癌症多组学联合降维带来可靠且可解释的生存预测

Vincent Le Goff[1, *]、Vincent Guillemot[2]、Cathy Philippe[3]、Wiam Ahdidou[1]、Gwendoline Mendes[1]、Jean-François Deleuze[1,4]、Edith Le Floch[1] 和 Arnaud Gloaguen[1] 1 法国国家人类基因组研究中心（CNRGH）、法国原子能委员会、巴黎萨克雷大学，埃夫里-库尔库龙 2 巴斯德研究所、巴黎西岱大学、生物信息学与生物统计学中心，巴黎 3 Neurospin、法国原子能委员会、巴黎萨克雷大学，吉夫叙尔伊韦特 4 人类多态性研究中心（CEPH）让·多塞基金会，巴黎，法国 * vincent.legoff@cnrgh.fr

摘要 多组学研究正变得越来越普遍，尤其适合解析癌症中发生的失调的复杂性。在此背景下，通过可解释的方法改进预后评估，有望实现更好的患者分层，并帮助识别新的预后生物标志物和治疗靶点。然而，多组学数据通常具有高维性（患者数量少、变量数量多）。应对高维性的一种可能方案是联合降维（JDR）方法，该方法根据多组学数据计算低维潜在空间。在本研究中，我们在18个不同的TCGA癌症数据集上对12种JDR方法进行了基准测试。生存预测通过两步过程实现：首先估计低维多组学潜在空间，随后在有或无额外临床数据的情况下进行生存预测。我们为其中4种方法赋予了预测新患者生存情况的能力。结果表明，JDR方法能够借助临床数据改善生存预测，并达到与单独使用临床数据相当的预测性能，但不同癌症之间的差异很大。在JDR方法中，JIVE和RGCCA表现出更优的性能。在后续的自助法分析中，JDR方法识别出跨癌症具有可靠性的预后生物标志物。我们提出了在低级别胶质瘤中对这类标志物进行具有临床相关性的解释。可复现代码见 https://github.com/vlegoff/MultiOBenchmark。

1 引言

测序成本的降低促进了在同一队列中获取多种组学模态，从而产生了多组学数据集。这类数据集有助于更好地理解疾病，因为它们能够识别在多种组学中失调的过程。在癌症研究领域，癌症基因组图谱（TCGA）（https://www.cancer.gov/tcga）已为数千份样本提供了涵盖多个癌症类型的多组学数据，包括基因组学（体细胞突变、拷贝数变异）、转录组学（基因表达、miRNA表达）、表观基因组学（DNA甲基化）和蛋白质组学信息。因此，TCGA是进一步加深对癌症理解的宝贵资源。许多癌症研究关注的主要结局通常是总生存期或无进展生存期。TCGA等资源以及其他多组学队列有助于提高此类预测的准确性，进而帮助构建更完善的分子层面疾病亚型分类，并发现新的预后生物标志物和治疗靶点。

然而，分析多组学数据是一项具有挑战性的任务， fi首要原因是其固有的高维特性。多组学数据集通常包含超过 10 000 甚至 100 000 个预测变量，数量远远超过样本量。为应对这一挑战，人们提出了多种方法， [1], 涵盖从经典机器学习到深度学习的方法。经典机器学习包括 Cox模型 [2] 及其正则化扩展：IPF-LASSO [3], Priority-LASSO [4], 基于 Cox模型的提升方法（包括 CoxBoost） [5] 以及随机森林方法，例如 BlockForest [6]。Herrmann 等人 [7], 对这些方法进行了基准测试，使用 18 个 TCGA数据集表明，在将临床变量与多组学数据结合使用时，平均而言，没有任何方法的表现显著优于仅使用临床数据的基线模型。值得注意的是，在汇总 18 个癌症数据集的结果后，BlockForest 似乎是表现最佳的方法之一。深度学习（DL）方法也已用于单组学生存分析，例如 DeepSurv [8] 和 Cox-nnet [9]。Wissel等人[10]在癌症预后预测背景下对这些方法在多组学分析中的应用进行了基准测试，发现其抗噪能力较差，且表现与 BlockForest 等非深度学习方法相近。人们还专门开发了用于整合多组学数据的深度学习架构，最著名的当属 CustOmics [11]。在同一背景下，类似 CustOmics 的深度学习模型在深度学习方法中取得了最佳预后精确度，但其表现仍与 BlockForest 相近 [12]。

尽管上述方法均未能可靠地优于仅基于临床数据的模型，Herrmann 等人 [7] 表明，利用多组学数据模态结构的方法优于传统方法。然而，这项研究遗漏了多组学整合领域中一类常用的结构化方法

这些方法已被证明表现良好，尤其是在聚类任务中 [13, 14]：联合降维（JDR）方法。这类方法旨在利用不同组学层之间的联合信号，估计一个能够涵盖多组学数据集中主要信息的低维（潜在）空间。此外，JDR方法易于解释，因为个体在潜在空间中的位置图示有助于理解所捕获的信息。而且，与各预测变量相关联的权重能够表征它们在该空间计算中的重要性。

正如 Cantini 等人 [14], 所示，JDR方法的潜在空间与生存数据显著相关。本研究量化了这类潜在空间对新患者生存情况的预测能力，而此前的研究尚未涉及这一点。然而，降维方法传统上属于无监督方法。为解决这一局限，有人提出将其作为预测任务 [15, 16, 17, 18]之前的第一步 fi。在此，我们将这一方法应用于生存预测，遵循 Chegraoui 等人 [18], 的做法，对通过联合降维确定的潜在空间进行 Cox回归。要预测新患者的生存情况，还需要将新个体映射到先前确定的潜在空间中；这要求对部分JDR方法进行方法学上的改进。采用这种方法，我们在 Herrmann 等人 [7], 使用的18个TCGA癌症数据集上，对12种JDR方法以及此前基准测试中表现最佳的方法进行基准测试，并与一个类似的基线进行比较，该基线仅使用临床变量估计Cox模型。除了基于临床和组学数据进行预测外，我们还考虑仅使用组学数据进行预测的情形。此外，我们证明了JDR方法的结果具有可解释性：通过自助法，我们展示了这些方法可用于识别可靠的预后生物标志物。进一步分析低级别胶质瘤中的这些多组学标志物后，可得出具有临床意义的解释，并得到现有文献的支持。

2 结果

2.1 已测试的 JDR 方法原理

本文中，x 表示标量，x 表示向量，X 表示矩阵。对于多组学数据，每种组学数据均表示为其自身的矩阵 $\mathbf{X}_k$，其大小为 $n \times p_k$，$k \in \{1, \ldots, K\}$。个体数量为 n，而 p_k 是组学 k 的变量数量。$\mathbf{X}_k$ 称为数据块或（组学）模态。

联合降维（JDR）方法是多数据块特异性降维（DR）方法，即它们计算一个所有所考虑数据块共有的低维潜在空间。我们评估的 JDR 方法可根据其降维策略分类：基于协方差的方法（Cov）、矩阵分解（MF）方法 [13]和张量分解（TF）方法 [14]。

矩阵分解方法旨在将 $\mathbf{X}_k$ 近似表示为 $\mathbf{Z}\mathbf{W}_k^\top$ 的形式，其中 $\mathbf{Z} \in \mathbb{R}^{n \times R}$ 是一个矩阵，用于描述每个个体在一个所有模态共有的 R 维潜在空间中的位置；$\mathbf{W}_k \in \mathbb{R}^{p_k \times R}$ 是一个针对每个数据块的权重矩阵，可用于根据 Z 重构 $\mathbf{X}_k$（见图 fi1A）。在矩阵分解方法中，我们考察了联合与个体变异解释（JIVE） [19], ，该方法同时考虑所有模态共有的潜在空间，以及每个数据块特有的不同潜在空间；整合式非负矩阵分解（IntNMF） [20], 在权重和潜在空间中加入非负性约束，从而具有良好的聚类特性；iCluster 是另一种专为聚类设计、采用最大似然法的方法； [21, 22]以及多组学因子分析（MOFA） [23, 24], ，这是一种能够处理多种数据分布的贝叶斯方法。

基于协方差的方法旨在为每个数据块寻找一个权重向量 $\mathbf{w}_k$ 和一个成分（潜变量）$\mathbf{z}_k = \mathbf{X}_k\mathbf{w}_k$ ，使数据块成分之间的协方差（或相关性）最大化（见图 1B）。在称为去除效应的过程中，这种最大化操作会重复 R 次；每次重复前，都会从相应数据块中去除已确定的 $\mathbf{z}_k^r$ 的影响。由此得到 R 个 $\times$K $\mathbf{w}_k^r$ 及其对应的 $\mathbf{z}_k^r$，共同构成潜在空间 Z。在基于协方差的方法中，我们考虑了正则化广义典型相关分析（RGCCA）[25, 26] ，该方法将典型相关分析 [27] 和偏最小二乘 [28] 推广到两个以上数据块，通过最大化用户指定的数据块成分对之间的协方差；稀疏广义典型相关分析（SGCCA） [29], ，它是 RGCCA 的一种变体，通过约束 $\mathbf{w}_k^r$ 来实现稀疏性；以及多重协惯量分析（MCIA） [30], ，它在采用与非对称对应分析（NSC） [31]相同预处理步骤的同时，最大化每个数据块成分与合成中心之间的协方差。

张量分解方法要求将组学数据组织成张量的形式，在本例中为一个三维表格 $\mathbf{X} \in \mathbb{R}^{n \times p \times K}$ ，其条目对应于个体、变量（例如基因）和组学的每一种组合。这要求每个数据集具有相同的特征集合，因此我们在基因层面对每种组学模态进行了汇总（详见方法 4.2）。在张量分解方法中，我们考虑了张量主成分分析（tPCA） [32],，这是主成分分析在张量数据上的扩展。该方法旨在寻找三个半正交矩阵：潜在空间 $\mathbf{Z} \in \mathbb{R}^{n \times R}$、所有组学共有的权重矩阵 $\mathbf{W} \in \mathbb{R}^{p \times R}$ ，以及 $\mathbf{O} \in \mathbb{R}^{K \times R}$ ，其中每个元素描述了

组学模态 k 对成分 r 的贡献。随后，可以用 $Z\Delta_k W^\top$近似数据块 X_k，其中 Δ_k 是一个对角矩阵，包含 O 的第 k 行。与非张量矩阵分解所需的 R(n ＋p＋K) 个参数相比，该方法仅使用 R(n ＋Kp) 个参数重构数据，从而进一步增强了数据压缩效果，并提高了降维的降噪能力。该模型将重点提取跨组学共同受影响的基因所形成的分子特征，从而增强模型的可解释性。除 tPCA 外，我们还考虑了张量独立成分分析（tICA），该方法计算旋转，使 tPCA 成分不仅彼此不相关，而且相互独立 [33]。截至目前所介绍的每种 JDR方法背后的详细方程见补充方法。

虽然 JDR方法通常是无监督的，但对于 RGCCA、SGCCA 和 JIVE，可以利用生存信息对潜在空间的构建进行监督。对于 RGCCA 和 SGCCA，监督是通过最大化各数据块与响应块 y 之间的协方差来实现的。在生存预测中，y 是空（不含变量）Cox模型的偏差残差 [34, 18, 35]。监督式联合变异分解（sJIVE） [36] 提出，将原始 JIVE 矩阵分解问题与最小二乘准则进行凸组合并使其最小化；该准则对应于将常见潜在空间和数据块特异潜在空间同时对 y 进行线性回归。我们使用的响应变量与 RGCCA 和 SGCCA 使用的偏差残差相同。因此，所得潜在空间旨在捕获更多与生存相关的信息。

除了 JDR方法之外，我们还考虑了两种主成分分析作为对照，以比较经典降维（DR）方法与 JDR 方法。第一种是在应用主成分分析之前将所有模态拼接起来，对应于早期整合（EI） [10]。第二种则是对所有模态分别应用主成分分析，再将潜在空间拼接起来，对应于晚期整合（LI）；而大多数 JDR方法会被归类为中期整合。

我们还考虑了 Herrmann 等人 [7] 基准测试中的部分方法：被其归类为朴素方法的方法（即不考虑多组学数据的数据块结构），包括最小绝对收缩和选择算子和生存随机森林（ranger） [37, 38]；结构化方法则包括 IPF-LASSO、Priority-LASSO、临床优先型 Priority-LASSO 和 BlockForest。最后，为了将 JDR 方法的性能与基线进行比较，我们采用了这项先前基准测试 [7] 中相同的参照方法：仅使用临床数据的 Cox模型，以及用于估计生存曲线的非参数估计量 Kaplan-Meier 估计量。此处测试的所有方法均以 R 包的形式提供：所用的软件包和函数见表 1。

2.2 基于 JDR方法的生存预测

JDR方法传统上属于无监督方法，并非专门为预测任务设计。文献中已有研究采用降维方法生成的潜在空间作为监督模型的输入 [15, 17, 16, 18], ，我们将这种方法称为序列建模。此处，为了预测生存情况，我们使用 Cox模型，这是一种用于删失生存时间数据的半参数线性模型。与 Herrmann 等人 [7] 的方法（始终将临床数据与组学数据一并纳入）不同，我们还在不使用临床数据的情况下预测生存情况，以评估组学数据本身的预测能力。在纳入临床数据时，我们会在拟合 Cox模型时将其加入潜在空间 fi，以模拟 Herrmann 等人 [7] 中表现良好的临床优先方法。为缓解基于协方差的方法在潜在空间中因多重共线性引发的问题，我们对所有方法均使用经 l_2-正则化的 Cox模型 [39] 。为控制正则化对基线的影响，我们在参照方法中加入了仅使用临床变量的 l_2-正则化 Cox模型。

为了对这种序列建模方法进行基准测试，我们需要在不重新训练模型的情况下，预测测试集中每个个体在估计所得潜在空间中的位置。对于大多数 JDR方法而言，这并非标准任务。RGCCA、JIVE、sJIVE 和 tPCA（+tFOBI）虽可提供预测，但我们必须为 IntNMF、iCluster 和 MCIA 开发预测方法。对于 MOFA，我们采用了 MOTL [40], 这一专为该方法开发的迁移学习方法（更多详情见方法4.3和补充方法）。

我们在嵌套交叉验证（CV）循环中运行这种序列建模方法，其中内层采用 10 折交叉验证来选择最佳超参数，尤其是潜变量的数量和 Cox模型的正则化参数（嵌套交叉验证示意图见 fi图 1C）。外层采用重复的 5×10 折交叉验证，通过两项指标评估生存预测表现：Uno 一致性指数 [41], （判别指标）和积分Brier评分（IBS） [42], （与预测误差相近）（有关交叉验证流程和 JDR方法超参数的更多详情，请参见方法4.4、4.5；Herrmann 等人 [7] 提出的方法的超参数则按照其研究中的描述进行调优）。

该交叉验证设置及所有后续实验均在与 Herrmann 等人相同的 18 个 TCGA数据集上进行[7], ，采用 4 种不同的组学模态：患者内标准化的基因（RNA）和微 RNA（miRNA）表达、拷贝数变异（CNV）以及体细胞突变（突变）。每个数据集中的个体数量从 35（急性髓系白血病）到 735（BRCA）不等，变量总数则从 90 151 到 103 366 不等。嵌套交叉验证

在每个数据集上独立运行，所采用的框架与 Herrmann 等人类似 [7] ，并进行了大幅更新（有关数据集和计算框架的详细信息，请参见方法 4.1 和 4.6）。

2.3 J降维在 18 个 TCGA 癌症数据集上带来可靠的生存预测 s

通过此次交叉验证获得的各方法性能见于 fi图 2。对于每种癌症，将 50 次生存预测的性能在重抽样中取中位数进行汇总。因此，每个箱线图由 18 个值绘制而成，每个值对应一个癌症数据集（未汇总的箱线图，即每种癌症对应一个箱线图，见补充材料中的 fi图 S9 至 S26）。我们未观察到正则化对基线（临床数据上的非正则化 Cox 模型）有任何影响。总体而言，在使用临床数据时，我们观察到 JDR方法在 C 指数上与基线性能相当，并且在 IBS 上能够显著优于基线。对于主成分分析，也能得到相同的观察结果。在 BlockForest 中测试的结构化方法和朴素方法中，BlockForest 已被反复证明是这一情境下表现最佳的方法 [7, 10, 12, 43], ，也是唯一一个其中位数几乎与参考模型相当的方法，尽管其结果并未显著更好。在不使用临床数据时，所有方法的性能均有所下降，无论是在 C 指数还是 IBS 上；然而，大多数结构化方法和 JDR方法在 IBS 上仍与基线性能相当。值得注意的是，BlockForest、sJIVE、S/RGCCA Sup. 和 SGCCA Unsup. 在 IBS 和 C 指数上均未显著劣于基线。

几乎看不出明显倾向于某一种特定 JDR 策略。张量分解方法在临床数据下未取得任何显著结果，在不使用临床数据进行评估时也未达到基线水平。此外，在 tPCA 后应用或不应用张量独立成分分析（此处使用 tFOBI）似乎并不会影响结果。对于基于协方差的方法，除使用临床数据时的 C 指数外，MCIA 的中位性能结果与 S/RGCCA 相近。在所评估的基于协方差的方法中，主要差异在于 MCIA 的预处理步骤，该步骤最初是为分析列联表而引入的。SGCCA 的中位性能结果接近其非稀疏对应方法 RGCCA，但似乎略低一些；不过，在使用临床数据时的 C 指数中，监督版本除外。这一趋势与文献中关于 L1 与 L2 正则化的研究结果一致。对于矩阵分解方法，MOFA 和 iCluster 的中位性能结果系统性地低于 JIVE/sJIVE。intNMF 也存在这种情况，但仅限于 C 指数。这些模型之间的系统性差异在于，JIVE/sJIVE 能够提取特异性的组学效应以及共同效应，而后者是 intNMF/MOFA/iCluster 唯一关注的部分。比较主成分分析方法之间的性能差异可以发现，这些特异性效应确实十分重要，因为除不使用临床数据时的 IBS 外，LI-PCA 的中位性能结果始终优于 EI-PCA；在 IBS 中，两者表现相当。文献中已多次观察到 LI 优于 EI 的这一趋势 [11, 10]。

最后，在所有比较的方法中，LI-PCA、JIVE 和 S/RGCCA 系列始终表现出良好且相近的性能结果，在大多数情况下达到或超过基线。这些观察结果与 Cantini 等人 [14], 的基准测试一致；该测试得出结论：RGCCA、JIVE 和 MCIA（未测试 LI-PCA）是生存预测的最佳方法，不过其评估并未通过交叉验证进行。

表 2 展示了每种设置（临床数据有无与指标的组合，例如使用临床数据时的 Uno 一致性指数）和癌症类型下表现最佳的非基线方法。每种测试方法至少有一次被评为最佳方法。在最佳方法中，JDR方法的出现次数仅略多于非 JDR 方法。因此，如果要寻找表现最佳的方法，不应排除任何方法。唯一一个在将组学数据与临床数据结合使用时性能下降的癌症是 UCEC，在 C 指数和 IBS 指标上均如此。在 72 种设置与癌症的组合中，有 19 种观察到超过 10% 的提升，涉及 18 种癌症中的 7 种。LAML 是唯一一个 4 种测试设置均受到影响的癌症，其 C 指数在有和无临床数据时的提升幅度分别为 21% 至 28%。COAD 和 SARC 在 4 种设置中的 3 种下有所提升。对于 C 指数和 IBS，HNSC/KIRP 以及 LGG/LIHC 在有和无临床数据时均实现了超过 10% 的提升（KIRP 甚至达到 30%）。在这些提升中，11 例归因于 JDR方法，3 例归因于非 JDR 结构化方法，5 例归因于朴素方法。

我们在图 3 中展示了针对每种癌症、各项指标以及是否存在临床数据时，每种方法与基线之间相对差异的热图。对于每种情况，我们分别对癌症和方法独立计算了层次聚类。我们注意到，癌症对相对差异有明显影响，其作用超过了方法的影响。在有临床数据时，对于两项指标，相较于基线，大多数方法在 LGG、SARC、HNSC、COAD 和 LAML 中都能提升性能；而 KIRP 仅在 C 指数中有所提升，LIHC/PAAD 仅在 IBS 中有所提升。对于所有其他癌症，性能均与基线相当。在没有临床数据时，对于两项指标，大多数方法在 HNSC、COAD 和 KIRP 中能提升性能，而 LGG/SARC/LIHC/LAML 仅在 IBS 中有所提升。此外，对于两项指标，SKCM、KIRC、LUAD、STAD 和 UCEC 的预后精度大多下降；其他所有癌症（LAML/LUSC/LGG 除外）仅在 C 指数中出现下降。文中未提及的所有其他癌症

在 IBS 中，患者的预后与基线相当。总体而言，在大多数方法中，HNSC 和 COAD 在所有条件下都观察到性能提升，而 LGG/SARC/KIRP/LAML 在 4 种条件中的 3 种下观察到性能提升。这些数据集的大小差异很大，其中 LAML 是最小的数据集（35 名患者），而 LGG 和 HNSC 属于最大的数据集之列（分别有 419 名和 443 名患者）。预后精度的下降仅出现在没有临床数据的情况下，并且对于两项指标，仅涉及 5 种癌症（患者数量为 249 至 426 人）。尽管样本量似乎不会影响每种方法与基线之间的相对差异，但会影响方法的绝对性能（仅 IBS 具有统计学显著性，见补充结果 2.1）。

就方法而言，基于协方差的方法在 COAD 上的表现似乎略优于其他类型的方法。LASSO，以及表现稍逊的 ranger，其表现似乎也低于其他方法。这与 Herrmann 等人 [7] 得出的结论一致，即结构化方法优于朴素方法。有趣的是，在全部 4 种设置中，MOFA、iCluster 和 IntNMF 都聚在一起；除不含临床数据的 IBS 外，它们在所有设置中也都与 EI-PCA 聚在一起。由于这些方法只提取所有组学数据共有的 eff效应，这或许可以解释它们在不同癌症中的表现为何如此相似。

有关这些方法的技术信息可在表 1 中查阅错误率，并可在补充结果 2.2 中查阅软件可用性和运行时间。值得注意的是，JDR方法比 Herr- rmann 等人 [7]的研究中的大多数方法运行得更慢（BlockForest 除外）；而后者本身又比参考方法运行得更慢（见 fi图 S8）。在 JDR方法中，矩阵和张量分解方法通常比基于协方差的方法运行得更慢。不过，两种 PCA 的运行时间与先前研究中的方法相当。

2.4 组学数据对 JDR方法预测的贡献

为估算预测中来自组学数据或临床数据的部分，我们针对每种癌症、每个外层折和每种 JDR方法，根据相应的 Cox模型计算组学贡献（取值范围为 0 到 1）。随后对各外层折的贡献取平均值，并展示于 figure 4 中（详情请参阅方法4.7）。从箱线图中可以看到，对于大多数 JDR方法以及 LI-PCA，约有一半的预测来自组学数据。然而，对于 iCluster、MOFA 和 IntNMF，组学数据对预测的贡献较小（∼25%）。除了它们都属于矩阵分解模型外，它们也是仅有的严格聚焦于常见效应的 JDR方法，这或许可以解释这些结果。采用较少组学数据的维度约简模型 EI-PCA 也呈现出相同情况；与之相对的是只聚焦于特定效应的 LI-PCA。有趣的是，EI-PCA 的组学潜在空间对生存预测的贡献似乎高于 iCluster/MOFA/IntNMF，这或许可以用主成分分析对潜在空间施加的特定正交约束来解释。相反，监督式 S/RGCCA 的各个版本几乎完全使用组学数据进行预测。sJIVE 的表现有所不同，其组学数据对预测的贡献与无监督版本 JIVE 相近。这或许可以用监督策略的差异（回归与协方差）来解释。

从各癌症和方法的组学贡献热图中可以看到，不同癌症之间存在显著差异。大多数方法（EI-PCA、IntNMF、iCluster 和 MOFA 除外）在 BRCA、HNSC、结肠腺癌、KIRP 和肝细胞癌中使用了更多组学数据，而在胃腺癌、食管癌和肺腺癌中使用较少。S/RGCCA Sup. 是唯一不符合这一趋势的方法，但它在胰腺腺癌和肾透明细胞癌中使用的组学数据较少。除组学数据使用量较高外，HNSC、结肠腺癌和 KIRP 的预测表现也始终优于基线，无论是否纳入临床数据（见 fi图 3）。因此，这些癌症中的组学数据可能包含与预后相关的信息；这些信息与临床数据部分重叠，同时也能对其起到补充作用。相比之下，在胃腺癌和肺腺癌中，不纳入临床数据时的预测表现始终低于基线，而纳入临床数据时则与基线相近。

2.5 JDR方法的自助法识别预后生物标志物 s

为展示降维方法的可解释性，我们进行了自助法实验，以估计对于每种方法而言，与 JDR模型相关的变量权重以及 Cox 模型中的潜在成分权重的置信区间。我们选择了监督式和非监督式 RGCCA 、 JIVE 、 sJIVE，以及在所有癌症上应用的 LI-PCA 来进行这些实验，因为这些方法的表现始终良好。由于稀疏性会导致自助抽样样本之间产生额外的不稳定性，而这需要通过专门的程序进行研究，因此在本分析中排除了 SGCCA。因此，对于每种癌症，每个模型均进行了 fi 次有放回重抽样，共 99 次，以计算每个变量和成分的置信区间及相应的 P 值（详见方法4.7）。图 5 展示了 Cox模型中显著潜变量的比例，以及每种模型中显著基因的比例。我们计算了每个数据块和每种癌症中显著变量的比例。对于 JDR模型，当某个数据块包含多个成分时，我们计算了与该数据块相关的各成分之间显著变量比例的平均值。最后，我们展示了每种方法和每个数据块的均值及 95% 正态置信区间。

均值及每种方法和每个数据块的 95% 正态置信区间。对于 Cox模型，我们观察到，当将临床数据加入潜在空间后，临床变量中有一个稳定的比例（介于 14.8% 和 18.1% 之间）是生存的显著预测变量。在这些临床变量中，最常被检出的变量是诊断时年龄；由于该变量存在于全部 18 种癌症中，因此在 18 种癌症的 30 个 Cox模型中均被检出，其中包括 $\times 5$ JDR方法 $= 90$ Cox模型。癌症分期可能更系统地与生存显著相关（见表 S8），但其在不同癌症中的存在情况和命名方式各不相同。有趣的是，SKCM 和 UCEC 均有一个临床变量在全部 5 个模型中达到显著：分别是 Breslow厚度（黑色素瘤厚度）以及肿瘤对子宫肌层（子宫壁的肌肉层）的浸润百分比。这两者似乎都承载着宏观尺度的信息，难以直接通过分子数据捕捉，但与预后高度相关。这或许解释了为何对于 SKCM 和 UCEC，在不包含临床数据时，所有模型和指标的预测结果均系统性下降。

由组学数据计算得出的成分在不同方法间的使用情况较不一致，且总体占比也较低；监督式RGCCA除外，其组学成分在Cox模型中有较高比例达到显著（比例介于 76.7%和 93.3 %之间），这与图4相符。就JDR模型而言，无论是否纳入临床数据，所有数据块中都发现了显著基因。LI-PCA中显著变量的总体占比较高，而RGCCA中的占比较低。对于大多数方法，显著基因占比最高的数据块是RNA区块和miRNA区块。除纳入临床数据的JIVE外，所有方法的突变数据块中显著变量的占比均低于其他数据块。无论是否纳入临床数据，JIVE中显著突变的方差都较高。

2.6 JDR模型的解释揭示了与低级别胶质瘤生存相关的多组学特征

随后，我们选取对Cox模型具有显著意义的成分，并对与这些成分相关的自助法比率进行基因集富集分析（GSEA）（见方法4.7）。对于RNA组分，我们使用MSigDB标志集作为基因集 [44] ；对于miRNA组分，我们使用TAM 2.0功能集 [45]。为简洁起见，我们选择展示4种不同模型的结果：在低级别胶质瘤（LGG）上，使用临床数据的 JIVE（JIVE 临床版），以及不使用临床数据的JIVE、监督式RGCCA和RGCCA 非监督版。选择该数据集有多方面原因：(i) 在不使用临床数据时，监督式RGCCA、RGCCA 非监督版、LI-PCA、JIVE和sJIVE在IBS方面优于基线；(ii) 当存在临床数据时，所有采用自助法的方法在预测LGG生存时都高度使用组学数据（fi图4）。对于RGCCA模型，使用和不使用临床数据时的超参数完全相同（正则化参数除外），因此在解释时我们优先选择不使用临床数据的模型（见表S3和S4）。

对于这些模型中的每一个，与每个显著的 RNA 和 miRNA 成分相关、根据自助法Cox模型确定的前10名显著富集基因集（按基因比例排序，Q值阈值为0.05）见图6。对于 RNA 成分，至少有一个 JIVE 模型和一个 RGCCA模型鉴定出12个基因集，且这些基因集均对生存产生有害影响。在这12个基因集中，6个被归类为免疫基因集：同种异体移植物排斥、凝血、IL6/Jak/STAT3信号传导、炎症反应、干扰素-α响应和干扰素-γ响应；1个属于发育类：上皮-间质转化；1个属于通路类：细胞凋亡；2个属于增殖类：E2F靶基因和 G2M检查点；另有2个属于信号传导类：通过 NF-κB 的 TNF-α信号传导和 mTORC1信号传导。因此，LGG 转录组学数据中的大部分预后信号似乎来自免疫过程，且这些标志特征的过度表达会带来不良影响。在这些免疫过程中，干扰素-α和干扰素-γ响应在所有成分和模型中均表现出富集，而 IL6/Jak/STAT3信号传导和同种异体移植物排斥则在除一个成分外的所有成分中富集。已有研究表明，IL6 和 STAT3 相关基因的表达与胶质瘤的肿瘤分级和患者生存相关 [46, 47, 48] ，而 I L6/Jak/STAT3通路已被证实可促进肿瘤生长和免疫逃逸 [49, 50]。持续的干扰素信号传导也可促进肿瘤进展和免疫抑制。值得注意的是，持续的干扰素-α信号传导会影响 STAT3信号传导，而持续的干扰素-γ信号传导则与更高的 IL6表达相关 [51, 49]。间质转化是除一个成分外所有成分均鉴定出的另一项标志特征，也是生存的重要驱动因素，因为它与复发性及治疗耐药肿瘤相关 [52]。尤其值得关注的是，间质转化会受到其他富集通路的影响，即通过 NF-κB 的干扰素-α响应以及 IL6/Jak/STAT3信号传导 [49]。我们可以推测，这些成分捕获了某一胶质瘤亚型的特征，该亚型表现出与胶质母细胞瘤间质亚型相似的机制，而后者与不良预后相关 [53]。补充结果2.3对这一点进行了进一步探讨，其中我们展示了 JIVE 临床版能够识别低级别胶质瘤的3种已知亚型 [54], , 其中包括更接近胶质母细胞瘤的 IDH -野生型胶质瘤 [55, 56]。最后，E2F靶基因和 G2M检查点标志特征参与细胞周期调控，而细胞周期调控是癌症中常见的失调过程 [57]。I

n miRNA，15 个基因集 JIVE 和 RGCCA 模型均发现了以下通路：脂肪细胞分化、骨再 -

生成、细胞分裂、DNA损伤反应、卵泡发生、葡萄糖代谢、粒细胞生成、先天免疫、神经元分化、致癌微小RNA、成骨细胞分化、氧化应激、腹腔稳态、T细胞分化和毒性。除神经元分化外，这些通路都与风险增加相关；大多数模型显示，神经元分化与风险降低相关。粒细胞生成、T细胞分化和先天免疫均属于免疫过程，可能与我们在RNA 组分中发现的免疫特征有关。细胞分裂和葡萄糖代谢可能与肿瘤增殖有关，而神经元分化大多表现出的保护作用，可能与一种类似于胶质母细胞瘤前神经元亚型的亚型有关；与胶质母细胞瘤的其他亚型相比，该亚型的预后更好 [53]。

为进一步解读低级别胶质瘤，我们展示了JIVE、JIVE 临床版、监督式RGCCA和RGCCA 非监督版的拷贝数变异组分的自助法比率，见图7。除RGCCA 非监督版的cnv.1成分外，所有成分对生存预测均具有显著性；尽管如此，我们仍保留了该成分，因为按照设计，它与对应的RNA和miRNA成分高度相关。所有模型均识别出1号染色体短臂和19号染色体长臂上的强信号，这对应于文献中报道的常见1p/19q共缺失。对于大多数成分而言，这是一种保护性信号，与Wen等人的研究结果一致。在JIVE 临床版的第三个常见成分中，归因于该共缺失的影响却是有害的，这与其他所有成分及文献报道相反。补充结果2.3对这一影响进行了探讨，其可能由每个JIVE 临床版成分区分亚型的方式所解释。文献中观察到的另一个事件是10号染色体的有害缺失，该事件通常与7号染色体增益共同发生；我们可以将所有成分识别出的信号与之关联起来。最后一个值得关注的信号是9p染色体上的峰值，该区域包含CDKN2A/B基因；已有研究表明，该区域的缺失与较差的预后相关，这与此处观察到的信号一致。

对于 JIVE、JIVE 临床版、监督式RGCCA和RGCCA 非监督版，其突变成分的自助法比率见图 8。与拷贝数变异一样，所有成分均显著，但 RGCCA 非监督版的 mutation.1 成分除外。JIVE 的 mutation.1 成分未识别出任何显著基因。JIVE 临床版的所有共同成分均将 IDH1 识别为保护性基因；其前两个共同成分将 FUPB1 和 CIC 识别为保护性基因，而第二个成分将 NOTCH1 和 IDH2突变识别为保护性，这与 Wen等人 [54]定义的 I型胶质瘤一致。对于有害突变基因，三个成分均识别出 EGFR，成分 2 和 3 识别出 PTEN，成分 3 还识别出 NF1，这些结果均与 III型胶质瘤一致 [54]。II型胶质瘤的标志物，即 TP53 和 ATRX，在成分 2 和 3 中呈现相反效应。CIC 和 FUBP1 也是在这两个成分中被识别为具有相反效应的基因。这些符号变化与 JIVE 临床版识别出的亚型一致，相关探讨见补充材料 2.3。RGCCA 非监督版识别出 FUBP1、IDH1、EGFR、NOTCH1、NF1 和 CIC，且其效应方向一致。其他识别出的基因在胶质瘤文献中没有已知的预后相关突变：JIVE 临床版的第二个共同成分识别出 TTN 和 FAT2，第三个共同成分识别出 PIK3CA；RGCCA 非监督版还识别出 11 个基因。监督式RGCCA识别出大量显著基因（43 个），其中大多数在文献中尚无相关报道。对此应谨慎解读，因为将仅拟合突变数据块的监督式正则化广义典型相关分析模型与拟合所有数据块的模型进行比较后发现，监督式RGCCA在突变数据块上过拟合（补充图 S59）。因此，在这种情况下，突变数据块对预测的贡献最大（补充图 S55；关于该图的构建方法，详见方法4.7）。我们推测，监督式RGCCA利用罕见突变来区分生存时间特定的个体。尽管如此，IDH1/2、CIC 和 EGFR 的效应方向仍与文献一致。

3 讨论

在本研究中，我们针对生存预测情境，对 12 种 JDR 方法进行了严格的基准测试。这类方法广泛应用于多组学研究领域，但 Herrmann 等人 [7] 的基准测试并未涵盖这些方法。通过使用相同的数据集，并采用与这项先前研究相似的框架，我们使结果更易于比较。此外，我们拓展了 Cantini 等人 [14]的工作，通过在预测任务中测试 JDR 来评估其表现；尤为重要的是，我们通过交叉验证，在模型未见过的个体上使用生存特异性指标评估预测能力。为此，我们必须为部分受测算法开发特定的预测方法；这一方面在侧重无监督分析的 JDR 方法中往往未被探索。得益于这种预测导向，我们的工作既有助于在研究队列中识别新的生物标志物，也适用于临床常规中的预后检测。尽管我们侧重于生存预测，但借助上述新型预测方法以及序列建模的灵活性，我们的工作也可拓展至其他预测任务。

从这项基准测试来看，使用 JDR方法结合序列建模来整合组学数据和临床数据，平均而言，其生存预测表现优于仅使用临床数据的基线方法。Herrmann 等人测试的方法则不符合这一结论 [7]（无论是否采用结构化方法），这与他们的原始研究结果一致，

尽管我们的汇总策略有所不同。若不使用临床数据，BlockForest、RGCCA 和 JIVE模型是仅有的能够在两项指标上达到临床数据预测表现的方法。这对发现新的生物标志物尤其有希望。不过，似乎没有一种方法适用于所有癌症；在针对特定癌症寻找最佳表现时，也不能先验地排除任何方法（无论是否为 JDR方法、是否采用结构化方法）。我们还注意到，不同癌症对各方法相对于基线的表现有很大影响，表现提升或下降的幅度从样本量最小的癌症到样本量最大的癌症都有。我们推测，对于某些癌症，组学数据与生存之间的关联更强，或者至少更容易通过本文介绍的方法提取。例如，某种癌症可能存在生存差异明显的亚型，而特定模型能够捕捉到这些差异，低级别胶质瘤便是如此。

JDR方法的部分性能提升可能来自降维，而非联合特性，因为 LI-PCA 并未考虑组学之间的联系，却优于其中一些方法。此外，LI-PCA 也优于 EI-PCA，这似乎表明，与仅提取联合效应相比，仅提取组学特异性效应是一种更优的策略。在序列建模框架中，先提取组学特异性效应，再拟合 Cox模型，可以让这一生存模型辨别涉及哪些组学以及其影响程度。这与 Wissel等人 [12] 的研究相呼应；他们指出，在类似情境下，执行多组学整合的最佳策略是根据生存任务进行组学选择。

尽管 LI-PCA 的表现与部分 JDR方法相近，但表现最稳定出色的方法是 JIVE 和 S/RGCCA，我们建议在测试的方法中采用这两种方法。这与 Cantini 等人 [14] 的研究结果一致。正如他们的研究所指出的，这些方法的共同点是能够计算各模态特异性的潜变量，从而放宽了大多数 JDR方法的限制——即只能包含所有数据块共有的成分。然而，它们的理论方法有所不同：一方面，JIVE 既能提取共有成分，也能提取各数据块独有且彼此不相关的特异性效应。另一方面，RGCCA 通过最大化协方差，寻找既能解释所属数据块部分方差、又能与其关联数据块相关的成分。因此，一个成分可以同时包含共有和特异性效应。它们提取这两类效应的灵活性或许可以解释其相较于 LI-PCA 的表现。除了各模态共有或特异的成分外，一些方法还会考虑仅与一部分组学模态相关的成分（例如，仅根据 RNA 和 miRNA 计算得出的成分） [59, 60]。然而，这类方法在当前技术发展水平中仍不常见，且计算成本高昂，因为它们必须考虑组学模态所有可能组合构成的完整空间。

监督式JDR方法可作为一种有用的解释工具，通过增强潜在空间中的生存信号来发挥作用；例如，RGCCA的监督式方法产生了更高比例的与生存显著相关的潜变量。不过，JIVE的监督式版本并未呈现这一结果，原因可能是其归一化方式不同于常规JIVE，也可能是其收敛阈值不如JIVE严格。具体而言，RGCCA的监督式版本在性能上与非监督式版本相当，同时所需计算的成分更少，收敛也更快，因此在计算时间方面具有优势。然而，正如低级别胶质瘤的案例所示，监督式RGCCA可能会在至少某些数据块上出现过拟合。这削弱了监督方法的一些优势，因为泛化能力较差的过拟合模型不仅性能会下降，也更难解释。偏差残差是一种简单有效的方法，可用于监督生存数据的降维方法 [34, 35]，但该方法依赖于对Cox似然的近似。因此，监督式降维方法可以受益于一种精确方法，例如使用Cox部分似然或其他生存特异性指标。

对于某些癌症，JDR方法能够很好地利用组学数据预测生存，从而在有无临床数据的情况下达到或优于基线模型。这表明，进一步解释这些结果的获得方式有助于识别预后生物标志物。我们表明，自助法在此情境下是一种有用的工具，可用于同时量化每个潜变量对生存预测的重要性，以及单个基因在潜变量中的重要性。富集分析也可用于识别失调通路和机制。在低级别胶质瘤这一特fi定情形中，我们能够识别出文献中已知与预后相关的染色体事件和突变基因。我们还识fi别出若干基因集（标志特征），其表达与较差预后相关，且似乎与已知的癌症机制相符。尽管如此，这些结果必须谨慎解释，因为并非每一个对生存预测成分有贡献的基因本身都与生存相关，而更可能与潜在的疾病亚型相关，低级别胶质瘤便是如此。在实践中，我们建议比较至少两个采用不ff同整合策略的模型结果，以得出更清晰的解释。对于低级别胶质瘤，JIVE 临床版在所测试的模型中识fi别出最清晰的特征。对于该癌症，交叉验证仅为JIVE 临床版选择了共同成分。由于所有组学均对共同成分有贡献，这或许可以解释此类成分更易于解释的原因。这进一步支持如下观点：能够同时提取组学特异fi性与联合效ff应的JDR应被优先采用，因为这种 fi灵活性可提高预测性能，而联合成分

增强可解释性。

4 方法

4.1 数据集

为评估联合降维方法的预后性能，我们使用了与Herrmann 等人 [7], 相同的18个TCGA数据集，以便于各项研究之间的比较。对于每个数据集，可获得四种分子数据类型：信使 RNA和miRNA表达、按基因计的拷贝数变异（CNV）（这三者均来自肿瘤样本），以及基因层面的体细胞突变。Herrmann 等人 [7]的作者在第32版（Gencode v22）发布之前，使用R包TCGAbiolinks（函数GDCquery）从TCGA下载了数据。对于每种癌症类型，原始信使 RNA计数（data.category = "转录组分析"，data.type = "基因表达定量"，workflow.type = "HTSeq - Counts"）经 fi滤以去除低表达基因（在 $\geq 95\%$的患者中< 0.5 每百万读数）。随后使用limma 包中的voom函数获得log2 每百万计数（log2-CPM）。对于miRNA，选取了每百万读数值（data.category= "转录组分析"，data.type = "微小 RNA 表达定量"），并与信使 RNA类似，去除低表达微小 RNA（在 $\geq 95\%$的患者中< 0.5 每百万读数），再用voom函数获得log2-CPM。对于拷贝数变异，检索了所有人类 Ensembl ID的片段均值（data.category ="拷贝数变异"，data.type = "拷贝数片段"），并 fi滤除了所有患者中的空值。对于突变（此次使用函数GDCquery_Maf，流程为 = "mutect2"），仅保留在至少一名患者中存在至少一次体细胞突变的基因。若某个基因带有至少一次体细胞突变，则其基因值设为1，否则设为0。每种癌症还可获得数量不等的临床数据，例如每位患者的年龄、性别和癌症分期。每个数据集的个体数从35（急性髓系白血病）到735（BRCA）不等，变量总数介于90 151与103 366之间，其中约60%来自拷贝数变异数据块，25%来自RNA，15%来自突变，而miRNA与临床数据占变量总数的比例不足1%（见补充表 S2）。在这18个数据集中没有任何一个数值缺失。

4.2 张量分解方法的张量构建

如结果部分所述，张量分解方法要求每种模态具有相同的特征集合。此处，RNA、突变和拷贝数变异数据集已按基因层面汇总，因此我们保留这三种模态交集中的特征。由于 Herrmann 等人 [7] 仅列出了在某种癌症类型中至少有一名患者观察到体细胞突变的基因，突变数据块可能相当小（例如，急性髓系白血病有 2176 个基因，见表 S2）。因此，我们将 RNA 数据集中存在、但突变数据集中缺失的基因替换为全零列，以明确计入这些基因中不存在体细胞突变的情况，最终保留尽可能多的基因来构建张量。如果不进行这种插补，我们会丢失许多基因，因为突变数据块通常是最小的。我们本可以将拷贝数变异数据集中存在、但突变数据集中缺失的基因替换为全零列。然而，根据拷贝数变异添加突变列，就必须在 RNA 数据块中添加相应的列，因为拷贝数变异数据块中的许多基因都是非编码基因。但 RNA 数据块的插补更为复杂，需要对这些非编码基因的表达作出强假设。成熟微小 RNA 仍需按基因层面汇总。我们使用 R包 miRBaseConverter [61]，根据前体 ID 获取成熟微小 RNA ID。基于 MONTI [62], 的方法，我们计算了靶向每个基因的微小 RNA 的平均表达值。经验证的靶标通过 multimir R 包 [63]获取。由于一个前体 miRNA 可能对应多个成熟 miRNA，其靶标集合被定义为相应成熟 miRNA 靶标集合的并集 fi。

4.3 序贯 Cox 建模

几乎所有 JDR 方法都是无监督方法，并非为预测任务而设计。在文献中，这一问题通过将 JDR 方法确定的潜在空间用作预测任务的变量来处理，例如，在生存预测中使用 JIVE [16] ，以及在分类 fi或预后 [18]研究中使用 R/SGCCA [29] 。为了根据 JDR 方法的潜在空间 Z 预测生存，我们采用相同的方法，称为序列建模：对于有监督和无监督 JDR 方法，我们均将其潜在空间用作正则化 Cox 回归 [2], 的变量。Cox模型是一种用于事件发生时间删失数据的半参数线性模型。我们选择 Cox模型，因为它是用于删失数据的最简单模型之一，而且有助于与参考模型进行比较。Herrmann 等人 [7] 仅考虑了使用临床数据以及

组学数据的方法，以确定组学数据能否改善预后检测。由于我们也关注在不考虑临床变量的情况下潜在空间与生存数据之间的关系，因此针对这种序列建模方法，我们考虑了两种可能性：是否将临床变量添加到潜在空间中。在适用情况下，临床数据会在 Cox 回归阶段加入，而不是纳入 JDR 方法，以模拟 Herrmann 等人 [7]中表现良好的、偏向临床数据的方法。

为了对这种序列建模方法进行基准测试，我们需要在不重新训练模型的情况下，对测试集中的个体进行预测。对于大多数 JDR 方法，这并非标准任务，因为它们属于无监督方法。对于主成分分析，预测新个体的位置很简单：由于满足正交约束，权重矩阵是一个投影矩阵，因此我们可以轻松地将新的观测值映射到训练过程中确定的潜在空间中。然而，对于此处介绍的大多数 JDR 方法，正交约束并不成立。在可行的情况下，我们使用了软件包提供的函数（RGCCA、JIVE、sJIVE、tPCA 和 tPCA+tFOBI）。在无法使用这些函数时（IntNMF、iCluster、MOFA、MCIA），我们设计了一种预测方法：固定变量的权重 fi，同时尽可能贴近原始优化问题，从而确定新个体的位置。对于 IntNMF，我们使用软件包的内部函数求解新个体的位置。对于 iCluster，我们在 R 中复现了原始 Fortran E-M 算法中的期望步骤。我们最初也考虑了 iClusterPlus [64] ，因为它可以考虑特定的数据分布 fi（二项分布、多项分布、高斯分布和泊松分布），而 iCluster 只能考虑高斯分布。然而，我们未能为该方法开发出可行的预测函数：iClusterPlus 在 E 步中使用 Metropolis-Hastings 算法，而非 iCluster 的闭式解，并且我们无法使该算法在新数据上收敛。对于 MOFA，我们采用了一种已发表的迁移学习方法 MOTL [40], ；该方法使用与 MOFA 相同的贝叶斯框架，在固定权重的同时 fi计算新个体的位置。最后，对于 MCIA，我们必须对新个体应用非对称对应分析 [31]预处理步骤。我们测试了两种方法，并选用了表现较好的一种。随后，预测 MCIA 中新个体的位置便很直接，与 RGCCA 类似。有关这些预测方法的更多详情，请参阅补充方法。

4.4 JDR方法的交叉验证设置

为了严格评估序列建模方法的性能，并选择最佳超参数（最值得注意的是潜变量的数量以及 Cox模型的 l_2 正则化参数），我们采用嵌套交叉验证方法，方式与 Herrmann 等人基准测试的部分方法类似[7], 例如 cv.glmnet（嵌套交叉验证示意图见 fi图 1C）。在 fi第一层中，我们采用 10 折交叉验证来选择最佳超参数组合，但 iCluster 除外，由于其计算负担较重，仅采用 3 折。该 fi第一层的评估指标为 Cox部分对数似然，并采用 Verweij等人提出的近似方法计算 [65], ，这是 glmnet 的默认方法，也可视为 Cox似然泛化误差的近似值。这一 fi第一层（内层）交叉验证在外层重复 10×5折交叉验证的每次迭代中运行，后者通过下文介绍的两项生存指标来评估性能。每种癌症的重复 10×5折交叉验证意味着，综合所有癌症，每个模型都要运行 18 个癌症 $\times$ 5 折 $\times$ 10 次重复 $= 900$ 折；而对于每一折，还要运行 10 个内层折，并针对测试的每种超参数组合重复这一过程。

我们通过两项针对生存预测的fi专用指标来衡量模型的表现：Uno 的 C 指数 [41]和积分Brier评分（IBS） [42]。直观而言，C 指数将基于 Cox模型预测风险得出的个体排序，与基于实际死亡时间得出的真实排序进行比较：它大致相当于患者配对中排序正确的比例。IBS 会针对每位患者计算模型输出的生存曲线与理想生存曲线之间的距离；理想生存曲线是一个阶梯函数，在观察到的死亡时间之前取值为 1，之后取值为 0。因此，对于不会影响排序的微小预测变化，它更为敏感；在模型选择中，它应优先于 C 指数作为评估指标，不过后者更容易解释。我们的两项生存指标都需要估计删失分布，因此对研究中较晚时间点被删失的个体高度敏感 [66, 67]。为减轻这一ff影响，我们按照 Wissel等人 [10]的方法，将两项指标的计算时间范围都截断于一个时间阈值，该阈值为删失时间的 0.9 分位数。最后，对于执行出错的重抽样，如果此类重抽样占比低于 20%（10 次重抽样），则用该癌症对应重抽样结果的中位数替代其表现；否则，C 指数和 IBS 分别替换为 0.5 和 0.25 [7] ，这两个数值代表随机预测的表现。

4.5 JDR方法考虑的超参数

我们纳入的 JDR方法各有一组超参数。不过，它们都共享两类超参数：一类是由于采用序列建模方法而需要设置的 Cox 模型正则化强度，我们使用与 glmnet 相同的流程来确定该参数；另一类是潜变量数量，这是我们针对 EI-PCA、LI-PCA、IntNMF、MOFA 和 MCIA 测试的唯一超参数。我们测试了{1、5、 10} 或 {2、5、 10}这几种不同的潜变量数量；后一种情况适用于 IntNMF、MOFA、MCIA 和 tFOBI，因为这些方法不允许潜变量数量为 1。提取成分的上限设为 10，因为我们认为这是信号提取量与计算成本之间的合理平衡。Cantini 等人 [14]的研究也使用了 10 个成分。请注意，对于基于协方差的方法，这些数字指的是每种模态的潜变量数量，因此实际的潜在空间包含其四倍数量的变量（对应四个组学数据块）。为了确定正则化参数网格，我们在整个外层训练集上运行 glmnet [39]，流程与 cv.glmnet 类似。因此，正则化参数的最大值是一个最小值，在该值下 Cox模型不会选择任何变量 [68], ；最小值则比最大值低 100 倍。我们在这两个值之间按对数尺度等距取 100 个值。接下来，对于 JDR方法的超参数，如果其中一些可以针对每种模态设置不同的值，为降低交叉验证的计算成本，我们选择将这些超参数ff固定为四种模态完全相同的值。因此，对于 JIVE，我们测试了共同成分数量为 0、1、5 或 10，与个体成分数量为 0、1、5 或 10 的各种组合（不包括完全不提取成分的情况，总计 15 组超参数组合，不含正则化参数）。对于 sJIVE，我们加入了 $\eta \in \{0.1、 0.5、 0.9\}$的组合，以调节监督学习的程度（总计 75 组；更多细节见补充方法 η）。对于 RGCCA，我们考察了潜变量数量（{1、5、 10}）与若干 $\tau \in \{0.1、 1\}$取值的组合，以控制相关性与协方差之间的权衡（共 6 组；更多细节见补充方法 τ）。对于 SGCCA， τ 设为 1，稀疏性超参数 s 则需要调优。我们将 s 设为其最大值的 25%、50% 或 75%；最大值取决于每个数据块中的变量数量，达到最大值时模型不稀疏。与潜变量数量组合后共形成 9 组。对于 iCluster，为减轻计算负担，我们只考虑两种潜变量数量：1 或 5，并设置稀疏性超参数，使大约 0%、33% 或 66% 的变量被选中（共 6 组）。这些数值是通过在每个数据集的所有样本上运行 iCluster，凭经验确定的。内层交叉验证循环为每个模型选出的各超参数组合的频率见 fi图 S27 至 S40。

4.6 基准测试的实现

除包含 sJIVE 的 sup.r.jive 以 GitHub 仓库形式提供外，所有使用的程序包均可在 CRAN 或 Bioconductor 上获取。我们对两个程序包进行了修改：IntNMF，用于修复一个影响多次初始化时最佳因子分解选择的错误；iClusterPlus，用于允许 Fortran 代码在 conda 环境中编译。由于 R 实现不会保存 MOTL 所需的全部参数，为了使用 MOTL，我们通过 reticulate 直接调用 mofapy2 Python 程序包运行 MOFA。iCluster 在这些数据集上运行需要大量随机存取存储器（每个中央处理器约 100 GB）。由于集群中符合这一要求的中央处理器很少，我们不得不使用 3 个内层交叉验证折（而非 10 个）来纳入该方法，如上文所述。

为实施此基准测试，我们在 R 中使用了 mlr3 框架，因为 Herrmann 等人使用的 mlr 包[7] 现已弃用。因此，我们为每种方法开发了相应的 mlr3学习器。这些 mlr3学习器实现了序列建模以及内层交叉验证。外层交叉验证和指标计算部分由 mlr3 管理。每种 JDR方法的学习器均可在一个 R包中获得。对于 BlockForest、IPF-LASSO 和 Priority-LASSO，我们改编了 Herrmann 等人已为 mlr 开发的学习器。该基准测试在本地计算集群上运行，该集群由来自不同架构的共计 2307 个处理器核心组成。我们使用 batchtools R包将 mlr3 框架与集群连接起来。我们开发了一个软件包，用于自动化这一过程的部分环节。更多详情请参见第 5 节。

4.7 解释

组学贡献值根据权重向量计算 $\boldsymbol{\beta}$，这些权重向量来自 fi在 JDR 方法的潜在空间上拟合的 Cox模型。更具体地说，在除潜在组学变量外还纳入临床变量时，我们对 $\boldsymbol{\beta}$: $\boldsymbol{\beta}_{\text{norm}} = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}$ 进行标准化，其中 S 是潜在组学变量的集合，我们计算了$\overline{\sum_{i \in S} \beta^2_{\text{norm},\, i}}$ 针对每个模型和每种癌症的每一折进行计算，从而估算组学数据对生存预测的贡献程度，结果介于 0 和 1 之间。随后，按癌症取均值，对这些值进行汇总。

对 JDR 方法进行自助抽样时，必须考虑各方法的不变性 [34]。事实上，JIVE 和主成分分析具有旋转不变性，这意味着 JIVE/PCA优化问题的一个解经过任意旋转后，仍然是同样好的解。为减轻这种效ff应对自助抽样样本的影响，我们将所有潜在空间和权重矩阵重新对齐至在原始数据集上拟合的模型 fi；具体做法是，针对每个自助抽样样本，确定其权重矩阵与原始权重矩阵之间的普鲁克分析旋转。RGCCA 具有符号不变性，这是旋转不变性的一种特殊情况。必要时乘以 -1，以确保自助抽样模型与原始模型的潜变量（及其对应权重）之间的相关性为正。LI-PCA、sJIVE 和 RGCCA 会从数据集中移除零方差变量；为将这一点纳入考虑，每个自助抽样样本的旋转和相关性仅针对自助抽样模型与原始模型共有的变量进行计算。在每次重抽样中，被移除变量的权重均设为 0。P 值根据自助抽样样本使用自助比率计算，即某个系fi数的均值与其标准差之比 [69]。在零假设下，自助比率服从标准正态分布。使用 Benjamini-Hochberg 程序对 P 值进行校正。该程序分别进行两次：一次针对所有与 JDR模型相关的 P 值（共 100 458 240 个），另一次针对所有与 Cox模型相关的 P 值（共 4014 个）。

组学贡献根据 99 次自助法重抽样计算，方法与此前计算组学与临床数据贡献时类似；这次则利用上述重新对齐方法，计算每个数据块的贡献。相关贡献见图 S41 至 S58。

自助法方法的超参数通过以下方式选定：针对每种癌症，选择在内层交叉验证中最常被评为最佳的参数组合（见表 S3-S7）。若两个或更多超参数组合并列，则选择 Cox 对数似然中位数最佳的组合（使用 Verweij 和 Van Houwelingen 近似法计算 [65]）。

在富集分析中，我们通过 R包 ClusterProfiler [71]，使用快速基因集富集分析（FGSEA） [70]，依据自助比率对变量进行排序，并将 MSigDB[44] 中的标志特征作为基因集。我们通过纳入所有通路、所有成分、所有方法和所有癌症来计算 q值 [72]，以控制错误发现率（FDR）。

5 代码可用性

运行基准测试所开发和使用的所有脚本，以及为本分析开发的软件包，均可在 GitHub 上获取：https://github.com/vlegoff/ MultiOBenchmark。本基准测试使用的所有方法和软件包均安装在 conda环境中。同一 GitHub 仓库还包含创建这些环境所需的全部 conda 配置文件（列有软件包及其版本）和安装脚本。

6 数据可用性

本分析使用的所有数据及产生的中间结果均可在 Zenodo 上获取：https://zenodo.org/ records/24402988。

表 1：用于运行各方法的 R 软件包和函数，以及预测函数的来源（自定义或由软件包提供）和各方法的错误百分比。调用次数对应于癌症数量（18）× 外层折数量（50）× 超参数组合数量 ×是否存在临床数据（如适用）。

Method name	Method type	Package::function	Prediction	Errors	Calls	Error rate
Kaplan-Meier [73]	Reference	survival::survfit	From package	0	900	0.00%
Cox clin. [2]		survival::coxph		0	900	0.00%
Cox L2 clin. [39]		glmnet::glmnet		0	900	0.00%
LASSO [39]	Naive			0	1,800	0.00%
ranger [37, 38]		ranger::ranger		0	1,800	0.00%
IPF-LASSO [3, 74]	Structured	ipflasso::cvr.adaptive.ipflasso		420	1,800	23.33%
Priority-LASSO [4]		prioritylasso::prioritylasso		74	1,800	4.11%
Priority-LASSO clin. fav. [4]				8	900	0.89%
BlockForest [6]		blockForest::blockfor		0	1,800	0.00%
PCA Joint [75]	PCA	ade4::dudi.pca		0	5,400	0.00%
PCA Indiv. [75]				0	5,400	0.00%
JIVE [19]	JDR - MF	r.jive::jive		3	27,000	0.01%
sJIVE [36]		sup.r.jive::sJIVE		5,109	81,000	6.31%
IntNMF [20]		IntNMF::nmf.mnnals	Custom	0	5,400	0.00%
iCluster [21, 22]		iClusterPlus::iCluster2	Custom	1,022	10,800	9.46%
MOFA [23, 24]		Python package mofapy2	MOTL [40]	0	5,400	0.00%
tPCA [32]	JDR - TF	tensorBSS::tPCA	From package	0	5,400	0.00%
tPCA+tFOBI [33]		tensorBSS::tFOBI		0	5,400	0.00%
RGCCA Unsup. [25, 26]	JDR - Cov	RGCCA::rgcca		693	10,800	6.42%
SGCCA Unsup. [29]				1,042	16,200	6.43%
RGCCA Sup. [25, 26, 34, 18]				665	10,800	6.16%
SGCCA Sup. [29, 34, 18]				1,088	16,200	6.72%
MCIA [30]		omicade4::mcia	Custom	10	5,400	0.19%

Table 2: **Best method by cancer** Best methods and their median performance are given by cancer, for both metrics, with and without clinical data. Median performance of Cox clin. is given as reference. The name of the method is in bold if it is a JDR method. Cancers are sorted from the lowest number of patients (LAML) to the highest (BRCA). A + as exponent indicates that for the specific metric and setting, the best method also outperforms Cox clin.

	Cox clin.		With clinical data				Without clinical data			
	C-Index	IBS	C-Index		IBS		C-Index		IBS	
Cancer	Value	Value	Method	Value	Method	Value	Method	Value	Method	Value
LAML	0.585	0.238	BlockForest^{+}	0.754	**iCluster** $^{+}$	0.179	BlockForest^{+}	0.707	ranger^{+}	0.185
ESCA	0.575	0.202	**RGCCA Unsup.**$^{+}$	0.592	IPF-LASSO^{+}	0.186	**RGCCA Sup.**	0.560	**MOFA+MOTL**$^{+}$ & EI-PCA^{+}	0.193
PAAD	0.674	0.172	**sJIVE**$^{+}$	0.711	**sJIVE** $^{+}$	0.158	**SGCCA Sup.**	0.654	EI-PCA^{+}	0.172
SARC	0.677	0.203	**RGCCA Sup.** $^{+}$	0.752	**RGCCA Sup.** $^{+}$	0.154	LASSO^{+}	0.702	**sJIVE** $^{+}$	0.168
LIHC	0.596	0.196	**RGCCA Sup.**$^{+}$	0.614	ranger^{+}	0.164	**tPCA**	0.588	ranger^{+}	0.160
KIRP	0.526	0.133	LI-PCA^{+}	0.693	BlockForest^{+}	0.122	**JIVE**$^{+}$	0.683	**sJIVE** $^{+}$	0.123
COAD	0.620	0.142	**SGCCA Sup.** $^{+}$	0.747	**RGCCA Unsup.** $^{+}$	0.123	**SGCCA Sup.** $^{+}$	0.712	**SGCCA Unsup.** $^{+}$	0.121
OV	0.611	0.168	Priority-LASSO^{+}	0.615	**IntNMF**$^{+}$	0.165	**IntNMF**$^{+}$	0.614	**IntNMF**$^{+}$	0.165
KIRC	0.752	0.138	LI-PCA^{+}	0.785	EI-PCA^{+}	0.133	**RGCCA Sup.**	0.709	EI-PCA	0.149
SKCM	0.645	0.176	LI-PCA^{+}	0.657	EI-PCA^{+}	0.172	ranger	0.574	LI-PCA	0.192
STAD	0.657	0.179	**SGCCA Unsup.**$^{+}$	0.660	**IntNMF** $^{+}$	0.174	**IntNMF**	0.559	**tPCA+tFOBI**	0.194
BLCA	0.659	0.194	Priority-LASSO^{+}	0.681	**tPCA** $^{+}$	0.178	**RGCCA Sup.**	0.640	**RGCCA Sup.**$^{+}$	0.191
UCEC	0.750	0.0903	**IntNMF**	0.741	**IntNMF**	0.0904	LI-PCA	0.677	ranger	0.0989
LUSC	0.551	0.207	LI-PCA^{+}	0.566	**sJIVE** $^{+}$	0.202	EI-PCA^{+}	0.565	**sJIVE**$^{+}$	0.204
LGG	0.733	0.124	LASSO^{+}	0.792	LASSO^{+}	0.100	LASSO^{+}	0.749	LASSO^{+}	0.110
LUAD	0.690	0.148	Priority-LASSO^{+}	0.691	Priority-LASSO clin. fav.$^{+}$	0.148	**MCIA**	0.570	**MCIA**	0.166
HNSC	0.545	0.203	**RGCCA Sup.** $^{+}$	0.632	**RGCCA Sup.** $^{+}$	0.188	**SGCCA Sup.** $^{+}$	0.615	**tPCA+tFOBI** $^{+}$	0.192
BRCA	0.670	0.0823	**sJIVE** $^{+}$	0.708	**sJIVE** $^{+}$	0.0774	LI-PCA^{+}	0.675	LI-PCA^{+}	0.0797

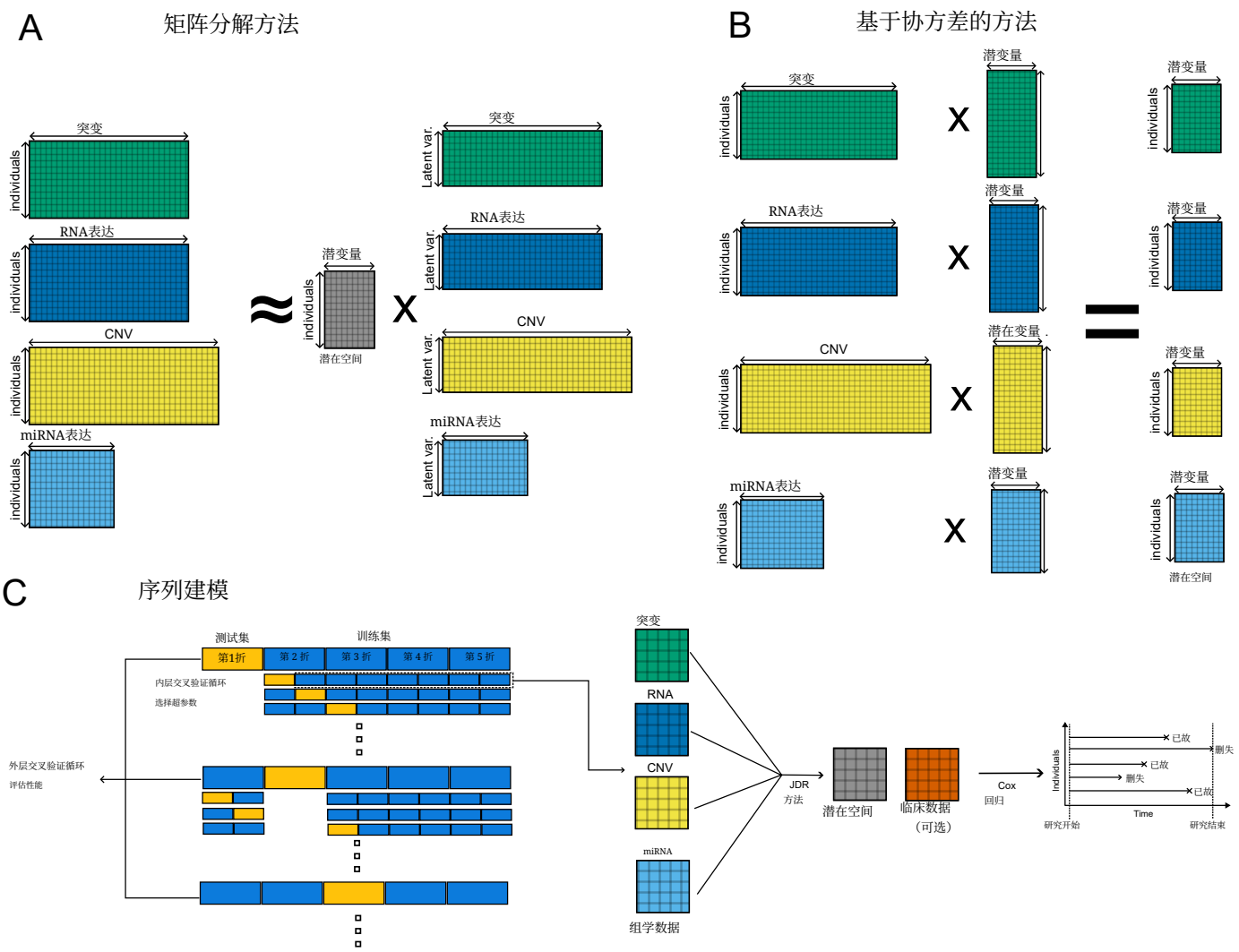

图 1：用于对 JDR 方法进行基准测试的 JDR 方法和交叉验证概览。以下方法的简化表示：A - 矩阵分解方法，B - 基于协方差的方法，以及 C - 生存数据的序列建模及其交叉验证。

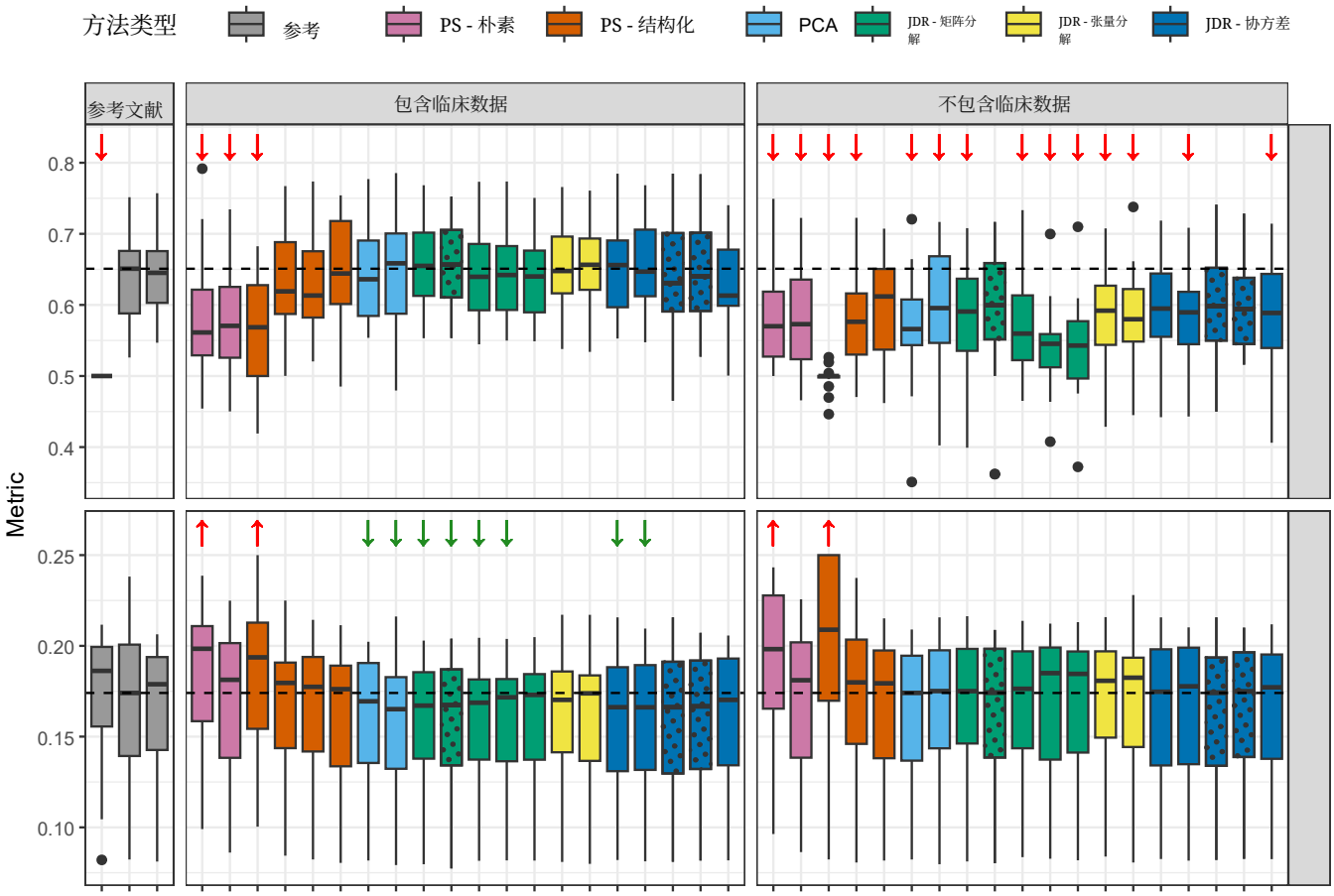

图 2：各测试方法在不同癌症中的表现。对于每种癌症，将 50 个外层折的结果取中位数进行汇总。顶行对应 C 指数（越高越好），底行对应 IBS（越低越好）。各列依次展示：不使用任何变量（Kaplan-Meier 估计量）或仅使用临床变量（Cox 模型和正则化 Cox 模型）的参照方法、使用临床数据的方法，以及不使用临床数据的方法。有监督 JDR 方法以点线表示。使用配对 Wilcoxon 检验，将各方法与仅使用临床数据的非正则化 Cox 模型基线进行比较。按癌症进行配对。使用 Benjamini-Hochberg 程序校正 P 值。当校正后的 P 值小于 0.05 时，会在相应箱线图上方标出指示效应方向的箭头。绿色箭头表示性能提升，红色箭头表示性能下降。PS：先前研究 [7]。

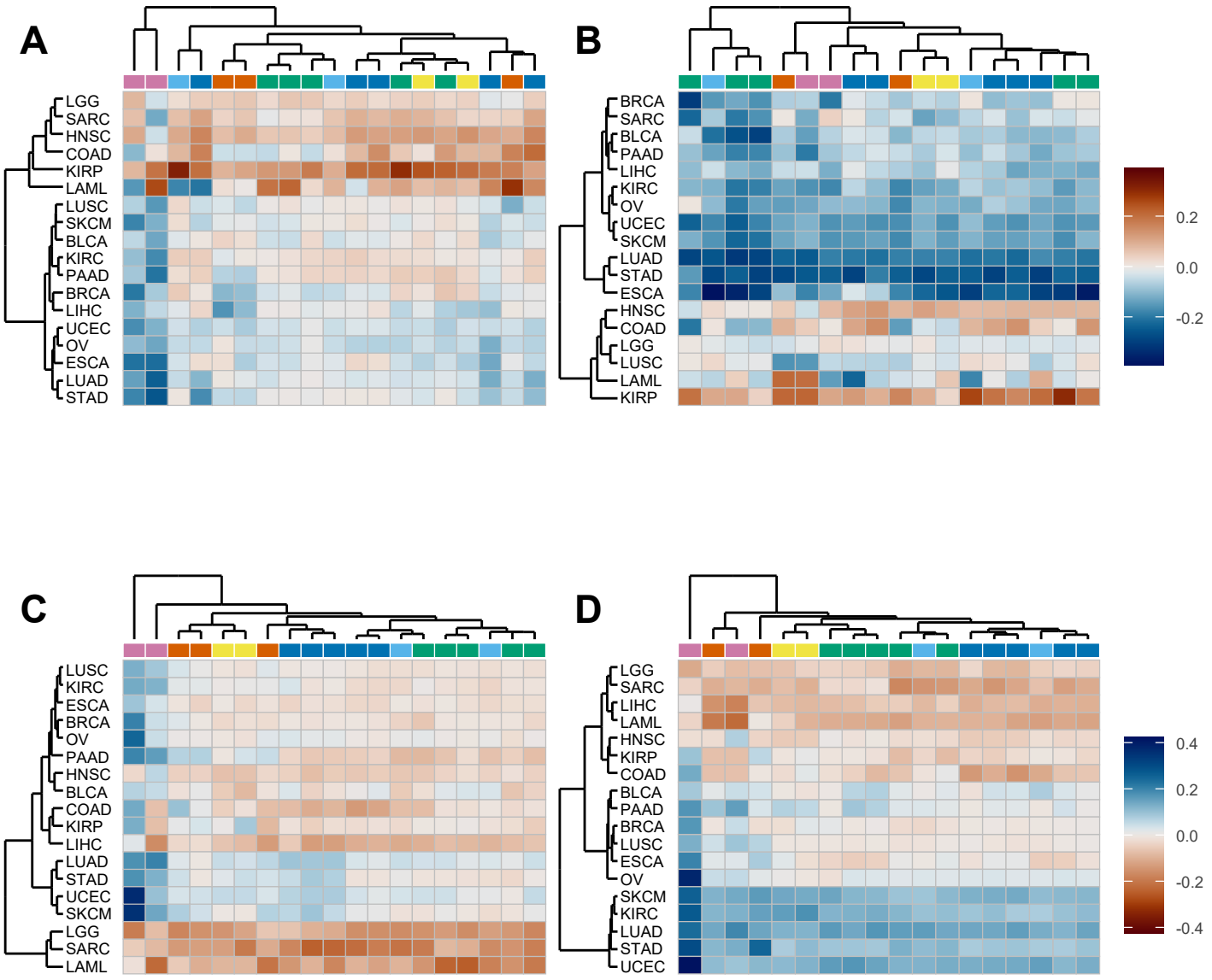

图3：各癌症中每种方法与 Cox临床模型之间相对中位性能的热图。A——使用临床数据的 Uno 一致性指数。B——不使用临床数据的 Uno 一致性指数。C——使用临床数据的 IBS。D——不使用临床数据的 IBS。红色表示性能提升，蓝色表示性能下降。热图上方的颜色对应于 fi图2中的方法类型。聚类采用 ward.D2连接法和欧氏距离。由于以下两个原因，IPF-LASSO 未纳入此分析：(i) 其性能持续低下，导致其他效应更难观察；这由 (ii) 其较高的错误率所解释，进而导致缺失的生存性能值被填补为默认值：C 指数为 0.5，IBS 为 0.25（更多详情请参见方法4.4）。

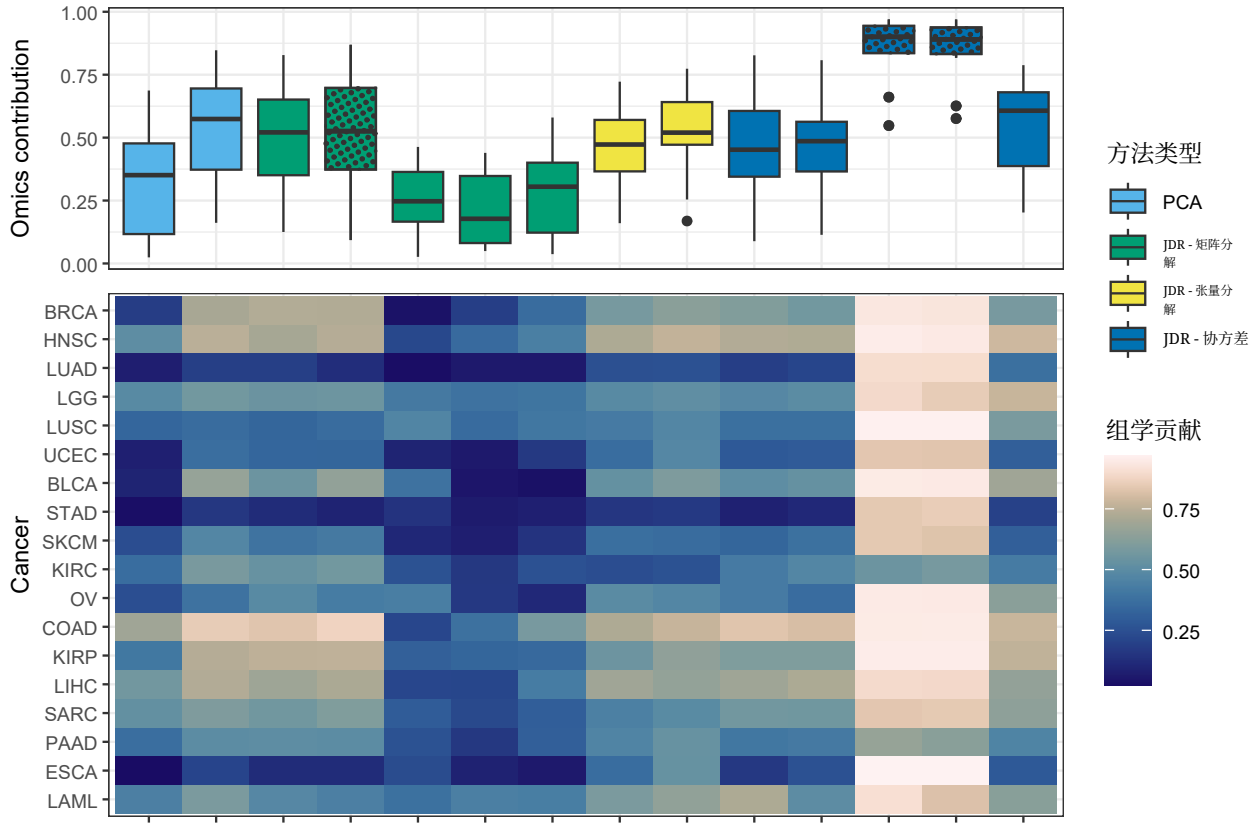

图4：存在临床数据时，组学数据对 Cox模型预测的贡献。上图：各癌症在 50 个外层折上的平均贡献箱线图。下图：各癌症和方法的平均贡献热图。监督方法以点状表示。

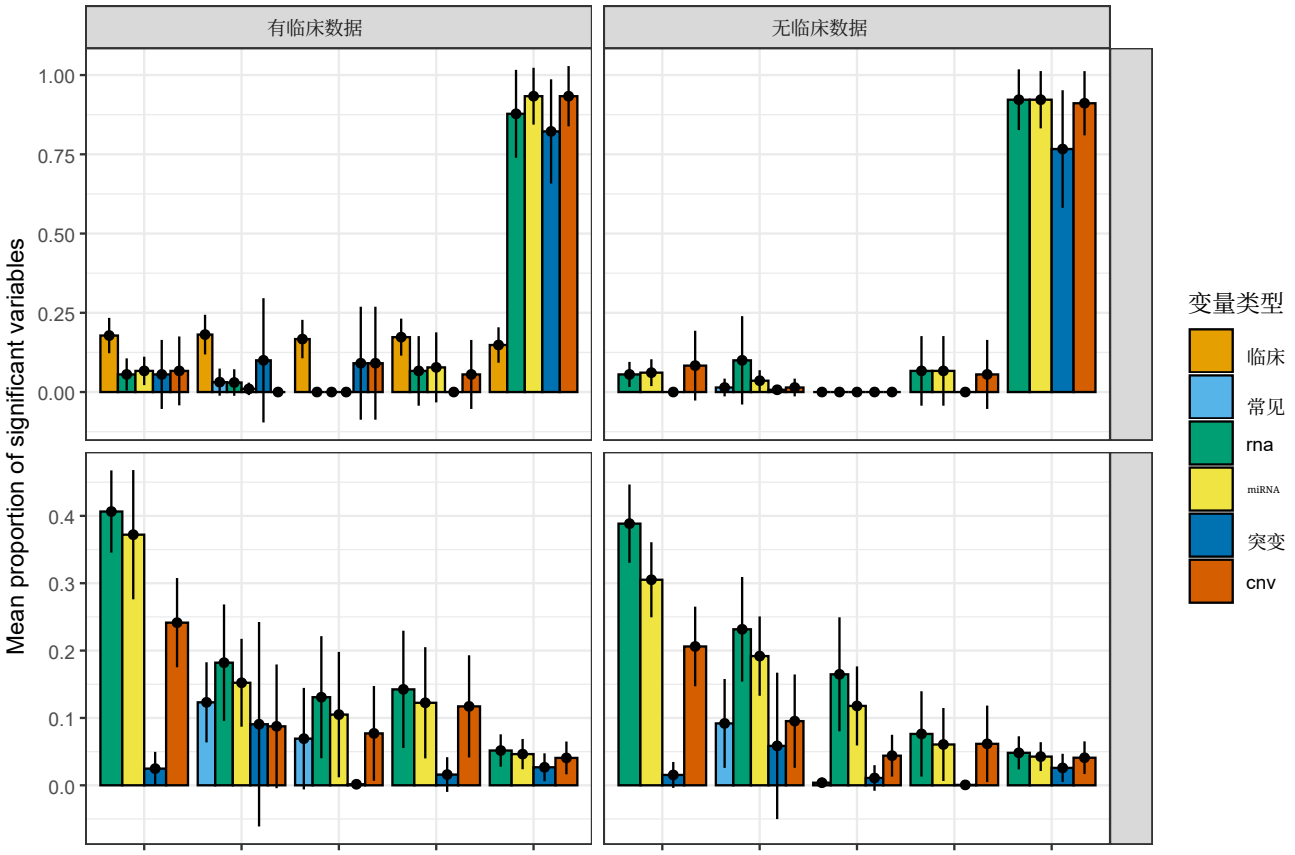

图5：采用自助法方法时，18种癌症中显著变量平均比例。上排和下排分别表示Cox模型和JDR模型中显著变量的比例，而左列和右列分别表示将临床数据添加到潜在空间和未添加临床数据到潜在空间的模型。误差条表示18种癌症均值的95%正态置信区间。

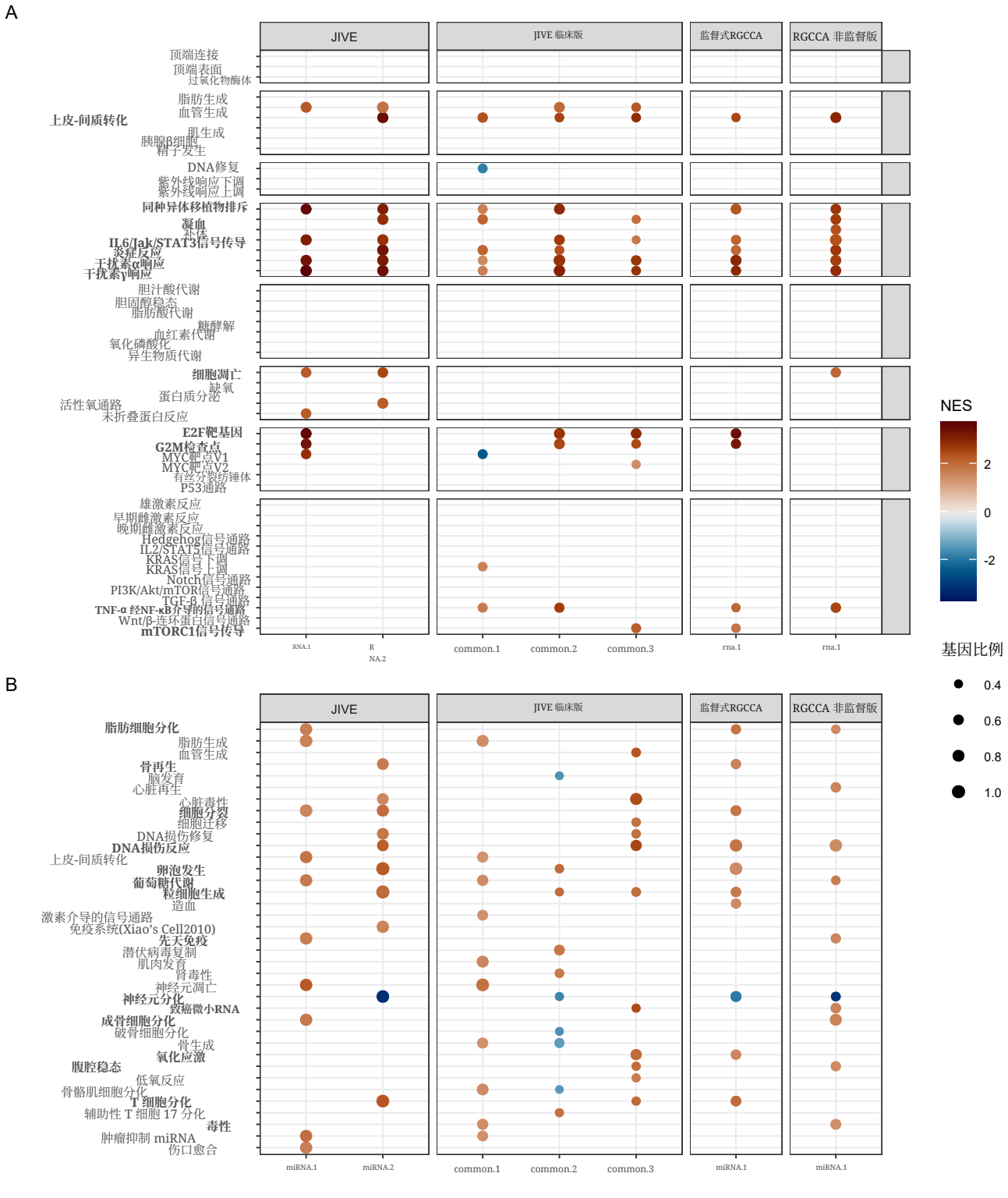

图6：LGG 中 JIVE、JIVE 临床版、监督式RGCCA 和 RGCCA 非监督版各成分富集的前10名基因集，按基因比例排序。A - RNA组分的 MSigDB 标志集富集结果。B - miRNA 组分的 TAM 2.0 功能基因集富集结果。仅保留显fi著富集的通路。block.n 表示这是该数据块的第 n 个成分。根据相关成分的 Cox模型权重调整标准化富集分数（NES）的符号，使正分数对应对生存有害的影响。仅对 Cox模型中显fi著的成分进行富集分析。粗体基因集是至少一个 RGCCA 模型和一个 JIVE 模型共同检出的。标志类别来自 Liberzon等人 [44]。NES：标准化富集分数；CC：细胞组分；develop.：发育；dmg：DNA损伤；prolif.：增殖。20

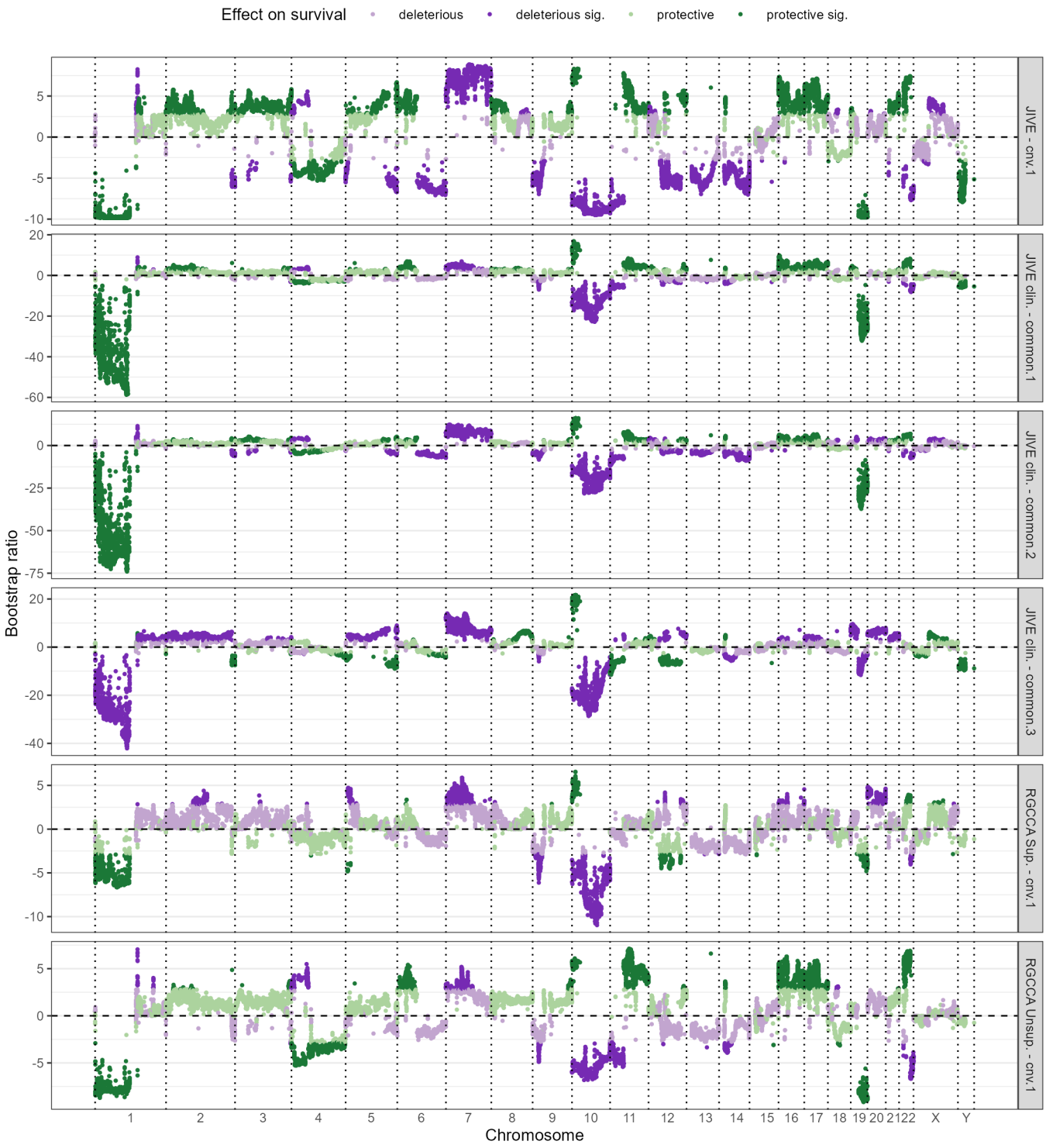

图 7：JIVE、JIVE 临床版、监督式RGCCA 和 RGCCA 非监督版的拷贝数变异组分自助比率。对于每个拷贝数变异变量，其自助比率的符号设为其均值的符号。随后，通过将该变量在低级别胶质瘤数据集中的拷贝数变异均值（主要事件为缺失或扩增）的符号、其自助比率的符号（该变量对整个成分的贡献）以及 Cox模型赋予该成分的权重（整个成分对生存的贡献）相乘，来确定其对生存的影响。最终，乘积为正（分别为负）表示有害（分别为保护性）影响。显著性采用 0.05 的 Q值阈值确定。

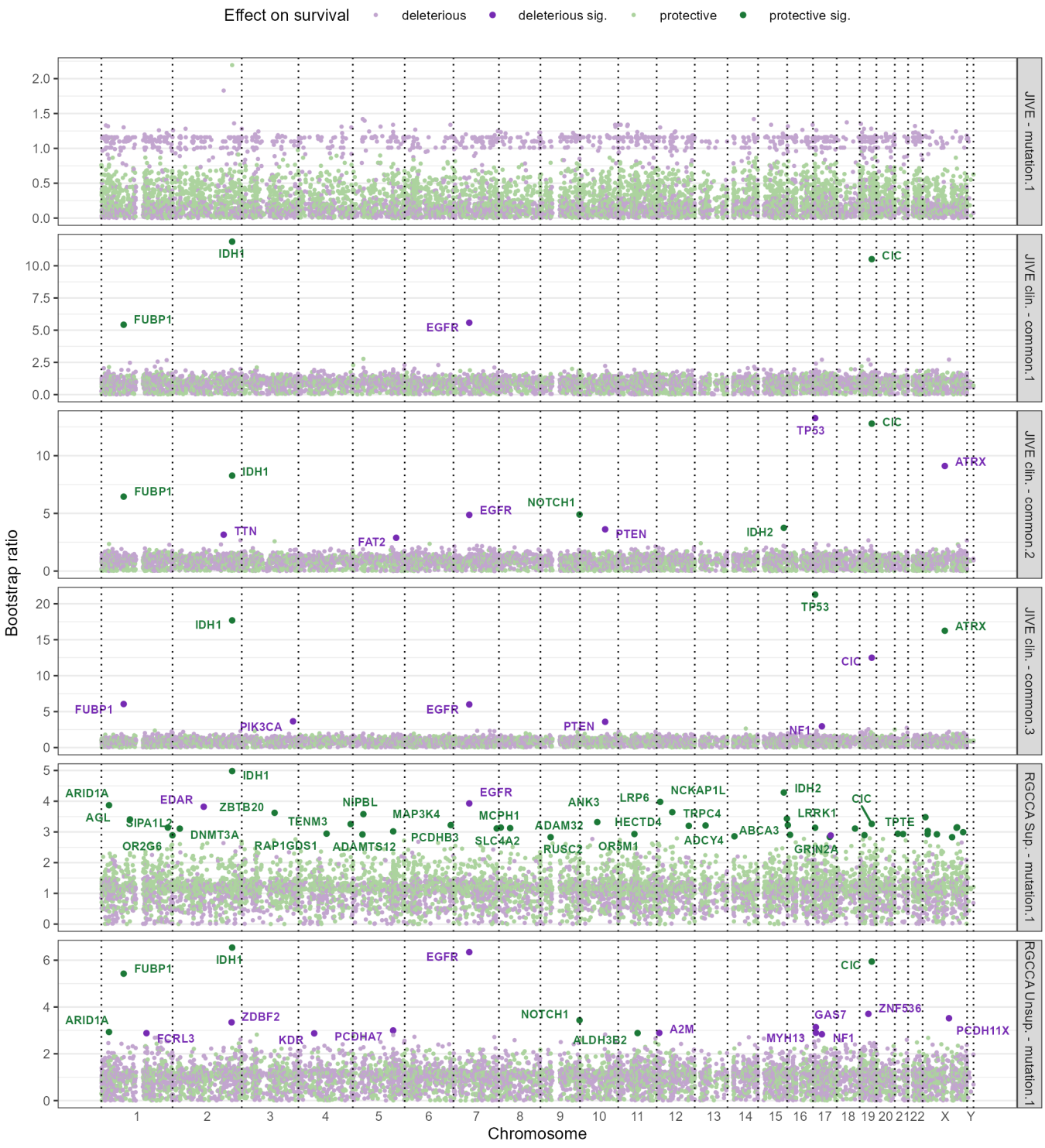

图 8：JIVE、JIVE 临床版、监督式RGCCA 和 RGCCA 非监督版的突变成分自助比率。图中显示每个基因的自助比率绝对值。随后，通过将其自助比率的符号（该变量对整个成分的贡献）与 Cox模型赋予该成分的权重（整个成分对生存的贡献）相乘，确定其对生存的影响。最终，乘积为正（分别为负）表示有害（分别为保护性）影响。显著性通过 Q值阈值 0.05 确定。

参考文献

[1] D. Tran、H. Nguyen、V.-D. Pham、P. Nguyen、H. Nguyen Luu、L. Minh Phan、C. Blair DeStefano、S.-C. Jim Yeung 和 T. Nguyen。“使用多组学整合和临床变量进行癌症生存预测的全面综述”。载于：《生物信息学简报》26.2（2025），bbaf150。doi: 10.1093/bib/bbaf150。[2] D. R. Cox。“部分似然”。载于：Biometrika 62.2（1975），第269–276页。doi: 10.1093/biomet/62.2.269。[3] A.-L. Boulesteix、R. De Bin、X. Jiang 和 M. Fuchs。“IPF-LASSO：基于多组学数据、带惩罚因子的整合式 L1惩罚回归，用于预测”。英文。载于：Computational and Mathematical Methods in Medicine 2017（2017），第7691937页。doi: 10.1155/2017/7691937。[4] S. Klau、V. Jurinovic、R. Hornung、T. Herold 和 A.-L. Boulesteix。“Priority-Lasso：使用多组学数据预测临床结局的简单层次化方法”。英文。载于：BMC Bioinformatics 19.1（2018），第322页。doi: 10.1186/s12859-018-2344-6。[5] J. Zobolas 和 H. Binder。CoxBoost：通过基于似然的提升方法，为单一生存终点或竞争风险建立 Cox模型。2025。[6] R. Hornung 和 M. N. Wright。“Block Forests：用于临床与组学协变量数据块的随机森林”。英文。载于：BMC Bioinformatics 20.1（2019），第358页。doi: 10.1186/s12859-019-2942-y。[7] M. Herrmann、P. Probst、R. Hornung、V. Jurinovic 和 A.-L. Boulesteix。“使用多组学数据的生存预测方法的大规模基准研究”。英文。载于：《生物信息学简报》22.3（2021），bbaa167。doi: 10.1093/bib/bbaa167。[8] J. Katzman、U. Shaham、J. Bates、A. Cloninger、T. Jiang 和 Y. Kluger。“DeepSurv：使用 Cox比例风险深度神经网络的个体化治疗推荐系统”。英文。载于：BMC Medical Research Methodology 18.1（2018）。arXiv:1606。00931 [stat.ML], 第24页。doi: 10.1186/s12874-018-0482-1。[9] T. Ching、X. Zhu 和 L. X. Garmire。“Cox-nnet：一种用于高通量组学数据预后预测的人工神经网络方法”。英文。载于：PLOS Computational Biology 14.4（2018），e1006076。doi: 10.1371/journal.pcbi.1006076。[10] D. Wissel、D. Rowson 和 V. Boeva。“多组学生存模型的系统比较揭示了普遍缺乏噪声鲁棒性”。载于：Cell Reports Methods 3.4（2023），第100461页。doi: 10 .1016 / j . crmeth.2023.100461。[11] H. Benkirane、Y. Pradat、S. Michiels 和 P.-H. Cournède。“CustOmics：一种用于多组学整合的多功能基于深度学习的策略”。英文。载于：PLOS Computational Biology 19.3（2023），e1010921。doi: 10. 1371/journal.pcbi.1010921。[12] D. Wissel、D. Rowson 和 V. Boeva。“基于层次自编码器的整合通过软模态选择提升多组学癌症生存模型的性能”。载于：（2021）。doi: 10.1101/2021.09.16. 460589。[13] M. Pierre-Jean、J.-F. Deleuze、E. Le Floch 和 F. Mauger。“多组学数据整合中13种无监督方法的聚类与变量选择评估”。英文。载于：《生物信息学简报》21.6（2020），第2011–2030页。doi: 10.1093/bib/bbz138。[14] L. Cantini、P. Zakeri、C. Hernandez、A. Naldi、D. Thieffry、E. Remy 和 A. Baudot。“用于癌症研究的联合多组学降维方法基准测试”。英文。载于：《自然·通讯》12.1（2021）。编号：1，第124页。doi: 10.1038/s41467-020-20430-7。[15] E. Bair 和 R. Tibshirani。“通过基因表达数据预测患者生存的半监督方法”。英文。载于：PLOS Biology 2.4（2004）。编辑：Todd Golub，e108。doi: 10.1371/journal.pbio.0020108。[16] A. Kaplan 和 E. F. Lock。“通过对多个分子数据源进行降维预测患者生存”。载于：Cancer Informatics 16（2017），第1176935117718517页。doi: 10.1177/1176935117718517。[17] A. Tenenhaus 和 M. Tenenhaus。“用于多块或多组数据分析的正则化广义典型相关分析”。载于：European Journal of Operational Research 238.2（2014），第391–403页。doi: 10.1016/j.ejor.2014.01.008。[18] H. Chegraoui、V. Guillemot、A. Rebei、A. Gloaguen、J. Grill、C. Philippe 和 V. Frouin。“整合多组学与先验知识：Graphnet惩罚影响的研究”。载于：Bioinformatics 39.8（2023），btad454。doi: 10.1093/bioinformatics/btad454。

[19] E. F. Lock、K. A. Hoadley、J. Marron 和 A. B. Nobel。“用于多个数据类型整合分析的联合与个体变异解释（JIVE）”。载于：《应用统计年鉴》7.1（2013），第523–542页。doi: 10.1214/12-AOAS597。[20] P. Chalise 和 B. L. Fridley。“基于非负矩阵分解算法的多层组学数据整合聚类”。英文。载于：PloS One 12.5（2017），e0176278。doi: 10.1371/journal.pone.0176278。[21] R. Shen、A. B. Olshen 和 M. Ladanyi。“使用联合潜变量模型整合多个基因组数据类型，并应用于乳腺癌和肺癌亚型分析”。英文。载于：Bioinformatics 25.22（2009），第2906–2912页。doi: 10.1093/bioinformatics/btp543。[22] R. Shen、Q. Mo、N. Schultz、V. E. Seshan、A. B. Olshen、J. Huse、M. Ladanyi 和 C. Sander。“使用 iCluster 发现胶质母细胞瘤中的整合性亚型”。载于：PLOS ONE 7.4（2012），e35236。doi: 10.1371/ journal.pone.0035236。[23] R. Argelaguet、B. Velten、D. Arnol、S. Dietrich、T. Zenz、J. C. Marioni、W. Huber、F. Buettner 和 O. Stegle。多组学因子分析——一种用于无监督整合多组学数据集的框架。英文。预印本。Bioinformatics，2017。doi: 10.1101/217554。[24] R. Argelaguet、D. Arnol、D. Bredikhin、Y. Deloro、B. Velten、J. C. Marioni 和 O. Stegle。“MOFA+：用于全面整合多模态单细胞数据的统计框架”。载于：《基因组生物学》21.1（2020），第111页。doi: 10.1186/s13059-020-02015-1。[25] A. Tenenhaus 和 M. Tenenhaus。“正则化广义典型相关分析”。英文。载于：Psychometrika 76.2（2011），第257–284页。doi: 10.1007/s11336-011-9206-8。[26] M. Tenenhaus、A. Tenenhaus 和 P. J. F. Groenen。“正则化广义典型相关分析：序贯多块成分方法的框架”。英文。载于：Psychometrika 82.3（2017），第737–777页。doi: 10.1007/s11336-017-9573-x。[27] H. Hotelling。“两组变量之间的关系”。载于：Biometrika 28.3-4（1936），第321–377页。doi: 10. 1093/biomet/28.3-4.321。[28] S. Wold、M. Sjöström 和 L. Eriksson。“PLS-regression：化学计量学的基础工具”。载于：Chemometrics and Intelligent Laboratory Systems。PLS Methods 58.2（2001），第109–130页。doi: 10.1016/S0169-7439(01)00155-1。[29] A. Tenenhaus、C. Philippe、V. Guillemot、K.-A. Le Cao、J. Grill 和 V. Frouin。“广义典型相关分析中的变量选择”。载于：Biostatistics 15.3（2014），第569–583页。doi: 10.1093/biostatistics/kxu001。[30] C. Meng、B. Kuster、A. C. Culhane 和 A. M. Gholami。“一种整合多组学数据集的多变量方法”。英文。载于：BMC bioinformatics 15（2014），第162页。doi: 10.1186/1471-2105-15-162。[31] P. M. Kroonenberg 和 R. Lombardo。“非对称对应分析：分析具有依赖结构的列联表的工具”。载于：Multivariate Behavioral Research 34.3（1999）。_电子预印本：https://doi.org/10.1207/S15327906MBR3403_4，第367–396页。doi: 10.1207/S15327906MBR3403_4。[32] J. Virta、S. Taskinen 和 K. Nordhausen。“将完整张量独立成分分析应用于功能性磁共振成像数据”。载于：2016 IEEE Signal Processing in Medicine and Biology Symposium（SPMB）（2016）。doi: 10.1109/SPMB.2016.7846858。[33] J. Virta、B. Li、K. Nordhausen 和 H. Oja。“张量值数据的独立成分分析”。载于：Journal of Multivariate Analysis 162（2017），第172–192页。doi: 10.1016/j.jmva.2017.09.008。[34] P. Bastien。“基于偏差残差的删失数据高维情境下 PLS回归”。载于：Chemometrics and Intelligent Laboratory Systems。《化学计量学大会“CHIMIOMETRIE 2006”精选论文，法国巴黎，2006年11月30日–12月1日》91.1（2008），第78–86页。doi: 10. 1016/j.chemolab.2007.09.009。[35] P. Salguero、A. Buendía-Galera 和 S. Tarazona。“Coxmos：面向高维和多组学数据的可解释生存模型”。英文。载于：BioData Mining 19.1（2026），第53页。doi: 10.1186/s13040-026-00559-9。[36] E. F. Palzer、C. H. Wendt、R. P. Bowler、C. P. Hersh、S. E. Safo 和 E. F. Lock。“sJIVE：监督式联合与个体变异解释”。载于：Computational statistics & data analysis 175（2022），第107547页。doi: 10.1016/j.csda.2022.107547。[37] M. N. Wright 和 A. Ziegler。“ranger：在 C++ 和 R 中快速实现用于高维数据的随机森林”。英文。载于：《统计软件期刊》77（2017），第1–17页。doi: 10.18637/jss.v077.i01。

[38] H. Ishwaran、U. B. Kogalur、E. H. Blackstone 和 M. S. Lauer。《随机生存森林》。英文版。载于：《应用统计年鉴》2.3（2008），第 841–860 页。doi：10.1214/08-AOAS169。[39] N. Simon、J. H. Friedman、T. Hastie 和 R. Tibshirani。《通过坐标下降求解 Cox 比例风险模型的正则化路径》。英文版。载于：《统计软件期刊》39（2011），第 1–13 页。doi：10.18637/ jss.v039.i05。[40] D. P. Hirst、M. Térézol、L. Cantini、P. Villoutreix、M. Vignes 和 A. Baudot。《MOTL：利用迁移学习增强多组学矩阵分解》。载于：《基因组生物学》26（2025），第 224 页。doi：10.1186/s13059- 025-03675-7。[41] H. Uno、T. Cai、M. J. Pencina、R. B. D' Agostino 和 L. J. Wei。《关于使用删失生存数据评估风险预测程序整体适 adequacy 的 C统计量》。英文版。载于：《医学统计学》30.10（2011），第 1105–1117 页。doi：10.1002/sim.4154。[42] E. Graf、C. Schmoor、W. Sauerbrei 和 M. Schumacher。《生存数据预后分类方案的评估与比较》。英文版。载于：《医学统计学》18.17-18（1999）。_电子预印本：https://onlinelibrary.wiley

.com/doi/p

0258%2819990915/30%2918%3A17/18%3C2529%3A%3AAID-SIM274%3E3.0.CO%3B2-5, 第2529–2545页。doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5。[43] D. Wissel、N. Janakarajan、A. Grover、E. Toniato、M. R. Martínez 和 V. Boeva。“SurvBoard：多组学癌症生存模型的标准化基准测试”。英文版。载于：《生物信息学简报》26.5（2025），bbaf521。doi: 10.1093/bib/bbaf521。[44] A. Liberzon、C. Birger、H. Thorvaldsdóttir、M. Ghandi、J. P. Mesirov 和 P. Tamayo。“分子特征数据库（MSigDB）标志性基因集合集”。载于：《细胞系统》1.6（2015），第417–425页。doi: 10.1016/j.cels.2015.12.004。[45] J. Li、X. Han、Y. Wan、S. Zhang、Y. Zhao、R. Fan、Q. Cui 和 Y. Zhou。“TAM 2.0：微小RNA集分析工具”。英文版。载于：《核酸研究》46.W1（2018），W180–W185。doi: 10.1093/nar/gky509。[46] A. J. West、V. Tsui、S. S. Stylli、H. P. Nguyen、A. P. Morokoff、A. H. Kaye 和 R. B. Luwor。“白细胞介素-6-STAT3信号传导在胶质母细胞瘤中的作用”。载于：《肿瘤学快报》16.4（2018），第4095–4104页。doi: 10.3892/ol.2018.9227。[47] Y. Li、F. Jiang、S. Zhu、H. Jia 和 C. Li。“STAT3通过调节STAT1、FOXO1和MYC的表达，推动低级别胶质瘤的恶性进展”。载于：《分子生物科学前沿》11（2024），第1419072页。doi: 10.3389/fmolb.2024.1419072。[48] M. S. Y. Tan、E. Sandanaraj、Y. K. Chong、S. W. Lim、L. W. H. Koh、W. H. Ng、N. S. Tan、P. Tan、B. T. Ang 和 C. Tang。“基于STAT3的基因特征对胶质瘤患者进行分层，以指导靶向治疗”。英文版。载于：《自然·通讯》10.1（2019），第3601页。doi: 10.1038/s41467-019-11614-x。[49] M. Yi、T. Li、M. Niu、H. Zhang、Y. Wu、K. Wu 和 Z. Dai。“靶向细胞因子和趋化因子信号通路进行癌症治疗”。英文版。载于：《信号转导与靶向治疗》9.1（2024），第176页。doi: 10.1038/s41392-024-01868-3。[50] D. E. Johnson、R. A. O'Keefe 和 J. R. Grandis。“靶向癌症中的IL-6/JAK/STAT3信号轴”。英文版。载于：《自然综述：临床肿瘤学》15.4（2018），第234–248页。doi: 10.1038/nrclinonc.2018.8。[51] C. Wen、Q. Lan、Y. Wang、Y. Ni、A. S. T. Wong 和 D. Liu。“健康与疾病中的干扰素信号通路”。载于：《分子生物医学》6（2025），第135页。doi: 10.1186/s43556-025-00381-5。[52] A. Segerman、M. Niklasson、C. Haglund、T. Bergström、M. Jarvius、Y. Xie、A. Westermark、D. Sönmez、A. Hermansson、M. Kastemar、Z. Naimaie-Ali、F. Nyberg、M. Berglund、M. Sundström、G. Hesselager、L. Uhrbom、M. Gustafsson、R. Larsson、M. Fryknäs、B. Segerman 和 B. Westermark。“胶质瘤起始细胞之间药物和辐射反应的克隆变异与前神经型-间质型转化相关”。载于：《细胞报告》17.11（2016），第2994–3009页。doi: 10.1016/j.celrep.2016.11.056。[53] C. Xu、P. Hou、X. Li、M. Xiao、Z. Zhang、Z. Li、J. Xu、G. Liu、Y. Tan 和 C. Fang。“胶质母细胞瘤分子表型的全面理解：分类、特征与转化”。载于：《癌症生物学与医学》21.5（2024），第363–381页。doi: 10.20892/j.issn.2095-3941.2023.0510。[54] P. Y. Wen 和 D. A. Reardon。“胶质瘤诊断、分类与治疗进展”。英文版。载于：《自然综述：神经病学》12.2（2016），第69–70页。doi: 10.1038/nrneurol.2015.242。

[55] D. Stichel、A. Ebrahimi、D. Reuss、D. Schrimpf、T. Ono、M. Shirahata、G. Reifenberger、M. Weller、D. Hänggi、W. Wick、C. Herold-Mende、M. Westphal、S. Brandner、S. M. Pfister、D. Capper、F. Sahm 和 A. von Deimling。“脑肿瘤中EGFR扩增、7号染色体增益与10号染色体缺失联合出现以及TERT启动子突变的分布，及其将IDH野生型星形细胞瘤重新分类为胶质母细胞瘤的潜力”。英文版。载于：《神经病理学学报》136.5（2018），第793–803页。doi: 10.1007/s00401-018-1905-0。[56] D. J. Brat、K. Aldape、H. Colman、E. C. Holland、D. N. Louis、R. B. Jenkins、B. K. Kleinschmidt-DeMasters、A. Perry、G. Reifenberger、R. Stupp、A. von Deimling 和 M. Weller。“cIMPACT-NOW更新3：‘弥漫性星形细胞胶质瘤，IDH野生型，具有胶质母细胞瘤分子特征，WHO IV级’的推荐诊断标准”。英文版。载于：《神经病理学学报》136.5（2018），第805–810页。doi: 10.1007/s00401-018-1913-0。[57] D. Hanahan。“癌症标志：新维度”。载于：《癌症发现》12.1（2022），第31–46页。doi: 10.1158/2159-8290.CD-21-1059。[58] D. Noack、J. Wach、A. Barrantes-Freer、N. H. Nicolay、E. Güresir 和 C. Seidel。“低级别和高级别胶质瘤中的纯合CDKN2A/B缺失：个体患者数据的荟萃分析及p16免疫组织化学检测的预测价值”。载于：《神经病理学通讯》12（2024），第180页。doi: 10.1186/s40478-024-01889-7。[59] I. Gaynanova 和 G. Li。“多视图数据的结构学习与整合分解”。载于：《生物统计学》75.4（2019），第1121–1132页。doi: 10.1111/biom.13108。[60] S. Yi、R. K. W. Wong 和 I. Gaynanova。“多视图数据整合的分层核范数惩罚”。载于：《生物统计学》79.4（2023），第2933–2946页。doi: 10.1111/biom.13893。[61] T. Xu、N. Su、L. Liu、J. Zhang、H. Wang、W. Zhang、J. Gui、K. Yu、J. Li 和 T. D. Le。“miRBaseC-onverter：用于转换和检索不同版本miRBase中miRNA名称、登录号、序列及家族信息的R/Bioconductor软件包”。载于：《BMC生物信息学》19.Suppl 19（2018），第514页。doi: 10.1186/s12859-018-2531-5。[62] I. Jung、M. Kim、S. Rhee、S. Lim 和 S. Kim。“MONTI：用于基因层面整合分析的多组学非负张量分解框架”。载于：《遗传学前沿》12（2021）。doi: 10.3389/fgene. 2021.682841。[63] Y. Ru、K. J. Kechris、B. Tabakoff、P. Hoffman、R. A. Radcliffe、R. Bowler、S. Mahaffey、S. Rossi、G. A. Calin、L. Bemis 和 D. Theodorescu。“multiMiR R包和数据库：整合微小RNA—靶标相互作用及其疾病和药物关联”。载于：《核酸研究》42.17（2014），e133–e133。doi: 10.1093/nar/gku631。[64] Q. Mo、S. Wang、V. E. Seshan、A. B. Olshen、N. Schultz、C. Sander、R. S. Powers、M. Ladanyi 和 R. Shen。“整合癌症基因组数据中的模式发现与癌症基因识别”。英文版。载于：《美国国家科学院院刊》110.11（2013），第4245–4250页。doi: 10.1073/pnas.1208949110。[65] P. J. Verweij 和 H. C. Van Houwelingen。“生存分析中的交叉验证”。英文版。载于：《医学统计学》12.24（1993），第2305–2314页。doi: 10.1002/sim.4780122407。[66] H. Kvamme 和 Ø. Borgan。“行政删失下的Brier评分：问题与解决方案”。载于：《机器学习研究期刊》24.2（2023），第1–26页。[67] R. Sonabend、J. Zobolas、R. B. Bin、P. Kopper、L. Burk 和 A. Bender。《使用平方损失和对数损失检验生存模型外部验证中的边际适当性》。arXiv:2212.05260 [math.ST]版本：3。2025。doi: 10.48550/arXiv.2212.05260。[68] T. Hastie、J. Qian 和 K. Tay。《“glmnet”简介》。英文版。2026。网址：https://glmnet.stanford.edu/ articles/glmnet.html（访问日期：2026年11月23日）。[69] A. R. McIntosh 和 N. J. Lobaugh。“神经影像数据的偏最小二乘分析：应用与进展”。英文版。载于：《神经影像》23 Suppl 1（2004），S250–263。doi: 10.1016/j.neuroimage.2004.07.020。[70] G. Korotkevich、V. Sukhov、N. Budin、B. Shpak、M. N. Artyomov 和 A. Sergushichev。《快速基因集富集分析》。英文版。页码：060012；栏目：最新成果。2021。doi: 10.1101/060012。[71] T. Wu、E. Hu、S. Xu、M. Chen、P. Guo、Z. Dai、T. Feng、L. Zhou、W. Tang、L. Zhan、X. Fu、S. Liu、X. Bo 和 G. Yu。“clusterProfiler 4.0：用于解释组学数据的通用富集工具”。载于：《创新》2.3（2021），第100141页。doi: 10.1016/j.xinn.2021.100141。[72] J. D. Storey。“错误发现率的直接方法”。载于：《英国皇家统计学会杂志B辑：统计方法学》64.3（2002），第479–498页。doi: 10.1111/1467-9868.00346。

[73] E. L. Kaplan 和 P. Meier。《基于不完整观测的非参数估计》。载于：《美国统计学会会刊》53.282（1958），第 457–481 页。doi：10.1080/01621459.1958.10501452。[74] G. Schulze。《基于多组学数据的临床结局预测》。英语。硕士论文。2017 年。doi：10.5282/ubm/ epub.59092。[75] F. Husson、S. Le 和 J. Pagès。《使用 R 进行探索性多变量分析：实例讲解》。第 2 版。纽约：Chapman and Hall/CRC，2017 年。doi：10.1201/b21874。