arXiv:2302.03991v1 [q-bio.GN] 8 Feb 2023

部分缺失多组学协变量数据的预测方法：文献综述与实证比较研究

Roman Hornung[1] ,*，Frederik Ludwigs[1],Jonas Hagenberg[1,2,3,4],和 Anne-Laure Boulesteix [1]

[1]医学信息处理、生物统计与流行病学研究所，慕尼黑大学，德国慕尼黑[2]精神病学转化研究系，马克斯·普朗克精神病学研究所，德国慕尼黑[3]计算生物学研究所，亥姆霍兹慕尼黑中心，德国诺伊赫贝格[4]国际马克斯·普朗克精神病学转化研究学院，德国慕尼黑[*]通讯作者：Roman Hornung，hornung@ibe.med.uni-muenchen.de

2023年2月9日

摘要

近年来，随着组学数据可用性的提高，产生了更多多组学数据，即由多种类型的高维分子数据构成的数据，例如基因组、转录组或蛋白质组数据，且这些数据均来自同一批患者。这类数据适合作为自动结局预测中的协变量，因为每种组学类型都可能提供独特的信息，相较于仅使用一种组学数据类型，或许能够提升预测效果。然而，在训练数据以及需要应用自动预测规则的数据（即测试数据）中，不同类型的组学数据往往并非对所有患者都可用。我们将这类数据称为分块缺失的多组学数据。首先，我们对适用于此类数据的现有预测方法进行文献综述。随后，我们利用由 13 个公开可用的多组学数据集组成的数据集，比较其中若干种方法在不同块状缺失模式下的预测性能。最后，我们讨论这项实证比较研究的结果，并得出一些初步结论。

1 引言

各种类型组学数据的生成正变得越来越迅速且成本效益越来越高。因此，可获取的所谓多组学数据也越来越多，即针对同一批患者测量的多种类型高维分子数据，例如基因组、转录组或蛋白质组数据。过去几年中，人们开发了多种利用这些数据预测患者结局的方法（详见 Hornung和Wright（2019）的文献综述）。然而，近期人们开始质疑，相较于简单的临床模型，使用多组学数据是否确有益处（Herrmann 等，2020）。

无论多组学数据对于预测是否有用，出于各种原因，针对同一预测问题使用的不同来源的多组学数据，往往并不包含完全相同的数据类型。最重要的是，需要进行预测的数据，即测试数据，通常与用于制定预测规则的数据，即训练数据，所包含的数据类型并不相同（Krautenbacher 等人，2019年）。训练数据也常由来自不同来源（例如不同医院）的子集组成，而这些子集通常包含不同组合的组学数据

类型，如前所述。若考虑至少在一个观测值中可用的所有组学数据类型，并将不同观测值中不可用的数据类型视为缺失数据，则可以将所有观测值对应的数据拼接起来，形成一个部分缺失的大型数据集（示例见图1）。下文将这种形式拼接而成的数据称为“分块缺失的多组学数据”，其中不同的组学数据类型称为“数据块”。为简便起见，数据集中共享相同观测数据类型组合的观测值群体称为“子集”。请注意，实践中除组学数据外，通常还可获得临床协变量（例如年龄或疾病分期），其中往往包含大量预测信息。本文假设所有患者始终都能获得临床协变量。下文将这类数据称为“临床数据块”。

我们首先介绍适用于分块缺失多组学数据的当前最先进预测方法，随后通过一项实证基准比较研究，评估其中若干方法的表现。在本研究中，我们使用了大量公开可用的多组学数据集，并人为生成了缺失数据。我们将 TP53 突变的有无作为响应变量。尽管建立预测 TP53 突变是否存在的模型在临床上没有意义，但这些突变与不良临床结局有关（Wang 和 Sun，2017年）。基于此，我们在基准研究中将 TP53 作为表型结局的替代指标。

2 现有的部分缺失多组学协变量数据预测方法

能够处理分块缺失数据的现有方法大致可分为三类：朴素方法、用于预测建模前的插补方法，以及在预测建模过程中处理缺失结构的方法。下文将介绍这些方法，第3.2.1节将更详细地阐述纳入实证比较研究的方法。

2.1 朴素方法

最简单的方法是完整案例法。训练在所有具有完整观测的观测值子集上进行。为了增加训练观测值的数量，仅使用测试数据中也包含的数据块是合理的。该方法的主要缺点是会导致丢弃训练观测值中可能很重要的一部分。在极端情况下，甚至可能不存在任何具有完整数据的患者。例如，一个研究中心可能只收集了转录组数据和蛋白质组数据，而另一个研究中心只收集了转录组数据和基因组数据；对于转录组、蛋白质组和基因组数据的组合而言，没有任何患者属于完整案例。

另一种朴素方法是单块方法，即仅使用训练数据和测试数据中均可获得的一个数据块来训练模型。在这种情况下，使用交叉验证预测性能或袋外预测性能最高的数据块是高效的。该方法的主要缺点是会导致丢弃预测变量中可能很重要的一部分。完整案例法和朴素方法具有一个共同优势：在训练时，可以使用任何适用于高维协变量数据的预测方法。

2.2 基于插补的方法

最简单的情况下，可以像传统的缺失数据插补那样，使用插补方法填补缺失值。这里可以使用 missForest（Stekhoven 和 B¨uhlmann，2011年）或半监督学习方法（Lan 等人，2022年）。TOBMI（Dong 等人，2019年）也采用类似的方法，但利用块结构来定义观测值之间的相似性。与后一类方法以及上一小节介绍的方法一样，可以使用任何适用于高维协变量的预测方法。

有几种方法使用矩阵补全算法来填补缺失值。Thung 等（2014年）首先通过删除噪声协变量和冗余协变量，并选取能够较好代表测试数据中观测值的观测值，来缩减训练数据的规模。其次，从测试数据中删除相同的协变量后，对合并后的训练数据和测试数据中的缺失数据进行插补。Cai 等（2016年）假设分块缺失数据对应的矩阵近似为低秩矩阵，并提出了一种结构化矩阵补全算法。Linder 和 Zhang（2019年）以此为基础，进一步允许单个协变量存在缺失值，而不局限于分块缺失结构。

其他插补方法包括 Zhu 等（2020）提出的方法，该方法假设数据是来自指数族的实现值，并在考虑组学类型内部及彼此之间关系的基础上估计缺失值的参数。在多重分块插补方法（Xue 和 Qu，2021）中，数据根据缺失模式划分为互不重叠的组，并对缺失值进行多次插补。随后，利用插补后的数据集生成估计方程，并将不同的估计量结合起来，得出一个预测结果。MI-GAN（Dai 等，2021）针对不同的缺失模式学习生成对抗网络，以对缺失值进行多重插补。多块数据驱动稀疏偏最小二乘（mdd-sPLS）（Lorenzo 等人，2019）首先对每个数据块分别进行偏最小二乘（PLS）回归。然后结合各个 PLS 模型的预测变量，对缺失值进行均值插补，再以期望最大化的方式应用 PLS 程序，从而获得更好的插补结果。

一些方法利用潜在结构进行插补。Zhang 等（2020）估计一个因子模型，然后利用该模型插补缺失值。Yang 等（2020）首先使用自编码器学习每种数据类型 A 与其余数据类型之间的成对映射。其次，对这些成对映射进行正则化，使其彼此保持一致；将每个成对映射应用于 A 插补缺失值；最后，对使用不同成对映射获得的插补结果取平均值。需要注意的是，尽管该方法并非专为多组学数据设计，但仍可用于此类数据。

Hieke 等（2016）构建不同的惩罚回归模型，并将一个模型的预测结果用作下一个模型的偏移量。对于因缺失值而没有某个模型预测结果的观测值，仅对偏移量进行插补，而不对缺失值本身进行插补。priority-LASSO-impute（pL-imp（可用））（Hagenberg，2020）采用并推广了这一思路。该方法基于一个数据块拟合序贯 LASSO 模型，并将前一个模型的预测结果作为偏移量。使用其他数据块中的所有可用数据对缺失的偏移量进行插补。

2.3 处理缺失模式的方法

另一种方法是根据缺失模式将数据划分为若干组，然后在不填补缺失值的情况下处理这种结构。此类方法中最早的一种由 Ingalhalikar 等人（2012）提出，他们将数据划分为若干子集，使每个子集针对特定的数据块组合都包含完整观测值。针对每个子集训练一个模型，并根据预测的预期误差对预测结果进行反向加权。多源随机森林也是一种类似的方法，Ludwigs（2020）在一项规模较小的实证研究中对其进行了介绍和评估。该方法在数据子集上分别训练随机森林，每个子集都包含具有特定数据块组合的观测值。与 Ingalhalikar 等人（2012）的方法类似，这些森林的预测结果根据其袋外预测性能加权（Breiman，2001）。最初，多源森林的设计目的与使用部分缺失的多组学协变量数据进行预测不同，详情请参见第3.2.1节。Krautenbacher 等人（2019）也采用了类似思路，即针对每个数据块训练一个模型。

上一段所述的方法在训练数据的各个子集上独立学习模型。不过，也有一些方法会在针对数据不同子集所获得的模型之间共享信息。这样做的优点是能够提高各个模型的稳健性。Yuan 等人（2012）提出了不完整多源特征（iMSF）学习方法。首先，将数据划分为互不重叠的子集，其中每个子集都包含具有特定数据块组合（且不含其他数据块）的观测值。随后，在每个子集上拟合一个正则化回归模型。这些模型受到如下约束：在所有共享某一数据块的模型中，该数据块内必须选择相同的协变量。不过，同一数据块的系数无需取相同的值。与上一段讨论的方法不同，iMSF 无需对测试数据应用不同的模型并获得加权预测。相反，为了获得预测，

会应用具有正确数据块组合的模型。例如，假设训练数据包含数据块 A、B 和 C。在这种情况下，利用训练数据，针对每一种可能的组合分别学习一个模型，即使用以下组合：$\{A, B\}$、$\{A, C\}$、$\{B, C\}$ 和 $\{A, B, C\}$。现在假设某个测试观测缺少数据块 B，因此仅包含数据块 A 和 C。此时，要获得预测，就将基于组合 $\{A, C\}$ 学习得到的模型应用于该测试观测。

Xiang 等人（2014年）进一步发展了 iMFS 方法，提出了不完整源特征选择（iSFS）模型。同样，根据缺失模式形成不同的子集，但与 iMFS 不同的是，这些子集彼此重叠。每个子集被定义为包含某一特定数据块组合中的所有数据块的全部观测值，同时也可能包含其他数据块。例如，如果有三个数据块 A、B 和 C，则与组合 A 和 C 对应的子集包含仅具有 A 和 C 的所有观测值，也包含除 A 和 C 外还具有 B 的所有观测值。与 iMFS 的情况相比，这一过程的优点是每种组合所考虑的观测值更多。此外，与 iMFS 不同的是，并不会为每种组合分别学习一个系数向量。相反，会针对所有数据块中的变量学习一个通用系数向量；为了获得各个组合的系数向量，再将通用系数向量中的各项分别乘以数据块和组合特定权重。这样做的优点是，iSFS 需要估计的参数更少，因而应能得到更稳定的模型。Lan 和 Jiang（2021年）将 iMSF 方法重新表述为一个多任务学习问题。

Liu 等人（2017年）也以与 iMSF 类似的方式，根据缺失模式将观测值划分为多个子集；不过，与 iMSF 相比，这种方法能更充分地利用可用数据。随后，在每个子集上学习一个超图，同时纳入高阶关系。将不同的超图结合起来，用于训练分类器。Dong 等人（2021年）采用了类似的方法。与 Liu 等人（2017年）不同的是，他们在数据的低秩表示上学习超图。之后，为每个数据块训练一个支持向量机分类器，并将预测结果合并。另一种采用超图的方法是基于异构图的多模态融合（Chen 和 Zhang，2020）。首先根据缺失模式将数据划分为多个子集，并构建异构超节点图。在每个图上学习不同数据类型之间的关系，并利用这些信息构建新的超图。随后，学习不同缺失模式之间的相互作用，并将不同数据类型的信息融合到一个嵌入中。

Chen 和 Zhang（2020）、Ingalhalikar 等人（2012）以及 Lan 和 Jiang（2021）提出的方法并非专为多组学情形设计。不过，这些方法仍可应用于此类数据。

3 部分缺失多组学协变量数据处理方法的实证比较研究

3.1 中立性声明

本研究旨在保持“中立”，也就是说，我们着重于比较，而非推广某种特定的新方法（Boulesteix 等人，2017a,b）。不过，我们对所有方法的熟悉程度并不相同。具体而言，纳入研究的两种方法——多源随机森林和优先级 LASSO 插补——由本文部分作者开发。尽管我们对这些方法的熟悉有助于恰当地设置它们，但我们始终致力于进行公平比较：我们既没有比其他方法投入更多精力来优化这两种方法，也没有将研究设计成偏向其中任何一种方法。

3.2 比较研究的设计

3.2.1 比较的方法及其配置

在本小节中，我们将更详细地介绍实证研究中比较的各类方法及其具体配置。从第 2 节所述的方法中选取这些方法时，我们确保第 2.1 至第 2.3 节介绍的三类方法中每类至少包含两种。此外，我们只选择了已在公开 R 包中实现，或能够以合理工作量实现的方法。

本研究考察了以下方法：完整案例法、单块方法、基于 TOBMI 的插补、mdd-sPLS、分块随机森林、多源随机森林和优先级 LASSO 插补。如第 2.1 节所述，前两种是朴素方法。它们作为基线，用于与其他更复杂的方法进行比较。其余方法中有三种基于插补（基于 TOBMI 的插补和 mdd-sPLS，以及优先级 LASSO 插补，见第 2.2 节），另有两种不采用插补程序，而是处理缺失模式（分块随机森林和多源随机森林，见 refsec:pattern 节）。为简便起见，下文将使用“训练数据”和“测试数据”，分别代替“分块缺失多组学训练数据”和“分块缺失多组学测试数据”。

完整案例法（ComplcRF）在传统的完整病例分析中，所有包含缺失值的观测值都会被删除。然而，存在分块缺失时，这种做法往往不可行或效果不佳，因为可能没有或只有极少数观测值的所有数据块均被观测到。因此，为了增加不含缺失数据的训练观测值数量，我们首先从训练数据中删除测试数据中未出现的所有数据块。随后，从训练数据中删除所有包含缺失值的观测值。最后，使用训练数据构建随机森林预测规则（Breiman，2001）。请注意，与完整案例法一样，在下文介绍的大多数方法中，只有在了解测试数据的缺失数据结构后，才能训练预测规则。随机森林未进行调参优化：调参值采用 R 包 randomForestSRC（2.9.2 版）的默认值。例如，每次分裂所抽取的协变量数量 $mtry$ 设为协变量数量的平方根，并且每个森林构建 500 棵树。对于单块方法、基于 TOBMI 的插补方法以及分块随机森林（均在下文介绍）中构建的随机森林，我们也采用相同的做法。

单块方法（SingleBlRF）即使可用数据集可能包含多个数据块，训练也可以仅使用单个数据块进行。在某些情况下，这种做法可能更有优势：某个数据块包含了大部分可用的预测信息，或者不同数据块包含的预测信息存在冗余。单块方法的第一步是从训练数据中移除测试数据中未包含的所有数据块。随后，我们在每个数据块上训练一个随机森林，并使用袋外预测（Breiman，2001）和响应变量的真实值计算这些随机森林的性能指标值。我们使用受试者工作特征曲线下面积（AUC）作为性能指标。最后，在计算出的 AUC 值最大的那个数据块上训练随机森林。

基于 TOBMI 的插补方法（ImpRF）与完整案例法相比，一种更复杂的处理缺失数据的方法是使用数据驱动程序插补缺失值。与完整案例法相比，插补方法的一个重要优势是，训练预测规则时无需排除任何观测值。

TOBMI 专为以下情形设计：存在两个（组学）数据块 A 和 B，数据块 A 对所有观测值均可用，而部分观测值的数据块 B 缺失。首先，计算所有观测值之间的马氏距离矩阵M，其中需要特别注意的是，仅使用数据块 A 中的测量值。随后，按以下方式对每个数据块 B 缺失数据的观测值 i进行插补：1) 在数据块 B 中有测量值的观测值子集中，找出根据 M 与 i 最接近的 k 个观测值（再次注意，后者的距离矩阵仅使用数据块 A 得出）。这些 k 个观测值可能与 i 的行为相似，因此在第二步中用于插补 i 的缺失值；2) 将 i 中每个缺失的数据块 B 数值替换为步骤 1) 确定的 k 个最近邻的加权均值。权重采用从 i 计算的马氏距离的倒数。对于 k，取数据块 A 和 B 均有测量值的观测值数量的平方根（向下取整）。

由于如上一段所述，TOBMI 不适用于一般的分块缺失模式，因此在本比较研究中，我们按以下步骤插补训练数据。首先，我们将所有观测值均无缺失值的数据块拼接起来。随后，我们反复使用 TOBMI，每次插补一个不同的数据块中的缺失值。此处，拼接后的

无缺失值数据块充当上一段中的“数据块 A”，而当前这次迭代中待插补的数据块则充当“数据块 B”。插补是在各个子集之间进行的。请注意，采用此流程时，至少需要有一个数据块对所有观测值均可用，因为我们至少需要一个数据块来计算所有观测值之间的马氏距离（即矩阵 M）。实际应用中大多数情况下都能满足这一条件，因为通常所有患者都有临床信息。

随后，我们从训练数据中移除了测试数据集中不可用的所有数据块，并使用经过这种处理的训练数据构建了随机森林。

多块数据驱动稀疏偏最小二乘（mdd-sPLS）（MddsPLS）与上述插补方法类似，Lorenzo 等（2019）提出的 mdd-sPLS 涉及对数据集进行插补。与基于 TOBMI 的插补方法不同，该方法不仅对训练数据集进行插补，也对测试数据集进行插补，并且分别进行插补。插补后，训练数据和测试数据中不再有缺失值，因此所有数据块都可用于预测。mdd-sPLS 算法是一种基于偏最小二乘（PLS）技术的复杂方法。简单来说，mdd-sPLS 分别在各数据块上执行 PLS 类程序，随后整合信息并预测结果。对于 mdd-sPLS 在预测中使用的协变量，其训练数据和测试数据中的缺失值均使用额外的基于偏最小二乘的模型进行插补。插补时分别使用结果变量和其他协变量。构建预测模型和插补缺失值这两个步骤会重复进行，直至相关潜变量收敛。

我们采用 10 折交叉验证，通过对 10 个取值进行网格搜索，确定相关矩阵的最优正则化参数。此外，对于所涉及的矩阵分解，我们使用一个分量，并根据每个数据块中所选协变量的数量，对各数据块的分量进行反向加权。详情请参阅 Lorenzo 等（2019）。

分块随机森林（BlwRF）与上述方法不同，Krautenbacher（2018）提出的方法（我们称之为“分块随机森林”）使用训练数据集中所有可用的测量值，且不进行插补。该方法为每个数据块分别构建一个随机森林；在每种情况下，都会使用训练数据中包含相应数据块测量值的所有观测值。为获得测试数据的预测，首先将相应的随机森林应用于测试数据中每个可用的数据块，并在每种情况下计算 $Y = 2$ 的预测概率，其中 $Y \in \{1, 2\}$ 表示二元响应。其次，根据相应随机森林通过袋外数据计算出的曲线下面积，为测试数据中各数据块所得的预测概率赋予成比例的权重，并计算其加权平均。请注意，通常也可以采用其他权重，例如依据曲线下面积以外的性能指标确定的权重，或为所有数据块赋予相同权重；后一种做法对应于不加权平均。

多源随机森林（MultisRF）该方法最初是为以下情形设计的：存在多个训练数据集，这些数据集具有不同且可能相互重叠的协变量集，但共享一个响应变量。本文所考虑的情形与此不同：在 MultisRF 最初针对的情形中，不存在成块缺失的数据，但缺失模式更加分散，即每个数据集缺失的协变量各不相同（或按照本文所采用的术语，缺失的“子集”不同）。测试数据包含训练数据集中可用协变量集的一个子集。

为训练预测规则，在每个训练数据集上构建一个标准随机森林。进行预测时，首先按以下方式剪枝这些森林中的每棵树：从划分完整（经自助法或子抽样得到的）数据集的第一个分裂开始，沿树中的每个分支继续，并检查所遇到的每个分裂是否使用了测试数据中可用的协变量；如果没有，则截断该分支。这样可以确保森林中的树只使用测试数据中可用的协变量。随后，与 Krautenbacher（2018）的情形类似，将每个森林应用于测试数据，并按照相应森林的袋外 AUC 值对每个预测进行比例加权。

上述过程可直接应用于使用部分缺失的多组学协变量数据进行预测。在这里，对于协变量集各不相同的数据集，会使用数据的子集；每个子集都包含具有特定数据块组合的观测值。该方法已实现

于可在 GitHub 上获取的 R 包 'multisForest'（https://github.com/RomanHornung/multisForest, 0.1.0版）中。由于该方法的计算负担相对较重，我们在每个森林中仅使用了 250 棵树，而其他基于随机森林的方法使用的是 500 棵树。其余调优参数采用 R 包 'randomForestSRC'（2.9.2 版）中的默认值。

优先级 LASSO 插补（pL-imp（可用））（PrLasso）pL-imp（可用）算法由 Hagenberg（2020）开发，是多组学预测方法 priority-LASSO（Klau 等人，2018）针对分块缺失的多组学数据的扩展。该方法可通过 GitHub 上的 R 包 'priority-lasso' 获取（https://github.com/jonas-hag/prioritylasso）（Klau 等人，2020）。priority-LASSO 是一种基于最小绝对收缩与选择算子（LASSO）（Tibshirani，1996）的方法，允许研究人员为数据块指定优先级排序。通常会进行优先级排序，因为有些数据块比其他数据块更成熟或更容易获取。获得优先级 LASSO 预测规则的第一步，是仅使用优先级最高的数据块中的协变量拟合 LASSO 模型。第二步再次拟合 LASSO 模型，不过这次仅使用优先级第二高的数据块中的协变量，并将第一步拟合的 LASSO 模型的线性预测量作为模型方程中的偏移量。通过纳入后一个偏移量，仅使用优先级第二高的数据块中所含、而优先级最高的数据块中不包含的预测信息。按照数据块的优先级顺序，对所有数据块迭代执行这一过程，从而获得所有协变量的模型系数估计值。

对于分块缺失的多组学数据，这种估计方案并不适用，因为在每一步中，偏移量仅对相应数据块可用的观测值可用。Hagenberg（2020）针对这一问题考虑了几种解决方案。本文仅采用其中一种，即 pL-imp（可用）算法，因为它在 Hagenberg（2020）中表现良好，而且从概念角度看也似乎最有前景。采用这种方法时，使用 LASSO 模型对缺失的偏移量进行填补。简而言之，在每一步中，对于每个子集，该方法首先使用偏移量可用的观测值训练一个 LASSO 模型。在这个 LASSO 模型中，偏移量是响应变量，其他数据块则包含协变量。随后使用拟合后的模型预测缺失的偏移量。

pL-imp（可用）算法并非为处理测试数据中的缺失块而设计。为解决这一问题，我们在将模型拟合到训练数据之前，排除了测试数据中不可用的所有数据块。此外，与原始 priority-LASSO 算法类似，pL-imp（可用）算法要求用户提供可用数据块的优先级排序。由于本文比较研究所用数据集没有可用的生物学信息，因此优先级排序按以下方式确定：1）对每个数据块分别拟合 LASSO 模型，并使用 5折交叉验证（交叉验证）估计各模型对应的偏差；2）将最低交叉验证偏差值对应的数据块赋予最高优先级，将第二低交叉验证偏差值对应的数据块赋予第二高优先级，依此类推；如果两个或更多数据块对应相同的交叉验证偏差值，则随机确定这些数据块之间的优先顺序。

priority-LASSO 和 priority-LASSO-impute 估计程序中所涉及的 LASSO 模型，其收缩参数通过网格搜索和 10折交叉验证确定。

3.2.2 数据

数据材料由来自癌症基因组图谱（TCGA）项目的13个公开可用的多组学数据集组成（Weinstein等，2013）。这些数据是此前在Hornung和Wright（2019）中使用的21个数据集的一个子集。在这21个数据集中，有18个包含所有四个作为协变量的组学数据块（见下文）、临床数据块和响应变量，即是否存在TP53突变。在剩余的18个数据集中，我们剔除了不平衡数据集，这些数据集中较少的响应变量类别所占观测值比例不到15%。最终得到13个数据集，用于比较研究。协变量由以下四个数据块组成：临床数据块、拷贝数变异数据块、miRNA数据块和RNA数据块。表1概述了所用的数据集。

表1：数据集概览。第二列显示观测值的数量。第三列显示具有 TP53突变的观测值比例。第四列至第七列显示相应数据块中的协变量数量（clin：临床协变量，cnv：拷贝数变异，mirna：miRNA，rna：RNA）。

标签	n	TP53比例	clin	cnv	miRNA	rna
BLCA	310	0.49	4	57964	825	23081
BRCA	863	0.34	8	57964	835	22694
COAD	350	0.56	5	57964	802	22210
ESCA	121	0.83	4	57964	763	25494
HNSC	411	0.69	5	57964	793	21520
LGG	454	0.46	3	57964	645	22297
LIHC	298	0.29	4	57964	776	20994
LUAD	424	0.49	6	57964	799	23681
LUSC	365	0.85	7	57964	895	23524
PAAD	142	0.63	4	57964	612	22348
SARC	183	0.36	2	57964	778	22842
STAD	284	0.47	6	57967	787	26027
UCEC	503	0.36	3	57447	866	23978

3.2.3 分块缺失的生成与性能评估

通过随机删除数据集中的部分内容，生成分块缺失模式。首先，按3:1的比例将数据集分为训练数据和测试数据。其次，如下一段所述，分别在训练数据和测试数据中引入分块缺失模式；训练数据集有五种不同模式，测试数据集有四种，见图1。下文分别将训练数据分块缺失模式简称为“trbmp”，将测试数据分块缺失模式简称为“tebmp”。如图1所示，每种trbmp由一个、两个、三个或八个观测子集组成。对于数据集、trbmp和tebmp的每种组合，上述过程均重复五次。

每次将数据分为训练数据和测试数据后，都会随机分配训练观测值的子集归属，并且对于每个训练数据集，各子集大小相同（再次注意，数据按 3:1 的比例分为训练数据和测试数据）。在对组学数据块进行随机置换后，我们引入缺失模式，每次重复使用不同的置换。训练数据和测试数据分别执行这一过程。首先，随机置换组学数据块（临床数据块始终位于第一位）。其次，根据所考虑的相应分块缺失模式删除数据矩阵中的值（图1）。分别对训练数据和测试数据完成这些步骤后，再次对数据块重新排序，使其恢复原始顺序，从而确保训练数据和测试数据中的数据块顺序相同。如果没有上述对组学数据块的随机置换，那么对于不同的缺失模式，每个数据块被观测到的频率就会不相等。这样一来，当不同缺失模式（trbmps 和 tebmps）的结果存在差异时，就无法判断这些差异究竟是由缺失模式造成的，还是因为某些具有影响力的数据块在不同缺失模式中的出现程度不同。置换过程确保了在考虑相同 trmps 和 tebmps 时，不同重复中的子集所缺失的数据块也各不相同。例如，考虑 trbmp 2 和 tebmp 3；在第一次划分训练数据和测试数据时，训练数据中的第一个集合可能包含 RNA 和 miRNA 数据，而第二个集合仅包含突变数据；测试数据可能包含 RNA 和突变数据，但不包含 miRNA 数据。在第二次划分中，训练数据中的第一个集合可能仅包含突变和 miRNA 数据，而第二个集合仅包含 RNA 数据；测试数据则可能包含突变和 miRNA 数据，但不包含 RNA 数据。

对于每个划分，第3.2.1节所述的方法均在训练数据上学习，随后应用于测试数据，并根据三项指标衡量其在测试数据上的表现：布里尔分数、受试者工作特征曲线下面积（AUC）和准确率。布里尔分数衡量区分度和校准，是一种严格评分规则，因为它衡量的是

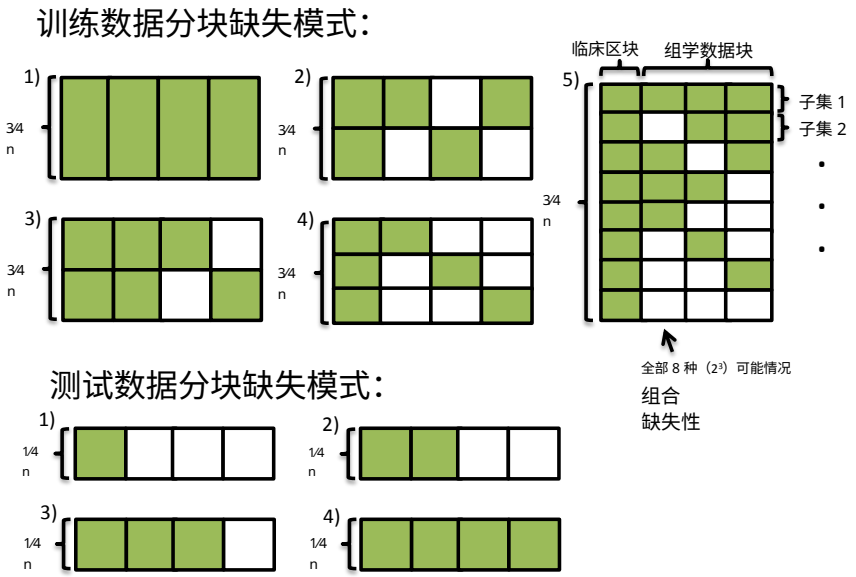

图1：训练和测试数据中的分块缺失模式。彩色/空白方框分别表示相应数据块存在/缺失。第一列始终表示临床数据块。

类别概率预测的准确性（例如，在类别A和类别B这两个类别中，属于类别B的概率）。准确率不是一种严格评分规则，因为它只评估类别预测的准确性，而不评估类别概率预测。因此，与布里尔分数相比，它利用的预测信息更少。AUC只衡量区分度。更准确地说，它衡量模型依据预测概率对不同受试者进行正确排序的能力。请注意，即使模型校准不佳，也可能具有较高的AUC值。例如，模型输出的预测概率系统性地偏低，但类别B预测概率较高的观测值仍然比预测概率较低的观测值更常属于类别B，就会出现这种情况。

3.2.4 代码可用性

用于执行和评估分析（包括数据预处理）的所有R代码均可在GitHub（https://github.com/RomanHornung/bwm_article）上获取。预处理后的数据集以CSV文件形式存放于在线开放获取仓库figshare（http://doi.org/10.6084/m9.figshare.20060201.v1）。有关预处理的详情，请参见Hornung和Wright（2019）。

3.3 结果

首先需要说明的是，对于某些 trbmps 和 tebmps，有些方法只能给出少量成功预测，甚至无法成功预测。例如，在考虑训练区块缺失模式2、3或4时，ComplcRF 不适用于 tebmp 4：对于 tebmp 4，测试数据中没有缺失值，因此训练数据中的变量均不会被删除（参见第3.2.1节）；正因如此，对于 trbmps 2、3或4，训练数据中都不存在完整案例（参见图1），因此完整案例法 ComplcRF 不适用。MultisRF 不适用于训练区块缺失模式1，因为实现 MultisRF 的 R 包 ‘multisForest’（0.1.0版）不支持缺失值的训练数据集。此外，PrLasso 在少数情况下会出错。具体而言，PrLasso 总共进行了1300次重复，其中10次出现错误，除一次外，均发生在 ESCA 数据集上。其余方法在所有情况下都给出了预测结果。ComplcRF、MultisRF 和 PrLasso 出现结果缺失的重复比例分别为25.0%、34.1%和0.8%。关于预测失败原因的详细说明，请参阅补充材料 A 节。

3.3.1 整体性能比较

图2展示了各方法在所有 trbmps、tebmps 和数据集上的汇总排名。请注意，这里我们仅纳入了七种待比较方法均给出结果的重复。如果将每种方法所有可用的重复都纳入，方法间的比较就会有失公平。如上所述，这是因为某些方法无法针对特定的 trmps 和 tebmps 给出预测结果。鉴于不同 trmps 和 tebmps 下的预测性能通常有所不同，如果采用所有可用结果，方法间的比较就会受到 trmps 和 tebmps 的混杂影响。

总体而言，不同方法的性能差异并不明显。在布里尔分数和准确率方面，PrLasso表现最佳；而在曲线下面积方面，各方法之间没有明显的优胜者。BlwRF在三项性能指标上的表现都属于最差之列，而MddsPLS仅在布里尔分数和曲线下面积方面表现不佳。

如上所述，图2仅纳入了七种方法均有结果的重复，因此排除了许多重复。这主要是由于ComplcRF和MultisRF造成的。例如，如上所述，由于MultisRF未能提供该方法的结果，因此排除了所有采用trbmp 1的重复。因此，补充图S1和S2分别展示了排除ComplcRF和MultisRF后上述结果。排除ComplcRF和MultisRF后，我们能够纳入更多重复，因为无需分别排除所有缺少ComplcRF和MultisRF结果的重复。分别排除ComplcRF和MultisRF后，结果没有明显差异。

补充图S3、S4和S5展示了各项指标的原始值，这些结果证实，不同方法所得结果之间的差异并不显著。

3.3.2 按训练区块缺失模式分别考察性能

图3展示了各方法在不同 trbmp 下的布里尔分数排名。为使表述清晰，正文中没有展示所有三项性能指标的结果。AUC 和准确率的相应结果见补充图S6和S7。之所以选择布里尔分数，是因为它可以被视为最重要的指标：它同时衡量区分度和校准，而 AUC 和准确率各自只衡量其中一项。下文将重点讨论布里尔分数，同时也会描述另外两项性能指标中观察到的差异。

PrLasso 在图3所示的四种 trbmp 下均表现最佳。请注意，此处未纳入训练块缺失模式1，因为 MultisRF 不适用于训练块缺失模式1；此外，我们再次仅考虑了七种方法均有结果的重复。对于训练块缺失模式2至5，BlwRF 和 MddsPLS 再次属于表现最差的方法。对于训练块缺失模式2、4和5，ImpRF 也属于表现最差的方法，但在训练区块缺失模式3下并非如此。ImpRF 在训练区块缺失模式3下并非表现最差，这一观察结果很可能可以用 ImpRF 的一个特性来解释。如第3.2.1节所述，ImpRF 使用所有观测值中均有观测的那些数据块进行拼接，以计算用于插补的距离矩阵。对于训练区块缺失模式3，所有观测值中均有两个数据块可用；而对于其他 trbmp，只有一个数据块满足这一条件（不包括没有缺失观测值的训练块缺失模式1）。因此，可以预期，训练区块缺失模式3下计算出的距离矩阵能更好地反映观测值之间的真实距离，这或许可以解释为何 ImpRF 在训练区块缺失模式3下的表现优于其他 trbmp。不过，在 AUC 和准确率方面（补充图S6和S7），我们并未观察到这一现象；相较于布里尔分数所得结果，ImpRF 总体表现更好。

对于 trbmp 5，除 PrLasso 和 SingleBlRF 外，所有方法的表现都同样不佳。与其他 trmps 相比，ComplcRF 在 trbmp 5 上的表现可能更差，因为该 trbmp 中完整观测的数量要少得多。MultisRF 在 trbmp 5 上表现更差，可能是因为其子集的规模远小于其他 trbmps。基于这些小型子集训练的随机森林，其预测性能可能因此受到影响。同样，就曲线下面积和准确率而言，ImpRF 在 trbmp 5 上表现更好。

排除 ComplcRF 和 MultisRF 后得到的结果再次非常相似（补充图 S8 至 S13）。由于 MultisRF 不适用于 trbmp 1，因此只有排除该方法后，才能研究 trbmp 1 的结果（补充图 S11）。这些结果相似

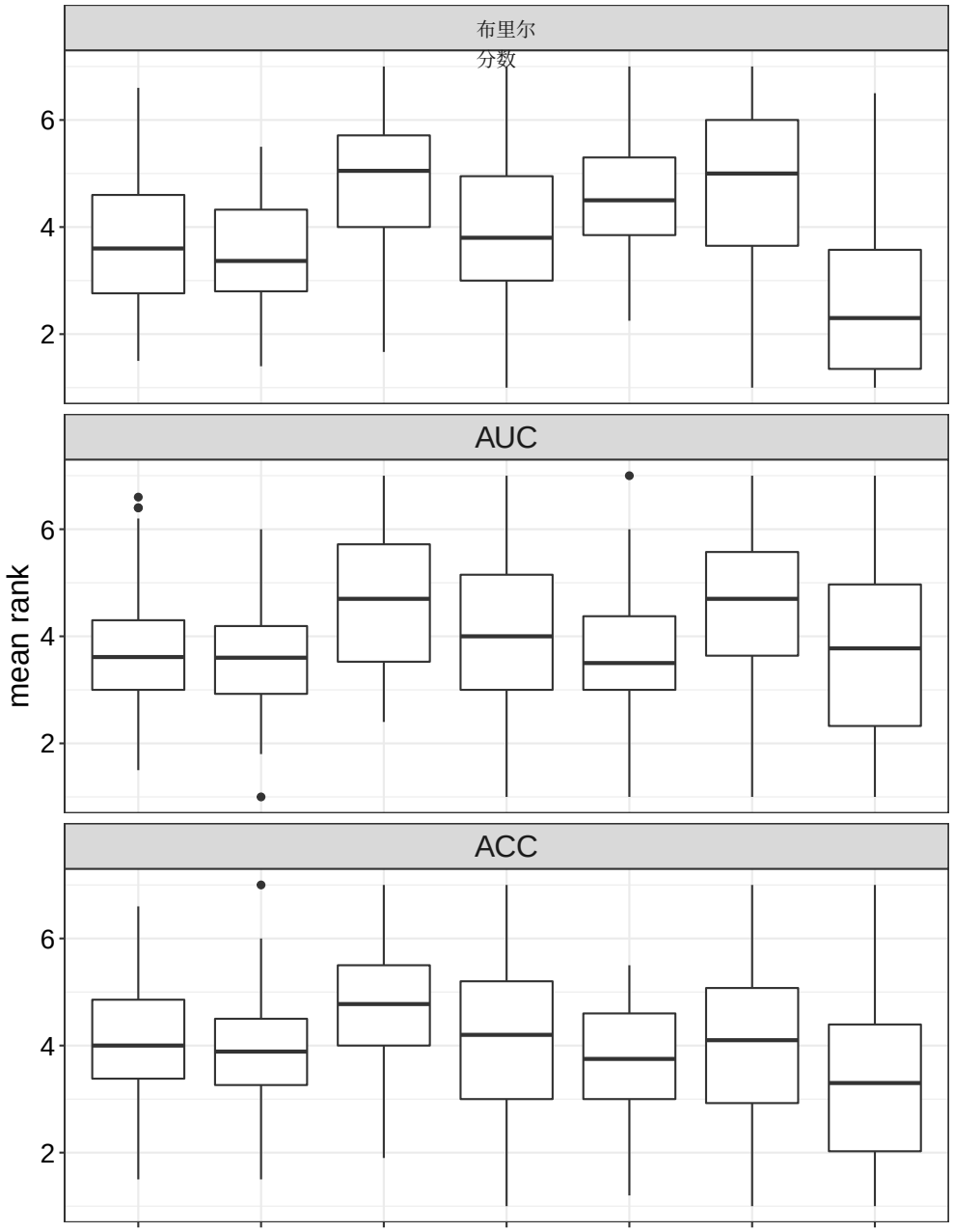

图2：各方法在三项性能指标方面相较于其他方法所取得的排名——整体性能。针对每种 trbmp、tebmp、数据集和重复的组合计算排名，仅纳入七种纳入考察的方法均有结果的重复。随后对各重复的排名取平均值。排名越小，表示性能越好。

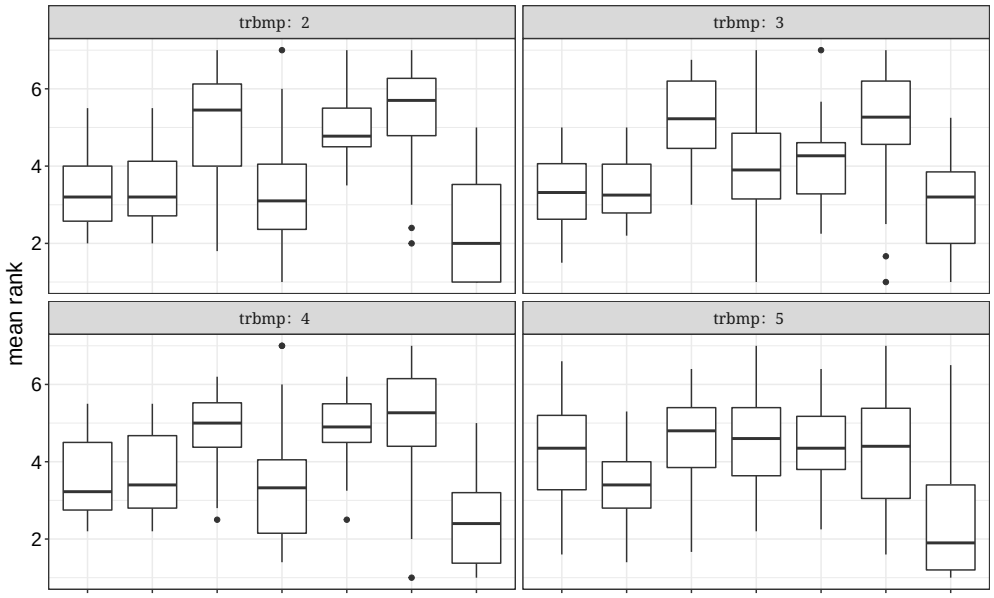

图3：各方法在布里尔分数方面相对于其他方法的排名，按 trbmp 分别列出。排名针对 trbmp、tebmp、数据集和重复的每种组合进行计算；仅纳入所有七种方法均有结果的重复。随后对各重复的排名取平均值。排名越小，表示性能越好。

与 trbmp 2、3 和 4 的结果相比，BlwRF 在这里的表现明显最差。不过，在曲线下面积方面，MdssPLS 的表现也同样差（补充图S12）。请注意，如果训练数据中没有缺失块（即 trbmp 1），ComplcRF 和 ImpRF 完全相同。

总体而言，方法之间在曲线下面积和准确率方面的性能差异小于布里尔分数方面的差异（补充图S6、S7、S9、S10、S12和S13）。按 trbmp 分析所得的性能指标原始值见补充图S14至S22。

3.3.3 按测试块缺失模式分别考察性能

为了按 tebmp 分别比较不同方法的结果，我们再次重点关注布里尔分数（图4）。除 tebmp 2 外，PrLasso 在所有 tebmp 下的表现均最佳；在 tebmp 2 下，ComplcRF 和 SingleBlRF 的表现则同样出色。ComplcRF 的这一良好表现或许可以解释为：在 tembmp 2 下，测试数据中观测到的数据块较少，因此许多数据块会从训练数据中移除，进而增加训练数据中的完整观测数量。不过请注意，在曲线下面积和准确率方面（补充图S23和S24），PrLasso 并非明显最佳的方法。

对于 tebmp 1，ComplcRF、SingleBlRF、BlwRF 和 ImpRF 的表现同样出色。这可以解释为，对于该 tebmp，测试数据中只有临床数据块可用，而对于这些方法，训练数据中所有在测试数据中不可用的数据块都会被移除。因此，这四种方法在 tebmp 1 中的运行方式完全相同，因为它们都只使用临床数据块构建标准随机森林。请注意，对于 tebmp 1，PrLasso 相当于对临床数据块拟合一个简单的 Lasso 模型。鉴于此，有趣的是，PrLasso 的表现仍优于上述四种基于随机森林的方法（曲线下面积指标除外，见补充图S23）。这意味着，使用临床数据块作为协变量数据时，标准 Lasso 的表现优于标准随机森林。tebmp 1 也是 MddsPLS 的表现优于上述四种基于随机森林的方法的唯一情形——

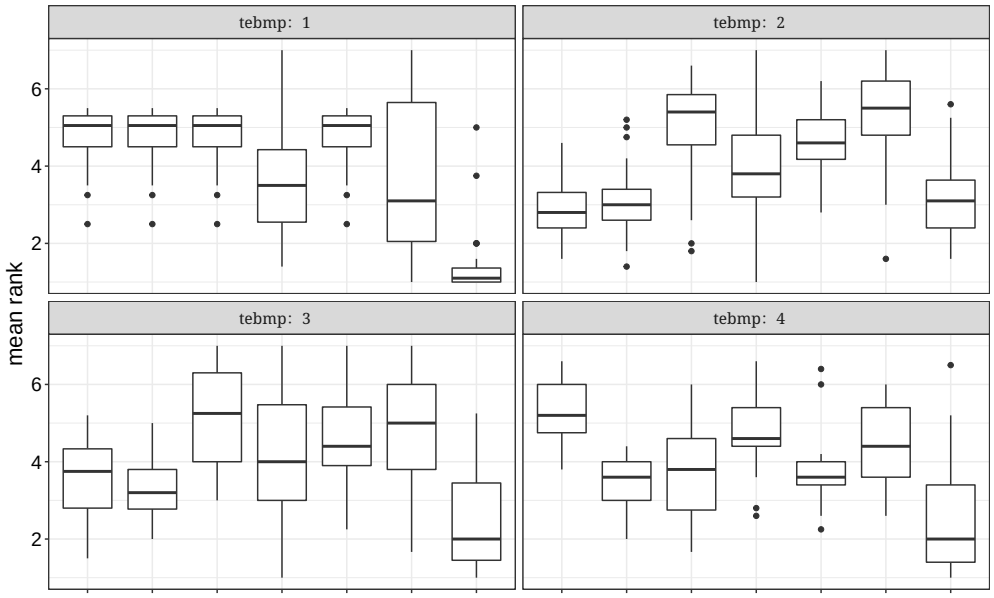

图4：各方法在布里尔分数方面相对于其他方法所取得的排名，按 tebmp 分别展示。针对 trbmp、tebmp、数据集和重复次数的每种组合，计算各方法的排名；仅纳入七种方法均有结果的重复次数。随后对这些重复次数的排名取平均。排名越小，表示性能越好。

基于随机森林的方法。鉴于 tebmp 1 是唯一一个测试数据中除临床数据块外的所有数据块均缺失的 tebmp，MdssPLS 在 tebmp 1 中表现较好，可能是因为该方法是唯一会对测试数据中的缺失块进行插补的方法。然而，对于曲线下面积指标，MdssPLS 在 tebmp 1 中的表现也逊于上述四种基于随机森林的方法；在这一指标上，MultisRF 的表现也较差。

解读其余 tebmps 的结果时，必须考虑到图4中 tebmp 4 展示的结果仅为 trbmp 5 下获得的结果。与之前一样，本图仅展示七种纳入考察的方法均得出结果的重复次数。正如上文所述，对于 tebmp 4，除 trbmp 5 外，所有 trbmps 下的结果均缺少 ComplcRF（trbmps 2、3 和 4）或 MultisRF（trbmp 1）的结果。因此，解读 tebmp 4 下获得的结果时，我们必须参考补充图 S25，该图展示了排除 ComplcRF 后获得的相应结果。对于 tebmps 2、3 和 4，BlwRF 和 MddsPLS 仍属于表现最差的方法。对于 tebmps 3 和 4（tebmp 4 参见补充图 S25），ImpRF 也属于表现最差的方法；但在考察 trbmp 与 tebmp 各种组合的结果时（第 3.3.4 节），这一点不那么明显；而对于曲线下面积和准确率，完全看不出这一点（补充图 S23、S24、S26、S27、S29 和 S30）。对于在布里尔分数方面观察到 ImpRF 在 tebmps 3 和 4 下表现略差，一种可能的解释如下：测试数据中的数据块数量越多，训练数据中保留的数据块也越多，因此缺失值也越多，这意味着需要填补的缺失值比例更大。

与前文各节一样，各方法在曲线下面积和准确率方面的性能差异较小（补充图 S23、S24、S26、S27、S29 和 S30）。再次排除 ComplcRF 和 MultisRF（补充图 S25 至 S30）并未显著改变结果，但 tebmp 4 除外，详见上文。性能指标的原始值见补充图 S31 至 S39。

3.3.4 按 trbmp 和 tebmp 的各个组合分别考察性能

在前两个小节中，我们分别按 trbmp 和 tebmp 研究了各方法的性能，但没有针对 trbmp 与 tebmp 的各种组合分别进行研究。图5展示了各方法在不同 trbmp 与 tebmp 组合下相对于其他方法的布里尔分数排名。由于某些方法不适用于特定的 trbmp 和 tebmp（组合），要解读所有 trbmp 与 tebmp 组合下的结果，我们还需参考补充图 S42 和 S45；这两幅图分别展示了排除 ComplcRF 和 MultisRF 后的结果。

在考察的三项性能指标中，一个共同的观察结果是：对于 tebmp 4，即测试数据中没有缺失观测值的情形，ComplcRF 在 trbmp 5 下的表现远差于 trbmp 1（补充图 S45 至 S47）。这很可能是因为在 trbmp 1 中，所有观测值都是完整的，而在 trbmp 5 中，只有第一个子集中的观测值是完整的。因此，在 trbmp 1 下，ComplcRF 可用于训练的观测值要多得多。在第 3.3.2 节中，我们已经观察到 ComplcRF 在 trbmp 5 下表现较差；但在此处聚焦于 tebmp 4 时，这种性能下降要明显得多，因为如果测试数据中的所有数据块都可用（即 tebmp 4），那么在 trbmp 5 下，完整观测始终仅限于第一个子集。

不过，对于布里尔分数（图5）和准确率（补充图 S41），而非曲线下面积（补充图 S40），还可以观察到：在 tebmp 1 的情况下，MultisRF 在训练区块缺失模式3下的表现不如在 trbmps 2、4、5 下。这或许可以解释为：从最初的分裂点开始，MultisRF 会剪除所有树，在使用了测试集中不存在的变量的分裂点处截断树枝。与组学协变量相比，临床协变量数量很少，因此最初的分裂点使用临床协变量的可能性极低。正因如此，如果测试数据中只有临床协变量（即 tebmp 1 的情况），森林中的许多树就会被完全移除，这自然会导致预测性能下降。最后，MultisRF 在训练区块缺失模式3下的表现不如其他 trbmp（2、3 和 4），原因可能是 trbmp 3 中观察到的数据块更多，因此最初的分裂点尤其不太可能使用临床协变量。

对于 trmbp 1 和 tebmp 4 的组合，PrLasso 并未明显优于 ComplcRF（补充图 S45 至 S47）。值得注意的是，第 3.3.3 节的结果显示，在 tebmp 1 的情况下，PrLasso 优于 ComplcRF（以及 SingleBlRF、BlwRF 和 ImpRF）；此时仅使用临床数据块作为协变量数据。这表明，PrLasso 可能只有在仅使用临床协变量数据时才优于 ComplcRF（即标准随机森林），而在多组学数据的情况下则未必如此。

AUC 和准确率的结果与布里尔分数的结果大多非常相似；不过，前两项性能指标下各方法性能之间的差异往往仍然较小（补充图 S40、S41、S43、S44、S46 和 S47）。除少数例外外，排除 ComplcRF 和 MultisRF 几乎不会改变其余方法所得的结果（补充图 S42 至 S47）。性能指标的原始值见补充图 S48 至 S56。

4 讨论

对于我们考察的大多数不同 trbmps 和 tebmps，PrLasso 在布里尔分数方面表现最佳。在第 3.3.3 节中，我们获得了有力证据，表明与基于随机森林的方法相比，PrLasso 更好地利用了临床协变量中包含的预测信息。鉴于已有研究发现，临床协变量对于多组学数据的预测非常重要（Herrmann 等，2020），这可能是 PrLasso 在我们的比较研究中表现优异的一个重要原因。

在大多数设置下，BlwRF 和 MddsPLS 的表现最差。这两种方法的共同点是将各个数据块彼此独立地处理，而在进行预测时会使用所有数据块。将不同数据块分开处理可能并不有效，因为其中包含的预测信息存在重叠。令人意外的是，朴素方法总体上并不属于表现最差的方法，在某些设置下甚至属于表现最佳的方法。相比之下，PrLasso 展现出更稳健的行为，即

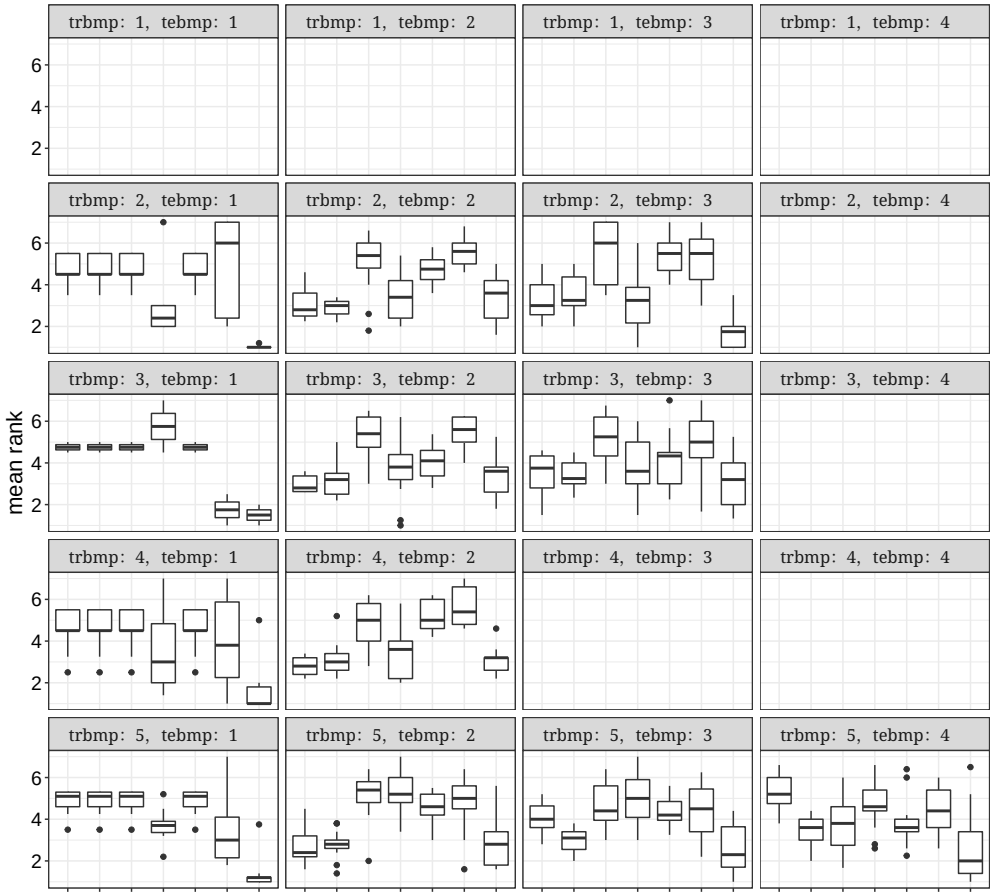

图5：各方法在布里尔分数方面相对于其他方法所取得的排名，按 trbmp 和 tebmp 的每种组合分别列出。针对每种 tebmp、tebmp、数据集和重复组合计算排名，仅纳入七种方法的结果均可用的重复。随后对各重复的排名取平均值。排名越小，表现越好。

它始终是表现最好的方法之一。正如预期的那样，如果训练数据中只有少量完整观测，完整案例法 ComplcRF 的表现并不好。

就目前的形式而言，MultisRF 并不十分适用于多组学情形。一个重要原因是预测前执行的剪枝过程（参见第3.2.1节）。如果测试数据包含较小的数据块，这一过程预计会导致使用 MultisRF 构建的随机森林中有很大比例的树被移除。我们在比较研究中观察到了这一点：如第 3.3.4 节所述，如果测试数据仅包含临床协变量，但同时包含许多组学数据块，其预测性能就会下降。通过将使用不同子集构建的标准随机森林替换为能考虑多组学结构的预测方法（例如块森林方法（Hornung 和 Wright，2019年）），也可以让 MultisRF 更适用于多组学情形。

不同性能指标得出的结果有所不同。例如，尽管 PrLasso 在布里尔分数方面的表现明显优于其他方法，但在曲线下面积方面，各方法之间并没有明显的优胜者。总的来说，解读我们的结果时必须格外谨慎。与所有基于真实数据集的研究一样，如果我们以其他真实数据集为基础（采用其他纳入标准，或使用来自其他数据库的数据），并针对每个真实数据集进行更多次随机重复，所得结果在一定程度上可能会有所不同。此外，尽管我们的数据集数量多于许多仅基于少数（例如四个或五个）数据集的基准研究，但纳入的数据集数量仍然有限。Nießl 等人（2022年）表明，即使使用大量数据集，基准研究的结果通常也存在变异，并且会受到分析选择的显著影响。鉴于此，为避免得出无法复现的结论，我们格外谨慎地为观测结果提供合理解释。在 ComplcRF 和 MultisRF 的某些 trmpbs 和 tempbs 中，没有相应的结果。一般而言，这类系统性缺失值是实证基准研究中的一个问题，可能导致结果偏倚。不过，我们在可视化中只纳入了通过全部七种方法获得的重复，因此确保了在考虑这一问题的情况下恰当地呈现结果。此外，分别排除 ComplcRF 和 MultisRF 后提供的补充图表表明，我们的结果对于缺失值造成的偏倚相当稳健。

我们的研究还有若干局限，有待未来研究加以解决。为使研究清晰易懂，我们只考察了数量有限的简化缺失模式，这些模式可能无法涵盖现实中遇到的所有潜在情形。例如，实际情形可能呈现更多“非规则模式”，其中患者子集的规模差异很大。此外，个别协变量也可能存在缺失值，即除患者子集中整类组学数据缺失之外，还可能出现其他缺失值。值得注意的是，实践中，包含不同数据块组合的子集中的数据往往来自不同来源（例如，在不同时间点生成或使用不同机器采集）。这会导致数据子集之间存在系统性差异，通常称为批次效应（Li 等人，2009年）。我们实证研究中的训练数据和测试数据子集是从同一数据集随机划分而来，因此不存在批次效应。在训练数据中，可以使用 ComBat（Johnson 等人，2006年）等批次效应去除方法校正此类批次效应。相应地，校正测试数据时可采用附加批次效应去除（Hornung 等人，2016年），将测试数据转换为与训练数据分布相似的形式。这种附加批次效应去除有助于提升预测性能；存在批次效应时，预测性能往往会下降（Hornung 等人，2017年）。

我们假设在训练模型之前，测试数据中的缺失模式已知。原则上，虽然可以根据相应测试数据的缺失模式，在预测之前重新训练模型，但在实践中这并不总是可行的。将模型应用于具有不同缺失模式的测试数据，与本研究所考虑的情形存在根本差异，并且需要采用不同的、可能更复杂的技术。

尽管从本文采用的纯方法论视角来看，将 TP53突变选作疾病结局替代指标的二元响应变量是可以接受的，但考虑其他响应变量，或将研究扩展至其他类型的响应变量（例如删失生存时间），或许也有意义。然而，我们研究中考虑的并非所有方法都能处理这种情况。我们还可以将研究扩展至纳入更多方法，尤其是未在 R 中实现、因而未纳入本次比较的方法。

5 结论

我们首先概述了有关分块缺失的多组学协变量数据预测方法的前沿文献。随后，我们对其中一些方法开展了大规模基准比较研究。研究结果或可帮助面临分块缺失的多组学数据的应用研究人员选择合适的方法。不过，鉴于基准研究的结果通常存在较大的变异性，因此不应过度解读本研究结果的具体细节。除了应用研究人员，文献综述和基准研究也可能有助于方法学研究人员开发更强大的新方法，既汲取最有前景方法的优势，又弥补其不足。

致谢

作者感谢 Anna Jacob 对语言表达进行的宝贵润色。本研究得到德国科学基金会的资助：授予 Roman Hornung 的 HO6422/1-2 号项目 [；授予 Anne-Laure Boulesteix 的 BO3139/6-2 和 BO3139/4-3 号项目]。

补充材料和 R代码

补充材料可通过以下链接获取：http://www.ibe.med.uni-muenchen. de/organisation/mitarbeiter/070_drittmittel/hornung/missmultiomics_suppfiles/suppmat_hornungetal2023.pdf。本研究所用数据可在 figshare 上公开获取，网址为 http://doi.org/10.6084/m9.figshare.20060201.v1。所有用于执行和评估分析（包括数据预处理）的 R代码均可在 GitHub 上获取（https://github.com/RomanHornung/bwm_article）。

参考文献

A.-L. Boulesteix、H. Binder、B. Abrahamowicz、W. Sauerbrei 等。论比较统计方法的研究之必要性与设计。《生物计量学杂志》。*Biometrische Zeitschrift*，60(1):216–218，2017年a。

A.-L. Boulesteix、R. Wilson 和 A. Hapfelmeier。迈向循证计算统计学：从临床研究中汲取关于真实数据基准研究的作用与设计的经验。《*BMC*医学研究方法学》，17(1):1–12，2017年b。

L. Breiman. 随机森林。*Mach Learn*, 45(1):5–32, 2001. T. Cai, T. T. Cai, and A. Zhang. 结构化矩阵补全及其在基因组数据整合中的应用。《美国统计学会会刊》, 111:621–633, 2016. J. Chen and A. Zhang. Hgmf：面向不完整多模态数据的异构图融合。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1295–1305页，2020年。Z. Dai, Z. Bu, and Q. Long. 通过生成对抗网络对高维区块缺失值问题进行多重插补。载于*2021*年第*20*届*IEEE*机器学习与应用国际会议（*ICMLA*），第791–798页。IEEE，2021年。A. Dong, Z. Li, M. Wang, D. Shen, and M. Liu. 用于多模态痴呆诊断的高阶拉普拉斯正则化低秩表示。*Frontiers in Neuroscience*, 15:634124, 2021年。X. Dong, L. Lin, R. Zhang, Y. Zhao, D. C. Christiani, Y. Wei, and F. Chen. TOBMI：采用k近邻加权方法进行跨组学区块缺失数据插补。《*Bioinformatics*》，35(8):1278–1283, 2019年。

J. Hagenberg. 基于多组学数据中区块缺失值的预后建模惩罚回归方法。硕士论文，慕尼黑大学，2020年。M. Herrmann, P. Probst, R. Hornung, V. Jurinovic, and A.-L. Boulesteix. 使用多组学数据的生存预测方法大规模基准研究。《*Briefings in Bioinformatics*》, 2020年。数字对象标识符：10.1093/bib/bbaa167. bbaa167. S. Hieke, A. Benner, R. F. Schlenl, M. Schumacher, L. Bullinger, and H. Binder. 通过提取互补信息，将多种分子来源整合为临床风险预测特征。《*BMC Bioinformatics*》, 17:327, 2016年。R. Hornung and M. N. Wright. 分块森林：适用于临床与组学协变量数据块的随机森林。《*BMC Bioinformatics*》, 20:358, 2019年。R. Hornung, A.-L. Boulesteix, and D. Causeur. 将位置与尺度批次效应校正与潜在因子校正数据清理相结合。《*BMC Bioinformatics*》, 17:27, 2016年。R. Hornung, D. Causeur, C. Bernau, and A.-L. Boulesteix. 通过附加批次效应校正或附加标准化改进跨研究预测。《*Bioinformatics*》, 33(3):397–404, 2017年。M. Ingalhalikar, W. A. Parker, L. Bloy, T. P. Roberts, and R. Verma. 利用含缺失特征的多参数数据学习病理模式。载于国际医学图像计算与计算机辅助干预会议，第468–475页，柏林，海德堡，2012。Springer Berlin Heidelberg。W. E. Johnson, C. Li, and A. Rabinovic. 使用经验贝叶斯方法校正微阵列表达数据中的批次效应。《*Biostatistics*》, 8(1):118–127, 2006年。S. Klau, V. Jurinovic, R. Hornung, T. Herold, and A.-L. Boulesteix. Priority-lasso：一种利用多组学数据预测临床结局的简单层次化方法。《*BMC Bioinformatics*》, 19:322, 2018。S. Klau, R. Hornung, and A. Bauer. *prioritylasso*：采用偏移量方法分析多组学数据，2020年。网址：https://github.com/RomanHornung/prioritylasso。R 包版本 0.2.4。N. Krautenbacher. 复杂、有偏及大数据学习：以儿童哮喘为例，在流行病学研究和基因组医学中进行疾病风险预测。博士论文，慕尼黑工业大学，2018年。网址：http://mediatum.ub.tum.de/doc/1446834/document.pdf。N. Krautenbacher, N. Flach, A. B¨ock, K. Laubhahn, M. Laimighofer, F. J. Theis, D. P. Ankerst, C. Fuchs, and B. Schaub. 一种高维多变量分析策略，可根据遗传、免疫学和环境因素对儿童哮喘表型进行分类。《*Allergy*》, 74(7):1364–1373, 2019年。Q. Lan and S. Jiang. 一种基于分块缺失数据的信用评估建模方法。《*Applied Intelligence*》, 51(10):6859–6880, 2021年。W. Lan, X. Chen, T. Zou, and C.-L. Tsai. 通过半监督学习方法对协变量中的高缺失率数据进行插补。《*Journal of Business & Economic Statistics*》, 40(3):1282–1290, 2022年。J. Li, P. Bushel, Chu T-M, and R. D. Wolfinger. 主方差成分分析：估计微阵列基因表达数据中的批次效应。载于 A. Scherer 编，《微阵列实验中的批次效应与噪声：来源与解决方案》，第141–154页。John Wiley & Sons，奇切斯特，英国，2009年。H. Linder and Y. Zhang. 针对缺失数据和通路模型的迭代集成插补及其在乳腺癌亚型中的应用。《*Communications for Statistical Applications and Methods*》, 26:411–430, 2019年。M. Liu, Y. Gao, P.-T. Yap, and D. Shen. 面向不完整多模态数据的多超图学习。《*IEEE Journal of Biomedical and Health Informatics*》, 22:1197–1208, 2017年。H. Lorenzo, J. Saracco, and R. Thi´ebaut. 面向多块不完整数据的监督学习。arXiv:1901.04380, 2019年。

F. Ludwigs. 针对多个训练数据集及含分块缺失值测试数据的预测方法比较研究。硕士论文，慕尼黑大学，2020年。C. Nießl、M. Herrmann、C. Wiedemann、G. Casalicchio 和 A.-L. Boulesteix。基准研究中的过度乐观，以及解读其结果时设计与分析选项的多重性。*Wiley*《数据挖掘与知识发现跨学科评论》，12:e1441，2022年。D. J. Stekhoven 和 P. B¨uhlmann。Missforest——一种针对混合类型数据的非参数缺失值填补方法。《*Bioinformatics*》，28(1):112–118，2011年。K.-H. Thung、C.-Y. Wee、P.-T. Yap、D. Shen 和 A. D. N. Initiative。通过矩阵收缩与补全，利用不完整多模态数据诊断神经退行性疾病。《神经影像》，91:386–400，2014年。R. Tibshirani。通过 LASSO 实现回归收缩与变量选择。《英国皇家统计学会会刊：*B*辑（方法论）》，58(1):267–288，1996年。X. Wang 和 Q. Sun。人类癌症中的 TP53突变、表达及相互作用网络。《肿瘤靶点》，8(1):624–643，2017年。J. N. Weinstein、E. A. Collisson、G. B. Mills、K. R. Shaw 和 B. A. Ozenberger。癌症基因组图谱泛癌分析项目。《自然·遗传学》，45(10):1113–20，2013年。S. Xiang、L. Yuan、W. Fan、Y. Wang、P. M. Thompson、J. Ye 和 I. Alzheimer's Disease Neuroimaging。针对异质块级缺失数据的双层多源学习。《神经影像》，102 Pt 1:192–206，2014年。F. Xue 和 A. Qu。模型选择中多源分块缺失数据的整合。《美国统计学会会刊》，116(536):1914–1927，2021年。X. Yang、Y.-J. Kim、M. Taub、R. Azevedo 和 M. Chi。Prime：智能辅导系统中多模态的块级缺失处理。载于国际多媒体建模会议，第63–75页。施普林格，2020年。L. Yuan、Y. Wang、P. M. Thompson、V. A. Narayan、J. Ye 和 A. D. N. Initiative。用于联合分析不完整的多源异质神经影像数据的多源特征学习。《神经影像》，61: 622–632，2012。Y. Zhang、N. Tang 和 A. Qu。针对高维块级缺失数据的填补因子回归。《中华统计学志》，30(2):631–651，2020年。H. Zhu、G. Li 和 E. F. Lock。针对具有块级缺失结构的多类型数据的广义整合主成分分析。《*Biostatistics*》，21:302–318，2020年。