MethConvTransformer：用于跨组织阿尔茨海默病检测的深度学习框架

屈刚、Gu anghao Li、赵忠明，代表阿尔茨海默病神经影像学倡议*

arXiv:2601.00143v1 [q-bio.GN] 1 Jan 2026

摘要—**阿尔茨海默病（AD）是一种多因素神经退行性疾病，其特征是认知能力逐渐下降以及脑内广泛的表观遗传失调。DNA甲基化作为一种稳定而动态的表观遗传修饰，有望成为早期检测阿尔茨海默病的无创生物标志物。然而，不同组织和研究中的甲基化特征差异显著，限制了结果的可重复性及转化应用价值。为应对这些挑战，我们开发了MethConvTransformer，这是一种基于Transformer的深度学习框架，可整合脑和外周组织的DNA甲基化谱，以支持生物标志物发现。该模型将逐CpG位点线性投影与卷积层和自注意力层相结合，以捕捉CpG位点间的局部及长程依赖关系，同时纳入受试者层面协变量和组织嵌入，从而区分共有的甲基化效应与区域特异性效应。在六个 GEO 数据集和一个独立的 ADNI 验证队列中进行的实验表明，该模型始终优于传统机器学习基线方法，在判别能力和泛化能力方面表现更佳。此外，采用线性投影、SHAP 和 Grad-CAM++ 进行的可解释性分析揭示了具有生物学意义、且与阿尔茨海默病相关通路相一致的甲基化模式，包括免疫受体信号传导、糖基化、脂质代谢以及内膜系统（内质网/高尔基体）组织。总之，这些结果表明，MethConvTransformer 能够提供稳健的跨组织阿尔茨海默病表观遗传生物标志物，并实现多分辨率可解释性，从而推动可重复的甲基化诊断方法发展，并为疾病机制提供可检验的假设。**

索引词—**阿尔茨海默病（AD）、生物标志物、跨组织分析、深度学习、DNA甲基化、可解释人工智能（XAI）、Transformer模型。**

一、引言

阿尔茨海默病（AD）是最常见的痴呆类型，也是一项日益严峻的公共卫生挑战。根据阿尔茨海默病协会2025年报告 [1]，目前美国估计有720万名65岁及以上的成年人患有阿尔茨海默病；如果没有有效的预防或治愈措施，到2060年，这一数字预计将接近翻倍，达到1380万。阿尔茨海默病是导致死亡的主要原因之一，2022年记录的死亡人数超过12万，使其成为美国第七大死因 [2]。阿尔茨海默病相关的认知能力持续下降，不仅给患者带来沉重负担，也给家庭照护者和医疗保健系统造成巨大压力。鉴于这些挑战，亟需开发灵敏且易于获取的生物标志物，以便在阿尔茨海默病前驱期检测出该病。

表观遗传标记（如DNA甲基化）日益被认为是很有前景的候选指标。DNA甲基化通常发生在CpG二核苷酸中的胞嘧啶残基上，即甲基共价添加到胞嘧啶的5’位，形成5-甲基胞嘧啶（5mC）。这种修饰可通过招募抑制蛋白或阻断转录因子结合来调节基因表达 [3]。在发育过程中，细胞会获得谱系和组织特异性甲基化模式 [4] ，这些模式足够稳定，可以作为组织特性和疾病过程的分子特征 [5]。这些特性使DNA甲基化成为包括阿尔茨海默病在内的神经退行性疾病颇具吸引力的候选生物标志物。

既往研究表明，阿尔茨海默病与脑中的甲基化变化相关 [6], [7], ，但这些变化的幅度和一致性在不同脑区及研究之间存在显著差异 [8], [9]。在阿尔茨海默病背景下，大规模全表观基因组关联研究（EWAS）在皮层组织中发现了稳健的甲基化差异，其中许多差异定位于参与免疫调节、突触功能和胶质细胞反应的基因 [10]–[12], ，从而与阿尔茨海默病中已知的病理过程相一致。相比之下，血液来源的甲基化数据研究结果一致性较低：尽管部分研究报告了血液甲基化与脑脊液生物标志物之间的相关性 [13],，但在独立队列中的重复验证仍然面临挑战，这可能是由于组织特异性、队列构成和分析方法存在差异 [14], [15]。这种组织特异性差异凸显了生物标志物开发中的转化困境：脑来源的甲基化谱与疾病病理密切相关，但只能在死后或通过侵入性操作获得；而血液更易用于临床应用，却可能产生较弱或重复性较差的特征。

鉴于这些组织特异性差异，人们开发了计算模型，以改进基于甲基化的生物标志物的识别。传统的机器学习分类器，如支持向量机和逻辑回归，通常依赖于选定的CpG子集或

G.Qu、G.Li和Z.Zhao来自德克萨斯大学休斯敦健康科学中心麦克威廉姆斯生物医学信息学学院和公共卫生学院，美国得克萨斯州休斯敦，邮编77030。通讯作者：Zhongming Zhao。Zhongming.Zhao@uth.tmc.edu *本文准备过程中使用的数据来自阿尔茨海默病神经影像学倡议（ADNI）数据库（adni.loni.usc.edu）。因此，ADNI的研究人员参与了ADNI的设计与实施和/或提供了数据，但未参与本报告的分析或撰写。ADNI研究人员的完整名单请见：http://adni.loni.usc.edu/wp-content/uploads/how to apply/ ADNI_Acknowledgement List.pdf

降维特征，但受限于线性决策规则，难以捕捉数十万个基因座之间更高阶的非线性依赖关系 [16]。在脑甲基化研究领域，EWASplus 利用源自皮层 EWAS 数据的 CpG 特征，成功区分了阿尔茨海默病病例与对照组 [17]。近期，深度学习已应用于纵向甲基化数据，以预测阿尔茨海默病进展，并在建模复杂模式方面展现出更优的性能 [18]。然而，这些模型通常基于单组织数据集（例如脑或血液）开发，未能利用组织之间可获得的互补信息，从而限制了其转化应用潜力。为解决这一局限，Silva 等人 [19]探索了血液与脑甲基化数据的跨组织整合，并识别出阿尔茨海默病相关 CpG 位点；然而，一致信号的数量较少，预测性能仍然有限。由于当前的跨组织策略主要基于荟萃分析或回归，因此无法充分建模异质性和组织特异性调控机制。这凸显了开发新型整合框架的必要性，以捕捉复杂的跨组织关系，从而改进阿尔茨海默病中的生物标志物发现和机制性认识。

基于上述局限性，我们提出了一种用于阿尔茨海默病检测的多组织表观遗传整合框架。该模型采用基于 Transformer 的 [20] 架构，能够捕捉 CpG 位点之间的长程非线性依赖关系。模型纳入年龄和性别等协变量，使用基于索引的位置编码来表示 CpG 位点，并学习组织嵌入，以区分共享的甲基化模式与组织特异性甲基化模式。通过在涵盖脑和外周组织的异质性队列上训练，我们的框架旨在提升基于甲基化的诊断的敏感性和泛化能力，同时揭示阿尔茨海默病相关跨组织表观遗传过程的机制。重要的是，我们的框架整合了一个可解释性模块，用于量化各个 CpG 位点对模型预测的贡献，并支持在不同组织间统一评估特征重要性。这一能力超越了传统的差异甲基化分析；后者通常依赖单个位点水平的统计显著性阈值。差异分析虽然能够识别稳健的位点特异性变化，但往往会忽略可能具有生物学意义的微弱效应或组合效应。我们的方法通过捕捉更高阶的跨组织模式，并以连续且可解释的方式对 CpG 位点的重要性进行排序，对此类分析加以补充和扩展，从而将预测建模与机制性洞见联系起来。

II. 材料与方法

A. 数据获取与队列

本研究将使用通过 Illumina 450K 或 EPIC 芯片生成的公开 DNA 甲基化数据集。我们将收集来自多种组织（例如脑区和外周血）的数据，以便进行跨组织建模。具体而言，我们汇编了多个来自基因表达综合数据库（GEO）的公开 DNA 甲基化数据集，涵盖脑组织和血液组织。表 I 总结了本研究纳入的数据集、组织区域、样本量、平台及可用数据格式。在各队列中，我们从 GEO 获取了通过 Illumina 450K 或 EPIC 芯片生成的公开 DNA 甲基化数据集。纳入标准要求提供原始 IDAT 配对文件（或等效的探针水平强度）以及足以确定组织/区域和疾病状态的核心样本元数据。我们重点关注脑区和外周血；缺乏明确组织注释或可协调标签的数据集予以排除。若同一受试者—区域存在重复检测或技术重复，则在采集时保留所有记录（本轮未进行去重）。

所有数据集均从 GEO 获取，并经过筛选，仅纳入具有明确诊断标签（阿尔茨海默病、轻度认知障碍或认知正常对照组）的受试者。对于多脑区研究，仅保留元数据完整的样本。我们排除了：

- 缺少年龄、性别或诊断信息的样本，
- 原始 GEO 记录中标记为技术重复或低质量芯片的样本，
- IDAT/β-矩阵文件缺失或损坏的样本。

我们整合了上述六项GEO研究的数据集，共纳入1,656个经过质量控制的样本，涵盖十种组织类型（额叶皮层、内嗅皮层、颞上回、小脑、全血、前额叶皮层、背外侧前额叶皮层、海马、颞中回和颞叶皮层）。我们将不同队列中的诊断组统一为阿尔茨海默病（AD）和认知正常对照组（CN），最终得到908例AD病例和748名对照。除此之外，我们还使用来自ADNI的独立DNA甲基化数据集评估模型的泛化能力。该数据集提供基于血液的甲基化谱，包含统一的诊断类别和详细的临床元数据。（本文准备过程中使用的数据来自ADNI数据库（adni.loni.usc.edu）。ADNI于2003年启动，是一项公私合作项目，由首席研究员Michael W. Weiner医学博士领导。ADNI的主要目标是检验能否结合连续的磁共振成像（MRI）、正电子发射断层成像（PET）、其他生物标志物以及临床和神经心理学评估，来衡量轻度认知障碍和早期阿尔茨海默病的进展。）

B. 质量控制与标准化

所有原始甲基化数据（IDAT 文件）均使用 ChAMP（芯片甲基化分析流程）R软件包[25], 进行处理，该软件包为 Illumina 450K 和 EPIC 芯片提供标准化工作流程。该流程将在不同批次和组织中一致应用，以确保可比性。首先，使用 champ.load 函数导入 IDAT 文件，随后进行探针过滤，移除：(i) 检测 p-值 > 0.01的探针，(ii) 微珠计数 < 3的探针，(iii) 非CpG探针，(iv) 与已知单核苷酸多态性重叠的探针，以及 (v) 在 [26]中注释的交叉反应探针。质量控制后，使用BMIQ方法对剩余探针进行标准化，以校正 I型和II型探针偏倚。

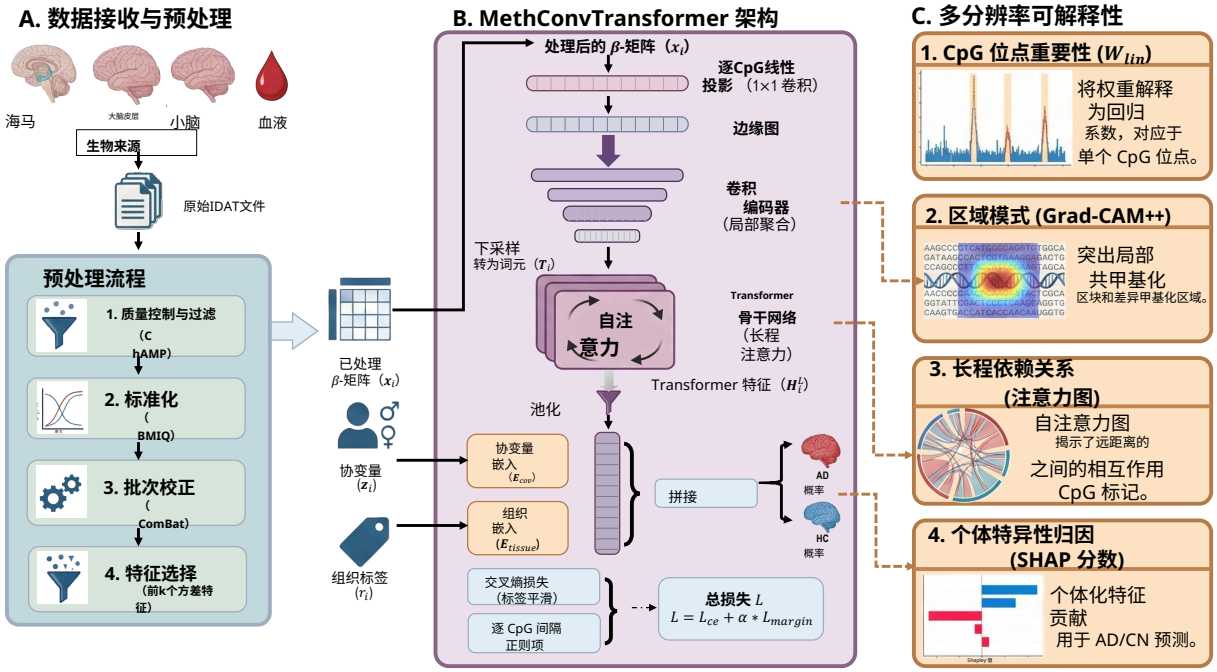


图 1：MethConvTransformer DNA 甲基化分析框架概览：(A) 预处理流程：将来自多个脑区和血液的原始 IDAT 文件转换为经标准化和批次校正的 β 矩阵，并选取 CpG 特征；(B) MethConvTransformer 架构结合逐 CpG 线性投影、局部卷积编码和基于 Transformer 的长程注意力，并与协变量和组织嵌入联合整合，用于 AD 预测；(C) 多分辨率可解释性模块，涵盖 CpG 位点重要性、区域共甲基化模式（Grad-CAM++）、长程依赖关系（注意力图）以及个体特异性特征归因（SHAP）。

表 I：本研究纳入的 DNA 甲基化数据集概述。

数据集	组织 / 区域	样本	平台	数据格式	样本组
GSE125895 [21]	内嗅皮层 (ERC)；髓-海马；背外侧前额叶皮层；小脑-lum	269	Illumina 450K	IDAT	82 例阿尔茨海默病，187 例认知正常
GSE134379 [22]	颞中回；小脑-小脑	404	Illumina 450K	IDAT	225 例阿尔茨海默病，179 例认知正常
GSE66351 [12]	额叶皮层神经元和胶质细胞	192	Illumina 450K	IDAT	106例阿尔茨海默病，84例认知正常
GSE59685 [6]	内嗅皮层；颞上回；前额叶皮层；小脑；全血	531	Illumina 450K	β-矩阵	207 例阿尔茨海默病，103 例认知正常
GSE80970 [23]	前额叶皮层；上部颞回	286	Illumina 450K	β-矩阵	148 例阿尔茨海默病，138 例认知正常
GSE144858 [24]	全血	300	Illumina 450K	β-矩阵	93 名阿尔茨海默病患者、111 名轻度认知障碍患者、96 名认知正常者

表 II：合并 GEO 数据集和 ADNI 验证数据集的参与者人口统计学特征及临床特征。

数据集	样本	#受试者	年龄（均值 ±标准差）	性别（男/女）
合并	AD	908	83.81 ± 8.57	333 / 575
	对照组	748	74.64 ± 13.01	401 / 347
ADNI	AD	169	73.84 ±7.25	105 / 64
	对照组	212	74.00 ± 5.97	104 / 108

为整合不同批次和组织中的信息，所有数据集的 β-值矩阵均对齐到一组共同的CpG探针，这组探针由各批次的交集定义。随后，通过拼接对齐后的矩阵创建合并的β-矩阵，并组装相应的表型表。使用sva软件包中的ComBat评估并校正批次效应，且将研究来源视为批次变量包，并将研究来源视为批次变量。

在整合之前，会对年龄、性别和组织类型等协变量进行统一处理：年龄经过 z-score 标准化，性别编码为二元指示变量，组织或脑区标签则赋予指定的数值代码。将根据芯片间相关性和主成分分析识别离群样本，并将其排除在下游建模之外。

为降低维度并保留具有信息量的基因座，我们计算每种组织内每个 CpG 位点的方差，并选择方差最高的前 $k = 5000$ 个位点。建模特征集定义为在不同组织中均表现出高方差的 CpG位点的并集。该策略能够富集具有显著个体间变异的基因座，同时保留组织特异性信号。因此，所得处理后的数据集提供了一种经过协调、质量控制且降维后的表示

适用于跨组织机器学习分析的 DNA甲基化表示。

C. 模型设计

为应对DNA甲基化数据建模所面临的挑战——这类数据涵盖异质性脑组织和血液组织中的数十万个CpG位点——我们开发了卷积–Transformer框架 **MethConvTransformer**。不同于对每个位点独立评估且依赖严格 p值阈值的传统差异甲基化分析，我们的方法旨在利用CpG位点之间的局部和远程相关性。卷积编码器首先聚合短程依赖关系并降低维度，而Transformer主干网络则捕获远距离位点之间更高阶的非线性关系。值得注意的是，我们引入了逐CpG线性投影层，可直接量化位点层面的贡献，为经典统计检验与现代深度学习搭建了自然的桥梁。这一设计不仅提升了阿尔茨海默病分类的预测性能，还能在CpG和组织层面提供可解释的特征归因，从而支持对表观遗传改变进行跨组织比较。

形式化地，令 $x_i \in \mathbb{R}^P$ 表示受试者 i 在 P 个CpG位点上的甲基化谱， $z_i \in \mathbb{R}^K$ 表示协变量向量（例如年龄和性别）r_i 表示分类组织标签。为保留CpG层面的可解释性，我们首先应用CpG 逐位线性投影，为每个位点生成间隔分数:

$$\boldsymbol{h}_i = \boldsymbol{W}_{\text{lin}} x_i + \boldsymbol{b}_{\text{lin}}, \quad \boldsymbol{h}_i \in \mathbb{R}^P, \tag{1}$$

其中， W_{lin} 为每个 CpG 位点分配一个权重， b_{lin} 为偏置项。该投影既可作为压缩后的输入表示，也能直接用于归因单个 CpG 位点的预测重要性。

随后，卷积编码器 $f_{\text{conv}}(\cdot)$ 对投影信号 h_i 进行处理，聚合局部依赖关系，并将高维向量下采样为词元序列

$$\boldsymbol{T}_i = f_{\text{conv}}(\boldsymbol{h}_i) \in \mathbb{R}^{L \times d}, \quad L \ll P, \tag{2}$$

从而在保留短程 CpG 位点相关性的同时降低计算成本。

词元序列随后输入多层 Transformer 骨干网络 $f_{\text{trans}}(\cdot)$:

$$\boldsymbol{H}_i^{(l+1)} = f_{\text{trans}}^{(l)}(\boldsymbol{H}_i^{(l)}), \quad \boldsymbol{H}_i^{(0)} = T_i, \tag{3}$$

其中，前面的模块采用高效的注意力机制，最后一个模块则保留完整的注意力权重，以提高可解释性。为将甲基化衍生特征与受试者元数据相结合，我们将最终表示构造为

$$\boldsymbol{u}_i = \text{Pool}(\boldsymbol{H}_i^{(L)}) \parallel E_{\text{cov}}(\boldsymbol{z}_i) \parallel E_{\text{tissue}}(r_i), \tag{4}$$

其中， $H_i^{(L)} \in \mathbb{R}^{L \times D}$ 是来自最后一个 Transformer 块的标记矩阵，Pool(·) 表示沿标记维度进行平均池化，从而生成一个 D 维的 CpG 依赖关系摘要。协变量嵌入 $E_{\text{cov}}(\boldsymbol{z}_i) \in \mathbb{R}^{d_{\text{cov}}}$将观测到的受试者层面协变量 $\boldsymbol{z}_i$ （例如年龄和性别）映射到紧凑的潜在空间，而组织嵌入$E_{\text{tissue}}(r_i) \in \mathbb{R}^{d_r}$ 则将区域身份编码为一种学习得到的固定效应。拼接后的向量 $\boldsymbol{u}_i \in \mathbb{R}^{D+d_{\text{cov}}+d_r}$ 作为用于分类的联合表示，确保预测以统一的方式纳入 CpG 特征、协变量和组织背景。

随后得到诊断预测:

$$\hat{\boldsymbol{y}}_i = \text{softmax}(\boldsymbol{W}_c \boldsymbol{u}_i + \boldsymbol{b}_c), \tag{5}$$

其中， u_i 表示第 i 个输入的表征， W_c 和b_c 分别为可学习的权重矩阵和偏置向量，而 softmax(·) 将逻辑值映射为类别上的概率分布。

该架构整合了 CpG 逐位线性投影、卷积下采样和 Transformer 注意力，以共同应对跨组织甲基化分析中维度、异质性和可解释性方面的挑战。线性投影为单个 CpG 位点分配明确的权重，提供位点级重要性评分，类似于差异甲基化分析，但无需依赖任意设定的显著性阈值。随后，卷积层在保留局部相关性的同时压缩极高维输入，从而实现高效的表征学习。Transformer 主干网络基于压缩后的序列，捕获可能反映阿尔茨海默病中分布式表观遗传失调的长程跨 CpG 依赖关系。通过将学习得到的协变量嵌入和组织嵌入与池化后的 Transformer 输出拼接，该模型明确考虑人口统计学差异和组织背景，从而支持统一的跨组织生物标志物发现。

D. 可解释性与生物标志物识别

为确保模型预测能够追溯到具有生物学意义的实体，我们采用了多种与所实现架构相匹配的归因机制。这些机制涵盖单个 CpG 位点、局部区域、受试者特异性贡献以及序列级依赖关系。

在 CpG层面，我们采用一个深度卷积层，该层专为可解释性而设计，而非用于评估模型性能。该线性投影为每个 CpG 输入独立分配权重，将原始甲基化谱 $x_i \in \mathbb{R}^{1 \times P}$ 转换为边缘图，如公式 1 所示。所得系数 W_{lin} 为各个 CpG位点提供连续且可直接解释的重要性评分。由于该操作沿 CpG 轴严格呈线性，这些权重可类似于回归系数进行解读，从而阐明网络在预测过程中优先考虑了哪些 CpG 位点。

虽然单个位点权重可以识别单个 CpG 位点，但表观遗传调控往往源于 CpG 位点的空间相关区块。为揭示此类局部模式，我们在选定的卷积特征图上使用 Grad– CAM++ [27], [28] 。对于类别 c ，其逻辑值为 y^c ，来自通道 k 的激活值为 $A^k = (A_1^k, \ldots, A_{L'}^k)$，令 $G_t^k = \partial y^c / \partial A_t^k$表示类别得分相对于激活值 A_t^k的梯度。Grad–CAM++会为各位置分配权重

$$\alpha_{k,t}^c = \frac{(\boldsymbol{G}_t^k)^2}{2(\boldsymbol{G}_t^k)^2 + \sum_{u=1}^{L'} \boldsymbol{A}_u^k (\boldsymbol{G}_u^k)^3 + \varepsilon}, \tag{6}$$

其中

$$\boldsymbol{w}_k^c \;=\; \sum_{t=1}^{L'} \alpha_{k,t}^c \,\mathrm{ReLU}(\boldsymbol{G}_t^k), \tag{7}$$

并构建类别激活图

$$\boldsymbol{M}^c(t) \;=\; \mathrm{ReLU}\Big(\sum_{k=1}^{C} \boldsymbol{w}_k^c\, \boldsymbol{A}_t^k\Big), \qquad t = 1, \ldots, L'. \tag{8}$$

其中， L' 表示卷积/池化后的序列长度， C表示通道数， $\boldsymbol{A}_t^k$ 表示通道 k在位置 t处的激活。权重$\alpha_{k,t}^c$ 突出显示激活 $\boldsymbol{A}_t^k$及其敏感度 $\boldsymbol{G}_t^k$均为类别 c提供一致正向证据的位置，而分母通过汇总同一通道不同位置上的高阶项，使加权结果保持稳定。通道权重 $\boldsymbol{w}_k^c$随后汇总各位置的贡献，最终映射$\boldsymbol{M}^c$ 则强调对预测提供正向支持的区域。将结果上采样 $\boldsymbol{M}^c$ 回 CpG 分辨率后，即可得到覆盖 P个位点的显著性曲线，从而直接进行生物学解读。这种方法的一项重要优势是能够灵活适用于不同的表征尺度：将 Grad–CAM ++ 应用于较早的卷积层，可突出靠近输入的精细 CpG 级信号；而针对更深层，则可突出经层级池化后形成的、更抽象的区域级模式。这种多尺度可解释性尤其适用于DNA甲基化数据，因为单个位点的变异和区块级共甲基化都具有生物学意义。

为补充全局特征权重，并提供针对具体受试者的解释，我们采用 SHAP（Shapley 加性解释）。其核心思想是将每个 CpG 位点视为合作博弈中的一名“参与者”，模型对某位受试者的预测则对应博弈收益。随后，将 CpG 的贡献定义为其在所有可能特征联盟中的平均边际效应，从而确保归因既公平又具有可加性。

对于类别 c，受试者 i 的 CpG j 的 Shapley 值定义为

$$\phi_j^{(c)}(\boldsymbol{x}_i) = \sum_{S \subseteq F \setminus \{j\}} w(S)\, \Delta_S^j v_c(\boldsymbol{x}_i), \tag{9}$$

其中， F 表示 CpG 特征集。这里，$w(S) = \frac{|S|!\,(|F|-|S|-1)!}{|F|!}$ 是联盟 S 的 Shapley 权重，而 $\Delta_S^j v_c(\boldsymbol{x}_i) = v_c(S \cup \{j\};\, \boldsymbol{x}_i) - v_c(S;\, \boldsymbol{x}_i)$ 则量化了给定 S 时 CpG j 的边际贡献。

价值函数通过对不属于 S 的特征相对于背景集 $\mathcal{B}$ 进行边缘化来定义：

$$v_c(S; \boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{x}' \in \mathcal{B}}\left[f_c\big(\boldsymbol{x}_i|_S \,\oplus\, \boldsymbol{x}'|_{F\setminus S}\big)\right], \tag{10}$$

其中， $f_c(\cdot)$ 是模型的第 c 类 softmax 输出，而 $\oplus$表示将 $\boldsymbol{x}_i$ 中缺失坐标替换为背景实例 $\boldsymbol{x}'$ 中对应坐标的操作。在实际应用中， $\mathcal{B}$ 被设为单个参考样本，并由包装器固定区域相关特征和协变量相关特征。因此， $\phi_j^{(c)}(\boldsymbol{x}_i)$ 仅针对 CpG 特征提供特定于类别的概率归因。SHAP 使我们能够针对每位受试者确定哪些 CpG 位点促使模型倾向于某一给定类别，从而在全局重要性评分之外提供个体化的可解释性。

最后，在序列层面，最后一个 Transformer 块会生成多头自注意力矩阵，揭示 CpG 标记之间的依赖关系。对于具有查询矩阵 $\boldsymbol{Q}^{(h)}$、 $\boldsymbol{K}^{(h)} \in \mathbb{R}^{L\times d_h}$ 和键矩阵的第 h 个头，

$$\boldsymbol{A}^{(h)} \;=\; \mathrm{softmax}\!\left(\frac{\boldsymbol{Q}^{(h)}\boldsymbol{K}^{(h)\top}}{\sqrt{d_h}}\right) \in \mathbb{R}^{L\times L}, \tag{11}$$

以及对 H 个头的注意力图取平均后得到的图，

$$\bar{\boldsymbol{A}} \;=\; \tfrac{1}{H}\sum_{h=1}^{H} \boldsymbol{A}^{(h)}, \tag{12}$$

会保留下来用于解释。这些矩阵量化了一个样本中 CpG 标记彼此关注的程度，捕捉由 Transformer 主干网络引入的长程依赖关系。由于区域嵌入仅添加在分类器头中，因此注意力图表示的是纯粹的 CpG 标记关系，不受组织或协变量影响。

综合来看，这些可解释性机制从多个分辨率层面呈现了模型决策。逐CpG线性投影将模型参数直接关联到各个甲基化位点，从而提供全局效应大小的解释。Grad–CAM++ 则通过突出局部相关性区块予以补充，这些区块更能体现差异甲基化区域等生物学实际情况。最后，SHAP 提供个体特异性归因，解释为何某个特定样本会被分到某一类别。通过结合全局位点级权重、区域模式和个体化特征贡献，我们的框架提供了连贯且具有生物学依据的可解释性分析，有助于在高维甲基化数据中稳健地发现生物标志物。

E. 模型训练

使用分层抽样（默认80/10/10）将数据集划分为训练集、验证集和测试集，以保持各类别的比例。使用 Adam 优化模型，并分别对 CpG 投影层和网络其余部分应用不同的权重衰减。根据验证集 AUC 进行提前停止。

训练目标结合了标签平滑交叉熵损失和逐 CpG 间隔正则项，以增强预测稳健性和可解释性。形式上，对于一个小批次 $\mathcal{B}$ ，我们最小化

$$\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \Big[\mathcal{L}_{\mathrm{CE}}(\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_i) \;+\; \alpha\, \mathcal{R}_{\mathrm{margin}}(\boldsymbol{y}_i^{\pm 1}, \boldsymbol{m}_i)\Big], \tag{13}$$

其中， $\mathcal{L}_{\mathrm{CE}}$ 表示采用标签平滑的交叉熵， $|\mathcal{B}|$表示小批次 $\mathcal{B}$的大小， $\mathcal{R}_{\mathrm{margin}}$ 是 CpG 特异性边际惩罚项， $\mathcal{L}_{\mathrm{CE}}$ 是定义如下的交叉熵项

$$\mathcal{L}_{\mathrm{CE}}(\tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_i) = -\sum_{c=1}^{C} \tilde{\boldsymbol{y}}_{ic}\, \log \hat{\boldsymbol{y}}_{ic},$$

其中，$\hat{\boldsymbol{y}}_i$ =Softmax函数($\boldsymbol{z}_i$) 是由逻辑值 $\boldsymbol{z}_i$计算得到的预测类别概率。我们进一步在交叉熵损失中应用标签平滑，以避免过度自信并提升

泛化能力。受试者 i 和类别c的平滑目标定义为

$$\tilde{\boldsymbol{y}}_{ic} = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{C}, & \text{if } c=\boldsymbol{y}_i, \\ \dfrac{\varepsilon}{C}, & \text{if } c\neq \boldsymbol{y}_i, \end{cases} \tag{14}$$

其中， $\boldsymbol{y}_i$是样本 i的真实类别标签， $c\in\{1,\dots,C\}$表示一个候选类别， C 是类别总数， $\varepsilon\in[0,1)$ 是标签平滑因子。当 $\varepsilon=0$时，目标将退化为标准独热编码。

间隔正则项作用于逐 CpG位点的间隔分数 $\boldsymbol{m}_i=(m_{i,1},\dots,m_{i,P})$，促使模型在特征层面区分类别。其定义如下：

$$\mathcal{R}_{\text{margin}}(\boldsymbol{y}_i^{\pm1},\boldsymbol{m}_i)=\frac{1}{k}\sum_{t\in\text{Top-}k}\max\bigl(0,\,1-\boldsymbol{y}_i^{\pm1}m_{i,t}\bigr), \tag{15}$$

其中， $\boldsymbol{y}_i^{\pm1}=2\boldsymbol{y}_i-1\in\{-1,+1\}$ 是二元标签编码，求和范围限定为具有最大合页残差 $r_{i,t}=\max(0,\,1-\boldsymbol{y}_i^{\pm1}m_{i,t})$ 的 k 个 CpG位点。这种“困难位点”筛选确保损失聚焦于区分度最高或分类效果较差的 CpG位点，而不是将梯度分散到所有位点上。权衡参数 $\alpha>0$ 控制正则化强度。结合标签平滑的交叉熵能够稳定训练并缓解过拟合，尤其是在类别不平衡的情况下。逐 CpG 间隔正则项还带来两项额外优势：(i) 它在位点层面施加具有生物学意义的间隔，使特征重要性更加稳健；(ii) top-k 设计将监督集中于最具挑战性的 CpG位点，从而提升收敛效率和可解释性。

学习率、隐藏维度、层数、丢弃、标签平滑和权重衰减等超参数通过贝叶斯优化（Optuna） [29]进行调整，采用树结构帕岑估计器（TPE）采样器，并通过中位数剪枝提前停止表现不佳的试验。搜索空间涵盖离散范围和连续范围，并根据验证集 AUC 选出最佳配置。使用曲线下面积、准确率和 F1 分数评估留出的测试集上的性能。为确保稳健性，每项实验均使用 10 个不同的随机种子重复进行，并报告所有指标在各次运行中的均值和标准差。

III. 结果与分析

A. 各数据集中差异甲基化有限

我们使用 ChAMP 流程，分别对每个 GEO 数据集进行差异甲基化分析，并进行假发现率（FDR）校正。如图 2 所示，我们对每个数据集中的差异甲基化位点（DMPs）和差异甲基化区域（DMRs）进行了系统评估。几乎在所有数据集中，逐 CpG 位点的差异检验都未发现多少可重复的信号。在 GSE125895（内嗅皮层、海马、背外侧前额叶皮层、小脑）中，没有位点通过 FDR 校正，名义关联在不同脑区间的重叠度极低，方向也不一致。GSE134379（颞中回、小脑）和 GSE80970（前额叶皮层、颞上皮层）也呈现相同模式：火山图紧密聚集于中心围绕 $\Delta\beta\approx0$，只有少数离群点；排名靠前的基因座具有脑区和数据集特异性，且未能在不同脑区间重复验证。GSE66351 中分选出的神经元和胶质细胞组分在进行协变量调整后，同样未发现稳健的 CpG 关联。在多组织研究 GSE59685 中，少数皮层基因座在单个组织内接近名义显著性，但未能在不同组织间重复出现；配对的全血组则基本呈现无信号模式。全血队列 GSE144858 也呈现对称、近乎无信号的火山图，没有 CpG 位点达到 FDR 阈值。

如图 2 A-D 所示，分布证据与这些火山图模式一致。各组的 β值密度仍按预期呈双峰分布，且诊断组之间高度重叠；未见表明存在广泛甲基化差异的整体偏移或方差变化。综合来看，这些观察结果表明，效应量的绝对值较小，名义显著性结果在不同脑区间不稳定，血液中的信号尤其微弱。因此，在这种情况下，直接进行位点级差异分析只能得到有限且无法重复验证的 CpG信号，这凸显了仅凭单变量检验来筛选阿尔茨海默病外周甲基化生物标志物的难度。

差异甲基化区域分析（如图 2 E–H 所示）在部分数据集中鉴定出略多的候选区域。在 *ADNI*全血中，差异甲基化区域的基因集富集分析突出了 Polycomb/双价性程序（例如 BENPORATH_SUZ12_TARGETS、BENPORATH_EED_TARGETS 和 BENPORATH_PRC2_TARGETS），以及一般性发育和癌症甲基化特征（例如 HATADA_METHYLATED_IN_LUNG_CANCER_UP），表明其反映的是外周细胞状态调控，而非神经元病理。其他血液数据集也呈现出相同模式：*GSE66351* 富集于造血/miRNA靶标模块（例如 miR-223、miR-150）和广泛的细胞内程序，而 *GSE80970* 则主要表现为构成性和印记信号（例如 chrXq27 以及 LOPEZ_MBD_TARGETS_IMPRINTED_AND_X_LINKED）。相比之下，脑组织数据集呈现出神经与突触调控主题：*GSE125895* 显示出神经谱系和双价标记（MEISSNER_BRAIN_HCP、H3K27me3/H3K4me3）的显著富集，以及神经元相关微RNA（例如 miR-138、miR-133b）的富集；*GSE134379*则呈现出类似的 PRC2/双价性与神经/胶质细胞相关微RNA模块组合。这些差异与以下情况一致：早期 tau 易感的皮层区域在脑内经历突触退化和广泛的转录重编程，而外周甲基化主要反映全身免疫和炎症过程 [12]。然而，这些结果仍具有高度的数据集特异性，不同研究之间几乎没有重叠。

总体而言，这些研究结果表明，传统的单变量差异甲基化分析所提供的信号极其有限：大多数数据集均未发现显著的CpG位点，而全血数据集的表现最差。这种缺乏可重复性和跨组织一致性的现象凸显了直接进行位点级检验的内在局限性，并促使人们采用整合建模策略，以发现与阿尔茨海默病相关的细微组合甲基化特征

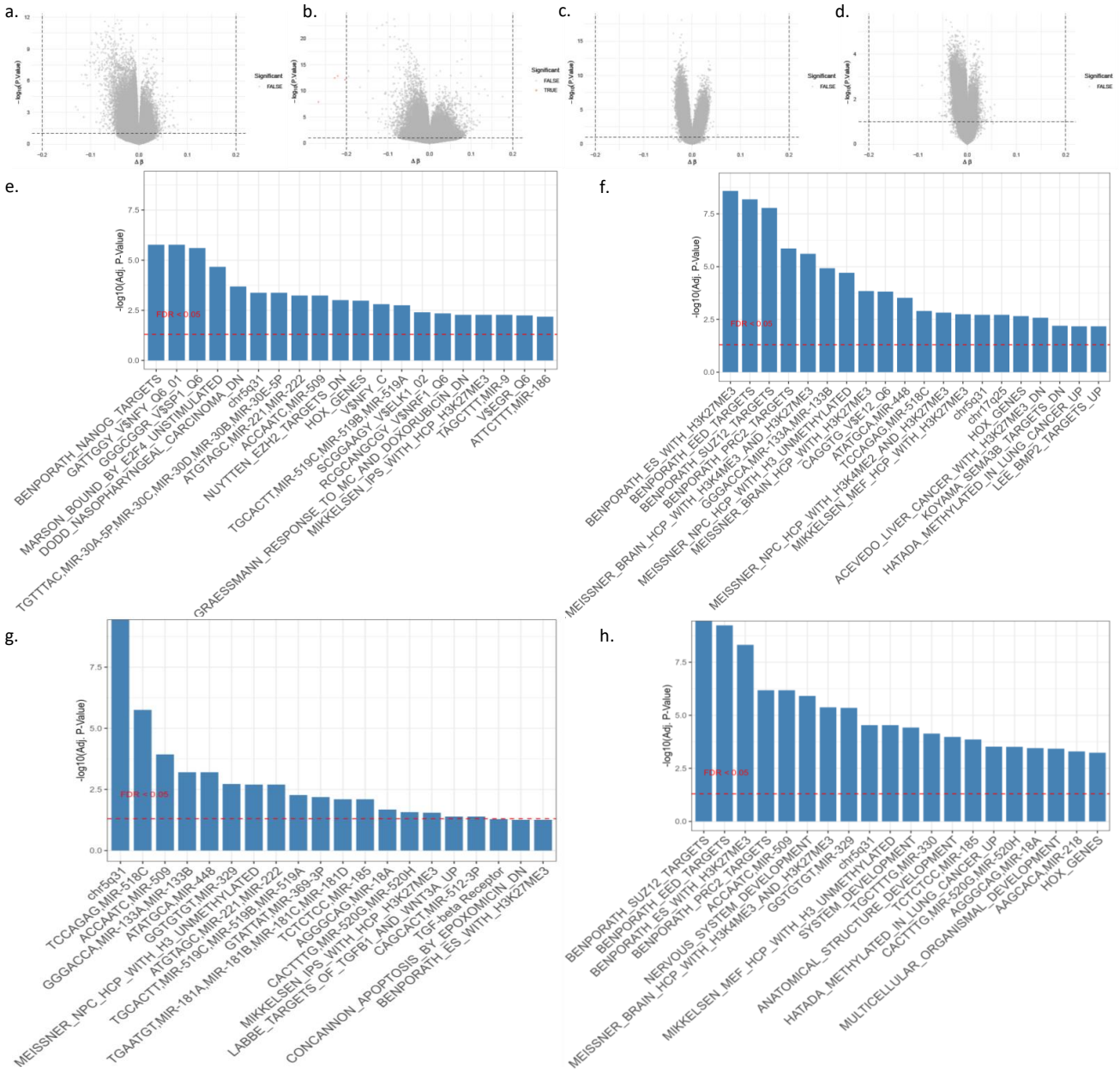


图 2：**各个数据集的差异甲基化谱**。(a-d)：火山图展示了位点级差异甲基化（原始 p 值，$|\Delta\beta| > 0.2$），并表明全基因组显著差异甲基化位点十分稀少。(e-h)：基于差异甲基化区域的基因集富集分析（$-\log_{10}$ 错误发现率）揭示了各数据集间存在异质性的通路富集。总体而言，这些结果表明，单独进行位点级检验只能获得有限且不一致的信号，因此有必要采用整合建模。

疾病。

B. 模型性能

1) 单个数据集揭示组织区域特异性的表观遗传敏感性：我们在多个相互独立的 DNA 甲基化数据集上评估所提出的框架，用于区分阿尔茨海默病（AD）与认知正常对照组，涵盖脑组织和外周组织。所有数据集均采用一致的配置进行处理，未针对特定数据集进行微调，以确保可比性，并反映生物学信号和样本的内在异质性

构成。分类结果如表III所示，揭示出明显的组织依赖性模式。源自易受阿尔茨海默病病理影响的皮层和海马区域的数据集表现出最强的判别性能，曲线下面积约为0.90–0.97，平均准确率高于0.85。这些脑区的特征包括早期tau蛋白积累、突触退化以及广泛的转录重编程，而这些变化均伴随着显著的DNA甲基化改变 [11]。相比之下，ADNI和GSE144858等全血数据集的分类能力明显较弱，曲线下面积约为

0.55–0.66。这种性能下降与以下事实一致：外周甲基化在很大程度上反映的是全身免疫和炎症过程，而非直接的神经元病理 [30]。

在解剖结构相似的皮层区域所衍生的数据集之间，也观察到显著的性能差异。这种差异可能由细胞类型组成的不同所致，因为神经元富集样本与胶质细胞富集样本呈现出不同的甲基化谱；也可能与测量平台和标准化流程等技术因素有关[31]。单细胞甲基组研究表明，神经元和非神经元细胞群之间的甲基化和染色质组织存在显著差异 [32]，这表明，组织整体测量反映的是多个功能各异的细胞贡献的综合结果。尽管本研究分析未采用针对特定数据集的优化，但不同区域和组织间观察到的梯度具有明确的生物学意义。与阿尔茨海默病相关的甲基化变化，在发生直接神经退行性变和胶质细胞重塑的皮层和海马区域最为显著 [12]，而外周组织则包含较弱但互补的系统性信息。这些结果表明，与阿尔茨海默病相关的表观遗传改变在不同组织中的分布并不均匀，某些区域对疾病分类的贡献尤为突出。这一观察促使我们开展后续的组合数据集实验，将所有组织来源纳入联合建模，以捕捉共有的和区域特异性的甲基化特征，并定量评估它们对阿尔茨海默病病理的相对贡献。

2）组合甲基化基准测试：为定量评估整合多种组织来源是否能提升稳健性和泛化能力，我们接下来进行了组合数据集实验，在统一特征空间中对所有脑部和外周甲基组进行联合建模。在此设置下，来自不同队列和组织类型的样本经协调一致的预处理和标准化后合并，使模型能够学习差异甲基化的整体模式，同时保留区域特异性背景。为进行比较评估，我们训练了一系列基线分类器，包括概率型分类器（高斯朴素贝叶斯）、线性分类器（逻辑回归、线性判别分析和支持向量机）以及非线性集成方法（随机森林和梯度提升），并使用相同的特征表示和交叉验证划分。此设计确保观察到的性能差异反映的是模型能力，而非数据不平衡或预处理伪影。

与单一数据集分析相比，合并框架在区分 AD 与对照样本方面表现出显著更高的判别能力。如表 IV 所总结，MethConvTransformer 模型在所有评估指标上的平均表现均为最高，其 ROC 曲线下面积（曲线下面积）为 **0。842 ± 0. 021**，分类准确率为 **0。774 ± 0.022**，F1 分数为 **0。803 ± 0. 017**。逻辑回归、线性判别分析（LDA）和线性支持向量机等经典线性基线模型展现出中等但稳定的预测能力，而基于树的集成模型（梯度提升和随机森林）的表现波动更大，且泛化能力相对较弱。根据 Welch 双样本 检验，MethConvTransformer 的性能显著高于大多数基线模型（）。唯一的例外是 LinearSVM 和 L2 正则化逻辑回归，其曲线下面积差异未达到统计显著性，尽管准确率和 F1 分数均取得了显著提升，这凸显了整合建模在揭示复杂的多区室表观遗传特征以及提升疾病分类性能方面的价值。

C. 可解释性分析

1）通过线性归因和*SHAP*归因剖析单个*CpG*的相关性：为阐明甲基化特征如何促进疾病分类，我们采用了双重可解释性策略，整合模型内在分析与事后归因分析。在CpG层面，我们引入了线性投影层，为每个位点分配明确的系数权重，从而直接估算其对模型决策边界的贡献。这些系数量化了每个CpG影响的方向和大小，但并不意味着其具有显著的个体效应；相反，它们反映了多个微弱且分布于不同位置的信号如何在线性网络中汇总。与此同时，我们计算了基于梯度的SHAP值，以捕捉扰动每个CpG对输出概率的边际影响，并将基因座、组织和协变量之间的非线性依赖关系纳入考量。结合这两种方法，可以区分结构编码的特征相关性（来自投影权重）与依赖上下文的预测影响（来自SHAP），从而全面表征模型可解释性。

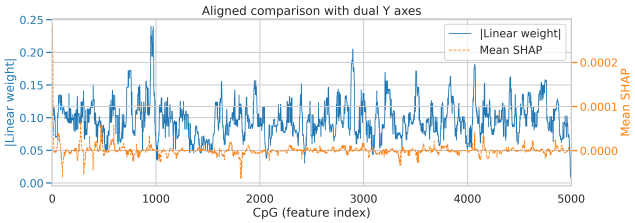


图3：**CpG线性投影权重与基于SHAP的特征归因比较。**蓝色曲线表示线性投影层的绝对系数，反映网络内部固有的特征相关性。橙色虚线表示跨样本的平均SHAP值，概括了依赖上下文的边际效应。两种曲线均表明，大多数CpG位点的个体效应较弱，预测信息主要来自分布式、协同作用的甲基化模式，而非孤立的基因座。

2）位点间注意力模式与组织情境调节：如图3所示，线性权重的绝对值呈现高频波动，仅有少数局部峰值，表明大多数 CpG 位点的个体效应有限，只有一小部分位点对模型判别有中等贡献。平均 SHAP 曲线在零值附近呈现较平滑的振荡，反映出当跨队列取平均时，单个 CpG 位点的边际效应总体较弱。

表 III：不同数据集和组织区域的性能汇总。

数据集	组织区域	AUC	ACC	F1
ADNI	全血	0.55 ± 0.07	0.62 ± 0.02	0.37 ± 0.14
GSE125895	内嗅皮层；海马；背外侧前额叶皮层；小脑	0.95 ± 0.04	0.91 ± 0.09	0.79 ± 0.29
GSE134379	颞中回；小脑	0.62 ± 0.04	0.63 ± 0.03	0.69 ± 0.05
GSE144858	全血	0.66 ± 0.11	0.69 ± 0.06	0.58 ± 0.15
GSE59685	内嗅皮层；颞上回；前额叶皮层；小脑；全血	0.95 ± 0.09	0.94 ± 0.07	0.96 ± 0.05
GSE66351	额叶皮层神经元和胶质细胞	0.74 ± 0.11	0.78 ± 0.07	0.84 ± 0.05
GSE80970	前额叶皮层；颞上回	0.90 ± 0.08	0.85 ± 0.12	0.86 ± 0.10

注。 每个数据集均使用 $n = 10$ 个随机种子进行评估（均值 ± 标准差）。模型在统一配置下进行训练，且未针对每个数据集进行微调。缩写：ERC = 内嗅皮层；DLPFC = 背外侧前额叶皮层；STG = 颞上回；PFC = 前额叶皮层；MTG = 颞中回。

表 IV：MethConvTransformer 与基线模型的比较（均值 ± 标准差及 Welch t 检验 p 值）。

模型	AUC	p	ACC	p	F1	p
高斯朴素贝叶斯	0.554 ± 0.010	1.78e-29	0.536 ± 0.026	4.37e-14	0.490 ± 0.085	8.57e-07
梯度提升	0.741 ± 0.015	6.94e-13	0.668 ± 0.011	2.18e-16	0.706 ± 0.011	6.36e-16
KNN	0.651 ± 0.007	3.38e-24	0.624 ± 0.011	2.50e-20	0.666 ± 0.021	1.02e-11
LDA	0.781 ± 0.011	2.96e-10	0.708 ± 0.016	4.06e-09	0.740 ± 0.012	4.83e-11
LinearSVM	0.829 ± 0.010	2.77e-01	0.733 ± 0.028	1.67e-03	0.778 ± 0.019	8.19e-03
L1 正则化逻辑回归	0.769 ± 0.011	3.04e-12	0.691 ± 0.019	1.32e-09	0.724 ± 0.023	2.03e-07
L2 正则化逻辑回归	0.825 ± 0.011	3.32e-01	0.741 ± 0.015	1.09e-02	0.770 ± 0.013	1.24e-02
MLP	0.498 ± 0.002	1.30e-29	0.526 ± 0.031	1.03e-14	0.398 ± 0.213	2.84e-06
随机森林	0.701 ± 0.011	1.55e-20	0.640 ± 0.016	7.17e-18	0.676 ± 0.024	1.06e-12
MethConvTransformer	**0.842 ± 0.021**	–	**0.774 ± 0.022**	–	**0.803 ± 0.017**	–

注。 数值表示跨随机种子的均值 ± 标准差。采用 Welch 双样本 t 检验评估各基线模型与 MethConvTransformer 之间性能差异的统计显著性。粗体表示每项指标的最高平均性能。较小的 p 值表示差异具有统计显著性的证据更强（$p < 0.05$）。LinearSVM 和 L2 正则化逻辑回归在曲线下面积方面与 MethConvTransformer 的差异不显著（$p > 0.05$）。

这一模式与图2中经典差异甲基化分析的结果一致：只有少数 CpG 位点超过统计阈值，这凸显了单个位点变化本身不足以实现稳健的组间区分。相反，模型整合了大量细微且相关的甲基化变异，以形成预测性表征。两种归因指标均呈一致升高的区域表明，CpG 簇在内在和外在可解释性分析中均有稳定贡献，说明这些位点可能协同发挥作用，但并非主要预测位点。在更高的表征层面，观察到组织嵌入会调节注意力图，突出显示组织间甲基化变异较高的 CpG 簇，从而将分子异质性锚定于解剖学情境。

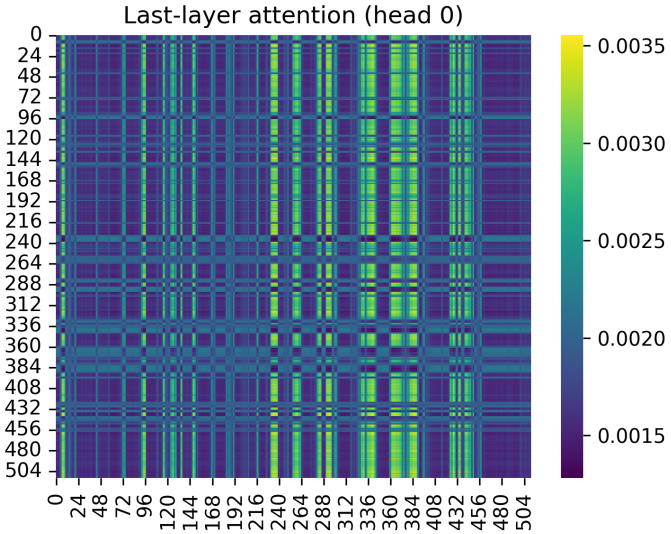


图 4：**最后一个 Transformer 层中的注意力模式**。一个代表性注意力头的自注意力权重热图，展示了 CpG 嵌入之间的成对依赖关系。大多数注意力值接近于零，表明相互作用稀疏且局部化。周期性出现的十字形热点突显了彼此增强表征的 CpG 簇，这表明模型捕捉到了区域特异性的共甲基化或共享功能调控。

为进一步探究模型如何整合不同 CpG 位点的信息，我们检查了最后一个 Transformer 层的注意力权重。图 4 展示了一个具有代表性的注意力头的注意力矩阵，显示了所有输入位点之间的成对依赖强度。由此形成的模式呈现出稀疏的网格状结构，只有少数 CpG 位点对表现出较强的相互注意力，而大多数位点对的权重接近于零。这表明，模型会选择性地关注少量相互作用，而不是将注意力均匀分配到所有位点。观察到的高强度对角线和周期性交叉连接表明，可能存在局部且局限于特定区域的注意力，这或许与已知的基因组

模块化特征相符——例如共调控CpG簇或染色质邻域效应——而非远距离、无结构的依赖关系。

从定量角度看，最终层的整体注意力熵始终较低，这表明随着模型收敛，注意力变得更加集中。这种聚焦行为说明，较高层捕捉到越来越具体的特征关系，并直接用于分类。纳入组织嵌入后，这些注意力热点的幅度和空间分布范围会随组织而系统性变化，反映出模型如何在不同生物学环境下对甲基化相互作用进行情境化表征。总之，这些发现表明，模型学到了具有生物学意义且具备情境感知能力的共甲基化结构，其中只有少数 CpG 相互作用对稳健预测至关重要。

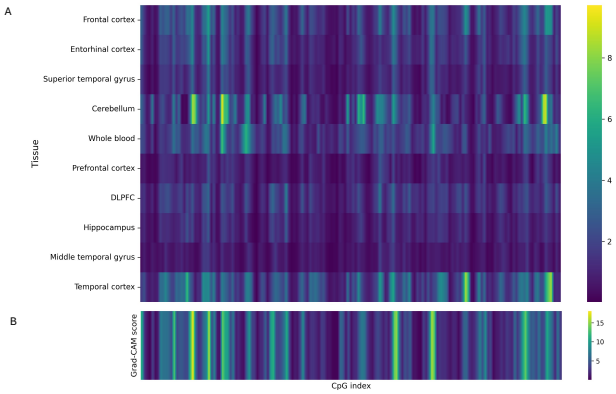
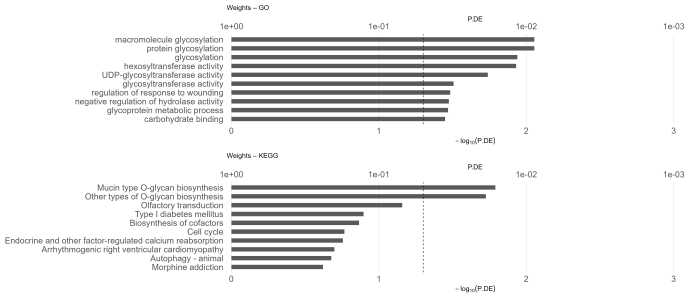


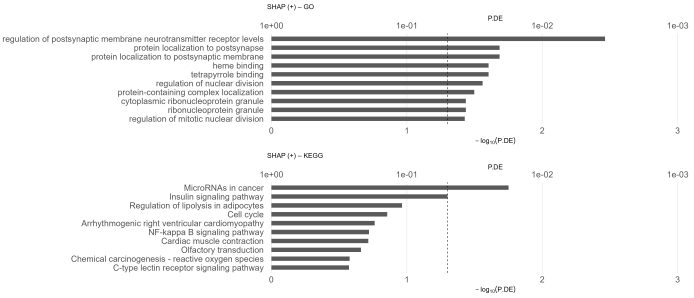
图 5：**Grad-CAM++ 基于甲基化的阿尔茨海默病分类的可解释性**。(A) 组织特异性 Grad-CAM++激活图，展示了 10 种组织中归因幅度的平均值。行对应组织，列对应区域层面的 CpG索引。(B) 汇总各组织的整体 Grad-CAM++ 重要性分布，突出显示了有助于阿尔茨海默病分类的 CpG 簇。

3) 组织特异性 *Grad-CAM++* 归因揭示区域性表观遗传脆弱性：Grad-CAM++ 分析刻画了驱动阿尔茨海默病分类的区域甲基化模式（图 5）。与 SHAP 或线性投影分析不同，后者将重要性归因于单个 CpG 位点；Grad-CAM++则作用于卷积特征图，并量化局部 CpG 邻域中的空间整合激活。因此，它反映的是区域甲基化模式的整体贡献，而非孤立的单个位点效应。

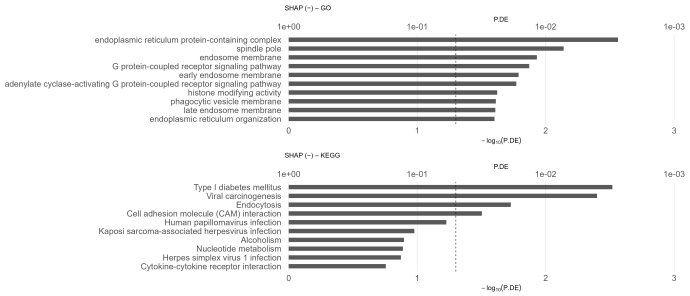
结果（图 5 A）显示，小脑表现出最强的Grad-CAM++激活值，其次是颞叶皮层、额叶皮层、内嗅皮层和全血。所有其他组织的激活模式均相对较弱。小脑激活值的突出表现表明，该模型认为该组织中的区域性甲基化变化具有较高的预测相关性。这一观察结果与先前的证据一致：尽管小脑相对未受明显神经退行性变的影响，但在阿尔茨海默病中会经历广泛的代谢和表观遗传重塑 [7], [33]。颞叶皮层和额叶皮层同样表现出升高的



(a) 使用线性投影权重绝对值最高的CpG位点进行富集分析，突出显示免疫受体信号传导和聚糖生物合成通路。



(b) 使用SHAP值最高的正值对应的CpG位点进行富集分析，显示脂质代谢、高尔基体组织和能量产生过程。



(c) 使用SHAP值最负的CpG位点进行富集分析，这些位点与水解酶活性、GPCR信号传导和内质网组织相关。

图 6：基于 MethConvTransformer 模型不同可解释性层所优先识别的 CpG位点进行的基因本体（GO）和 KEGG 通路富集分析。富集结果表明，从免疫激活和糖基化到脂质代谢及细胞应激反应等不同生物过程，共同促成了与阿尔茨海默病相关的甲基化特征。

Grad-CAM++ 得分与突触丢失和认知障碍所影响的区域相符 [34], [35], ，而中等程度的内嗅皮层激活则与其在早期 tau 病理中的既定作用相一致。全血中较弱的 Grad-CAM++ 信号表明，其判别能力低于脑组织，但仍显示出可测量的外周反映，体现了疾病相关的甲基化改变 [36], [37]。需要注意的是，不同组织的样本量差异显著，这可能会影响 Grad-CAM++ 激活值的大小，因为样本较多的组织会在平均图中贡献更多梯度信息。此外，Grad-CAM++强度较低，并不意味着这些组织中的CpG位点

对阿尔茨海默病分类没有参考价值。相反，这表明仅依赖有限的CpG位点子集难以实现准确预测，且疾病相关的甲基化模式分布在多个区域。整体 Grad-CAM++ 特征（图 5B）突出了在不同组织中持续有助于分类的 CpG 簇，它们可能代表共享的调控模块，例如神经炎症和突触过程。

4) 富集分析：我们对模型不同可解释性层所优先筛选的 CpG 位点进行基因本体（GO）和 KEGG 通路富集分析 [38], [39] 。我们分析三组 CpG 位点：按模型权重绝对值排名前 5% 的位点、具有正向贡献的前 5% 位点，以及具有负向贡献的前 5% 位点。使用 Illumina HumanMethylation450k注释 [40], 将 CpG位点映射至基因，并从中选取与我们的特征空间相匹配的背景 CpG 位点作为检验全集。采用 missMethyl软件包进行富集分析，并使用非中心超几何（Wallenius）模型校正 CpG 探针数量不均等的问题 [41], [42]。为考虑多重比较问题，采用 Benjamini–Hochberg 程序控制假发现率，同时报告原始 p 值和错误发现率（FDR） q值 [43]。

图 6 总结了基于权重衍生重要性以及正、负 SHAP贡献所识别出的主要生物过程和通路。具有较高 Linger 投影权重衍生重要性的特征主要富集于免疫受体信号传导和聚糖生物合成通路 [44], [45]。诸如激活免疫反应/调节免疫反应的细胞表面受体信号传导、蛋白酪氨酸激酶活性以及蛋白质自磷酸化等 GO 术语具有高度显著性，同时还富集于包括黏蛋白型*O*-聚糖生物合成、其他类型*O*-聚糖生物合成、细胞周期和内分泌及其他因子调节的钙重吸收在内的 KEGG 通路。这些富集结果表明，我们的模型捕捉到了与免疫调节和翻译后糖基化过程相关的甲基化特征 [46], [47]。具有正向 SHAP贡献的 CpG位点富集于脂质代谢过程 [48],、高尔基体组织和细胞能量产生。代表性 GO 术语包括糖鞘脂代谢过程、作用于酯键的水解酶活性、高尔基体池、高尔基体堆和前体代谢物和能量的生成；而包括癌症中的微*RNA*、胰岛素信号传导、脂肪细胞脂解调节和*NF–κB* 信号传导在内的 KEGG 通路则突出了代谢和炎症信号网络。这些发现表明，脂质处理改变和内膜系统动态是阿尔茨海默病跨组织甲基化改变的关键组成部分。

相反，具有负 SHAP值的CpG位点与作用于碳氮键的水解酶活性、G蛋白偶联受体（GPCR）信号传导以及内质网（ER）组织相关。富集的GO术语包括作用于碳氮键（非肽键）的水解酶活性、紫外光反应、跨膜运输负调控和内质网组织。相关KEGG通路涉及*1*型糖尿病、病毒致癌、内吞作用、细胞黏附分子以及多种病毒感染通路（例如，人乳头瘤病毒感染 和单纯疱疹病毒感染）。负权重特征中出现代谢和传染病通路，说明与全身性合并症相关的甲基化特征 [49]可能会抵消阿尔茨海默病相关信号，或反映保护性影响或混杂因素。

总体而言，这些富集结果表明，我们的基于Transformer的模型识别出广泛的生物学过程，涵盖免疫激活、糖基化、脂质代谢、信号转导和细胞应激反应。由注意力权重和SHAP值得出的互补特征显示，注意力机制侧重于免疫和糖基化机制，而SHAP则凸显了会以相反方向调节疾病概率的代谢和信号传导过程。总之，这些发现提升了模型的可解释性，并为跨组织DNA甲基化如何参与阿尔茨海默病病理生理学提供了具有生物学依据的假设。

IV. 讨论

A. 主要贡献

本研究提出**MethConvTransformer**，这是一种跨组织Transformer架构，旨在对CpG位点之间的局部和长程依赖关系进行建模，同时显式纳入协变量和组织背景。在对六个公开可用的GEO队列以及一个源自阿尔茨海默病神经影像学倡议（ADNI）的独立验证集进行基准测试时，该模型始终优于经典机器学习基线方法和当代深度学习方法，展现出在不同预处理流程、组织取样和队列构成下的稳健性。这些结果使MethConvTransformer跻身首批专为跨组织推断和多分辨率可解释性而优化的基于Transformer的甲基化模型之列——这两个方面制约着现有的大多数表观遗传预测框架。

与当前甲基化疾病预测领域的先进方法相比——该领域主要由弹性网络特征、基于树的集成模型、卷积神经网络（CNNs）以及近年来出现的基于图和基于Transformer的方法主导——我们的方法实现了三项主要进展。首先，我们引入了逐*CpG*线性投影，使深度学习表征与表观遗传流行病学中传统采用的效应量推理相一致，从而生成在不同组织间仍具可解释性和可比性的重要性评分。其次，卷积下采样模块在减少标记长度的同时保留短程共甲基化结构，使Transformer注意力层能够将容量用于捕捉更高阶的依赖关系——这是相较于传统卷积神经网络或循环神经网络模型的一项架构优势。第三，使用组织嵌入和协变量编码器，使模型能够将阿尔茨海默病的共享性、全身性甲基化特征，与组织特异性效应或人口统计学效应区分开来，而无需进行按组织分层训练。这种统一建模能力直接解决了跨组织甲基化研究中的一个主要缺口

在这类研究中，既往研究在很大程度上依赖荟萃分析或事后协调化，而不是集成式架构。

B. 生物学洞见与可解释性概述

除了分类性能之外，MethConvTransformer 还通过整合全局 CpG权重、基于 SHAP 的受试者特异性边际效应，以及针对区域性 CpG邻域的 Grad-CAM++ 注意力图，构建了一个多层次可解释性框架。这些可解释性层共同指向生物学上连贯的模块。全局权重凸显了免疫受体信号传导和糖基化通路——这些过程已被反复证实与神经炎症、免疫–脑串扰及阿尔茨海默病病理有关 [10]。SHAP 分析进一步识别出代谢、囊泡运输以及内质网/高尔基体组织相关程序；这些程序的组织特异性变化会调节类别概率，这与近期证据相符：脂质重塑、能量代谢缺陷和内膜系统应激会影响神经元韧性及小胶质细胞状态。相较之下，负向 SHAP 特征指向合并症相关程序（例如糖尿病和病毒应答），提示全身性暴露可能会干扰与阿尔茨海默病相关的外周甲基化特征。

综合来看，这些发现支持一种日益得到大规模表观基因组范围关联研究证实的范式：阿尔茨海默病相关甲基化特征具有多基因特性，分布于多个组织，涉及许多效应较小的 CpG 位点，而非少数效应较大的基因座 [36]。MethConvTransformer 能够在血液和脑组织中识别此类分布式模式，这表明了一个更广泛的原则：当 Transformer 架构采用可解释的归纳偏置（如逐 CpG 投影、卷积下采样和组织嵌入）进行结构化设计时，即便存在组织异质性、芯片差异和复杂的混杂因素，也能捕捉与疾病相关的生物学特征。

C. 临床与信息学意义

从转化医学的角度来看，我们的结果表明，单靠外周甲基化特征，其区分能力可能难以媲美脑源性甲基化谱（如大脑皮层或海马），但将其纳入综合风险模型后，仍可能带来显著价值；这类模型可结合外周甲基化与人口统计学变量、认知测试、影像学或体液生物标志物。此处发现的可解释 CpG 模块为后续验证提供了可检验的假设，有助于构建精简的靶向甲基化检测面板（如亚硫酸氢盐扩增子检测），并支持针对特定细胞类型的机制研究。在实际应用中，可解释的跨组织甲基化模型有望支持早期风险分层、优化预防或早期干预试验的患者筛选，并促进对治疗反应的纵向监测。

从生物医学信息学的角度来看，这项研究展示了如何利用先进架构（Transformer）对高维表观遗传数据（约含 100,000+ 个 CpG 特征）进行建模，同时保留可解释性——这是临床应用的必要前提。通过将经典的效应量推理与现代深度学习相结合，并纳入明确的组织/情境嵌入，MethConvTransformer 为健康信息学领域未来的高维生物标志物建模确立了蓝图。

D. 局限性与未来工作

尽管所提出的 MethConvTransformer 框架表现出良好的性能和可解释性，但仍存在若干局限。首先，本研究分析的是回顾性公开队列，这些队列在预处理、芯片平台和采样方面各不相同，即使经过协调化，仍可能存在残余批次效应。其次，尽管该模型整合了多种组织，但匹配的脑—血液样本数量有限，制约了对组织间对应关系的直接推断，也限制了因果解释。第三，组织整体甲基化数据会对异质性细胞群体进行平均；纳入单细胞或分选细胞甲基组应能提高生物学分辨率，并阐明细胞类型特异性机制。第四，作为一种数据驱动方法，结果取决于超参数、优化动态和初始化，因此有必要开展更广泛的稳健性探测（例如跨实验室重复验证、扰动测试和校准分析）。最后，目前的工作侧重于横断面分类；拓展至纵向预测 [50] 转化、衰退速率或治疗反应，并整合多组学或影像学模态，应能提升其转化应用价值。未来的研究方向包括：(i) 在不同研究中心对候选 CpG 面板和通路模块进行前瞻性验证；(ii) 拓展多视图方法，将甲基化与转录组、蛋白质组或影像学特征联合编码；(iii) 纳入显式的细胞类型去卷积或单细胞引导；以及 (iv) 加入因果发现组件 [51]（例如不变风险或反事实目标），以区分疾病生物学与队列及共病效应。

V. 结论

我们提出了 MethConvTransformer，这是一种可解释的跨组织 Transformer，通过统一逐CpG线性归因、卷积上下文和自注意力，捕捉与阿尔茨海默病相关的分布式甲基化依赖关系。该模型在异质性队列中取得了当前最先进的性能，并产生了稳定且具有生物学一致性的归因结果，这些结果聚焦于免疫信号传导、糖基化、脂质代谢以及内质网/高尔基体组织。通过协调预测准确率与多分辨率可解释性，该框架推动了可复现的表观遗传生物标志物发现，并提出了可检验的通路层面假设。这些特征使 MethConvTransformer 成为阿尔茨海默病前瞻性验证、靶向检测设计和纵向监测的实用基础，也为其他复杂疾病中的跨组织甲基化建模提供了可泛化的范本。

VI. 资助

本研究获得了美国国立卫生研究院（NIH）授予 Z.Z.] 的资助，项目编号为 [U01AG079847 和 R01LM012806。G.Q. 获得了

由德克萨斯州癌症预防与研究所（CPRIT RP210045）资助的生物医学信息学、基因组学与转化癌症研究培训计划（BIG-TCR）提供的CPRIT博士后奖学金。感谢 UTHealth癌症基因组学核心提供技术支持（CPRIT RP240610）。

VII. 致谢

阿尔茨海默病神经影像学倡议（ADNI）的数据收集与共享由美国国立卫生研究院国家老龄化研究所资助（资助编号U19AG024904）。受资助机构为北加利福尼亚研究与教育研究所。此前，ADNI 还曾获得美国国家生物医学成像与生物工程研究所、加拿大卫生研究院以及通过美国国立卫生研究院基金会（FNIH）提供的私营部门捐款的资助，其中包括以下机构的慷慨捐助：艾伯维、阿尔茨海默病协会、阿尔茨海默病药物发现基金会、Araclon Biotech、BioClinica, Inc.、Biogen、Bristol Myers Squibb Company、CereSpir, Inc.、Cogstate、卫材株式会社、Elan Pharmaceuticals, Inc.、礼来公司、EuroImmun、F. Hoffmann-La Roche Ltd 及其关联公司 Genentech, Inc.、Fujirebio、GE Healthcare、IXICO Ltd.、Janssen Alzheimer Immunotherapy Research & Development, LLC.、Johnson & Johnson Pharmaceutical Research & Development LLC.、Lumosity、Lundbeck、Merck & Co., Inc.、Meso Scale Diagnostics, LLC.、NeuroRx Research、Neurotrack Technologies、Novartis Pharmaceuticals Corporation、辉瑞公司、Piramal Imaging、施维雅、武田药品工业株式会社以及 Transition Therapeutics。

VIII. 利益冲突

作者声明不存在利益冲突。

IX. 作者贡献

G.Q. 构思了这项研究、开展了实验、分析了结果并撰写了论文初稿。G.L. 协助进行数据预处理和分析，并参与框架开发。Z.Z. 共同构思了这项研究，为研究设计提供指导，并对论文进行了重要修订。所有作者均已阅读并批准论文最终版本。

参考文献

[1] A. Association，“2025年阿尔茨海默病事实与数据”，*Alzheimer's & Dementia*，第21卷，第4期，e70235页，2025年。 [在线]。可访问：https://alz-journals.onlinelibrary.wiley.com/doi/abs/ 10 1002 70235 。/alz。

[2] J. Xu、S. L. Murphy、K. D. Kochanek 和 E. Arias，“死亡：2022年最终数据”，国家生命统计报告，第4期，第1页，2025年。[3] L. D. Moore、T. Le 和 G. Fan，“DNA甲基化及其基本功能”，神经精神药理学，第38卷，第1期，第23–38页，2013年。[4] S. Horvath，“人体组织和细胞类型的DNA甲基化年龄”，基因组生物学，第14卷，第10期，3156页，2013年。[5] C. Sarnowski 等，“多组织表观遗传分析揭示了CPT1A位点与胰岛素抵抗和阿尔茨海默病相关的不同关联”，临床表观遗传学，第15卷，第1期，173页，2023年。[6] K. Lunnon 等，“甲基组分析表明阿尔茨海默病中ANK1的皮层失调”，自然神经科学，第17卷，第9期，第1164–1170页，2014年。[7] P. L. De Jager 等，“阿尔茨海默病：脑DNA甲基化在ANK1、BIN1、RHBDF2及其他基因座的早期变化”，自然神经科学，第17卷，第9期，第1156–1163页，2014年。

[8] A. S. Yokoyama、J. C. Rutledge 和 V. Medici，《阿尔茨海默病中的DNA甲基化改变》，Environmental Epigenetics，第3卷，第2期，文章编号dvx008，2017年。[9] A.-L. Lang 等人，《老年人脑中阿尔茨海默病神经病理改变的甲基化差异》，Acta Neuropathologica Communications，第10卷，第1期，文章编号174，2022年。[10] L. Zhang 等人，《前额叶皮层DNA甲基化差异的表观基因组范围荟萃分析提示阿尔茨海默病中的免疫过程》，Nature Communications，第11卷，文章编号6114，2020年，第1期。[11] G. Shireby 等人，《皮层中阿尔茨海默病神经病理学的DNA甲基化特征主要由非神经元细胞类型的差异驱动》，Nature Communications，第13卷，文章编号5620，2022年，第1期。[12] G. Gasparoni 等人，《对纯化神经元和胶质细胞进行DNA甲基化分析，揭示人类大脑皮层中与年龄及阿尔茨海默病相关的特异性改变》，Epigenetics & Chromatin，第11卷，第1期，文章编号41，2018年。[13] R. G. Smith 等人，《EMIF-AD研究中与阿尔茨海默病脑脊液生物标志物相关的血液DNA甲基组特征》，Alzheimer's & Dementia，第20卷，第10期，第6722–6739页，2024。[14] P. D. Fransquet、P. Lacaze、R. Saffery、J. McNeil、R. Woods 和 J. Ryan，《血液DNA甲基化作为痴呆潜在生物标志物：系统综述》，Alzheimer's & Dementia，第14卷，文章编号103，2018年，第1期，第81页。[15] T. Kaleck 等人，《与脑脊液中YKL-40和NFL生物标志物水平相关的血液DNA甲基组特征的重复验证》，Alzheimer's & Dementia，第21卷，第9期，文章编号e70647，2025年。[16] J. Ren、B. Zhang、D. Wei 和 Z. Zhang，《基于机器学习识别阿尔茨海默病患者的甲基化基因生物标志物》，BioMed Research International，第2020卷，文章编号8348147，2020年，第1期。[17] Y. Huang 等人，《一种机器学习方法对脑表观遗传学进行分析，揭示与阿尔茨海默病相关的激酶》，Nature Communications，第12卷，第1期，文章编号4472，2021年。[18] L. Chen、A. J. Saykin、B. Yao、F. Zhao 和 A. D. N. Initiative，《利用外周血时间序列DNA甲基化数据预测阿尔茨海默病进展的多任务深度自编码器》，Computational and Structural Biotechnology Journal，第20卷，第5761–5774页，2022年。[19] T. C. Silva 等人，《阿尔茨海默病血液与脑表观基因组范围关联研究的跨组织分析》，Nature Communications，第13卷，第1期，文章编号4852，2022年。[20] G. Qu 等人，《可解释的认知能力预测：用于分析功能性脑网络的综合门控图变换器框架》，IEEE Transactions on Medical Imaging，第43卷，第4期，第1568–1578页，2023年。[21] S. A. Semick 等人，《多个脑区的DNA甲基化和基因表达整合分析提示阿尔茨海默病中的新基因》，Acta Neuropathologica，第137卷，第4期，第557–569页，2019年。[22] D. L. Brokaw 等人，《阿尔茨海默病中的细胞死亡与存活通路：利用组学数据集进行整合性假设检验》，Neurobiology of Aging，第95卷，第15–25页，2020年。[23] R. G. Smith 等人，《跨越HOXA基因簇的48 kb区域DNA甲基化升高与阿尔茨海默病神经病理学相关》，Alzheimer's & Dementia，第14卷，第12期，第1580–1588页，2018年。[24] J. A. Roubroeks 等人，《阿尔茨海默病血液表观基因组范围关联研究发现HOXB6基因中显著的DNA高甲基化》，Neurobiology of Aging，第95卷，第26–45页，2020年。[25] Y. Tian 等人，《ChAMP：更新的Illumina BeadChips甲基化分析流程》，Bioinformatics，第33卷，文章编号3984，2017年，第24期，第3982页。[26] W. Zhou、P. W. Laird 和 H. Shen，《Infinium DNA甲基化BeadChip探针的全面表征、注释与创新应用》，Nucleic Acids Research，第45卷，第4期，第e22–e22页，2017年。[27] A. Chattopadhay、A. Sarkar、P. Howlader 和 V. N. Balasubramanian，《Grad-CAM++：深度卷积网络的通用梯度视觉解释》，载于2018年IEEE冬季计算机视觉应用会议（WACV），IEEE，2018年，第839–847页。[28] G. Qu 等人，《用于认知能力预测的集成流形正则化多模态图卷积网络》，IEEE Transactions on Biomedical Engineering，第68卷，第12期，第3564–3573页，2021年。[29] T. Akiba、S. Sano、T. Yanase、T. Ohta 和 M. Koyama，《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，2019年，第2623–2631页。

[30] A. Ramakrishnan 等人，《阿尔茨海默病外周免疫中的表观遗传失调》，Neuron，第112卷，第8期，第1235–1248页，2024。[31] V. C. Alves、E. Carro 和 J. Figueiro-Silva，《揭示阿尔茨海默病中的DNA甲基化：基于芯片的人脑研究综述》，Neural Regeneration Research，第19卷，第11期，第2365–2376页，2024。[32] W. Tian 等人，《人脑中的单细胞DNA甲基化与三维基因组架构》，Science，第382卷，第6667期，文章编号eadf5357，2023年。[33] R. Lardenoije 等人，《衰老与神经退行性变的表观遗传学》，Progress in Neurobiology，第131卷，第21–64页，2015年。[34] H. Braak 和 E. Braak，《阿尔茨海默病相关改变的神经病理学分期》，Acta Neuropathologica，第82卷，第4期，第239–259页，1991年。[35] S. E. Arnold、B. T. Hyman、J. Flory、A. R. Damasio 和 G. W. Van Hoesen，《阿尔茨海默病患者大脑皮层中神经原纤维缠结和神经炎性斑块的地形学与神经解剖学分布》，Cerebral Cortex，第1卷，第1期，第103–116页，1991年。[36] R. G. Smith 等人，《阿尔茨海默病表观基因组范围关联研究的荟萃分析揭示大脑皮层中新的差异甲基化位点》，Nature Communications，第12卷，文章编号3517，2021年，第1期。[37] M. E. Levine 等人，《用于评估寿命和健康寿命的衰老表观遗传生物标志物》，Aging（纽约州奥尔巴尼），第10卷，第4期，文章编号573，2018年。[38] T. G. O. Consortium，《基因本体：用于统一生物学的工具》，Nature Genetics，第25卷，第1期，第25–29页，2000年。[39] M. Kanehisa 和 S. Goto，《KEGG：京都基因与基因组百科全书》，Nucleic Acids Research，第28卷，第1期，第27–30页，2000年。[40] J. Sandoval 等人，《用于检测人类基因组中450,000个CpG位点的DNA甲基化微阵列验证》，Epigenetics，第6卷，第6期，第692–702页，2011年。[41] B. Phipson、J. Maksimovic 和 A. Oshlack，《missMethyl：用于分析Illumina HumanMethylation450微阵列数据的R软件包》，Bioinformatics，第32卷，第2期，第286–288页，2016年。[42] A. Fog，《Wallenius非中心超几何分布的计算方法》，Communications in Statistics—Simulation and Computation®，第37卷，第2期，第258–273页，2008年。[43] Y. Benjamini 和 Y. Hochberg，《控制假发现率：一种实用且有效的多重检验方法》，Journal of the Royal Statistical Society: Series B (Methodological)，第57卷，第1期，第289–300页，1995年。[44] H. Haukedal 和 K. K. Freude，《糖基化对阿尔茨海默病的影响》，Frontiers in Neuroscience，第14卷，文章编号625348，2021年。[45] Y. Kang、Q. Zhang、S. Xu 和 Y. Yu，《阿尔茨海默病中糖缀合物的改变及其作用》，Frontiers in Aging Neuroscience，第16卷，文章编号1398641，2024年。[46] C. P. Boix、I. Lopez-Font、I. Cuchillo-Iba˜nez 和 J. S´aez-Valero，《阿尔茨海默病患者脑中的淀粉样前体蛋白糖基化发生改变》，Alzheimer’s Research & Therapy，第12卷，第1期，文章编号96，2020年。[47] W. Zhang、T. Chen、H. Zhao 和 S. Ren，《衰老与神经退行性疾病中的糖基化：衰老与神经退行性变中的糖基化》，Acta Biochimica et Biophysica Sinica，第56卷，文章编号1208，2024年，第8期。[48] K. Yuyama 和 Y. Igarashi，《将糖鞘脂与阿尔茨海默病淀粉样蛋白-ß联系起来：细胞外囊泡与功能性植物材料》，Glycoconjugate Journal，第39卷，第5期，第613–618页，2022年。[49] W. Zhang 等人，《用于认知健康的生活方式韧性指数的DNA甲基化特征》，Alzheimer’s Research & Therapy，第17卷，第1期，文章编号88，2025年。[50] L. Chen 等人，《可解释的时空图演化学习及其在发育期间动态脑网络分析中的应用》，NeuroImage，第298卷，文章编号120771，2024年。[51] Y. Wang 等人，《用于研究大脑发育期间动态有效连接变化的深度动态因果学习模型》，IEEE Transactions on Biomedical Engineering，第71卷，第12期，第3390–3401页

，2024年。