邻边同时更新的MPNN网络，如图[[fig:mpnn\_flow]](#fig:mpnn_flow)所示。每一层MPNN网络中，结点特征更新由结点以及结点周围邻边特征共同确定，如示意图中的结点特征矩阵隐层的数据流所示，过程3是表示通过结点的邻边特征汇聚到一起，再结合过程4结点本身继承上一步的特征，其具体的计算方法如公式[[equ:MineMPNNnup]](#equ:MineMPNNnup)所示。邻边特征更新由邻边以及邻边两侧端点的结点特征共同确定，如示意图中的邻边特征矩阵隐层的数据流所示，过程2表示邻边特征由其源点和汇点特征共同计算，过程1表示邻边特征继承上一步的特征，其具体的计算方法如公式[[equ:MineMPNNeup]](#equ:MineMPNNeup)所示。

消息传递过程中隐层维度设置为64维，两层消息传递之后，采用了对结点特征MaxPooling和MeanPooling的池化方法以及对邻边MaxPooling和MeanPooling的池化方法。一共可以得出四个池化结果，如图中concat到一起之后的四色特征所示，对结果进行拼接之后，形成的256维的特征代表图整体特征。后续分别为两层感知器和softmax层得到图的分类预测，两层感知器和sigmoid层得到图的评分预测。感知器的隐层维度均为64维。

### 筛选模型算法

本节介绍基于点边消息传递网络的复合物筛选模型的具体算法，包括数据处理，模型训练，样本筛选与结果比对。 具体如算法[[alg:mpnngcn-screen]](#alg:mpnngcn-screen)所示。

: train protein complexes; : protein complexes before screen; : golden bench protein complexes; : protein-protein interaction network : predicted protein complexes F1/SPA metrix after and before screen; :PIN node number;:PIN edge number;; ; ; : feated subgraph extract algorithm; ;  ; ; ; ;; ; ;; ; optimization ; : feated subgraph extract algorithm, see in [2.3](#section:featSubNetworkConstruct:allSample); ;; ; ; ;

模型算法中输入数据包括本文[2.3](#section:featSubNetworkConstruct:allSample)节构造的训练数据集、待筛选数据集以及蛋白质相互作用网络图结构数据，其中训练数据集由蛋白质复合物的正样本数据集，中间样本数据集和负样本数据集构成，训练数据集中每一个蛋白质符合物样本均具有分类标注与评分标注。

模型的输出数据为待筛选数据集筛选之前的所有样本与筛选之后剩余样本的评价指标，评价指标需要参照真实复合物样本集计算，计算的指标包括F1值和综合评价指标SPA。最终的输出为筛选后F1评价指标、筛选前F1评价指标、筛选后SPA评价指标、筛选前SPA评价指标，如算法中输出部分所示。

算法中第2行代表基于蛋白质特征嵌入方法的结点特征，其中包括基于图自编码器的16维特征、基于深度随机游走的64维特征和基于蛋白质自有特征嵌入的2维特征，这些特征具体的计算方法如[2.2.2](#subsection:featPPINetwork:nodeFeatConstruct)所示,最后经过Concat处理拼接为82维的结点特征矩阵，其中n表示蛋白质相互作用网络中所有结点的数量。第3行代表基于多种蛋白质相似性嵌入方法的邻边特征，其中包括基于GO相似性得到的2维特征，基于蛋白质拓扑域相似性得到的7维特征、基于蛋白质亚细胞定位得到的2维特征和基于拓扑相似性的1维特征，这些特征具体的计算方法如[2.2.3](#subsection:featPPINetwork:edgeFeatConstruct)所示，最后经过Concat处理拼接为12维的邻边特征矩阵，其中m表示蛋白质相互作用网络中的所有连边数量。

算法中第4行和第17行表示结合互作网络，特征数据和蛋白质复合物中的蛋白质集合提取取蛋白质复合物特征子图的方法，其中表示训练数据集提取的所有特征子图，为待筛选数据集提取的所有特征子图，其具体实现如[2.3](#section:featSubNetworkConstruct:allSample)所示。

第7行表示特征的初始化，由于结点初始特征和邻边初始特征维度不匹配，为了结点和邻边之间能够进行特征融合，首先需要将其特征统一的转换为64维。 第9行表示基于邻边特征更新结点特征的过程，其中邻边特征进行特征转换与激活，结点特征为其所有邻边特征的均值。 第10行表示基于结点特征更新邻边特征的过程，邻边特征由其两个端点的特征共同确定，分别汇点特征与源点特征形成的数据流以及源点特征自生的更新。 第11行表示特征更新过程中会基于最大值池化保留一部分结点原有特征和邻边原有特征。 算法中第9至11行共同确定了单层的点边消息传递网络更新方式，其具体计算过程如[5.3.1](#section:MPNN:update)所示。

算法中第13行表示图读出过程，由于在该模型中邻边特征初始化为了结点特征，同时结点特征经过两层邻边卷积神经网络已经和子图拓扑结构进行了融合，此时结点特征具有较强的表达能力，隐层隐层图读出采用所有结点隐层特征的最大池化和平均池化拼接作为子图的特征。

最后第14行表示以基于子图特征的分类预测与评分预测。 第20行至第24行表示基于以训练模型筛选蛋白质复合物样本以及评价阶段。其具体细节如算法[[alg:nodegcn-screen]](#alg:nodegcn-screen)描述所述。

## 实验设计及结果分析

为了验证基于点边消息传递网络的模型的提升，本文对比了前两章分别设计的基于图卷积神经网络的复合物筛选模型和基于邻边卷积网络的复合物筛选模型。

基于图卷积神经网络的复合物筛选模型是基于研究蛋白质相互作用网络的拓扑结构展开，利用了结点的拓扑特征并基于结点的图卷积神经网络对拓扑特征进行融合，一定程度上体现了网络中结点嵌入对复合物预测的作用，证实了复合物中蛋白质在互作网络中相关性对复合物的形成具有一定的影响。

基于邻边卷积网络的复合物筛选模型是基于研究蛋白质之间相似性特征展开，利用了邻边的相似性特征并基于邻边卷积对数据进行融合。生物的相似性数据对复合物预测以及复合物分类具有较明显的提升。

消息传递模型将PIN网络全局特征以及生物特征融合到一起，分别以结点和邻边的形式嵌入到复合物子图中。相较于单独利用其中一方面的模型具有更好的预测能力。为了验证模型的有效性，本章进行了如下的实验。

###### 实验方案（一）

稀疏网络中基于消息传递筛选模型实验

在本文介绍的四个网络中，DIP网络的密度最低，为0.0014，其平均度数最低6.98，具有一定的稀疏性。且DIP网络的结点数量较多，为4928个，仅次于Biogrid网络，数据完整性较高。因此方案（一）选择DIP网络作为稀疏网络进行相关实验。

0.2cm

[fig:dip\_fusion]

C2cmC2cmC2cmC2cmC2cm **F1值** & **原始数据** & **全局特征GCN筛选** & **邻边特征GCN筛选** & **融合模型筛选**  
Dpclus算法 & 0.414 & 0.434 & 0.484 & 0.501  
Clique算法 & 0.296 & 0.477 & 0.453 & 0.487  
IPCA算法 & 0.291 & 0.336 & 0.420 & 0.408

C2cmC2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **全局特征GCN筛选** & **邻边特征GCN筛选** & **融合模型筛选**  
Dpclus算法 & 1.333 & 1.425 & 1.490 & 1.498  
Clique算法 & 0.658 & 0.815 & 0.731 & 0.729  
IPCA算法 & 0.794 & 0.922 & 0.838 & 0.848

[tab:result/DIP/fusion]

依据[3.4](#section:NodeConv:experience)部分中实验方案（一）中提到的方法，针对DIP网络构造了具有四类分类的训练特征子图数据集，基于三种方法生成了待点边消息传递网络的评价模型。模型训练完成之后，分别对三个待筛选特征子图样本数据集中的样本进行了测试并将其中符合要求的复合物保留下来，作为筛选后样本数据集。接下来分别对比了筛选前后样本数据集F1值和综合评价指标。

按照以上的实验方法本章对比了第三章提出的基于图卷积神经网络的复合物筛选模型和第四章提出的基于邻边卷积的蛋白质复合物筛选模型，基于消息传递的模型保留了全局结点特征嵌入和相似性特征嵌入，融合模型具有更精确的蛋白质复合物的评价能力，本章总体的实验结果如下所示。 图[[fig:dip\_fusion]](#fig:dip_fusion)为在DIP网络中，基于结点的全局特征模型、基于邻边的生物特征模型以及特征融合模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/DIP/fusion]](#tab:result/DIP/fusion)为实验结果具体数据。

###### 实验方案（二）

稠密网络中基于消息传递筛选模型实验

在本文介绍的四个网络中，Biogrid网络的密度最高，为0.0038，其平均度数最低21.40，远高于其他网络，相较于其他网络Biogrid网络稠密性较高。且Biogrid网络的具有最高的结点数量，数据为5573。因此方案（二）选择Biogrid网络作为稠密网络进行相关实验。

0.2cm

[fig:biogrid\_fusion]

C2cmC2cmC2cmC2cmC2cm **F1值** & **原始数据** & **全局特征GCN筛选** & **邻边特征GCN筛选** & **融合模型筛选**  
Dpclus算法 & 0.425 & 0.470 & 0.529 & 0.521  
Clique算法 & 0.203 & 0.373 & 0.510 & 0.542  
IPCA算法 & 0.294 & 0.366 & 0.373 & 0.364

C2cmC2cmC2cmC2cmC2cm **SPA值** & **原始数据** & **全局特征GCN筛选** & **邻边特征GCN筛选** & **融合模型筛选**  
Dpclus算法 & 1.796 & 1.871 & 1.950 & 1.948  
Clique算法 & 0.901 & 1.009 & 1.011 & 1.041  
IPCA算法 & 0.859 & 0.970 & 0.891 & 0.834

[tab:result/Biogrid/fusion]

依据[3.4](#section:NodeConv:experience)部分中实验方案（二）中提到的方法，针对Bigrid网络构造了具有四类分类的训练特征子图数据集，基于三种方法生成了待筛选复合物数据集。实验中训练了基于邻边卷积网络的评价模型，并分别对三个待筛选特征子图样本数据集中的样本进行了测试，获取了筛选后样本数据集，最后对比了筛选前后样本数据集F1值和综合评价指标。本章对比了基于图卷积神经网络的复合物筛选模型和基于邻边卷积的蛋白质复合物筛选模型，本章总体的实验结果如下所示。

图[[fig:biogrid\_fusion]](#fig:biogrid_fusion)为在Biogrid网络中，基于结点的全局特征模型、基于邻边的生物特征模型以及特征融合模型筛选之后结果的对比，包含了F1值对比图和SPA值对比图。表[[tab:result/Biogrid/fusion]](#tab:result/Biogrid/fusion)为实验结果具体数据。

###### 实验结论

从实验方案（一）从图[[fig:dip\_fusion]](#fig:dip_fusion)中可以看出，在DIP网络中，MPNN融合模型的F1指标在Dpclus和Clique的数据样本中均得到了最优结果，SPA指标在Clique和IPCA算法中有一定降低，而Dpclus算法中SPA指标达到了最优。 和实验方案（二）从图[[fig:biogrid\_fusion]](#fig:biogrid_fusion)中可以看出，在Clique样本集上，其SPA指标和F1值均达到了最优结果，在Dpclus的样本集中SPA指标也得到了最优结果。

基于结点和邻边特征融合的复合物筛选模型，对比基于结点特征的图卷积模型和基于邻边特征的邻边卷积模型，其模型更为复杂，筛选的结果中准确率更高。模型在Dpclus和Clique算法中均取得了一定的评价指标效果提升，结果表明了融合特征模型的有效性。

## 本章小结

本章基于特征融合出发，探讨了融合全局特征以及生物相似性特征的情况下，模型设计以及进行了相关的实验。 在特征子图具有GAE和Deepwalk特征等结点特征、多种蛋白质相似性嵌入邻居特征的情况下，本章提出了改进的MPNN更新方法。提出了结点与邻边交替融合更新的模型，详细的阐述了特征子图中交替更新的计算过程以及整体的模型算法。最后本章对比了利用结点特征的基于图卷积的模型，利用相似性特征的基于邻边卷积的模型，实验结果表明融合方法对于数据集中的结果具有一定的提升效果。

# 结论与展望

## 总结

本文总结了现有蛋白质复合物预测方法的局限性，包括无法充分利用蛋白质复合物的先验知识，不能充分融合蛋白质特征和蛋白质相互作用特征，预测结果准确度较低等等。针对现有方法的不足，本文提出了蛋白质复合物筛选模型，利用已知的蛋白质复合物和蛋白质复合物样本生成算法构成特征子图数据集，训练蛋白质复合物评价模型，对待筛选复合物预测结果进行评分与过滤，从而提高预测结果评价指标。

本文具体完成了如下的研究。

1）本文构建了具有结点特征和邻边特征的蛋白质相互作用网络，即特征互作网络。在蛋白质相互作用网络中，基于图自编码器和深度随机游走嵌入结点特征，表达了蛋白质的全局拓扑信息；基于蛋白质生物数据的相似性挖掘算法嵌入多维度邻边特征，表达了蛋白质功能特性。

基于特征互作网络，本文提出了特征子图的概念并构造了蛋白质复合物特征子图数据集。为了有效的学习蛋白质复合物的生成规律，本文结合特征互作网络将蛋白质复合物原始样本映射为对应的特征子图，从而将不可学习的蛋白质集合数据转换为可学习的结构化数据。本文构建了特征子图训练数据集和待筛选数据集，其中训练数据集包括基于已知但蛋白质融合的正样本数据集、基于COACH算法生成的中间样本数据集、基于受限随机游走的负样本数据集，待筛选数据集为基于核心附属蛋白质复合物预测算法的待筛选样本数据集。

2）本文提出了基于图卷积神经网络的复合物筛选模型。模型从拓扑数据出发，基于图卷积神经网络的方法提取了深层次的蛋白质结点特征。基于图自编码器和深度随机游走算法，该模型在蛋白质互作网络中嵌入具有全局特性的结点特征。在蛋白质特征子图中，基于图神经网络的方法将结点特征进行动态融合与嵌入转换。实验阶段，该模型与无全局特征的模型、基于子图拓扑统计特征模型进行了对比与分析，表明融合全局特征的图卷积神经网络方法的有效性。

3）本文提出了基于邻边卷积的复合物筛选模型。该模型充分的嵌入了生物数据，基于邻边卷积算法融合生物数据，解决了已有算法中蛋白质复合物生物数据抽象程度较低的问题。本章在蛋白质互作网络的邻边中嵌入了三种蛋白质关联相似性特征，包括GO注释特征、拓扑域特征和亚细胞定位特征。在实验中该模型对比了无邻边特征情况下的邻边卷积复合物筛选模型，实验结果表明生物数据邻边嵌入以及邻边卷积融合方法的有效性。

4）本文提出了基于点边消息传递网络的复合物筛选模型。针对蛋白质全局拓扑特征和蛋白质生物相似性特征的融合，本文提出了基于点边消息传递网络的复合物筛选模型。该模型采用消息传递的更新方法在复合物特征子图中交替更新结点特征和邻边特征，实现了全局拓扑特征与生物功能特征的动态更新和深度融合。实验中该模型对比了基于图卷积神经网络筛选模型和基于邻边卷积筛选模型，在多个对比实验中点边消息传递模型对评价指标的提升达到了最优，表明特征融合模型能更好的挖掘蛋白质复合物的形成规律。

## 展望

本文提出了特征蛋白质互作网络构建，复合物特征样本集集以蛋白质复合物的评分模型。在此基础上，还有一部分研究可以进一步展开。

1）动态蛋白质复合物子图样本生成。细胞是构成生物体的基本结构，是代谢和功能的基本单元，具有动态性。蛋白质相互作用网络会随着生物进程动态变化，因此构建动态蛋白质相互作用网络更符合研究需求。同时在网络中动态的生成蛋白质复合物样本可以更深层次的研究蛋白质复合物的生成规律。然而动态网络的构建以及动态子图的生成均具有一定的挑战性，也是该领域的研究热点和难点。

2）融合多源的蛋白质特征。本文融合了基础的蛋白质特征和GO注释、拓扑域等相似性特征。蛋白质序列数据和蛋白质的三维结构数据是蛋白质特征的直观的体现，序列数据和三维结构的研究也是蛋白质研究领域的重点。如何将蛋白质序列数据、三维结构数据和蛋白质复合物预测问题结合起来仍然需要进一步研究。