ELSEVIER

特点

生物医学数据中的二重奏效应如何迷惑机器学习？

王立荣[a] , 王立顺 [b,c], 吴文斌 [d,e,⇑]

机器学习（ML）模型已被越来越多地用于药物开发，以更快地确定潜在目标。交叉验证技术通常被用来评估这些模型。然而，这种验证方法的可靠性可能会受到数据二重性的影响。当独立得出的数据彼此非常相似时，就会出现数据二重性，导致模型无论如何训练都会表现良好（即二重性效应）。尽管在生物医学数据中存在大量的数据二重性和它们的非flationary效应，但它们仍然没有被描述出来。我们展示了它们在生物医学数据中的普遍性，证明了二重奏是如何产生的，并提供了它们混杂效应的证据。为了减轻二重奏的影响，我们建议在训练-验证分离之前识别数据的二重奏者。

关键词计算生物学；数据科学；多普勒效应；机器学习

简介

ML模型已被越来越多地用于药物发现中，以加快药物开发速度。ML以多种方式提高药物发现的效率。ML模型可以更快地筛选出更好的候选药物（焦油），减少用于发现和测试的时间。它们还可以确定现有的美国食品和药物管理局（FDA）批准的药物用于治疗其他疾病（药物再利用），大大降低

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

了药物开发的成本。[1]这两种方法在最近几年都显示出了前景。Exscientia的 "Centaur Chemist "人工智能（AI）平台对一种新的抗癌候选药物EXS21546进行了检测。

8个月，目前正在进行临床试验（NCT04727138）。[2]几个由ML识别的治疗2019年冠状病毒（COVID-19）的药物和药物组合也已进入临床试验，如由基于网络的方法[3]识别的褪黑激素和托瑞米芬的药物组合（NCT04531748）和由BenevolentAI的知识图谱识别[4]的baricitinib（NCT04373044和NCT04401579）。

基于ML和AI的分类模型也提高了药物开发的效率。分类器（训练有素的模型）已被用于预测新药与疾病的相互作用[5,6]和可能的药物不良反应。[7]鉴于其昂贵的如果要进行有效的药物测试，必须对这些分类器进行适当的培训和测试，以确定合适的候选药物。

在ML中已经确立的是，在评估分类器的性能时，训练和测试数据集应该是独立得出的。然而，独立得出的训练和测试数据集仍然可能产生不可靠的验证结果。例如，在数据二重体上训练和验证的模型（由于机会或其他原因，训练和验证集高度相似）可能会表现良好，而不管训练的质量如何。[8]当一个分类器因为数据的存在而错误地表现良好时

https://doi.org/10.1016/j.drudis.2021.10.017

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

我们说有一个观察到的 "二重奏 "效应。早些时候，我们说过，当样本在其测量中出现相似时，就会出现数据二重性。然而，这可能并不能保证二重性效应的存在。因此，同时产生多普勒效应（混淆ML结果）的数据多普勒效应被称为功能性多普勒效应。尽管有一些记录在案的数据二重性的例子（见 "生物数据中数据二重性的丰富程度"），但检查样本训练-评估对是否独立和/或不相似的做法仍然不常见。此外，数据二重性及其伴随的下游分析效应（二重性效应）没有得到很好的记录和理解。在此，我们想更好地了解疑似功能性掺杂物之间的相似程度以及验证集中可接受的功能性掺杂物的比例。尽管存在一些建议的识别数据二重性的方法（见 "改善数据二重性"），但大多数方法都不是可生成的或足够强大的。因此，当务之急是研究数据二重性的性质并提出改进的二重性识别方法。[9]我们使用一个具有适当设计的CON-TROLS的肾细胞癌基准数据集，在此说明在这种生物医学数据中，功能二重性的普遍性，数据二重性对ML的影响，以及消除二重性效应的方法。

特写

生物数据中的数据二重性的丰度

在现代生物信息学中已经观察到了数据二重性。在一个值得注意的案例中，Cao和Fullwood对现有的染色质相互作用预测系统进行了详细评估。[10]他们的工作显示，这些系统的性能被夸大了，因为在报告这些系统时，评估方法有问题。特别是，这些系统是在与训练集具有高度相似性的测试集上进行评估的。Goh和Wong也观察到了数据二重性的存在，即某些验证数据被保证了良好的性能。即使所选的特征是随机的，但在给定的特定训练数据中，数据二重性也是存在的。[11]数据二重性也存在于生物信息学的既定领域：在蛋白质功能预测中，具有相似序列的蛋白质被推断为是同一祖先蛋白质的后代，从而继承了该祖先的功能（即这两种蛋白质被推断为功能相似）。这种对归纳推理的天真应用在大多数情况下是正确的（数据二重性的情况），给我们一个高度准确预测的错误印象。然而，在更大的启发下，我们意识到这种方法无法正确预测序列不太相似但功能相似的蛋白质的功能，如暮光之城的同源物[12]和序列总体上不相似但活性部位残基相似的酶。[13]药物发现中也有类似的例子：定量结构-活性关系（QSAR）模型是分类和回归ML模型，通过分子的结构特性来预测其生物活性。[14]QSAR模型假定结构相似的分子具有相似的活性。在大多数情况下，这个假设是真实的（数据掺假的情况）。将具有相似活性的相似分子分为训练集和验证集（在时间分割验证或随机测试集选择过程中的偶然性）[15]混淆了模型验证，因为训练不良的模型（在无信息的结构特性上训练）在这些分子上可能仍然表现良好。[16]我们只能通过在具有不同活性的类似分子上测试它们的表现来区分训练不良的模型和训练良好的模型（SAR悖论）。如果该悖论是结构上的微小变化造成的，而这些变化对结合力有很大的影响，那么[17]理论上训练有素的模型在这些情况下仍然表现良好，因为它们是根据信息结构特性训练的，因此能够检测到这些微小的变化，而训练不良的模型则无法识别真正的生物活性。尽管生物医学数据科学界似乎越来越意识到这种数据二重性的问题，但令人惊讶的是，消除或尽量减少测试和研究之间的相似性的程序，却没有得到证实。

数据二重性的识别

鉴于二重性效应有可能造成混淆，因此在验证前能够识别训练集和验证集之间存在的数据二重性是至关重要的。识别数据二重性的一个合理方法是使用排序方法（如主成分分析）或嵌入方法（如t-SNE），再加上散点图，以了解样本在减维空间的分布情况。然而，我们发现这样的方法是不可行的，因为数据的二重性

在减维空间

中并不一定是

可区分的（在线补充资料中的图S1）。

早期对类似问题的研究也提出了识别数据二重性的措施。其中一种方法，dupChecker，通过比较其CEL文件的MD5指纹来识别重复的样本。[18]相同的MD5指纹表明样本是重复的（基本上是复制的，因此表明有泄漏问题）。因此，dupChecker并不能检测出真正的数据二重性，即独立得出的样本偶然相似。另一项措施，成对皮尔逊相关系数（PPCC），捕捉不同数据集的样本对之间的关系。[19]一个异常高的PPCC值表明一对样本构成了PPCC数据的二重性（注意，不可能确定这对样本中哪一个是原始的）。尽管合理和直观，但原始PPCC论文的主要局限性在于它从未在PPCC数据二重性和它们混淆ML任务的能力之间建立起联系（即具有功能效应，因此，作为功能二重性）。在重新分析他们的数据时，我们也意识到，他们报告的二重体实际上是泄漏的结果（在样本复制之间），因此，并不构成真正的数据二重体。然而，PPCC作为一种量化指标的基本设计在方法学上是合理的。因此，我们用它来识别潜在的功能二重性（从PPCC数据二重性中提取）。

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

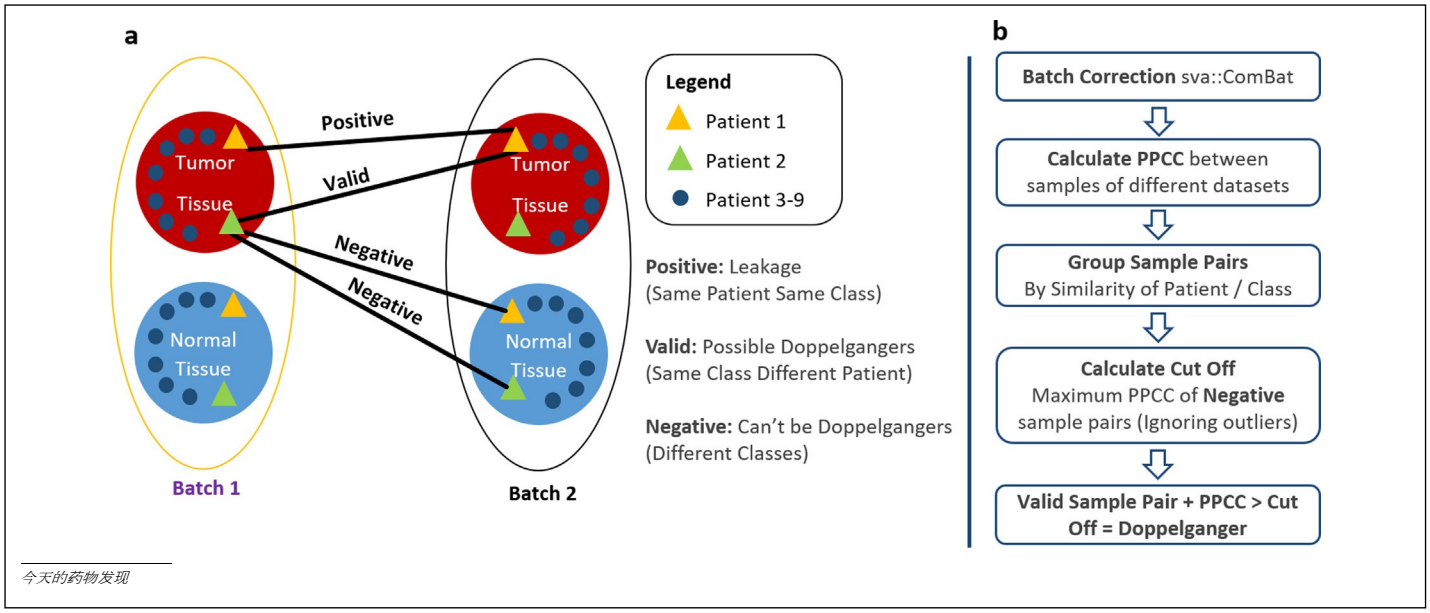


图1

说明成对皮尔逊相关系数（PPCC）数据二重性识别方法的图示。(a) 根据不同类型的样本对的病人和等级的相似性的命名规则。(b) PPCC数据二重性识别的过程。PPCC数据二重性被定义为有效的样本对，其PPCC值大于所有负的样本对。

pelgängers），来自构建的基准情景。

为了构建基准情景，我们使用了Guo等人的肾细胞癌（RCC）亲基因组学数据，该数据[9]取自NetProt软件库[20]（见在线补充信息中的补充方法）。选择RCC是因为它在构建明确的场景方面的效用。(i) 负面的情况，在这种情况下，通过构建不同类别标签的样本对，不允许有掺假者；(ii) 有效的情况，在这种情况下，通过构建分配给同一类别标签但来自不同样本的样本对，允许有掺假者。然后，这些影响可以与积极的情况（通过采取来自同一样本的技术复制构建的对；这些构成明显的泄漏问题，因此，不被视为二重奏）进行比较（图1a）。我们在RCC数据集的两批中模拟了这些情况（见在线补充资料中的补充方法）。

我们根据有效情景与消极和积极情景的PPCC分布情况来识别PPCC数据二重性。令人惊讶的是，我们观察到高比例的PPCC数据二重性（一半的样本与至少一个其他样本是PPCC数据二重性；图2c）。在有效的情况下，PPCC的分布是一个广泛的连续体，没有明显的断裂。这表明，使用离群检测方法（如PPCC原始论文中所建议的）将不够敏感。它还表明，数据二重性是作为样本之间相似性光谱的一部分而自然存在的（而不是壮观的异常）。至于为什么会出现这种情况，我们不能确定这是否是PPCC本身的问题，还是因为基因的转录特征在大多数情况下是正相关的。[21]我们检查了相同和不同组织对之间的PPCC分布情况（图2b）。相同组织对的PPCC值总体上仍然很高，表明样本之间的相关性很高，即使它们来自不同的病人。鉴于许多基因有共同的调节器，这并不令人惊讶。然而，如果我们比较不同的组织对，PPCC的分布肯定会更低，因为其中一定也存在着类效应。相比之下，当我们考虑来自同一样品或组织的复制时，PPCCs也是非常高的。这些评价表明，PPCC具有有意义的区分价值。

PPCC数据二重性的混杂影响

在确定了RCC中的PPCC数据二重性后，我们探讨了它们对不同的随机训练分类器（训练分类器是一个从训练数据中"学习"出来的ML模型）的估值准确性的影响。这将确定PPCC数据二重性是否作为功能性二重性，对ML性能有明显的影响（见在线补充信息中的补充方法）。

我们注意到，训练和验证数据中PPCC数据二重性的存在会影响ML的性能，即使这些特征是随机选择的（因此是无意义的；换句话说，模型在验证期间应该表现不佳）。这一发现在不同的训练和验证数据集（图3）和不同的ML模型上是一致可重复的。此外，在训练和验证集中代表的二重奏对越多，ML的性能就越不稳定。这表明PPCC数据二重性者的数量与二重性者效应的大小之间存在着基于剂量的关系。当所有正确训练的模型的验证精度

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

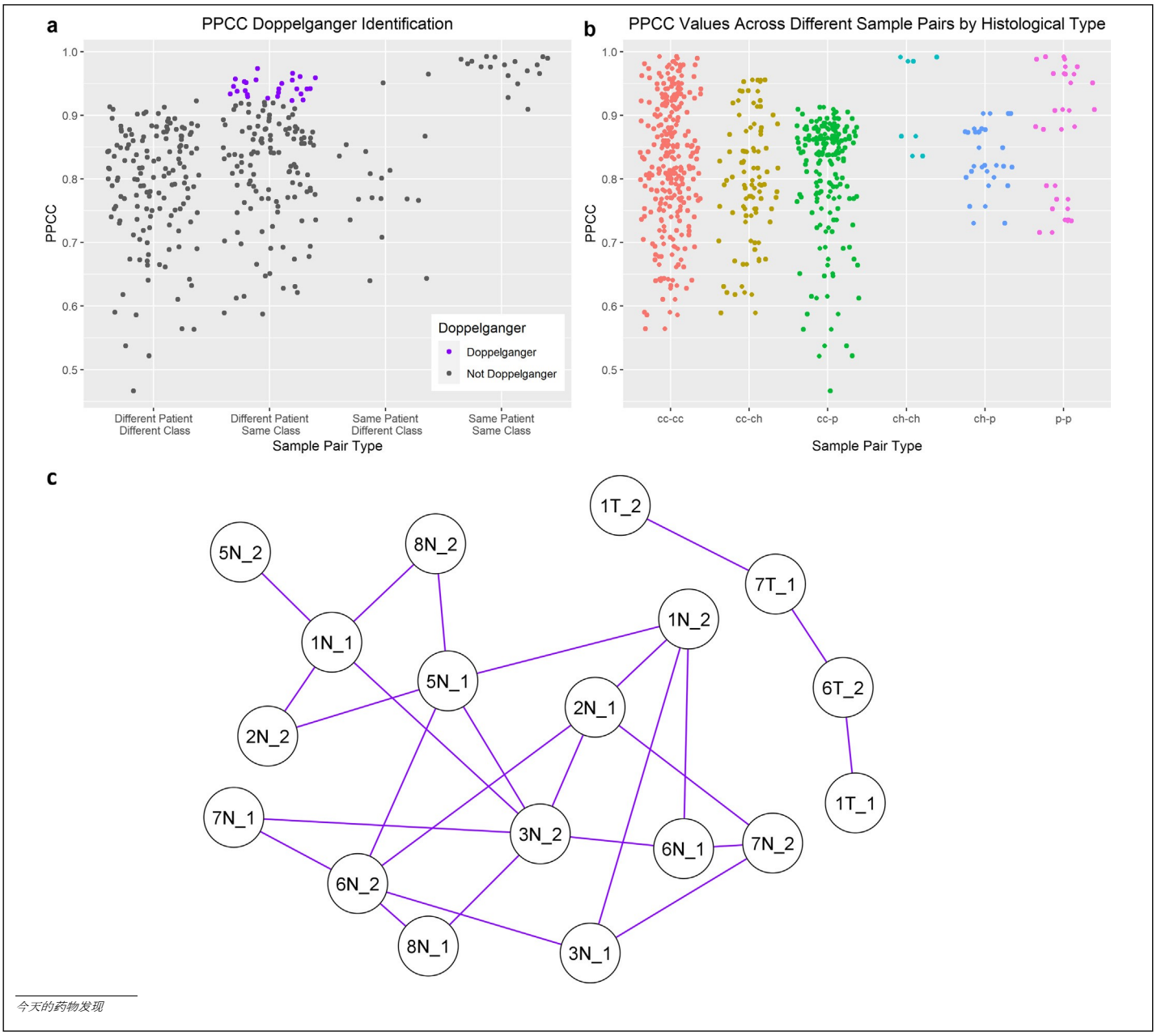


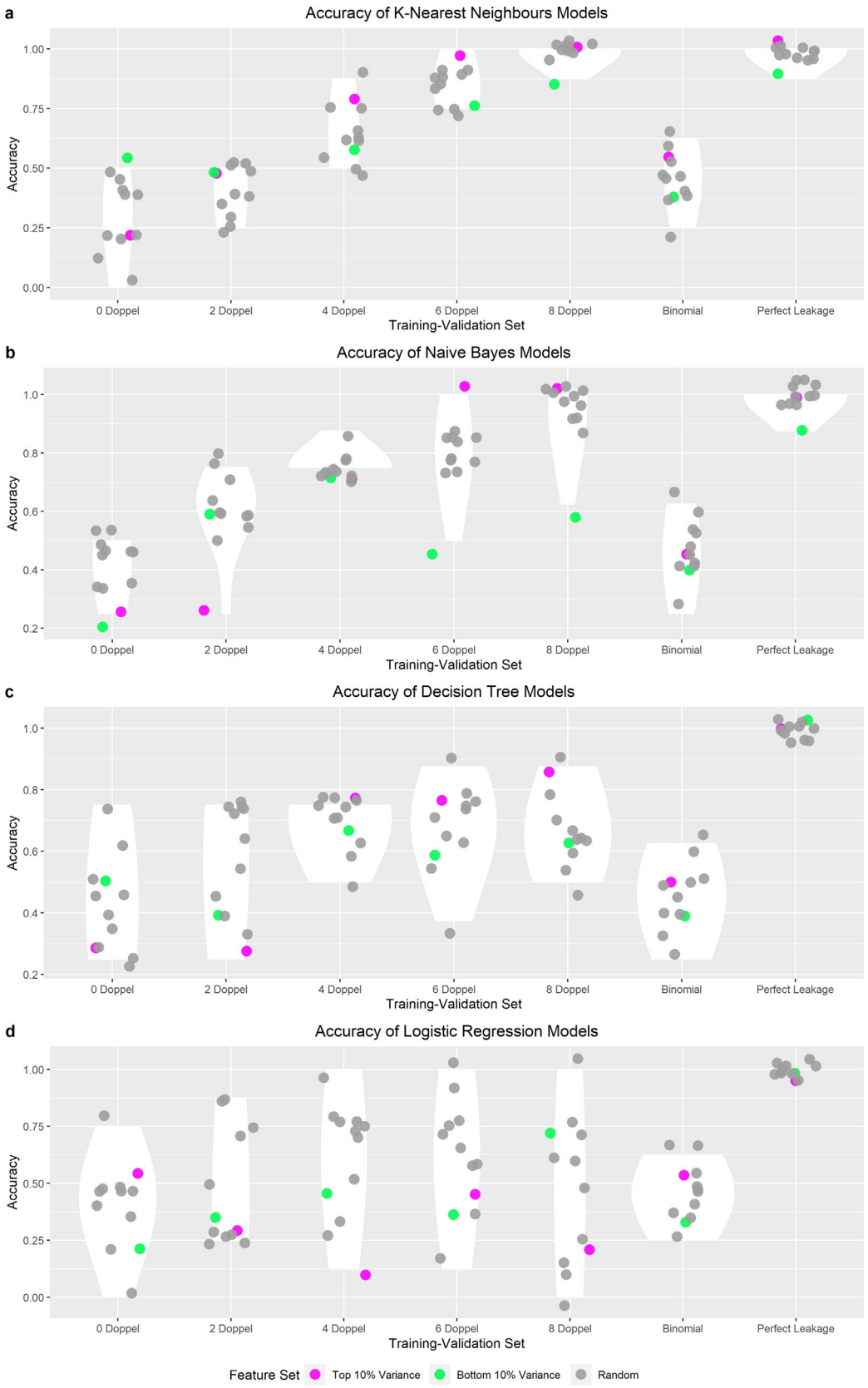
图2

(a) 不同样本对的成对皮尔逊相关系数（PPCCs）的分布。X轴表示按其病人和等级的相似性分组的样本对的类型。Y轴表示PPCC（即两个样本之间的皮尔逊相关系数）。26个PPCC数据二维码用紫色标示。(b) 不同样本对的PPCC值按其组织学类型的分布。X轴表示按组织学类型对的样本对的类型。透明细胞肾细胞癌（RCC）用cc表示，嗜铬细胞RCC用ch表示，乳头状RCC用p表示，Y轴表示PPCC。(c) 26个PPCC数据的二重性可视化为一个图表。每个节点代表一个不同的样本，第一个数字表示病人编号，下面的字母代表类别（N，正常；T，肿瘤），连字符后的数字代表样本的批次（例如，'5N_2'表示来自第五个病人的正常样本的第二个复制）。每个节点/样本之间存在一条边，意味着这两个样本是PPCC数据的二重体。该图中有18个节点（即在总共36个样本中，有18个样本与至少一个其他样本是二重性的）。

(有'前10%差异'的特征集），对'Doppel 4'（训练-验证集）被分层为PPCC数据二重奏者和非PPCC数据二重性阶层，所有ML模型在PPCC数据二重性上的表现都比非PPCC数据二重性上的表现高。

PPCC数据二重性（在线补充信息中的表S2）。这一结果反映了我们先前关于蛋白质序列功能预测的观点：在有许多类似例子（许多数据二重性）的地方，良好的准确性很容易得到，但事实上却不能保证对不太相似的例子的普遍性。然而，当类似的例子很少时（数据二重性很少），模型中的差距就会暴露出来，因此，模型的表现往往会不佳。

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。



今天的药物发现

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

这一结果表明，PPCC数据的二重性（基于成对的相关性）作为功能性的二重性（混淆ML结果），产生了类似于数据泄露的flationary效应。在我们使用k-近邻（kNN）模型的实验中，二重性影响和泄漏之间的相似性是显而易见的，在验证中带有8个二重性的训练-验证集显示出与完全泄漏的训练-验证集相同的准确性分布（图3a）。然而，并不是所有的模型都受到同样的影响：与决策树和逻辑回归模型相比，kNN和天真贝叶斯模型在性能变化和二重奏剂量之间有更明显的关系。

将所有的二重奏者放在训练集中，准确率下降到~0.5，这是一个在随机签名上训练的模型的预期准确率。很明显，当所有PPCC数据的二重性被放在一起训练集时，二重性效应被消除了。这为避免二重奏效应提供了一种可能的方法。然而，将PPCC数据的二重性限制在训练集或验证集是次要的解决方案。在前者中，当训练集的大小被限定时（因此，每一个被纳入的数据二重性者都会导致一个不太相似的样本被排除在训练集之外），它导致的模型可能不能很好地概括，因为该模型
缺乏知识。在后者中，你可能会出现壮观的赢家通吃的情况（二重奏者都会被预测正确或错误）。

改善数据二重性的问题

数据二重性会对ML产生不良的影响。这就提出了一个问题：二重奏的效果如何？可以进行管理。在上一节中，我们认为在训练集或验证集中强制将多巴哥人放在一起是次优的解决方案。Cao和Fullwood呼吁根据所分析的数据的特定背景，采取更全面和严格的评估策略。[10]这可以通过根据单个染色体分割训练和测试数据（而不是将所有染色体放在一起），以及使用不同的细胞类型来生成训练-评估对来实现，从而在该领域建立一个良好的实践/标准。然而，这在实践中是很难做到的，因为它的前提是要有先验知识和高质量的背景/基准数据。

在使用PPCC离群检测包doppelgangR（见"数据二重性的识别"）来识别二重性的研究中，PPCC数据二重性可以被移除以减轻其影响。[22,23]然而，这种方法不适用于具有高比例PPCC数据二重性的小型数据集，如RCC，因为去除PPCC数据二重性会使数据减少到无法使用的程度（至于所有生物医学数据实例中数据二重性的发生率是多少，我们认为这将是一个联盟级的努力，鉴于二重性的影响对高质量的生物医学ML模型构成明显的障碍，这可能是必要的）。

因此，我们也试图用一些不会导致样本量大幅减少或需要大量背景数据的方法来缓解二重奏效应，尽管我们的尝试到目前为止都是失败的。例如，我们试图通过以下方法来修剪数据移除对数据二重性效应有很大贡献的变量（在线补充信息中的图S2）。然而，我们观察到在去除相关变量后，PPCC数据二重性效应没有变化。这一观察结果暗示了二重奏效应的极端复杂性，因为样本对之间的高相关性的原因不能简单地用高度相关的变量子集来解释。我们现在正着眼于新的特征工程和诺玛化方法来解决这个问题，这可能被证明是更成功的。

建议

虽然直接从数据中去除数据二重性被证明是难以实现的，但我们仍然需要防范二重性效应。

我们的第一个建议是，以元数据为指导，进行仔细的交叉检查。在这里，我们使用RCC中的元数据来构建负面和正面的案例。这使我们能够预测PPCC的得分范围，即不可能存在二重体的情况（不同类别；阴性案例）和存在泄漏的情况（基于复制的同患者和同类别；阳性案例）。值得关注的合理的数据二重性是来自同一类别但不同病人的样本。有了这些来自元数据的信息，我们就能够识别潜在的二重性，并将它们全部归入训练集或验证集，有效地防止二重性效应，并允许对ML性能进行相对更客观的评估。与此类似，来自同一样本的技术复制也应该得到类似的处理。这一建议与生物信息学中的指导方针相类似，即当ML模型被训练于

3

图3

不同机器学习（ML）模型对训练-验证集的预测性能，验证集中成对的皮尔逊相关系数（PPCC）数据二重性的数量不同。评估的ML模型包括：K-近邻（kNN）（a），天真贝叶斯（b），决策树（c），和逻辑回归（d）模型。X轴表示验证集的类型。i Doppel "指的是在训练集中有*i*个PPCC数据二重性的验证集（其中*i*=0、2、4、6和8），"二项式 "指的是由12个（特征集的数量）二项式分布产生的准确率，*N*=8（因为验证集中有8个样本），*P*=0。5（每个验证样本猜出正确标签的概率）（阴性对照），'完美泄漏'指的是验证集与训练集有8个重复（阳性对照）。Y轴表示ML模型在8个样本的验证集上的准确度，最低准确度为0，最高准确度为1。"前10%方差 "是指由方差最大的蛋白质组成的特征集（即数据集中蛋白质总数中的前10%）。底部10%差异 "是指由数据集中蛋白质总数的10%的最低差异的蛋白质组成的特征集，"随机 "是指由数据集中蛋白质总数的10%的随机选择的蛋白质组成的特征集。

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

对于来自生物序列的数据，研究人员应确保训练和测试样本不是重复的或相似度高的样本。[24]

我们的第二个建议是对数据进行分层。与其在整个测试数据上评估模型性能，我们可以将数据分层为不同相似度的阶层（例如，PPCC数据二重性者和非PPCC数据二重性者，并分别评估每个阶层的模型性能）。假设每个分层都与现实世界中的已知比例相吻合，那么在对某一分层的性能进行互测时，我们仍然能够通过考虑该分层在现实世界中的流行率来了解分类器的现实性能。更重要的是，模型性能较差的阶层能指出分类器中的差距。在RCC中，用于分层性能评估的非PPCC二元组也正好是乳头状RCC样本。鉴于每个组织中肾癌细胞的比例是已知的（乳头状RCC占肾癌细胞的10%），[25]分类器在乳头状RCC上的不良表现表明，这10%的肾癌细胞样本是我们分类器的一个薄弱环节，

需要 进一步改进。[21]

我们的第三个建议是进行极其强大的独立验证检查，涉及尽可能多的数据集（分歧验证）。[8]尽管不是直接对冲数据二重性，但分歧验证技术可以告知分类器的客观性。尽管训练集中可能存在数据二重性，但它也能告知模型的可生成性（就现实世界的使用而言）。

未来的研究可以探索其他不严重依赖元数据的功能二重奏识别方法。在这种方法中，我们可以直接确认功能二重性。例如，我们可以寻找验证集的子集，这些子集无论使用何种ML方法都能预测正确。这些子集是训练集的潜在功能二重性（训练集和验证集之间的样本对，无论我们如何训练模型，都会影响模型的准确性）。随后，将这种方法与PPCC进一步配对，可能使我们能够分辨出二重奏的部分训练集中测试集样本的子集（或者反过来说，有趣的问题是，不相似的集子，或者非数据二重性，是否也可以作为功能二重性）。在模型评估过程中，应该避免这些子集，因为它们作为功能二重奏，对不同模型的相对性能没有什么启示。

结论性意见

ML模型的性能通常是通过在验证数据上测试模型的准确性来评估的。只有当验证数据独立于训练数据时，这种模型验证的方法才是有效的。然而，这个假设通常被认为是真实的，不需要事先检查。在存在二重奏效应的情况下，这个被广泛接受的假设可能不成立。我们发现，在我们的测试数据中，二重奏是相当普遍的，而且它对ML的准确性有直接的影响。这反过来又降低了ML对表型分析和后续潜在药物线索识别的作用。我们还注意到，这种影响的程度取决于两个主要因素：功能二重体的相似性和功能二重体在验证集中的比例。不幸的是，二重奏效应不容易分析解决。因此，为了避免性能下降，在训练和验证数据分类之前，检查数据中潜在的二重性是很重要的。

作者贡献

L.R.W.进行了分析并撰写了手稿。L.W.提供了关键的反馈。W.W.B.G.监督并共同撰写了手稿。

鸣谢

本研究/项目由新加坡国家研究基金会根据其产业联盟基金--定位前（IAF-PP）资助倡议提供支持。本资料中表达的任何意见、发现和结论或建议均为作者本人的观点，不代表新加坡国家研究基金会的观点。W.W.B.G.还感谢新加坡教育部（MOE）一级拨款（RG35/20号拨款）的支持。

利益声明

作者宣称有竞争性的利益。

附录A。补充材料 本文的补充数据可在网上找到：https://doi.org/10.1016/j.drudis.2021.10.017。

参考文献

1 Y.Zhou,F.Wang,J.Tang,R.Nussinov,F.Cheng, ArtificialintelligenceinCOVID-19drugrepurposing, LancetDigitalHealth2(2020)e667-e676.
2 Savage N. Tapping into the drug discovery potential of AI.自然》杂志。2021年5月27日在线发表。2021年9月2日访问。https://doi.org/10。1038/d43747-021-00045-7.
3 F.Cheng, S. Rao, R. Mehra, COVID-19治疗：使用基于网络的方法结合抗炎和抗病毒治疗，Cleve Clin J Med (2021), https://doi.org/10.3949/ccjm.87a. ccc037 [2021年6月30日在线发表].
4 P.Richardson,I.Griffin,C.Tucker,D.Smith,O. Oechsle,A.Phelan,etal.,Baricitinibaspotential treatmentfor2019-nCoVacuterespiratorydisease, Lancet395(2020)e30.
5 J.-Y.Shi, X.-Q.Shang, K. Gao, S.-W.Zhang, S.-M.Yiu, An integrated local classification model of predicting drug-drug interactions via Dempster-Shafer theory of evidence,SciRep8(2018)1-11.
6 M.Oh, J. Ahn, Y. Yoon, A network-based classification modelforderivingnoveldrug-diseaseassociations andassessingtheirmolecularactions,PLoSONE9 (2014)e111668.
7 Y.Hwang, M. Oh, G. Jang, T. Lee, C. Park, J. Ahn, et al., Identifyingthecommongeneticnetworksof ADR (adversedrugreaction)clustersanddevelopingan ADRclassificationmodel,MolBiosyst13(2017)1788-1796.
8 S.Y. Ho,K.Phua,LWong,W.W.B. Goh,扩展外部验证以检查学习模型的可解释性和可概括性,模式1(2020) 100129.
9 T.Guo, P. Kouvonen, C.C. Koh, L.C. Gillet, W.E. Wolski, H.L.Röst,etal.,Rapidmassspectrometricconversion of tissue biopsy samples into permanent quantitative digitalproteomemaps,NatMed21(2015)407-413.
10 F.Cao,M.J.Fullwood, Inflated performance measures inenhancer-promoterinteraction-prediction methods,NatGenet51(2019)1196-1198.
11 W.W.B.Goh,L.Wong,Turningstrawintogold: buildingrobustnessintogenesignatureinference, DrugDiscovToday24(2019)31-36.
12 M.N.Wass,M.J.Sternberg,ConFunc: 暮光之城的功能注释,生物信息学24 (2008)798-806.
13 I.Friedberg,Automatedproteinfunction prediction-thegenomicchallenge,BriefBioinform 7(2006)225-242.
14 D.Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade,Artificialintelligenceindrugdiscoveryand development,DrugDiscovToday26(2021)80-93.
15 E.N. Muratov, J. Bajorath, R.P. Sheridan, I.V. Tetko, D. Filimonov, V. Poroikov, 等，QSAR without borders, ChemSocRev49(2020)3525-3564。
16 A.Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, et al., QSAR modeling: where have you been?你要去哪里？, J Med Chem 57 (2014)4977-5010
17 Q.Chen, L. Wu, W. Liu, L. Xing, X. Fan, 通过整合结构和基因表达信息增强QSAR模型性能,Molecules18(2013) 10789-107801.

特写

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。

18 Q.Sheng,Y.Shyr,X.Chen,DupChecker:a bioconductorpackageforcheckinghigh throughputgenomicdataredundancyinmeta-analysis,BMCBioinform15(2014)323.

19 L.Waldron,M.Riester,M.Ramos,G.Parmigiani,M. Birrer, The Doppelgänger effect: hidden duplicates in databases of transcriptome profiles, J Natl Cancer Inst 108(2016)djw146.

20 W.W.B.Goh,L.Wong,NetProt:基于复杂的特征选择，JProteomeRes16(2017)3102-3112.

21 D.Venet, J.E. Dumont, V. Detours, Most random gene expressionsignaturesaresignificantlyassociated withbreastcanceroutcome,PLoSComputBiol7 (2011)e1002240.

22 K.Lakiotaki,N.Vorniotakis,M.Tsagris,G. Georgakopoulos,I.Tsamardinos,BioDataome:a collectionofuniformlypreprocessedand 自动注释的数据集用于数据驱动的生物学，数据库 2018（2018）bay011。

23 S.Ma,S.Ogino,P.Parsana,R.Nishihara,Z.Qian,J. Shen,etal.,Continuityoftranscriptomesamong colorectalcancersubtypesbasedonmeta-analysis, GenomeBiol19(2018)1-14.

24 生物信息学。作者须知（机器学习） 。 https://academic.oup.com/bioinformatics/pages/instructions_for_authors#General%20Policies。[Accessed October 21, 2021].

25 V.F.Muglia,A.Prando, Renal cellcarcinoma: histologicalclassificationandcorrelationwith imaging findings,RadiolBras48(2015)166-174.

李荣旺 [a], 王林森 [b,c], 吴文彬 [d,e,*]

[a]计算机科学和工程学院。新加坡南洋理工大学

[b]新加坡国立大学计算机科学系，新加坡

[c]新加坡国立大学病理学系，新加坡

[d]新加坡南洋理工大学李光前医学院

[e]南洋理工大学生物科学学院，新加坡

* *通讯作者在。新加坡南洋理工大学李光前医学院。*

特写

请将这篇文章作为新闻引用。L.R. Wang等，Drug Discovery Today（2021），https://doi.org/10.1016/j.drudis.2021.10.017。