Advances in Applied Mathematics 应用数学进展, 2025, 14(12), 48-58
Published Online December 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.1412484

Hans汉斯

IMMO：面向不完整多组学数据的整合分析框架

李佳惠

青岛大学数学与统计学院，山东 青岛

收稿日期：2025年11月4日；录用日期：2025年11月29日；发布日期：2025年12月5日

摘　要

随着多组学数据整合方法的发展，微生物多组学数据在揭示复杂疾病机制方面发挥着越来越重要的作用。然而，在实际研究中，多组学数据普遍存在样本不完全匹配、模态广泛缺失以及高维稀疏性等问题，严重制约了组学信息之间的有效整合。为应对这些挑战，本文提出一种基于动态掩码机制的多组学整合模型(IMMO)，采用联合自编码器架构，引入动态掩码策略，在训练过程中自适应地处理缺失数据，实现对不完全多组学数据的表征学习与数据重构。在炎症性肠病(IBD)和糖尿病数据集上的实验结果表明，IMMO在数据重建和疾病分类任务中表现出良好的性能，且所学习到的潜在特征能够有效捕捉与疾病相关的关键微生物模式。这为不完全多组学数据的整合分析提供了一种稳定、高效且可解释的方案。

关键词

动态掩码，不完整数据整合，深度学习，自编码器

IMMO: An Integrated Analysis Framework for Incomplete Multi-Omic Data

Jiahui Li

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: November 4, 2025; accepted: November 29, 2025; published: December 5, 2025

Abstract

Microbiome multi-omics data are increasingly valuable for deciphering complex diseases. However, their integration is often hindered by incomplete sample matching, widespread missing modalities, and high-dimensional sparsity. To address these challenges, we propose IMMO (Integration Model

文章引用: 李佳惠. IMMO: 面向不完整多组学数据的整合分析框架[J]. 应用数学进展, 2025, 14(12): 48-58.
DOI: 10.12677/aam.2025.1412484

for Incomplete Multi-Omics), a joint autoencoder-based framework that incorporates a dynamic masking mechanism to adaptively handle missing data during training. Evaluated on inflammatory bowel disease (IBD) and diabetes cohorts, IMMO demonstrates strong performance in both data reconstruction and disease classification, with latent representations capturing disease-relevant microbial patterns. Our approach offers a robust, efficient, and interpretable solution for integrative analysis of incomplete multi-omics data.

Keywords

Dynamic Masking, Incomplete Data Integration, Deep Learning, Autoencoder

1. 引言

肠道微生物群在调节宿主营养代谢、免疫稳态及屏障功能中发挥关键作用，其结构与功能的失衡已被广泛证实与多种慢性疾病密切相关[1] [2]。因此，肠道微生物组被视为揭示疾病发生机制、实现早期诊断与动态监测的重要生物资源[3] [4]。

随着高通量测序技术的发展，宏基因组学、元转录组学和代谢组学等多组学手段能够从物种组成、基因表达、代谢产物等方面系统性地解析微生物生态系统[5] [6]，这些数据为理解炎症性肠病(IBD) [7] [8]、糖尿病(T2D) [9]和结直肠癌[10] [11]等复杂疾病的发病机制提供了关键线索。然而，实际应用中，多组学数据往往面临样本不完全匹配、不同模态间数据缺失、高维稀疏以及样本量有限等挑战[12]，这些问题严重制约了跨组学信息的有效融合与生物学意义的深入解析。

传统的数据填补或样本剔除策略虽然能够一定程度上缓解上述问题，但易导致信息失真或统计效能下降。现有主流多组学整合方法，如 MOFA+ [13]、联合潜变量模型[14]和基于图卷积网络的 MoGCN [15]，通常假设输入数据完整，难以有效应对实际研究中常见的不完整数据结构。深度学习模型如 ProgCAE [16]和 TMO-Net [17]虽然在微生物分类任务中表现出良好性能，但其对缺失模态的鲁棒性仍不足。近年来，掩码学习(masked learning)策略在自然语言处理[18]和计算机视觉[19] [20]等领域展现出强大的表示学习与缺失数据建模能力，为多组学整合提供了新思路，但其静态掩码机制在训练过程中缺乏灵活性，限制了模型对复杂生物数据的学习效率。

为了解决上述问题，本文提出一种基于动态掩码的多组学整合模型(IMMO-integration)。该模型基于自编码器(AE)，引入随训练进程自适应调整的动态掩码机制，结合共享解码结构与加权重构损失，旨在提高对不完全微生物多组学数据的鲁棒表征学习与高效重构能力。通过在炎症性肠病多组学数据库(IBDMDB) [8]和斯坦福大学个体化医疗项目(IPOP)糖尿病[9]相关多组学数据集上的实验验证，本研究展示了 IMMO-integration 在重构性能与下游分类任务中的优越表现，为其在精准医学领域的应用奠定了基础。

2. 数据来源与预处理

研究所用数据源自炎症性肠病多组学数据库(IBDMDB) (https://www.ibdmdb.org)的不完整多组学数据集[20]，该数据集可预测炎症性肠病(IBD)状态，包含三组学观察：宏基因组学(mg)、代谢组学(mb)和元转录组学(mt)。由于原始数据特征维度过大且缺失过多，为确保数据质量便于后续建模，在实验中首先对其进行清洗和标准化处理，经预处理后，数据集的样本重叠情况及标签信息如表 1 所示。

Table 1. Sample overlap and label distribution across omics modalities
表 1. 不同组学样本重叠情况及标签信息

Datasets	Samples	NonIBD	IBD
IBD_common	324	88	236
bol_tra_common	795	195	600
tra_bio_common	324	88	236
bol_bio_common	470	123	347
tra_bol_all	1588	414	1174
tra_bio_all	1020	243	777
bol_bio_all	1661	425	1236
IBD_all	1664	426	1238

除炎症性肠病数据集外，本研究还整合了斯坦福大学个体化医疗项目(http://med.stanford.edu/ipop.html)数据集，该数据集涵盖了代谢组学、蛋白质组学、肠道微生物 16S 测序)以及 RNA 测序丰度。主要用于研究与糖尿病相关的生物标志物。经预处理和特征过滤后，每个组学的具体信息如表 2 所示。

Table 2. Sample overlap and label distribution in the diabetes dataset
表 2. 糖尿病数据集样本重叠情况及标签信息

Datasets	Total Samples	Diabetes/Pre-diabetes	Control
Common Samples	606	479	49
RNA_Proteomics Common	805	629	79
RNA_bolomics Common	820	646	82
Proteomics_bolomics Common	903	698	94
RNA_Proteomics All	1008	782	101
All	1140	906	105

3. 实验设置

3.1. 实验环境设置

实验环境设置如表 3 所示。

Table 3. Experimental environment setup
表 3. 实验环境设置

软件包	版本号
TensorFlow	2.8.0
Scikit-learn	1.0.2
Pandas	1.3.5
NumPy	1.21.2
Matplotlib	3.4.3
SciPy	1.7.1

3.2. IMMO 模型架构设计

模型框架如图 1 所示。模型采用分阶段处理架构。

在进行数据预处理后，通过动态掩码[19]生成策略为所有组学生成随 epoch 演化的随机掩码，有效提升模型对数据缺失场景的适应能力；随机掩码之后通过独立编码网络提取各组学深层特征，利用批归一化与随机失活技术增强特征鲁棒性；中端构建跨模态融合模块，将异构特征映射至同一潜在空间；后端设计共享式解码器，通过联合重构机制实现多模态数据同步恢复。训练过程引入加权重构损失与自适应学习率策略，结合梯度裁剪与早停机制，在保证优化稳定性的同时促进跨模态特征均衡学习。

该框架通过动态掩码扰动与参数共享的协同优化，显著提升了多组学表征的可解释性，为复杂生物系统的深度解析提供了新方法。具体实施过程包括以下要点。

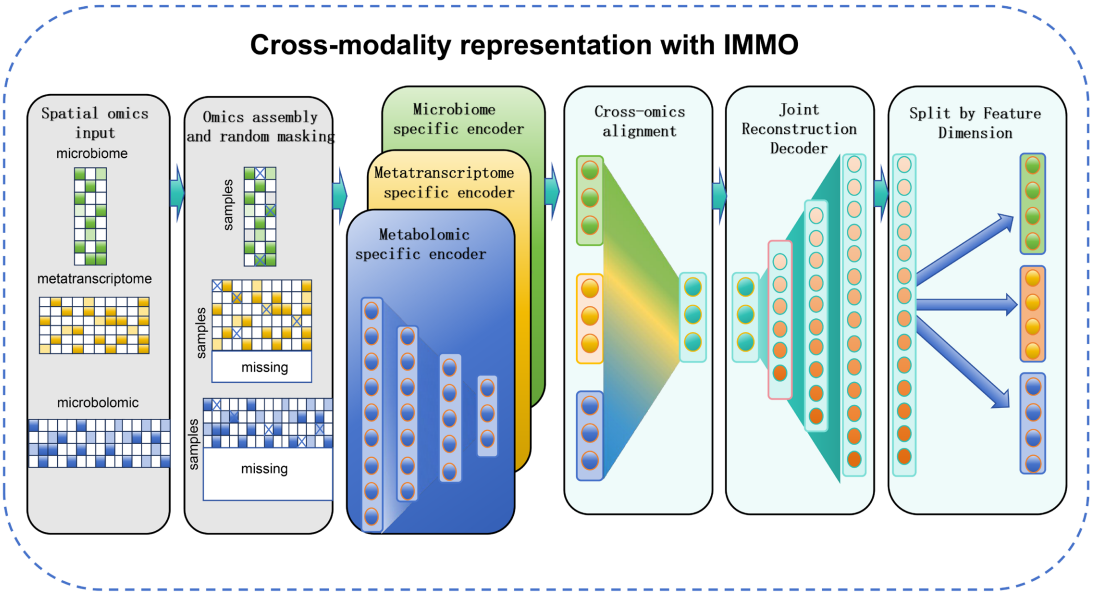

Figure 1. Overall framework of IMMO
图 1. IMMO 基本框架

(1) 动态掩码机制

在深度学习模型训练中，掩码技术被广泛应用于数据增强、正则化以及处理缺失数据等场景[21]。动态掩码策略即为在训练过程中，输入特征的一部分将被随机掩码(即设置为零)，并且保留概率 $P(t)$随着 epoch 指数增长，公式如下：

$$P(t)=P_0\cdot(1+r)^t,\ \text{subject to}\ P_0<P(t)\le P_{\max} \tag{1}$$

这里，P_0是初始保留概率，r 是增长率，t 表示训练 epoch 数，并且存在一个最大上限 $P_{\max}$。这意味着随着训练的进行，保留概率逐渐增加至一个预设的最大值，从而确保模型逐步暴露于更完整的数据中，同时保持其对不完整数据的鲁棒性。在每个 epoch 开始时，针对每种组学模态独立生成一个二进制动态掩码矩阵 $M(t)$，其中每个元素遵循伯努利分布：

$$M_{ij}^{dyn}(t)\sim\text{Bernoulli}\left(P(t)\right) \tag{2}$$

M_{ij}^{dyn} 指第 i 个样本的第 j 个特征是否被保留(=1)或被掩码(=0)。初始保留概率 P_0。对于不同的组学模

态是特定的，反映了各组学数据集的结构特性。

为了综合考虑固有的缺失值和训练过程中产生的随机掩码，在损失计算期间使用的最终有效掩码定义为：

$$M_{total}(t) = M_{true} \otimes M_{dyn}(t) \tag{3}$$

这里 M_{true} 是原始数据矩阵，通过与掩码矩阵进行逐元素乘法实现掩码。此机制确保只有在动态掩码方案下未被掩码且实际存在的数据才参与到损失计算中。

(2) 编码结构设计

IMMO-integration 采用多分支独立编码器架构，分别为各组学模态构建各自的全连接编码网络。每个编码器由多个全连接层构成，每层后接非线性激活 Swish 函数，定义为：

$$\text{Swish}(x) = x \cdot \sigma(x) \quad \sigma(x) = \frac{x}{1+\mathrm{e}^{-x}} \tag{4}$$

$\sigma(x)$ 表示 sigmoid 函数，接下来对每个小批量数据进行标准化处理，批归一化定义为：

$$BN(x) = \alpha\left(\frac{x-\mu_B B}{\sqrt{\sigma_B^2+\varepsilon}}\right)+\beta \tag{5}$$

其中 μ_B 和 σ_B^2 为批内均值与方差，α 、β 为可学习参数。

各编码器输出的低维潜在向量经进一步压缩后映射至共享潜在空间。该过程通过一个额外的全连接层实现维度对齐，最终形成一个融合多组学信息的紧凑表征。

(3) 解码结构设计

解码阶段作为编码过程的逆映射，目标是从共享潜在表示中重构原始多组学数据空间。模型采用共享解码器架构，包含若干全连接层，每层同样集成 Swish 激活、批归一化与 Dropout 机制，以保障梯度流动顺畅并抑制过拟合。解码器最后一层使用线性激活函数，输出维度等于所有组学特征维度之和，生成对原始输入的连续值预测。通过最小化重构误差，模型学习从潜在空间到输入空间的可逆映射，从而增强潜在表示的信息密度与生物学可解释性。

(4) 隐藏层定义与非线性变换

模型中所有隐藏层均为全连接结构，第 l 层输出定义为：

$$h^{(l)} = \text{Swish}\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \tag{6}$$

其中 $W^{(l)}$ 和 $b^{(l)}$ 分别为权重矩阵与偏置向量。

(5) 损失函数设置

在模型训练与潜在空间重构任务中，采用加权均方误差作为核心损失函数与重构性能度量指标，以应对不同组学间维度差异大及生物学重要性不同的问题。其定义如下：

$$WMSE = \sum_{i=1}^{N}\alpha_i \cdot \frac{1}{|X_i|}\sum_{x\in X_i} M(x)\cdot\left(x-\hat{x}\right)^2 \tag{7}$$

其中 N 是组学模态的数量，α_i 是分配给第 i 个模态的权重，X_i 是第 i 个模态的数据点数量，$M(x)$是一个二值掩码，用于指示数据点 x 是否被观测到：1 表示存在(观测到)，0 表示缺失。

3.3. 参数调优

为选择模型的最优超参数，本研究采用基于高斯过程的贝叶斯优化(Gaussian Process-based Bayesian

Optimization)。实施过程中将贝叶斯优化的目标函数即为模型损失函数。优化过程重点调整了六类关键超参数，包括潜在空间维度、编码器中 Dropout 层的失活率、优化器初始学习率、动态掩码机制中的初始特征保留概率以及学习率指数衰减因子和各组学模态在加权重构损失中的权重分配。

经过 50 轮贝叶斯优化迭代，算法收敛于一组稳定的超参数配置，最终确定的最优参数组合如表 1 所示。进一步分析发现最优解倾向于采用较高的初始特征保留概率(initial_p = 0.9)和较慢的学习率衰减(decay_rate = 0.99)。这在理论上是合理的：较高的初始保留率有助于模型在训练早期充分利用可观测数据，稳定编码器的初始化状态，避免因信息严重缺失而导致梯度不稳定。但在实际实验设置中，本研究将动态掩码机制的初始保留概率设定为 0.75，而非贝叶斯优化推荐的 0.9。这一调整并非对优化结果的否定，而是基于对 IBD 多组学数据特性的所作出的策略性选择。原始整合数据集中存在高达约 60%的跨组学缺失值，且缺失模式呈现非随机性，若采用 initial_p = 0.9 进行训练，则模型在初始阶段所面对的“人工缺失”强度远低于真实数据中的缺失水平，可能导致其对实际缺失机制的建模能力不足，进而削弱在真实应用场景下的泛化性能。为此，我们主动将初始保留概率下调至 0.75，使训练初期的掩码强度更贴近真实数据的观测稀疏性。这一设定有效增强了模型对高缺失率场景的鲁棒性，并促使编码器在早期即学习到更具恢复能力的潜在表示。后续的消融实验(见 4.1 节)进一步验证，相较于静态掩码或过高保留率的动态策略，p = 0.75 在重构误差、下游分类准确率及潜在表示的生物学一致性等多维度指标上均表现出更优的综合性能。潜在维度最终确定为 110，表明该维度足以捕获多组学数据的核心变异模式，同时避免过度参数化。损失权重最优组合赋予代谢组数据最高权重，反映出其在 IBD 表型关联中可能具有更强的判别能力，这与已有研究中代谢物作为功能输出端更贴近表型的结论一致。具体超参数调优过程如表 4 所示。

Table 4. Hyperparameter optimization process
表 4. 超参数调优过程

Hyperparameter	Range	Optimal Value
latent_dim	[32, 128]	110
dropout_rate	[0.1, 0.5]	0.47329
initial_learning_rate	[0.0001, 0.01]	0.00183
initial_p	[0.3, 0.9]	0.9
decay_rate	[0.9, 0.99]	0.99
loss_weights	[0.3_0.3_0.4, 0.4_0.3_0.3,0.5_0.25_0.25]	0.3_0.3_0.4

4. 实验结果

4.1. 降维重构动态掩码 VS 静态掩码

为充分验证动态掩码的有效性，本研究对比了静、动态掩码的训练损失函数。此处设置静态掩码的固定掩码概率为 0.75。两种掩码的损失函数如图 2 所示。从损失函数的变化来看，动态策略在前 10 个 epoch 损失下降 61.5%，静态同期下降 34.3%，整个过程中动态策略持续优化至 72 epochs，静态掩码在 24 epochs 提前终止，且动态掩码在后期训练过程中仍能保持较快下降速度，表明模型在动态遮挡模式下能够更有效学习数据特征，在 IBD 预测任务中展现出独特的优势。

4.2. 隐藏表示预测分类结果

为理解肠道微生物组与 IBD 之间的具体机制联系，在对原始多组学数据进行整合以后，将隐藏层用于预测样本是否为 IBD。三种组学数据共选择了 6 种有监督的学习方法，分别为深度神经网络(DNN)、

支持向量机(SVM)、随机森林(RF)、梯度提升树(GBDT)以及线性判别分析(LDA)和弹性网络(ElasticNet)。

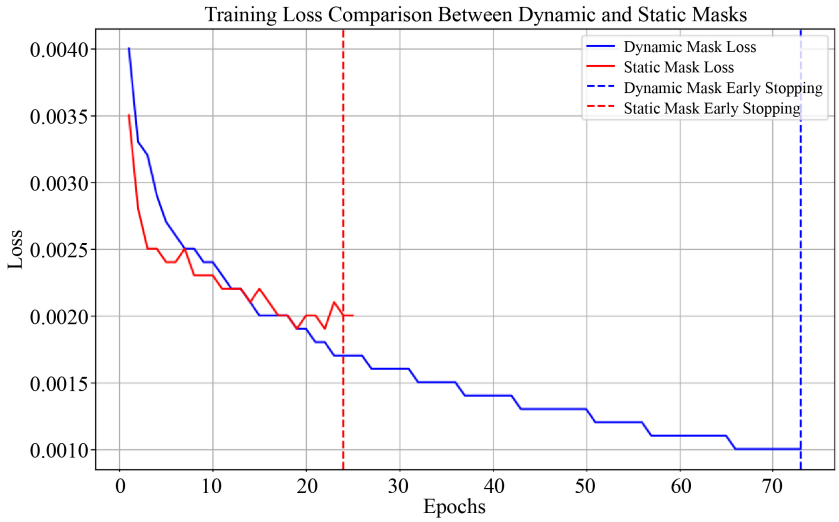

Figure 2. Training curves comparing dynamic and static masking strategies
图 2. 动态掩码 VS 静态掩码损失函数变化图

六种方法分类结果的 AUC 如表 5 所示，神经网络(DNN)展现出卓越的非线性建模能力，其 AUC 指标在大部分数据集显著优于其它方法，在 bol_bio_commo 数据集达到 0.976 的峰值。虽然支持向量机(SVM)在个别数据集上表现更优，但在其余数据集出现显著波动(0.864~0.962)，却易受特征冗余干扰。

传统线性模型包括线性判别分析(LDA)与岭回归(Ridge Regression)在实验中显现出明显局限性，其平均 AUC 均低于 0.885，故鉴于神经网络(Neural Network)的 AUC 指标均高于其它方法，将神经网络作为预测方法。

Table 5. Predictive performance AUC from different methods
表 5. 不同方法在隐藏层上预测 AUC

Datasets/Model	SVM	GBDT	Neural Network	LDA	Ridge Regression	RF
IBD_common	0.962	0.945	0.952	0.871	0.906	0.967
bol_tra_common	0.864	0.868	0.935	0.849	0.853	0.832
tra_bio_common	0.962	0.949	0.960	0.871	0.906	0.954
bol_bio_common	0.961	0.926	0.976	0.937	0.950	0.946
tra_bol_all	0.911	0.913	0.960	0.868	0.869	0.933
tra_bio_all	0.916	0.904	0.934	0.892	0.894	0.936
bol_bio_all	0.882	0.913	0.947	0.873	0.873	0.924
IBD_all	0.885	0.912	0.929	0.859	0.861	0.932

为评估动态掩码机制在多源异构数据整合中的有效性，本研究不仅采用隐藏层表示(Latent_110)作为输入进行预测，还以原始数据作为对照组进行实验。如图 3、图 4 所示，分别展示了基于隐藏层特征与原始数据在不同数据集上的模型预测性能(AUC 值)。结果表明，在八组不同数据组合中，Latent_110 模型

AUC 有六组显著高于原始数据，实现了性能提升，体现出其在提升预测性能方面的优越性。且使用原始数据直接建模时，各模型 AUC 波动较大，尤其在复杂数据集上表现不稳定；相比之下，基于隐藏层特征的模型展现出更高的鲁棒性，进一步验证了动态掩码降维在提取关键信息以及增强跨模态特征融合方面的有效性。

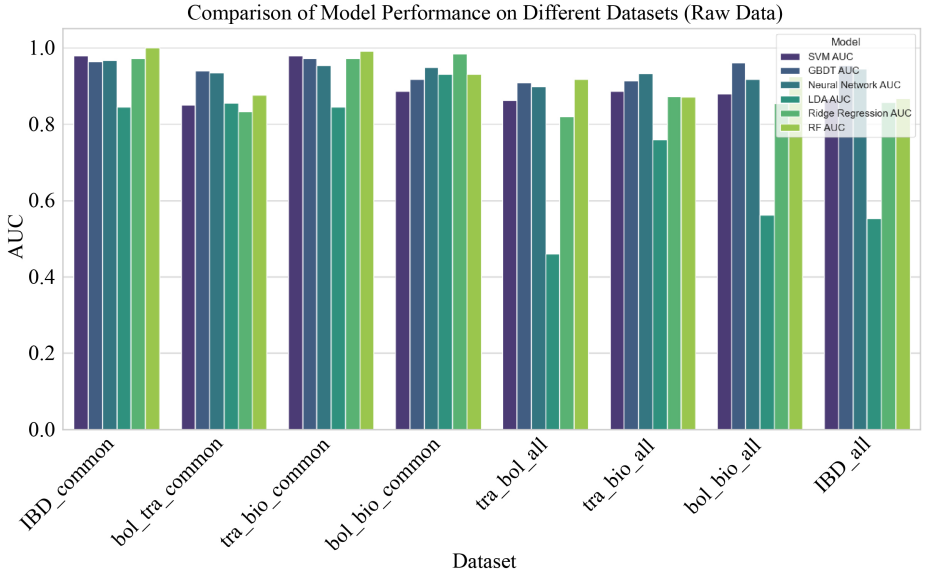

Figure 3. Comparison of model performance on different datasets (raw data)
图 3. 不同数据集上模型性能的比较(原始数据)

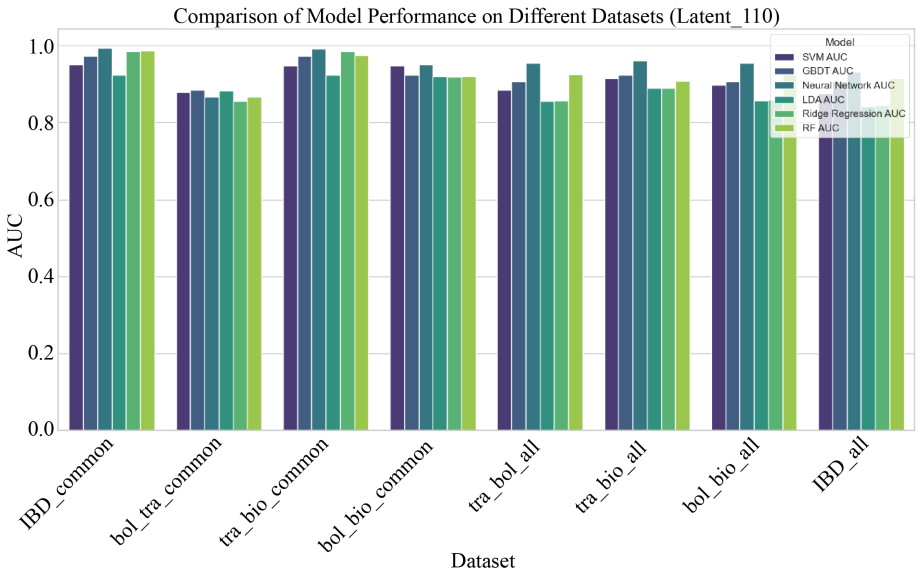

Figure 4. Comparison of model performance on different datasets (Latent_110)
图 4. 不同数据集上模型性能的比较(Latent_110)

图 3、图 4 显示了随着组学类型数量的减少，各类有监督的学习方法 AUC 的变化。图中显示随着组学类型数量的减少，虽然各种有监督学习方法 AUC 均有所下降，但如图 5，折线图较为平缓，尤其是神经网络和随机森林。神经网络在所有类别中都保持了相对稳定的性能，其在 2 类组学中的平均 AUC 为

0.955，与完整 3 类组学数据的 0.993 相比，差异不足 4%；即使在许多仅含 1 类组学数据的 IBD_all 中，仍能保持 0.933 的 AUC 值，显著优于传统 LDA 方法。这表明它对不完整数据具有鲁棒性。这表明使用模型降维整合后的神经网络方法能够充分利用样本数据，避免因数据缺失导致的信息损失。

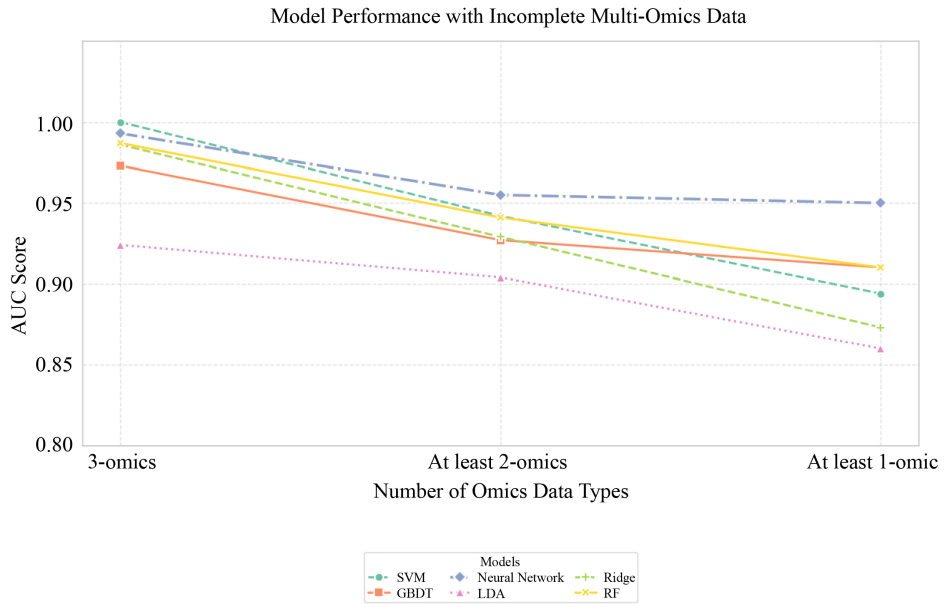

Figure 5. Model performance with incomplete multi-omics data
图 5. 不完整多组学数据下的模型性能

4.3. 潜在表征的生物学可解释性与特征归因分析

为深入理解模型预测背后的生物学逻辑，我们采用基于梯度的归因方法，系统评估了各组学特征对 IBD 分类的贡献程度。

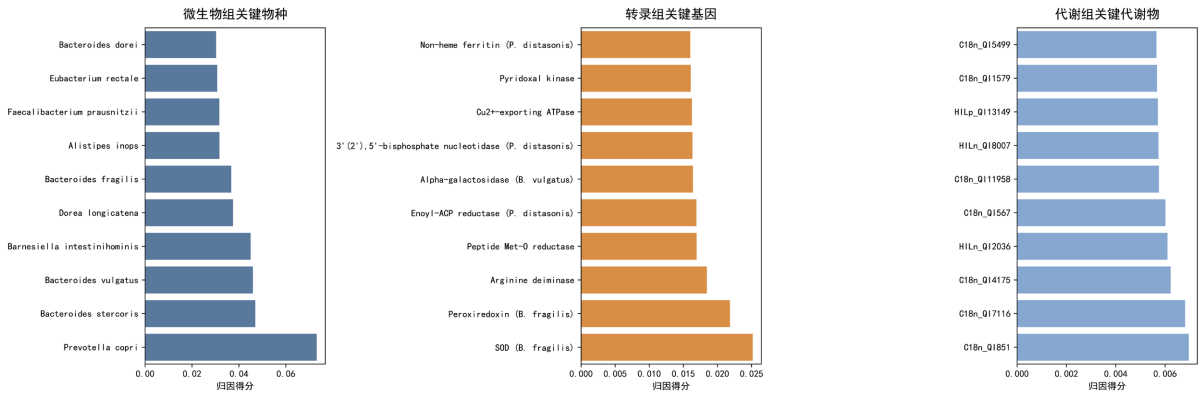

Figure 6. Top 10 most important features and attribution scores across the three omics layers
图 6. 三组学的前 10 重要特征及归因得分

如图 6 所示，在微生物组层面，Prevotella copri 的归因得分最高(0.073)，显著高于其他物种，多个拟杆菌属成员(如 *Bacteroides stercoris*、*B. vulgatus* 和 *B. fragilis*)以及以产丁酸著称的 *Faecalibacterium prausnitzii* 同样表现出较强的归因信号。这些菌种在维持肠道屏障完整性、调节宿主免疫反应等方面已有

广泛报道，其重要性进一步证实了模型捕捉到的生物学相关性。转录组层面的高贡献特征则集中于若干与氧化还原平衡和能量代谢相关的基因。例如，源自 *Bacteroides fragilis* 的超氧化物歧化酶(SOD)和过氧化物还原酶均显示出较高的归因值，说明肠道微生物可能通过调控局部氧化应激状态参与疾病进程。此外，精氨酸脱亚胺酶(arginine deiminase)等参与氨基酸代谢的酶类也被识别为关键因子，说明微生物在病理条件下可能发生代谢重编程。代谢组特征的单个归因值相对较低(最高约为 0.007)，但排名靠前的代谢物如 C18n_QI851 和 C18n_QI711 多属于脂质类化合物，且在多个样本中稳定出现。这一模式反映了宿主与微生物共同作用下脂质代谢通路的系统性扰动。

5. 总结

本研究针对微生物多组学数据中普遍存在的样本不完全匹配、模态缺失及高维稀疏性等挑战，提出了一种基于动态掩码的多组学数据整合模型 IMMO-integration。该模型通过引入随训练进程自适应调整的动态掩码机制，增强了对不完整输入数据的鲁棒性，避免了传统静态掩码可能导致的局部最优与泛化能力下降问题。结合参数共享的解码结构与加权重构损失函数，模型实现了跨组学特征的有效融合与均衡学习，在保证数值稳定性的同时提升了潜在表示的生物学可解释性。

在 IBDMDB 数据集上的实验结果表明，动态掩码策略显著优于静态掩码方案，训练损失持续下降至收敛，且在下游 IBD 状态预测任务中，基于潜在特征的神经网络分类在大部分数据子集上 AUC 超过 0.93，部分达到 0.99，验证了该模型在特征提取与判别建模方面的优越性能。进一步在斯坦福 IPOP 糖尿病相关多组学数据集上的外部验证显示，模型在不同疾病背景下仍保持稳定的分类表现，AUC 最高达 0.898，证明其具备良好的跨数据集泛化能力。

IMMO-integration 为不完全多组学数据的整合分析提供了一个高效、稳健且可扩展的计算框架。其不仅能够有效应对现实研究中常见的数据缺失问题，还能生成具有强判别力的低维表征，支持后续的生物标志物挖掘与疾病机制研究。此外，期望未来能够进一步拓展该模型至更多组学类型，并探索其在纵向多时点数据建模中的应用潜力，助力精准医学的发展。

参考文献

[1] Young, V.B. (2017) The Role of the Microbiome in Human Health and Disease: An Introduction for Clinicians. *British Medical Journal*, **356**, j831. https://doi.org/10.1136/bmj.j831

[2] Uebanso, T., Shimohata, T., Mawatari, K. and Takahashi, A. (2020) Functional Roles of B-Vitamins in the Gut and Gut Microbiome. *Molecular Nutrition & Food Research*, **64**, Article 2000426. https://doi.org/10.1002/mnfr.202000426

[3] Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., *et al*. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. *Science*, **312**, 1355-1359. https://doi.org/10.1126/science.1124234

[4] Glassner, K.L., Abraham, B.P. and Quigley, E.M.M. (2020) The Microbiome and Inflammatory Bowel Disease. *Journal of Allergy and Clinical Immunology*, **145**, 16-27. https://doi.org/10.1016/j.jaci.2019.11.003

[5] Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., *et al*. (2019) Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases. *Nature*, **569**, 655-662. https://doi.org/10.1038/s41586-019-1237-9

[6] Hu, X., Yu, C., He, Y., Zhu, S., Wang, S., Xu, Z., *et al*. (2024) Integrative Metagenomic Analysis Reveals Distinct Gut Microbial Signatures Related to Obesity. *BMC Microbiology*, **24**, Article No. 119. https://doi.org/10.1186/s12866-024-03278-5

[7] Jacobs, J.P., Lagishetty, V., Hauer, M.C., Labus, J.S., Dong, T.S., Toma, R., *et al*. (2023) Multi-Omics Profiles of the Intestinal Microbiome in Irritable Bowel Syndrome and Its Bowel Habit Subtypes. *Microbiome*, **11**, Article No. 5. https://doi.org/10.1186/s40168-022-01450-5

[8] Xu, J. and Yang, Y. (2021) Gut Microbiome and Its Meta-Omics Perspectives: Profound Implications for Cardiovascular Diseases. *Gut Microbes*, **13**, Article 1936379. https://doi.org/10.1080/19490976.2021.1936379

[9] Zhou, W., Sailani, M.R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S.R., *et al*. (2019) Longitudinal Multi-Omics of

Host-Microbe Dynamics in Prediabetes. *Nature*, **569**, 663-671. https://doi.org/10.1038/s41586-019-1236-x

[10] Deek, R.A., Ma, S., Lewis, J. and Li, H. (2024) Statistical and Computational Methods for Integrating Microbiome, Host Genomics, and Metabolomics Data. *eLife*, **13**, e88956. https://doi.org/10.7554/elife.88956

[11] Ronen, J., Hayat, S. and Akalin, A. (2019) Evaluation of Colorectal Cancer Subtypes and Cell Lines Using Deep Learning. *Life Science Alliance*, **2**, e201900517. https://doi.org/10.26508/lsa.201900517

[12] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., *et al*. (2020) MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data. *Genome Biology*, **21**, Article No. 111. https://doi.org/10.1186/s13059-020-02015-1

[13] Shen, R., Olshen, A.B. and Ladanyi, M. (2009) Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis. *Bioinformatics*, **25**, 2906-2912. https://doi.org/10.1093/bioinformatics/btp543

[14] Reel, P.S., Reel, S., Pearson, E., Trucco, E. and Jefferson, E. (2021) Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnology Advances*, **49**, Article 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

[15] Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., *et al*. (2022) MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Frontiers in Genetics*, **13**, Article 806842. https://doi.org/10.3389/fgene.2022.806842

[16] Liu, Q. and Song, K. (2023) ProgCAE: A Deep Learning-Based Method That Integrates Multi-Omics Data to Predict Cancer Subtypes. *Briefings in Bioinformatics*, **24**, bbad196. https://doi.org/10.1093/bib/bbad196

[17] Wang, F.A., Zhuang, Z., Gao, F., He, R., Zhang, S., Wang, L., *et al*. (2024) TMO-Net: An Explainable Pretrained Multi-Omics Model for Multi-Task Learning in Oncology. *Genome Biology*, **25**, Article No. 149. https://doi.org/10.1186/s13059-024-03293-9

[18] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North*, Minneapolis, June 2019, 4171-4186.

[19] He, K., Chen, X., Xie, S., Li, Y., Dollar, P. and Girshick, R. (2022) Masked Autoencoders Are Scalable Vision Learners. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16000-16009. https://doi.org/10.1109/cvpr52688.2022.01553

[20] Hu, M., Zhu, J., Peng, G., Lu, W., Wang, H. and Xie, Z. (2023) IMOVNN: Incomplete Multi-Omics Data Integration Variational Neural Networks for Gut Microbiome Disease Prediction and Biomarker Identification. *Briefings in Bioinformatics*, **24**, bbad394. https://doi.org/10.1093/bib/bbad394

[21] Yao, X., Huang, Z., Hu, X., Yang, J. and Guo, Y. (2024) Masking the Unknown: Leveraging Masked Samples for Enhanced Data Augmentation. *Proceedings of the* 40*th Conference on Uncertainty in Artificial Intelligence* (*UAI*), Barcelona, 15-19 July 2024, 3597-3606.