arXiv:2611.30601v2 [cs.AI] 24 Nov 2026

从部分配对的小规模数据中学习多模态表征

Merel Kuijs[*1, 2] **Till Richter**[*1, 2] **Rushin H Gindra**[*1, 3, 4, 6] **Korbinian Traeuble**[*1, 2] **Christian Matek**[5] **Rebeka Luknárová**[1] **TingyingPeng**[1,2,6] **FabianJTheis**[1,2]

[1]计算生物学研究所，德国环境健康研究中心，慕尼黑亥姆霍兹中心，诺伊赫贝格，德国[2] 计算机科学系，慕尼黑工业大学计算、信息与技术学院，慕尼黑工业大学，加兴，德国[3] 转化癌症研究与实验癌症治疗研究所，伊萨尔河右岸诊所，慕尼黑工业大学医学与健康学院，慕尼黑工业大学，德国[4] 转化癌症研究中心，慕尼黑工业大学，慕尼黑，德国[5]病理学研究所，埃朗根大学医院，埃朗根，德国[6]亥姆霍兹人工智能，慕尼黑亥姆霍兹中心，诺伊赫贝格，德国 通信作者：fabian.theis@helmholtz-munich.de

摘要

对部分配对样本的多模态数据进行建模，对于推动生物医学等领域的发展至关重要。在这些领域，虽然拥有大量单模态数据集和基础模型，但配对数据仍然稀缺。现有融合方法依赖大规模配对数据集，因此在配对不完全的场景中应用受限。我们提出 CAMEO，这是一种基于对抗学习的模态融合框架，可整合来自小规模、部分配对数据集的模态。CAMEO 将任意预训练的单模态编码器与跨模态潜在空间对齐机制相结合，只需少量配对样本即可学习共享表示。在生态位分类和细胞类型组成预测等计算病理学任务上的评估结果表明，CAMEO 具有更高的数据效率；在配对数据量较少的情况下，其表现优于 CLIP 等对比学习方法，也凸显了配对注释稀缺时对抗对齐的优势。为促进后续研究，我们还发布了一个完整标注的 HuggingFace 数据集，其中包含三种器官的配对图像和基因表达模态。通过扩展融合方法以应对配对有限和数据集规模较小的问题，我们提供了一个推动多模态学习发展、拓宽其在现实世界生物医学问题中适用范围的框架。

1 引言

空间转录组学和组织病理学（HP）成像的最新进展推动了我们对组织结构的理解，为我们提供了细胞状态的互补视角（Larsson 等，2021年；Palla 等，2022年；Fischer 等，2023年）。空间转录组学在捕捉基因表达的同时保留空间信息，而组织病理学成像则以高分辨率揭示组织形态和细胞组织结构。整合这些模态有望进一步深入了解组织结构和细胞相互作用。

基础模型已成为分析生物医学模态的强大工具（Szałata 等人，2024年）。这些模型利用大规模单模态数据集进行自监督学习（Richter 等人，2024年）。scGPT（Cui 等，2024a）和 Geneformer（Cui 等，2024年（b））等模型通过掩码语言建模，从基因表达数据（即定量转录组谱）中学习基因间依赖关系。这类模型适用于各种细胞层面和基因层面的下游

[*]共同贡献。

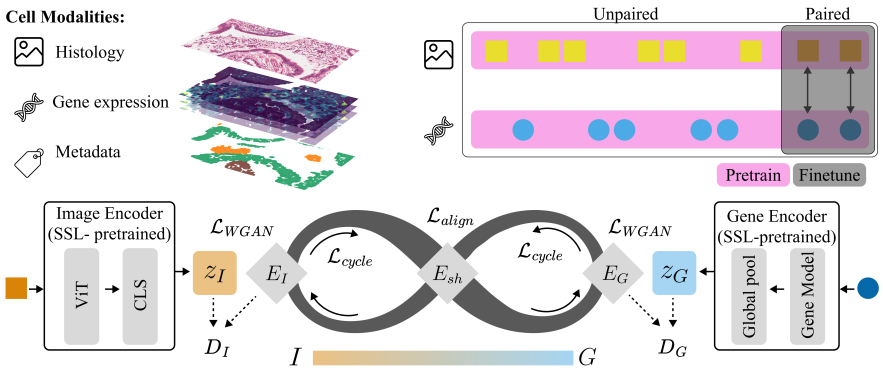


图 1：CAMEO 模型。左上：组织中的细胞由多种模态表示，包括组织学（视觉）、基因表达向量（功能生物学），以及我们用于下游任务的其他元数据（如专家注释）。右上：模型首先在视觉（上）和基因表达（下）数据仓库中广泛可获取的、未配对的单模态数据上进行预训练，然后在小规模配对数据集上进行微调。下方：单模态图像和基因表达编码器通过自监督学习进行训练，将图像映射到 z_I，并将基因表达向量映射到 z_G。随后，CAMEO 通过 E_{sh} 将这些生成的表示投影到共享多模态潜在空间中。这样即可根据图像表示预测基因表达表示，反之亦然。

任务，包括细胞注释和基因功能预测。UNI（Chen 等人，2024年）和 CTransPath（Wang 等人，2022年）等视觉模型使用对比目标，从 H&E 染色的全切片图像中学习形态模式。然而，这些方法没有利用多个模态中存在的潜在互补信息。尽管多模态模型已成功应用于语言和视觉等领域（Radford 等，2021），但它们通常需要大规模配对数据集，这限制了其在配对数据稀缺领域中的适用性。

我们提出了 CAMEO（图 1），一种对抗式模态融合框架，作为对比学习和朴素拼接方法的替代方案。在配对样本稀缺的生物医学场景中，CAMEO 的对抗机制仅需少量配对数据即可提供稳健的性能。我们的跨组织分析揭示了不同融合策略的互补优势：当配对数据有限时，对抗对齐具有更优的稳健性；同时，研究也强调了无论采用何种融合方法，都应系统地评估编码器组合的重要性。

我们的比较评估为计算病理学中的多模态整合策略选择提供了实用见解。总之，我们的贡献如下：

- 我们提出 CAMEO，这是一种对抗式模态融合框架，可在共享潜在空间中对齐预训练单模态编码器，从而有效利用规模较小且部分配对的数据集。
- 我们发布了一个经过对齐且由专家标注的 HuggingFace 数据集，整合了三种组织类型的图像和基因表达模态。
- 我们证明，多模态方法通常在小众分类任务上的表现优于单模态基线，而理想的编码器组合取决于数据集。
- 我们发现，单模态编码器的独立准确率难以预测其在多模态设置中的最终收益。这表明，有必要系统地评估编码器组合，而不能想当然地认为表现最佳的单模态模型会自动产生最佳融合模型。
- 我们展示了，与 CLIP 和拼接基线相比，CAMEO 对配对比例降低表现出更强的鲁棒性；在配对数据有限且存在类别分布偏移的场景中，这一点尤为重要。

2 相关工作

2.1 配对模态上的多模态训练

多模态表征学习算法通常依赖配对数据来对齐不同模态之间的嵌入。例如，CLIP（Radford 等，2021年）在包含 4 亿个图像-文本对的大规模语料库上进行训练，在零样本分类任务中展现出令人印象深刻的性能。同样，ImageBind（Girdhar 等人，2023年）在样本至少由一种共同模态表示的条件下对齐多种模态；当此类重叠样本稀缺时，这一条件可能具有较强限制性。BLIP-2（Li 等人，2023年）通过查询变换器将视觉特征与预训练语言模型对齐，并利用配对的图像-文本数据进行训练。类似地，LLaVa（Liu 等人，2023年）在配对的图像-文本数据上对预训练模型进行微调，以实现复杂推理。另一方面，CoCa（Yu 等人，2022年）从头开始训练，利用配对数据支持图像描述生成和跨模态对齐等多模态任务。

2.2 病理学中的多模态方法

在病理学领域，多模态方法通过整合影像与文本数据，提升诊断和预后任务的性能。这些方法通常分为两类：图像块级和切片级。图像块级视觉-语言方法（Huang 等人，2023年；Lu 等人，2024a；2024b；Zheng 等人，2024年；Ikezogwo 等人，2025年）通常采用 CLIP 风格的对比式预训练。相比之下，TITAN（Ding 等人，2024年）、PathAlign（Ahmed 等人，2024年）和 PRISM（Shaikovski 等人，2024年）等切片级方法，会在配对的全切片图像与文本数据上运用多实例学习（MIL）算法，以执行病理报告生成，以及组织学切片与临床注释的跨模态检索等任务。

除图像-文本数据外，图像-表达数据也已在多模态病理学方法中得到探索。Tangle（Jaume 等人，2024年）将组织学图像与配对的整体RNA测序数据相结合，用于少样本切片分类和检索；在切片级任务中，其性能优于单模态基线，体现了将高维测序数据与成像数据相结合的优势。RNA-CDM（Carrillo-Perez 等人，2024年）、HE2RNA（Schmauch 等人，2020年）和 BLEEP（Xie 等人，2023年）探索了成像模态与测序模态之间的转换，即根据一种模态的实例生成另一种模态。Yang 等人（2021年）则独特地利用未配对数据实现了这种转换。生存预测任务也显著受益于多模态方法，MOTCat（Xu 和 Chen，2023年）、CMTA（Zhou 和 Chen，2023年）、GPDBN（Wang 等人，2021年）和 HFBSurv（Li 等人，2022年）便是其中的代表。这些方法整合配对的成像数据和测序数据，构建预测模型，在临床结局预测方面优于单模态方法。较新的方法，如 MUSK（Xiang 等人，2025年），通过在大规模非配对数据集上进行预训练，降低了对完整配对数据的需求。尽管 MUSK 在精准肿瘤学任务中取得了优异性能，但其依赖大型语料库来学习稳健的跨模态嵌入；当可用数据量较少时，这种依赖可能会限制其泛化能力。

2.3 模态对齐与集成技术

近期的对齐研究尝试通过引入先验知识或因果推断技术，整合未配对的多模态样本。GLUE（Cao & Gao，2022年）利用知识图谱引导未配对多组学数据的整合，为组合不同模态提供了一个可解释的框架。然而，在构建合适的知识图谱不切实际或不可行时，其泛化能力会受到限制。Xi 等人（2024年）估算每种模态的倾向评分，以捕捉潜在状态与扰动之间的共享信息。尽管该方法能够在不同模态间匹配未配对样本，但仍需要稳健的数据整理。另一条研究路线聚焦于利用最优传输对齐未配对组学数据集。SCOT（Demetci 等人，2022年）和 GROT（Wang 等人，2024年）等方法利用共享的基因特征空间实现跨模态对齐，但基因表达与组织学图像缺少共同坐标系，因此难以定义二者之间具有生物学意义的传输成本。Nakada 等人（2023年）扩展了 CLIP 损失以纳入未配对数据，从而能够对从 MNIST 和 Fashion-MNIST 中抽取的配对图像进行对齐。这个合成数据集可能过度简化了多模态关系，从而限制了研究结果对更复杂自然数据分布的泛化能力。

我们的方法建立在不断发展的多模态对齐研究基础之上。不过，我们聚焦于临床病理学中的真实世界应用；由于获取配对样本成本高昂，此类应用中的数据有限，配对样本也十分稀缺。我们具体针对图像块级任务，而非切片级分析。由于后者依赖多实例学习算法，因此不适用于图像块级任务。不同于以往使用配对视觉语言数据集训练的图块级方法，我们的方法同时利用配对和未配对的视觉转录组学数据。通过使用预训练单模态编码器和跨模态对齐机制，我们能够以极少的配对标注样本，高效地适配下游任务。这一能力在医疗场景中尤为重要，因为收集大量配对多模态数据可能既耗时又昂贵。除了展示该方法在下游任务中的有效性外，我们还针对多种基因表达与图像编码器组合开展了系统的消融研究。我们发现，多模态模型通常比单模态基线取得更好的性能，但结果取决于所采用的编码器组合。重要的是，单独使用单模态编码器时的效果，并不能始终可靠地预测它们在多模态模型中的表现，这凸显了在为多模态方法选择编码器时进行全面评估的必要性。

3 方法

3.1 概述

CAMEO 采用模块化设计，兼容任何图像或基因表达编码器。它包含三个阶段：(1) 使用特定于模态的编码器进行初始数据嵌入，(2) 通过跨模态生成进行无监督表征对齐，以及 (3) 通过线性探测对下游任务进行监督评估。我们使用预计算的潜在嵌入，而非原始数据，从而降低计算复杂度和内存需求。CAMEO 的设计使我们能够整合强大的预训练编码器，例如在大规模组织学或单细胞数据上训练的基础模型，从而提升鲁棒性和泛化能力。我们的方法解决了多模态自监督学习中的若干关键挑战，包括数据可用性有限，这使得构建大型数据集成本高昂且技术难度大；配对数据稀缺，这限制了依赖完全配对示例的模型（如 CLIP）；以及批次效应，这些效应源自不同实验方案和设备引入的系统性差异。

接下来，我们将详细介绍模型的各个组成部分：1) 视觉编码器，2) 基因表达编码器，3) 使用多模态融合的预训练过程，以及 (4) 用于分类和回归的线性探测阶段。

3.2 视觉编码器

令 $x^I \in \mathbb{R}^{H\times W\times 3}$ 表示从组织切片中提取的图像块，其中 H 和 W 分别为图像块的高度和宽度，3 个通道对应 RGB 颜色强度。每张图像以 20 倍的平均放大倍率和每像素 0.5 微米（MPP）的分辨率展示其中细胞的形态结构。用于病理学的现代基础模型通常采用在大规模组织学切片数据集上预训练的视觉Transformer（ViT）（Dosovitskiy 等，2021）架构。为评估视觉编码器在多模态融合中的作用，我们使用五个预训练模型（四个基于 ViT 的编码器和一个卷积编码器）开展消融实验。这些编码器以即插即用的方式应用于 CAMEO 框架中。有关所使用具体编码器的更多细节，请参见第 4.3 节。

3.3 基因表达编码器

令 $X^G \in \mathbb{R}^{n\times p}$ 表示基因表达矩阵，其中 n 是细胞或空间点的数量，p 是测量基因的数量。每个输入 $x^G \in \mathbb{R}^p$ 都是一个高维向量，表示单个细胞或空间点的表达谱。基因表达编码器将这些向量转换为紧凑的潜在表示，从而捕获与下游任务相关的生物信号。

为系统评估不同基因编码策略对性能的影响，我们对五种基因表达编码器进行了消融实验。其中包括文献中的四种现有模型，以及我们自行训练的图注意力网络（GAT）（Richter 等，2023）。GAT 编码器遵循图注意力网络（Velickovic 等，2017）的公式，通过学习

学习到的注意力权重，从空间邻居聚合信息以生成节点嵌入。我们使用自监督掩码建模目标对 GAT 进行预训练：随机掩码每个表达向量的部分子集，并利用邻域结构根据未掩码的条目进行重建。所有编码器均以即插即用方式使用，其使用方法和评估设置的详细信息见第 4.3 节。

3.4 多模态融合

In t在这项工作中，我们提出了一种通用的对抗框架，用于学习跨模态表示 .

3.4.1 模型架构

对抗融合模块由两个跨模态生成器（$\mathcal{G}$）及其对应的判别器（$\mathcal{D}$）组成：

$$\begin{aligned} \mathcal{G}_{G\to I} &: \mathcal{X}_G \to \mathcal{X}_I & \mathcal{D}_I &: \mathcal{X}_I \to \mathbb{R} \\ \mathcal{G}_{I\to G} &: \mathcal{X}_I \to \mathcal{X}_G & \mathcal{D}_G &: \mathcal{X}_G \to \mathbb{R} \end{aligned}$$

其中， $\mathcal{X}_I \subset \mathbb{R}^{d_I}$ 和 $\mathcal{X}_G \subset \mathbb{R}^{d_G}$ 分别表示图像和基因表达的嵌入空间。生成器的输入为 $z_I \in \mathbb{R}^{d_I}$ 和 $z_G \in \mathbb{R}^{d_G}$，它们由模态特定的单模态编码器获得：

$$z_I = F_I(x^I), \quad z_G = F_G(x^G),$$

其中， F_I： $x^I \to \mathbb{R}^{d_I}$ 为图像编码器， F_G： $x^G \to \mathbb{R}^{d_G}$ 为基因表达编码器。每个生成器都采用共享瓶颈架构：

$$\begin{aligned} \mathcal{G}_{G\to I}(z_G) &= E_{sh}(E_G(z_G)) \\ \mathcal{G}_{I\to G}(z_I) &= E_{sh}(E_I(z_I)) \end{aligned}$$

其中， E_G 和 E_I 分别是投影层，用于调整表达数据和图像嵌入，然后将其输入共享层 E_{sh}。

3.4.2 训练目标

阶段 1：表示学习 在第一阶段，我们专注于从部分配对的图像和基因表达数据中学习共享表示。给定配对样本 $\{(x_i^I,\ x_i^G)\}_{i=1}^{n_p}$ 以及未配对样本 $\{x_j^I\}_{j=1}^{n_u^I}$、 $\{x_k^G\}_{k=1}^{n_u^G}$，我们的目标是学习模态之间的映射，同时保持语义关系。

CAMEO 的目标由三个部分组成：

1. **对抗损失**：对于每种模态，判别器学习区分真实嵌入和生成嵌入（Goodfellow 等人，2014年）：

$$\begin{aligned} \mathcal{L}_{WGAN} = &\mathbb{E}_{z_I\sim p_I}[\mathcal{D}_I(z_I)] - \mathbb{E}_{z_G\sim p_G}[\mathcal{D}_I(\mathcal{G}_{G\to I}(z_G))] \\ &+ \mathbb{E}_{z_G\sim p_G}[\mathcal{D}_G(z_G)] - \mathbb{E}_{z_I\sim p_I}[\mathcal{D}_G(\mathcal{G}_{I\to G}(z_I))] \end{aligned}$$

2. **循环一致性损失**：为确保跨模态翻译过程中的语义保持（Zhu 等人，2017年）：

$$\begin{aligned} \mathcal{L}_{cycle} = &\mathbb{E}_{z_I\sim p_I}[\|\mathcal{G}_{G\to I}(\mathcal{G}_{I\to G}(z_I)) - z_I\|_1] \\ &+ \mathbb{E}_{z_G\sim p_G}[\|\mathcal{G}_{I\to G}(\mathcal{G}_{G\to I}(z_G)) - z_G\|_1] \end{aligned}$$

3. **对齐损失**：对于配对示例集合 $\mathcal{P}$ 中的配对样本（z_I^p， z_G^p），嵌入会在共享潜在空间中进行显式对齐：

$$\mathcal{L}_{align} = \mathbb{E}_{(z_I^p, z_G^p)\sim\mathcal{P}}[\|E_{sh}(z_I^p) - E_{sh}(z_G^p)\|_2^2]$$

总损失由这些项结合权重因子构成：

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{cycle} + \lambda_2\mathcal{L}_{WGAN} + \lambda_3\mathcal{L}_{align}$$

形式上，阶段 1 的优化遵循极小极大形式：

$$\min_{\mathcal{G}_{I\to G},\mathcal{G}_{G\to I}} \left(\lambda_1 \mathcal{L}_{cycle} + \lambda_3 \mathcal{L}_{align}\right) + \min_{\mathcal{G}_{I\to G},\mathcal{G}_{G\to I}} \max_{\mathcal{D}_I,\mathcal{D}_G} \lambda_2 \mathcal{L}_{WGAN}.$$

在阶段 1 中，使用 $\mathcal{L}_{total}$更新生成器 $\mathcal{G}_{I\to G}$ 和 $\mathcal{G}_{G\to I}$，而判别器$\mathcal{D}_I$和 $\mathcal{D}_G$仅使用 $\mathcal{L}_{WGAN}$进行更新。

阶段 2：线性探测 在第二阶段，我们在配对数据上训练一个线性探测头。我们将生成器 $\mathcal{G}_{I\to G}, \mathcal{G}_{G\to I}$ 冻结。这些生成器仅用于在共享潜在空间中为每个基因表达样本和图像样本生成嵌入。随后，将两种模态产生的嵌入拼接，并输入线性探测头进行分类或回归：

$$\hat{y} = h_\theta([E_{sh}(E_I(z^I)); E_{sh}(E_G(z^G))])$$

在此阶段，仅更新线性探测头的参数 θ，并使用特定任务的损失函数：回归任务使用均方误差（MSE），分类任务使用交叉熵。附录 F 给出了用于评估的宏平均 F1 分数、微平均 F1 分数和 R^2 指标的定义。CAMEO 提供稳定、确定性的嵌入，能够适应有限的配对数据和不重合的流形结构（见附录 G）。

4 实验

本节评估各多模态框架从部分配对的小规模数据中学习的能力。我们的实验围绕两个关键问题展开：（1）这些表示是否优于仅利用图像或基因表达数据的单模态方法；（2）CAMEO 与需要更多配对数据的多模态基线方法相比表现如何。我们在广泛的编码器组合上评估性能，以即插即用方式改变基因表达和视觉骨干网络。此外，我们还考察不同配对比例下性能的变化，配对比例从完全配对到配对数据极其有限不等，从而能够在现实约束下分离跨模态对齐的影响。我们在第 4.1 节介绍数据集，在第 4.2 节介绍任务，在第 4.3 节介绍基线方法，并在第 4.4 节报告涵盖三种组织（乳腺、肺和胸腺）的定量结果。始终贯穿全文的是，我们强调编码器选择、数据可用性和任务特征在塑造多模态学习策略成效方面所发挥的作用。

4.1 数据

为了评估多模态学习框架和单模态基线，我们使用了两个 10x Xenium 空间转录组学队列和一个 10x Visium 空间转录组学队列。每个数据集都包含专家生成的真实标注，但由于类内异质性、类别不平衡，以及需要考虑空间和上下文依赖关系，预测这些标签颇具挑战。因此，这些数据集非常适合采用整合两种独立模态互补信息的方法。肺纤维化（LungPF）数据集包含 23 个 10x Xenium 样本（Vannan 等，2023年），这些样本用于研究肺纤维化的病理生物学。第二个 10x Xenium 队列包含来自 4 位患者的 7 个公开乳腺癌样本，托管于 10x Genomics 网站，并由我们的合作病理学家完成标注。第三个数据集是一个 10x Visium 队列，包含来自 11 位供体（6 位儿科供体、5 位胎儿供体）的 19 个胸腺样本（Yayon 等，2024年），用于研究胎儿期和出生后早期阶段的 T 细胞发育。为了构建用于多模态学习的 HuggingFace（HF）数据集，我们通过将每个图像块与其对应的基因表达谱关联起来，对基因表达和成像模态进行了对齐。所有数据集都将与研究界共享。我们按患者分层，将每个数据集划分为训练集、验证集和测试集（见附录 I）。

4.2 下游任务

4.2.1 生态位分类

微环境是具有生物学意义的组织区域，可调节细胞行为，通常涉及多种细胞类型和细胞外组分。生态位注释涉及根据

专家标注的真实标签识别这些区域，从而实现详细的组织表征。这对于理解区域特异性过程至关重要，例如免疫浸润或基质活性；这些过程与预后和治疗结局相关。在此任务中，图像块被视为生态位，并根据病理学家标注的组织形态分为 $n_{classes}$ 个不同类别。对于 LungPF 队列，共识别出 24 个组织学生态位，每个生态位代表一个独特的肺纤维化组织区域。同样，对于乳腺队列，共标注了 14 个不同的组织学特征。对于胸腺队列，则根据形态学标志标注了 11 个区域（Yayon 等，2024年）。有关标注的更多详细信息及其分布见附录 I。

4.2.2 细胞类型组成预测

许多疾病，包括癌症和自身免疫性疾病，都涉及细胞类型组成的变化。根据基因表达数据估算细胞类型组成，有助于识别与疾病相关的诊断生物标志物，并深入了解疾病微环境，这对开发靶向治疗至关重要。在细胞类型组成预测任务中，目标是估算图像块中各细胞类型所占的比例。我们评估模型预测由 LungPF 队列创建团队定义的 9 种广义细胞类型组成的能力。更多详情见附录 I。

表1：使用**100% 配对数据**对肺（$n_{classes} = 24$）、乳腺（$n_{classes} = 14$）和胸腺（$n_{classes} = 11$）数据集上的模型进行多类别生态位分类性能比较。星号（*）表示未在测试集上评估该模型。GAT 模型仅兼容单细胞空间数据。因此，我们在胸腺数据集上的实验中排除了 GAT 编码器，因为该数据集包含基于点位的基因表达测量。蓝色、橙色和绿色文本分别表示图像编码器、基因表达编码器和多模态方法。

数据类型	模型名称	Lung		乳腺		胸腺	
		宏平均 F1	微平均 F1	宏平均 F1	微平均 F1	宏平均 F1	微平均 F1
图像	CONCH	0.099	0.286	0.169	0.197	0.152	0.211
	CTransPath	0.105	0.304	0.128	0.188	0.127	0.218
	ResNet	0.096	0.246	0.104	0.156	0.136	0.205
	UNI	0.134	0.401	0.133	0.133	0.127	0.194
	Virchow	0.116	0.391	0.145	0.146	0.158	0.222
基因表达	CellPLM	0.183	0.199	0.130	0.369	0.134	0.227
	GAT	0.171	0.263	0.181	0.383	*	*
	Nicheformer	0.158	0.211	0.134	0.242	0.102	0.183
	scGPT	0.177	0.269	0.153	0.386	0.134	0.208
	scVI	0.125	0.234	0.144	0.312	0.164	0.253
多模态	CAMEO-CTransPath-GAT	0.166	0.177	0.096	0.335	*	*
	CAMEO-CTransPath-Nicheformer	0.075	0.064	0.080	0.233	0.193	**0.347**
	CAMEO-CTransPath-scVI	0.158	0.223	0.079	0.294	0.155	0.214
	CAMEO-CONCH-GAT	0.128	0.287	0.124	0.374	*	*
	CAMEO-CONCH-Nicheformer	0.092	0.253	0.186	0.277	0.207	0.290
	CAMEO-CONCH-scVI	0.164	0.278	0.154	**0.445**	0.212	0.293
	CAMEO-UNI-GAT	0.137	0.274	0.098	0.227	*	*
	CAMEO-UNI-Nicheformer	0.193	0.308	0.141	0.297	0.172	0.302
	CAMEO-UNI-scVI	0.137	0.249	0.181	0.411	0.185	0.267
	CLIP-CTransPath-GAT	0.199	**0.412**	0.204	0.407	*	*
	CLIP-CTransPath-Nicheformer	0.188	0.330	0.225	0.382	0.176	0.255
	CLIP-CTransPath-scVI	0.182	0.352	0.209	0.419	0.202	0.269
	CLIP-CONCH-GAT	0.180	0.358	0.195	0.341	*	*
	CLIP-CONCH-Nicheformer	0.158	0.309	0.196	0.341	0.173	0.247
	CLIP-CONCH-scVI	0.157	0.315	0.184	0.342	0.161	0.248
	CLIP-UNI-GAT	0.195	0.382	0.194	0.422	*	*
	CLIP-UNI-Nicheformer	0.177	0.334	0.133	0.332	0.170	0.251
	CLIP-UNI-scVI	0.194	0.335	0.170	0.305	0.188	0.259
	拼接-CTransPath-GAT	**0.214**	0.368	0.293	0.337	*	*
	拼接-CTransPath-Nicheformer	0.203	0.399	0.218	0.266	0.187	0.283
	拼接-CTransPath-scVI	0.165	0.300	0.240	0.241	**0.227**	0.301
	拼接-CONCH-GAT	0.162	0.350	**0.305**	0.359	*	*
	拼接-CONCH-Nicheformer	0.181	0.298	0.140	0.187	0.176	0.276
	拼接-CONCH-scVI	0.108	0.277	0.250	0.300	0.179	0.253
	拼接-UNI-GAT	0.145	0.392	0.152	0.233	*	*
	拼接-UNI-Nicheformer	0.130	0.398	0.140	0.141	0.148	0.198
	拼接-UNI-scVI	0.126	0.322	0.136	0.199	0.152	0.190

4.3 基线方法

我们评估三类基线方法：仅图像、仅基因表达和多模态。图像基线使用五种预训练编码器：UNI（Chen 等人，2024年）、CTransPath（Wang 等人，2022年）、ResNet-50（He 等人，2015年）、Virchow（Vorontsov 等人，2024年）和 CONCH（Lu 等人，2024a）。这些模型均处于冻结状态，并在其上训练线性分类器（线性探测），以评估其表征质量。基因表达基线以高维基因表达向量作为输入。我们对 CellPLM（Wen 等人，2023年）、scGPT（Cui 等，2024a）、Nicheformer（Schaar 等，2024）以及第 3.3 节所述的 GAT 编码器进行线性探测，并保持所有编码器冻结。对于 scVI（Lopez 等，2018），我们仅使用配对 HuggingFace 数据集的训练集划分从头训练编码器，并在学习到的嵌入上训练线性分类器。对于多模态基线方法，我们考虑两种框架：CLIP（Radford 等，2021）和“拼接”。在“拼接”框架中，单模态基因表达编码器和图像编码器的原始输出嵌入会直接拼接，不进行任何显式对齐或融合。我们通过不同的基因表达编码器与图像编码器组合，对每种多模态框架进行测试。所有多模态框架都会生成联合嵌入，并用其训练特定任务的线性头。所有基线方法的更多详情见附录 E。

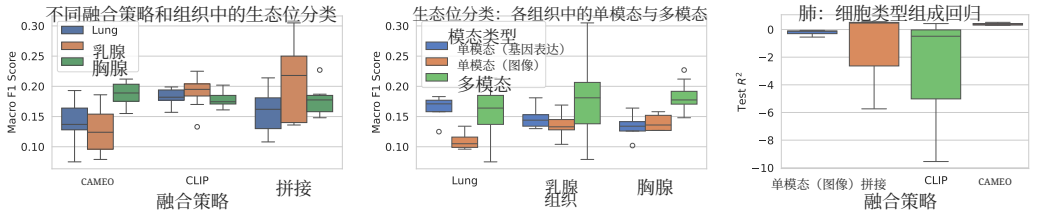


图 2：按模态和融合策略进行的性能比较。左：各融合方法在不同组织中使用的编码器组合。中：使用 100% 配对数据时，单模态模型与多模态模型的宏平均 F1 分数。右：细胞类型组成预测的回归测试 R^2，结果显示多模态模型优于单模态方法，其中 CAMEO 的性能最高。

4.4 结果

所有模型均使用相同的训练集、验证集和测试集划分进行评估（按患者分层）。由于存在显著的类别不平衡（见附录 I），我们使用宏平均 F1 分数作为生态位分类的主要指标，并使用 R^2 分数进行细胞类型组成回归。

我们在三种组织类型（肺、乳腺和胸腺）上进行的实验得出了一些一致的发现。首先，多模态方法优于其单模态方法，表明整合图像和基因表达数据中的互补信息具有重要价值，如图2（中）所示。在细胞类型组成预测任务中，这一优势得到了进一步验证，其中 CAMEO 获得了最高的 R^2 分数（图2，右）。然而，性能提升因组织类型和编码器组合而异，如图2（左）所示。在肺部数据集中，Concat–CTransPath–GAT 的宏平均 F1 分数最高（0.214）。对于乳腺组织，Concat–CONCH–GAT 表现最佳（0.305）；而在胸腺组织中，Concat–CTransPath–scVI 的结果最高（0.227）。有趣的是，在乳腺和胸腺组织中，CAMEO 的微平均 F1 分数均优于所有 Concat 变体：乳腺组织中 CAMEO–CONCH–scVI 达到 0.445，胸腺组织中 CAMEO–CTransPath–Nicheformer 达到 0.347（表1）。尽管在使用 100% 配对数据训练时，Concat 模型表现出乎意料地出色（宏平均 F1），但随着配对数据比例降低，其性能也随之下降，下一节将对此进行展示。

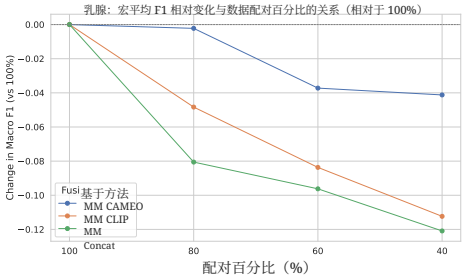


图 3：随着配对百分比降低，乳腺宏平均 F1 分数的相对变化。与 CLIP 和拼接相比，CAMEO 表现出更强的稳健性。肺部也呈现出类似的模式，见附录图 5。

表 2：使用默认 UNI 和 scVI 编码器时，在减少数据配对场景下的生态位分类性能。

组织	方法	100%	80%	60%	40% (Δ)
Lung	CAMEO	0.137	**0.153**	**0.158**	**0.143** (Δ **4.4%**)
	CLIP	**0.194**	0.146	0.128	0.133 (Δ-31.5%)
	拼接	0.126	0.125	0.117	0.123（Δ-2.2%）
乳腺	CAMEO	**0.181**	**0.157**	0.086	**0.103** (Δ-43.3%)
	CLIP	0.170	0.103	0.108	0.078（Δ-54.4%）
	拼接	0.136	0.119	**0.122**	0.102（Δ-**25.1%**）

配对百分比与性能之间的关系

性能揭 ls that whil的性f 性能 typically

随着配对数据减少，性能有所下降，但不同方法的下降幅度差异显著。值得注意的是，与 CLIP 相比，CAMEO 在配对比例降低时表现出更强的鲁棒性，尤其是在乳腺数据集上（图3）。肺部数据集也呈现出类似趋势（附录图 5）。这种鲁棒性在分类和回归任务中均保持一致：随着配对数据减少，CAMEO 仍能维持更强的性能（附录图 5）。这些趋势反映了不同编码器组合下的平均性能。表 2 使用 UNI 和 scVI 编码器提供了详细结果：对于肺部数据集，在配对比例为 40% 时，CAMEO 的性能提升了 4.4%，而 CLIP 和拼接的性能分别下降了 31.5% 和 2.2%；对于乳腺组织，即使配对数据减少，CAMEO 仍保持最高性能（0.103）。我们认为，CAMEO 的鲁棒性源于其对抗式对齐流程，该流程能够有效利用未配对数据来保持表征质量。

4.5 消融研究

4.5.1 编码器选择的影响因多模态框架而异

我们的结果表明，编码器选择会对性能产生关键影响，但不同多模态框架的最优组合各不相同。例如，在肺部数据集中，表现最佳的 CAMEO 变体是 CAMEO–UNI–Nicheformer（0.193），而 Concat 则通过 Concat–CTransPath–GAT 获得最高分（0.214）。这种不一致性同样延伸至其他组织，其中一个框架表现最佳的编码器组合很少会与另一个框架的最优选择一致（表1）。这一发现凸显了在实施任何多模态方法时开展全面编码器评估的重要性。

4.5.2 单模态性能无法预测多模态成功

一个出人意料的发现是，单模态编码器的性能无法很好地预测其在多模态设置中的有效性，如乳腺数据集的图4所示。例如，尽管 GAT 在该数据集的基因表达编码器中取得了最高的宏平均 F1 分数（0.181），但 CLIP 和 CAMEO 中表现最佳的多模态组合却使用了 Nicheformer。这一模式在不同组织中持续存在（附录图6），表明编码器间的互补性可能比单个编码器的强度更为重要。这种非预测性关系对
仅根据编码器的独立性能来选择编码器是一种常见做法；这凸显了在多模态框架中系统评估编码器组合的必要性。

4.6 建议

根据我们的研究结果，我们针对多模态框架的开发提出以下最佳实践建议：(i) 系统地评估多种编码器组合，因为基因表达与图像编码器的最佳配对可能因组织类型和多模态整合方法的不同而有所变化；(ii) 将拼接作为具有竞争力的基线方法，因为它通常表现良好；(iii) 在配对数据有限或存在标签不平衡的情况下，考虑采用 CAMEO 等对抗方法，因为它们能够有效利用非配对样本。

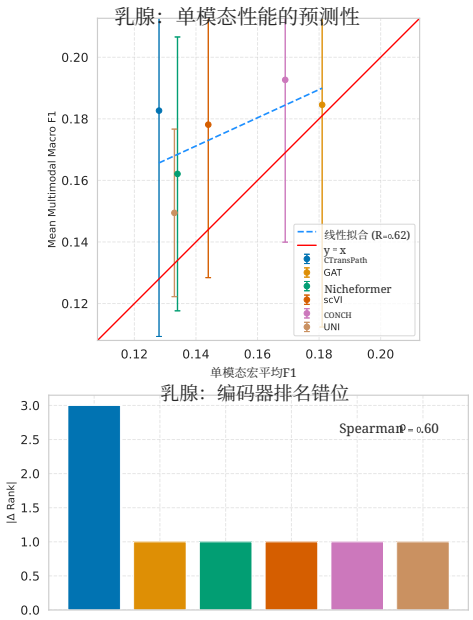


图4：乳腺数据集中单模态性能对多模态成功的预测性。相关性中等（Spearman $\rho = 0.60$），多模态模型之间的方差较大，且排名错位显著（底部面板）。结果 f

5 讨论

我们的研究提出，在配对样本稀缺的生物医学场景中，对抗对齐是现有多模态融合方法的一种有效替代方案。CAMEO 在配对数据减少时仍表现出更强的稳健性，证明了对抗机制在利用未配对样本方面的价值，

这在计算病理学中尤为重要，因为单模态数据丰富，但生成配对数据集仍然很困难。出乎意料的是，我们发现单模态编码器的性能并不能很好地预测多模态成功，这挑战了模型选择中的传统观点，并凸显了编码器互补性的重要性

互补性。这表明，在融合框架中编码器之间的交互方式可能比它们各自的性能更为重要，因此在实施多模态方法时，需要评估多种组合。虽然我们的工作聚焦于空间组学和组织病理学，但本文探讨的对抗对齐原则也可以扩展到其他配对数据有限的领域。通过发布专家标注的多模态数据集，并为模态融合策略提供实践见解，我们既推动了方法学的发展，也提供了相关资源，以促进在数据受限环境中进一步探索多模态学习。

致谢

K.T.、T.R. 和 R.H.G. 得到了亥姆霍兹联合会在联合研究学院“慕尼黑数据科学学院 – MUDS”框架下的支持。T.R. 还获得了慕尼黑机器学习中心（MCML）的额外支持，该中心由联邦研究、技术与航天部（BMFTR）和巴伐利亚自由州资助。

参考文献

Faruk Ahmed、Andrew Sellergren、Lin Yang、Shawn Xu、Boris Babenko、Abbi Ward、Niels Olson、Arash Mohtashamian、Yossi Matias、Greg S. Corrado、Quang Duong、Dale R. Webster、Shravya Shetty、Daniel Golden、Yun Liu、David F. Steiner 和 Ellery Wulczyn。《PathAlign：用于组织病理学全切片图像的视觉语言模型》，2024年。网址：https://arxiv.org/abs/2406.19578。

Mustapha Bounoua、Giulio Franzese 和 Pietro Michiardi。利用最小熵耦合学习匹配未配对数据。载于 第*42*届国际机器学习会议论文集（*ICML 2025*），2025年。网址：https://arxiv.org/abs/2503.08501。

Zhi-Jie Cao 和 Ge Gao。利用图链接嵌入进行多组学单细胞数据整合和调控推断。《自然·生物技术》，40(10)：1458–1466，2022年10月。

FranciscoCarrillo-Perez、MarijaPizurica、YuanningZheng、TarakNathNandi、RaviMadduri、Jeanne Shen 和 Olivier Gevaert。通过级联扩散模型从RNA测序数据生成肿瘤合成全切片图像瓦片。《自然·生物医学工程》，2024年3月。

Richard J. Chen、Tong Ding、Ming Y. Lu、Drew F. K. Williamson、Guillaume Jaume、Andrew H. Song、Bowen Chen、Andrew Zhang、Daniel Shao、Muhammad Shaban、Mane Williams、Lukas Oldenburg、Luca L. Weishaupt、Judy J. Wang、Anurag Vaidya、Long Phi Le、Georg Gerber、Sharifa Sahai、Walt Williams 和 Faisal Mahmood。迈向用于计算病理学的通用基础模型。《自然·医学》，30(3)：850–862，2024年。

Haotian Cui、Chloe Wang、Hassaan Maan、Kuan Pang、Fengning Luo、Nan Duan 和 Bo Wang。scgpt：利用生成式人工智能构建单细胞多组学基础模型。《自然·方法》，第1–11页，2024a。

Zhanbei Cui、Tongda Xu、Jia Wang、Yu Liao 和 Yan Wang。Geneformer：使用基于Transformer的上下文建模进行基因压缩。在*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第8035–8039页。IEEE，2024年（b）。

Pinar Demetci、Rebecca Santorella、Björn Sandstede、William Stafford Noble 和 Ritambhara Singh。通过SCOT进行单细胞多组学整合。《计算生物学杂志》，29(1)：19–22，2022年1月。doi：10.1089/cmb.2021.0446。URL https://doi.org/10.1089/ cmb.2021.0446。

Tong Ding、Sophia J. Wagner、Andrew H. Song、Richard J. Chen、Ming Y. Lu、Andrew Zhang、Anurag J. Vaidya、Guillaume Jaume、Muhammad Shaban、Ahrong Kim、Drew F. K. Williamson、Bowen Chen、Cristina Almagro-Perez、Paul Doucet、Sharifa Sahai、Chengkuan Chen、Daisuke Komura、Akihiro Kawabe、Shumpei Ishikawa、Georg Gerber、Tingying Peng、Long Phi Le 和 Faisal Mahmood。《病理学多模态全视野切片基础模型》，2024年。URL https: //arxiv.org/abs/2411.19666。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。《一张图像胜过 16x16 个词：用于大规模图像识别的 Transformer》，2021年。URL https://arxiv.org/abs/2010.11929。

C.W. ELSTON 和 I.O. ELLIS。乳腺癌的病理学预后因素。I. 组织学分级在乳腺癌中的价值：一项大型研究的长期随访经验。组织病理学，19(5):403–410，1991年。

David S Fischer、Anna C Schaar 和 Fabian J Theis。利用细胞空间图对组织中的细胞间通信进行建模。《自然·生物技术》，41(3):332–336，2023年。

Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：一个将所有内容绑定在一起的嵌入空间，2023年。网址：https://arxiv.org/abs/2305.05665。

Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。神经信息处理进展，27，2014年。

何恺明、张祥宇、任少卿和孙剑。用于图像识别的深度残差学习，2015年。URL https://arxiv.org/abs/1512.03385。黄智、Federico Bianchi、Mert Yuksekgonul、Thomas J. Montine 和 James Zou。利用医学推特进行病理图像分析的视觉-语言基础模型。《自然·医学》，29(9):2307–2316，2023年。Wisdom Oluchi Ikezogwo、Mehmet Saygin Seyfioglu、Fatemeh Ghezloo、Dylan Stefan Chan Geva、Fatwir Sheikh Mohammed、Pavan Kumar Anand、Ranjay Krishna 和 Linda Shapiro。Quilt-1M：用于组织病理学的一百万个图像-文本对，2025年。URL https://arxiv.org/abs/2306.11207。Guillaume Jaume、Lukas Oldenburg、Anurag Vaidya、Richard J. Chen、Drew F. K. Williamson、Thomas Peeters、Andrew H. Song 和 Faisal Mahmood。计算病理学中由转录组学引导的切片表征学习，2024年。URL https://arxiv.org/abs/2405. 11618。Ludvig Larsson、Jonas Frisén 和 Joakim Lundeberg。空间分辨转录组学为基因组学增添了新的维度。《自然·方法》，18(1):15–18，2021年。李俊南、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型引导语言图像预训练，2023年。URL https://arxiv.org/abs/2301.12597。

Ruiqing Li、吴兴奇、Ao Li 和 Minghui Wang。HFBSurv：采用因子化双线性模型的层次化多模态融合用于癌症生存预测。*Bioinformatics*，38(9)：2587–2594，2022年4月。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。《视觉指令微调》，2023年。网址：https://arxiv.org/abs/2304.08485。Romain Lopez、Jeffrey Regier、Michael B Cole、Michael I Jordan 和 Nir Yosef。《单细胞转录组学的深度生成建模》。《自然·方法》，15(12):1053–1058，2018年12月。Ming Y Lu、Bowen Chen、Drew F K Williamson、Richard J Chen、Ivy Liang、Tong Ding、Guillaume Jaume、Igor Odintsov、Long Phi Le、Georg Gerber、Anil V Parwani、Andrew Zhang 和 Faisal Mahmood。《用于计算病理学的视觉语言基础模型》。《自然·医学》，30(3):863–874，2024年3月。Ming Y. Lu、Bowen Chen、Drew F. K. Williamson、Richard J. Chen、Melissa Zhao、Aaron K. Chow、Kenji Ikemura、Ahrong Kim、Dimitra Pouli、Ankush Patel、Amr Soliman、Chengkuan Chen、Tong Ding、Judy J. Wang、Georg Gerber、Ivy Liang、Long Phi Le、Anil V. Parwani、Luca L. Weishaupt 和 Faisal Mahmood。《用于人类病理学的多模态生成式人工智能副驾驶》。*Nature*，634(8033):466–473，2024年（b）。Luca Marconato、Giovanni Palla、Kevin A. Yamauchi、Isaac Virshup、Elyas Heidari、Tim Treis、Wouter-Michiel Vierdag、Marcella Toth、Sonja Stockhaus、Rahul B. Shrestha、Benjamin Rombaut、Lotte Pollaris、Laurens Lehner、Harald Vöhringer、Ilia Kats、Yvan Saeys、Sinem K. Saka、Wolfgang Huber、Moritz Gerstung、Josh Moore、Fabian J. Theis 和 Oliver Stegle。《Spatialdata：用于空间组学的开放通用数据框架》。《自然·方法》，22(1):58–62，2025年。Ryumei Nakada、Halil Ibrahim Gulluk、Zhun Deng、Wenlong Ji、James Zou 和 Linjun Zhang。《理解多模态对比学习并纳入未配对数据》，2023年。网址：https://arxiv.org/abs/2302.06232。Giovanni Palla、Hannah Spitzer、Michal Klein、David Fischer、Anna Christina Schaar、Louis Benedikt Kuemmerle、Sergei Rybakov、Ignacio L Ibarra、Olle Holmberg、Isaac Virshup 等。《Squidpy：用于空间组学分析的可扩展框架》。《自然·方法》，19(2):171–178，2022年。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。《从自然语言监督中学习可迁移视觉模型》，2021年。网址：https: //arxiv.org/abs/2103.00020。

Till Richter、Anna Schaar、Francesca Drummer、Cheng-Wei Liao、Leopold Endres 和 Fabian J Theis。《Spatialssl：利用自监督学习进行小鼠脑全脑空间转录组学》。载于*NeurIPS 2023* 科学人工智能研讨会，2023年。

Till Richter、Mojtaba Bahrami、Yufan Xia、David S Fischer 和 Fabian J Theis。阐明自监督学习在单细胞基因组学中的有效应用。《自然·机器智能》，第 1–11 页，2024年。

Anna C Schaar、Alejandro Tejada-Lapuerta、Giovanni Palla、Robert Gutgesell、Lennard Halle、Mariia Minaeva、Larsen Vornholz、Leander Dony、Francesca Drummer、Mojtaba Bahrami 等。Nicheformer：用于单细胞和空间组学的基础模型。*bioRxiv*，第 2024–04 页，2024年。

Benoît Schmauch、Alberto Romagnoni、Elodie Pronier、Charlie Saillard、Pascale Maillé、Julien Calderaro、Aurélie Kamoun、Meriem Sefta、Sylvain Toldo、Mikhail Zaslavskiy、Thomas Clozel、Matahi Moarii、Pierre Courtiol 和 Gilles Wainrib。利用全切片图像预测肿瘤 RNA测序表达的深度学习模型。*Nature Communications*，11(1):3877，2020年8月。

George Shaikovski、Adam Casson、Kristen Severson、Eric Zimmermann、Yi Kan Wang、Jeremy D. Kunz、Juan A. Retamero、Gerard Oakley、David Klimstra、Christopher Kanan、Matthew Hanna、Michal Zelechowski、Julian Viret、Neil Tenenholtz、James Hall、Nicolo Fusi、Razik Yousfi、Peter Hamilton、William A. Moye、Eugene Vorontsov、Siqi Liu 和 Thomas J. Fuchs。Prism：用于切片级组织病理学的多模态生成式基础模型，2024年。网址：https: //arxiv.org/abs/2405.10254。

Akash Singh、Kirti Biharie、Marcel J. T. Reinders、Ahmed Mahfouz 和 Tamim Abdelaal。scTopoGAN：单细胞数据的无监督流形对齐。《生物信息学进展》，3(1)：vbad171，2023年1月。doi：10.1093/bioadv/vbad171。

Artur Szałata、Karin Hrovatin、Sören Becker、Alejandro Tejada-Lapuerta、Haotian Cui、Bo Wang 和 FabianJ Theis。《单细胞组学中的 Transformer：综述与新视角》。《自然·方法》，21(8):1430–1443，2024年。

Annika Vannan、Ruqian Lyu、Arianna L. Williams、Nicholas M. Negretti、Evan D. Mee、Joseph Hirsh、Samuel Hirsh、David S. Nichols、Carla L. Calvi、Chase J. Taylor、Vasiliy. V. Polosukhin、Ana PM Serezani、A. Scott McCall、Jason J. Gokey、Heejung Shim、Lorraine B. Ware、Matthew J. Bacchetta、Ciara M. Shaver、Timothy S. Blackwell、Rajat Walia、JenniferMS Sucre、JonathanA. Kropski、Davis J McCarthy 和 Nicholas E. Banovich。基于图像的空间转录组学揭示了与肺纤维化中远端肺重塑相关的分子生态位失调。*bioRxiv*，2023年。

Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio、Yoshua Bengio 等。图注意力网络。*stat*，1050(20):10–48550，2017年。

Eugene Vorontsov、Alican Bozkurt、Adam Casson、George Shaikovski、Michal Zelechowski、Kristen Severson、Eric Zimmermann、James Hall、Neil Tenenholtz、Nicolo Fusi、Ellen Yang、Philippe Mathieu、Alexander van Eck、Donghun Lee、Julian Viret、Eric Robert、Yi Kan Wang、Jeremy D Kunz、Matthew C H Lee、Jan H Bernhard、Ran A Godrich、Gerard Oakley、Ewan Millar、Matthew Hanna、Hannah Wen、Juan A Retamero、William A Moye、Razik Yousfi、Christopher Kanan、David S Klimstra、Brandon Rothrock、Siqi Liu 和 Thomas J Fuchs。用于临床级计算病理学和罕见癌症检测的基础模型。《自然·医学》，30(10):2924–2935，2024年10月。

Xiyue Wang、Sen Yang、Jun Zhang、Minghui Wang、Jing Zhang、Wei Yang、Junzhou Huang 和 Xiao Han。用于组织病理图像分类的基于Transformer的无监督对比学习。*Medical Image Analysis*，81:102559，2022年。

Zexuan Wang、Qipeng Zhan 和 Li Shen。用于单细胞数据整合的图正则化最优传输。*bioRxiv*，2024年。doi: 10.1101/2024.10.30.621072。URL https://www.biorxiv. org/content/early/2024/11/02/2024.10.30.621072。

Zhiqin Wang、Ruiqing Li、Minghui Wang 和 Ao Li。GPDBN：整合基因组数据与病理图像、用于乳腺癌预后预测的深度双线性网络。*Bioinformatics*，37(18):2963–2970，2021年9月。

Hongzhi Wen、Wenzhuo Tang、Xinnan Dai、Jiayuan Ding、Wei Jin、Yuying Xie 和 Jiliang Tang。Cellplm：超越单细胞的细胞语言模型预训练。*bioRxiv*，pp.2023–10，2023年。

Johnny Xi、Jana Osea、Zuheng Xu 和 Jason Hartford。非配对多模态数据的倾向评分对齐，2024年。网址：https://arxiv.org/abs/2404.01595。

Jinxi Xiang、Xiyue Wang、Xiaoming Zhang、Yinghua Xi、Feyisope Eweje、Yijiang Chen、Yuchen Li、Colin Bergstrom、Matthew Gopaulchan、Ted Kim、Kun-Hsing Yu、Sierra Willens、Francesca Maria Olguin、Jeffrey J Nirschl、Joel Neal、Maximilian Diehn、Sen Yang 和 Ruijiang Li。用于精准肿瘤学的视觉—语言基础模型。*Nature*，2025年1月。

Ronald Xie、Kuan Pang、Sai W. Chung、Catia T. Perciani、Sonya A. MacParland、Bo Wang 和 Gary D. Bader。通过双模态对比学习从 H&E 组织学图像预测空间分辨基因表达，2023年。网址：https://arxiv.org/abs/2306.01859。

Yingxue Xu 和 Hao Chen。基于多模态最优传输的协同注意力变换器与全局结构一致性，用于生存预测，2023年。网址：https://arxiv.org/abs/2306. 08330。

Karren Dai Yang、Anastasiya Belyaeva、Saradha Venkatachalapathy、Karthik Damodaran、Abigail Katcoff、Adityanarayanan Radhakrishnan、G V Shivashankar 和 Caroline Uhler。使用自编码器实现单细胞成像与测序数据之间的多域转换。*Nature Communications*，12(1):31，2021年1月。

Nadav Yayon、Veronika R. Kedlian、Lena Boehme、Chenqu Suo、Brianna T. Wachter、Rebecca T. Beuschel、Oren Amsalem、Krzysztof Polanski、Simon Koplev、Elizabeth Tuck、Emma Dann、Jolien Van Hulle、Shani Perera、Tom Putteman、Alexander V. Predeus、Monika Dabrowska、Laura Richardson、Catherine Tudor、Alexandra Y. Kreins、Justin Engelbert、Emily Stephenson、Vitalii Kleshchevnikov、Fabrizio De Rita、David Crossland、Marita Bosticardo、Francesca Pala、Elena Prigmore、Nana-Jane Chipampe、Martin Prete、Lijiang Fei、Ken To、Roger A. Barker、Xiaoling He、Filip Van Nieuwerburgh、Omer Ali Bayraktar、Minal Patel、E. Graham Davies、Muzlifah A. Haniffa、Virginie Uhlmann、Luigi D. Notarangelo、Ronald N. Germain、Andrea J. Radtke、John C. Marioni、Tom Taghon 和 Sarah A. Teichmann。映射至连续组织轴的人类胸腺细胞空间图谱。*Nature*，635(8039):708–718，2024年。

Jiahui Yu、Zirui Wang、Vijay Vasudevan、Legg Yeung、Mojtaba Seyedhosseini 和 Yonghui Wu。CoCa：对比式图像描述器是图文基础模型，2022年。URL https://arxiv. org/abs/2205.01917。

Sunyi Zheng、Xiaonan Cui、Yuxuan Sun、Jingxiong Li、Honglin Li、Yunlong Zhang、Pingyi Chen、Xueping Jing、Zhaoxiang Ye 和 Lin Yang。PathCLIP 病理图像分析基准测试，2024年。URL https://arxiv.org/abs/2401.02651。

Fengtao Zhou 和 Hao Chen。用于生存分析的跨模态翻译与对齐，2023年。URL https://arxiv.org/abs/2309.12855。

Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。使用循环一致性对抗网络进行非配对图像到图像翻译。载于*IEEE* 国际计算机视觉会议论文集，第 2223–2232 页，2017年。

A 数据与代码可用性

所有用于复现所报告结果的必要数据和代码将在论文发表后发布。

B 大语言模型的使用

我们使用大语言模型（LLM）协助编辑文本和调整图表格式。大语言模型仅用于提升内容的清晰度和呈现效果；所有科学内容、分析和解释均由作者完成。

C 局限性

CAMEO 基于这样一个假设：两种模态可以提供彼此不同但互补的 y info信息。这一假设带来了一些局限性：

- **表征质量**。 CAMEO 对齐通过自监督单模态编码器学习的表征。对于下游任务而言，其有意义地对齐这些表征的能力受单模态表征质量的限制。但这一限制同时也是一个机遇，因为模块化设计允许用户用改进后的模型替换编码器。

- **模态转换**。 互补性这一假设限制了对不共享共同概念的模态进行对齐。如果一种模态独立于另一种模态，例如，则转换结果会是随机的。

- **下游任务适用性**。 我们在计算病理学任务上测试了该方法，这些任务因其固有的多模态性质而被选中，其中组织学和基因表达数据都具有信息价值。该方法可能并不适用于每一项任务。具体而言，对于某一种模态几乎没有或完全没有带来额外收益的任务，该方法可能无法泛化。

- **模态特异性信息的损失**。 对齐过程可能会减少模态特异性信息，因为它优先考虑共享表示。未来的一个潜在方向是同时学习共享的潜在因子和模态特异性的潜在因子，在保留每种模态独有信息的同时，继续从它们的对齐中获益。

D 附加结果

本节提供了补充结果和可视化内容，以补充第 4.4 节中呈现的主要发现。

D.1 配对比例降低的影响

图5展示了随着配对数据比例降低，肺部数据集的相对性能变化，与图3中呈现的乳腺数据集结果相互补充。

D.2 单模态性能的预测性

图6将图4中的分析扩展至肺部数据集和胸腺数据集。较弱的相关性（Spearman ρ 值分别为 0.37 和 0.31）证实，乳腺数据集中观察到的模式（$\rho = 0.60$）在不同组织中具有一致性，但强度有所不同。

E 模型细节

E.1 图像基线

图像基线利用预训练编码器从组织病理图像中提取嵌入。这些编码器处于冻结状态，并在其基础上训练任务专用头。

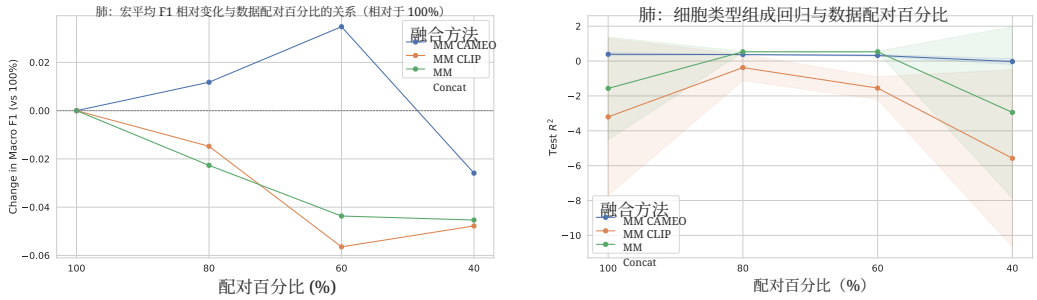


图5：肺部数据集：与100%配对相比，宏平均 F1 分数的相对变化（左）。肺部数据集与乳腺（图3）呈现相似的模式。用于细胞类型组成预测的绝对 R^2 分数的均值和标准差（右）。

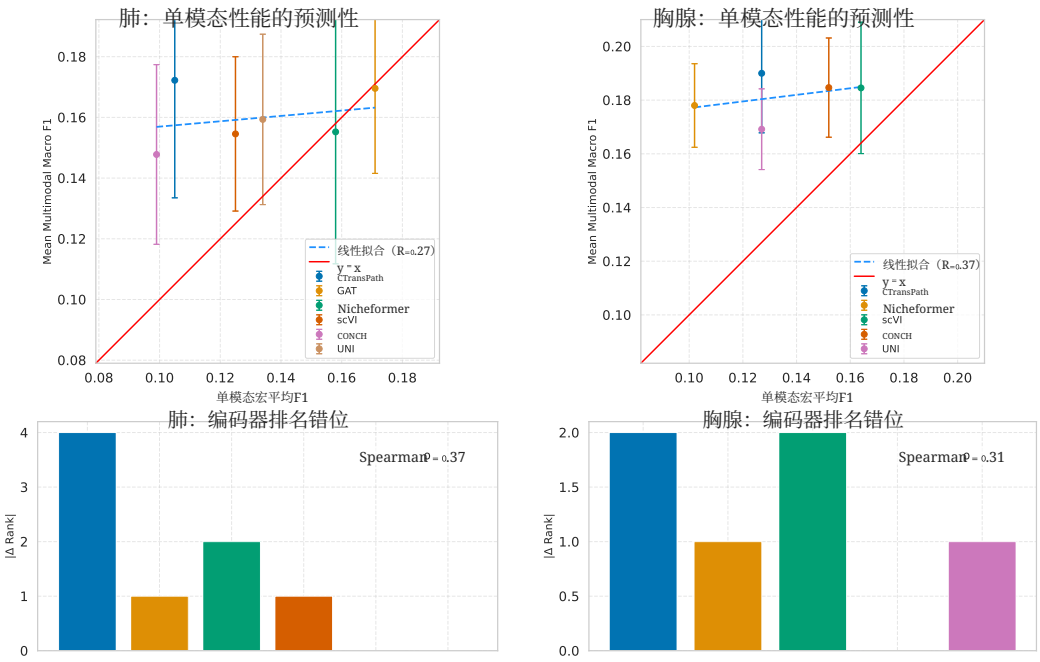


图6：肺部（左）和胸腺（右）数据集中单模态性能对多模态成功的预测性。两种组织均显示出较弱的相关性（分别为 $\rho = 0.37$ 和 $\rho = 0.31$），以及显著的排名错位，进一步支持了这一发现：单模态性能无法很好地预测多模态成功。

• **UNI**（Chen 等人，2024年）：UNI 基于 ViT-Large（ViT-L）架构构建，并使用从覆盖 20 种主要组织类型的 100,000 张苏木精和伊红（H&E）染色全切片图像（WSIs）中获取的超过 1 亿个图像块进行预训练。
• **CTransPath**（Wang 等人，2022年）：将卷积神经网络与多尺度 Swin Transformer 架构相结合。该模型在来自癌症基因组图谱（TCGA）和 PAIP数据集的1500万个未标注图像块上进行了预训练。

• **ResNet-50**（He 等人，2015年）：一种在 ImageNet-1k 上进行预训练的卷积神经网络。
• **Virchow**（Vorontsov 等人，2024年）：在310万张全切片组织病理学图像上进行预训练的自监督视觉Transformer。
• **CONCH**（Lu 等人，2024a）：一种通过在 117 万组组织病理学专用图像-标题对上进行对比学习训练的视觉语言模型。

E.2 基因表达基线

基因表达基线方法处理高维组学数据，其中每个样本都表示为一个由基因表达值构成的向量。

• **GAT**：图注意力网络（GAT）（Velickovic 等，2017年）通过利用注意力权重动态调整邻居节点的影响来学习节点表示。注意力机制会对邻居节点的特征进行加权聚合，其中注意力系数根据节点特征的相似性学习得到。这使模型能够关注更相关的邻居节点，因此 GAT 特别适用于图结构数据，例如空间转录组学测量数据。在此背景下，α_{uv} 表示节点 u 与节点 v之间的注意力权重，该权重根据两个节点的特征并结合图结构 G学习得到。虽然 GAT 通常以无监督方式训练，但研究表明，自监督学习（SSL）能够提升其性能，尤其是在应用于大型数据集时（Richter 等，2023年）。模型采用掩码建模进行预训练，其中基因表达谱 x_v 的部分内容会被掩码，目标是根据未被掩码的内容预测被掩码的条目。令$f_\theta(x_v)$表示单模态编码器，该编码器将基因表达向量 x_v 映射为嵌入 z_v，随后在融合模块中使用。训练目标是最小化掩码建模损失；对于一组被掩码的节点 $\mathcal{M}$，该损失为：

$$\mathcal{L}_{\text{SSL}} = \sum_{v \in \mathcal{M}} ||(1_p - M_v) \odot f_\theta(M_v \odot x_v) - \hat{x}_v||_2^2, \tag{1}$$

其中 $M_v \in \{0,1\}^p$ 是细胞 v 的基因掩码，而（$1_p - M_v$）确保只有被掩码的条目会参与损失函数计算。这种自监督学习预训练使模型无需标注数据，便能从大规模单模态数据集中学习有意义的表示。通过捕捉基因表达模式中的复杂空间依赖关系，所学习的嵌入为下游任务提供了坚实基础。

• **scVI**（Lopez 等，2018）：一种用于分析基因表达数据的概率框架。它采用变分自编码器（VAE）来学习基因表达谱的低维潜在表示。

• **scGPT**（Cui 等，2024a）：通过将基因视为词（标记），将细胞视为句子，对 GPT 架构进行改造，以处理基因表达数据。它在 3300 万个跨组织人类细胞上进行了预训练。

• **CellPLM**（Wen 等，2023）：一种显式编码细胞空间关系的空间转录组学基础模型。CellPLM 在 900 万个 scRNA-seq细胞和 200 万个空间转录组学细胞上进行训练。

• **Nicheformer**（Schaar 等，2024年）：一种专为空间单细胞组学设计的基于Transformer的基础模型。该模型在来自73种人类和小鼠组织的5,700万个解离细胞和5,300万个空间分辨细胞上进行了预训练。

E.2.1 多模态基线方法

• **CLIP**（Radford 等，2021年）：一种通过对比学习训练的多模态模型。每种模态都经过独立编码，并投影到模态特定潜在空间中。对比损失函数对齐配对样本，从而实现联合表示学习。

• **拼接**：将基因表达和图像编码器生成的嵌入进行拼接，不采用任何显式的对齐或融合机制。

E.3 CAMEO 架构

- **输入维度：**
  – 图像嵌入：$d_{\text{img}} = 1024$（UNI）或 768（CTransPath）或 512（CONCH）– 基因表达嵌入：$d_{\text{gex}} = 128$（GAT）或 128（scVI）或 512（Nicheformer）– 公共潜在空间：$d = \min(d_{img},\ d_{gex})$
- **生成器架构：**

- Encoder layers: $d \to d \to d/2$
- Decoder layers: $d/2 \to d \to d$
- Shared encoder layer: $d/2 \to d/4$
- Shared decoder layer: $d/4 \to d/2$
- All layers include BatchNorm1d and LeakyReLU ($\alpha = 0.2$)

• **判别器架构：**

– 输入层： $d \to d/4$– 隐藏层：

BatchNorm1d +LeakyReLU– 输出层：

$d/4 \to 1$

E.4 CAMEO训练配置

• 共享层使用 LeakyReLU 激活函数（$\alpha = 0.2$）
• 判别器中的批归一化
• 权重裁剪范围： $[-0.01, 0.01]$ ，适用于 Wasserstein生成对抗网络
• 优化器：RMSprop（Wasserstein生成对抗网络）或 AdamW（标准生成对抗网络）

E.5 训练协议

基因表达编码器预训练 GAT 模型使用掩码基因表达预测任务进行预训练：

• 掩码比例：0.4（40% 的表达值被掩码）
• 学习率：0.001
• 批大小：1（每个批次包含一个完整图块）
• 早停耐心值：200 个训练轮次
• 最大训练轮数：5000
• 梯度裁剪：1.0
• 损失函数：均方误差或负二项分布（当 predict_nb=True 时，负二项分布的效果更好）

scVI 训练 肺和乳腺数据集的 scVI 模型使用 scVI-tools 流程中的默认参数进行预训练。对于胸腺，模型按照作者的建议（Yayon 等，2024）进行预训练，并使用样本 ID 作为批次协变量。

CLIP 和 CAMEO 训练 对于 CLIP 和 CAMEO 模型，我们评估了 10 个训练轮次的预训练时长。针对 CAMEO，我们专门优化了 λ_{align} ，以平衡对齐损失项的贡献。

线性探测阶段 我们采用线性探测设置，在不更新编码器权重的情况下，单独评估预训练表示的质量。线性探测器使用 AdamW 优化器进行训练，学习率为 1e-3，共训练 10 个轮次。我们发现，该学习率在不同模型初始化和任务中均表现稳健。

CAMEO 特定参数

• λ_{align}：控制对齐损失的权重
• 生成器学习率：1e-4
• 判别器学习率：5e-5
• 权重裁剪值：0.01

F 评估指标

令 $y \in \{1, \ldots, C\}$ 为真实类别标签，$\hat{y}$ 为 N 个样本的预测类别标签。

宏平均 F1 分数。类别 c的 F1 分数定义为

$$\mathrm{F1}_c = \frac{2\,\mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c},$$

其中

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}.$$

宏平均 F1 分数对所有 C 类别进行等权平均:

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c.$$

宏平均 F1 将所有类别视为同等重要，并且对少数类别上的性能较为敏感。

微平均 F1 分数。 微平均 F1 分数首先汇总所有类别的 TP、FP 和 FN，然后使用这些汇总计数应用 F1 公式来计算精确率和召回率。微平均 F1 根据类别频率加权，反映模型的整体性能。

R^2 **分数**。 对于回归任务，决定系数定义为

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

其中， $\bar{y}$ 是 $\{y_i\}_{i=1}^{N}$的样本均值。 R^2 衡量模型所解释的方差比例。

G 概率对齐方法比较

We 在组织病理学-基因表达对齐的背景下比较两种主要的概率对齐方法： e
exp表达对齐:

DDMEC （Bounoua 等，2025 年）使用带有 KL 正则化“软边缘分布”的扩散模型，学习模态之间的随机条件分布。该方法提供了具有理论依据的不确定性建模，但训练扩散模型并从中进行采样需要大量计算资源。

scTopoGAN （Singh 等，2023 年）将拓扑保持自编码器与基于 GAN 的潜空间对齐相结合。该方法假设模态之间具有共享拓扑结构——这一假设适用于分子测量（例如 ATAC-seq、CITE-seq），但当不同模态在不同尺度上捕获本质不同的生物学特征时，其适用性较低。

对于组织病理学-基因表达对齐，关键的实际考量包括:

1. 大型空间斑块的计算效率 2. 模态之间不重合的流形
结构 3. 配对锚点的可用性，即使数量有限 4. 下游任务
对稳定嵌入的要求

这些限制使得 CAMEO 这类确定性方法更适用于这一特定用例，不过，在假设成立且计算资源允许的情况下，概率方法仍然很有价值。

H 计算资源

模型使用 Python 3.10 在一台 GPU 服务器上训练，服务器规格如下:

- GPU：30 个节点，每个节点配备 96 个中央处理器、768GB 内存和 4 个 NVIDIA H100 GPU，每个 GPU 配备 94GB HBM2 显存。

• GPU：5 个节点，每个节点配备 96 个中央处理器、1TB 内存和 4 个 NVIDIA A100 GPU，每个 GPU 配备 80GB HBM2 显存。
• GPU：4个节点，每个节点配备252个中央处理器、2 TB内存、8个NVIDIA A100 GPU，每个GPU配备80 GB HBM2 显存。

每个模型的训练通常耗时不到6小时，而在大多数情况下不到2小时。

I 数据集

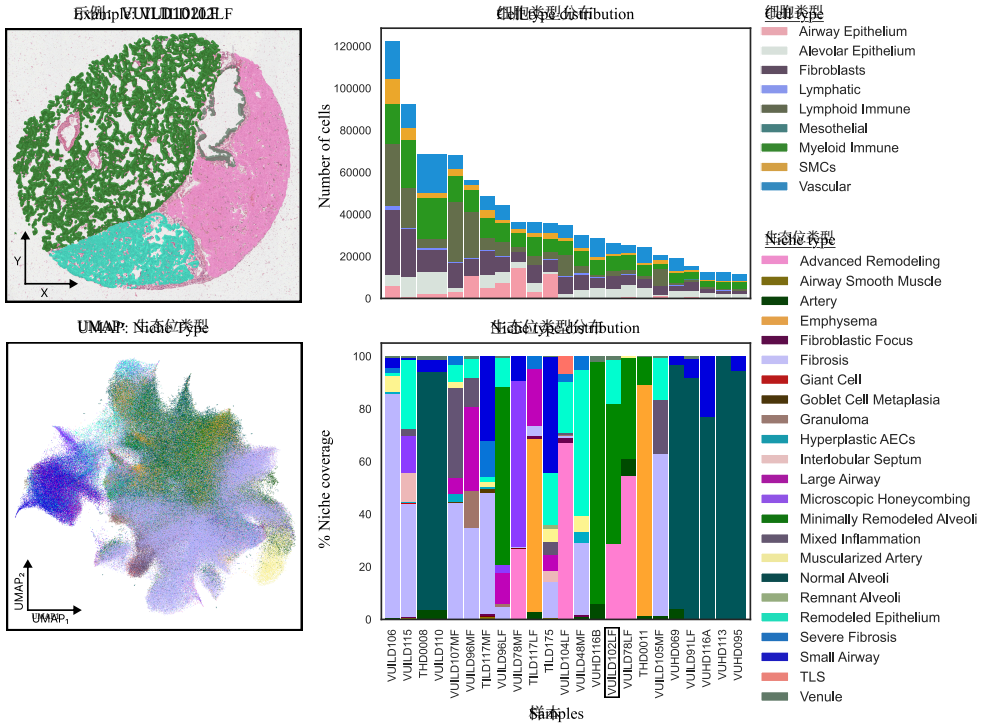


图7：LungPF Xenium 队列的元数据概览（Vannan 等，2023年），展示细胞类型分布、生态位标签分布以及各样本间的标签不平衡。左上：叠加专家提供的生态位注释的示例组织学图像；该样本对应的生态位标签分布显示在柱状图中（用方框标出样本名称以供参考）。右上：所有23个样本的总体细胞类型分布，显示细胞类型组成存在轻微差异。下方：图像块级基因表达谱的UMAP及总体生态位标签分布，揭示了23种独特注释生态位类型之间存在显著不平衡。标记为“NOANNOT”的微环境不包含在任何可视化结果中。

配对空间转录组学数据集仍然十分稀缺。空间数据集通常仅发布基因组学组分及其对应的空间背景，而配对组织学数据则因发布大型图像的成本过高而被省略。尽管这一情况正在逐步改善，但收集大规模配对数据集仍然颇具挑战。为了支持对部分配对空间数据的研究，我们创建了一个 HuggingFace 数据集，其中包含生态位层面的组织学图像和基因表达数据。生态位定义为组织学切片上的一个 224×224 像素裁剪区域，并与该区域内细胞对应的基因表达值配对。此外，每组数据还包含专家提供的生态位注释，以及由 Xenium 默认分割估算的每个点位的细胞总数。
为确保组织学数据与转录组数据之间正确的空间对齐，我们使用了 SpatialData（Marconato 等，2025年）。生态位在每张切片上以镶嵌方式铺排，彼此不重叠；我们还应用质量控制过滤器，排除组织占比不足 50% 的生态位（即主要包含背景的生态位）。为简化处理并便于解读，部分纳入的细胞所含转录本均按该生态位内的整细胞数据处理。

肺部数据集包含来自 19 位患者的 23 个 Xenium 样本，涵盖健康组织及不同严重程度的肺纤维化。专家病理学家在 H&E染色图像上标注了生态位区域，而原研究作者则提供了多个粒度层级的单细胞注释。本研究采用最宽泛的细胞类型标签。该数据集总计包含约 100 万个细胞，分布于 $\sim 71K$ 图像块中（表3）。图 8a 展示了训练集、验证集和测试集中的生态位标签分布，突显了标签不平衡以及肺部样本之间的分布偏移。

乳腺癌数据集由 10x Genomics（Xenium 技术的制造商）生成并提供。该数据集包含来自 4 名患者的 7 个 Xenium 样本，涵盖受癌症影响的乳腺组织。该数据集包含分布于 $\sim 126K$ 个图像块中的 ~ 2.3 百万个细胞。我们的合作病理学家使用 Elston-Ellis评分系统（ELSTON & ELLIS，1991年）提供了生态位级别的注释。图 8b 展示了乳腺队列中训练集、验证集和测试集划分的生态位标签分布。

胸腺数据集由来自 11 名供体的 19 个 Visium 样本组成，涵盖胎儿期（受孕后第 11–21 周）和儿科阶段（新生儿至 3 岁）的组织。该数据集旨在绘制出生前和出生后早期阶段的 T 细胞发育图谱。该数据集使用皮质-髓质轴（CMA）进行注释，这是一种由数据集原始作者建立的通用坐标框架。由于胸腺数据集是基于斑点的 Visium 数据集，因此其中包含来自空间点的转录组，而非单个细胞的转录组。这些空间点分布于 $\sim 40K$ 个图像块中。儿科组织样本用于训练，而胎儿样本则留作测试（表3）。图 8c 展示了胸腺队列中训练集、验证集和测试集划分的生态位标签分布。

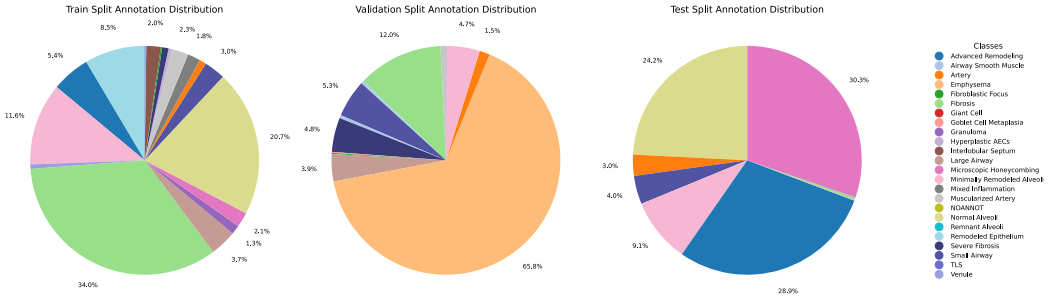
表3概述了全部三个队列的数据集构成，详细列出了每个样本的生态位数量、细胞类型、细胞计数、图像块数量以及数据划分。图7进一步展示了肺部队列中生态位类型及其对应细胞类型组成的分布，突显了数据集内部的异质性。

J 影响声明

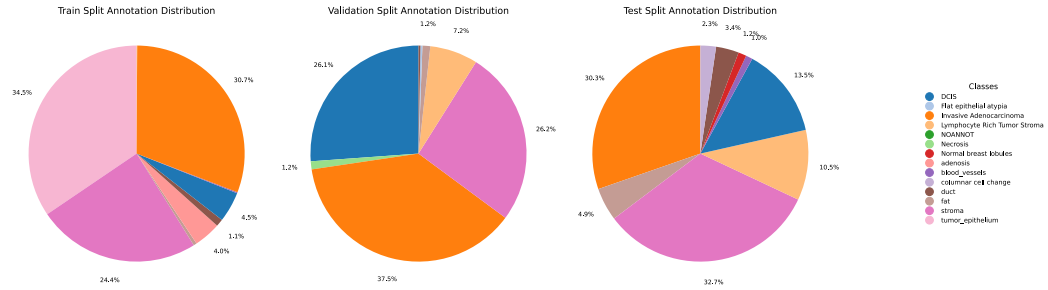
本研究推动了面向配对数据有限场景的多模态学习技术发展。通过发布经过标注的 HuggingFace 数据集和开源实现，我们希望促进多模态研究的可复现性。虽然我们的框架是为生物医学应用开发的，但其底层技术也有望惠及其他数据配对成本高昂或难以实现的领域。鉴于本研究涉及医疗领域，且患者来源的分子和组织学数据具有敏感性，应用时必须谨慎考虑数据隐私、知情同意和临床验证要求。多种数据模态的整合会进一步提高患者信息的可识别性，因此必须负责任地处理数据。

组织	患者	样本	#微环境	#细胞类型	#细胞	#图像块	数据划分
Lung	THD0008	THD0008	5	9	117，612	4，316	训练集
Lung	THD0011	THD0011	5	9	53，544	2，099	Val
Lung	TILD117	TILD117MF	11	9	86，728	3，831	Val
Lung	TILD117	TILD117LF	8	9	65，423	2，800	Val
Lung	TILD175	TILD175	11	10	59，824	2，489	训练集
Lung	VUHD069	VUHD069	4	9	36，862	1，673	Test
Lung	VUHD095	VUHD095	4	9	19，631	808	训练集
Lung	VUHD113	VUHD113	2	9	20，850	1，281	训练集
Lung	VUHD116	VUHD116B	4	9	48，213	1，898	训练集
Lung	VUHD116	VUHD116A	3	9	19，005	926	训练集
Lung	VUILD102	VUILD102LF	5	10	45，545	2，616	训练集
Lung	VUILD104	VUILD104LF	11	9	63，792	2，593	训练集
Lung	VUILD105	VUILD105MF	6	9	39，419	1，779	训练集
Lung	VUILD106	VUILD106	11	10	235，514	9，923	训练集
Lung	VUILD107	VUILD107MF	11	9	115，596	4，392	训练集
Lung	VUILD110	VUILD110	5	9	117，612	4，316	训练集
Lung	VUILD115	VUILD115	13	10	176，255	7，305	训练集
Lung	VUILD48	VUILD48MF	8	9	52，956	2，352	训练集
Lung	VUILD78	VUILD78LF	5	9	55，614	2，227	Test
Lung	VUILD78	VUILD78MF	7	10	73，135	2，722	Test
Lung	VUILD91	VUILD91LF	5	10	23，427	1，450	训练集
Lung	VUILD96	VUILD96LF	8	9	76，223	3，479	训练集
Lung	VUILD96	VUILD96MF	7	9	95，241	4，034	训练集
乳腺	患者1	Rep1	9	12	261，661	11，122	Val
乳腺	患者1	Rep2	8	12	170，749	7，381	Val
乳腺	患者2	样本2	10	16	221，766	9，101	Test
乳腺	患者3	TENX94	7	*	545,980	18,042	训练集
乳腺	患者3	TENX96	7	*	557,757	18,277	训练集
乳腺	患者4	TENX95	7	*	917,241	31,154	训练集
乳腺	患者4	TENX97	7	*	919,345	31,693	训练集
胸腺	Z1	TA11486161	11	*	*	1，445	训练集
胸腺	Z2	TA11556492	11	*	*	3，045	训练集
胸腺	Z2	TA11486164	11	*	*	2，827	训练集
胸腺	Z4	TA11486162	11	*	*	3，210	训练集
胸腺	Z4	TA11556493	11	*	*	2，963	训练集
胸腺	Z4	TA11556494	11	*	*	2，481	训练集
胸腺	Z6	TA11556496	11	*	*	2，295	训练集
胸腺	Z6	TA11556495	11	*	*	2，764	训练集
胸腺	F113	WSSS_F_IMMsp11604689	10	*	*	311	Val
胸腺	F114	WSSS_F_IMMsp11604687	11	*	*	1，264	Test
胸腺	F114	WSSS_F_IMMsp11604688	11	*	*	1，162	Test
胸腺	F130	WSSS_F_IMMsp11604686	11	*	*	2，121	Test
胸腺	F130	WSSS_F_IMMsp11604685	11	*	*	1，846	Test
胸腺	F133	WSSS_F_IMMsp11765867	11	*	*	1，606	Val
胸腺	F135	WSSS_F_IMMsp11765870	11	*	*	2，570	Val
胸腺	U09	WSSS THYst9518033_	11	*	*	2，125	训练集
胸腺	U11	WSSS THYst9518030_	11	*	*	2，384	训练集
胸腺	U11	WSSS THYst9142088_	11	*	*	2，035	训练集
胸腺	U11	WSSS THYst9142089_	11	*	*	2，190	训练集

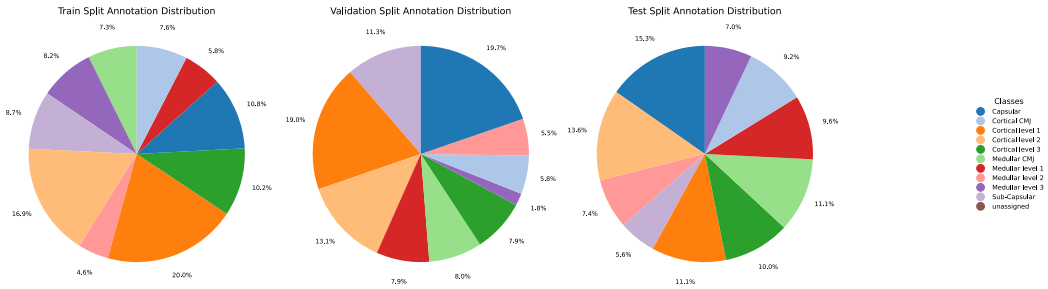
表3：肺、乳腺和胸腺数据集中的样本数、患者数、唯一生态位注释数（包括 NOANNOT）、唯一细胞类型数（包括 NOMAP）、细胞计数、图像块数量以及数据划分（训练集/验证集/测试集）汇总。单个细胞与多个图像块相交时，可能会被重复计数。对于胸腺数据集，由于其基于斑点而非基于细胞，细胞计数和细胞类型分类不适用，以 * 表示。同样，某些乳腺样本缺少细胞类型注释；在这些情况下，唯一细胞类型数也标记为 *。



(a) 肺部队列中基于患者分层划分的训练集、验证集和测试集划分的标签分布。可视化中不包括 NOANNOT 类别。各数据划分之间观察到较大的分布偏移。



(b) 乳腺队列中基于患者分层划分的训练集、验证集和测试集划分的标签分布。可视化中不包括 NOANNOT 类别。各数据划分之间观察到中等程度的分布偏移。



(c) 胸腺队列中基于患者分层划分的训练集、验证集和测试集划分的标签分布。可视化中不包括 NOANNOT 类别。各数据划分之间观察到较小的分布偏移。

图 8：肺部、乳腺和胸腺队列在训练集、验证集和测试集划分中的标签分布（按患者分层）。不包括 NOANNOT 类别。