OmicsCL: 用于癌症亚型发现和生存分层的无监督对比学习

Atahan Karagöz
Department of Computer Science
University of Basel
Basel, Switzerland
atahan.karagoez@stud.unibas.ch

摘要—无监督学习从多组学数据中学习疾病亚型为促进个性化医疗提供了重要机会。我们引入了组学 *CL*，这是一个模块化的对比学习框架，将异质性组学模态（如基因表达、DNA 甲基化和 miRNA 表达）联合嵌入到统一的潜在空间中。我们的方法采用了一种生存感知对比损失函数，鼓励模型学习与生存相关模式对齐的表示形式，而无需依赖标注结果。在 TCGA BRCA 数据集上进行评估时，组学 *CL* 发现了具有临床意义的聚类，并且在患者生存方面实现了强大的无监督一致性。该框架展示了在超参数配置上的鲁棒性，并可以调整以优先考虑亚型一致性和生存分层。消融研究表明，集成生存感知损失显著增强了学习嵌入的预测能力。这些结果突显了对比目标在高维、异质性组学数据中发现生物学洞察力的潜力。

***Index Terms*—多组学、对比学习、癌症亚型、生存分析、无监督学习**

I. 介绍

癌症是一种异质性疾病，通过基因组、表观基因组和转录组等不同生物学水平上的复杂分子改变表现出来。高通量测序技术的出现使研究人员能够收集多组学数据集，这些数据集共同提供了个体肿瘤的全面分子视图。整合这些异质性数据类型对于揭示潜在亚型和基于预后对患者进行分层至关重要。然而，由于各组学模式之间的高维度、噪声和不一致性，有效的多组学整合仍是一项具有挑战性的任务。

传统亚型发现方法依赖于使用预定义的癌症亚型标签进行监督学习，或者使用具有有限生物学解释能力的无监督聚类技术。最近，深度学习在学习组学数据的紧凑表示方面显示出潜力，然而许多模型要么是监督式的，要么需要复杂的架构和大量的注释数据。此外，那些不明确纳入生存结果的模型可能无法捕捉到亚型之间的临床相关差异。

为了解决这些挑战，我们提出**组学 CL**，一个无监督对比学习框架，旨在从没有亚型标签的多组学数据中学习联合嵌入。组学 *CL* 融合了跨组学特定编码器的对比目标，并结合了一种新颖的生存意识对比损失，鼓励具有相似生存结果的患者的嵌入在潜在空间中更接近。这种设计使模型能够在完全无监督的方式下学习生物上有意义且包含生存信息的表示。

II. 相关工作

多组学数据的整合在过去的十年中已被广泛研究，用于癌症亚型的发现。传统方法如相似性网络融合（SNF）[1] 和 iCluster [2] 结合异质性数据源以构建统一表示，通常随后进行无监督聚类。虽然有效，但这些方法依赖于预定义的相似性度量，并不直接优化下游生存相关性。

随着深度学习的兴起，基于自动编码器的方法已成为从高维组学数据中学习低维表示的流行方法。模型如 MOFA [3] 和 DCCA [4] 利用概率或典型相关性基础框架来联合嵌入多种模态。然而，这些方法通常假设配对数据分布，并且并非专门为生存意识表示学习而设计。

对比学习最近作为表示学习中的一种强大无监督方法，在计算机视觉和生物信息学等多个领域崭露头角。诸如 scCL [5] 和 CONAN [6] 等方法将对比目标应用于单细胞或多组学设置。这些方法大多数专注于学习模态不变特征或最大化数据视图之间的协议，但它们往往忽视了临床结果如患者生存时间的时间方面。

生存分析在深度学习中通常通过监督模型如 DeepSurv [7] 或 DeepHit [8] 来解决，这些模型需要标注的事件时间并通常直接预测生存函数。虽然这些模型已经取得了强大的性能，但它们需要大量的标注数据，并且不适合用于无监督分层任务。最近的大规模基准测试 [9] 强调了在多组学设置下监督生存模型的局限性，强化了对更灵活、无监督替代方案的需求。

我们的工作弥合了无监督表示学习与生存分析之间的差距。与之前模型不同，组学 *CL* 引入了一种生存感知对比损失，该损失在不依赖显式的生存监督或预定义亚型标签的情况下，将时间结果信息编码到嵌入空间中。这允许在完全无标签的设置下发现具有临床意义的癌症亚型，同时仍然保留用于患者分层的判别特征。

III. 方法论

A. 问题定义

令 $\mathcal{D} = \{(\boldsymbol{x}_i^{(g)}, \boldsymbol{x}_i^{(m)}, \boldsymbol{x}_i^{(r)}, t_i, e_i)\}_{i=1}^{N}$ 表示一个包含 N 名患者的多组学数据集，其中 $\boldsymbol{x}_i^{(g)}$、$\boldsymbol{x}_i^{(m)}$ 和 $\boldsymbol{x}_i^{(r)}$ 分别对应基因表达、DNA 甲基化和 microRNA 图谱。每位患者还与生存时间 $t_i \in \mathbb{R}^+$ 和一个事件指示器 $e_i \in \{0, 1\}$ 相关联，其中 1 表示死亡，0 表示删失。我们的目标是为每种组学模态学习紧凑的嵌入，并获得一种联合表示形式，使患者能够被有意义地聚类成预测生存结果的亚型，在训练过程中无需依赖亚型标签。

B. 模型架构

作为一个模块化的对比学习框架，组学 *CL* 学习了组学模态之间的视图特定和联合表示。对于每个模态，我们使用了一个由具有共享结构但独立权重的神经网络参数化的专用编码器 $f_\theta^{(v)}$。每个编码器都包含一个多层感知机，带有批归一化和 ReLU 激活函数，并跟随一个投影头，将特征映射到潜在空间 $\mathbb{R}^d$ 中，其中 d 是嵌入维度。这些投影被 ℓ_2 规范化，位于单位超球面上。

C. 组学间的对比目标

给定一批样本的小批量，我们为来自不同模态 $v \neq w$ 的每个 i 构造正对 $(z_i^{(v)}, z_i^{(w)})$，并为 $j \neq i$ 构造负对 $(z_i^{(v)}, z_j^{(w)})$。我们采用归一化温度缩放交叉熵损失（NT-Xent）[10]：

$$\mathcal{L}_{\text{NT-Xent}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^{(v)}, z_i^{(w)})/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{[j \neq i]} \exp(\text{sim}(z_i^{(v)}, z_j^{(w)})/\tau)}, \tag{1}$$

其中 $\text{sim}(a, b)$ 表示余弦相似度，τ 是一个温度参数。此目标鼓励来自同一患者不同组学视图的表示之间的协议。

D. 生存意识对比损失函数

为了将时间风险结构编码到嵌入空间中，我们引入了一种新颖的无监督方法*生存对比损失*。它惩罚具有不同生存时间（当两者均已去世时）患者的表示，并鼓励具有相似结果的嵌入之间的接近性。设 d_{ij} 为嵌入 z_i 和 z_j 之间的欧氏距离，$\Delta t_{ij} = |t_i - t_j|$ 为它们的生存时间差。损失定义为：

$$\begin{aligned}\mathcal{L}_{\text{surv}} = &\lambda_{\text{pull}} \cdot \mathbb{E}_{i,j}\left[\mathbb{1}_{[e_i=e_j=1]} \cdot \mathbb{1}_{[\Delta t_{ij}<\delta]} \cdot d_{ij}^2\right] \\ &+ \lambda_{\text{push}} \cdot \mathbb{E}_{i,j}\left[\mathbb{1}_{[\Delta t_{ij}\geq\delta]} \cdot \max(0, \delta - d_{ij})^2\right],\end{aligned} \tag{2}$$

其中，δ 是一个可调的时间裕量，$\lambda_{\text{pull}}, \lambda_{\text{push}}$ 是加权系数。值得注意的是，这种表述不依赖于监督风险标签，并在完全无监督的环境中运行，使其能够在不同癌症类型和数据分割中泛化。

E. 联合训练

最终训练目标是对比损失和生存感知正则化项的加权和：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NT-Xent}} + \alpha \cdot \mathcal{L}_{\text{surv}}, \tag{3}$$

其中 α 控制模态对齐和生存分层之间的权衡。在训练过程中，我们使用循环学习率调度器，并基于验证一致指数（C 指数）进行早期停止，该指数评估所学嵌入捕捉生存风险的程度。

F. 聚类与评估

训练后，我们将跨组学模态学习到的嵌入连接起来形成统一的患者表示。这些表示使用 KMeans 进行聚类。评估使用聚类和生存指标进行，我们将在第 IV-D 节中详细说明。

IV. 实验

A. 数据集

我们对来自多组学癌症基准测试 [11] 的 TCGA-BRCA 数据集进行了评估。该数据集包括三种主要的组学视图：基因表达（RNA-seq）、DNA 甲基化（450k array）和 miRNA 表达。为每位患者提供了生存信息，包括事件发生时间或删失时间和一个事件指示器。此外，还提供了 PAM50 分型注释用于评估聚类性能。

B. 数据预处理

我们将样本标识符在不同组学来源之间进行了统一，并移除了那些生存时间或事件状态缺失的患者。生存时间是从如 “overall_survival” 这样的临床字段中提取的，相应的二元死亡指标则来自于 “status”；这两者都被一致编码。具有缺失值的亚型标签被补全为 “Unknown”，并且在监督评估指标中予以排除。所有的预处理脚本作为我们流程的一部分发布以确保可重复性。

预处理后，数据集共包含 612 名患者，且所有三种组学模式和生存标签都可用。我们对每个组学视图应用了 z 分数标准化，并使用固定随机种子将数据集分割为 60%的训练集、20%的验证集和 20%的测试集以确保可重复性。

C. 实现细节

每个组学编码器都是一个两层 MLP，隐藏维度为 128，投影维度为 64。所有嵌入都被 ℓ_2 规范化。模型使用带有权重衰减 1×10^{-6} 的 Adam 优化器进行训练，并采用从 1×10^{-5} 到 1×10^{-3} 的循环学习率策略。NT-Xent 损失温度 τ 设置为 0.1。生存对比损失使用了 1.0 的边缘，并且基于网格搜索将权重系数 α 设置为 10.0。

训练最多进行了 1000 个周期，基于验证一致性指数提前停止，耐心为 20 个周期。训练过程中捕获的时间动态在图 1 中可视化。每次训练运行都设置了随机种子以保证可重复性。所有实验都在配备 Apple M1 Max CPU 和 64GB 内存的机器上运行。

D. 评估指标

为了评估学习到的表示的质量，我们评估了预测簇的聚类一致性和生存相关性。

a) 聚类度量标准： 为了评估预测的聚类与已知的 PAM50 亚型之间的一致性，我们报告了几种无监督聚类指标。*轮廓系数*测量样本在聚类内的凝聚度和分离度。*纯度*反映了基于每个聚类中的多数投票正确分配的样本比例。*调整兰德指数*（*ARI*）量化预测标签与真实标签之间的相似性，调整了偶然因素的影响。*归一化互信息*（*NMI*）测量聚类分配与地面实况亚型之间共享的信息。

b) 生存指标： 为了评估聚类对患者生存期分层的能力，我们使用特定的生存度量。*一致指数*（*C 指数*）评估预测风险评分与实际生存时间之间的一致性。*对数秩检验*提供一种统计度量，用于衡量跨聚类的生存分离程度。最后，*Kaplan-Meier 曲线*可视化每个预测聚类随时间变化的生存概率。

E. 基线

由于我们的主要目标是保持无监督，我们将方法的性能与在同一特征上训练的 Cox 比例风险（CoxPH）模型进行了比较。然而，需要注意的是，CoxPH 直接优化监督生存预测，而*组学聚类分析法*在没有访问标签的情况下推断出与生存相关的嵌入。

我们还将*组学 CL* 的版本与未使用生存感知对比正则化的训练版本进行了比较，展示了我们的设计选择对下游生存分析的影响。

图 1. 训练时期的综合生存直方图。此可视化图表捕捉了在训练过程中删失和死亡事件分布的演变，突出了模型编码到嵌入空间中的时间动态。

V. 结果

A. 生存分层性能

组学 CL 在生存预测中表现出强劲性能，这体现在测试集上的一致性指数（C 指数）为 **0.7512**。这一结果表明，尽管训练过程中没有标签监督，模型所学习的无监督嵌入能够有效地捕捉到患者群体中的风险相关结构。

Kaplan – Meier 曲线在图 2 中显示了由 KMeans 预测的聚类之间存在明显分离，支持了学习到的表示保留了临床上相关的生存差异这一假设。此外，多变量对数秩检验得到了一个 p 值为 0.0082 的结果，表明预测的聚类之间的生存分布具有统计学上的显著差异。

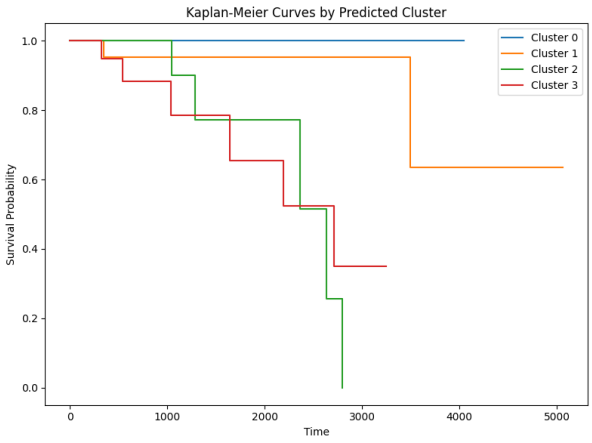

图 2. Kaplan – Meier 曲线按预测的聚类分层。观察到明显的生存分离，特别是在聚类 0 和聚类 2 之间。

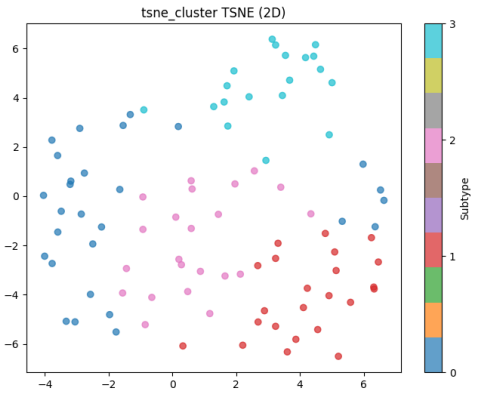

图 3. 嵌入的二维 t-SNE 可视化，按预测的聚类着色。尽管是无监督训练，但仍出现了不同的亚群。

B. 子类型发现和聚类质量

表 I 总结了聚类指标。组学 *CL* 在不使用亚型标签的情况下达到了 0.4022 的纯度。尽管 ARI 和 NMI 较低——反映出与 PAM50 的弱对齐——这是由于标签噪声和无监督设置所预期的。正如子节 V-F 中进一步讨论的，替代配置可以显著提高聚类性能。

表 I
使用 KMeans 在测试集上进行聚类评估指标 $k = 4$

度量	得分
Silhouette Score	0.0705
Accuracy	0.0000
Adjusted Rand Index (ARI)	-0.0013
Normalized Mutual Info (NMI)	0.0672
Purity	0.4022

C. 学习到的嵌入可视化

为了定性评估学习到的表示，我们将嵌入通过 UMAP 和 t-SNE 投影到了二维和三维空间。图 3 和 4 展示了潜在空间中的有意义结构。虽然聚类并未与 PAM50 标签完全对齐，但视觉上的分离表明模型捕获了替代的生物子结构或临床上相关的特征。为了增强探索性，我们还生成了 HTML 格式的交互式三维图，提供了更动态的簇几何视图。

D. 消融研究：生存对比损失的影响

我们进行了一项消融研究，以量化生存感知正则化的效果。从训练目标中移除生存对比损失项导致 C 指数显著下降至 0.617。这验证了我们的假设，即直接将时间生存动态纳入对比损失可以提高学习嵌入的生存区分能力。

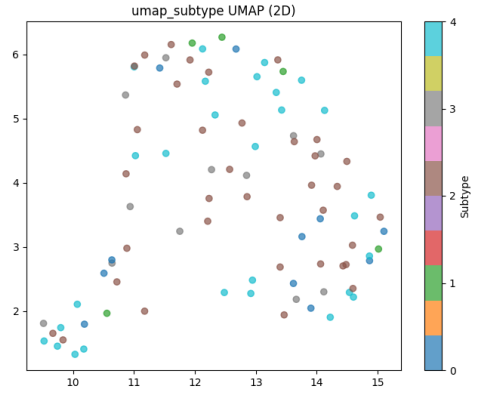

图 4. 基于 PAM50 亚型着色的嵌入 2D UMAP 可视化。模型在无监督的情况下部分恢复了亚型结构。

E. 与 *Cox* 比例风险模型的比较

为了提供一个基线比较，我们评估了在相同嵌入上训练的 Cox 比例风险（CoxPH）模型的生存分层性能。我们评估了不同聚类配置下的一致指数（C 指数），得分范围从 0.4570（2 个集群）到 0.7541（9 个集群）。组学 *CL* 在仅使用 4 个集群的情况下达到了 0.7512 的 C 指数，在低于 9 个集群的所有配置中始终优于 CoxPH。尽管 CoxPH 在 9 个集群时达到了稍高的峰值，但我们的方法展示了无需依赖监督生存建模的竞争力性能，突显了组学 *CL* 在从多组学数据捕捉与生存相关的信息方面的有效性。

F. 可在生存率和亚型指标之间进行配置的权衡

虽然我们的主要配置优化了生存分层，从而达到了 0.7512 的高 C 指数，组学 *CL* 也展示了适应其他目标的能力。具体来说，通过调整如 KMeans 中的聚类数量（k）

和嵌入维度等超参数，我们观察到了与亚型相关的聚类指标的改进。

例如，将聚类的数量增加到 $k = 9$ 导致纯度分数显著提高到了 **0.5217**，这表明与已知的 PAM50 子类型有更好的对齐。然而，这种配置产生了较低的 C 指数，突显了在无监督多组学表示学习中生物子类型的发现和生存差异之间的内在权衡。

这种可配置性表明*组学 CL* 并不是严格绑定到单一目标，而是可以根据下游应用的需求调整以优先考虑特定的临床或生物学目标。

VI. 讨论

组学 CL 在不同配置中的表现揭示了无监督多组学学习的几个重要特征。首先，虽然生存意识对比损失显然增强了患者结局的分层，但它可能会抑制潜在空间中亚型特异性结构。这表明某些与生存相关的模式可能跨越已知的亚型边界或捕捉正交生物信号。

我们还探索了几种改进生存感知表示学习的增强方法，包括无监督边界调度、多视图一致性惩罚和时间感知难样本挖掘。然而，这些修改并没有持续提高性能，在某些情况下，引入了噪声到学习动态中。有趣的是，通过 $\tanh(t_i - t_j)$ 稳定化的时间相似性加权被证明是有益的，将 C 指数推近到了中级配置下的 0.73 范围。

值得注意的是，*组学 CL* 的架构故意保持简单——每个组学模态都有一个轻量级 MLP 编码器，并且具有共享的对比目标。尽管如此简单，它在无监督生存建模中仍能胜过许多更为复杂的方案。这表明有意义地整合和对齐组学视角，结合原理性的目标（如对比损失和生存感知损失），可以生成架构负担极小的强大模型。

这些发现突出了对比目标在多组学设置中的复杂行为，其中对一个生物轴的优化可能会掩盖其他轴。在各种配置中观察到的表现强调了平衡任务特定目标与方法论简洁性的必要性，凸显了对比框架作为无监督生物医学表示学习的有前景工具。深入了解嵌入空间内生物信号如何相互作用仍然是推进可解释且临床稳健模型的关键。

VII. 限制和未来工作

尽管*组学聚类分析*的结果很有前景，但仍有一些局限性需要考虑。首先，虽然我们的方法在生存预测方面表现出色，但在揭示生物学上有意义的亚型方面的有效性仍然取决于特定的超参数配置。观察到的基于生存的聚类与亚型纯度之间的权衡表明，没有单一的配置能够最优地平衡所有下游目标。这突显了需要更多原则性的多目标优化策略或针对生物医学任务量身定制的模型选择标准。

其次，当前的模型架构基于每种组学模态独立的编码器，然后在表示层进行平均融合。虽然有效，但这种简单的晚期融合可能无法捕捉到跨模态的高阶相互依赖关系。未来的工作可以研究可学习的注意力机制融合、跨模态转换器或共享编码层以实现更丰富的组学特异性信号整合。

第三，尽管模型是完全以无监督的方式训练的，它仍然通过生存感知对比损失间接依赖于生存时间和审查标签。虽然这不构成有监督的亚型学习，但它从生存结果中引入了弱监督。探索完全标签无关的训练方案或自监督预训练任务可以将该框架的通用性扩展到更嘈杂或注释较少的数据集。

最后，本研究专注于单一癌症类型（TCGA BRCA），这可能限制了其普遍适用性。将*组学 CL* 应用于具有不同组学特征和生存模式的其他队列对于验证其稳健性和广泛应用至关重要。此外，将临床变量或影像数据整合到对比训练过程中仍然是构建更全面患者模型的一个开放方向。

在未来迭代中，我们还旨在探索直接针对一致性指数优化的可微生存损失代理，并结合未标记数据与稀疏亚型注释的*组学 CL* 半监督扩展。

VIII. 结论

组学 CL 提供了一种灵活且有效的方法，用于在不依赖亚型标签的情况下揭示多组学癌症数据中临床上相关的结构。通过将多视图表示学习与生存感知对比正则化器相结合，我们的方法有效地将基因表达、DNA 甲基化和 miRNA 图谱整合到统一的嵌入中，捕捉分子相似性和生存异质性。

通过对 TCGA BRCA 数据集的广泛实验，我们证明*组学 CL* 达到了 0.7512 的强大无监督一致指数，并且 Kaplan-Meier 生存曲线在统计学上有显著分离（log-rank$p = 0.0082$）。这些结果验证了我们的方法在无标签设置下学习预后信息表示的能力。

此外，我们强调了我们的管道的灵活性：通过调整配置参数，如聚类数量、嵌入维度或生存损失权重，*组学 CL* 可以被调优以强调不同的评估标准，例如亚型纯度或轮廓系数。这种适应性在生物医学环境中尤其有价值，在这些环境中，不同的下游应用可能优先考虑可解释性、预后或亚型发现。

总体而言，*组学 CL* 为促进在高维生物数据中进行无监督发现的方法体系做出了贡献。其模块化结构、对监督的极小依赖以及强大的实证性能表明，它可以作为未来模型的基础，用于处理更复杂、异质性和临床细微的数据集。

参考文献

[1] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014. [Online]. Available: https://www.nature.com/articles/nmeth.2810

[2] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009. [Online]. Available: https://academic.oup.com/bioinformatics/article/25/22/2906/180866

[3] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, p. e8124, 2018. [Online]. Available: https://www.embopress.org/doi/10.15252/msb.20178124

[4] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, vol. 28, no. 3. PMLR, 2013, pp. 1247–1255. [Online]. Available: https://proceedings.mlr.press/v28/andrew13.html

[5] L. Du, R. Han, B. Liu, Y. Wang, and J. Li, "Scccl: Single-cell data clustering based on self-supervised contrastive learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 3, pp. 2233–2241, 2023. [Online]. Available: https://doi.org/10.1109/TCBB.2023.3241129

[6] G. Ke, Z. Hong, Z. Zeng, Z. Liu, Y. Sun, and Y. Xie, "Conan: Contrastive fusion networks for multi-view clustering," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 653–660. [Online]. Available: https://doi.org/10.1109/BigData52589.2021.9671851

[7] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, 2018. [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1

[8] C. Lee, W. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 2314–2321. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11842

[9] M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix, "Large-scale benchmark study of survival prediction methods using multi-omics data," *Briefings in Bioinformatics*, vol. 22, no. 3, p. bbaa167, 2021. [Online]. Available: https://academic.oup.com/bib/article/22/3/bbaa167/5893227

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[11] D. Leng, L. Zheng, Y. Wen, Y. Zhang, L. Wu, J. Wang, M. Wang, Z. Zhang, S. He, and X. Bo, "A benchmark study of deep learning-based multi-omics data fusion methods for cancer," *Genome Biology*, vol. 23, no. 1, p. 171, 2022. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02739-2