MAFS：通过过滤方法深度融合实现高维数据的多头注意力特征选择

Xiaoyan Sun[1, †], **Qingyu Meng**[1, †], **Yalu Wen**[1, *]

[1]奥克兰大学统计系，新西兰奥克兰

[†]这些作者对本文贡献相同。[*]通讯作者： [y.wen@auckland.ac.nz]

arXiv:2601.02668v1 [cs.LG] 6 Jan 2026

摘要

特征选择对于高维生物医学数据至关重要，可在精准医学应用中提升预测性能、降低计算成本并增强可解释性。现有方法面临诸多挑战。过滤方法具有高度可扩展性，但无法捕捉复杂关系或消除冗余。基于深度学习的方法能够建模非线性模式，但通常在大规模应用中缺乏稳定性、可解释性和效率。单头注意力提升了可解释性，但在捕捉多层级依赖关系方面存在局限，并且仍对初始化敏感，从而降低了可重复性。现有大多数方法很少将统计可解释性与深度学习的表征能力相结合，尤其是在超高维设置下。为此，我们提出 MAFS（基于多头注意力的特征选择），一种将统计先验与深度学习能力相结合的混合框架。MAFS 首先利用基于过滤式方法的先验进行稳定初始化并引导学习。随后，它使用多头注意力从多个视角并行考察特征，从而捕捉复杂的非线性关系和交互。最后，重排序模块整合各注意力头的输出，解决冲突并最大限度地减少信息损失，以生成稳健且一致的特征排名。该设计将统计引导与深度建模能力相结合，在最大限度保留信息性信号的同时生成具有可解释性的重要性得分。在模拟数据集和真实世界数据集上，包括癌症基因表达数据和阿尔茨海默病数据，MAFS 始终较现有基于过滤式方法和基于深度学习的替代方法实现更优的覆盖率和稳定性，为高维生物医学数据中的特征选择提供了一种可扩展、可解释且稳健的解决方案。

1 引言

精准医学旨在根据个体的遗传、环境和生活方式因素制定有针对性的预防与治疗策略[1]。高通量生物技术和大规模生物样本库的快速扩展，使得多组学数据得以收集，其中涵盖基因组学、转录组学、表观基因组学及其他分子层面信息，同时还包括临床和生活方式信息[2, 3, 4]。这些数据集具有超高维、异质和多模态等特点，能够捕捉跨越多个分子尺度的复杂生物过程。从此类数据中提取有意义的生物学信号需要高效的特征选择，这能够提升下游模型性能、降低计算复杂度，并增强可解释性和可重复性[5, 6, ,]，为精准诊断、药物开发和个性化治疗奠定基础。

特征选择通常分为筛选器、包装法和嵌入法[7, 8, 9]。对于超高维数据，包装法通常在计算上不可行[10, 11, ,]，而嵌入法（例如，最小绝对收缩和选择算子和基于树的模型[12, 13]）将特征选择整合到模型训练中，但在此类场景下仍受可扩展性限制。相比之下，筛选方法无需进行模型训练，而是利用统计或相关性度量独立评估特征相关性，因此在初步筛选中具有较高的可扩展性和效率。例如，确定性独立筛选（SIS）[14] 利用边际相关性高效识别有潜力的特征；其扩展方法（如 DC-SIS）引入距离相关以捕捉非线性依赖[15,]，而 CSIS 则考虑条件贡献，以减少混杂[16]。包括肯德尔秩相关[17] 和基于球相关的确定性独立筛选（BCor-SIS）[18,] 在内的稳健替代方法，进一步增强了对非正态性和非线性效应的适应能力。总体而言，筛选方法能够以较低成本提供多视角的先验信息，但无法消除冗余特征，并且仍受其假设的特征-结果关系所约束，这可能导致遗漏关键特征和复杂交互[9]。

深度学习天生能够捕捉复杂的非线性关系，因此是特征选择的有力工具[19]。根据架构和机制的不同，基于深度学习的方法包括：基于图的方法（如 GRACES），利用样本与特征之间的关系，但可能面临可扩展性方面的挑战[2]；梯度归因方法（如 DeepLIFT 和 GradientShap），计算效率高且易于直观解释，但对参考基线和梯度噪声较为敏感[20, 21]；基于扰动的方法（如局部可解释模型无关解释（LIME）和特征消融），与模型无关且能够衡量局部重要性，但计算开销较大，并且对采样策略较为敏感[22, 23]；架构层方法（如 CancelOut 和自编码器），能够在训练过程中学习紧凑的表示，但可能存在稳定性问题，且容易受到正则化或初始化的影响[24, 25, 26]。

在深度学习方法中，注意力机制特别适用于特征选择，因为它们能够学习重要性权重，而无需进行显式的特征工程或依赖特定领域的假设[27]。注意力机制最初为神经机器翻译[28] 而开发，随后在 Transformer架构[27,] 中得到进一步发展，能够有选择地关注最相关的特征。外部注意力机制[29] 利用共享记忆单元将复杂度降至线性规模，进一步提高了计算效率，因此非常适用于特征选择。例如，Gui 等人首次提出了基于注意力机制的监督式特征选择方法 AFS[30,]。EAR-FS[31] 通过将外部注意力机制融入特征选择，进一步推进了这一方法，并展现出从高维基因组数据中识别相关特征的卓越能力。尽管具有这些优势，传统的单一注意力机制通常只沿单一维度捕捉特征关系，限制了其全面表征生物数据中复杂交互和多层级依赖关系的能力。此外，与其他深度学习方法一样，注意力机制在稳定性方面也面临挑战，包括对初始化的敏感性，以及在相同数据集上特征重要性排名的波动。这些问题对于确保可靠的生物标志物发现和稳健的

高维生物医学应用中的特征选择尤为关键。

为应对这些挑战，我们提出了 MAFS（基于多头注意力的特征选择），这是一种新颖的框架，将过滤方法与多头注意力机制相结合，并通过重排序模块识别重要特征。MAFS 首先进行基于过滤的筛选，以生成初始特征先验，从而降低对初始化的敏感性并加快收敛。随后，它运用多头注意力从多个视角并行处理特征，克服单注意力机制的局限性。最后，我们设计了一个重排序模块，用于优化各注意力头的特征集，以最大限度地减少信息损失并解决冲突。这种深度融合设计将过滤方法的稳定性和效率与深度学习的表征能力相结合，在提供可解释的特征重要性得分的同时，通过先验知识引导和增强的重排序选择策略，尽可能保留关键特征。

2 结果

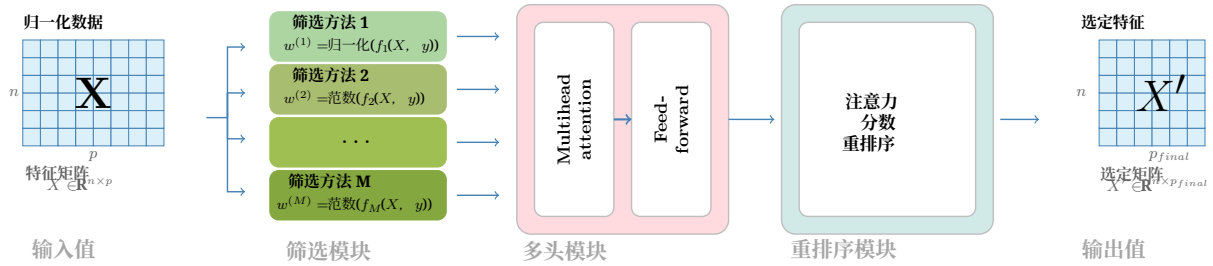
2.1 方法概述与模拟

我们开发了 MAFS（图 1a），这是一种深度学习框架，将统计过滤与基于注意力的表征学习相结合，以实现对超高维生物医学数据中高效、可扩展且可解释的特征选择。MAFS 包含三个关键步骤。首先，它应用多种经典过滤式方法生成初步相关性估计，在特定的特征—结局假设下捕捉特征重要性。其次，这些由过滤方法得出的估计值用于指导多头外部注意力机制，从多个不同视角对特征重要性进行互补评估（图 1b）。最后，通过重新排序策略整合各个注意力头的输出，以识别信息量最大的特征（图 1c）。与现有的基于深度学习的方法相比，我们的设计显著提升了特征选择准确性和计算效率。

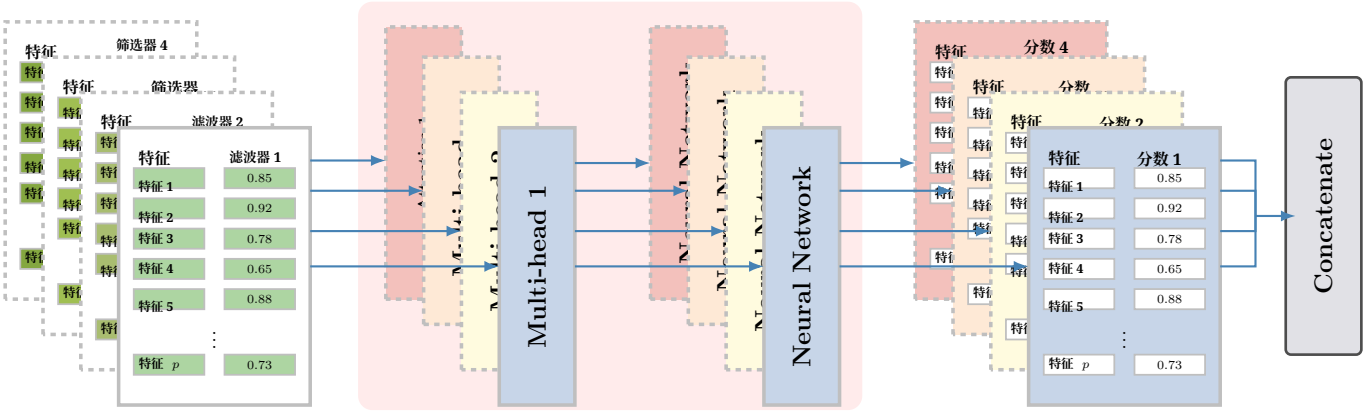
2.2 关系特异性特征选择性能

对于高维且样本量适中的数据，当预先指定的特征选择比率从 0.5% 增加至 2% 时，关系特异性特征选择性能（$n = 2,\ 000,\ p = 100,\ 000$）如图 2 和图 3 所示，分别对应连续型结局和二元结局。不同样本量（$n = 500$ 和 $n = 2,\ 000$）、维度（$p = 25,\ 000$，$p = 50,\ 000$ 以及 $p = 100,\ 000$）和结果类型（连续型和二元）的其他结果见补充图 S1–S10。在所有场景中，MAFS 在所有方法中均表现出一致的优越性能。此外，与其他方法（±2.5–3.1%）相比，MAFS 的置信区间（±2.2%）更窄，表明其在重复试验中的稳定性更高。

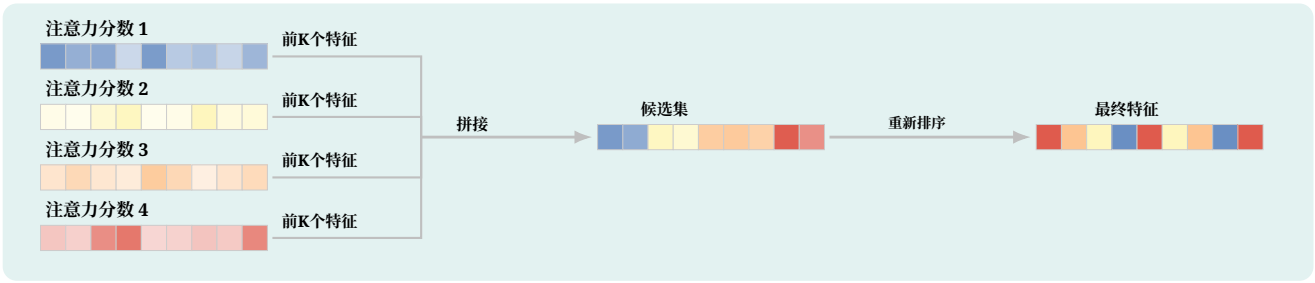
如预期，当预先指定的特征选择比例从 0.5% 增加至 2% 时，选定特征的数量也相应增加，从而使所有方法和函数形式的覆盖率提高（图 2 和 3）。然而，不同方法的改进幅度有所差异，反映出它们在选择过程早期优先识别信息性特征方面的能力不同。例如，对于连续型结局，尽管 MAFS 在所有预先指定的选择比例下均实现了最高覆盖率，但随着选择比例从 0.5% 提高至 2%，其覆盖率平均仅增加约 7.85%。这表明 MAFS 能够在早期有效识别最具信息量的特征，而扩大特征集所带来的额外收益有限。相比之下，GRACES、EAR-FS 和 DeepLIFT 的覆盖率分别提高了 21.8%、29.1% 和 79.1%，说明这些方法需要更大的特征子集才能找回更多因果特征，并且与 MAFS 相比，它们在早期进行特征优先级排序方面的效果较差。对于 CancelOut，尽管



(a) MAFS的整体架构



(b) 多头注意力模块



(c) 重排序模块

图 1 | MAFS 概览。 (a) 所提出的基于多头注意力的特征选择（MAFS）框架的整体架构，展示了三个主要模块：筛选模块、多头注意力模块和重排序模块。(b) 多头注意力模块，其中多个注意力头并行处理输入。每个注意力头都有自己的一组可学习参数，用于捕捉不同的特征模式和交互。(c) 重排序模块，该模块汇总各个注意力头的输出，以识别信息性最强的特征。

尽管取得了显著改进，但其覆盖率始终低于 0.06，反映出其在该设置下的有效性有限。

特征选择方法的相对性能在特征与结果之间不同的函数关系下存在显著差异。以连续特征和连续结果为例，在线性设置中，排名靠前的方法表现总体相近：MAFS 的覆盖率最高（0.86），GRACES（0.84）和 EAR-FS（0.70）紧随其后，而 DeepLIFT（0.26）和 CancelOut（0.05）的表现则明显较差。随着潜在关系变得更加非线性，性能差距显著扩大。在对数关系中，MAFS 达到了近乎完美的覆盖率（1.0），远超 GRACES（0.25）、EAR-FS（0.19）、DeepLIFT（0.09）和 CancelOut（0.05）。在余弦关系中也出现了类似模式：MAFS 的覆盖率达到 0.93，而最强的竞争方法 GRACES 和 EAR-FS 分别为 0.33 和 0.27。即使对于中等复杂度的关系，例如指数关系（0.91 vs. 0.88 vs. 0.74）、幂关系（0.95 vs. 0.91 vs. 0.81）和交互效应（0.56 vs. 0.43 vs. 0.37），MAFS 也始终优于 GRACES 和 EAR-FS。上述结果表明，尽管强大的基线方法在简单设置中具有竞争力，但随着函数复杂度的增加，MAFS 表现出显著更强的稳健性。

我们进一步发现，特征–结局关系的复杂程度会影响增加特征选择深度所带来的益处。对于对数函数或余弦函数等关系，增加选定特征的数量并未显著提高 MAFS 的覆盖率；即使选择比例仅为 0.5%，其覆盖率也已接近完美（$\geq$0.80）。相比之下，对于交互等更复杂的关系，扩大特征子集能为 MAFS 带来更大的提升：选择比例从 0.5% 增至 2% 时，覆盖率从 0.48 提高到 0.56。其他关系则呈现介于两者之间的趋势，覆盖率适度提高了 0.05 至 0.12。这些发现表明，当预期特征与结果之间存在更复杂的函数依赖时，选定更大的特征子集可能会有所助益。

特征选择方法的表现还受到特征和结果变量类型的影响。在考察连续型结果时，与分类特征相比，各方法通常在连续特征下取得更高的覆盖率。对于连续特征，MAFS 的覆盖率为 0.806，大幅优于 GRACES（0.594）、EAR-FS（0.488）、DeepLIFT（0.204）和 CancelOut（0.051）。当特征为分类特征时，各方法之间的表现差距有所缩小。例如，MAFS 在分类特征下的覆盖率为 0.653，而 GRACES 和 EAR-FS 分别达到 0.487 和 0.432，表明与连续特征设置相比，其相对优势有所减弱。尽管 GRACES 专为小样本设置设计，MAFS 即使在这些条件下也表现相当或更优（补充图 S1–S6）。在其他维度设置和更小样本量下，结果仍保持一致，在不同特征类型与结果类型组合中呈现出相似的相对表现模式。



图 2 | 连续型结局的关系特异性特征选择结果。 在高维设置（$n = 2,000, p = 100,000$）中，七种不同特征–响应关系在选择比例为0.5%、1%、1.5%和2%时，连续型结局的覆盖率。

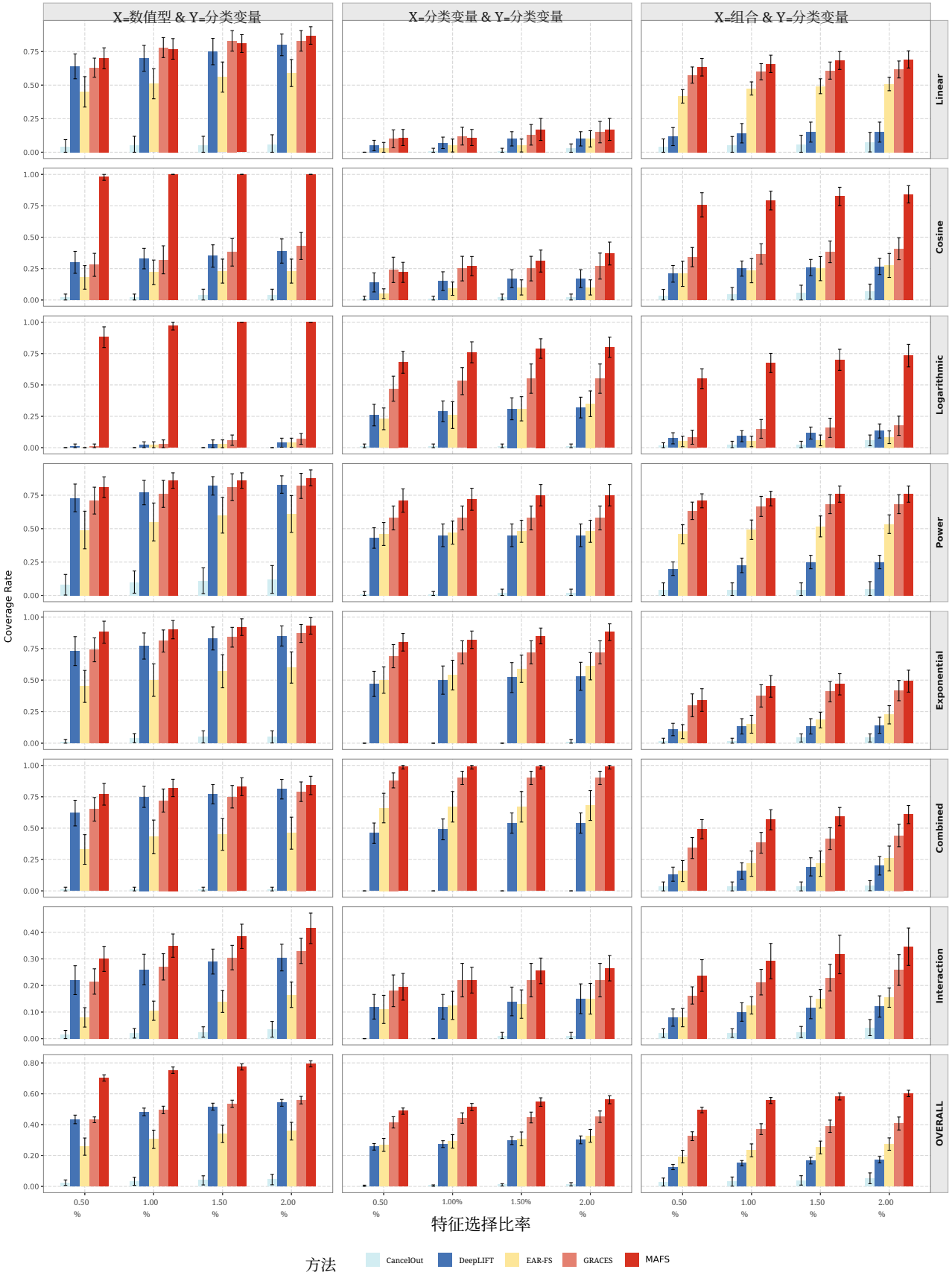


图 3 | 二元结果的关系特异性特征选择结果。 在高维设置（$n = 2,\ 000, p = 100,\ 000$）下，针对二元结果，在特征选择比例为 0.5%、1%、1.5% 和 2% 时，七种不同特征–响应关系下的覆盖率。

2.3 特征维度的影响

2.3.1 特征维度的影响：特征选择准确性

图 4 和补充图S11展示了当输入特征数量从25,000增加到 $n = 2,$,000以及 $n = 500$时各特征选择方法的性能。在所有输入维度下，无论样本量、特征类型或结果类型如何，所提出的MAFS始终具有最高的覆盖率。在大多数情况下，其性能依次仅次于GRACES、EAR-FS、DeepLIFT和CancelOut。例外情况是特征为连续型且结果为二元型，此时DeepLIFT的性能超过EAR-FS，排序为MAFS、GRACES、DeepLIFT、EAR-FS和CancelOut。

当选取排名靠前特征的2%时，对于输入维度为25,000、50,000和100,000的情况，选定特征的数量分别为500、1,000和2,000。如图 4A和补充图S11.A所示，随着维度增加，大多数方法的覆盖率保持稳定或略有提升。这种稳定性可能是因为选定特征的数量随维度成比例增长，从而有效缓解了维度灾难。

相比之下，当选择固定数量的排名靠前的特征时（即排名前100和排名前500的特征），所有方法的覆盖率都呈现出预期的下降趋势，其中 MAFS 仍是表现最佳的方法，并且覆盖率的百分比降幅最小。当选择范围限制在排名前100的特征时，随着输入维度的增加，各方法的覆盖率均有所下降。例如，当特征和结果均为连续变量且样本量为2,000时，随着输入特征数量从25,000增加到100,000，MAFS 的覆盖率从0.6下降至0.475，相当于覆盖率降低了21%。对于 GRACES（从0.375降至0.20）、EAR-FS（从0.325降至0.1）、DeepLIFT（从0.125降至0.025）和CancelOut（从0.075降至0），相应的百分比降幅分别为47%、69%、80%和100%。当特征为连续变量且结果为二元变量时，也观察到了类似的模式。在相同设置下，MAFS（从0.575降至0.375）、GRACES（从0.40降至0.225）、EAR-FS（从0.375降至0.05）、DEEPLift（从0.40降至0.20）和 CancelOut（从0.10降至0）的覆盖率分别下降了38%、44%、87%、50%和100%。尽管某些方法（例如 CancelOut）的覆盖率绝对变化可能相对较小，但其比例降幅通常较大，且整体覆盖率水平极低（例如低于5%），因此在高维设置下并不适合用于降维。当选择排名前500的特征时，也观察到了类似的下降趋势，不过随着输入维度增加，降幅更加温和。

这些模拟结果表明，维度增加会给所有特征选择方法带来巨大挑战。尽管在所有考察的输入维度下，MAFS 始终表现最佳，但在高维场景中选择稍大的特征子集，有助于找回更多与结果相关的变量，这可能有利于下游任务。

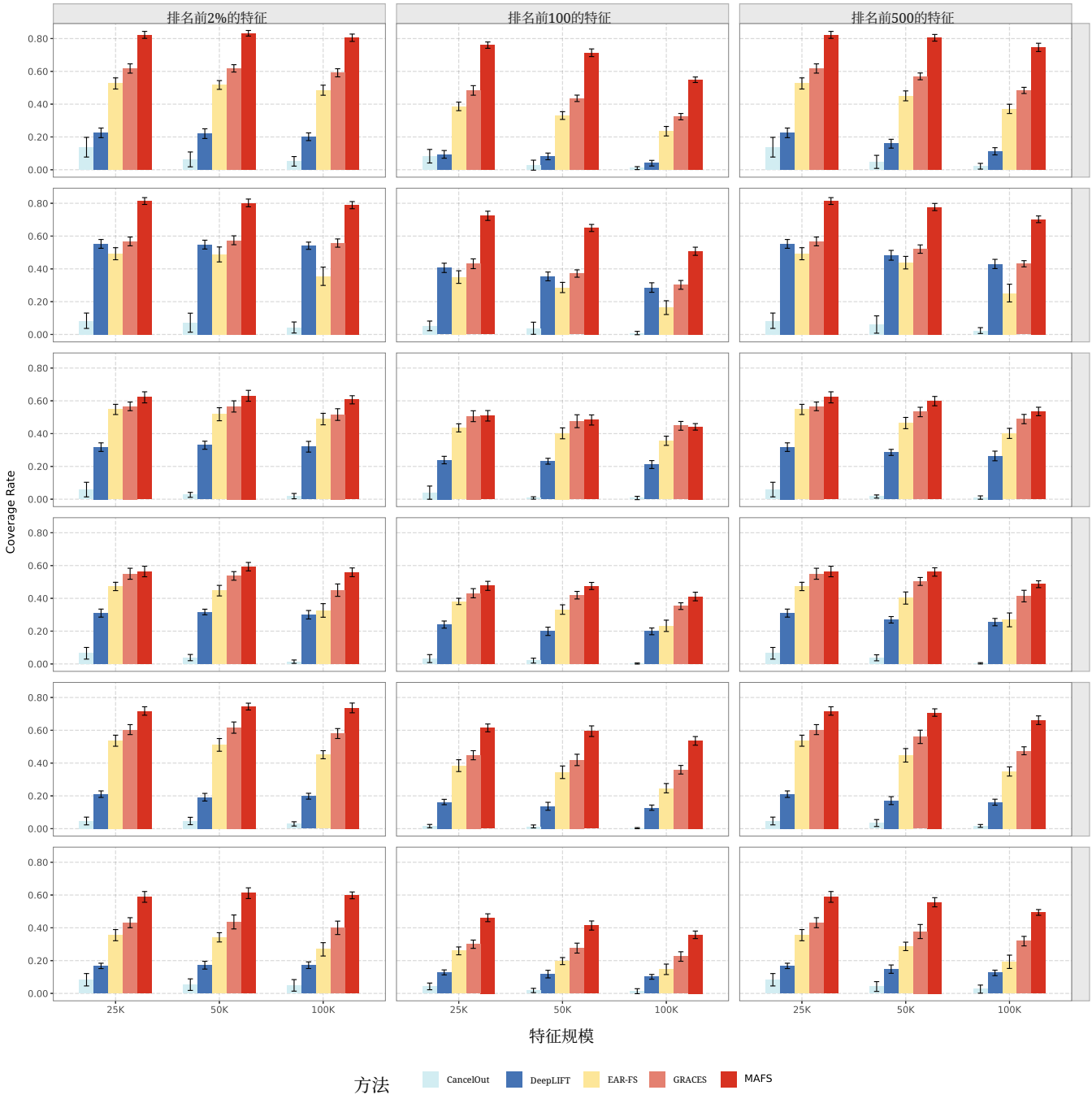


图 4 | 不同特征维度下的特征选择性能。 中等样本量（$n = 2,\ 000$）下不同输入特征维度的覆盖率。每个子图比较了三种特征维度（25K、50K 和 100K 个特征）下的五种特征选择方法。行对应于由特征类型（连续型、分类变量和组合型）与响应类型（连续型和二元）定义的六种数据类型组合。列对应于三种选择标准：前 2%、排名前 100 和排名前500的特征。

2.3.2 特征维度的影响：计算效率

表1 展示了五种特征选择方法在不同样本量（$n = 500$ 和 $n = 2,\ 000$）和特征维度（$p = 25,\ 000$、$50,\ 000$ 和 $100,\ 000$）下的运行时间分布。所有实验均在 REANNZ 高性能计算平台上进行[32], , 运行时间以分钟计。

连续型结局与二元结果之间的运行时间略有差异。对于连续型结局，MAFS 需要更长的计算时间，且随着维度增加，这一差距进一步扩大。例如，在 $p = 100,000$ 和 $n = 2,000$ 维度下，MAFS 处理连续型结局分别需要 117.6 分钟，而处理二元结果需要 88.5 分钟。其余方法在不同结果类型之间仅表现出边际差异。

在高维场景（$n = 2,000$、 $p = 100,000$）下，不同方法的运行时间差异显著。对于连续型结局，CancelOut、DeepLIFT 和 EAR-FS 的计算效率仍然较高，所需时间分别为 20.2、13.2 和 29.1 分钟，而 MAFS 需要 117.6 分钟。作为最强的基线比较方法，GRACES 的计算成本显著更高，达到 1,569.9 分钟。二元结果也呈现出类似模式：CancelOut、DeepLIFT、EAR-FS、MAFS 和 GRACES 的运行时间分别为 25.4、14.3、28.4、88.5 和 1,540.0 分钟。

尽管 MAFS 的速度慢于最快的轻量级基线方法，但它在计算效率与选择性能之间实现了良好平衡。值得注意的是，MAFS 的速度比表现最佳的基线方法 GRACES 快一个数量级以上，同时在所有评估场景中均显著提高了覆盖率。

结果	方法	样本量					
		$n = 500$			$n = 2000$		
		25k	50k	100k	25k	50k	100k
连续	MAFS	5.9 [5.6, 6.2]	11.3 [10.9, 11.7]	44.2 [38.4, 50.0]	16.7 [15.5, 18.0]	47.8 [42.7, 53.0]	117.6 [110.1, 125.0]
	GRACES	83.0 [74.6, 91.4]	176.1 [161.5, 190.6]	380.2 [348.5, 412.0]	253.4 [234.6, 272.1]	601.3 [567.8, 634.9]	1569.9 [1492.4, 1647.4]
	EAR-FS	1.0 [1.0, 1.1]	2.0 [1.8, 2.1]	4.4 [3.5, 5.4]	1.8 [1.5, 2.0]	4.2 [3.6, 4.7]	29.1 [23.7, 34.6]
	DeepLIFT	0.5 [0.4, 0.6]	3.4 [2.6, 4.2]	4.2 [3.2, 5.2]	1.3 [1.1, 1.4]	7.0 [5.5, 8.5]	13.2 [11.4, 15.1]
	CancelOut	0.6 [0.5, 0.7]	1.3 [1.1, 1.5]	4.1 [3.4, 4.8]	2.3 [2.0, 2.6]	3.9 [3.4, 4.3]	20.2 [15.0, 25.4]
二元	MAFS	5.5 [5.2, 5.8]	11.2 [9.7, 12.7]	40.6 [33.7, 47.6]	11.3 [9.9, 12.7]	26.4 [23.1, 29.8]	88.5 [81.8, 95.2]
	GRACES	86.9 [79.6, 94.3]	169.8 [130.6, 209.1]	392.8 [374.0, 411.6]	285.7 [260.6, 310.8]	555.7 [522.4, 589.0]	1540.0 [1421.6, 1658.4]
	EAR-FS	1.0 [1.0, 1.1]	2.0 [1.8, 2.1]	6.1 [5.0, 7.1]	1.7 [1.5, 1.8]	4.7 [4.2, 5.2]	28.4 [24.7, 32.1]
	DeepLIFT	0.6 [0.5, 0.7]	2.9 [2.3, 3.5]	6.4 [4.9, 7.8]	1.6 [1.5, 1.8]	7.9 [6.5, 9.3]	14.3 [12.5, 16.1]
	CancelOut	0.8 [0.6, 0.9]	1.4 [1.2, 1.7]	4.4 [3.5, 5.3]	1.9 [1.6, 2.1]	3.5 [3.0, 4.0]	25.4 [22.7, 28.1]

表。 1 | 五种特征选择方法在不同样本量和特征维度下的运行时间（分钟），并报告95%置信区间。

2.4 真实数据分析

2.4.1 癌症基因表达数据集分析

图 5 展示了所有方法在六个癌症数据集上的AUROC。总体而言，多层感知机的分类性能最高，其次是支持向量机和K近邻，不过这三种分类器之间的差异较小。值得注意的是，特征选择方法的相对性能因数据集而异，但在不同分类器下表现一致。在容易分离的数据集（即白血病和前列腺癌）上，大多数方法均取得了较高的性能（AUROC ≥ 0.95）。在难度适中的数据集（胶质瘤）上，MAFS在支持向量机和K近邻下优于所有基线方法，但在多层感知机下与GRACES表现相当（AUROC ≈ 0.95）。在更具挑战性的数据集（结肠癌和肺癌）上，MAFS显著优于所有基线方法。在多层感知机下选取20个特征时，MAFS在结肠癌数据集上的AUROC达到0.922，比GRACES高0.026，比EAR-FS高0.067；在肺癌数据集上达到0.785，比GRACES高0.022，比其他方法高出0.030以上。

MAFS 在所有三种分类器中都保持了稳定的相对表现。尽管不同分类器的绝对表现有所差异，但各方法的相对排序始终稳定。在由 3 种分类器和 6 个癌症数据集组成的 18 个测试场景中，MAFS 在 18 个场景中的 17 个取得了最高或并列最高的表现。唯一的例外是使用支持向量机作为分类器的前列腺癌数据集场景；在该场景中，MAFS 的表现与 DeepLIFT 相当，同时仍优于所有

其余对比方法。这种跨分类器的一致性表明，性能差异反映的是特征选择的有效性，而非对特定分类器架构的敏感性。

MAFS 还展现出高效的特征利用能力，这表明与基线方法相比，MAFS 能更快地收敛到最优特征子集。例如，在白血病数据集上，MAFS 仅需 2-3 个特征即可达到 AUROC > 0.95，而 GRACES 需要 4-6 个特征，EAR-FS 则需要 8-10 个特征才能达到相近的性能。在结肠数据集上，MAFS 使用 5 个特征便达到 0.90 的 AUROC，而基线方法需要 10-15 个特征。在其他数据集上也观察到类似的规律，这表明 MAFS 使用规模较小的特征子集即可实现高性能，从而可能降低计算需求，并增强下游分析的可解释性。

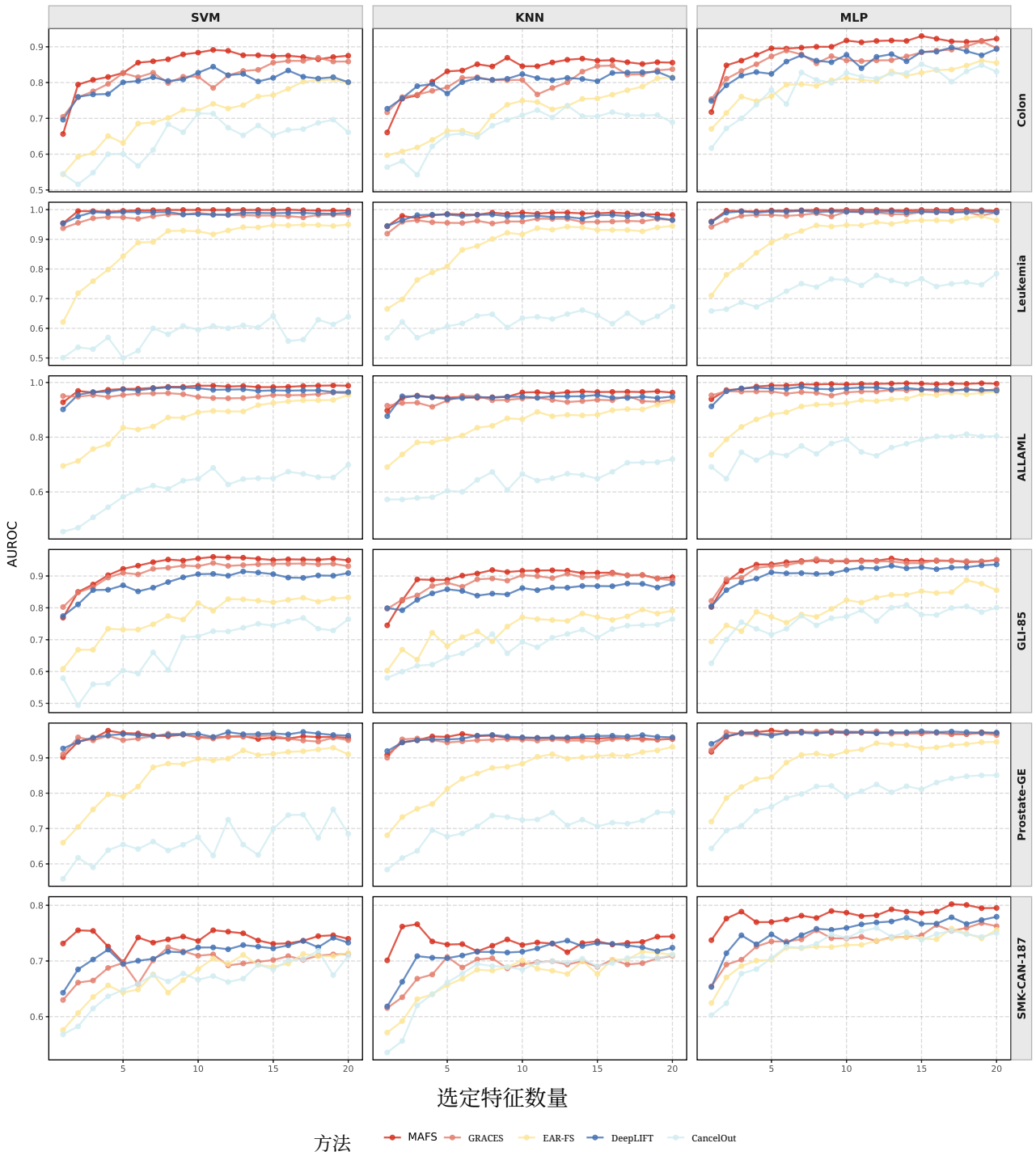


图 5 | 癌症基因表达数据集上的评估。 六个癌症数据集中各特征选择方法的 AUROC 性能比较。每行对应一个数据集，每列代表一种不同的分类器：支持向量机（左）、K近邻（中）和多层感知机（右）。

2.4.2 阿尔茨海默病神经影像学计划数据集分析

图 6 展示了所有特征选择方法在九种基于图像的表型上的皮尔逊相关系数。总体而言，相关性数值在不同脑区之间存在差异，这表明基因表达的影响具有脑区特异性。在大多数脑区和回归器中，MAFS 通常取得高于或相当于基线方法的相关性，表明其能够有效选择

用于连续表型预测的预测特征。

不同脑区的相关性数值存在差异。在选取 50 个特征时，脑干和苍白球中的相关性较高，MAFS 分别达到 0.440 和 0.395（MLP），而 GRACES 分别为 0.418 和 0.375，EAR-FS 分别为 0.409 和 0.329，DeepLIFT 分别为 0.228 和 0.190，CancelOut 分别为 0.288 和 0.222。在杏仁核、第四脑室、壳核和丘脑中观察到中等程度的相关性，其中 MAFS 始终达到或略高于基线方法。例如，在第四脑室中，MAFS 的相关性为 0.372，而 GRACES 为 0.344，DeepLIFT 为 0.192，EAR-FS 为 0.175，CancelOut 为 0.214。在伏隔核区、尾状核和海马中发现较低的相关性，但 MAFS 仍在各方法中达到最高或相当的数值（例如，伏隔核区中为 0.219，而 GRACES 为 0.158，DeepLIFT 为 0.125，EAR-FS 为 0.100，CancelOut 为 0.131）。

MAFS 在不同回归器下均保持了稳定的相对性能。尽管绝对相关值有所变化，但各方法的排序总体上保持一致。在 27 个测试场景（9 个区域$\times$ 3 回归器）中，MAFS 在大多数场景下排名第一，并在其余场景中表现出竞争力，这表明其性能优势源于有效的特征选择，而非对特定回归器架构的敏感性。MAFS 还展现出高效的特征利用能力，以少于基线方法的特征数量达到较高的相关值。这种高效的特征使用方式可能降低计算成本，并有助于下游神经影像分析中的解释。

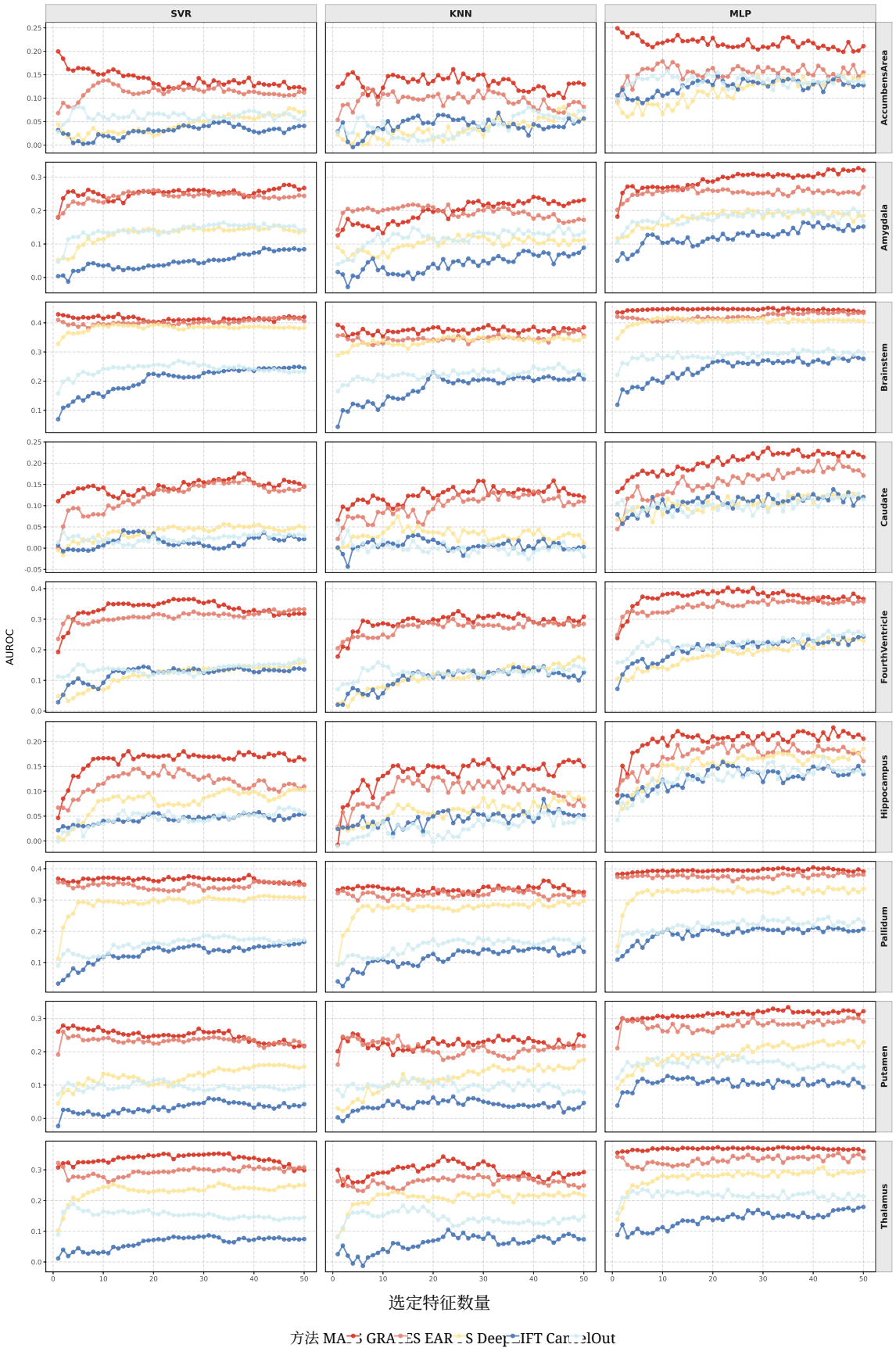


图 6 | ADNI 数据集上的评估。 比较不同特征选择方法在 ADNI 数据集上的相关性表现。每一行对应一个脑区，每一列表示使用不同回归模型所获得的结果：支持向量回归（左）、K近邻回归（中）和多层感知机回归（右）。

3 讨论

在本研究中，我们提出了 MAFS，这是一种统一的特征选择框架，旨在应对高维生物医学数据特征选择中的三项关键挑战：1）传统过滤器–包装器流程在初步筛选过程中造成的不可逆信号损失；2）单一标准方法对特征相关性的评估范围狭窄；3）深度学习方法常见的初始化敏感性问题，容易导致模型不稳定。MAFS 将多种过滤方法整合到多头注意力架构中，将统计相关性得分转化为各注意力头的软性、可学习先验，从而解决这些问题。这些先验引导优化过程，但不会充当硬阈值，并且能够通过反向传播进行调整。该设计既保留了过滤方法的计算效率和可解释性，又利用注意力机制的表征能力和多视角建模能力，捕捉高维分子数据中的复杂依赖关系。不同于现有通过多数投票汇总多视角信息的嵌入方法，我们设计了一个重排序模块，即使某些重要特征只被部分视角识别，也能将其捕捉到。通过开展广泛的模拟，并分析多个具有分类变量和连续型结局的真实数据集，我们证明，MAFS 的性能优于或媲美表现最佳的常用深度学习特征选择方法。

模拟实验揭示了两个值得注意的趋势。首先，尽管随着特征维度增加，所有方法的覆盖率都呈下降趋势（图 4），但即使在最具挑战性的设置下，MAFS 仍保持较高的准确性。这种稳健性很大程度上归功于信息引导初始化，与随机或均匀的初始值相比，它能带来更有利的优化景观。信息引导初始化使模型更接近最优参数区域，从而加快收敛，并帮助注意力机制从训练初期就聚焦于信息性特征[33]。其次，在七种特征–结局关系中，MAFS 在几乎所有配置下都取得了最高的覆盖率，尤其是在对数变换关系及其他非线性依赖关系中优势显著（图 2 和 3）。多头架构使互补的统计视角能够并行发挥作用，从而减轻单一评判标准的局限性。采用不同过滤方法初始化的各个头能够捕捉共同及独特的特征–结果依赖关系，降低因单一评判标准而遗漏重要特征的风险。结合深度神经网络学习非线性效应和高阶效应的固有能力，MAFS 非常适用于数据维度高、样本量有限[2]，且生物学关系[34] 复杂多样的生物医学应用。尽管多头注意力会略微增加计算开销（表 1），但 MAFS 的速度仍远快于表现最佳的基于深度学习的特征选择方法（即基于图的 GRACES），在准确性和效率之间实现了良好平衡。

在真实数据集上的应用进一步凸显了 MAFS 的实用价值及其相较于现有方法的优势。在癌症基因表达数据集上，MAFS 的表现与 GRACES 和 DeepLIFT 等强基线方法相当或更优，展现了其在高维、小样本量场景中的稳健性。在易于区分的数据集（白血病和前列腺癌数据集）上，性能差异很小，这可能是因为强烈的特征–结局关联使大多数方法都能识别判别信号，从而降低了精细排序的影响。相比之下，在更具挑战性的数据集（即结肠癌和肺癌数据集）上，MAFS 通过信息引导初始化和多头注意力实现的多视角学习，能够更有效地识别具有不同关系的信息性特征，显著提升下游分类性能。在对 ADNI 神经影像学结局的分析中也观察到了类似模式，其中信号较弱，预测任务也更困难。要选择能够预测连续认知评分的特征，需要捕捉细微的渐变关联，而较高的噪声会放大特征排序不理想所带来的影响。在各脑区中，MAFS 始终

优于所有基线方法，这凸显了其三个核心模块——引导式初始化、多头注意力和重排序——在弱信号情形下生成稳定且具有生物学意义的特征排名方面所作出的贡献。

值得注意的是，仅我们提出的滤波器权重引导的初始化就能提升基于注意力的特征选择性能。这与先前的研究结果一致，即深度网络采用更优的初始值不仅能够提升性能，还能加速其向最优解收敛[33]。在模拟数据（补充图 S12 - S17）、癌症基因表达数据集（补充图 S18）和 ADNI 神经影像数据（补充图 S19）上，用基于筛选器的值替代 EAR-FS 中的均匀初始化，均能持续提升性能。尽管多头注意力带来了额外收益，但这种无需修改架构或训练方案的简单初始化变更能够持续改善性能，这一事实表明，对于基于注意力的方法而言，初始化质量可能与架构复杂性同等重要。这些发现表明，有原则的初始化是一种具有可迁移性且成本低廉的策略，可用于增强现有的基于注意力的框架，而无需进行完整的架构重新设计。

尽管 MAFS 相较于现有方法展现出明显优势，但仍有若干局限性值得考虑。首先，虽然采用了归一化以确保不同筛选器得到的权重处于可比尺度，但简单的归一化仍可能引入尺度失配和排序不稳定性。探索替代性的融合方案，例如可学习尺度参数或基于排名的聚合，或许有助于解决这一问题。其次，MAFS 需要完整的特征矩阵，这可能会排除含有缺失值的样本，并在生物医学数据集中引入选择偏差。引入插补或缺失数据感知机制是未来研究的一个重要方向。

总之，MAFS 将传统过滤方法与深度学习相结合，以应对高维生物医学数据中的特征选择挑战。通过信息引导初始化、多头注意力和有效的重排序，它在不同数据类型、函数关系和维度设置下实现了更优的特征优先级排序，并展现出稳健的性能。

4 方法

4.1 模型

所提出的 MAFS模型通过包含三个核心组件的渐进式架构设计，解决了超高维生物数据特征选择中的关键挑战（图1）。筛选模块采用多种基于过滤的特征选择方法，根据各自假设的特征与结果关系，对特征重要性进行初步评估。由此得到的权重为下游学习提供先验知识，从而增强模型的稳定性并促进收敛。多头注意力核心模块从多个视角评估特征，其中每个注意力头专门处理一种筛选方法的输出，同时从互补的角度评估特征重要性。生成的注意力权重增强了模型的可解释性和透明度，使人们能够更清楚地理解特征重要性和选择依据。此外，多头模块采用外部注意力替代自注意力，将计算复杂度从 $O(n^2p)$ 降至 $O(np)$，从而显著提升效率。重排序模块引入了一种改进的细化机制，对各注意力头的特征集进行重排序和整合，大幅减少信息损失和头间冲突。所提出的 MAFS 的伪代码见算法1。

算法 1 MAFS 特征选择 1：输入：注意矩阵 $X \in \mathbb{R}^{n\times d}$、目标向量 $y \in \mathbb{R}^n$、过滤方法集 $F = \{f_1, f_2, \ldots, f_k\}$2：输出：所选特征子集索引 3：模块1：过滤权重初始化 4：对于 $j = 1$ 到 k，执行▷ 计算滤波方法 j 的特征重要性 5：$w^{(j)} \leftarrow f_j(X, y)$ 6：$w^{(j)}_{\text{norm}} \leftarrow (w^{(j)} - \mu_j)/\sigma_j$ ▷ 归一化处理 7：结束循环 8：模块2：多头注意力处理 9：初始化注意力头参数 $\{W_h, b_h\}$，对于 $h = 1$、...、M，其中 $M = |F|$ 10：对于训练轮次 $= 1$ 到最大训练轮次，执行 11：对于 $h = 1$ 到 M，执行 12：$\boldsymbol{\alpha}^{(h)} \leftarrow W_L^{(h)T}\text{ReLU}\Big(\cdots\text{ReLU}\Big(W_1^{(h)T}\boldsymbol{w}^{(h)}_{\text{norm}} + b_1^{(h)}\Big)\cdots\Big) + b_L^{(h)}$ ▷ 注意力权重 13：$X^{(h)} \leftarrow X \odot \boldsymbol{\alpha}^{(h)}$ ▷ 软特征选择 14：结束循环 15：$L^{(h)}$ 总计 $= \begin{cases} -\sum_{i=1}^{n}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}^{(h)}) + \lambda\cdot\sum_{k=1}^{d}\tau_k^{(h)}|\alpha_k^{(h)}| & \text{用于分类任务（^）} \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{(h)})^2 + \lambda\cdot\sum_{k=1}^{d}\tau_k^{(h)}|\alpha_k^{(h)}| & \text{用于回归任务} \end{cases}$ 16：结束循环 17：模块3：特征重排序 18：$S \leftarrow \emptyset$ 19：对于 $h = 1$ 到 M，执行 20：$S \leftarrow S \cup \text{TopK}(\boldsymbol{\alpha}^{(h)}, K)$ 21：结束循环 22：训练重排序模型 (S, y) 以获得特征重要性得分 23：最终特征 ← 排序并选择最终特征子集 24：返回最终特征

4.1.1 模块1：过滤权重初始化模块

过滤权重初始化模块旨在整合多种基于过滤的特征选择方法所提供的先验知识，为下游多头注意力模块提供稳定且有效的初始权重。该模块通过采用多种基于过滤的特征选择方法（例如 SIS、BCor-SIS 和肯德尔秩相关系数），捕捉多样化的特征–结局关系，每种方法都代表一种不同的统计视角。该模块从一组过滤方法 $F = \{f_1, f_2, ..., f_k\}$中提取先验特征重要性信息，其中每种过滤方法 f_j 独立计算特征重要性向量 $w^{(j)} = f_j(X, y) \in \mathbb{R}^d$。这里，$X \in \mathbb{R}^{n\times d}$ 表示特征矩阵，$y \in \mathbb{R}^n$ 表示目标向量。

每个筛选器向量 $w^{(j)}$ 随后进行归一化：$w^{(j)}_{\text{norm}} = \frac{w^{(j)} - \mu_j}{\sigma_j}$，其中 μ_j 和 σ_j 分别表示来自第 j 个过滤方法的权重均值和标准差。归一化可避免取值范围较小的筛选权重产生退化的注意力输入。尽管归一化可能会略微扭曲原始权重分布[35]，，并可能削弱某些过滤方法所表达的强偏好信号，但它能确保多头注意力模块中的每个注意力头都接收到有意义且可比较的输入信号，同时保留各过滤方法的独特特性。这些归一化后的筛选权重提供了一个结构化的起点，取代随机初始化，从而有助于后续注意力模块实现更稳定、高效的优化。

4.1.2 模块2：多头注意力模块

多头注意力核心以筛选模块为基础，从多个角度并行评估特征。通过让每个头专注于不同的先验知识来源（例如 SIS、BCor-SIS 和肯德尔秩相关系数的权重），模型能够同时关注线性、非线性以及基于秩的依赖关系。与逐一评估特征的传统过滤方法不同，注意力模块中的每个头都会考虑多个特征的联合贡献，从而捕捉高阶交互。通过整合各头的输出，我们的多头设计能够提取特征相关性的验证性信号和互补信号，提升稳健性，并捕捉单头注意力经常忽略的多维依赖关系。

图 1b 展示了多头注意力模块的架构。注意力头的数量等于模块1中的过滤方法数量（即 $M = |F|$）。每个注意力头通过专用的多层感知机独立处理其分配到的滤波权重，从而避免不同先验知识来源之间的干扰。所有注意力头的变换表示都会被输入各自的前馈网络，以聚合样本级信息，并为每个注意力头生成特征级重要性得分。对于第 h 个注意力头（$h \in \{1, 2, \ldots, M\}$），注意力权重 $\boldsymbol{\alpha}^{(h)}$ 通过一个 L层多层感知机计算如下：

$$\boldsymbol{\alpha}^{(h)} = W_L^{(h)T}\mathrm{ReLU}\left(W_{L-1}^{(h)T}\mathrm{ReLU}\left(\cdots\mathrm{ReLU}\left(W_1^{(h)T}\mathbf{w}_{\mathrm{norm}}^{(h)} + \mathbf{b}_1^{(h)}\right)\cdots + \mathbf{b}_{L-1}^{(h)}\right)\right) + \mathbf{b}_L^{(h)}, \tag{1}$$

其中，$\{W_l^{(h)}, \mathbf{b}_l^{(h)}\}_{l=1}^{L}$ 表示第 h 个注意力头的可学习参数，L 为可配置的超参数。令 $\alpha_k^{(h)}$ 表示 k 的第 $\boldsymbol{\alpha}^{(h)}$个元素，代表根据注意力头 h对特征 k 的重要性得分。

The 软特征选择机制通过逐元素乘法实现 :

$$\mathbf{X}^{(h)} = \mathbf{X} \odot \boldsymbol{\alpha}^{(h)}, \tag{2}$$

其中，$\odot$ 表示哈达玛积。这种方法能够进行连续的特征加权，而非分类变量选择，使模型能够捕捉细微的重要性关系。由此得到的注意力权重分布 $\boldsymbol{\alpha}^{(h)}$ 能够以透明且可解释的方式评估特征重要性。通过比较不同头的权重分布，用户可以了解各种特征选择视角之间的一致性和差异。

第 h 个头的总体损失（$h \in \{1, 2, \ldots, M\}$）将主要任务损失与特征选择正则化项结合起来：

$$L_{\mathrm{total}}^{(h)} = L_{\mathrm{pred}}^{(h)} + \lambda \sum_{k=1}^{d} \tau_k^{(h)} |\alpha_k^{(h)}|, \tag{3}$$

其中，$L^{(h)}$ 预测 表示头 h 的预测损失（分类任务使用交叉熵，回归任务使用均方误差），d是特征总数，λ 是用于控制预测性能与特征稀疏性之间权衡的正则化强度，$\tau_k^{(h)}$ 是我们设计的特征—头特定自适应权重，其思路与自适应套索类似。具体而言，为利用每种滤波方法中的先验知识，并促使不相关特征的效应为零，我们根据初始滤波权重将 $\tau_k^{(h)}$ 定义为：

$$\tau_k^{(h)} = \min\left(\frac{1}{(|w_{\mathrm{norm},k}^{(h)}| + \epsilon)^{\gamma}}, \tau_{\max}\right), \tag{4}$$

其中，$w^{(h)}$ 归一化 k 是来自模块1的归一化过滤权重向量$\mathbf{w}_{\mathrm{norm}}^{(h)}$ 的第 k个元素，表示分配给头的滤波方法中第 k个特征的归一化初始权重

h， ϵ 是一个用于防止数值不稳定性的小常数， γ 控制惩罚项的自适应程度， $\tau_{\max}$ 则限制惩罚项的最大值，以防止出现极端值。这种自适应机制会对先验支持较强（ $|w^{(h)}$ 归一化 $_k|$ 较高）的特征施加较小的惩罚，对先验支持较弱的特征施加较大的惩罚，使每个头都能根据其特定的先验知识来源调整特征重要性，同时通过 L_1 正则化保持稀疏性。

请注意，尽管 MAFS 采用多头架构，但 M 个注意力头并行运行，彼此之间不进行融合。每个注意力头都会独立生成自己的特征选择结果和预测，从而能够同时探索多样化的特征子集。这种设计便于从不同统计视角进行比较分析，并通过类似集成的评估支持稳健性评估。

4.1.3 模块3：特征重排序模块

特征重排序模块是 MAFS 的最终优化阶段。其目标是在尽可能减少信息损失的同时，整合多个注意力头识别出的信息性特征。传统方法通常通过简单投票或取平均的方式聚合基于筛选器的方法的输出，但这种策略可能会遗漏仅在特定统计准则下具有高度信息性的特征。为克服这一局限，MAFS 采用候选集构建与重新排序策略（见图1c），使模型能够从多个视角捕捉互补信号，而不会过早舍弃可能重要的特征。

对于每个注意力头 $h \in \{1, \ldots, M\}$，我们根据其学习到的注意力权重选取排名前 K 个特征。候选集 S 由所有注意力头选出的特征合并而成：

$$S = \bigcup_{h=1}^{M} \mathrm{TopK}(\boldsymbol{\alpha}^{(h)}, K), \tag{5}$$

其中，TopK($\boldsymbol{\alpha}^{(h)}$， K) 返回对应注意力头中注意力权重最高的 K 个特征的索引。由此得到的候选集满足 $K \le |S| \le M \cdot K$，其实际大小取决于不同注意力头之间的特征重叠情况。

S 中的特征会使用任何能够从 ($\mathbf{X}_S$， $\mathbf{y}$) 估算特征重要性的算法重新排序，其中 $\mathbf{X}_S$ 表示仅包含 S 中特征的数据矩阵。将超高维特征筛选至集合 S 后，即可应用大多数现有的特征重要性算法，包括包装式方法。适用的方法包括基于模型的重要性（例如随机森林、极端随机树）和基于扰动的方法（例如置换重要性、SHAP 值）。请注意，为避免评估偏差，不应在此阶段重复使用模块1中采用的基于过滤器的方法。在我们的实现中，我们采用了基于树的模型，并根据基尼不纯度对特征重新排序：

$$\beta_j = \frac{1}{N_{\text{trees}}} \sum_{t=1}^{N_{\text{trees}}} \mathrm{Gini}_j^{(t)}, \tag{6}$$

其中，$\mathrm{Gini}_j^{(t)}$ 表示树 t 中归因于特征 j 的基尼不纯度下降幅度。特征按 $\{\beta_j\}$ 降序排列，得到最终排序 $\mathcal{R}$，再根据应用需求或验证集标准选取排名前 ℓ 的特征。

4.2 常用的基于深度学习的特征选择方法

我们将所提出的方法与四种广泛使用的基于深度学习的特征选择方法进行了比较：GRACES[2], E AR-FS[31], DeepLIFT [20], 和 CancelOut[24]。选择这些方法进行比较，

是因为它们代表了基于深度学习的特征选择的多种范式，涵盖基于层的门控、基于梯度的归因、注意力机制和基于图的学习。

CancelOut 提出了一种基于可学习门控层的特征选择方法[24]。该方法将定制的特征选择层集成到深度神经网络中，在模型训练期间自动学习每个输入特征的重要性权重。该层通过逐元素加权筛选输入特征，其输出定义如下：

$$\mathrm{CancelOut}(\mathbf{X}) = \mathbf{X} \odot \sigma(\mathbf{W}_{CO}), \tag{7}$$

其中，$\mathbf{X}$ 是输入矩阵（$n \times p$），$\mathbf{W}_{CO}$ 是可学习的权重向量，$\sigma(\cdot)$ 是 Sigmoid 激活函数，而 $\odot$ 表示逐元素乘法。在训练期间，CancelOut 层通过在损失函数中加入两个正则化项来约束权重分布，从而实现特征选择：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} - \lambda_1 \cdot \mathrm{var}\left(\frac{\mathbf{W}_{CO}}{p}\right) + \lambda_2 \left\|\frac{\mathbf{W}_{CO}}{p}\right\|_1, \tag{8}$$

其中，$\mathcal{L}_{\text{pred}}$ 是损失，p 是特征数量，λ_1 和 λ_2 分别控制方差项和稀疏项的权重。训练完成后，通过对学习到的权重 $\mathbf{W}_{CO}$应用 Sigmoid 激活函数来估计特征重要性，并根据 $\sigma(w_{CO,\ i})$对特征进行排序。在所有分析中，我们均采用默认设置，并将 λ_1 和 λ_2 视为需要调优的超参数。

DeepLIFT 是一种基于梯度的归因方法，它通过将输出分解为各输入特征的贡献来解释模型预测[20]。该方法将每个神经元的激活与其参考激活之间的差值作为特征重要性分数，然后通过改进的反向传播规则在网络中反向传播。根据归因分数的绝对值对特征进行排序。在我们的分析中，参考输入使用默认的零向量。

EAR-FS 是一种基于注意力的特征选择方法[31]。它通过嵌入在多层感知机网络中的外部注意力机制学习特征重要性。注意力模块是一个可学习向量 $\mathbf{a} \in \mathbb{R}^p$（其中 p 表示特征数量），经过 sigmoid 激活后转换为 $\mathbf{a}^* = \sigma(\mathbf{a})$，并与输入特征进行逐元素相乘。在训练过程中，网络通过包含正则化项的损失函数进行优化：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \frac{\lambda}{\sum_i ((a_i^* - 0.5)^2)}, \tag{9}$$

其中，第一项为损失，λ为正则化系数，第二项则对接近 0.5 的注意力权重施加惩罚，以促使重要特征与不重要特征之间形成明确区分。训练完成后，根据注意力值 a_i^*对特征进行排序。在我们的分析中，注意力权重采用默认设置进行初始化，并对超参数 λ 进行了调优。

GRACES 是一种基于图的特征选择方法[2]。它通过在动态构建的样本相似度图上训练的图卷积网络（GCN）的输入梯度来学习特征重要性。在构建图时，仅当样本对的余弦相似度超过第 α 百分位阈值时，才会连接这些样本。GCN 通过最小化特定任务的损失进行训练，并通过随机梯度估计来估算特征重要性：对训练好的网络执行 $m = 10$ 次独立的丢弃法操作（丢弃率为 $= 0.5$），每次操作中都向 GraphSAGE 层的权重添加高斯白噪声。计算得到的平均梯度矩阵为：

$$G = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \mathcal{L}}{\partial W_{\text{input}}^{(i)}}, \tag{10}$$

其中，$W^{(i)}$ 输入 表示第 i 次丢弃法操作中的输入权重，$m = 10$。基于梯度的特征分数通过计算每个特征对应的梯度列得到：$g_j = \|G_{\cdot j}\|_2$。为增强稳健性，GRACES 将基于梯度的分数与 ANOVA F-统计量 f_j 进行加权融合：

$$s_j = f_{\text{correct}} \cdot g_j + (1 - f_{\text{correct}}) \cdot f_j, \tag{11}$$

其中，$f_{\text{correct}} \in [0, 1]$ 控制梯度证据和统计证据的相对贡献。通过选择 $j^* = \arg\max_{j \notin S} s_j$，迭代选取特征，并使用更新后的特征子集重新训练网络，直至达到目标数量。在我们的分析中，对超参数 α 和 f_{correct}进行了调优。

4.3 模拟

为评估 MAFS 的性能，我们开展了两项模拟实验：(1) 函数关系分析系统地评估了在预先设定的不同特征选择比例下，模型性能如何随不同特征-结果关系（线性、非线性和基于交互的关系）而变化。(2) 维度可扩展性分析考察了模型性能在不同特征维度下的变化。两项模拟均旨在反映生物医学数据的现实特征，包括高维度、异质的特征类型以及复杂的函数关系。

我们通过组合三种特征类型（连续基因表达、分类 SNP 基因型或二者组合）与两种结果类型（连续表型或二元疾病状态），生成了六种特征–结果组合。此 3×2 设计反映了常见的生物医学研究场景：使用连续特征和结果进行生物标志物定量；使用连续或组合特征及分类结局进行疾病分类；以及使用分类特征和结果进行基因型–表型关联。

4.3.1 模拟数据

为反映真实情境，我们使用 OmicsSIMLA 软件模拟了高维基因表达（连续型）和 SNP（分类变量）数据[36]，。该软件是一款功能强大的模拟器，能够生成真实的多组学数据集，同时保留底层数据结构及组学间的依赖关系。

对于 SNP 数据，OmicsSIMLA 需要由序列模拟器生成的外部参考单倍型。在此，我们使用 HAPGEN2[37]，根据千人基因组计划中的 CEPH 人群模拟了一个参考面板。该参考面板体现了欧洲人群真实的连锁不平衡模式和等位基因频率分布。我们在所有常染色体上模拟个体基因型，并将二倍体基因型编码为 0、1 或 2，分别表示次要等位基因的数量。常染色体基因组中约有 2400 万个 SNP。我们进行了标准质量控制（QC），并排除了符合以下条件的 SNP：（1）在所有样本中均无变异；或（2）与滑动窗口内的邻近变异存在较强连锁不平衡（$r^2 > 0.2$），窗口大小为 200 个 SNP（步长为 5 个 SNP）。质量控制后，约保留 35 万个 SNP，筛除率为 98.5%。

对于 RNA 测序基因表达数据，OmicsSIMLA 使用负二项分布对读数进行建模，分布参数根据真实转录组数据学习得到。具体而言，我们对基因平均表达量和离散参数进行校准，使其与癌症基因组图谱（TCGA）中的乳腺浸润性癌（BRCA）队列相匹配。这确保模拟数据能够捕捉真实组织中观察到的生物变异性。为模拟基因型与基因表达之间的真实调控关系，我们纳入了表达数量性状位点（eQTL）。eQTL 目录源自 GTEx 乳腺组织[38]。我们检索了显著的

顺式表达数量性状位点，满足双重阈值：基于置换的 $p < 10^{-8}$ 和基于 beta 近似的置换 $p < 0.05$。筛选出常染色体表达数量性状位点，并将等位基因和坐标与基于 CEPH 的 SNP 面板进行协调后，我们将生成的基因-SNP 映射提供给 OmicsSIMLA。这使模拟的基因表达能够由相应的基因型明确调节。按照标准的 RNA-seq 预处理流程，我们获得了归一化表达矩阵，并剔除了样本间无变异的基因。最终数据集包含约 26,000 个基因，其表达值在归一化后呈近似正态分布。这与典型的对数转换 RNA-seq 数据一致。

我们使用七种不同的函数形式对特征-结果关系进行建模：线性、余弦、对数、三次、指数、复合以及余弦-指数交互。每个样本的均值定义见公式 12：

$$\begin{aligned}\mu_i = {} & \beta_1 \sum_{a\in A} x_{i,a} + \beta_2 \sum_{b\in B} \cos(x_{i,b}) + \beta_3 \sum_{c\in C} \log(|x_{i,c}| + 10^{-6}) + \beta_4 \sum_{d\in D} x_{i,d}^3 \\ & + \beta_5 \sum_{e\in E} \exp(x_{i,e}) + \beta_6 \sum_{(g,k)\in(G,K)} \cos(x_{i,g_k}) \cdot \exp(x_{i,h_k}) \\ & + \beta_7 \Big[\sum_{f\in F} \cos(x_{i,f}) + \sum_{f\in F} \log(|x_{i,f}| + 10^{-6}) + \sum_{f\in F} x_{i,f}^3 + \sum_{f\in F} \exp(x_{i,f}) \Big], \end{aligned} \tag{12}$$

其中，$\{A, \ldots, H\}$ 表示将因果特征划分为互不相交的特征索引集，并为每个集合指定一种函数形式，以避免重叠。这一设计确保对每种关系类型进行独立评估。为提高对数变换的数值稳定性，我们添加了一个较小的常数 10^{-6} 。对于连续或分类特征场景，我们从可用特征池中选取了 40 个因果特征。其中，六种非交互函数形式各包含 5 个特征，交互项包含 10 个特征。对于同时涉及连续和分类特征的组合场景，我们选取了 48 个因果特征，包括 24 个连续特征和 24 个分类特征。每种特征类型在六种非交互形式中各贡献 3 个特征，并在交互项中贡献 6 个特征。我们考虑了连续型和分类响应变量。对于连续型结果，我们模拟为 $Y_i \sim N(\mu_i,\ 1)$。对于二元结果，为确保病例-对照设计均衡，我们首先对均值进行中心化（记为 $\tilde{\mu}_i$），然后通过二项抽样生成结果：$Y_i \sim \text{Binomial}(1,\ p_i)$，其中 $p_i = (1 + \exp(-\tilde{\mu}_i))^{-1}$。我们的模拟考虑了两种样本量，分别对应小规模探索性研究（$n = 500$）和中等规模研究（$n = 2000$）。不同函数关系的效应量汇总于补充表 S1。

4.3.2 模拟场景

为系统评估 MAFS 的特征选择性能，我们开展了两项互补的模拟实验，分别侧重于不同的评估方面。在两项模拟中，均采用覆盖率作为主要性能指标，其定义为：在所选子集中正确识别出的、属于各函数形式的真实因果特征所占的比例。为确保结果稳健，每种模拟设置均重复 20 次。数据被划分为训练集（80%）和验证集（20%），其中验证集用于超参数调优。报告了平均覆盖率，以及包括特征选择和超参数优化在内的计算资源消耗。

模拟1旨在评估该方法检测关系特异性特征的能力。我们采用多个预定义阈值（即特征总数的 0.5% 、 1% 、 1.5% 和 2% ），以确定每种函数形式的信息性特征是在选择过程中较早获得优先考虑，还是只有在考虑较大的子集时才能被选出。这项分析提供了

有关 MAFS 在各种函数依赖下的敏感性和稳健性的见解，并为确定下游分析所需的排名靠前特征数量提供了实用指导。

模拟2旨在评估随着特征维度不断增加，MAFS 的可扩展性。特征总数设为 25,000、50,000 和 100,000，分别对应典型的转录组学/蛋白质组数据集、多组学整合场景，以及超高维基因组学或表观基因组学研究。由于 OmicsSIMLA 的默认输出与目标维度不匹配，我们对连续数据进行了维度扩增，对分类数据进行了维度截断。对于连续特征，当所需维度大于 OmicsSIMLA 的默认维度时，通过调整均值参数并拼接多个特征集来生成额外变量。对于分类特征，则从经过质量控制的 SNP 集中依次选取 SNP，以匹配所需维度。对于混合类型特征，按 50%/50% 的比例拼接基因表达和 SNP 基因型，以达到目标维度。本组模拟采用了两种特征选择策略：固定比例策略（即特征总数的 2%），适用于大规模搜索空间；以及固定数量策略（即排名前 100 和前 500 的特征），反映了资源受限的验证研究所面临的限制。总之，这些模拟评估了随着特征空间维度增加，MAFS 的稳健性和可扩展性。

通过总运行时间（分钟）评估了不同样本量（$n = 500$ 和 $n = 2,000$）、特征维度（25K、50K 和 100K）以及结果类型（连续型和二元型）下的计算效率。值得注意的是，GRACES 通过迭代方式选择特征，而非生成全局重要性得分，因此在计算上无法评估所有特征。因此，对于 GRACES，报告了其选择最多占输入特征总数 2% 的特征时的运行时间。MAFS 的计算成本包括全部三个模块：（1）用于权重初始化的三种基于筛选器的方法；（2）多头注意力深度神经网络；以及（3）采用树算法的重排序模块。

4.3.3 模拟中基于深度学习的特征选择方法的参数

图 7 展示了所有方法的网络架构。为确保比较公平，所有嵌入式特征选择方法（CancelOut、EAR-FS、GRACES 和 MAFS）均采用一致的网络架构实现。每个模型均使用具有两个隐藏层的前馈神经网络。每层的宽度由输入维度除以缩放因子确定。每一层均由线性变换、批归一化、ReLU激活和丢弃法组成。对于作为事后归因方法的 DeepLIFT，我们首先使用相同的架构训练前馈网络，然后应用 DeepLIFT 计算特征重要性得分。

所有模型均使用 Adam优化器进行训练，批量大小为 32，最多训练 100 轮。当验证性能连续 10 轮未提升时，便采用早停。我们使用 Optuna 对关键超参数进行优化，包括方法特定的调优参数、学习率、丢弃率、批量大小和缩放因子；对于每种方法和每种数据配置，均进行了 100 次搜索迭代[39]。详细的超参数搜索空间和最优值见补充表 S2。

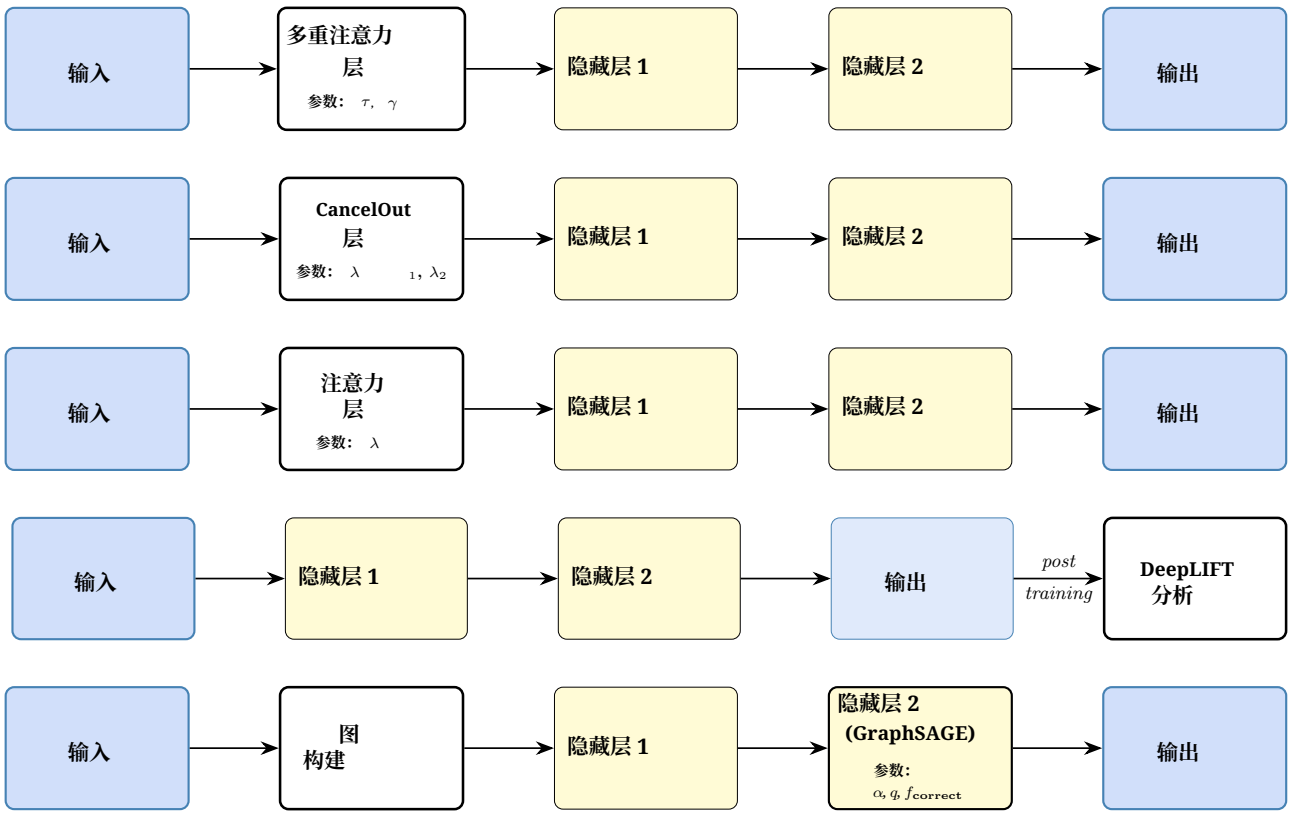


图 7 | 基线方法的网络架构。从上到下依次为：MAFS、CancelOut、EAR- FS、DeepLIFT 和 GRACES。所有方法都采用相当的容量，并突出显示各方法特有的组件。

4.4 真实数据应用

为评估 MAFS 的实际效用，我们分析了两类具有代表性的真实世界数据集。第一类包括六个公开可用的基因表达癌症数据集，具有二元结果（例如胶质瘤与对照组），这些数据集是基于分类的特征选择领域公认的基准[2]。第二类采用阿尔茨海默病神经影像学倡议（ADNI）多组学数据集，其中包含连续响应变量，可用于评估回归场景下的特征选择性能 [40]。

4.4.1 癌症基因表达数据集

这六个癌症基因表达数据集被广泛用作特征选择的基准数据集。它们涵盖血液系统恶性肿瘤和实体瘤，样本量（62–187）和维度（2,000–22,283）范围广泛。特征类型包括连续测量值（以实数表示的标准化表达值）和离散变量（采用序数编码的类别）。所有数据集均代表二元分类任务。详细特征见补充表 S4。每个数据集均以预处理后的形式从相应的原始发表文献或 GEO数据库中获取，且不含缺失值。我们直接使用所提供的数据，未进行额外预处理。

4.4.2 阿尔茨海默病神经影像学计划数据集

ADNI是一项多中心纵向研究，旨在识别和验证用于阿尔茨海默病早期检测与监测的临床、影像学、遗传学和生化生物标志物。在本研究中，我们重点选择用于预测影像衍生体积表型的基因表达特征，包括伏隔核、杏仁核、脑干、尾状核、第四脑室、海马、苍白球、壳核和丘脑。

使用了来自 ADNI 网站的影像衍生体积表型和基因表达特征[40]。为确保观测值彼此独立，当同一参与者存在多条记录时，我们将分析范围限定为基线访视。我们仅保留 RNA完整性数值（RIN）≥ 6 以确保 RNA质量满足可靠转录组测量的要求[41]。任一基因表达或表型数据存在缺失值的参与者均被排除在分析之外。最终分析数据集由 449 名参与者和 49,386 个基因表达特征组成。

4.4.3 基于深度学习的特征选择方法的参数及真实数据应用的评估指标

为了评估 MAFS 在真实世界数据集上的特征选择能力，我们采用了与模拟数据中相同的一组比较方法。所有基于深度学习的特征选择方法均采用与模拟研究中相同的网络架构（图 7）、训练设置和超参数搜索范围，并通过 Optuna（100 次迭代）对特定于模型的超参数进行优化。搜索空间的详细说明见补充表 S2。

由于真实世界数据集中的真实相关特征未知，因此使用预测准确率作为评估特征选择性能的替代指标。在癌症数据集上，遵循标准基准测试惯例[2], ，保留排名前 20 的特征用于后续分类分析。使用了三种分类器：支持向量机（SVM）、K近邻（KNN）和多层感知机（MLP）；该多层感知机包含两个隐藏层（分别有 64 和 32 个单元），并采用 ReLU激活。使用受试者工作特征曲线下面积（AUROC）评估分类性能。考虑到癌症基因表达数据集的样本量有限，我们采用 5 折交叉验证框架，以最大限度地利用训练数据。每个数据集划分为两部分：20% 留作测试集，其余 80% 用于特征选择和模型训练。

对于每个影像衍生结局，保留排名前 50 的特征用于回归分析。该阈值兼顾模型复杂度与预测性能，同时避免在中等样本量的设置下过拟合（$n = 449$）。使用了三种回归器：支持向量回归（SVR）、K近邻（KNN）和多层感知机（MLP）；后者采用相同的双隐藏层架构（分别有 64 和 32 个神经元，并采用 ReLU激活）。通过预测值与观测值之间的皮尔逊相关系数评估预测性能。鉴于样本量充足，采用固定的 60:20:20 训练集–验证集–测试集划分，并报告测试集上的性能。

所有下游预测模型的超参数均通过网格搜索进行优化（补充表 S3）。为确保研究结果的稳健性，所有实验均使用相同的数据划分和随机种子重复进行 20 次。

数据可用性

本研究使用的六个癌症基因表达数据集（Colon、Leukemia、ALLAML、GLI 85、Prostate GE 和 SMK CAN 187）均为公开可用的基准数据集，可通过 https://jundongl.github.io/scikit-feature/datasets.html获取。阿尔茨海默病神经影像学倡议提供的影像衍生体积表型和基因表达数据

（ADNI）取自ADNI数据库（https://adni.loni.usc.edu/）。访问ADNI数据需要注册并获得ADNI数据档案库的批准。本研究生成的模拟数据以及图中所用的源数据，可根据合理请求向通讯作者索取。

代码可用性

用于实现 MAFS 并复现本研究结果的所有代码均可在 https://github.com/xsun768/mafs_package 公开获取。该实现采用 Python 编写，包含模型训练和性能评估脚本。仓库中提供了依赖项和安装说明。

参考文献

1. Ashley, E. A. 迈向精准医学。《自然评论：遗传学》**17**, 507–522 (2016)。2. Chen, C., Weiss, S. T. & Liu, Y.-Y. 基于图卷积网络的高维小样本数据特征选择。《生物信息学》**39**, btad135 (2023)。3. Uffelmann, E.等全基因组关联研究。*Nat. Rev. Methods Primers* **1**, 59 (2021)。4. Motsinger-Reif, A. A. 等精准环境健康框架下的基因-环境相互作用。《细胞基因组学》**4** (2024)。5. Yang, P., Huang, H. & Liu, C. 单细胞时代的特征选择再探。《基因组生物学》**22**, 321 (2021)。6. Wang, H. 等人工智能时代的科学发现。《自然》**620**, 47–60 (2023)。7. Liu, H., Zhou, M. & Liu, Q. 一种用于不平衡数据分类的嵌入式特征选择方法。《*IEEE*/中国自动化学会自动化学报》**6**, 703–715 (2019)。8. Saeys, Y., Inza, I. & Larra˜naga, P. 生物信息学中特征选择技术综述。《生物信息学》**23**, 2507–2517 (2007)。9. Guyon, I. & Elisseeff, A. 变量与特征选择导论。《机器学习研究杂志》**3**, 1157–1182 (2003)。10. Bommert, A., Welchowski, T., Schmid, M. & Rahnenf¨uhrer, J. 高维基因表达生存数据特征选择过滤方法基准评估。《生物信息学简报》**23**, bbab354 (2022)。11. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. 使用支持向量机进行癌症分类的基因选择。《机器学习》**46**, 389–422 (2002)。12. Tibshirani, R. 通过最小绝对收缩和选择算子进行回归收缩与选择。《英国皇家统计学会杂志*B*辑》**58**, 267–288 (1996)。13. Clark, L. A. & Pregibon. 基于树的模型。载于《*S*语言中的统计模型》377–419 (Routledge, 2017)。14. Fan, J. & Lv, J. 超高维特征空间的确定性独立筛选。《英国皇家统计学会杂志*B*辑》**70**, 849–911 (2008)。

15. Li, R., Zhong, W. & Zhu, L. 基于距离相关学习的特征筛选。《美国统计学会杂志》107, 1129–1139 (2012)。16. Barut, E., Fan, J. & Verhasselt, A. 条件确定性独立筛选。《美国统计学会杂志》111, 1266–1277 (2016)。17. Hastie, T., Tibshirani, R. & Friedman, J. 《统计学习导论》(2009)。18. Pan, W., Wang, X., Xiao, W. & Zhu, H. 通用确定性独立筛选程序。《美国统计学会杂志》(2019)。19. Huang, H., Liu, C., Wagle, M. M. & Yang, P. 深度学习特征选择在单细胞RNA测序数据分析中的评估。《基因组生物学》24, 259 (2023)。20. Shrikumar, A., Greenside, P. & Kundaje, A. 通过传播激活差异学习重要特征。载于国际机器学习会议 3145–3153 (PMLR, 2017)。21. Lundberg, S. M. & Lee, S.-I. 一种解释模型预测的统一方法。《神经信息处理系统进展》30 (2017)。22. Ribeiro, M. T., Singh, S. & Guestrin, C. “我为什么应该相信你？”解释任意分类器的预测。载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集 1135–1144 (2016)。23. Kokhlikyan, N. 等。Captum：一个统一、通用的PyTorch模型可解释性库。预印本见 https://arxiv.org/abs/2009.07896 (2020)。24. Borisov, V., Haug, J. & Kasneci, G. Cancelout：用于深度神经网络特征选择的层。载于国际人工神经网络会议 72–83 (施普林格, 2019)。25. Bietti, A. & Mairal, J. 群不变性、对形变的稳定性与深度卷积表示的复杂度。《机器学习研究杂志》20, 1–49 (2019)。26. Achille, A. & Soatto, S. 深度表示中不变性与解耦的涌现。《机器学习研究杂志》19, 1–34 (2018)。27. Vaswani, A. 等。注意力就是你所需要的一切。《神经信息处理系统进展》30 (2017)。28. Bahdanau, D., Cho, K. & Bengio, Y. 通过联合学习对齐与翻译实现神经机器翻译。预印本见 https://arxiv.org/abs/1409.0473 (2014)。29. Guo, M.-H., Liu, Z.-N., Mu, T.-J. & Hu, S.-M. 超越自注意力：用于视觉任务的双线性层外部注意力。《IEEE模式分析与机器智能汇刊》45, 5436–5447 (2022)。30. Gui, N., Ge, D. & Hu, Z. AFS：一种用于监督式特征选择的基于注意力的机制。载于AAAI人工智能会议论文集 33, 3705–3713 (2019)。31. Xue, Y., Zhang, C., Neri, F., Gabbouj, M. & Zhang, Y. 一种用于大规模特征选择的基于外部注意力的特征排序器。《知识型系统》281, 111084 (2023)。32. 新西兰研究与教育先进网络（REANNZ）。REANNZ。 https://www.reannz.co.nz/ (2025)。

33. Glorot, X. & Bengio, Y. 理解深度前馈神经网络训练的难点。在 第*13*届人工智能与统计国际会议论文集 249–256（JMLR，2010年）。34. Reay, W. R. & Cairns, M. J. 推进全基因组关联研究在药物再利用中的应用。《自然评论：遗传学》**22**，658–671（2021年）。35. Qiu, X., Brooks, A. I., Klebanov, L. & Yakovlev, A. 归一化对微阵列数据相关结构的影响。《*BMC*生物信息学》**6**，120（2005年）。36. Chung, R.-H. & Kang, C.-Y. 用于复杂疾病研究的多组学数据模拟器及其在评估疾病分类多组学数据分析方法中的应用。《*GigaScience*》**8**，giz045（2019年）。37. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2：多种疾病相关单核苷酸多态性的模拟。《生物信息学》**27**，2304–2305（2011年）。38. GTEx联盟。GTEx联盟人体组织遗传调控效应图谱。《科学》**369**，1318–1330（2020年）。39. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna：新一代超参数优化框架。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集 2623–2631（2019年）。40. 阿尔茨海默病神经影像学倡议。ADNI。https://adni.loni.usc.edu/ （2024年）。41. Fleige, S. & Pfaffl, M. W. RNA完整性及其对实时定量逆转录PCR性能的影响。《分子医学评论》**27**，126–139（2006年）。

致谢

作者谨此感谢使用由王室公司新西兰研究与教育先进网络（REANNZ）有限公司托管、由商业、创新与就业部资助的电子研究基础设施平台。网址：https://www.reannz.co.nz

作者贡献

X.S. 和 Q.M. 对本研究贡献相同。X.S. 提出研究构想、开发算法、开展模拟并分析真实数据。Q.M. 协助生成模拟数据并分析真实数据。X.S. 和 Y.W. 撰写了手稿。Y.W. 负责指导本研究。

通讯作者

有关通信及材料索取事宜，请联系 Yalu Wen。

利益冲突

作者声明不存在任何利益冲突。