Pure Mathematics 理论数学, 2023, 13(7), 1911-1924
Published Online July 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.137196

Hans汉斯

基于图神经网络的miRNA-疾病关联预测研究综述

张语凡

山东科技大学，数学与系统科学学院，山东 青岛

收稿日期：2023年6月4日；录用日期：2023年7月7日；发布日期：2023年7月14日

摘　要

miRNA与复杂疾病之间的关联有着重要意义，研究复杂疾病组织中的miRNA的非正常表达为人类攻克相关复杂疾病提供了一个可行的解决方案。将已有的复杂疾病与miRNA的关联数据进行数学抽象建模，并设计合理、高效的算法实现预测未知关联预测，已经成为复杂疾病与miRNA的关联研究的主要内容。本文对图神经网络在miRNA-疾病关联预测中所涉及的关键技术，包括基于图神经网络的表示学习算法研究和miRNA疾病关联预测框架的研究现状、存在的问题以及面临的挑战进行系统综述。

关键词

miRNA，疾病，关联预测，图神经网络

Prediction Review of miRNA-Disease Association Based on Graph Neural Network

Yufan Zhang

College of Mathematics and System Science, Shandong University of Science and Technology, Qingdao Shandong

Received: Jun. 4th, 2023; accepted: Jul. 7th, 2023; published: Jul. 14th, 2023

Abstract

The association between miRNAs and complex diseases is of great significance, and studying the non-normal expression of miRNAs in complex disease tissues provides a feasible solution for human to overcome the related complex diseases. Modeling the existing association data of complex diseases and miRNAs by mathematical abstraction and designing reasonable and efficient algo-

文章引用: 张语凡. 基于图神经网络的 miRNA-疾病关联预测研究综述[J]. 理论数学, 2023, 13(7): 1911-1924.
DOI: 10.12677/pm.2023.137196

rithms to achieve prediction of unknown association prediction have become the main content of association studies of complex diseases and miRNAs. This paper provides a systematic review of the key technologies involved in graph neural networks in miRNA-disease association prediction, including the current status, problems and challenges of research on graph neural network-based representation learning algorithm research and miRNA disease association prediction framework.

Keywords

miRNA, Disease, Associations Prediction, Graph Neural Network

1. 引言

复杂疾病缺乏清晰的遗传模式，所以无法估计个体得病的风险，并导致其难以诊断和治疗。近些年来，随着分子测序技术的不断发展，分子生物学征向数据驱动的方向发展，需要研究人员从数据的角度出发重新认识各类复杂疾病，并研究其病理、生理基础，从而提高复杂疾病的早期诊断、分期、治疗及预后的技术[1]。现代生物技术的发展为我们提供了不同层面的生物数据，将这些不同层面、不同维度的数据加以整合、关联、分析并应用于复杂疾病致病机制的研究都是复杂疾病病理生物学研究的一个重要方面[2]。

研究表明在多种肿瘤细胞中存在 miRNA 的非正常表达，这说明 miRNA 与肿瘤的发生有着密不可分的作用。在肿瘤相关基因或缺陷位点区域存在了 50%的 miRNA，这也为 miRNA 和癌症的关联给出了一个有力的证据。据估算，每个 miRNA 大概能控制数十个基因的表达，而每个基因的表达也受到多重 miRNA 的协同调控。研究还表明，每一种特定的癌组织都存在一种特定的 miRNA 对其起了关键的作用。

在最新版的人类 miRNA 与疾病关联数据库中，收录了与 800 多种疾病有关联的 1200 个人类 miRNA [3]，这些 miRNA 的数目还不到已发现的人类 miRNA 数目的一半。因此，复杂疾病与 miRNA 的关联问题亟待深入研究。然而，通过生物实验和临床试验去揭示疾病与 miRNA 的关联是相对漫长的过程，并且需要大量的人力物力成本[4]，如果没有可靠的 miRNA 作为候选验证目标，还会面临巨大的失败风险。因此，如果能够用高效且准确的计算方法，为下游实验提供有希望的候选 miRNA 参考目标，那么就可以提高验证效率，缩短发现周期，加快整个领域对复杂疾病的研究进展。特别是当前数据资源较为丰富、硬件算力较为充足的情况下，设计合理且实用的计算模型来预测潜在的复杂疾病与 miRNA 关联这一做法切实可行。随着大数据时代的来临，数学中的很多经典理论和方法如图论和组合优化等，在实际场景中得到了广泛的应用，尤其机器学习、神经网络等理论框架在各个领域大放异彩。

本文安排如下，第一部分对基于图神经网络的表示学习算法研究进行了系统的总结和归纳，第二部分对 MiRNA-疾病关联预测框架进行了分析。最后根据当前研究工作中存在的问题与不足，展望分析未来基于图神经网络的 miRNA-疾病关联预测的研究方向。

2. 基于图神经网络的表示学习算法研究

1997 年 Sperduti 等人首次将神经网络应用于有向无环图，引发了研究者们对图神经网络进行的早期研究[5]。Gori 等人(2005)最初提出了图神经网络的概念，Scarselli 等人(2009)和 Gallicchio 等人(2010)进一

步阐述了这一概念。这些早期的研究属于循环图神经网络(RecGNNs)的范畴。它们是通过迭代来传播节点的邻域信息的一种方式来学习目标节点的表达，直到它不再发生变化。后期经过许多研究者的努力，关于图神经网络的问题被不断地完善和发展。

一般图神经网络可以划分为五大类别：图卷积网络(Graph Convolution Networks, GCN)、图注意力网络(Graph Attention Networks)、图自编码器(Graph Autoencoders)、图生成网络(Graph Generative Networks)和图时空网络(Graph Spatial-temporal Networks)。

2.1. 图卷积网络

GCN 方法又可以分为两大类，基于谱(spectral-based)和基于空间(spatial-based)。基于谱的方法从图信号处理的角度引入滤波器来定义图卷积，其中图卷积操作被解释为从图信号中去除噪声。基于空间的方法将图卷积表示为从邻域聚合特征信息，当图卷积网络的算法在节点层次运行时，图池化模块可以与图卷积层交错，将图粗化为高级子结构。如图 1 所示，这种架构设计可用于提取图的各级表示和执行图分类任务。

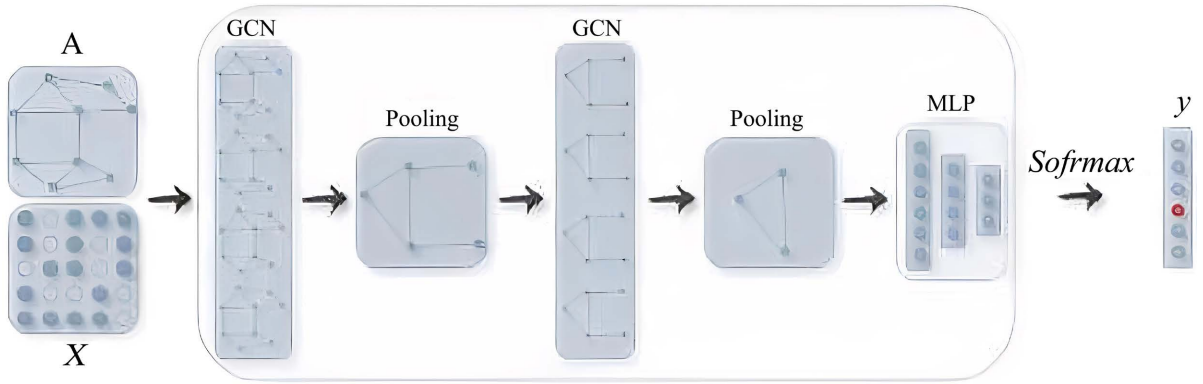

Figure 1. Prediction framework of graphical convolutional neural network [6]
图 1. 图卷积神经网络预测框架[6]

1) 基于频域

基于谱的方法从图信号处理的角度引入滤波器来定义图卷积，其中图卷积操作被解释为从图信号中去除噪声。在基于谱的图神经网络中，图被假定为无向图，无向图的一种鲁棒数学表示是正则化图拉普拉斯矩阵。现有的基于谱的图卷积网络模型有以下这些：Spectral CNN、Chebyshev Spectral CNN (ChebNet)、Adaptive Graph Convolution Network (AGCN) [6]。基于谱的图卷积神经网络方法的一个常见缺点是，它们需要将整个图加载到内存中以执行图卷积，这在处理大型图时是不高效的。

2) 基于空间域

基于空间的方法将图卷积表示为从邻域聚合特征信息，当图卷积网络的算法在节点层次运行时，图池化模块可以与图卷积层交错，将图粗化为高级子结构。基于空间的图卷积神经网络的思想主要源自于传统卷积神经网络对图像的卷积运算，不同的是基于空间的图卷积神经网络是基于节点的空间关系来定义图卷积的。基于空间的 GCN 可以进一步分为两类：recurrent-based 和 composition-based 的空间 GCN。recurrent-based 的方法使用相同的图卷积层来更新隐藏表示，composition-based 的方法使用不同的图卷积层来更新隐藏表示。作为最早的图卷积网络，基于谱的模型在许多与图相关的分析任务中取得了令人印象深刻的结果。这些模型在图信号处理方面有一定的理论基础。通过设计新的图信号滤波器可以从理论

上设计新的图卷积网络[6]。然而，基于谱的模型有着一些难以克服的缺点：

- 在效率方面，基于谱的模型的计算成本随着图的大小而急剧增加，因为它们要么需要执行特征向量计算，要么同时处理整个图，这使得它们很难适用于大型图。基于空间的模型有潜力处理大型图，因为它们通过聚集相邻节点直接在图域中执行卷积。计算可以在一批节点中执行，而不是在整个图中执行。当相邻节点数量增加时，可以引入采样技术来提高效率。
- 在一般性方面，基于谱的模型假定一个固定的图，使得它们很难在图中添加新的节点。另一方面，基于空间的模型在每个节点本地执行图卷积，可以轻松地在不同的位置和结构之间共享权重。
- 在灵活性方面，基于谱的模型仅限于在无向图上工作，有向图上的拉普拉斯矩阵没有明确的定义，因此将基于谱的模型应用于有向图的唯一方法是将有向图转换为无向图。基于空间的模型更灵活地处理多源输入，这些输入可以合并到聚合函数中。因此，近年来空间模型越来越受到关注。

2.2. 图注意力网络

图注意力网络(GAT)是一种基于空间的图卷积网络，它的注意机制是在聚合特征信息时，将注意机制用于确定节点邻域的权重。GAT 的图卷积运算定义为：

$$h_i^t = \sigma\left(\sum_{j\in N_i}\alpha\left(h_i^{t-1}, h_j^{t-1}\right)W^{t-1}h_j^{t-1}\right)$$

其中 $\alpha(\bullet)$是一个注意力函数，它自适应地控制相邻节点 j 对节点 i 的贡献。为了学习不同子空间中的注意力权重，GAT 还可以使用多注意力：

$$h_i^t = \|_{k=1}^K \sigma\left(\sum_{j\in N_i}\alpha_k\left(h_i^{t-1}, h_j^{t-1}\right)W_k^{t-1}h_j^{t-1}\right)$$

注意力机制如今已经被广泛地应用到了基于序列的任务中，它的优点是能够放大数据中最重要的部分的影响。这个特性已经被证明对许多任务有用，例如机器翻译和自然语言理解。如今融入注意力机制的模型数量正在持续增加，图神经网络也受益于此，它在聚合过程中使用注意力，整合多个模型的输出，并生成面向重要目标的随机行走。

图注意力网络优化了图卷积神经网络的几个缺陷：1) 图卷积神经网络擅长处理 transductive 任务，无法完成 inductive 任务。图卷积神经网络进行图卷积操作时需要拉普拉斯矩阵，而拉普拉斯矩阵需要知道整个图的结构，故无法完成 inductive 任务，而图注意力网络仅需要一阶邻居节点的信息(transductive 指的是训练、测试使用同一个图数据，inductive 是指训练、测试使用不同的图数据)；2) 图卷积神经网络对于同一个节点的不同邻居在卷积操作时使用的是相同的权重，而图注意力网络则可以通过注意力机制针对不同的邻居学习不同的权重。

除此之外，比较常用的注意力网络还有门控注意力网络(GaAN)和图形注意力模型(GAM)。

1) 门控注意力网络。GaAN 不同于传统的多头注意机制(它均衡的消耗所有的注意头)，它使用一个卷积子网络来控制每个注意头的重要性。门控注意力网络(GAAN)还采用了多头注意力机制来更新节点的隐藏状态。然而，GAAN 并没有给每个 head 部分配相等的权重，而是引入了一种自注意机制，该机制为每个 head 计算不同的权重。更新规则定义为其中 $\varnothing_o$ 是反馈神经网络，而 g_i^k 是第 k 个注意力 head 的注意力权重。

$$h_i^t = \varnothing_o\left(x_i \oplus \|_{k=1}^K g_i^k \sum_{j\in N_i}\alpha_k\left(h_i^{t-1}, h_j^{t-1}\right)\varnothing_v\left(h_j^{t-1}\right)\right)$$

2) 图注意力模型。图形注意力模型(GAM)提供了一个循环神经网络模型，以解决图形分类问题，通过自适应地访问一个重要节点的序列来处理图的信息。图形注意力模型(GAM)提供了一个循环神经网络模型，以解决图形分类问题，通过自适应地访问一个重要节点的序列来处理图的信息。GAM 模型被定义为下式，其中 f_h 是一个 LSTM 网络，f_s 是一个 step network，它会优先访问当前节点 v_{t-1} 优先级高的邻居并将它们的信息进行聚合。

$$h_t = f_h\left(f_s\left(r_{t-1}, v_{t-1}, g; \theta_s\right), h_{t-1}; \theta_h\right)$$

除了在聚集特征信息时将注意力权重分配给不同的邻居节点，还可以根据注意力权重将多个模型集合起来，以及使用注意力权重引导随机行走。尽管 GAT 和 GAAN 在图注意网络的框架下进行了分类，但它们也可以同时被视为基于空间的图形卷积网络。GAT 和 GAAN 的优势在于，它们能够自适应地学习邻居的重要性权重。然而，计算成本和内存消耗随着每对邻居之间的注意权重的计算而迅速增加。

2.3. 图自编码器

图自动编码器是一类图嵌入方法，其目的是利用神经网络结构将图的顶点表示为低维向量，其结构如图 2 所示。自编码器的是通过减少隐藏层神经元个数来实现重构样本，自编码器为了尽可能复现输入数据，其隐藏层必须捕捉输入数据的重要特征，从而找到能够代表原数据的主要成分。其主要是为图中节点找寻合适的 Embedding 向量，并通过 Embedding 向量实现图重构。其中获取到的节点 Embedding 可以用于支撑下游任务。

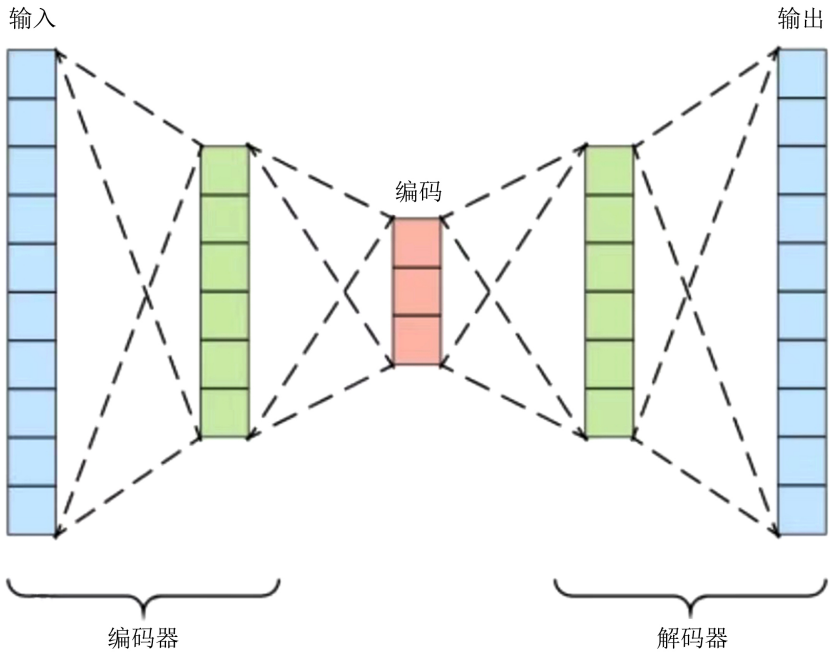

Figure 2. Self-encoder framework [7]
图 2. 自编码器框架[7]

随着图数据的逐渐增多，在图领域也运用到了大量的非概率模型的图自编码器,。最近，研究人员已经探索了将 GCN 作为编码器的用途，将 GCN 与 GAN 结合起来，或将 LSTM 与 GAN 结合起来设计图自动编码器。目前基于 GCN 的自编码器的方法主要有：Graph Autoencoder (GAE)和 Adversarially Regu-

larized Graph Autoencoder (ARGA)。

- Graph Autoencoder，GAE 是将 GCN 和自编码器的结合，其公式为：

$$\hat{A} = \sigma\left(ZZ^T\right)$$

$$Z = \mathrm{GCN}\left(X, A\right)$$

- ARGA 将对抗训练方案作为一个额外的正则化项纳入 GAE。整个架构图 3 所示。具体来说，编码器用作生成器，判别器的目的是区分潜在表示是来自生成器还是来自先验分布。这样，自动编码器就被强制匹配先验分布以作为正则化。

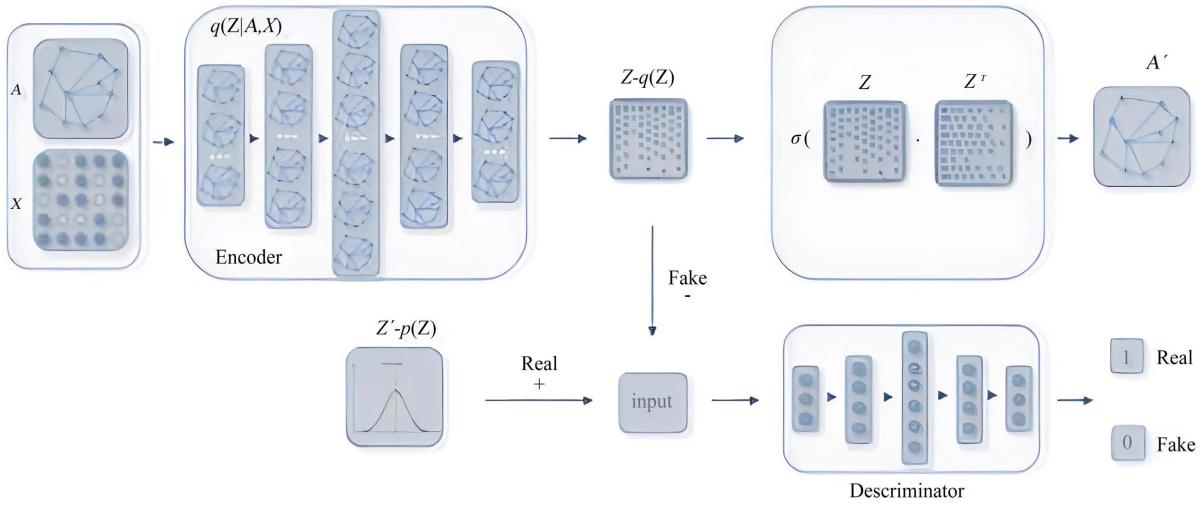

Figure 3. GAE graph autoencoder
图 3. 图自编码器框架

图自编码器的其它变体有：Network Representations with Adversarially Regularized Autoencoders (NetRA)，Deep Neural Networks for Graph Representations (DNGR)，Structural Deep Network Embedding (SDNE)，Deep Recursive Network Embedding (DRNE)。DNGR 和 SDNE 学习仅给出拓扑结构的节点嵌入，而 GAE、ARGA、NetRA、DRNE 用于学习当拓扑信息和节点内容特征都存在时的节点嵌入。图自动编码器的一个挑战是邻接矩阵 A 的稀疏性，这使得解码器的正条目数远远小于负条目数。为了解决这个问题，DNGR 重构了一个更密集的矩阵，即 PPMI 矩阵，SDNE 对邻接矩阵的零项进行惩罚，GAE 对邻接矩阵中的项进行重加权，NetRA 将图线性化为序列。

2.4. 图生成对抗网络

图表示学习，也称为网络嵌入，目的是将图(网络)中的每个顶点表示为低维向量，这有助于对顶点和边缘进行网络分析和预测。学习到的嵌入能够帮助广泛的现实应用程序，如链路预测、节点分类、推荐、可视化、知识图表示等。图表示学习的目的是将图中的每个顶点嵌入到一个低维向量空间中。现有的图形表示学习方法可分为两类：学习图中潜在连通性分布的生成模型，以及预测一对顶点之间存在边的概率的判别模型。图生成对抗网络将上述两类方法结合在一起，其中生成模型和判别模型是一种极大极小决策的博弈。此外，在考虑生成模型的实现时，提出了一种新的图形 SoftMax 来克服传统 SoftMax 函数的局限性，它能满足规范化、图结构感知和计算效率的要求。其模型框架如图 4 所示：

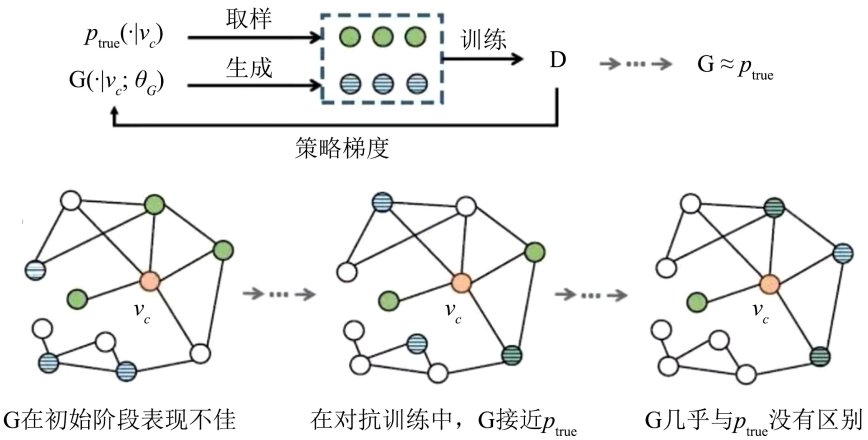

Figure 4. Graph Adversarial Methods
图 4. 图对抗网络框架

图生成网络的目标是在给定一组观察到的图的情况下生成新的图。图生成网络的许多方法都是特定于领域的。例如，在分子图生成中，一些工作模拟了称为 SMILES 的分子图的字符串表示。在自然语言处理中，生成语义图或知识图通常以给定的句子为条件。最近，人们提出了几种通用的方法。一些工作将生成过程作为节点和边的交替形成因素，而另一些则采用生成对抗训练。这类方法要么使用 GCN 作为构建基块，要么使用不同的架构。基于 GCN 的图生成网络主要有：

Molecular Generative Adversarial Networks (MolGAN)：将 relational GCN、改进的 GAN 和强化学习(RL)目标集成在一起，以生成具有所需属性的图。GAN 由一个生成器和一个鉴别器组成，它们相互竞争以提高生成器的真实性。在 MolGAN 中，生成器试图提出一个伪图及其特征矩阵，而鉴别器的目标是区分伪样本和经验数据。此外，还引入了一个与鉴别器并行的奖励网络，以鼓励生成的图根据外部评价器具有某些属性。

Deep Generative Models of Graphs (DGMG)：利用基于空间的图卷积网络来获得现有图的隐藏表示。生成节点和边的决策过程是以整个图的表示为基础的。简而言之，DGMG 递归地在一个图中产生一个节点，直到达到某个停止条件。在添加新节点后的每一步，DGMG 都会反复决定是否向添加的节点添加边，直到决策的判定结果变为假。如果决策为真，则评估将新添加节点连接到所有现有节点的概率分布，并从概率分布中抽取一个节点。将新节点及其边添加到现有图形后，DGMG 将更新图的表示。

其它架构的图生成网络主要有：

GraphRNN：通过两个层次的循环神经网络的深度图生成模型。图层次的 RNN 每次向节点序列添加一个新节点，而边层次 RNN 生成一个二进制序列，指示新添加的节点与序列中以前生成的节点之间的连接。为了将一个图线性化为一系列节点来训练图层次的 RNN，GraphRNN 采用了广度优先搜索(BFS)策略。为了建立训练边层次的 RNN 的二元序列模型，GraphRNN 假定序列服从多元伯努利分布或条件伯努利分布。

NetGAN：Netgan 将 LSTM 与 Wasserstein-GAN 结合在一起，使用基于随机行走的方法生成图形。GAN 框架由两个模块组成，一个生成器和一个鉴别器。生成器尽最大努力在 LSTM 网络中生成合理的随机行走序列，而鉴别器则试图区分伪造的随机行走序列和真实的随机行走序列。训练完成后，对一组随机行走中节点的共现矩阵进行正则化，我们可以得到一个新的图。

2.5. 图时空网络

图时空网络同时捕捉时空图的时空相关性。时空图具有全局图结构，每个节点的输入随时间变化。

例如，在交通网络中，每个传感器作为一个节点连续记录某条道路的交通速度，其中交通网络的边由传感器对之间的距离决定。图形时空网络的目标可以是预测未来的节点值或标签，或者预测时空图标签。最近的研究仅仅探讨了 GCNs 的使用，GCNs 与 RNN 或 CNN 的结合，以及根据图结构定制的循环体系结构。

目前图时空网络的模型主要有：Diffusion Convolutional Recurrent Neural Network (DCRNN)，CNN-GCN，Spatial Temporal GCN (ST-GCN)，Structural-RNN。

3. MiRNA-疾病关联预测框架

具体来说，miRNA-疾病潜在关联预测模型可分为四类，即基于分数函数的模型、基于复杂网络算法的模型、基于机器学习的模型和基于多种生物信息的模型。基于评分函数的模型对 miRNA 和疾病相关的训练数据采用概率分布或统计分析，以构建评分函数，对潜在的 miRNA-疾病关联进行排序，如图 5。基于复杂网络算法的模型主要基于不同角度的 miRNA 相似网络和疾病相似网络。基于机器学习的预测模型旨在通过提取有效特征或解决特定优化问题，利用强大的机器学习算法进行可靠预测[8]。多个基于生物信息的模型考虑了 miRNA 相关基因和疾病相关的多种类型，如 miRNA 基因和疾病－蛋白质关联，并试图通过这些中间介质协会构建 miRNA 与疾病之间的关联。

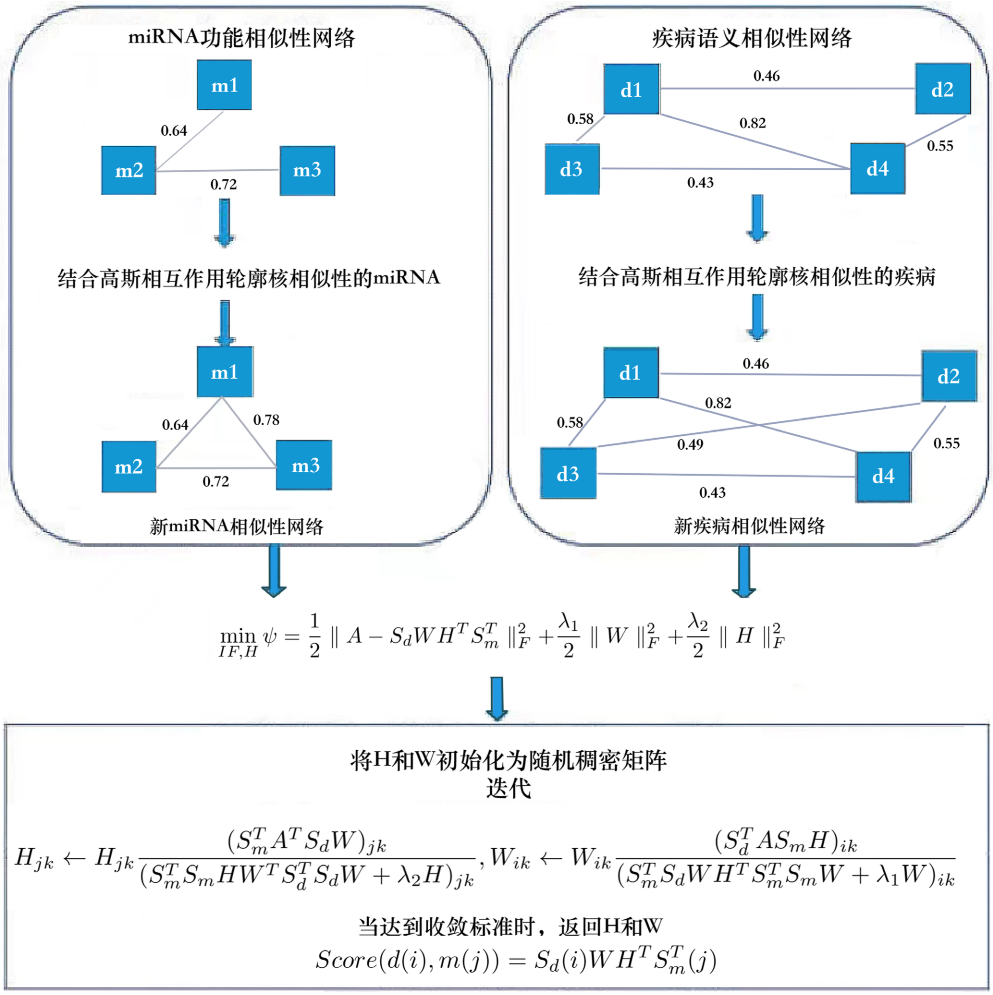

Figure 5. IMCMDA flowchart
图 5. IMCMDA 流程图

3.1. 基于分数的模型

Jiang 等人提出了一种新的计算方法，通过对 miRNA 功能相似性网络和人类表型 miRNA 组网络应用评分系统来评估 miRNA 可能参与特定疾病的概率，从而预测潜在的 miRNA-疾病关联。Shiet 等通过考虑蛋白质－蛋白质相互作用网络中 miRNA 靶点和疾病基因之间的功能关联，提出了一个计算模型[9]。miRNA 靶点和疾病基因被用作在蛋白质－蛋白质相互作用网络上实现随机游走的种子，以计算 P 值并评估 miRNA 与疾病之间的潜在关联。Chen 等人开发了一种基于 RNA-疾病关联预测(WBSMDA)的 miRNA 与疾病之间评分模型[10]。该模型通过定义 miRNA 和疾病对之间的“Within-Scores”和“Between-Scores”，并整合两个分数获得潜在 miRNA 疾病关联推断的最终分数。Pasquier 和 Garde’s (2016)提出了 MiRAI 模型，以确定潜在的 miRNA-疾病关联[11]。对于每个 miRNA，MiRAI 利用了五个关键信息：其已知的相关疾病、其靶 mRNAs、其家族成员、与邻居的距离以及文本格式的相关研究摘要来构建高维向量空间。此外，疾病和 miRNA 在载体空间中由载体表示。在降维后，MiRAI 可以通过计算与疾病载体的距离来获得与疾病相关的 miRNA 的排序列表。Zhu 等提出了基于路径的 MiRNA 疾病关联(PBMDA)预测模型[12]。该模型构建了一个由三个子图组成的异构图，并进一步采用深度优先搜索算法来推断潜在的 miRNA-疾病关联。该模型将 miRNA 与疾病之间的所有路径得分相加，计算关联可能性，根据得分获得最有可能的候选基因。Chen 等提出了一种新的用于 MiRNA-疾病关联预测的诱导矩阵补全模型(IMCMDA) [13]。主要思想是基于已知的关联以及整合的 miRNA 相似性和疾病相似性来补全缺失的 miRNA-疾病关联。综上所述，相似性得分计算方法的主题是构建一个网络模型，并使用不同的方法来度量网络中节点之间的相似性，以预测 miRNA 与疾病的相互作用，其中大多数受到所构建网络模型的质量和节点之间不完全关系的限制。

3.2. 基于机器学习的模型

该类方法主要涉及了机器学习或深度学习领域的一些方法论，利用丰富的 miRNA、基因、疾病等相关数据作为特征来设计预测方法。Li 等人基于图卷积神经网络设计了一种神经诱导矩阵补全模型(NIMCGCN)来预测 miRNA 与疾病的未知关联[14]。该方法首先利用图卷积网络从 miRNA 和疾病类似性网络中学习 miRNA 和疾病的潜在特征表示，然后将学习到的特征输入到神经诱导矩阵补全(NIMC)模型中，生成完备的 miRNA 与疾病关联矩阵。该模型中的参数是基于已知的 miRNA 疾病关联数据，以有监督的端到端方式学习得到的。Liang 等人提出了一种基于自适应多视图多标签学习(AMVML)的新方法来预测与疾病相关的候选 miRNA，并且从理论上证明了 AMVML 方法的收敛性及收敛速度[15]。

Chen 等人开发了一种基于半监督学习的正则化最小二乘法计算模型(RLSMDA)，用于推断人类 MiRNA-疾病关联[16]。在 RLSMDA 模型中，假定的 miRNA-疾病关联是通过疾病和 miRNA 空间中的组合分类器产生的。Chen 等人开发了基于 k-最近邻的 MiRNA-疾病关联预测计算模型(RKNNMDA)，通过集成 k-最近邻(KNN)算法和 SVM 排序模型来预测潜在的 MiRNA-疾病关联[17]。具体来说，他们引入了 SVM 排序模型，这是 SVM 算法的一个变体，通过从训练数据集中提取特殊特征对先前排序的邻居进行排序，如图 6。Chen 等人基于用于推断多种类型 MiRNA-疾病关联的限制性 Boltzmann 机器(RBM)，开发了用于多种类型 MiRNA-疾病关联预测的限制性 Boltzmann 机器模型(RBMMMDA) [18]。RBMMMDA 使用 RBM (深度学习的核心)提供一个自包含的框架来直接获取竞争分类器。Pasquieret 等人开发了基于奇异值分解(基于 SVD)的向量空间模型，通过考虑多个 miRNA 相关信息源来推断 miRNA 与疾病的关联[11]。该方法整合了来自多种 miRNA 相关信息的五个不同矩阵，包括 miRNA-disease，miRNA-neighbor，miRNA-target，miRNA-word 和 miRNA-family 关联。使用组合矩阵上的 SVD 提取特征向量。最后，该模

型通过优先考虑 miRNA 载体与疾病载体的余弦距离，得出了与疾病相关的 miRNA 的排序列表。Luo 等人提出一种名为 KRLSM 的预测模型，利用基于异质性组学数据的 Kronecker RLS 预测 mirna 与疾病的关联[8]。他们首先采用了 Kronecker 乘积的代数性质，并将 miRNA 空间和疾病空间组合成一个完整的 miRNA-疾病空间，以便使用 Kronecker 乘积相似矩阵进行预测。

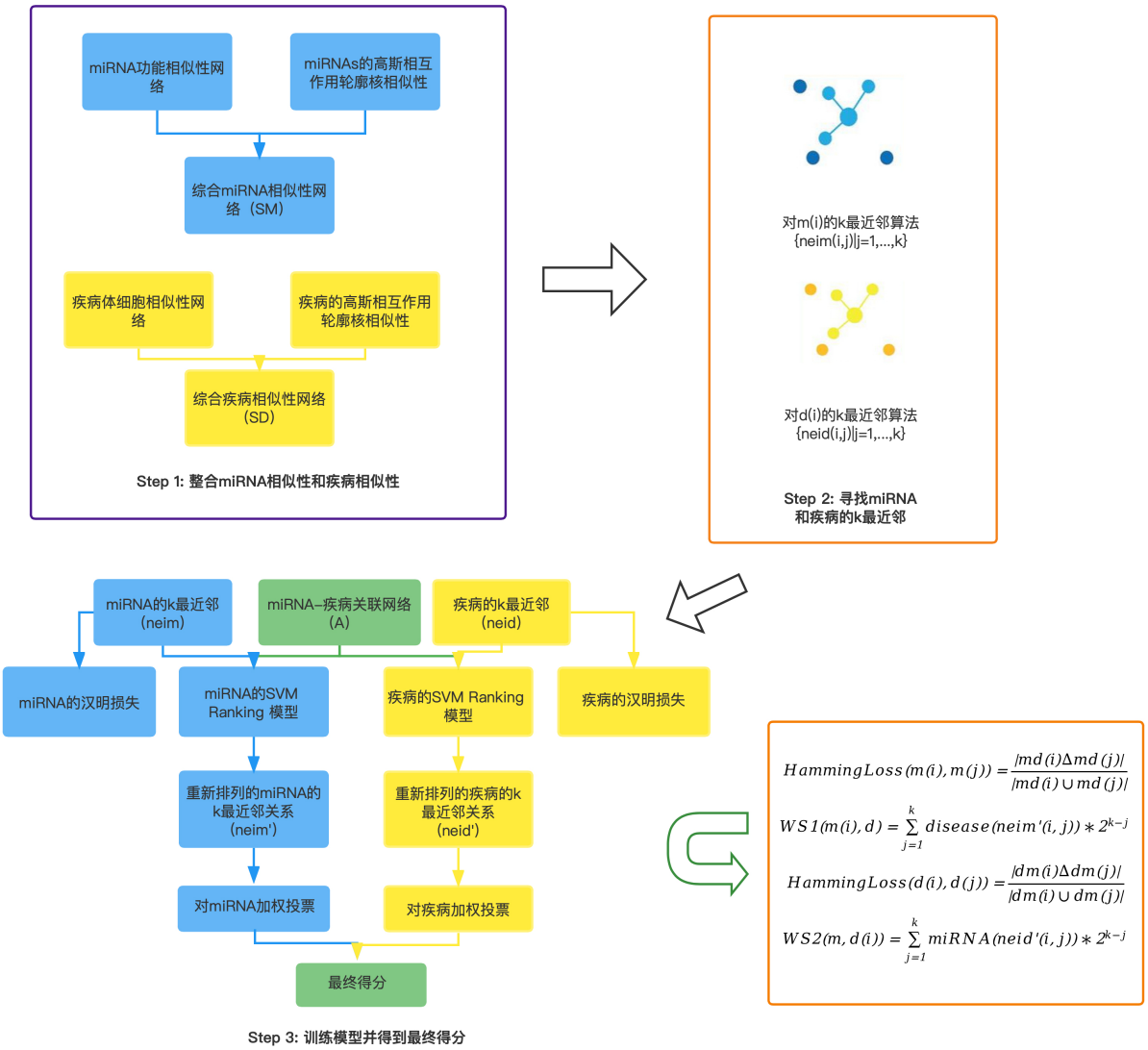

Figure 6. RKNNMDA flowchart
图 6. RKNNMDA 流程图

3.3. 基于深度学习的模型

近年来，深度学习越来越多地应用于这一领域。Xuan 等人[19]提出了一种基于双卷积神经网络(CNN)的模型，称为 CNNMDA，用于预测。两种 CNN 模型用于从原始结构和全局网络中提取特征，而后者具有通过非负矩阵分解(NMF)方法获得的 miRNA 和疾病的低维特征。最后的分数与这两部分结合在一起。Penge 等人引入了一个带有疾病和 miRNA 网络的基因层，通过自动编码器学习所有 miRNA 和疾病对的表示，然后应用 CNN 预测最终得分[20]。Zhang 等人提出了一种称为 VAEMDA 的无监督学习模型，用

于预测潜在的微相关疾病[21]。V-AEMDA 首先通过将人类 miRNA 疾病关联矩阵与 miRNA 相似矩阵和疾病相似矩阵剪接，构建了两个矩阵。之后，VAEMDA 应用两个变分自动编码器(VAE)模型学习隐藏 miRNA 和疾病表示。最后，VAEMDA 通过 VAE 模型将两个分数组合在各自的重建矩阵中，如图 7。Li 等提出了一个基于 GCN 的模型，名为 NIMCGCN，用于提取疾病和 miRNA 的代表性，然后应用神经诱导矩阵复合模型来预测 miRNA 与疾病之间的潜在联系[22]。Chen 等人提出了一种基于深度表示的计算模型，称为 DRMDA，该模型通过 SVM 分类器推断潜在的关联[23]。DRMDA 构建了一个堆叠式自动编码器，从集成的相似性网络中提取已知 miRNA 疾病对的特征。

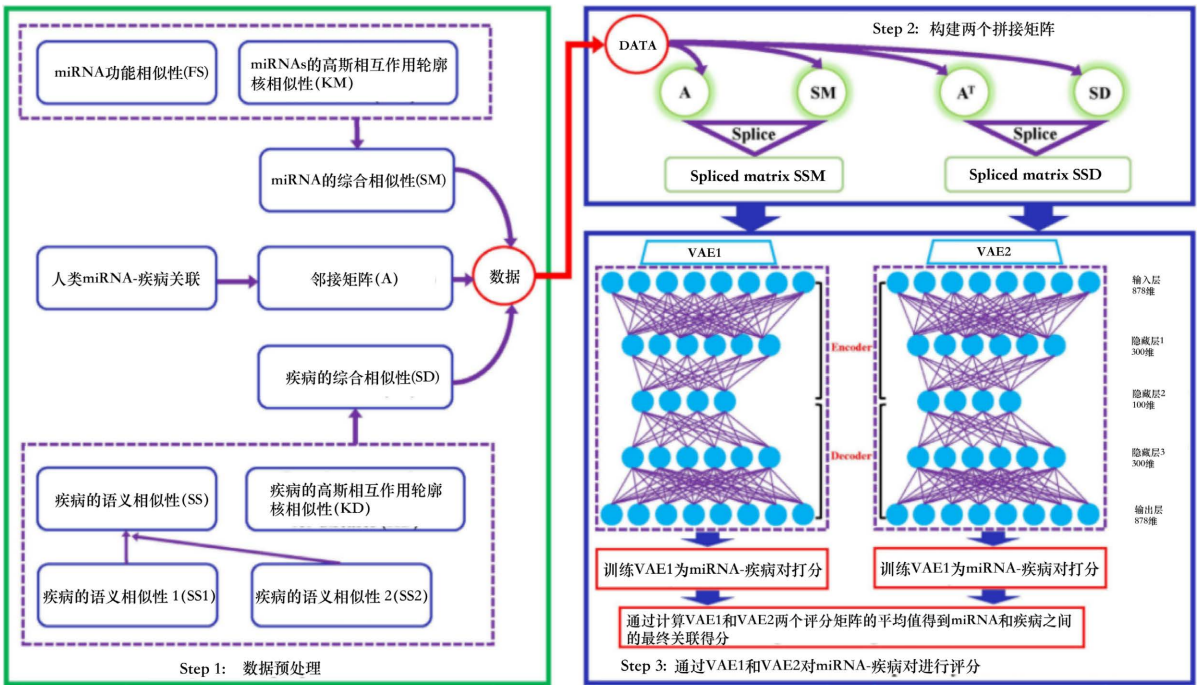

Figure 7. VAEMDA flowchart
图 7. VAEMDA 流程图

3.4. 基于多种生物信息的模型

网络驱动方法主要是依据 miRNA、疾病、基因、环境、因子、蛋白质和小分子化合物等之间的关联数据，构建相应的生物关联网络模型，然后依据网络的拓扑结构，结合领域知识，设计预测算法。Chen 等人提出了一个包含 miRNA、lncRNA 和疾病的三层异质关联网络模型(TLHNMDA)来预测可能的 miRNA 与疾病之间的关联[24]。作者将多源的数据信息进行整合，根据 miRNA 与 lncRNA、疾病与 miRNA 之间的关联数据以及对应的相似性信息，构建出一个三层异质关联网络，并基于网络拓扑结构(路径信息)设计了全局优化算法预测 miRNA 与疾病之间的未知关联。Zeng 等人提出了一种双层网络的结构扰动方法来预测 miRNA 与疾病之间的未知关联，利用 HumanNet 数据库中的对数似然评分构建了疾病之间的类似性网络，利用 miRTarBase 数据库中的 miRNA 靶点信息构建了 miRNA 之间的类似性网络，在构建的双层网络模型中先用结构一致性指标去评估连边的可预测性，然后设计了结构扰动算法(SPM)进行关联预测[25]。Sun 等人考虑到已有的计算模型过度依赖领域中的多种类似性信息，而很少关注网络本身的拓扑结构，因此提出了一种完全依赖网络拓扑结构的类似性计算方法(NTSMDA)来预测潜在 miRNA 与疾病的关联[26]。

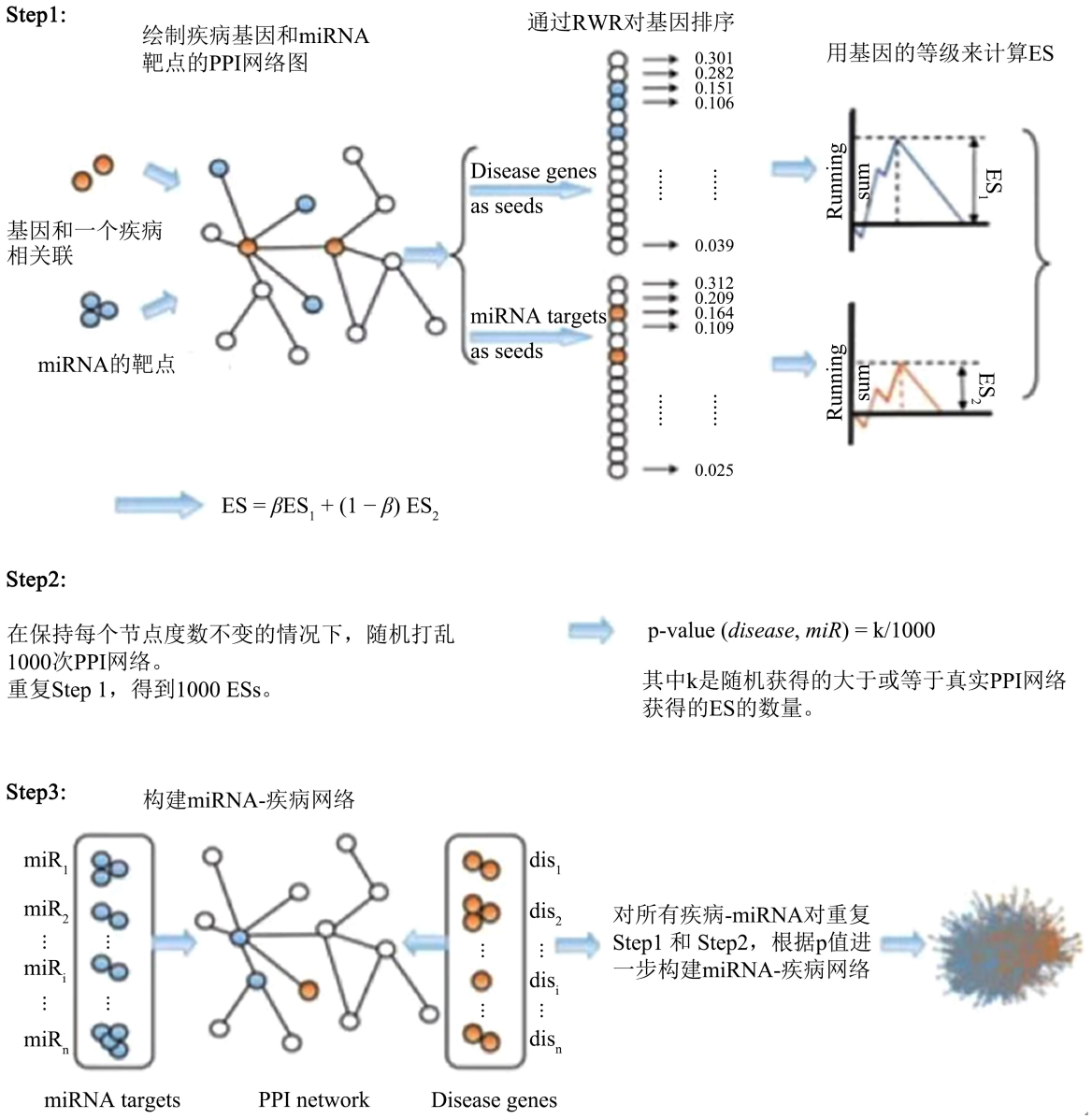

Figure 8. Flowchart for joining the PPI network
图 8. 加入 PPI 网络的流程图

上述三小节中回顾的两种 miRNA-疾病关联预测模型仅使用与 miRNA 或疾病直接相关的单一信息，包括实验支持的 miRNA-疾病关联。然而，由于实验鉴定的困难，已知的 miRNA-疾病的数量仍然不足。考虑到这一有限的数据量，有一些预测模型是通过考虑其他类型的先前生物信息而提出的，如蛋白质和靶基因相关网络，它们可以为推断 miRNA 与疾病的关联提供有价值的见解。

在不使用任何已知的 miRNA 疾病关联的情况下，Zhao 等人创新性地构建了一个 miRNA-lncRNA 疾病网络(DCSMDA)，该网络整合了 miRNA-lncRNA 关联和 lncRNA 疾病关联，以间接预测 miRNA 疾病关联，如图 8 [27]。Mork 等人提出了基于蛋白质驱动的 miRNA-疾病关联预测模型(miRPD) [28]。其中 miRNA-蛋白质－疾病关联被明确推断。除了将 miRNA 与疾病联系起来外，它还直接暗示了相关的潜在蛋白质，这些蛋白质可以用来形成可以通过实验验证的假设。miRNA 与疾病的推论是通过将已知和预测

的miRNA-蛋白质关联与从文献中挖掘的蛋白质－疾病关联文本进行耦合而得出的。并提出了评分方案，使之能够根据可靠性对从治疗和预测的miRNA靶点推断的miRNA-疾病关联进行排序，从而创建关联的高和中等置信集。Shietd等提出了一个推断miRNA与疾病关联的预测模型，该模型主要考虑蛋白质－蛋白质相互作用(PPI)网络中miRNA靶点与疾病基因之间的关联，该模型将疾病基因和miRNA靶点定位到PPI网络上，并分别以疾病基因和miRNA靶点作为种子节点，在PPI网络上实现RWR，测量富集分数[9]。Xuet等人提出了一个预测模型，通过构建miRNA与靶基因之间以及靶基因与疾病之间的相互作用网络，对与多种疾病相关的最有潜力的miRNA进行优先排序和识别[29]。具体而言，该模型从TCGA和GEO数据库收集miRNA-mRNA相互作用，并通过实施七种预测算法，将具有反向相关性的相互作用进一步用于筛选上下文相关的miRNA-靶基因相互作用。为了获得miRNA-疾病关联的相关性得分，该模型整合了三种类型的生物信息，包括GO子本体生物过程、KEGG路径信息和蛋白质相互作用网络中的平均最短路径(ASP)。Lanet等提出了一个名为KBMFMDI的计算框架。通过整合多种数据资源来衡量疾病相似性和miRNA相似性，从而推断miRNA与疾病之间的关系。此外，基于多核学习的全局方法用于预测miRNA与疾病的潜在关系。

4. 总结与展望

目前深度学习在miRNA-疾病潜在关联预测领域的应用比较多，但是其模型性能还有待于提升。基于图卷积神经网络的miRNA-疾病关联预测方法研究有着非常广阔的应用前景，技能提高，性能又能节约生物实验成本。在基于图卷积神经网络的miRNA-疾病关联预测方法研究方面，以下问题亟待解决：

1) 异质图简化技术有待提高。在以往的研究中，在利用miRNA相似度网络、疾病相似度网络和已知miRNA-疾病关联网络构建异质图后，并未对网络进行简化。但是许多miRNA之间的相似度、疾病之间的相似度特别小，这样的边本质上是无效的，因此应该去除。但是以往的研究往往忽略了这一重要部分，在理论分析上是欠缺的。

2) 边信息学习方法效果有待提升。图的两个基本元素是节点和边，节点携带着自身的特征信息，边同样携带着部分信息。在以往的研究中，大多重点都放在了节点上，边的信息都被忽略掉。

3) 有效获取miRNA及疾病嵌入图的节点特征。使用图神经网络处理异构图旨在获得节点嵌入表示，获得嵌入旨在通过保留图的网络拓扑结构和节点内容信息，将图中顶点表示为低维向量，以便使用简单的机器学习算法进行处理。图的结构一般来说是十分不规则的，可以认为是无限维的一种数据，所以它没有平移不变性。如何有效提取图中节点及边的特征也是该领域研究的重点之一。

参考文献

[1] Röst, H.L., Liu, Y., D'Agostino, G., *et al*. (2016) TRIC: An Automated Alignment Strategy for Reproducible Protein Quantification in Targeted Proteomics. *Nature Methods*, **13**, 777-783. https://doi.org/10.1038/nmeth.3954

[2] Jr Lowe, W.L. and Reddy, T.E. (2015) Genomic Approaches for Understanding the Genetics of Complex Disease. *Genome Research*, **25**, 1432-1441. https://doi.org/10.1101/gr.190603.115

[3] Li, Y., Qiu, C. and Tu, J. (2014) HMDD v2.0: A Database for Experimentally Supported Human microRNA and Disease Associations. *Nucleic Acids Research*, **42**, D1070-D1074. https://doi.org/10.1093/nar/gkt1023

[4] Chen, X., Huang, L., Xie, D. and Zhao, Q. (2018) EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association Prediction. *Cell Death & Disease*, **9**, Article No. 3. https://doi.org/10.1038/s41419-017-0003-x

[5] Sperduti, A. and Starita, A. (1997) Supervised Neural Networks for the Classification of Structures. *IEEE Transactions on Neural Networks*, **8**, 714-735. https://doi.org/10.1109/72.572108

[6] Wu Z., Pan, S., Chen, F., *et al*. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. https://doi.org/10.1109/TNNLS.2020.2978386

[7] https://zhuanlan.zhihu.com/p/46067799

[8] Luo, J., Xiao, Q., Liang, C., *et al*. (2017) Predicting MicroRNA-Disease Associations Using Kronecker Regularized Least Squares Based on Heterogeneous Omics Data. *IEEE Access*, **5**, 2503-2513. https://doi.org/10.1109/ACCESS.2017.2672600

[9] Shi, H., Xu, J., Zhang, G., *et al*. (2013) Walking the Interactome to Identify Human miRNA-Disease Associations through the Functional Link between miRNA Targets and Disease Genes. *BMC Systems Biology*, **7**, Article No. 101. https://doi.org/10.1186/1752-0509-7-101

[10] Chen, X., Yan, C.C., Zhang, X., *et al*. (2016) WBSMDA: Within and Between Score for MiRNA-Disease Association Prediction. *Scientific Reports*, **6**, Article No. 21106. https://doi.org/10.1038/srep21106

[11] Pasquier, C. and Gardès, J. (2016) Prediction of miRNA-Disease Associations with a Vector Space Model. *Scientific Reports*, **6**, Article No. 27036. https://doi.org/10.1038/srep27036

[12] You, Z., Huang, Z.A., Zhu, Z.X., *et al*. (2017) PBMDA: A Novel and Effective Path-Based Computational Model for miRNA-Disease Association Prediction. *PLOS Computational Biology*, **13**, e1005455. https://doi.org/10.1371/journal.pcbi.1005455

[13] Xing, C., Lei, W., Jia, Q., *et al*. (2018) Predicting miRNA-Disease Association Based on Inductive Matrix Completion. *Bioinformatics*, **34**, 4256-4265.

[14] Jin, L., Sai, Z., Tao, L., *et al*. (2020) Neural Inductive Matrix Completion with Graph Convolutional Networks for miRNA-Disease Association Prediction. *Bioinformatics*, **36**, 2538-2546.

[15] Liang, C., Yu, S. and Luo, J. (2019) Adaptive Multi-View Multi-Label Learning for Identifying Disease-Associated Candidate miRNAs. *PLOS Computational Biology*, **15**, e1006931. https://doi.org/10.1371/journal.pcbi.1006931

[16] Chen, X. and Yan, G.Y. (2014) Semi-Supervised Learning for Potential Human microRNA-Disease Associations Inference. *Scientific Reports*, **4**, Article No. 5501. https://doi.org/10.1038/srep05501

[17] Chen, X., Wu, Q.F. and Yan, G.-Y. (2017) RKNNMDA: Ranking-Based KNN for MiRNA-Disease Association Prediction. *RNA Biology*, **14**, 952-962.

[18] Chen, X., Clarence, Y., Zhang, X., *et al*. (2015) RBMMMDA: Predicting Multiple Types of Disease-microRNA Associations. *Scientific Reports*, **5**, Article No. 13877. https://doi.org/10.1038/srep13877

[19] Xuan, P., Sun, H., Wang, X., Zhang, T. and Pan, S. (2019) Inferring the Disease-Associated miRNAs Based on Network Representation Learning and Convolutional Neural Networks. *International Journal of Molecular Sciences*, 20, Article ID: 3648. https://doi.org/10.3390/ijms20153648

[20] Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., Shang, X. and Wei, Z. (2019) A Learning-Based Framework for miRNA-Disease Association Identification Using Neural Networks. *Bioinformatics*, **35**, 4364-4371. https://doi.org/10.3390/ijms20153648

[21] Zhang, L., Chen, X. and Yin, J. (2019) Prediction of Potential miRNA-Disease Associations through a Novel Unsupervised Deep Learning Framework with Variational Autoencoder. *Cell*, **8**, Article ID: 1040. https://doi.org/10.3390/cells8091040

[22] Li, J., Zhang, S., Liu, T., Ning, C., Zhang, Z. and Zhou, W. (2020) Neural Inductive Matrix Completion with Graph Convolutional Networks for miRNA-Disease Association Prediction. *Bioinformatics*, **36**, 2538-2546. https://doi.org/10.1093/bioinformatics/btz965

[23] Chen, X., Gong, Y., Zhang, D.H., You, Z.H. and Li, Z.W. (2018) DRMDA: Deep Representations-Based miRNA-Disease Association Prediction. *Journal of Cellular and Molecular Medicine*, **22**, 472-485. https://doi.org/10.1111/jcmm.13336

[24] Chen, X., Qu, J. and Yin, J. (2018) TLHNMDA: Triple Layer Heterogeneous Network Based Inference for miRNA-Disease Association Prediction. *Frontiers in Genetics*, **9**, Article ID: 234. https://doi.org/10.3389/fgene.2018.00234

[25] Zeng, X., Liu, L., Lü, L. and Zou, Q. (2013) Prediction of Potential Disease-Associated microRNAs Using Structural Perturbation Method. *Bioinformatics*, **34**, 2425-2432.

[26] Sun, D., Ao, L., Feng, H., *et al*. (2016) NTSMDA: Prediction of miRNA-Disease Associations by Integrating Network Topological Similarity. *Molecular Biosystems*, **12**, Article ID: 2224. https://doi.org/10.1039/C6MB00049E

[27] Zhao, H., Kuang, L., Wang, L., *et al*. (2018) Prediction of microRNA-Disease Associations Based on Distance Correlation Set. *BMC Bioinformatics*, **19**, Article ID: 141. https://doi.org/10.1186/s12859-018-2146-x

[28] Søren, M., *et al*. (2014) Protein-Driven Inference of miRNA-Disease Associations. *Bioinformatics*, **30**, 392-397.

[29] Xu, C., Ping, Y., Xiang, L., *et al*. (2014) Prioritizing Candidate Disease miRNAs by Integrating Phenotype Associations of Multiple Diseases with Matched miRNA and mRNA Expression Profiles. *Molecular BioSystems*, **10**, 2800-2809. https://doi.org/10.1039/C4MB00353E