背景引入

miRNA在细胞凋亡、分化和发展等生物过程中发挥着重要作用。越来越多的关注它们与复杂疾病的联系进一步强调了其重要性，但是传统的实验方法通常需要广泛的实验室能力和分析，准确但通常是低效的，它们耗时且昂贵。

本文针对以往的方法很少从整体的角度出发，只关注孤立的miRNA和疾病对象，从而忽略了人类细胞负责多种关系的问题，先构建了RNA-疾病 的多视图，然后用图注意力神经网络学习每个视图的miRNA和疾病的图拓扑特征，用多尺度特征融合模块优化，用miRNA和疾病的先验属性知识解决冷启动问题

作者：湖南大学彭绍亮教授团队博士生 Boya Ji ，Haitao Zou



相关工作

MHGTMDA: Molecular heterogeneous graph transformer based on biological entity graph

for miRNA-disease associations prediction基于生物实体图的分子异质图变换器用于miRNA-疾病关联预测

AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification注意力感知的多视图相似网络和超图学习用于miRNA-疾病关联识别

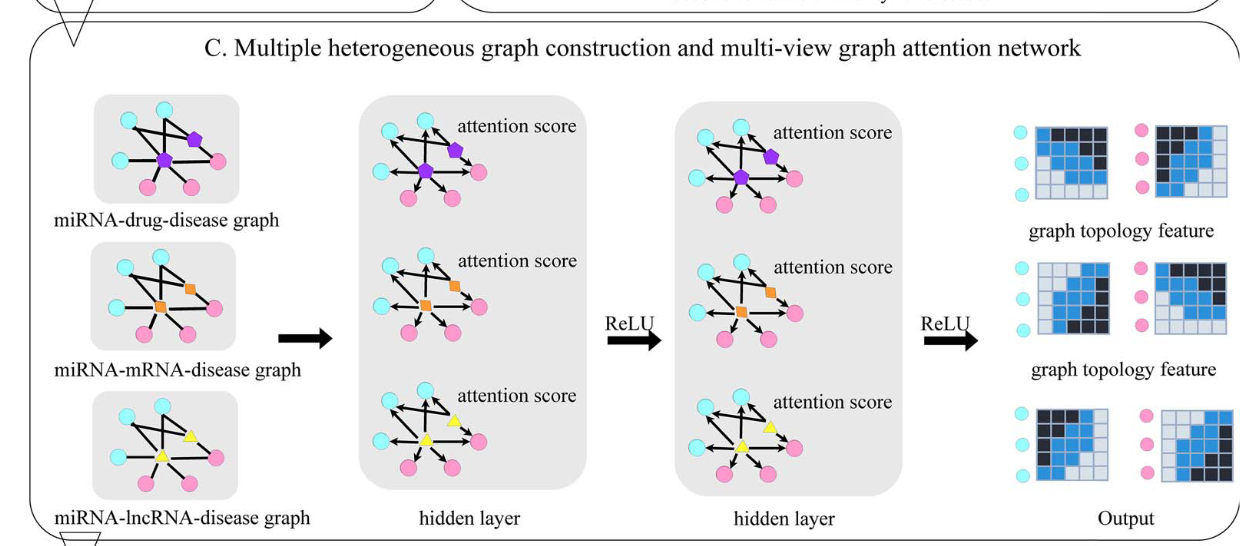
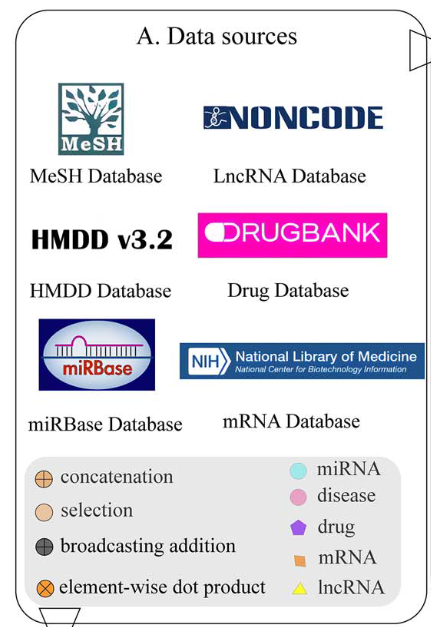
HGANMDA：Hierarchical graph attention network for miRNA-disease association prediction层次图注意力网络用于miRNA-疾病关联预测

NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion通过邻域约束矩阵补全预测miRNA-疾病关联

到目前为止，尽管已经使用了多种计算方法来预测MDA，但重要的是要注意，许多方法没有从整体角度出发，只关注孤立的miRNA和疾病目标，从而忽略了人类细胞负责多重关系的事实。

模型结构

数据集：人类microRNA疾病数据库（HMDDv3.2）-MDA、DrugBank数据库-异质性的miRNA-药物-疾病图、美国国家医学图书馆数据库-异质性的miRNA-mRNA-疾病、NONCODEV5数据库-异质性的miRNA-lncRNA-疾病图



特征提取：

1# miRNA 的基于序列的属性特征

使用k-mer方法将其转换为向量，该方法将序列划分为长度为k的子序列，从而获得一个长度为m-k+1的k-mer向量来表示长度为m的序列。随后，我们将这些频率值归一化，以创建一个64维向量，用于表示miRNA序列信息，从而捕获miRNA的序列属性特征。

2#基于语义相似性的疾病属性特征

医学主题词（MeSH，Medical Subject Heading，是由美国国家医学图书馆维护和更新的一个数据库）提供了一个全面的疾病分类系统。基于此，不同疾病之间的联系由有向无环图（DAG）表示。在这种图中，节点代表疾病的 MeSH 描述符，而有向边从更一般的实体连接到更具体的实体。例如，疾病 A 被表示为 DAG(A)=(D(A),E(A))，其中 D(A) 表示 A 和其祖先节点，E(A) 表示所有直接边。

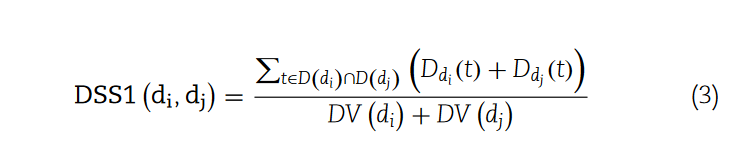
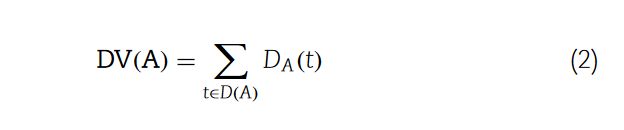
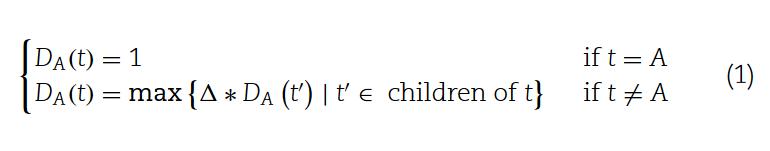
在这样的图中通过语意相似性特征来衡量两种疾病之间的关联度。

使用两种方法获取两种语意相似性特征 disease semantic similarity DSS1 / DSS2，采用两个语义相似性模型的总和作为疾病的属性特征，以实现更好的预测性能并解决冷启动问题。

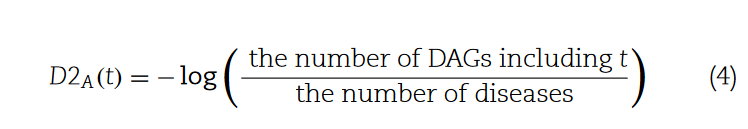
在此我的理解即为，冷启动问题源于模型在最开始时候还没学习到足够的特征，故而无法进行预测，那么我们通过已有的数据库MESH，为其提供基本的学习材料，来解决冷启动问题。

在以下算式中，A是一种疾病，t是MESH中的一种描述符

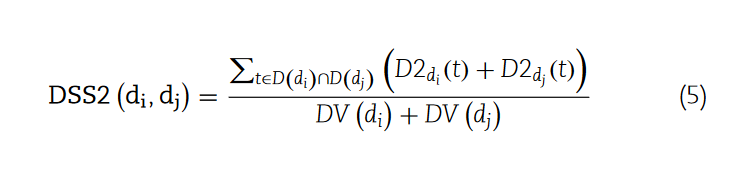
①DSS1获取方法：delta是语义贡献衰减因子，它减少了当疾病 t 不等于 A 时的影响。此外，疾病 A 本身保持了 1 的语义价值贡献。



②DSS2获取方法：考虑不同疾病对于语意价值具有不同的贡献，出现在较少DAG中的疾病应比出现在较多DAG中的疾病对A的影响较小。 疾病A的语义价值受以下方式影响：

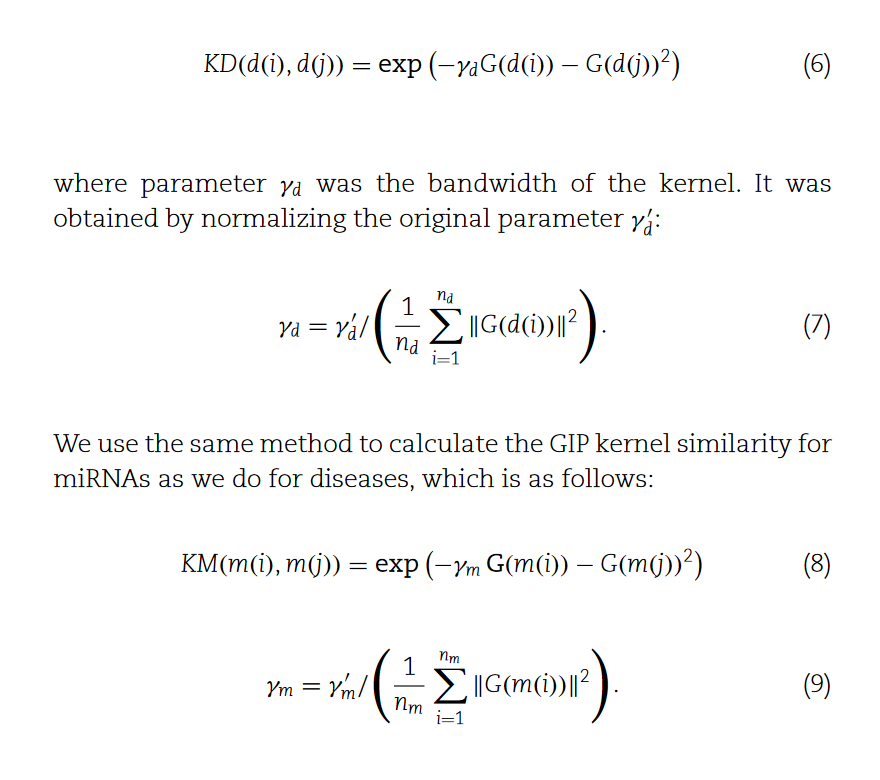


这样的方法与1不同之处在于考虑了不同类疾病



3#基于高斯交互作用特征轮廓核相似性的miRNA和疾病属性特征

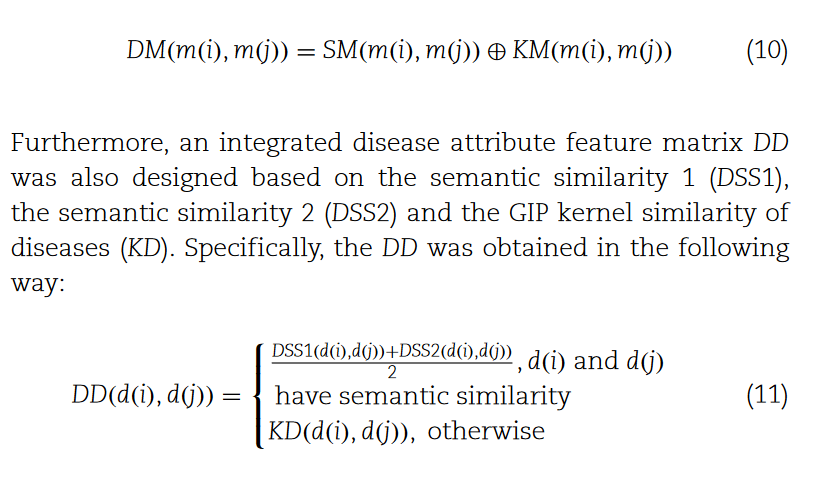
微小核糖核酸(microRNA,mRNA)具有相似功能往往与相似疾病有关，疾病d(i)和d(j)之间的GIP内核相似性 KD (d(i),d(j))定义如下：



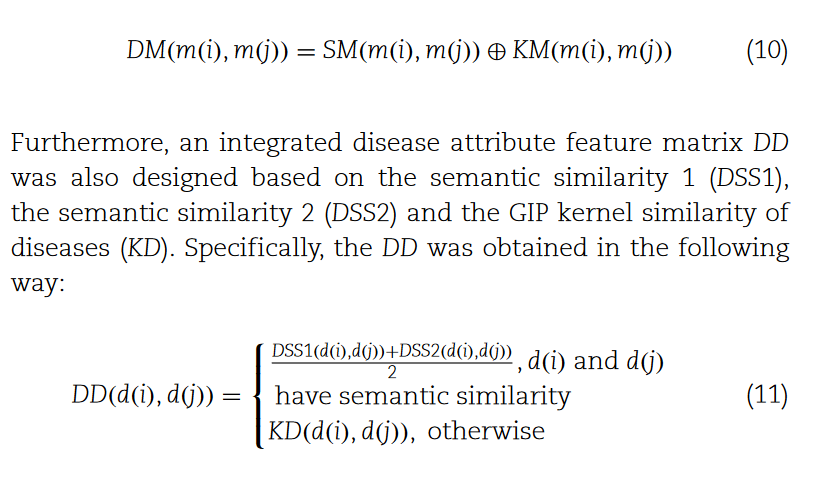
并且使用相同的方法来计算miRNA之间的相似性

以上所有工作的综合：

为了更全面地描述miRNA的属性特征，我们将基于序列的特征（SM）与基于GIP核相似性度量的特征（KM）连接起来，设计了一个集成miRNA属性特征矩阵DM。具体来说，DM通过以下方式获得：



此外，还基于语义相似性1（DSS1）、语义相似性2（DSS2）和疾病GIP内核相似度（KD）设计了一个综合疾病属性特征矩阵DD。具体来说，DD 的获取方式如下：



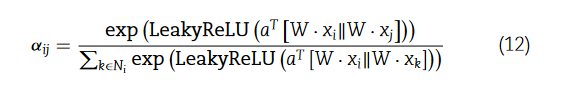
### 多视角异质图注意力网络

多视角异质图：由于疾病通路的复杂性和异质性，我们可能会忽略一些因素。因此，我们使用其他生物标志物作为中介来构建多视图图形，不同的中介物就是不同的视角。我们分别构建了三个异质图：miRNA-mRNA-disease 图、miRNA-lncRNA-disease 图和 miRNA-drug-disease 图。

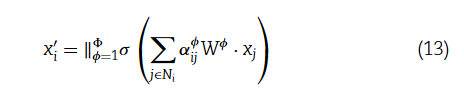
意思就是，不直接构建MDA的关联图，而是通过一个中间物质来构建MDA，比如原来是A-B，而现在写成A-C , C-B,所以A-C-B。

我们为每个视图使用了图注意力神经网络。最后，我们通过多尺度特征融合模块将三个视图中的每个miRNA和疾病特征融合为它们的最终图结构特征。下列介绍每个视图中的计算方法。

图注意力系数的计算：其中，W 是一个可学习的权重矩阵，Wi、Wj 和 Wk 分别表示节点 i、j 和 k 的特征向量。·、l 和 T 分别代表乘法、连接和转置操作，aT∈R2F 是单层前馈神经网络 a 的权重参数，LeakyReLU 是非线性激活函数，exp 是指数函数，Ni 表示节点 i 的所有邻居节点，αij 表示从节点 i 到节点 j 的注意力系数。为了增强模型的拟合能力，我们引入了多头注意力机制，即使用多个 W 矩阵同时计算各种注意力系数。



图节点特征的最终表示：其中，d 表示 W 矩阵的数量，α\_{φij} 表示第 φ 个注意力系数，σ 是激活函数，i 表示节点 i 的最终输出特征表示。当节点表示不再发生显著变化时，停止图注意力网络的学习过程，并获得最终的节点表示。



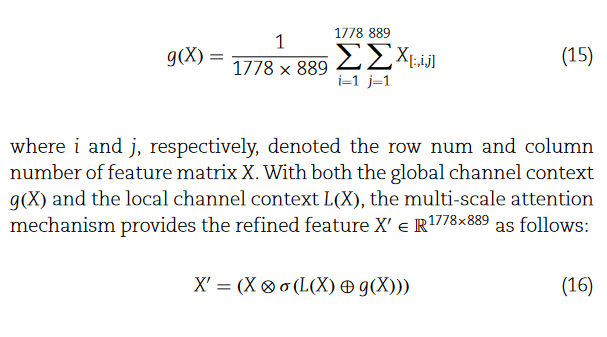
### 多尺度特征融合

为了有效地融合 miRNA 和疾病三种类型的图结构特征，以更准确地刻画它们并提高潜在关联预测的准确性，我们设计了一个多尺度注意力特征融合模块。具体来说，该模块聚合了 miRNA 和疾病三种类型图结构特征的局部和全局特征上下文。为了聚合这些特征的局部通道语境，我们采用了点卷积（PWConv）方法[18]，这是一种只考虑每个空间位置的点对通道交互的卷积形式，使用 1×1 核。通过瓶颈结构计算局部通道语境 L(X)：

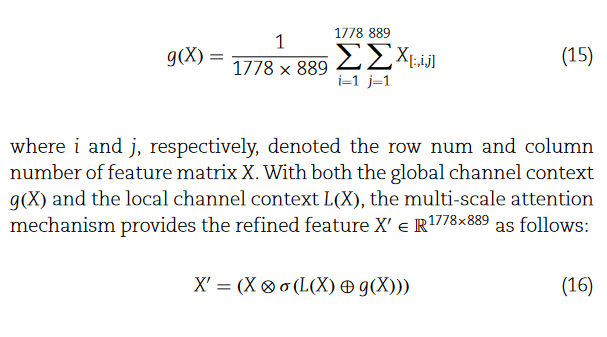
L(X) = B (PWConv2 (δ (B (PWConv1(X)))))

PWConv1 和 PWConv2 分别表示两个点卷积操作。PWConv1 和 PWConv2 的内核大小分别为 1 × 1778 × 889（通道减少层）和 3 × 1778 × 889（通道增加层）。B 表示批归一化 (BN)，而δ表示修正线性单元 (ReLU) 函数。

使用全局平均池化 (GAP) 如下：



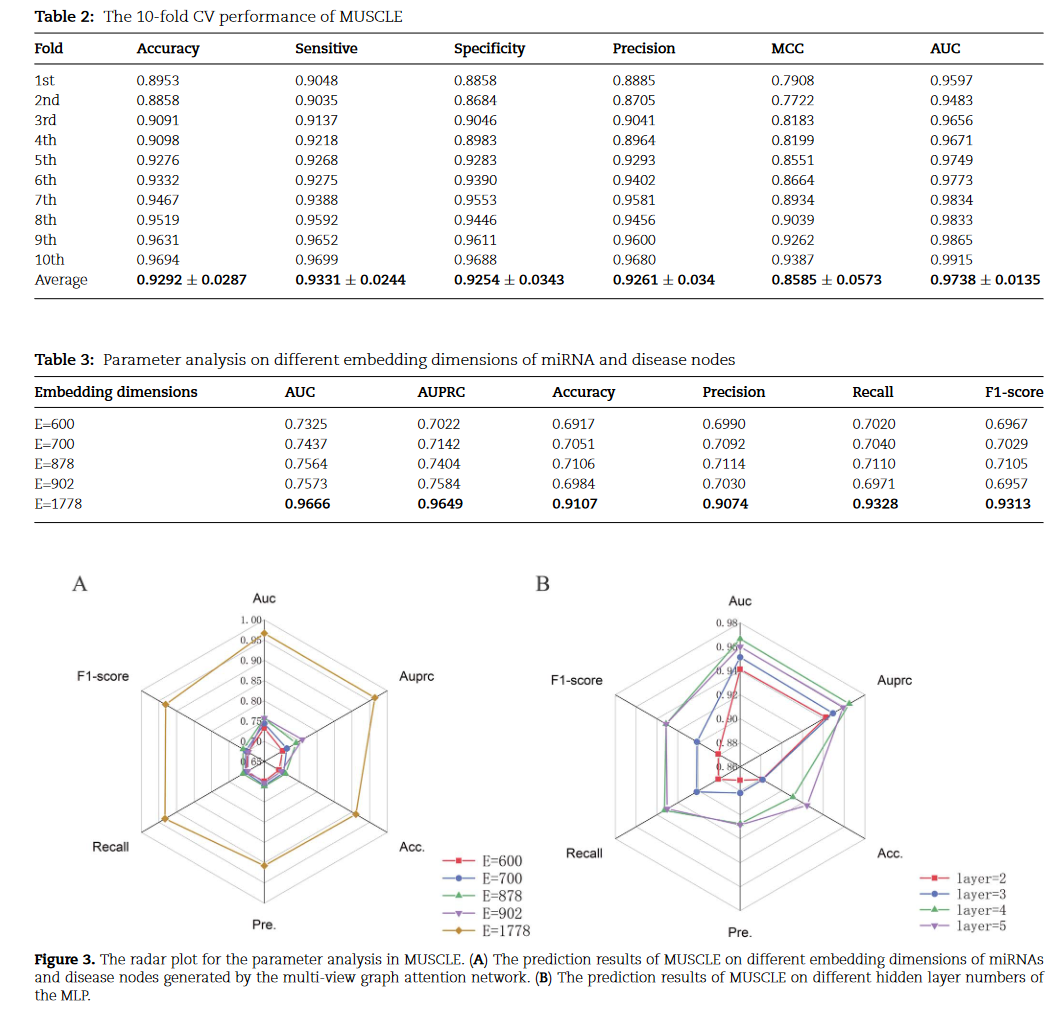
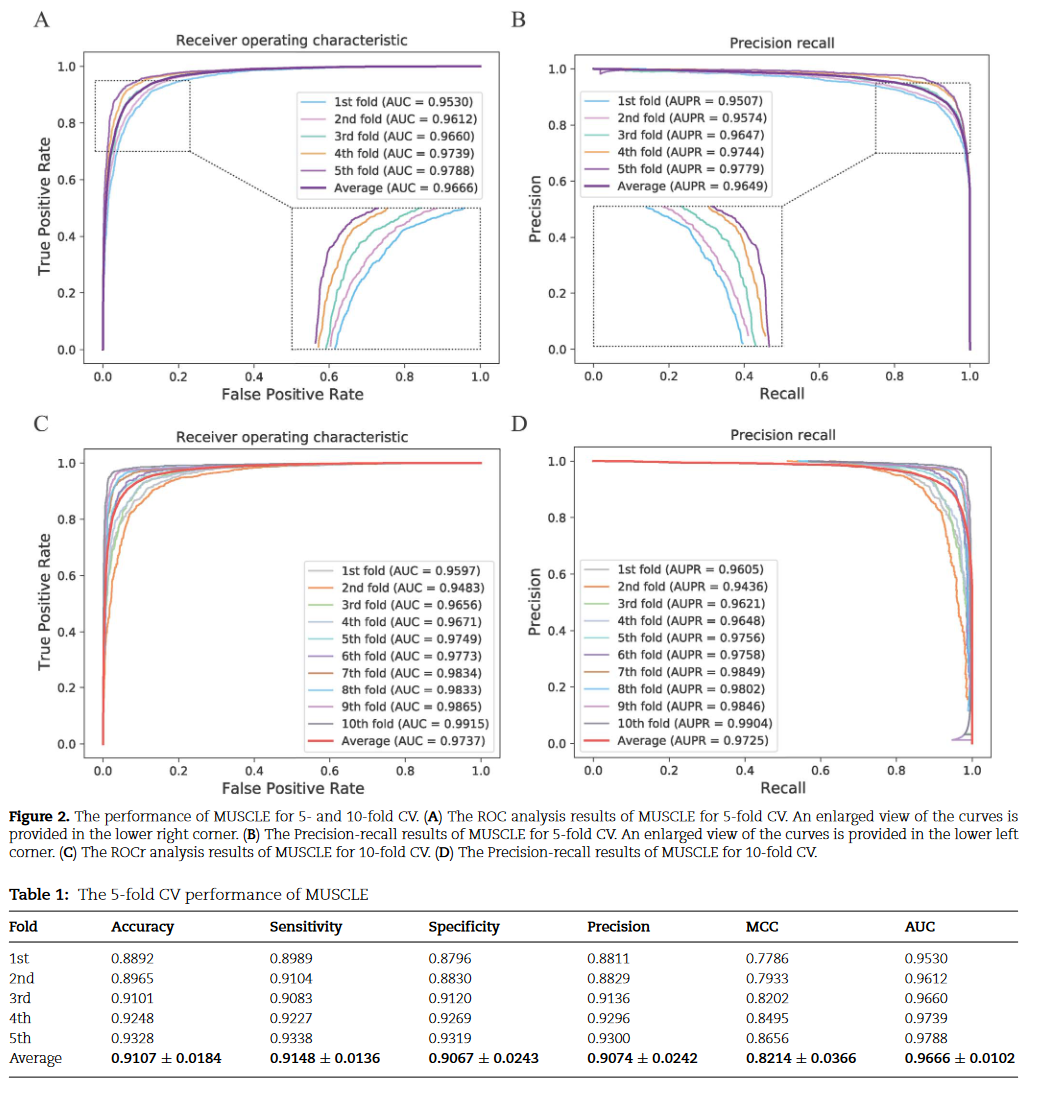
其中，i 和 j 分别表示特征矩阵 X 的行号和列号。通过全局通道上下文 g(X) 和局部通道上下文 L(X)，多尺度注意力机制提供了一个精炼的特征 X'∈R1778×889，如下所示：



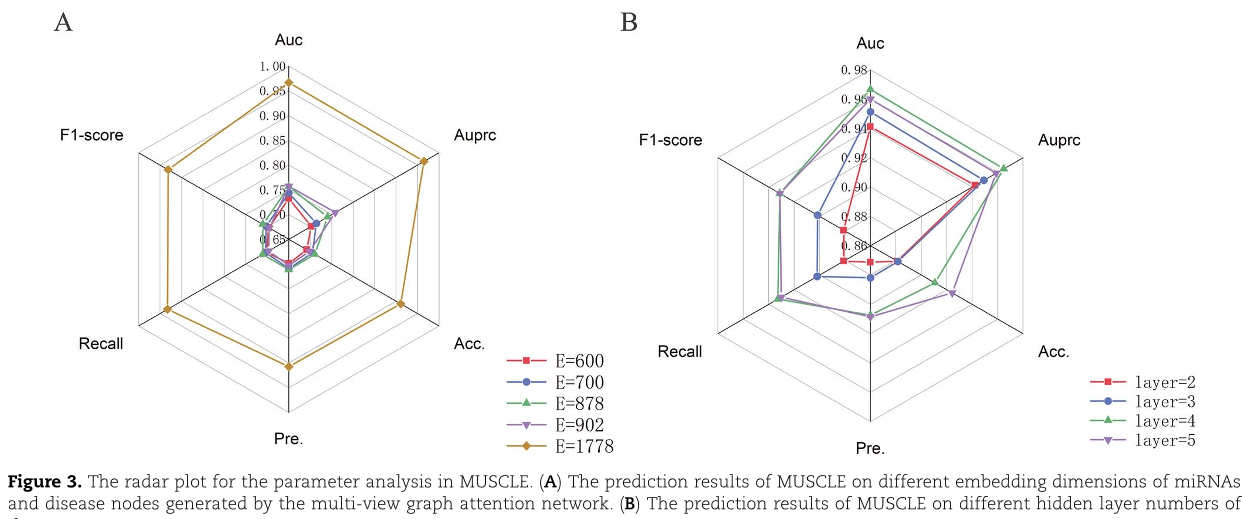
然后，将微小 RNA 和疾病的最终融合特征馈入一个多层感知器 (MLP)，并使用标准二元交叉熵损失函数进行训练和预测。

## 模型实验

### 交叉验证

在 5 折和 10折交叉验证下，MUSCLE 的平均 AUC 值分别为 0.9666 和 0.9737，标准差仅为 0.0102 和 0.0135。此外，MUSCLE 的平均 AUPR 值也分别达到 0.9649 和 0.9725。所有这些指标和可视化的结果证明了 MUSCLE 在预测潜在 MDA 时具有出色的表现和鲁棒性

### 参数分析

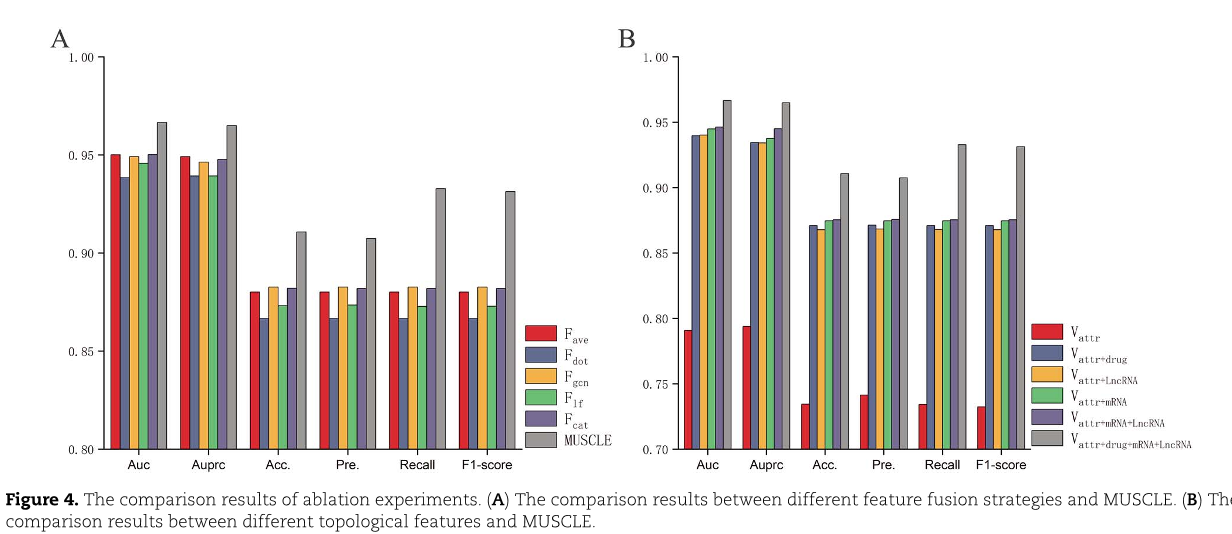


嵌入维度：随着维度的增加，模型预测的AUC持续上升。当维度达到 1778 时，结果得到了极大的改善。我们推测更大的特征维度提供了更多的信息，但是特征维度的持续增加实际上会增加噪声并导致性能下降，以及引入计算复杂度和时间开销等问题。

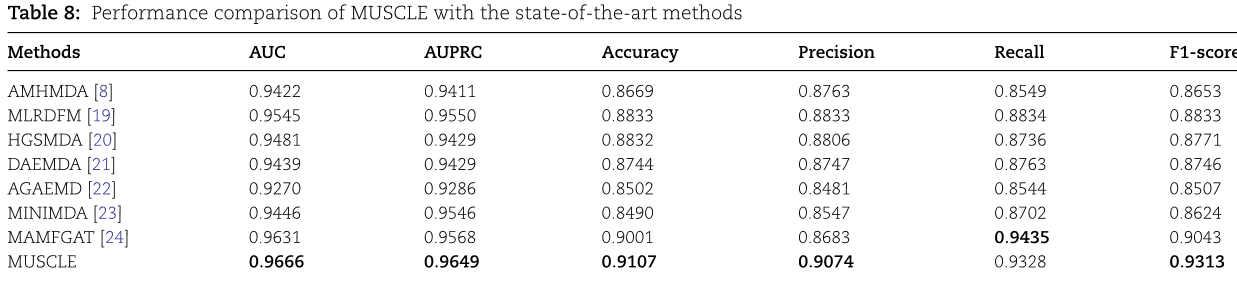
MLP层数：在保持其他参数不变的情况下，从2到5进行调整。图中显示了对这些参数获得的实验结果。随着隐藏层数量的增加，模型性能得到改善。然而，随着数量的继续增加，由于过拟合，性能下降。因此，我们设置MLP的数量为4

### 消融实验

不同的特征融合策略，包括基于平均值的特征融合策略（Fave）、基于点积的特征融合策略（Fdot）、基于图卷积神经网络（GCN）的特征融合策略（Fgcn）、在多尺度注意力特征融合模块中去除全局特征的策略（Flf）以及直接拼接特征融合策略（Fcat）。表6和图4A显示了不同特征融合策略的预测性能，其中我们的多尺度注意力特征融合模块（MUSCLE）取得了最佳性能。这种适当的特征融合策略使模型能够对潜在MDA具有更强的分类性能。



实验结论



我们方法的表现提升归功于两个因素。首先，信息量的增长提供了更多的方式来表征 miRNA 和疾病，从而帮助预测模型做出更准确的决策。其次，算法的改进使表征更加合理，并且可以从多个角度让预测模型从表征中获益更多。

在这项工作中，我们提出了一种计算方法（MUSCLE）来预测潜在的miRNA-疾病关联。首先，MUSCLE从miRNA与各种生物大分子之间的关系的整体视角出发，构建了多视图图，包括基于miRNA-药物-疾病、miRNA-mRNA-疾病和miRNA-lncRNA-疾病的关联图。然后，图注意力网络被用来为每个视图获取图拓扑特征。其次，MUSCLE有效地融合了多个图拓扑特征。此外，MUSCLE还同时考虑了miRNA和疾病 的先验属性知识，以获得更好的预测结果并解决冷启动问题。最后，学习到的表示被连接起来，并馈入MLP进行端到端训练和预测。为了评估MUSCLE的能力，我们分别进行了5-和10-折交叉验证实验。MUSCLE优于大多数当前最先进的模型。进一步的消融实验验证了我们特征组合和融合策略的有效性。

## 模型改进

此外，我们的方法仍然有很大的潜力可以改进一般化。例如，我们的方法可以扩展到一个多分类模型，它同时学习和预测多个MDA。此外，代谢组学信息、单细胞测序和空间转录组数据可以与miRNA和某些特定疾病联系起来。