

# 一种基于多头注意力机制的关键蛋白质靶点识别方法

杨文雅<sup>1</sup>, 关月<sup>1</sup>, 周予<sup>2</sup>

<sup>1</sup>贵州大学大数据与信息工程学院, 贵州 贵阳

<sup>2</sup>信阳职业技术学院数学与信息工程学院, 河南 信阳

收稿日期: 2023年9月20日; 录用日期: 2023年11月21日; 发布日期: 2023年11月28日

## 摘要

蛋白质是机体功能的重要执行者。特定的一些蛋白质对生物的生存、繁殖和生理调节尤为关键, 它们往往成为疾病发生、发展中的重要参与者, 这类蛋白质被称为关键蛋白质。因此, 在疾病预防和治疗过程中快速找到蛋白质关键靶点是尤为重要的。本文提出了一种借助多头注意力机制来解决这一难题的方法。该方法将蛋白质相互作用(PPI)网络的拓扑特征、基因表达谱特征以及同源特融合, 从而构建一个融合PPI网络。进而, 我们采用图注意力神经网络(GAT)模型, 学习融合PPI网络中节点的特征表示, 为了更好的地捕获蛋白质之间的关联关系, 我们引入多头注意力机制增强模型的学习效果。最终, 通过在DIP酵母蛋白数据集上的训练和测试, 实验结果证明了我们的方法相较于传统的基于拓扑的策略具有更高的识别精度。

## 关键词

多头注意力机制, 图神经网络, 蛋白质相互作用网, 基因表达谱, 同源相似性

# A Method for Essential Protein Target Recognition Based on a Multi-Head Attention Mechanism

Wenya Yang<sup>1</sup>, Yue Guan<sup>1</sup>, Yu Zhou<sup>2</sup>

<sup>1</sup>College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

<sup>2</sup>School of Mathematics and Information Engineering, Xinyang Vocational and Technical College, Xinyang Henan

Received: Sep. 20<sup>th</sup>, 2023; accepted: Nov. 21<sup>st</sup>, 2023; published: Nov. 28<sup>th</sup>, 2023

文章引用: 杨文雅, 关月, 周予. 一种基于多头注意力机制的关键蛋白质靶点识别方法[J]. 建模与仿真, 2023, 12(6): 5693-5702. DOI: 10.12677/mos.2023.126517

## Abstract

Proteins are important performers of organismal functions. Certain proteins are especially critical for organisms' survival, reproduction, and physiological regulation. They often play a significant role in the onset and progression of the disease, which are called essential proteins. Therefore, rapidly identifying the essential protein targets is of paramount importance in the prevention and treatment of diseases. In this paper, we propose an approach to address this challenge with the help of a multi-head attention mechanism. Our approach uses topological features of protein-protein interaction (PPI) networks, gene expression profile features, and homology features to construct a fusion PPI network. Further, we employ a graph attention neural network (GAT) model to learn the feature representations of the nodes in the fusion PPI network. To better capture the association relationships among proteins, we introduce a multi-head attention mechanism to enhance the learning effect of the model. Finally, through training and testing on DIP yeast protein datasets, the experimental results demonstrate the higher recognition accuracy of our method compared to the traditional topology-based strategy.

## Keywords

Multi-Head Attention Mechanism, Graph Neural Networks, PPI Networks, Gene Expression Profiles, Homology Similarity

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

蛋白质是基因表达的产物，参与几乎所有生物活动，是生物体中不可缺少的物质[1]。其中，某些蛋白质对生物的生命维持、繁衍及生理调控有着至关重要的影响，并在疾病的形成和进展中扮演着核心角色，我们通常将这些蛋白质称为关键蛋白质[2]。因此，在疾病的预防和治疗中，准确而迅速地定位这些关键蛋白质靶点变得格外关键。

传统的识别关键蛋白质的方法主要依靠生物学实验，例如 RNA 干扰[3]，单基因敲除[4]、条件性基因剔除[3]等，这些方法虽然准确性较高，但实验过程复杂、成本昂贵、实验周期长，而且适用于有限的物种。近年来，高通量蛋白质技术的发展为积累了大量的蛋白质相互作用(PPI)数据提供了条件，如蛋白质相互作用数据库(Database of Interacting Protein, DIP)，美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)等。这些数据库的建立为研究人员设计计算方法识别关键蛋白质奠定了数据基础。随着蛋白质数据的累积，PPI 网络的构建为理解蛋白质复杂性、生物通路和细胞组织功能提供了新的视角[5] [6]。一系列基于 PPI 拓扑结构的中心性方法被提出，例如，Degree Centrality (DC [7])方法采用度中心性，Subgraph Centrality (SC [8])采用子图中心性，Eigenvector Centrality (EC [9])采用特征向量中心性，Information Centrality (IC [10])引入信息中心性，Neighborhood Centrality (NC [11])引入邻居中心性，Local Average Connectivity (LAC [12])引入局部特征中心性。基于 PPI 网络拓扑特性提出的一系列方法虽然有效地提高了关键蛋白质识别率，但是蛋白质相互作用数据本身具有很高的假阳性，使得基于拓扑结构的方法存在局限性。为了解决上述问题，更多的研究者提出将蛋白质相互作用网络的拓扑特

征与生物特征相结合[13][14][15][16][17]。一种名为 RWHN [18]的方法通过整合加权蛋白质 - 蛋白质相互作用网络和域 - 域关联网络,构建了一个异构网络来识别必需蛋白质。Lei 等人提出了 RWEPI [19]算法,该算法首先利用基因表达信息和基因本体注释信息构建加权的蛋白质相互作用网络,然后采用随机游走算法来度量蛋白质的关键性。这些算法利用蛋白质内在的生物学特性优化网络结构,消除噪声对原始网络数据的影响,识别效果显著增强。

生物学网络分析已成为生物信息界和计算机界的热门话题。多头注意力机制作为一种强大的特征识别方法在生物网络分析中崭露头角。是自注意力机制的一种扩展,主要用于提高模型在处理不同位置间关系时的能力。同时,多头注意力机制允许模型自适应地关注网络中不同节点的信息,根据节点之间的相关性分配不同的注意力权重。这种机制使得模型能够更好地识别关键的节点和边,从而提高生物网络分析的准确性和解释性,更快地识别关键蛋白质。

与以往不同的是,本文提出一种基于多头注意力机制的关键蛋白质靶点识别方法(EPMHG)。该方法将基因表达谱和同源性特征与 PPI 网络拓扑特征融合,构建融合 PPI 网络,引入图注意力神经网络学习融合 PPI 网络中的节点表示,为了使得模型可以在多个表示空间中并行地捕获不同信息,增强模型的学习能力,多头注意力机制被采用,这可以有效的改进识别蛋白质的精度。与现存的基于拓扑特征和随机游走策略的方法相比,EPMHG 方法在识别精度上更胜一筹。

## 2. 关键蛋白质识别架构

### 2.1. 基于多头注意力机制的关键蛋白质识别框架

本文提出一种基于多头注意力机制的关键蛋白质靶点识别方法(EPMHG)。PPI 网络拓扑特征和基因表达谱、同源相似性相融合作为蛋白质的初始特征从而构建一个融合 PPI 网络。随后引入图注意力神经网络对融合 PPI 网络进行节点学习,为了更全面地捕获蛋白质之间的关联关系信息,我们采用多头注意力机制改善模型的学习能力,进而使得关键蛋白质识别精度得到改善。如图 1 所示,基于多头注意力机制的关键蛋白质靶点识别框架图。

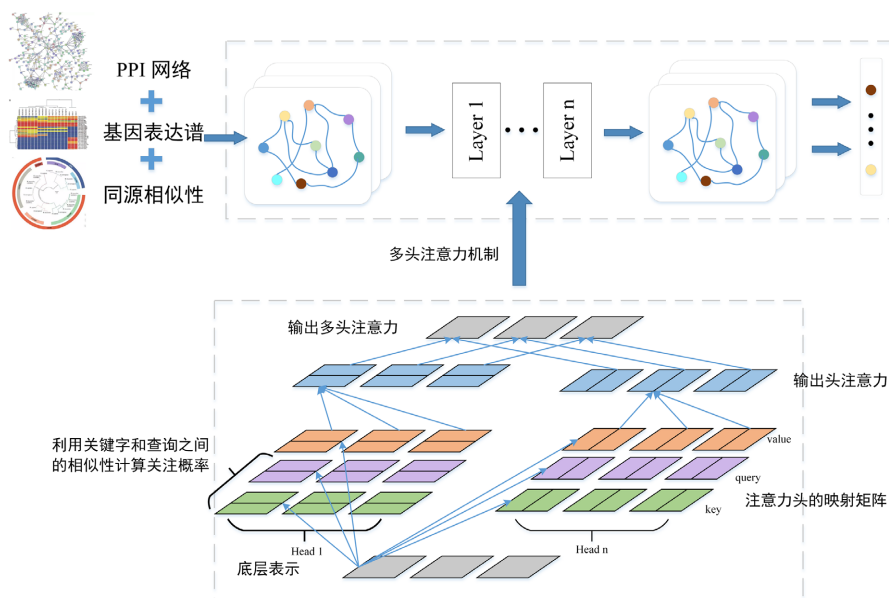


Figure 1. A framework for identifying essential proteins based on multi-head attention mechanisms

图 1. 基于多头注意力机制的关键蛋白质识别框架图

## 2.2. 构建 PPI 网络

细胞中的蛋白质不是独立的，它们相互作用，构成 PPI 网络。PPI 网络是一种重要的生物网络，满足复杂网络(如小世界、无标度和模块化)的拓扑特性。构建一个无向图  $G=(V,E)$ ，其中， $V$  为节点集， $E$  为边集。在这个网络中，节点集由不同的蛋白质组成。当且仅当蛋白质相互作用时，在 PPI 网络中每对蛋白质之间都有一条边。

## 2.3. 特征公式化

### 2.3.1. 计算边聚集系数

据研究表明，关键蛋白质更频繁的与其他关键蛋白质相连。通过分析关键蛋白质之间相互作用的拓扑特性，他们发现了一个几乎完全连通的指数网络，这意味着蛋白质的关键性与其相邻蛋白质的关键性之间存在着很强的相关性。

因此，我们将边聚集系数概念引入蛋白质相互作用网络中预测关键蛋白质[20]。边聚集系数描述了一条边的两个连接节点与其周围其他节点之间的紧密程度，该系数不仅可以评价边缘在 PPI 网络中的重要性，又可以描述两个蛋白质之间的紧密程度。我们定义  $u$  和  $v$  两个节点边距集系数公式如下：

$$Ecc(v_i, v_j) = \frac{z_{v_i, v_j}}{\min(d_{v_i} - 1, d_{v_j} - 1)} \quad (1)$$

其中， $z_{v_i, v_j}$  指的是相互作用的边  $v_i, v_j$  在 PPI 网络中形成的实际的三角形的数量， $\min(d_{v_i} - 1, d_{v_j} - 1)$  表示的是相互作用的边  $v_i, v_j$  在 PPI 网络中最大可能会构成的三角形数量。

### 2.3.2. 基因表达相似性

基因表达是一个基因被表达为功能性基因产物的过程，这些产物通常是蛋白质。基因表达也被广泛用于鉴定关键蛋白质。对于一个给定的基因  $u$ ，它在不同时间的基因表达可以通过一个载体来表达， $\exp(\mu, i)$  为基因  $\mu$  在  $i$  时刻的表达量。

PCC 是用来衡量两个变量之间线性相关的简单相关系数[21]，其值在  $[-1, 1]$  之间，我们引入 PCC 来表征基因共表达的相似性，它在自然科学中被广泛应用。对于  $u$  和  $v$  基因，他们之间的计算如下：

$$PCC_{v_i, v_j} = \frac{\sum_{k=1}^n \left( \frac{gen(v_i, k) - \overline{gen(v_i)}}{\sigma(v_i)} \right) \times \left( \frac{gen(v_j, k) - \overline{gen(v_j)}}{\sigma(v_j)} \right)}{n-1} \quad (2)$$

式中  $\overline{gen(v)}$  表示  $v$  基因在各个时刻的平均表达量， $\sigma(v_i)$  表示  $v$  基因在各个时刻的表达量的标准方差。如果为正，则基因  $v_i, v_j$  正相关，如果值为负，则基因  $v_i, v_j$  呈负相关。

### 2.3.3. 计算同源相似性

直系同源物是指源自共同祖先的同源蛋白质，它们通常具有高度相似的氨基酸序列，并保留相同或非常相似的功能。许多研究使用同源信息来识别 PPI 网络的进化信，发现蛋白质进化速率，推断蛋白质保守性[22]。

给定一个 PPI 网络  $G=(V,E)$ ，设  $S$  为参考物种集，用于获取  $v_i$  的同源信息， $s$  表示  $S$  集的元素。对于蛋白质  $v$ ，其同源分数  $OS_{v_i}$  被定义为蛋白质  $v_i$  在参考物种集中具有同源物的数量，其中  $v \in V (v=1, \dots, N)$ 。对同源相似性  $NOS_{v_i}$  进行归一化处理，公式为：

$$NOS_{v_i} = \frac{OS_{v_i}}{OS_{\max}} \quad (3)$$

其中,  $MAX\_OS$  表示  $G$  中所有蛋白质的  $os_{v_i}$  最大值。如果蛋白质  $v_i$  在  $S$  物种集中的所有生物都存在同源关系, 则同源得分为 1。相反, 如果蛋白质  $v_i$  没有和任何生物存在同源关系, 同源得分为 0。

### 3. EPMHG 方法原理

#### 3.1. 构建特征集合

EPMHG 方法通过融合 PPI 网络拓扑特征与基因表达特征、同源相似性特征, 突出蛋白质之间强关联关系。边聚集系数(ECC)强调蛋白质之间的紧密程度, 皮尔逊系数(PCC)衡量两个蛋白质之间线性相关性, 同源相似性(NOS)推断蛋白质的保守性。PPI 网络被定义为一个无向图  $G=(V,E)$ ,  $V$  表示节点集, 其中包含  $N$  个蛋白质节点。ECC, PCC, NOS 作为每个蛋白质节点的初始特征, 因此, 获得一个 3 维的蛋白质节点特征集合, 可表示为  $h = \{h_1, h_2, \dots, h_N\}$ ,  $h_i \in \mathbb{R}^F$ ,  $F=3$ 。

#### 3.2. 计算注意力分数

注意力层测目的是输出新的节点特征集,  $h = \{h'_1, h'_2, \dots, h'_N\}$ ,  $h'_i \in \mathbb{R}^{F'}$ 。在这个过程中特征向量的维度可能会发生改变, 即  $F \rightarrow F'$ 。为了保留足够的表达能力, 将输入特征转化为高阶特征, 至少需要一个可学习的线性变换。对于蛋白质节点  $v_i, v_j$ , 它们的特征  $h_i, h_j$  应用线性变换  $W \in \mathbb{R}^{F \times F'}$ , 从而  $F$  维转化为  $F'$  维新特征  $h'_i, h'_j$ :

$$e_{ij} = a(W h_i, W h_j) \quad (4)$$

在将输入特征运用线性变换转换为高阶特征后, 使用自注意力为每个节点分配注意力权重。其中,  $a$  表示一个共享注意力机制:  $\mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R}$ , 用于计算注意力系数  $e_{ij}$ , 也就是蛋白质节点  $v_i$  对  $v_j$  的影响力系数。

注意力机制的计算考虑了每个蛋白质节点对目标节点的影响。对于目标节点  $v_i$ , 只计算其邻域内的节点  $v_j \in N$  对目标节点的相似度  $e_{ij}$ 。

为了更好的在不同节点之间分配权重, 我们需要将目标节点与所有邻居计算出来的相关度进行统一的归一化处理, 这里用 softmax 归一化:

$$a_{ij} = \text{soft max}_j (e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \quad (5)$$

关于  $a$  的选择, 可以用向量的内积来定义一种无参形式的相关度计算  $\langle W h_i, W h_j \rangle$ , 也可以定义成一种带参的神经网络层, 只要满足  $a: \mathbb{R}^{d(l)} \times \mathbb{R}^{d(l+1)} \rightarrow \mathbb{R}$ , 即输出一个标量值表示二者的相关度即可。在论文实验中,  $a$  是一个单层前馈神经网络, 参数为权重向量  $a \in \mathbb{R}^{2F'}$ , 使用负半轴斜率为 0.2 的 LeakyReLU 作为非线性激活函数:

$$e_{ij} = \text{LeakyReLU}(a^T [W h_i \| W h_j]) \quad (6)$$

其中  $\|$  表示拼接操作。完整的权重系数计算公式为:

$$a_{ij} = \frac{\exp(\text{LeakyReLU}\{a^T [W h_i \| W h_j]\})}{\sum_{k \in N_i} \exp(\text{LeakyReLU}\{a^T [W h_i \| W h_k]\})} \quad (7)$$

得到归一化注意系数后, 计算其 `duyingtezheng1` 的线性组合, 通过非线性激活函数后, 每个接待的最后输出特征向量为:



$$\mathbf{h}'_i = \left( \sum_{j \in N_i} a_{ij} W \mathbf{h}_j \right) \quad (8)$$

### 3.3. 多头注意力机制

本文使用多头注意力机制来增强模型的学习能力，即对公式(8)调用  $K$  组相互独立的注意力机制，然后将输出结果拼接起来：

$$\mathbf{h}'_i = \parallel_{k=1}^K \sigma \left( \sum_{j \in N_i} a_{ij}^k W^k \mathbf{h}_j \right) \quad (9)$$

其中， $\parallel$ 表示拼接操作。 $a_{ij}^k$ 是第  $k$  组注意力机制计算出的权重系数， $W^k$ 是对应的输入线性变换矩阵，最终输出的节点特征向量  $\mathbf{h}'_i$  包含了  $KF'$  个特征。为了减少输出的特征向量的维度，也可以将拼接操作替换为平均操作。

$$\mathbf{h}'_i = \sigma \left( \frac{1}{K} \sum_{k=1}^K \sum_{j \in N_i} a_{ij}^k W^k \mathbf{h}_j \right) \quad (10)$$

## 4. 仿真实验

### 4.1. 实验数据

酵母蛋白质相互作用数据是目前使用最广泛的生物数据，实验以酵母蛋白质相互作用网络作为研究对象。本文选取来源于 DIP (Database of Interacting Protein) 数据库的蛋白质相互作用数据集来进行构建 PPI 网络。滤除数据集中重复的蛋白质，DIP 数据集中包含 5093 个蛋白质，1167 个关键蛋白质和 24,743 对相互作用边。本文在实验部分提出将 EPMHG 方法与现存的六种经典方法进行比较，分别是 Degree Centrality (DC [7]), Subgraph Centrality (SC [8]), Eigenvector Centrality (EC [9]), Information Centrality (IC [10]), Neighborhood Centrality (NC [11]), Local Average Connectivity (LAC [12])。

### 4.2. 实验结果对比及分析

#### 4.2.1. 识别数量

为了验证 EPMHG 算法在识别数量的表现，分别选择 DIP 数据集中蛋白质排名在 Top 1%、Top 5%、Top 10%、Top 15%、Top 20%、Top 25% 的蛋白质数量作为关键蛋白质，然后比较每种测度参数的预测结果中真实关键蛋白质的数量。如图 2 所示，EPMHG 算法在六个测度上的预测关键蛋白质数量分别为 50、213、373、514、636、723，根据数量比较，该方法在预测数量上均高于其余六种方法。这主要的原因是相比较的六种方法(DC, SC, IC, EC, LAC, NC)是基于 PPI 网络拓扑结构特征提出的，由于 PPI 网络本身具有较高的假阳性和假阴性，这使得预测到限制。相比较而言，EPMHG 算法是融合了 PPI 网络拓扑特征、基因表达特征和同源相似性特征，使得蛋白质节点信息更全面，因此预测的数量远高于其余六种方法。

#### 4.2.2. 精确度对比

为了进一步评估 EPMHG 算法预测关键蛋白质预测的准确度。本小节计算了敏感度(Sensitivity, SN)，正确率(Accuracy, ACC)，特异性(Specificity, SP)，阳性预测值(Positive Predictive Value, PPV)，F-测度(F-measure)，阴性预测值(Negative Predictive Value, NPV)六个指标对 EPMHG 算法与其余六种方法进行了比较。六种指标的计算公式表示如下：

$$SN = \frac{TP}{TP + FN} \quad (11)$$

$$SP = \frac{TN}{TN + FP} \quad (12)$$

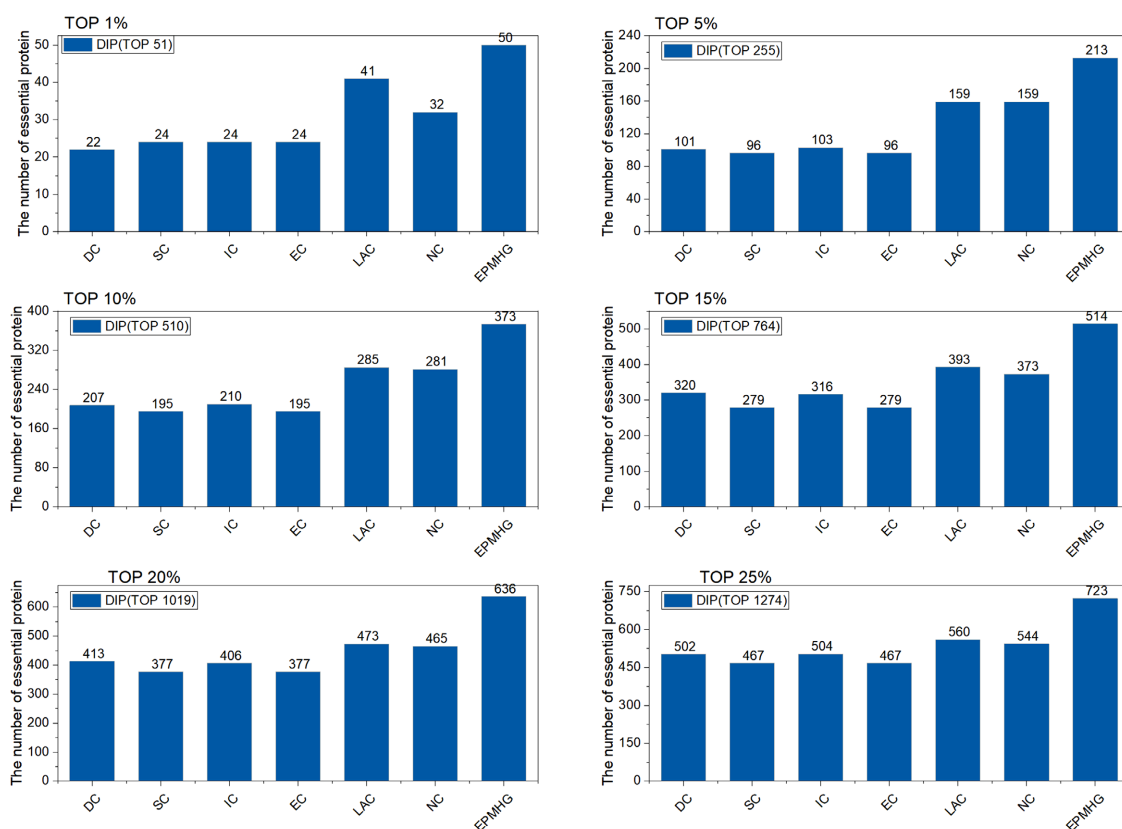
$$PPV = \frac{TP}{TP + FP} \quad (13)$$

$$NPV = \frac{TN}{TN + FN} \quad (14)$$

$$F\text{-measure} = \frac{2 \times SN \times PPV}{SN + PPV} \quad (15)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

其中, TP (true positive)表示关键蛋白质被正确预测为关键蛋白质的数量, FN (false negative)表示关键蛋白质被错误地预测为非关键蛋白质的数量, TN (true positive)表示关键蛋白质被正确预测为非关键蛋白质的数量, FN (false negative)表示非关键蛋白质被错误预测为关键蛋白质的数量。



**Figure 2.** The number of essential protein identifications  
**图 2.** 关键蛋白质识别数量图

该部分选择 DIP 数据集的 TOP 25% 作为计算六个指标的数据, 最终对比结果呈现在表 1 中。由表 1 可知, EPMHG 算法的准确率 ACC 为 0.775, 这超过了相比的其余六种方法。该方法采用了多头注意力机制增强了模型的学习能力, 使得模型可以更准确地预测关键蛋白质。

**Table 1.** Comparison of the six metrics in the DIP dataset  
**表 1.** 比较 DIP 数据集的六种指标

方法	SN	SP	PPV	NPV	F-measure	ACC
DC	0.4302	0.8034	0.3940	0.8259	0.4113	0.7178
SC	0.7944	0.7944	0.3666	0.8167	0.3826	0.7041
IC	0.4319	0.8039	0.3956	0.8264	0.4129	0.7186
EC	0.4002	0.7944	0.3666	0.8167	0.3826	0.7041
LAC	0.4799	0.8181	0.4396	0.8411	0.4588	0.7406
NC	0.4662	0.8141	0.4270	0.8369	0.4457	0.7343
PeC	0.4233	0.8013	0.3878	0.8238	0.4048	0.7147
EPMHG	0.6438	0.7875	0.2242	0.9586	0.3326	0.7750

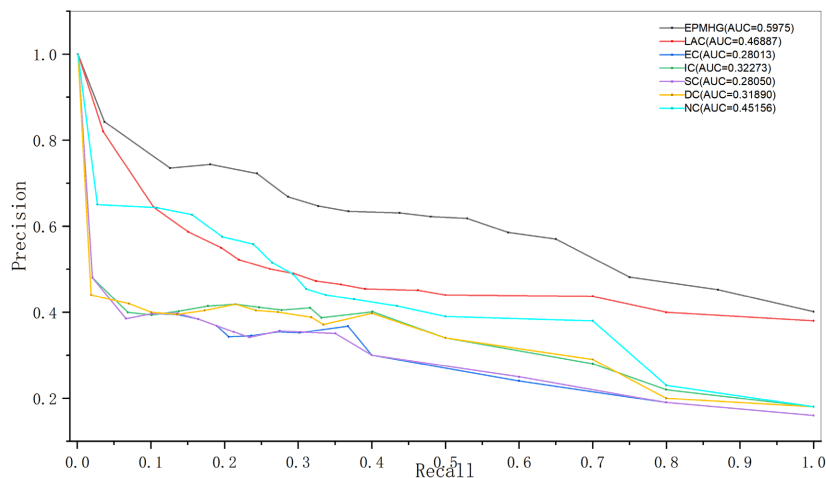
### 4.2.3. PR 曲线仿真

本文也研究了 EPMHG 算法的 Precision-recall (PR) 曲线, 并将其与其他六种方法进行了比较, 选择 DIP 数据集的 TOP 25% 作为验证数据。如图 3 所示, X 轴表示 Recall, Y 轴表示 Precision。Top n 个蛋白质的精确度和召回率被定义为:

$$\text{Precision} = \frac{\text{TP}(\text{Nu})}{\text{TP}(\text{Nu}) + \text{FP}(\text{Nu})} \quad (17)$$

$$\text{Recall} = \frac{\text{TP}(\text{Nu})}{\text{Total } P} \quad (18)$$

其中, TP(Nu) 表示从 Top n 个蛋白质里识别出的关键蛋白质数量, FP(Nu) 表示从 Top n 个蛋白质里识别出的非关键蛋白质数量, Total P 表示 DIP 数据集的关键蛋白质数量。同时也计算了 PR 曲线下面积, 这有助于更准确地比较各种方法在精确度的优势。从图 3 的仿真结果可知, EPMHG 算法的 PR 曲线高于其余六种方法, 这表明该方法有较高的识别准确率。EPMHG 算法引入图注意力神经, 该网络的学习策略是识别并重点关注与当前节点更相关或更重要的邻居, 进而更准确识别关键蛋白质。



**Figure 3.** The precision-recall simulation figure  
**图 3.** 精度 - 召回率(PR)仿真图



## 5. 结论

本文提出了一种结合多头注意力机制来识别关键蛋白质靶点的方法(EPMHG)。通过整合蛋白质相互作用网络的拓扑特征和生物信息构建融合 PPI 网络,并引入多头注意力机制增强图注意力神经网络模型的学习效果,最终,在 DIP 酵母蛋白数据集上成功地验证了该方法的识别精度。该方法的提出为蛋白质关键靶点的快速识别提供了一种更为有效的策略,并为未来的疾病预防和治疗提供了新视角。

## 参考文献

- [1] Giaever, G., *et al.* (2002) Functional Profiling of the *Saccharomyces Cerevisiae* Genome. *Nature*, **418**, 387-391.
- [2] Roemer, T., *et al.* (2003) Large-Scale Essential Gene Identification in *Candida albicans* and Applications to Antifungal Drug Discovery. *Molecular Microbiology*, **50**, 167-181. <https://doi.org/10.1046/j.1365-2958.2003.03697.x>
- [3] Cullen, L.M. and Arndt, G.M. (2005) Genome-Wide Screening for Gene Function Using RNAi in Mammalian Cells. *Immunology & Cell Biology*, **83**, 217-223. <https://doi.org/10.1111/j.1440-1711.2005.01332.x>
- [4] Li, M., Zhang, H., Wang, J.X. and Pan, Y. (2012) A New Essential Protein Discovery Method Based on the Integration of Protein-Protein Interaction and Gene Expression Data. *BMC Systems Biology*, **6**, Article No. 15. <https://doi.org/10.1186/1752-0509-6-15>
- [5] Pál, C., Papp, B. and Hurst, L.D. (2003) Genomic Function: Rate of Evolution and Gene Dispensability, *Nature*, **421**, 496-497. <https://doi.org/10.1038/421496b>
- [6] Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., and Li, X. (2017) Predicting Potential Drug-Drug Interactions by Integrating Chemical, Biological, Phenotypic and Network Data. *BMC Bioinformatics*, **18**, 1-12.
- [7] Vallabhajosyula, R.R., Chakravarti, D., Lutfekali, S., Ray, A. and Raval, A. (2009) Identifying Hubs in Protein Interaction Networks. *PLOS ONE*, **4**, e5344. <https://doi.org/10.1371/journal.pone.0005344>
- [8] Li, M., Wang, J., Chen, X., Wang, H. and Pan, Y. (2011) A Local Average-connectivity-Based Method for Identifying Essential Proteins from the Network Level. *Computational Biology and Chemistry*, **35**, 143-150. <https://doi.org/10.1016/j.compbiolchem.2011.04.002>
- [9] Bonacich, P. and Lloyd, P. (2001) Eigenvector-Like Measures of Centrality for Asymmetric Relations. *Social Networks*, **23**, 191-201. [https://doi.org/10.1016/S0378-8733\(01\)00038-7](https://doi.org/10.1016/S0378-8733(01)00038-7)
- [10] Stephenson, K. and Zelen, M. (1989) Rethinking Centrality: Methods and Examples. *Social Networks*, **11**, 1-37. [https://doi.org/10.1016/0378-8733\(89\)90016-6](https://doi.org/10.1016/0378-8733(89)90016-6)
- [11] Li, M., Wang, J., Wang, H. and Pan, Y. (2012) Identification of Essential Proteins Based on Edge Clustering Coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 1070-1080. <https://doi.org/10.1109/TCBB.2011.147>
- [12] Estrada, E. and Rodríguez-Velázquez, J.A. (2005) Subgraph Centrality in Complex Networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, **71**, 056103. <https://doi.org/10.1103/PhysRevE.71.056103>
- [13] Wuchty, S. and Stadler, P.F. (2003) Centers of Complex Networks. *Journal of Theoretical Biology*, **223**, 45-53. [https://doi.org/10.1016/S0022-5193\(03\)00071-7](https://doi.org/10.1016/S0022-5193(03)00071-7)
- [14] Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011) Network Medicine: A Network-Based Approach to Human Disease. *Nature Reviews Genetics*, **12**, 56-68. <https://doi.org/10.1038/nrg2918>
- [15] Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B. and Bussey, H. (1999) Functional Characterization of the *S. Cerevisiae* Genome by Gene Deletion and Parallel Analysis. *Science*, **285**, 901-906.
- [16] Zhang, X., Xu, J. and Xiao, W. (2013) A New Method for the Discovery of Essential Proteins. *PLOS ONE*, **8**, e58763. <https://doi.org/10.1371/journal.pone.0058763>
- [17] Tang, X., Wang, J., Zhong, J. and Pan, Y. (2014) Predicting Essential Proteins Based on Weighted Degree Centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **11**, 407-418. <https://doi.org/10.1109/TCBB.2013.2295318>
- [18] Peng, W., Wang, J., Cheng, Y. and Lu, Y. (2015) UDoNC: An Algorithm for Identifying Essential Proteins Based on Protein Domains and Protein-Protein Interaction Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**, 276-288. <https://doi.org/10.1109/TCBB.2014.2338317>
- [19] Peng, W., Wang, J.X., Wang, W., *et al.* (2012) Iteration Method for Predicting Essential Proteins Based on Orthology and Protein-Protein Interaction Networks. *BMC Systems Biology*, **6**, Article No. 87. <https://doi.org/10.1186/1752-0509-6-87>

- [20] Pereira-Leal, J.B., Audit, B., Peregrin-Alvarez, J.M. and Ouzounis, C.A. (2005) An Exponential Core in the Heart of the Yeast Protein Interaction Network. *Molecular Biology and Evolution*, **22**, 421-425.  
<https://doi.org/10.1093/molbev/msi024>
- [21] Jancura, P., Mavridou, E., Pontes, B. and Marchiori, E (2011) Describing the Orthology Signal in a PPI Network at a Functional, Complex Level. *International Symposium on Bioinformatics Research and Applications*, **6674**, 209-226.  
[https://doi.org/10.1007/978-3-642-21260-4\\_22](https://doi.org/10.1007/978-3-642-21260-4_22)
- [22] Wuchty, S., Barabasi, A.L., Ferdig, M.T. (2006) Stable Evolutionary Signal in a Yeast Protein Interaction Network. *BMC Evolutionary Biology*, **6**, 8. <https://doi.org/10.1186/1471-2148-6-8>