面向多组学数据的图对比学习

Nishant Rajadhyaksha, K.J
Somaiya College of Engineering 孟买 n.rajadhyaksha@somaiya.edu

Aarushi Chitkara St
Xaviers College 孟买
aarushi5812@gmail.com

arXiv:2301.02242v1 [q-bio.GN] 3 Jan 2023

摘要

与组学数据处理相关的技术不断进步，因此需要开发新的计算方法，以充分利用信息并帮助加深对人类疾病的认识。本文研究了引入图对比学习以利用图结构和信息，从而为多组学数据集的下游分类任务生成更优表示的效果。我们提出了一种名为多组学图对比学习器（MOGCL）的学习框架，其在多种用于整合多组学数据以完成监督学习任务的方法中表现更佳。我们表明，采用对比方法对图模型进行预训练，并以监督方式对其进行微调，是实现多组学数据分类的有效策略。

关键词 图对比学习 · 图神经网络 · 多组学

1 引言

组学是生物科学中的一个研究领域，其名称以“-组学”结尾，旨在对生物体或多个生物体中的生物分子集合进行整体表征和定量分析，从而揭示其结构、功能和动态变化。近年来，高通量生物医学研究方法的应用大幅增加，例如全基因组测序（WGS）、RNA测序（RNA-seq）、染色体构象捕获（Hi-C），以及与质谱联用的液相色谱 [1, 2]。整合来自不同分子来源的数据（例如微小RNA（miRNA）、信使RNA和DNA甲基化数据），对于深入了解不同疾病的分类和过程尤其有帮助。整合多组学数据需要高效的计算方法，以准确建模和解释这些关系。尽管每种组学技术只能呈现真实生物复杂性的一部分，但整合多种组学技术的数据有助于构建更全面的图景，从而提升分类任务的结果 [3–9]。

已有多种不同的方法被提出，用于整合多组学数据以完成分类任务。通常，不同组学的数据会被拼接在一起，或者采用集成学习，将每种组学类型的预测结果进行集成 [10, 11]。近年来，图神经网络（GNNs）作为一种高效的深度学习方法兴起，可用于对复杂系统建模，这促使研究人员探究其与多组学数据结合时的效果 [12, 13]。图对比学习是一种学习范式，利用数据对在图数据的正样本对与负样本对之间进行区分。对比学习可作为一种有效的预训练策略，用于在监督学习任务中训练图模型。本文介绍了一种名为 MOGCL 的框架，该框架从多组学数据构建图，并利用对比学习作为预训练策略，以助力下游分类任务。我们在基准数据集上将结果与多种不同的机器学习方法进行了比较。MOGCL 在 ROC曲线下面积（AUC）、准确率和 F1分数等多项指标上表现更佳。

2 文献综述

2.1 多组学数据的机器学习

已有大量研究探讨如何整合多组学数据以完成各种任务。许多方法侧重于对多组学数据进行半监督整合，而不利用数据集中真实标签所包含的信息 [14–17]。大多数自监督方法着重于为数据集中不同的组学组分配簇。随着生物技术的快速发展以及数据集得到详细整理，越来越多带有表型或性状标签的样本可供研究使用。这推动了监督学习模型的发展，这些模型在多组学数据集上表现更佳 [18–22]。核方法是强大的机器学习模型，在更高维的空间中运行，并可在给定的数据点之间找到线性插值。核方法常作为经典机器学习模型，用于分析和分类多组学数据。支持向量机（SVM） [23] 和偏最小二乘 [24] 是用于多组学数据的经典机器学习方法的例子。目前，深度学习方法已广泛用于多组学整合，以完成各种任务。Islam 等人 [25] 提出了一种方法，利用卷积神经网络学习多种组学类型的重要特征，最后将这些特征拼接起来，以预测乳腺癌亚型。Xu 等人 [26] 提出了一种使用堆叠稀疏自编码器（SAE）的方法，通过自编码器为每种组学数据生成表示，再将其输入深度灵活神经森林，以预测癌症亚型。

2.2 基于图的学习方法

图是复杂的数据结构，可用于对现实世界中的许多复杂现象建模。图神经网络（GNN）致力于将深度学习应用于图结构数据。图神经网络有多种应用，例如组合优化、神经机器翻译、蛋白质-蛋白质相互作用和药物发现 [27–34]。近年来，基于图的方法已用于多组学整合。Wang 等人 [35] 提出了一种将多组学数据转换为图的方法，以及一个名为 MOGONET 的模型。该模型由卷积网络（图卷积网络）层组成 [36] ，可生成经过改进的表示，用于在多个多组学数据集上执行下游分类任务，同时识别重要的生物标志物。Xiao 等人 [37] 提出了一个名为 MOGCN 的模型，用于基于多组学数据开展不同的癌症亚型分类任务。图对比学习是一种自监督训练方法，通过采用不同的图增强，利用数据集的结构信息和特征信息，生成更优的表示。常见的训练策略包括先预训练再微调 [38]，以及联合学习 [39]。Zhu 等人 [40] 提出了一个名为深度图对比表示学习（GRACE）的框架，该框架专门通过扰动生成两个图视图，并尝试通过最大化这两个视图中节点表示的一致性来学习节点表示。

3 方法

3.1 数据集

我们将模型应用于两个基准数据集，以验证 MOGCL 的有效性：ROSMAP [41]，该数据集描述阿尔茨海默病（AD）患者与正常对照组（NC）的相关信息；以及 BRCA，该数据集用于浸润性乳腺癌（BRCA）PAM50亚型分类。所使用的组学数据包括 DNA甲基化数据（甲基化）、miRNA表达数据（miRNA）和 mRNA表达数据（信使RNA）。数据集的详细信息进一步见表 1。

数据集名称	标签	特征数量（按） mRNA, meth, miRNA	用于训练的特征数量 mRNA, meth, miRNA
ROSMAP	正常对照组：169 ，阿尔茨海默病：182	55,889, 23,788, 309	200, 200, 200
BRCA	正常样：115 ， 基底样：131 ， HER2富集型：46 ， 管腔A型：436 ， 管腔B型：147	20,531, 20,106, 503	1000, 1000, 503

表 1：数据集摘要

3.2 从多组学数据构建图

本节介绍如何将多组学数据转换为图结构，以便利用强大的图神经网络模型进行后续处理。我们的任务可定义为构建图 $G = (V, E)$，其中 V、E 分别表示图的顶点和边。我们使用对每种组学数据进行预处理后得到的特征矩阵。每种组学类型的特征矩阵表示为 $X \in \mathbb{R}^{n \times d}$；对于 ROSMAP数据集，每种组学类型的 d 均为 200，n 为 351。类似地，对于 BRCA数据集，n 为 875；d 的取值范围为 1000（信使RNA和甲基化数据）至 503（miRNA数据）。图 G 的节点 V 表示收集组学数据的用户。我们构建邻接矩阵 $A \in \mathbb{R}^{n \times n}$ 来表示 G，邻接矩阵中的每个元素表示一个节点。图中从节点 i 到节点 j 的加权边，表示为 A 中第 i^{th} 行、第 j^{th} 列的元素。我们为每种组学数据分别构建这样的邻接矩阵。使用余弦相似度 [42] 作为距离度量，为特定组学数据集中的数据点构建成对距离矩阵。节点 i 与节点 j 之间的距离记为 t_{ij}。引入参数 k，表示每个节点要构建的边数。随后，通过从所有数据点之间的成对距离排序数组中选取 $n \times k^{th}$ 值，确定邻接参数。接着，根据数据点之间的距离小于邻接参数这一条件选取边 E。这可确保每个节点的边数为 k。若节点 i 到节点 j 的边属于所选边集合，则为其赋予权重 $1 - t_{ij}$。按照上述方法，为相应数据集中包含的每种组学类型构建邻接矩阵。

3.3 图对比学习

本节介绍我们采用图对比学习的训练方法。我们使用 GRACE [40]作为自监督模型。我们的对比学习方法分为两个阶段，即 i）数据增强和 ii）对比学习。我们使用随机删除边和特征掩蔽等增强函数来创建图的增强视图。为增强边，我们从原始图中随机删除一部分边。我们采样一个随机掩蔽矩阵 $\tilde{R} \in \{0, 1\}^{N \times N}$。其中，R˜的元素服从伯努利分布 $R \sim Bern(1 - pr)$，其中 pr 表示每条边被删除的概率。˜ ˜在本研究中，我们将 pr 设为 0.4。由此得到的邻接矩阵可表示为 $A = A \circ R$。为增强特征，我们随机掩蔽特征向量的各个维度，并将其替换为零。我们采样随机特征˜向量，根据伯努利分布构造矩阵 $M \in \{0, 1\}$，其大小与特征矩阵˜ ˜X相似。增强后的特征矩阵可表示为 $X = X \circ M$。我们使用 GCN [36] 模型作为编码器˜ ˜模型，以表示给定图的增强视图，并将其记为 f。设 $U = f(X_1, A_1)$ 和˜ ˜$V = f(X_2, A_2)$ 为使用我们的共享编码器模型处理两个图后生成的表示。我们的目标是最大化潜在空间中相似节点之间的一致性，并最小化其与其余对比节点之间的一致性。为此，我们采用归一化温度缩放交叉熵损失（NT-Xent） [43]。NT-Xent损失由公式 1 给出。

$$\ell(\boldsymbol{u}_i, \boldsymbol{v}_i) = \log \frac{e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau}}{e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_k)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{u}_i, \boldsymbol{u}_k)/\tau}}, \tag{1}$$

其中，u_i 和 v_i 分别表示特征矩阵U和V中的 i^{th} 特征向量。τ 表示温度参数。θ 是公式2中给出的相似度函数。

$$\theta(u, v) = c(n(u), n(v)) \tag{2}$$

其中，c(.,.)是余弦相似度函数，n(.)表示任意非线性函数，例如ReLU [44] 或LeakyReLU [45] 等。最后，我们基于NT-Xent损失优化共享编码器模型的权重。

使用给定数据集的标签，以监督方式进一步训练图卷积网络编码器。编码器模型通过先预训练、再微调的方式针对下游分类任务进行训练。在预训练阶段，我们首先以无监督方式对每种组学类型分别完整训练一个编码器模型。随后，我们使用给定数据集中的标签˜信息对这些模型进行微调。令 f 为预训练的图卷积网络编码器。我们将线性层与编码器模型生成的拼接特征结合使用，以生成预测标签 $Y = f(X, A)$。我们使用˜交叉熵损失计算预测标签 Y 与真实标签 Y 之间的损失，并最终依据该损失优化编码器模型。

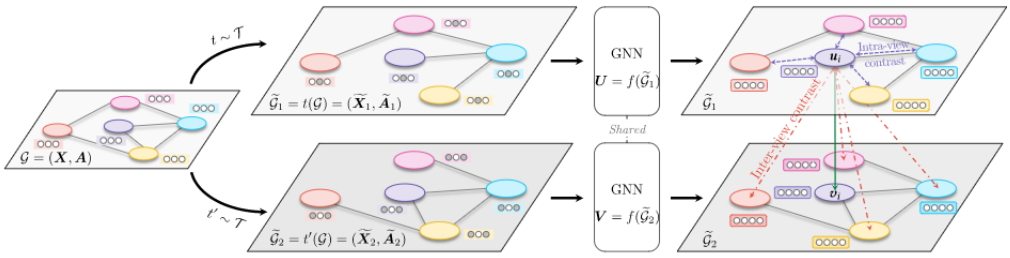

图 1：GNN编码器的对比学习

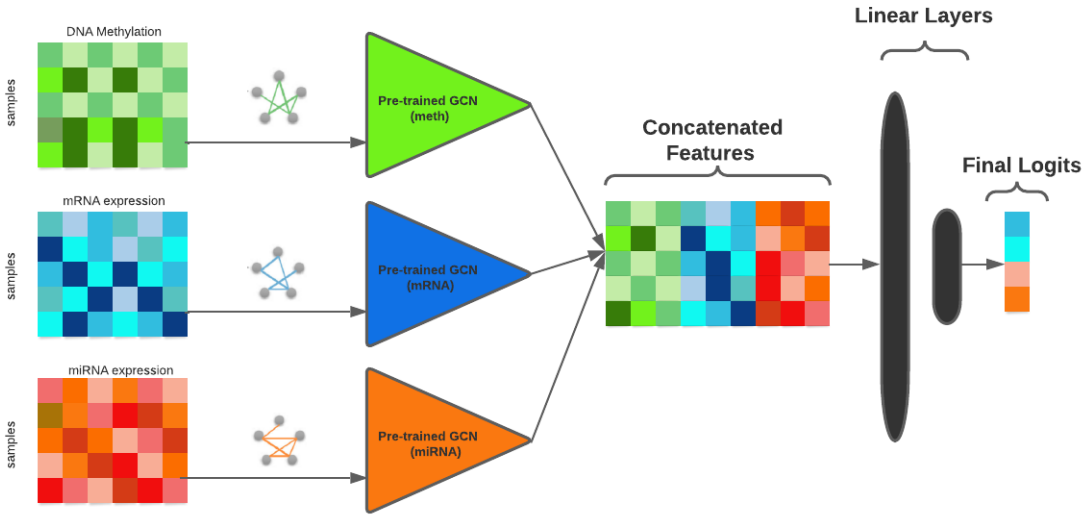

图 2：GNN编码器的下游监督训练

3.4 实验

本节介绍我们为评估 MOGCL 框架而开展的实验。我们首先为数据集中的每种组学类型构建图，并分别为每种类型训练一个独立的编码器模型。最后，我们拼接各编码器模型生成的特征，并采用先预训练、再微调的方法训练编码器模型。我们将分类结果与 [35] 中所述的结果进行比较，以评估在给定分类任务中引入对比学习方法的效果。通过对每种编码器模型组合进行 $r = 5$ 次随机权重初始化运行，计算其性能。我们使用准确率、F1分数和曲线下面积等指标评估模型在 ROSMAP 数据集上的表现，并使用准确率、加权 F1 分数和宏平均F1分数评估 BRCA 数据集。我们使用 PyTorch-geometric [46], PyGCL [47] 和 pytorch-lightning [48] 开展实验。所有实验均采用 Adam [49] 优化器，学习率设为 0.0001。我们使用两层图卷积网络作为图对比学习中的编码器模型。此外，我们还将两层线性层与编码器模型结合，利用给定的真实标签进行微调。所有实验中，我们都将特征向量压缩至 100 维潜在空间。我们通过可视化各组学类型的特征向量在经过编码器模型处理前后的变化，尝试展示预训练策略的效果。我们使用 t-SNE [50] 将特征向量压缩到二维空间，以生成可视化结果。

4 结果与讨论

分类结果　　ROSMAP和BRCA数据集的分类任务分别展示于表2和表3中　　.

表2：ROSMAP 上分类任务的结果

方法	准确率	F1	AUC
KNN	0.657 ± 0.036	0.671 ± 0.044	0.709 ± 0.045
SVM	0.770 ± 0.024	0.778 ± 0.016	0.770 ± 0.026
套索回归	0.694 ± 0.037	0.730 ± 0.033	0.770 ± 0.035
RF	0.726 ± 0.029	0.734 ± 0.021	0.811 ± 0.019
XGBoost	0.760 ± 0.046	0.772 ± 0.045	0.837 ± 0.030
NN	0.755 ± 0.021	0.764 ± 0.021	0.827 ± 0.025
GRridge	0.760 ± 0.034	0.769 ± 0.029	0.841 ± 0.023
分块偏最小二乘判别分析	0.742 ± 0.024	0.755 ± 0.023	0.830 ± 0.025
分块稀疏偏最小二乘判别分析	0.753 ± 0.033	0.764 ± 0.035	0.838 ± 0.021
NN_NN	0.766 ± 0.023	0.777 ± 0.019	0.819 ± 0.017
NN_VCDN	0.775 ± 0.026	0.790 ± 0.018	0.843 ± 0.021
MOGONET _NN	0.804 ± 0.016	0.808 ± 0.010	0.858 ± 0.024
MOGCL（我们的方法）	**0.818 ± 0.014**	**0.818 ± 0.014**	**0.866 ± 0.021**

表3：BRCA 上分类任务的结果。

方法	准确率	加权 F1	宏平均 F1
KNN	0.742 ± 0.024	0.730 ± 0.023	0.682 ± 0.025
SVM	0.729 ± 0.018	0.702 ± 0.015	0.640 ± 0.017
套索回归	0.732 ± 0.012	0.698 ± 0.015	0.642 ± 0.026
RF	0.754 ± 0.009	0.733 ± 0.010	0.649 ± 0.013
XGBoost	0.781 ± 0.008	0.764 ± 0.010	0.701 ± 0.017
NN	0.754 ± 0.028	0.740 ± 0.034	0.668 ± 0.047
GRridge	0.745 ± 0.016	0.726 ± 0.019	0.656 ± 0.025
分块偏最小二乘判别分析	0.642 ± 0.009	0.534 ± 0.014	0.369 ± 0.017
分块稀疏偏最小二乘判别分析	0.639 ± 0.008	0.522 ± 0.016	0.351 ± 0.022
NN_NN	0.796 ± 0.012	0.784 ± 0.014	0.723 ± 0.018
NN_VCDN	0.792 ± 0.010	0.781 ± 0.006	0.721 ± 0.018
MOGONET _NN	0.805 ± 0.017	0.782 ± 0.030	0.737 ± 0.038
MOGCL（我们的方法）	**0.853 ± 0.005**	**0.851 ± 0.010**	**0.823 ± 0.006**

MOGCL的性能与以下分类算法进行了比较：1）K近邻分类器（KNN）。评估期间，从训练数据中选取K近邻来进行标签预测。2）支持向量机分类器（SVM）。3）套索回归，即L1正则化线性回归。在套索回归中，为预测每个类别的概率分别训练一个模型，并选择预测概率最大的类别作为模型整体类别标签的最终预测。4）随机森林分类器（RF）。5）极端梯度提升（XGBoost）是一种分布式、可扩展的梯度提升决策树（GBDT）机器学习框架。6）全连接神经网络（NN）分类器。全连接神经网络的损失通过交叉熵损失计算。7）自适应组正则化岭回归（GRridge）。8）DIABLO [6]中提到的分块偏最小二乘判别分析。分块偏最小二乘判别分析通过潜变量判别分析（LDA）将多组学数据投影到潜在空间中。为了对离散结果进行分类，分块偏最小二乘判别分析整合在同一组样本上测量的多种组学数据类型。9）分块稀疏偏最小二乘判别分析。10）MOGONET_NN。MOGONET_NN在架构上与MOGCL相似，但不采用预训练策略。通过遵循我们的预训练方法，我们取得了显著的结果；在用于衡量结果的所有指标上，该方法的表现均优于其他模型。对于ROSMAP数据集，MOGCL的平均准确率为0.818，而MOGONET_NN为0.804。遵循这一趋势，MOGCL的F1分数和曲线下面积分别为0.818和0.866，而MOGONET_NN分别为0.808和0.856。对于BRCA数据集，MOGCL的准确率为0.853，而MOGONET_NN为0.805。MOGCL的加权F1分数和宏平均F1分数分别为0.851和0.823，而MOGONET_NN分别为0.782和0.737。这表明，在针对下游任务进行微调之外，采用基于图的半监督学习策略，是在多组学数据集上训练模型的有效训练策略。

我们通过分析3和4，展示了采用半监督方法进行训练的效果。我们利用 t-SNE 将数据点投影到二维平面上，从而可视化特征矩阵X。同样地，我们将 GCN编码器生成的特征向量映射到二维空间，并比较结果。在没有标签的情况下，MOGCL 尝试对嵌入进行聚类，以便在预训练阶段创建结构更清晰的表示。更好的表示有助于微调阶段，进而帮助获得更好的分类分数。

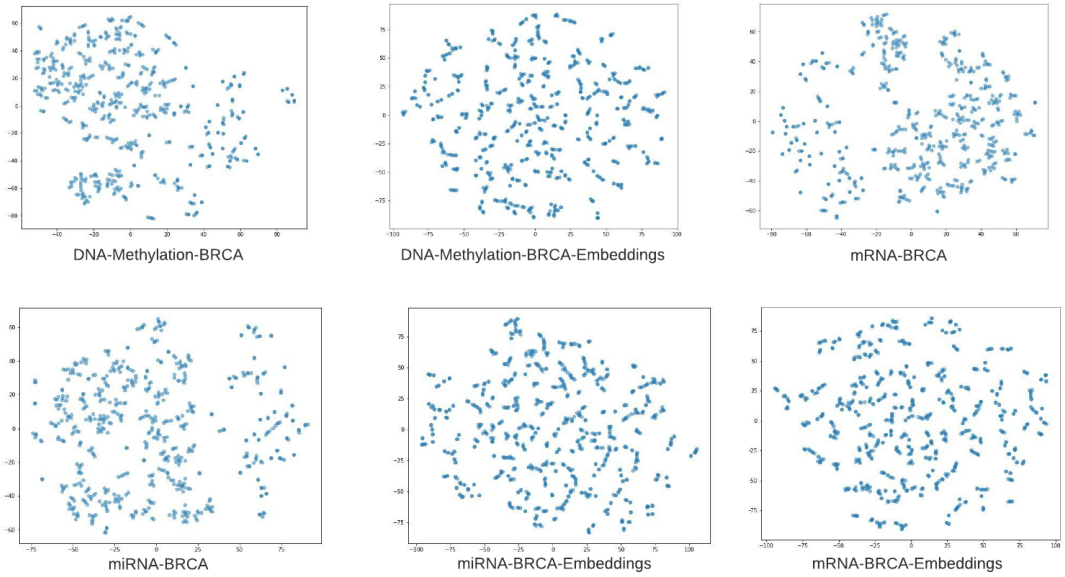

图 3：BRCA 嵌入

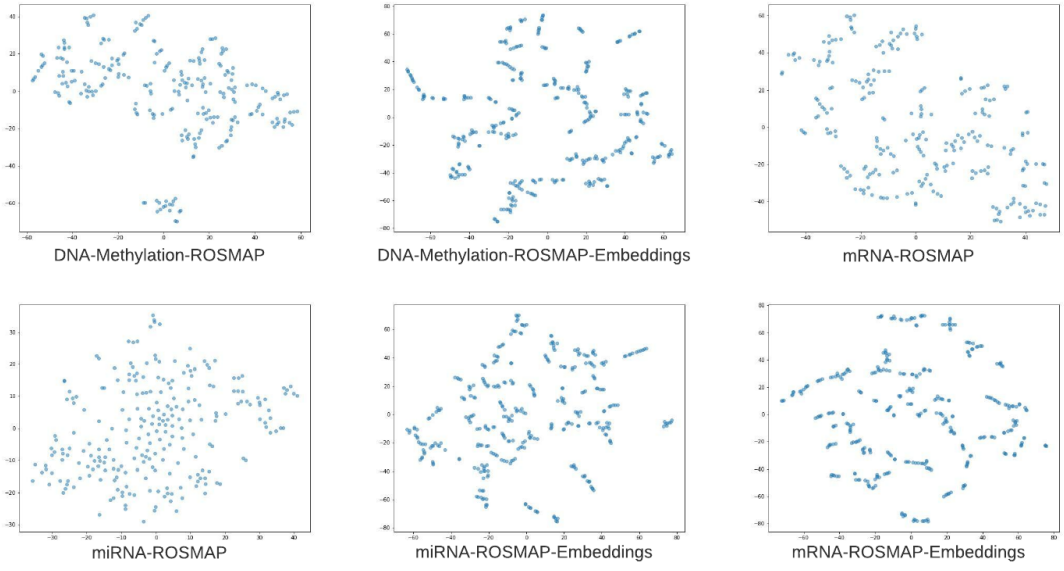

图 4：ROSMAP 嵌入

图5展示了不同组学类型组合经MOGCL处理后的性能。我们分别为研究中的所有组学类型预训练了三个编码器模型。为了计算性能，我们选取这些编码器模型的一种组合，并使用真实标签以监督方式对其进行训练。对于ROSMAP和BRCA数据集，将所有组学类型的信息拼接在一起输入时，MOGCL的表现最佳。对于BRCA数据集，信使RNA与DNA甲基化数据的组合效果次佳；而对于ROSMAP数据集，信使RNA与miRNA的组合效果次佳。对于ROSMAP和BRCA数据集，仅使用一种组学类型时效果最差。在BRCA和ROSMAP数据集中，仅使用DNA甲基化数据的效果最差，其次是miRNA和信使RNA数据。

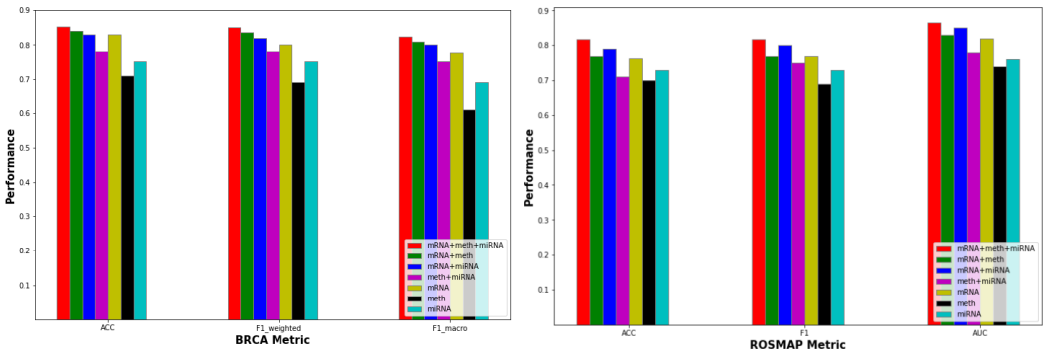

图5：组学类型组合的性能

5 结论

本文介绍了一种名为MOGCL的新型框架，该框架将图对比学习方法用于多组学数据分类。我们首先全面综述了机器学习领域中基于图的学习方法和多组学数据相关的研究。随后讨论了如何从多组学数据构建图。接着，我们介绍了MOGCL框架，该框架采用GRACE作为预训练方法，然后在监督设置下利用真实标签进行微调。我们讨论了该框架在BRCA和ROSMAP数据集上的结果，并表明其性能优于本研究采用的其他基线方法。我们通过分析不同指标下的性能，探讨了不同组学类型组合的使用。我们还通过可视化图编码器生成的嵌入，讨论了采用半监督预训练策略的影响。最后，我们得出结论：采用预训练方法是训练图模型以解决涉及多组学数据集的分类问题的一种高效方式。未来的工作可以尝试不同的对比学习方法，以确定哪种方法效率最高。还可以针对不同的图神经网络开展实验，例如图注意力网络（GAT）或图同构网络（GIN）等，以确定哪种网络最适合作为多组学数据集监督学习的编码器。

参考文献

[1] Angela P Fuentes-Pardo 和 Daniel E Ruzzante。用于保护生物学的全基因组测序方法：优势、局限性与实践建议。《分子生态学》，26(20):5369–5406，2017年10月。[2] Rui Chen 和 Michael Snyder。个性化组学在精准医学中的前景。《*Wiley*跨学科评论：系统生物学与医学》，5(1):73–82，2013年1月。[3] Zhi Huang、Xiaohui Zhan、Shunian Xiang、Travis S. Johnson、Bryan Helm、Christina Y. Yu、Jie Zhang、Paul Salama、Maher Rizkalla、Zhi Han 和 Kun Huang。Salmon：利用乳腺癌多组学神经网络进行生存分析学习。*Frontiers in Genetics*，10，2019年。[4] Sijia Huang、Kumardeep Chaudhary 和 Lana X. Garmire。越多越好：多组学数据整合方法的最新进展。*Frontiers in Genetics*，8，2017年。[5] Sebastian Canzler、Jana Schor、Wibke Busch、Kristin Schubert、Ulrike E. Rolle-Kampczyk、Hervé Seitz、Hennicke Kamp、Martin von Bergen、Roland Buesen 和 Jörg Hackermüller。毒理学中多组学数据整合的前景与挑战。《毒理学档案》，2020年2月。

[6] Amrit Singh、Casey P Shannon、Benoît Gautier、Florian Rohart、Michaël Vacher、Scott J Tebbutt 和 Kim-Anh Lê Cao。DIABLO：一种从多组学检测中识别关键分子驱动因素的整合方法。《生物信息学》，35(17):3055–3062，2019年9月。[7] Mark A. van de Wiel、Tonje G. Lien、Wina Verlaat、Wessel N. van Wieringen 和 Saskia M. Wilting。利用协变量数据提升预测效果：自适应组正则化岭回归。《医学统计学》，35(3):368–381，2016年。[8] Yingkun Zhu、Dengpan Bu 和 Lu Ma。多组学整合：系统性乳业研究向前迈进一步。《代谢物》，12(3)，2022年。[9] Indhupriya Subramanian、Srikant Verma、Shiva Kumar、Abhay Jere 和 Krishanpal Anamika。多组学数据整合、解读及其应用。《生物信息学与生物学见解》，14:1177932219899051，2020年1月。[10] Olivier B. Poirion、Zheng Jing、Kumardeep Chaudhary、Sijia Huang 和 Lana X. Garmire。Deepprog：一种利用多组学数据进行预后预测的深度学习与机器学习模型集成。《基因组医学》，13(1):112，2021年7月。[11] Shane O'Donnell、R. Ross 和 Catherine Stanton。多组学技术的进展：利用系统层面方法确定乳酸菌的功能。《微生物学前沿》，10:3084，2020年1月。[12] Chaoyi Yin、Yangkun Cao、Peishuo Sun、Hengyuan Zhang、Zhi Li、Ying Xu 和 Huiyan Sun。基于稳健图神经网络和多组学数据整合的癌症分子分型。《遗传学前沿》，13:884028，2022年5月。[13] Xiaohan Xing、Fan Yang、Hang Li、Jun Zhang、Yu Zhao、Mingxuan Gao、Junzhou Huang 和 Jianhua Yao。一种用于基于基因表达数据进行疾病诊断的可解释多层增强图注意力网络。载于2021年IEEE生物信息学与生物医学国际会议（BIBM），第556–561页。IEEE，2021年。[14] Ziwang Huang、Hua Chai、Ruoqi Wang、Haitao Wang、Yuedong Yang 和 Hejun Wu。通过自监督学习和Transformer整合图块特征，实现全切片图像生存分析。载于医学图像计算与计算机辅助干预国际会议，第561–570页。Springer，2021年。[15] Xiaoshuang Liu、Xian Xu、Xiao Xu、Xiang Li 和 Guotong Xie。利用异质基因调控网络学习多组学数据的表征。载于2021年IEEE生物信息学与生物医学国际会议（BIBM），第702–705页，2021年。[16] Ronglai Shen、Adam B. Olshen 和 Marc Ladanyi。利用联合潜变量模型对多种基因组数据类型进行整合聚类，并应用于乳腺癌和肺癌亚型分析。《生物信息学》，25(22):2906–2912，2009年9月。[17] Giulia Tini、Luca Marchetti、Corrado Priami 和 Marie-Pier Scott-Boyer。多组学整合：无监督聚类方法比较。《简明生物信息学》，20(4):1269–1279，2019年7月。[18] Lorin M. Towle-Miller、Jeffrey C. Miecznikowski、Fan Zhang 和 David L. Tritchler。Sumo-fil：网络分析前的监督式多组学过滤。《公共科学图书馆·综合》，16(8):1–20，2021年8月。[19] Otília Menyhárt 和 Balázs Gy˝orffy。多组学方法在癌症研究中的应用：肿瘤分型、预后和诊断。《计算与结构生物技术杂志》，19:949–960，2021年。[20] Kaiwen Tan、Weixian Huang、Jinlong Hu 和 Shoubin Dong。一种用于泛癌症临床结局终点预测的多组学监督式自编码器。《BMC医学信息学与决策制定》，20(3):129，2020年7月。[21] Zak Costello 和 Hector Garcia Martin。一种利用时间序列多组学数据预测代谢通路动态的机器学习方法。《npj系统生物学与应用》，4(1):19，2018年5月。[22] Milan Picard、Marie-Pier Scott-Boyer、Antoine Bodein、Olivier Périn 和 Arnaud Droit。用于机器学习分析的多组学数据整合策略。《计算与结构生物技术杂志》，19:3735–3746，2021年。[23] SungHwan Kim、Jae-Hwan Jhong、JungJun Lee 和 Ja-Yong Koo。用于整合多组学数据的元分析支持向量机。《生物数据挖掘》，10(1):2，2017年1月。[24]Candida S. Punla，https://orcid.org/0000-0002-1094-0018，cspunla@bpsu.edu.ph，Rosemarie C. Farro，https://orcid.org/0000-0002-3571-2716，rcfarro@bpsu.edu.ph，以及菲律宾巴丹省迪纳卢皮汉的巴丹半岛州立大学。我们到了吗？：对BPSU-DC BEED毕业生能力的分析。《国际多学科研究杂志》，4(3):50–59，2022年9月。[25] Md Mohaiminul Islam、Shujun Huang、Rasif Ajwad、Chen Chi、Yang Wang 和 Pingzhao Hu。一种用于乳腺癌分子亚型分类的整合式深度学习框架。《计算与结构生物技术杂志》，18:2185–2199，2020年8月。

[26] Jing Xu、Peng Wu、Yuehui Chen、Qingfang Meng、Hussain Dawood 和 Hassan Dawood。通过整合多组学数据进行癌症亚型分类的分层集成深度柔性神经森林框架。*BMC*生物信息学，20(1):527，2019年10月。[27] Martin J. A. Schuetz、J. Kyle Brubaker 和 Helmut G. Katzgraber。利用受物理启发的图神经网络进行组合优化。《自然·机器智能》，4(4):367–377，2022年4月。[28] Yun Peng、Byron Choi 和 Jianliang Xu。组合优化的图学习：前沿综述。《数据科学与工程》，6(2):119–141，2021年。[29] Binh Nguyen、Long Nguyen 和 Dien Dinh。用于神经机器翻译的多层社区感知图神经网络。载于第*29*届国际计算语言学会议论文集，第5021–5028页，韩国庆州，2022年10月。国际计算语言学委员会。[30] Xin Mei、Xiaoyan Cai、Libin Yang 和 Nanxin Wang。基于图变换器网络的文本表示。《神经计算》，463:91–100，2021年。[31] Elahe Nasiri、Kamal Berahmand、Mehrdad Rostami 和 Mohammad Dabiri。利用属性图嵌入预测蛋白质相互作用网络链接的新算法。《生物学与医学中的计算机》，137:104772，2021年。[32] Xinnan Dai、Fan Xu、Shike Wang、Piyushkumar A Mundra 和 Jie Zheng。Pike-r2p：基于蛋白质相互作用网络的知识嵌入与图神经网络，用于单细胞RNA到蛋白质预测。*BMC*生物信息学，22(6):1–17，2021年。[33] Zhengyang Wang、Meng Liu、Youzhi Luo、Zhao Xu、Yaochen Xie、Limei Wang、Lei Cai、Qi Qi、Zhuoning Yuan、Tianbao Yang 等。用于分子性质预测和药物发现的先进图与序列神经网络。《生物信息学》，38(9):2579–2586，2022年。[34] Pietro Bongini、Monica Bianchini 和 Franco Scarselli。用于药物发现的分子生成图神经网络。《神经计算》，450:242–252，2021年。[35] Tongxin Wang、Wei Shao、Zhi Huang、Haixu Tang、Jie Zhang、Zhengming Ding 和 Kun Huang。MOGONET利用图卷积网络整合多组学数据，实现患者分类和生物标志物识别。《自然·通讯》，12(1):3445，2021年6月。[36] Thomas N. Kipf 和 Max Welling。利用图卷积网络进行半监督分类，2016年。[37] Xiao Li、Jie Ma、Ling Leng、Mingfei Han、Mansheng Li、Fuchu He 和 Yunping Zhu。MOGCN：一种基于图卷积网络的癌症亚型分析多组学整合方法。*Frontiers in Genetics*，13，2022年。[38] Yuning You、Tianlong Chen、Yongduo Sui、Ting Chen、Zhangyang Wang 和 Yang Shen。通过数据增强进行图对比学习，2020年。[39] Han Zhao、Xu Yang、Zhenru Wang、Erkun Yang 和 Cheng Deng。通过联合表示聚类进行图去偏对比学习。载于 Zhi-Hua Zhou 编，第三十届国际人工智能联合会议论文集（*IJCAI-21*），第3434–3440页。国际人工智能联合会议组织，2021年8月。主会场。[40] Yanqiao Zhu、Yichen Xu、Feng Yu、Qiang Liu、Shu Wu 和 Liang Wang。深度图对比表示学习。载于*ICML*图表示学习及其拓展研讨会，2020年。[41] Philip L. De Jager、Yiyi Ma、Cristin McCabe、Jishu Xu、Badri N. Vardarajan、Daniel Felsky、Hans-Ulrich Klein、Charles C. White、Mette A. Peters、Ben Lodgson、Parham Nejad、Anna Tang、Lara M. Mangravite、Lei Yu、Chris Gaiteri、Sara Mostafavi、Julie A. Schneider 和 David A. Bennett。用于衰老和阿尔茨海默病研究的人类额叶皮层多组学图谱。《科学数据》，5(1):180142，2018年8月。[42] Alfirna Rizqi Lahitani、Adhistya Erna Permanasari 和 Noor Akhmad Setiawan。使用余弦相似度确定相似度度量：在线作文评估案例研究。载于*2016*年第四届国际网络与*IT*服务管理会议，第1–6页，2016年。[43] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表示对比学习的简单框架，2020年。[44] Kunihiko Fukushima。Cognitron：一种自组织多层神经网络。《生物控制论》，20(3–4):121–136，1975年9月。[45] Bing Xu、Naiyan Wang、Tianqi Chen 和 Mu Li。卷积网络中整流激活函数的实证评估，2015年。[46] Matthias Fey 和 Jan E. Lenssen。利用 PyTorch Geometric 实现快速图表示学习。载于*ICLR*图与流形表示学习研讨会，2019年。

[47] Yanqiao Zhu、Yichen Xu、Qiang Liu 和 Shu Wu。图对比学习的实证研究。*arXiv.org*，2021年9月。[48] William Falcon 等。PyTorch Lightning。*GitHub*。注：*https://github.com/PyTorchLightning/pytorch-lightning*，3(6)，2019年。[49] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun 编，第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。[50] Laurens van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。机器学习研究期刊，9:2579–2605，2008年。