



Cys-DHE: 基于双分支异构深度集成的半胱氨酸S-羧乙基化位点预测模型*

魏欣^{1)**} 胡思亲²⁾⁽¹⁾ 江西服装学院商学院, 南昌 330201; ⁽²⁾ 江西服装学院大数据学院, 南昌 330201

摘要 **目的** 针对半胱氨酸S-羧乙基化位点实验鉴定成本较高、周期较长以及现有预测模型对位置权重和多尺度序列信息利用不足的问题, 构建一种具有较好综合预测性能的计算模型。**方法** 提出基于双分支异构深度集成的半胱氨酸S-羧乙基化位点预测模型Cys-DHE。模型首先采用二元权重编码对以半胱氨酸为中心的序列进行表示, 并利用距离衰减机制增强中心邻域的位置信息。随后构建CNN-BiLSTM-Attention和CNN-Transformer两个异构分支, 分别提取局部基序、上下文依赖和长程关联特征, 并通过固定权重融合两个分支的预测概率。采用五折交叉验证和独立测试集对模型进行评估, 并与传统机器学习方法及单分支模型进行比较。**结果** 实验结果表明, 在独立测试集上, Cys-DHE的准确率(Acc)、敏感性(Sn)、特异性(Sp)、马修斯相关系数(MCC)、受试者工作特征曲线下面积(AUROC)和精确率-召回率曲线下面积(AUPR)分别达到89.60%、69.88%、91.57%、50.88%、86.54%和48.52%。与单分支模型相比, 固定权重集成进一步改善了整体判别效果, 表明两个异构分支能够提供具有互补性的序列特征表示。**结论** Cys-DHE能够有效融合不同尺度的序列信息, 提高半胱氨酸S-羧乙基化位点的预测能力, 可为候选位点筛选提供一种有效的计算方法。为便于研究使用, 在线预测服务器<http://www.lzzzlab.top/Cys-DHE/>已开放访问。

关键词 半胱氨酸S-羧乙基化, 深度学习, 异构集成, 二元权重编码

中图分类号 Q754, Q332

DOI: 10.3724/j.pibb.2026.0446

CSTR: 32369.14.pibb.20260446

蛋白质翻译后修饰(post-translational modification, PTM)是细胞调控蛋白质功能的重要方式之一。在蛋白质翻译完成后, 特定氨基酸残基可通过化学基团的引入、去除或转换发生结构改变, 从而影响蛋白质的空间构象、稳定性、定位方式以及分子间相互作用等多种生物学过程^[1-2]。由于PTM能够扩展蛋白质结构和功能的多样性, 被认为是连接基因表达信息与细胞功能状态的重要调控层面。半胱氨酸(cysteine, Cys)的侧链含有具有较高反应活性的巯基结构, 在多种蛋白质调控过程中发挥关键作用。该巯基不仅参与氧化还原反应和二硫键形成, 同时也是多种化学修饰发生的重要反应位点^[3]。已有研究发现, 半胱氨酸相关修饰广泛参与维持细胞氧化还原平衡、调控信号传递过程以及影响酶催化活性等生命活动^[4-6]。因此, 系统解析半胱氨酸修饰位点及其调控机制, 对于揭示蛋白质功能变化和疾病发生过程具有重要意义。

半胱氨酸S-羧乙基化(cysteine S-carboxyethylation)是近年来受到关注的新型半胱氨酸翻译后修饰, 其揭示了代谢异常与免疫调控之间潜在的分子联系。2023年, Zhai等^[7]通过蛋白质组学分析结合免疫学实验首次报道了半胱氨酸S-羧乙基化现象, 并发现代谢产物3-羟基丙酸能够作为羧乙基基团来源参与该修饰过程。研究表明, 整合素 α IIb中的Cys96位点发生羧乙基化后, 可影响相关蛋白稳定性, 并形成具有免疫识别能力的修饰肽段。这些异常修饰肽经过人类白细胞抗原(human leukocyte antigen, HLA)递呈后, 可激活自身反应性免疫反应, 并与强直性脊柱炎等自身免

* 江西省教育厅科学技术研究项目(GJJ2402711, GJJ2402712, GJJ2402708, GJJ2502709)资助。

** 通讯联系人。

Tel: 0791-87305677, E-mail: jci_wx@163.com

收稿日期: 2026-08-14, 接受日期: 2026-10-12

疫疾病相关。同年, Didonna^[8]进一步总结了半胱氨酸羧乙基化在自身免疫性关节炎中的潜在作用, 指出该修饰可能通过改变蛋白质性质、促进异常抗原产生以及影响免疫耐受参与疾病进展。随后, Zhai等^[9]在自身免疫疾病相关新抗原研究中进一步强调了代谢物驱动型蛋白质修饰的重要作用, 说明S-羧乙基化可能是连接代谢状态变化、蛋白质功能调节和免疫异常的重要桥梁。上述研究表明, S-羧乙基化具有明确的生物学意义, 同时也为探索代谢相关疾病机制提供了新的研究方向。

目前, S-羧乙基化位点的鉴定仍主要依赖实验技术, 其中质谱分析是研究蛋白质修饰的重要手段之一。高分辨率质谱技术能够在蛋白质组范围内检测修饰肽段, 并用于定位具体修饰残基, 为PTM研究提供了关键技术支持^[10]。然而, 质谱实验通常需要经历蛋白提取、酶解、修饰肽富集、色谱分离以及谱图解析等多个步骤, 实验流程复杂且对样本质量要求较高^[11]。此外, 对于低丰度修饰、复杂生物样本以及修饰位点精确定位等问题, 质谱技术仍面临一定限制^[12]。由于S-羧乙基化发现时间较短, 目前可用于研究的数据规模相对有限, 大范围实验筛查所有潜在修饰位点不仅需要较高成本, 也会受到实验条件限制。因此, 建立快速可靠的计算预测方法, 对候选S-羧乙基化位点进行优先筛选, 能够有效降低实验验证压力, 并促进该类新型修饰的功能研究。

计算方法的发展为蛋白质修饰位点研究提供了另一条技术路径。与逐位点实验鉴定相比, 计算模型能够利用已有序列和标注信息, 对潜在修饰位点进行快速筛选。He等^[13]对蛋白质翻译后修饰位点预测的研究进展进行了系统梳理, 指出数据集构建、序列表示、分类器设计和评价策略是影响模型性能的关键环节。Ramazi等^[14]进一步总结了PTM相关数据库、预测工具及常见特征表示方法, 说明基于序列信息的计算预测已成为实验研究的重要补充。早期研究通常依赖人工构建的序列描述符, 将蛋白质序列转换为数值特征后再进行分类。Chen等^[15]开发的iFeature集成了氨基酸组成、理化性质和进化信息等多类特征, 为蛋白质和肽段序列分析提供了较为统一的特征提取框架。随后, iLearn和iLearnPlus进一步将特征工程、机器学习建模及可视化分析整合到同一平台中^[16-17]。

传统机器学习方法在多类PTM预测任务中已经得到应用。Gnad等^[18]利用支持向量机完成赖氨酸

乙酰化位点预测, 说明经过合理设计的序列特征可以支持修饰位点分类。Ning等^[19]在赖氨酸琥珀酰化预测中进一步结合多种序列特征和集成SVM, 使模型获得更好的识别效果。此类方法计算流程较清晰, 但最终性能往往受到输入特征质量的直接影响。若修饰位点附近存在复杂的非线性关系或较长范围的残基依赖, 仅依靠预先设定的人工描述符可能难以充分保留原始序列中的有效信息。

随着深度学习方法进入蛋白质序列分析领域, 修饰位点预测逐渐由人工特征设计转向数据驱动的特征学习。Wang等^[20]提出MusiteDeep, 利用卷积神经网络直接从蛋白质局部序列中提取与修饰相关的模式, 展示了深度模型在PTM预测中的应用潜力。Meng等^[21]的综述表明, CNN、循环神经网络和注意力机制已广泛用于不同类型的翻译后修饰位点预测。LSTM通过门控结构对序列状态进行选择保留和更新, 能够缓解普通循环网络在长距离信息传递中的不足^[22]。在此基础上, 双向循环网络可同时整合序列前后两个方向的信息, 更适合处理以中心残基为核心的局部窗口^[23]。已有研究也表明, CNN与BiLSTM、注意力机制联合使用时, 可以兼顾局部模式提取、上下文建模和关键位置强化^[24]。

基于自注意力机制的模型进一步拓展了蛋白质序列表示的范围。Rives等^[25]通过对大规模蛋白质序列进行自监督学习发现, 序列模型能够学习到与蛋白质结构和功能相关的潜在表示。Elnaggar等^[26]提出ProtTrans后, 又从更大规模的蛋白质序列中验证了注意力模型在蛋白质表示学习中的有效性。与循环网络按顺序传递状态不同, 自注意力机制可以直接建立不同位置之间的联系, 因此在长程依赖建模方面具有自身特点。在半胱氨酸修饰研究中, Do等^[27]针对S-亚磺酰化位点构建深度预测模型, Ning等^[28]则开发GPS-Palm用于S-棕榈酰化位点识别。虽然这些工作针对的修饰类型与S-羧乙基化不同, 但其模型设计说明, 局部序列模式和更高层次上下文信息对于半胱氨酸修饰位点识别具有较高参考价值。需要强调的是, 这些工作研究的并非S-羧乙基化, 而是为新型半胱氨酸修饰预测提供可借鉴的计算建模思路。

针对半胱氨酸S-羧乙基化的专用计算预测研究目前仍然较少, DLBWE-Cys是该方向的重要代表性工作。2025年, Luo等^[29]提出DLBWE-Cys, 用于从蛋白质序列中识别S-羧乙基化位点。该模型以

中心半胱氨酸两侧的41个氨基酸残基作为序列窗口, 提出二元权重编码 (binary-weight encoding, BWE), 在传统二元编码基础上加入距离中心位点的指数衰减权重, 从而使靠近中心Cys的残基获得较高权值。随后, 模型通过CNN提取局部特征, 利用BiLSTM学习双向上下文信息, 并通过注意力机制强化关键特征。其结果表明, 在相同网络条件下, 引入位置权重后的BWE整体优于不加位置权重的二元编码, 并且在比较的多种编码方式中表现较好。该工作为S-羧乙基化位点预测建立了重要路线, 也证明了中心位置先验信息在该任务中的有效性。

不过, 现有方法仍有进一步优化空间。一方面, 不同编码方式对中心位点信息的利用程度并不相同。独热编码能够直接表示氨基酸类别, 但不能区分残基与中心Cys之间的相对距离。PSSM可以反映进化保守性, 但其数值来源于同源序列比对, 并不是围绕修饰中心构建的位置权重。因此, 在定长中心化序列中进一步突出不同位置的重要性, 仍具有一定研究价值。另一方面, 不同深度网络擅长处理的信息类型并不完全相同。CNN更适合提取局部连续模式, BiLSTM侧重顺序上下文, 而Transformer可以直接建模跨位置关系。若全部信息仅沿单一网络路径传递, 可能难以充分利用这些结构之间的差异。异构集成为解决这一问题提供了可行思路。2022年, Basith等^[30]在蛋白质修饰位点预测中通过堆叠方式整合多类分类器, 验证了不同模型联合决策的可行性。

基于上述考虑, 本文构建双分支异构深度集成模型Cys-DHE。模型首先采用BWE表示以中心半胱氨酸为核心的序列窗口, 通过距离衰减权重增强邻近残基的特征贡献。随后设置两条相互独立的深度学习路径。CNN-BiLSTM-Attention分支用于提取局部序列模式并整合双向上下文信息, 注意力机制进一步强化具有较高判别价值的位置。CNN-Transformer分支则在卷积特征基础上加入位置编码和多头自注意力, 用于描述不同残基之间的全局联系。两个分支分别形成具有差异性的预测结果, 最终通过固定权重完成概率融合。通过这种设计, 模型能够同时利用局部、上下文和长程关联信息, 从而提高S-羧乙基化位点的综合识别能力。本研究使用的数据集和源代码可从 <https://github.com/weixin7112/Cys-DHE> 获取。

1 材料与方法

1.1 基准数据集

本研究所采用的半胱氨酸S-羧乙基化位点基准数据集来源于Luo等^[29]公开发表的实验数据。正样本来源于已报道的S-羧乙基化位点, 共包含960条相关序列。负样本则从NCBI数据库中获取, 并以未发生该修饰的半胱氨酸位点为中心构建序列片段。随后采用CD-HIT-SET^[31]进行冗余控制, 去除了45%的重复序列以降低高度相似序列对模型训练的影响。本文采用以中心Cys为核心、两侧各20个残基组成的 $L=41$ 序列窗口。该窗口长度的选择参考Luo等^[29]对21、31、41、51和61不同序列长度的比较结果, 其中序列长度41条件下模型性能最佳, 因此本文沿用该长度作为统一输入长度, 以兼顾中心位点邻域信息的覆盖范围和模型预测性能。

经处理后, 训练集共包含1 204个样本, 其中正、负样本各602个, 用于五折交叉验证。同时构建包含913个样本的独立测试集, 其中正样本83个、负样本830个, 以评估模型在类别不平衡场景下的泛化能力。上述基准数据集的详细统计信息见表1。

Table 1 The Specific details of the benchmark dataset

Type	Positive	Negative
Training	602	602
Testing	83	830

1.2 提取特征方法

为表征半胱氨酸S-羧乙基化位点周围的序列组成及位置信息, 本文采用二元权重编码 (binary-weight encoding, BWE) 对长度为41的中心化肽段进行特征提取。该方法首先将20种标准氨基酸转换为20维二元向量, 再根据各残基与中心半胱氨酸之间的距离赋予不同的位置权重, 从而突出中心位点邻近区域的序列信息。

对于序列中第 i 个氨基酸, 其二元编码表示为:

$$B_i = (b_{i1}, b_{i2}, \dots, b_{i20}) \quad (1)$$

其中, 与当前氨基酸类型对应的维度取1, 其余维度均取0。随后, 根据残基与中心半胱氨酸之间的距离计算位置权重:

$$W_i = p \exp(-\alpha |d_i|) \quad (2)$$

式中, W_i 表示第 i 个位置的权重, p 为初始权重, a 为距离衰减系数, d_i 表示该残基与中心 Cys 之间的距离。本文取 $a=0.02$ 。中心 Cys 位于序列第 21 位, 其距离为 0, 因此权重最大。随着残基与中心位置距离增加, 其权重逐渐降低。最后, 将二元编码与位置权重相乘, 得到每个残基的最终 BWE 特征:

$$X_i = B_i \times W_i \quad (3)$$

将 41 个位置的特征按照序列顺序组合后, 每条样本最终形成一个 41×20 的特征矩阵, 并作为后续双分支深度学习模型的统一输入。该方法既保留了氨基酸类别信息, 又显式引入了残基与中心修饰位点之间的位置关系, 有助于模型重点学习与 S-羧乙基化相关的邻域序列特征。

1.3 模型构建

本研究提出一种双分支异构深度集成模型 Cys-DHE, 用于半胱氨酸 S-羧乙基化位点预测。模型以 BWE 特征矩阵作为统一输入, 通过两个结构不同但功能互补的深度分支分别提取序列特征, 并在决策层进行概率融合。整体框架兼顾局部序列模式、双向上下文关系和长程依赖信息, 从而增强模型对不同尺度特征的学习能力。Cys-DHE 主要由以下 3 个部分组成。

a. BWE 特征编码。对每条以半胱氨酸为中心长度 $L=41$ 的序列进行 BWE 编码, 将 20 种标准氨基酸转换为 20 维二元向量, 并依据残基与中心 Cys 之间的距离赋予不同的位置权重。经过编码后, 每条序列被表示为 41×20 的特征矩阵。该表示既保留了氨基酸类别信息, 又强化了中心位点邻近区域的特征贡献, 为后续深度特征提取提供统一输入。

b. 双分支异构深度特征提取。第一分支采用 CNN-BiLSTM-Attention 结构。首先利用卷积层提取局部序列模式, 再通过 BiLSTM 建模中心位点两侧的双向上下文关系, 随后利用注意力机制突出关键位置特征, 并获得 128 维上下文表示。第二分支采用 CNN-Transformer 结构。CNN 首先提取局部特征, 经维度映射和位置编码后输入 Transformer 编码器, 通过多头自注意力机制学习不同序列位置之间的长程依赖关系, 最后利用均值池化和全连接层得到分类特征。两个分支分别侧重于局部时序信息和局部全局信息的学习, 从不同角度读取 S-羧乙基化位点的序列特征。

c. 固定权重概率集成。两个深度分支分别输出样本属于 S-羧乙基化位点的预测概率。为充分利用异构网络之间的互补信息, 本文在决策层采用固定

权重进行融合。其中 CNN-BiLSTM-Attention 分支权重设为 0.75, CNN-Transformer 分支权重设为 0.25。该策略在保留主分支稳定预测能力的同时, 引入 Transformer 分支的全局依赖信息, 以降低单一模型结构带来的预测偏差。

Cys-DHE 的整体流程如图 1 所示。模型首先完成中心化序列的 BWE 特征表示, 随后分别通过 CNN-BiLSTM-Attention 和 CNN-Transformer 两个分支学习不同层次的序列特征, 最终在概率层完成异构集成。该设计通过双分支协同建模, 同时利用局部基序、双向上下文和长程依赖信息, 为半胱氨酸 S-羧乙基化位点的准确识别提供基础。

1.3.1 CNN+BiLSTM+Attention 模块

为增强模型对 S-羧乙基化位点局部序列模式和上下文依赖关系的联合表征能力, Cys-DHE 构建了由卷积神经网络、双向长短期记忆网络和注意力机制组成的特征提取分支。该分支以 BWE 生成的 41×20 特征矩阵作为输入, 通过层级化特征学习逐步提取与修饰位点判别相关的信息, 具体流程如图 1c 所示。

首先, 利用两层一维卷积对输入特征进行局部模式提取。第一层卷积将输入通道由 20 映射至 64, 第二层进一步扩展至 128, 两层卷积核大小均设为 4。随后通过池化操作压缩特征长度, 减少冗余信息并保留具有较强判别能力的局部响应。

卷积特征随后输入 BiLSTM, 以进一步建模序列两个方向的上下文依赖。BiLSTM 隐藏单元数设置为 64, 正向与反向隐藏状态拼接后形成 128 维特征表示。该结构能够同时利用中心 Cys 两侧的序列信息, 从而弥补单向序列建模对上下文利用不足的问题。

在 BiLSTM 输出的基础上, 引入注意力机制对不同位置的隐藏特征进行自适应加权, 使模型能够突出与 S-羧乙基化判别关系更密切的序列区域。加权后的特征汇聚为 128 维上下文向量, 并进一步输入全连接分类器得到该分支的预测概率。

通过上述过程, 该分支形成由局部特征提取、双向上下文建模和关键位置强化组成的层级表征路径, 为后续与 Transformer 分支进行异构集成提供互补的预测信息。

1.3.2 CNN + Transformer 模块

为进一步弥补局部卷积和循环结构对长程依赖关系建模能力的不足, Cys-DHE 引入 CNN-Transformer 分支, 从全局关联角度学习 S-羧乙基

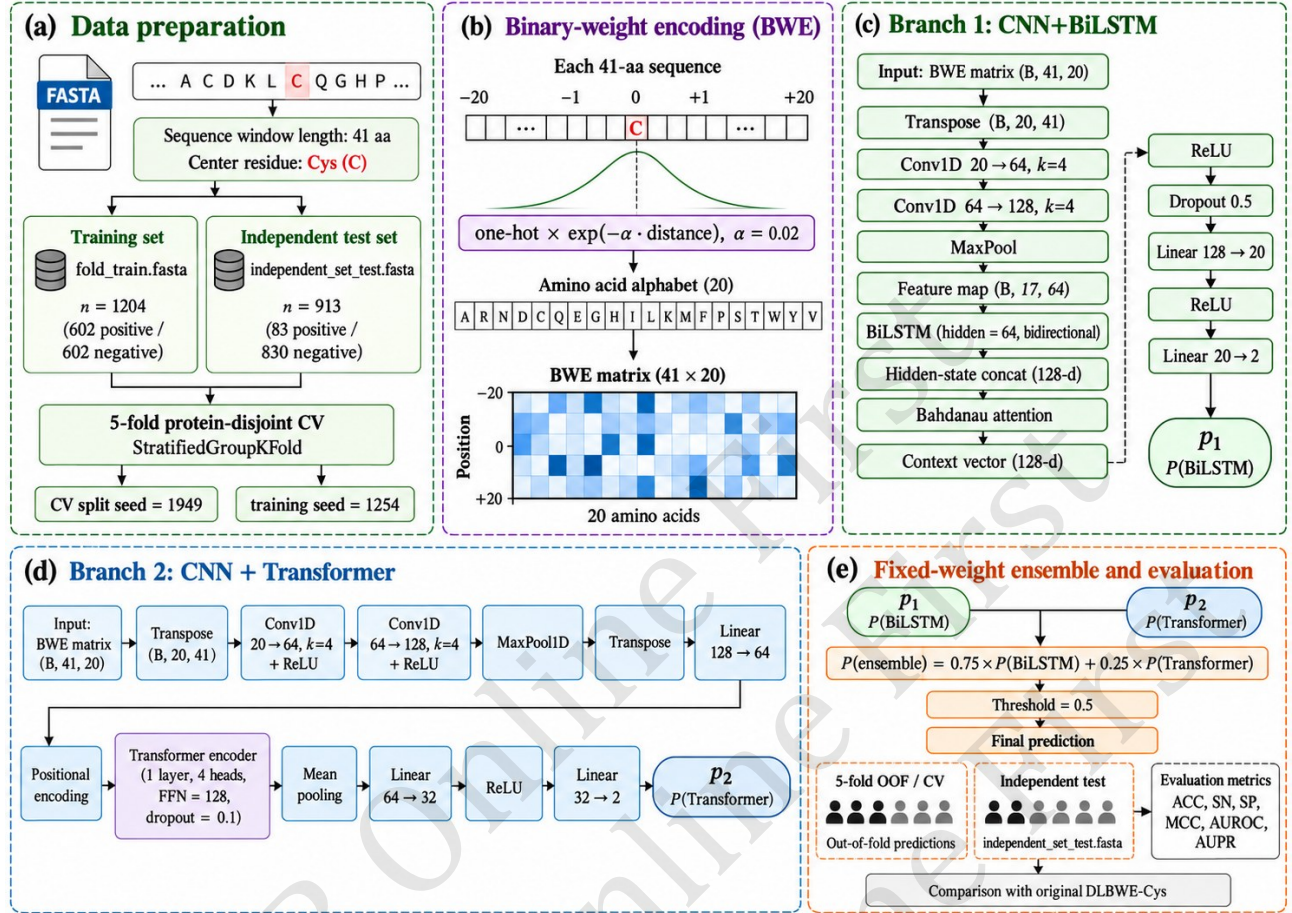


Fig. 1 Flowchart of the Cys-DHE model

(a) Data preparation. (b) Binary-weight encoding (BWE). (c) CNN-BiLSTM-Attention branch. (d) CNN-Transformer branch. (e) Fixed-weight ensemble and performance evaluation.

化位点周围的序列特征。该分支同样以 41×20 的 BWE 特征矩阵作为输入, 通过卷积特征提取与自注意力建模相结合, 获得与第一分支不同的特征表示, 具体流程如图 1d 所示。

输入特征首先经过两层一维卷积。第一层卷积将通道数由 20 扩展至 64, 第二层进一步扩展至 128, 两层卷积核大小均设置为 4。卷积操作后采用最大池化压缩特征长度, 并通过线性映射将特征维度转换为 64 维。

由于 Transformer 本身不具备序列位置信息建模能力, 因此在经过卷积模块后得到的特征中加入位置编码, 再将其输入 Transformer 编码器^[32]。本文采用 1 层 Transformer 编码器, 设置 4 个注意力头, 前馈网络维度为 128。多头自注意力机制能够从不同表示子空间建立序列位置之间的关联, 使模型在局部特征基础上进一步学习较远残基之间的依赖关系。

经过 Transformer 编码后, 对各位置特征进行均值池化, 得到 64 维全局特征向量。随后通过全连接层将其映射至 32 维, 并输出该分支的分类概率。与 CNN-BiLSTM-Attention 分支相比, 该分支更加侧重于跨位置特征交互和全局依赖信息的提取, 从而为两个异构分支之间的概率融合提供差异化的特征基础。

1.3.3 固定权重集成

为充分利用两个异构分支在特征学习方式上的差异, Cys-DHE 采用固定权重对两个分支输出的预测概率进行融合。具体流程如图 1e 所示。设 CNN-BiLSTM-Attention 分支输出的正类概率为 P_1 , CNN-Transformer 分支输出的正类概率为 P_2 , 则最终集成概率表示为:

$$P = 0.75 \times P_1 + 0.25 \times P_2 \quad (4)$$

其中, CNN-BiLSTM-Attention 分支被赋予较高权重, 以保留其在局部模式和上下文依赖建模方面的

稳定性,同时引入CNN-Transformer分支对长程关联信息的补充。最终以0.5作为分类阈值,当 $P \geq 0.5$ 时预测为S-羧乙基化位点,否则预测为非修饰位点。

1.3.4 模型训练与参数设置

为保证模型训练过程的一致性和可重复性,本文统一设置主要训练参数,具体如表2所示。两个分支均采用Adam优化器,初始学习率设为0.0005, batch size为64,最大训练轮数为100,并采用早停策略以降低过拟合风险。损失函数采用CrossEntropyLoss。

Table 2 The key hyperparameter settings.

Parameter	Setting
Optimizer	Adam
Learning rate	0.0005
Batch size	64
Max epochs	100
Early stopping patience	15
Loss function	CrossEntropyLoss

1.4 评价指标

为全面评价Cys-DHE对半胱氨酸S-羧乙基化位点的预测性能,本文采用准确率(accuracy, Acc)、敏感性(sensitivity, Sn)、特异性(specificity, Sp)和马修斯相关系数(Matthews correlation coefficient, MCC)作为评价指标。 MCC 能够综合考虑真正例、真负例、假正例和假负例,在类别分布不均衡的情况下仍具有较好的评价能力^[33]。具体计算如下:

$$Sn = \frac{TP}{TP + FN} \quad (5)$$

$$Sp = \frac{TN}{TN + FP} \quad (6)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$MCC =$$

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

其中, TP 表示正确预测为S-羧乙基化位点的样本数量, TN 表示正确预测为非修饰位点的样本数量, FP 表示错误预测为修饰位点的样本数量, FN 表示错误预测为非修饰位点的样本数量。

除上述指标外,本文进一步采用AUROC和AUPR评价模型在不同分类阈值下的整体判别能

力。AUROC表示受试者工作特征曲线(receiver operating characteristic curve)下面积,该曲线由真正率(TPR)和假阳性率(FPR)共同构成,可反映模型区分正负样本的整体能力。AUPR表示精确率-召回率曲线下面积,其中精确率($Precision$)和召回率($Recall$)分别描述模型预测结果的可靠性以及对阳性样本的覆盖能力^[34]。相比AUROC, AUPR对正负样本比例不均衡的数据集更加敏感,因此在蛋白质修饰位点预测任务中具有较高参考价值^[35]。

2 结果

2.1 不同序列编码方法的比较

评估输入表示方式对Cys-DHE预测结果的影响,本文保持数据划分及模型参数一致,分别采用BLOSUM62、ESM-2、1mer、2mer、3mer、1+2+3mer和BWE进行实验。BLOSUM62根据氨基酸替换矩阵将每个残基表示为20维向量,因此长度为41的序列形成 41×20 特征矩阵^[36]。ESM-2采用预训练模型esm2_t30_150M_UR50D对序列进行编码,每个残基原始表示维度为640,随后通过主成分分析(PCA)降至20维,最终得到 41×20 的输入矩阵^[37]。对于k-mer编码,1mer、2mer和3mer分别描述不同尺度的局部残基组合,并采用位置对齐的20维固定表示,因此均形成 41×20 特征矩阵^[38]。而1+2+3mer则对三种尺度的20维表示进行融合,最终同样得到 41×20 特征矩阵。BWE将20维二元编码与基于中心Cys距离的衰减权重相乘,所得输入维度同样为 41×20 。通过统一输入维度,可以在不改变后续网络结构及参数数量的条件下比较不同编码方式的表征能力。各编码方式在相同条件下的比较结果见表3。

不同编码方式对模型性能具有明显影响。BWE在 Sn 、 Sp 、 Acc 、 MCC 、AUROC和AUPR上

Table 3 Performance comparison of different position encoding methods on the train dataset

Methods	Sn	Sp	Acc	MCC	AUROC	AUPR
BLOSUM62	0.6412	0.8355	0.7384	0.4860	0.7997	0.8131
ESM-2	0.6262	0.7658	0.6960	0.3959	0.7470	0.7685
1mer	0.6379	0.7890	0.7135	0.4319	0.7607	0.7819
2mer	0.6445	0.6146	0.6296	0.2593	0.6624	0.6195
3mer	0.4967	0.5083	0.5025	0.0050	0.4913	0.5100
1+2+3mer	0.6595	0.4967	0.5781	0.1583	0.5951	0.5850
BWE	0.6803	0.8427	0.7519	0.5262	0.8257	0.8560

分别达到 0.680 3、0.842 7、0.751 9、0.526 2、0.825 7 和 0.856 0，整体表现最佳。BLOSUM62 的性能次之，其中 *AUROC* 为 0.799 7。ESM-2 和 1mer 仍具有一定判别能力，而 2mer、3mer 及 1+2+3mer 的整体性能明显下降，尤其 3mer 的 *MCC* 仅为 0.005 0，*AUROC* 为 0.491 3，接近随机分类水平。

结果表明，单纯依赖残基组成或高维上下文信息并不能充分描述 S-羧乙基化位点的局部位置特征。相比之下，BWE 通过引入残基与中心 Cys 之间的距离关系，能够更有效地突出中心邻域信息，因此本文最终选用 BWE 作为 Cys-DHE 的输入编码

方式。

2.2 不同固定融合权重的比较

为进一步说明双分支固定融合权重的确定依据，本文基于训练集五折交叉验证结果，对 CNN-BiLSTM-Attention 和 CNN-Transformer 两个分支的不同权重组合进行了比较。设 CNN-BiLSTM-Attention 分支的融合权重为 w_1 ，CNN-Transformer 分支的融合权重为 w_2 ，并满足 $w_1+w_2=1$ 。除两个单分支模型外，本文进一步比较了 w_1/w_2 分别为 0.25/0.75、0.50/0.50 和 0.75/0.25 的三种固定融合方案，结果如表 4 所示。

Table 4 Performance comparison of single-branch models and fixed-weight ensembles in five-fold cross-validation								
Methods	w_1	w_2	Sn	Sp	Acc	MCC	$AUROC$	$AUPR$
CNN-BiLSTM-Attention	1.00	0.00	0.675 6	0.846 2	0.751 0	0.526 5	0.825 3	0.855 0
CNN-Transformer	0.00	1.00	0.670 6	0.833 8	0.745 1	0.505 1	0.809 0	0.840 8
Fixed ensemble 1	0.25	0.75	0.672 8	0.831 0	0.745 5	0.511 8	0.817 7	0.852 0
Fixed ensemble 2	0.50	0.50	0.678 0	0.832 0	0.746 5	0.515 7	0.818 9	0.852 6
Cys-DHE	0.75	0.25	0.680 3	0.842 7	0.751 9	0.526 2	0.825 7	0.856 0

从结果可以看出，两个单分支模型具有不同的性能特点。CNN-BiLSTM-Attention 在 Sp 和 MCC 上分别达到 0.846 2 和 0.526 5，整体表现优于 CNN-Transformer，说明其在当前任务中具有较稳定的判别能力。CNN-Transformer 虽然单独预测性能相对较低，但能够从长程依赖角度提供不同于 BiLSTM-Attention 分支的预测信息。

在三种固定融合方案中，当 $w_1=0.75$ 、 $w_2=0.25$ 时，模型取得 $Sn=0.680 3$ 、 $Acc=0.751 9$ 、 $MCC=0.526 2$ 、 $AUROC=0.825 7$ 和 $AUPR=0.856 0$ 。其中， Sn 、 Acc 、 $AUROC$ 和 $AUPR$ 均高于另外两种融合比例， MCC 也由 0.511 8 和 0.515 7 提高至 0.526 2。相比之下，当 Transformer 分支权重提高至 0.75 时，整体性能有所下降，说明在当前数据条件下 CNN-BiLSTM-Attention 分支应承担主要预测作用，而 CNN-Transformer 分支更适合作为补充信息来源。

因此，综合五折交叉验证中的 MCC 、 $AUROC$ 、 $AUPR$ 及分类性能，本文最终将 CNN-BiLSTM-Attention 和 CNN-Transformer 的固定融合权重分别设为 0.75 和 0.25。该设置在保留主分支稳定预测能力的同时，引入 Transformer 分支的长程依赖信息，从而在多个评价指标之间获得较为均衡的性能。

2.3 与常见机器学习算法对比

为进一步评价 Cys-DHE 的预测性能，本文在

五折交叉验证设置下，将其与支持向量机 (support vector machine, SVM)^[39]、随机森林 (random forest, RF)^[40]、极端梯度提升 (eXtreme Gradient Boosting, XGBoost)^[41] 和轻量梯度提升机 (Light Gradient Boosting Machine, LightGBM)^[42] 进行比较。SVM 通过构建最大间隔分类超平面实现样本划分，RF 通过集成多棵随机决策树提高分类稳定性，XGBoost 和 LightGBM 则属于基于梯度提升决策树的集成学习方法。所有方法均采用相同的数据划分和评价指标，以保证结果具有可比性。结果如表 5 所示。

Table 5 Comparison with common machine learning performance on 5-fold cross validation						
Methods	Sn	Sp	Acc	MCC	$AUROC$	$AUPR$
SVM	0.651 2	0.662 8	0.657 0	0.314 0	0.709 1	0.729 2
RF	0.669 4	0.596 3	0.632 9	0.266 5	0.688 2	0.707 3
XGBoost	0.629 6	0.642 9	0.636 2	0.272 4	0.681 0	0.695 4
LightGBM	0.611 3	0.604 7	0.608 0	0.216 0	0.657 2	0.660 1
Cys-DHE	0.680 3	0.842 7	0.751 9	0.526 2	0.825 7	0.856 0

Cys-DHE 在 Sn 、 Sp 、 Acc 、 MCC 、 $AUROC$ 和 $AUPR$ 六项评价指标上均高于所比较的传统机器学习方法。相比之下，4 种传统机器学习方法中 SVM 整体表现较好，但其 Acc 和 MCC 分别仅为 0.657 0 和 0.314 0，与 Cys-DHE 仍存在较明显差距。RF 虽

然取得0.669 4的 Sn ，但 Sp 仅为0.596 3，说明其对阴性样本的识别能力相对有限。XGBoost 和 LightGBM 的整体性能同样低于Cys-DHE。

综合来看，传统机器学习方法依赖预先构建的固定特征进行分类，对复杂序列模式及不同位置之间依赖关系的刻画能力相对有限。Cys-DHE通过双分支深度网络对局部模式、上下文信息和长程依赖进行联合建模，因此在整体准确性、分类相关性以及曲线评价指标上均表现出更好的预测能力。

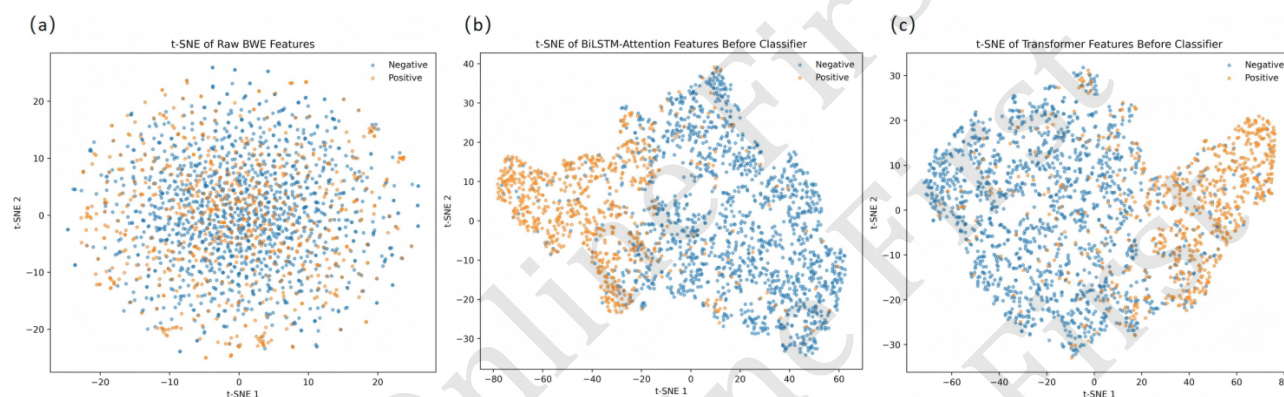


Fig. 2 Comparison of t-SNE visualizations between raw features and deep features extracted by different networks

(a) t-SNE projection of the raw BWE features. (b) t-SNE projection of the features extracted by the CNN-BiLSTM-Attention branch before the final classifier. (c) t-SNE projection of the features extracted by the CNN-Transformer branch before the final classifier.

经过CNN-BiLSTM-Attention分支学习后，正负样本的空间分布发生明显变化。阳性样本主要向左侧区域聚集，阴性样本则更多分布于中部和右侧区域，两类样本之间形成较为清晰的分离趋势。这说明该分支能够通过卷积、双向上下文建模和注意力加权进一步提取具有判别性的局部及上下文特征。

CNN-Transformer分支同样表现出较强的特征重构能力。与原始BWE特征相比，阳性样本更多集中于右侧区域，阴性样本主要分布在左侧和中部，两类样本的类间距离明显增大。该结果表明，Transformer结构能够在卷积局部特征的基础上进一步捕获不同位置之间的全局关联信息。

综合来看，两种网络均能显著改善原始BWE特征的可分性，但其形成的特征分布存在明显差异。这种差异说明CNN-BiLSTM-Attention和CNN-Transformer在特征学习机制上具有一定互补性，也为后续采用双分支异构集成提供了直观依据。

2.4 不同网络结构表征学习能力的验证

为进一步比较不同网络结构对序列特征的学习能力，本文采用t-SNE对原始BWE特征、CNN-BiLSTM-Attention分支特征和CNN-Transformer分支特征进行二维可视化。结果如图2所示，原始BWE特征中的阳性样本与阴性样本高度混杂，两类样本之间缺乏明显边界，说明初始编码本身的判别能力较为有限。

2.5 Cys-DHE与现有S-羧乙基化位点预测模型的比较

为进一步验证Cys-DHE在实际预测场景中的性能，本文将其与现有半胱氨酸S-羧乙基化位点预测模型DLBWE-Cys^[29]在独立测试集上进行比较（表6）。Cys-DHE在 Sp 、 Acc 和 MCC 上均优于DLBWE-Cys。其中， Sp 由0.853 0提高至0.915 7， Acc 由0.844 5提高至0.896 0， MCC 由0.437 7提高至0.508 8，这说明Cys-DHE在降低假阳性预测、提升整体分类准确性以及增强正负样本综合判别能力方面具有更明显的优势。

与此同时，为进一步直观评价Cys-DHE在不同分类阈值下的判别性能，本文绘制了独立测试集上的PR曲线和ROC曲线，如图3所示。Cys-DHE的AUROC达到0.865 4，ROC曲线整体明显高于随机分类基准线，表明模型具有较好的正负样本区分能力。考虑到独立测试集中正、负样本数量分别为83和830，类别分布明显不均衡，因此PR曲线能够更直接反映模型对少数类阳性位点的识别效果。Cys-DHE的AUPR为0.485 2，明显高于由阳性样本

比例决定的随机基线 0.090 9, 说明模型在类别不平衡条件下仍具有较好的阳性样本排序和识别能力。

Table 6 Comparison with Other Existing Advanced Models on the Test Dataset

Predictor model	<i>Sn</i>	<i>Sp</i>	<i>Acc</i>	<i>MCC</i>	<i>AUROC</i>	<i>AUPR</i>
DLBWE-Cys ^[29]	0.759 0	0.853 0	0.844 5	0.437 7	0.871 2	0.490 9
Cys-DHE	0.698 8	0.915 7	0.896 0	0.508 8	0.865 4	0.485 2

从独立测试集的样本构成来看, 阴性样本数量明显多于阳性样本, 因此模型对阴性位点的识别能力对实际预测结果影响较大。Cys-DHE 取得 0.915 7 的 *Sp*, 较 DLBWE-Cys 提升了 6.27%, 表明其能够更有效地排除非 S-羧乙基化位点, 减少错误阳性结果。与此同时, *MCC* 提升至 0.508 8, 说明这种性能提升并非单纯来源于类别分布偏向, 而是模型整体分类一致性得到改善。

Cys-DHE 的 *AUROC* 为 0.865 4, 与 DLBWE-Cys 的 0.871 2 较为接近, *AUPR* 分别为 0.485 2 和 0.490 9, 二者差异较小。*Sn* 方面, Cys-DHE 为 0.698 8, 低于 DLBWE-Cys 的 0.759 0。这表明 Cys-DHE 在提高特异性和整体准确性的同时, 对阳性

样本识别能力存在一定权衡。但综合 *Acc* 和 *MCC* 等指标来看, Cys-DHE 在当前类别不平衡测试集上取得了更加稳定的综合性能。

Cys-DHE 在部分指标上的性能改善可能与其双分支异构特征学习策略有关。CNN-BiLSTM-Attention 分支能够提取局部序列模式及双向上下文依赖, CNN-Transformer 分支进一步补充不同位置之间的长程关联信息。通过固定权重集成, 两类不同层次的序列表征在预测阶段实现互补, 从而降低单一网络结构的表征局限。结合前述编码方式、单分支比较和集成实验结果, 可以看出 Cys-DHE 在保持较强整体判别能力的基础上显著提升了独立测试集上的分类准确性和特异性, 表现出较好的泛化能力和实际应用潜力。

2.6 局部Cys组成与序列距离分析

为进一步分析中心 Cys 周围的局部序列特征, 并考察一级序列中局部 Cys 组成及其相对位置是否与 S-羧乙基化状态存在关联, 本文分别在完整 41 aa 序列窗口 (中心 Cys 及其两侧各 20 个残基)、中心 Cys 两侧 ± 10 个残基和 ± 5 个残基范围内, 统计除中心 Cys 外其他 Cys 残基的出现情况, 并比较阳性与阴性样本之间的差异 (表 7, 图 4)。

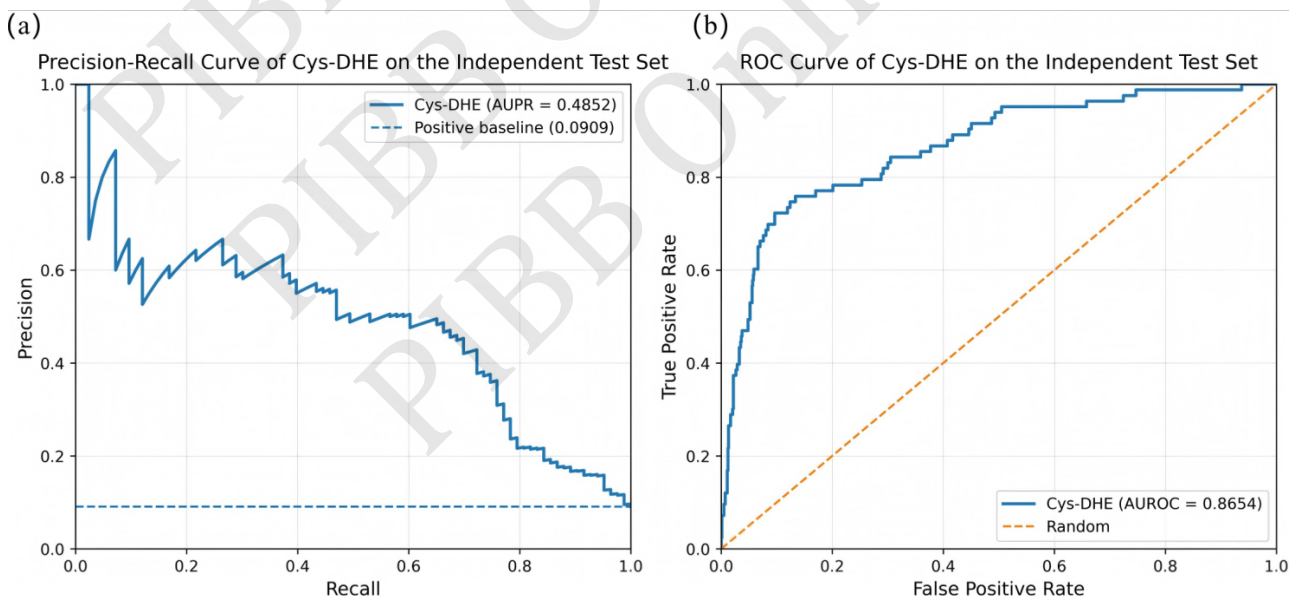


Fig. 3 Performance curves of Cys-DHE on the independent test set

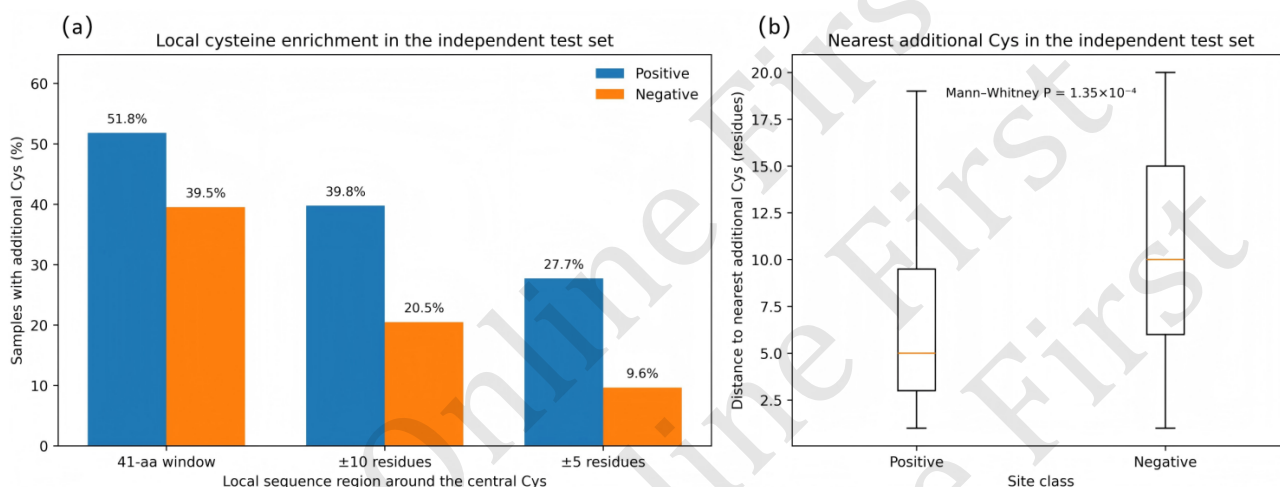
(a) Precision-recall curve. (b) Receiver operating characteristic curve.

在训练集中, 以完整 41 aa 序列窗口为分析范围时, 60.63% 的阳性样本中存在至少 1 个额外

Cys, 而阴性样本中的比例为 31.40%, 两组差异具有统计学意义 ($OR=3.365$, $P=2.10 \times 10^{-24}$)。当分

Table 7 Local cysteine enrichment around the central Cys in the training and independent test sets

Dataset	Local Cys definition	Positive with nearby Cys	Positive total	Positive rate/%	Negative with nearby Cys	Negative total	Negative rate/%	Odds ratio (positive vs negative)	Fisher exact <i>P</i>
Training	41-aa window	365	602	60.63	189	602	31.40	3.365	2.10×10^{-24}
Training	± 10 residues	259	602	43.02	82	602	13.62	4.788	1.94×10^{-30}
Training	± 5 residues	161	602	26.74	46	602	7.64	4.413	4.98×10^{-19}
Independent test	41-aa window	43	83	51.81	328	830	39.52	1.645	3.47×10^{-2}
Independent test	± 10 residues	33	83	39.76	170	830	20.48	2.562	1.50×10^{-4}
Independent test	± 5 residues	23	83	27.71	80	830	9.64	3.594	1.05×10^{-5}

**Fig. 4** Local cysteine enrichment and distance to the nearest additional Cys in the independent test set

(a) Proportion of samples containing at least one additional Cys within different local sequence regions around the central Cys. (b) Sequence distance from the central Cys to the nearest additional Cys in samples containing at least one additional Cys. The difference between positive and negative samples was evaluated using the Mann-Whitney U test.

析范围缩小至中心 Cys 两侧 ± 10 和 ± 5 个残基时, 阳性样本中存在其他 Cys 的比例分别为 43.02% 和 26.74%, 均高于阴性样本中的 13.62% 和 7.64%, 对应 *OR* 分别为 4.788 和 4.413, 差异均具有统计学意义。

相同趋势在独立测试集中仍然能够观察到。完整 41 aa 序列窗口中, 阳性样本中存在额外 Cys 的比例为 51.81%, 高于阴性样本的 39.52% (*OR*=1.645, *P*=0.034 7)。在 ± 10 个残基范围内, 阳性和阴性样本中的比例分别为 39.76% 和 20.48% (*OR*=2.562, *P*= 1.50×10^{-4})。在 ± 5 个残基范围内, 两组比例分别为 27.71% 和 9.64% (*OR*=3.594, *P*= 1.05×10^{-5})。图 4a 进一步直观显示, 在不同局部序列范围内, 阳性样本中邻近 Cys 的出现比例均高于阴性样本。

为进一步考察额外 Cys 与中心 Cys 之间的位置关系, 本文在完整 41 aa 序列窗口内存在至少 1 个

额外 Cys 的独立测试样本中, 比较最近额外 Cys 与中心 Cys 之间的一级序列距离。其中阳性组和阴性组分别包含 43 和 328 个样本。如图 4b 所示, 阳性样本中最近额外 Cys 的距离中位数约为 5 个残基, 而阴性样本约为 10 个残基, Mann - Whitney U 检验显示两组差异具有统计学意义 (*P*= 1.35×10^{-4})。该结果表明, 与阴性样本相比, S-羧乙基化阳性位点周围的额外 Cys 不仅出现频率更高, 而且在一级序列上更倾向于分布在靠近中心 Cys 的位置。

综合上述结果, 局部 Cys 组成及其相对于中心修饰位点的序列距离可能构成 S-羧乙基化相关序列微环境的一部分。由于 Cys-DHE 采用以中心 Cys 为核心的定长序列窗口, 并通过 BWE 引入位置权重, 因此模型可能利用这类局部残基组成及距离相关信息进行判别。需要强调的是, 一级序列上的邻近关系不能直接代表蛋白质三维空间中的邻近, 也不能据此判断二硫键形成。因此, 该分析应视为序列层

面的补充证据, 而非蛋白质结构或二硫键状态的直接验证。

2.7 Cys-DHE在线预测平台

为方便研究人员使用 Cys-DHE 开展半胱氨酸 S-羧乙基化位点预测, 本文搭建了配套的在线预测平台 (图 5), 其访问地址为 <http://www.lzzzlab.top/Cys-DHE/>。平台以 Cys-DHE 模型为核心, 将序列输入、特征编码、模型推理和结果展示集成到统一界面中, 用户无须配置本地运行环境即可完成预测。

平台支持蛋白质序列的在线提交。用户可按照

FASTA 格式输入待预测序列, 提交后系统自动完成序列预处理和位点识别, 并调用训练完成的 Cys-DHE 模型计算候选半胱氨酸位点的预测概率。预测完成后, 页面返回相应的位点信息及分类结果, 便于用户快速筛选潜在的 S-羧乙基化位点。

为降低模型使用门槛, 平台同时提供基本使用说明, 对输入格式、任务提交方式及结果含义进行说明。通过将模型封装为在线服务, Cys-DHE 能够由算法研究进一步扩展至实际序列分析场景, 为半胱氨酸 S-羧乙基化候选位点的快速筛选和后续实验验证提供便利。

Cys-DHE Server

Welcome to Cys-DHE Server

Cysteine S-carboxyethylation is an important protein post-translational modification, closely associated with protein functional regulation and related disease processes. This is the Cys-DHE Server, a dual-branch heterogeneous deep ensemble model for cysteine S-Carboxyethylation site prediction.

Prediction by inputting sequences:

Please enter the protein sequences in the box below.

>Example
VNNVVVRYGHSIALYKDKIYMYGGKIDSTGN

Submit Sequences

Reset Sequences

Prediction by uploading files:

Job Submission

Program name:

Please enter your project name

Email:

Please enter your email address

Protein file

Protein Sequences (FASTA format):

选择文件 未选择文件

Submit

Reset

Notice:

- Enter the sequences of the query proteins in FASTA format ([example](#)).
- The number of protein sequences submitted each time is limited to 5 when using the function "Prediction by input sequence".
- If you need to submit quantities of protein sequences at once, you can use the function "Prediction by uploading files". [More info...](#)

Xin Lab:Bioinformatics Team (Xin Lab)

Contact: jci_wx@163.com

Fig. 5 Screenshot of the Cys-DHE web server homepage

3 讨论

本研究构建的 Cys-DHE 将位置感知的 BWE 编码与两种结构不同的深度学习分支相结合, 其主要特点在于同时考虑中心 Cys 邻域的位置差异以及不同网络对序列信息的表征能力。固定权重实验表明, CNN-BiLSTM-Attention 与 CNN-Transformer 之间存在一定互补性, 但融合并未使所有评价指标同步提高。这说明异构集成的作用并不是简单叠加两个模型, 而是通过不同特征学习机制之间的互补,

在部分指标上改善整体判别能力。从模型结构上看, CNN-BiLSTM-Attention 更侧重局部模式和双向上下文关系, 而 Transformer 分支则能够补充跨位置关联信息, 这可能是融合模型获得较均衡性能的重要原因。

序列编码比较进一步表明, S-羧乙基化位点预测不仅依赖氨基酸组成, 还与残基相对于中心修饰位点的位置有关。BWE 通过距离衰减权重显式引入中心位置先验, 使靠近中心 Cys 的残基具有更高贡献, 这与位点预测任务强调局部序列环境的特点

较为一致。值得注意的是,本研究中ESM-2特征经过PCA降维后再输入统一网络,因此其当前表现受到特征压缩方式、数据规模及下游模型设置等多种因素影响,不能据此认为预训练蛋白质语言模型在该任务中缺乏价值。随着数据规模增加以及任务特异性微调方法的发展,预训练表示仍具有进一步探索的空间。

与传统机器学习方法相比,Cys-DHE表现出更好的综合判别能力,说明深度模型能够在BWE基础上进一步提取非线性序列关系。t-SNE结果中正负样本在深度特征空间中呈现出更清晰的分离趋势,也从可视化角度支持了这一点。不过,t-SNE主要用于直观展示高维特征的二维投影,其结果容易受到降维参数和投影过程影响,因此不宜将二维距离变化直接解释为分类性能提升,模型优劣仍以交叉验证和独立测试结果为主要依据。

独立测试结果显示,Cys-DHE在特异性、准确率和MCC方面表现较好,而敏感性及曲线评价指标相较DLBWE-Cys没有同步提高,反映出模型在降低假阳性和保持阳性召回之间存在一定权衡。对于S-羧乙基化候选位点筛选而言,这种特点具有一定实际意义。实际蛋白质序列中潜在未修饰Cys通常远多于真实修饰位点,较高的特异性有助于减少需要进一步验证的假阳性候选。但如果研究目的更强调尽可能发现潜在阳性位点,则仍需进一步优化分类阈值、损失函数或类别不平衡学习策略,以改善敏感性。

局部Cys分析提示,中心Cys邻域中的残基组成及其相对位置可能与S-羧乙基化状态存在一定联系。这一结果从序列层面支持了采用中心化窗口及位置加权策略进行建模的合理性,但其生物学意义需要谨慎解释。一级序列上的Cys邻近并不等同于蛋白质三维结构中的空间邻近,也不能据此判断是否形成二硫键。半胱氨酸的反应状态还可能受到溶剂可及性、局部构象、空间环境及二硫键状态等因素影响,因此当前局部Cys分析应视为序列层面的补充证据,而不能作为结构状态的直接验证。

本研究仍存在一定局限。首先,当前可用于S-羧乙基化建模的实验验证数据规模较小,虽然采用早停、五折交叉验证和独立测试降低了过拟合带来的影响,但模型的跨数据集泛化能力仍需更多独立样本验证。其次,Cys-DHE目前主要依赖一级序列信息,尚未显式整合蛋白质三维结构、进化保守性和功能背景等信息。此外,固定权重融合虽然简

单、稳定且易于复现,但无法根据不同样本动态调整两个分支的贡献。未来可进一步扩大实验数据规模,并结合残基溶剂可及性、空间邻域、二硫键注释、进化信息及预训练蛋白质表示,探索序列—结构联合建模和自适应融合策略,从而进一步提升模型的预测能力、生物学解释性和泛化性能。

4 结论

本研究提出了半胱氨酸S-羧乙基化位点预测模型Cys-DHE,通过将BWE位置权重编码与CNN-BiLSTM-Attention和CNN-Transformer两个异构深度学习分支相结合,实现了对局部序列模式、上下文依赖及跨位置关联信息的联合建模。

在独立测试集上,Cys-DHE的*Acc*、*Sn*、*Sp*、*MCC*、*AUROC*和*AUPR*分别达到0.896 0、0.698 8、0.915 7、0.508 8、0.865 4和0.485 2,表现出较好的综合预测能力。与现有DLBWE-Cys相比,模型在*Sp*、*Acc*和*MCC*方面具有优势,同时局部Cys分析进一步表明,中心Cys邻域中包含与S-羧乙基化状态相关的序列判别信息。

总体而言,Cys-DHE可作为一种基于一级序列的候选S-羧乙基化位点预筛选工具,为后续结构分析和实验验证提供参考。未来通过引入更多实验数据以及蛋白质结构、进化和预训练表示等多源信息,有望进一步提高模型的预测性能、解释性和泛化能力。

数据可用性声明 本文相关数据(DOI: 10.57760/sciencedb.014wi/CSTR: 31253.11.sciencedb.014wi)可在科学数据银行(Science Data Bank, ScienceDB)数据库中访问和获取,链接为: <https://www.scidb.cn/s/MfUVjy>。

参考文献

- [1] Walsh C T, Garneau-Tsodikova S, Gatto G J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed*, 2005, **44**(45): 7342-7372
- [2] Mann M, Jensen O N. Proteomic analysis of post-translational modifications. *Nat Biotechnol*, 2003, **21**(3): 255-261
- [3] Poole L B. The basics of thiols and cysteines in redox biology and chemistry. *Free Radic Biol Med*, 2015, **80**: 148-157
- [4] Reddie K G, Carroll K S. Expanding the functional diversity of proteins through cysteine oxidation. *Curr Opin Chem Biol*, 2008, **12**(6): 746-754
- [5] Wani R, Nagata A, Murray B W. Protein redox chemistry: post-translational cysteine modifications that regulate signal

- transduction and drug pharmacology. *Front Pharmacol*, 2014, **5**: 224
- [6] Chung H S, Wang S B, Venkatraman V, *et al*. Cysteine oxidative posttranslational modifications: emerging regulation in the cardiovascular system. *Circ Res*, 2013, **112**(2): 382-392
- [7] Zhai Y, Chen L, Zhao Q, *et al*. Cysteine carboxyethylation generates neoantigens to induce HLA-restricted autoimmunity. *Science*, 2023, **379**(6637): eabg2482
- [8] Didonna A. Cysteine carboxyethylation: a novel post-translational modification associated with autoimmune arthritis. *Sig Transduct Target Ther*, 2023, **8**(1): 274
- [9] Zhai Y, Wu K, Lin Q, *et al*. Post-translational modified neoantigens in autoimmune diseases: challenges of immune tolerance. *Adv Sci*, 2025, **12**(34): e01766
- [10] Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature*, 2016, **537**(7620): 347-355
- [11] Jensen O N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol*, 2004, **8**(1): 33-41
- [12] Na S, Paek E. Software eyes for protein post-translational modifications. *Mass Spectrom Rev*, 2015, **34**(2): 133-147
- [13] He W, Wei L, Zou Q. Research progress in protein posttranslational modification site prediction. *Brief Funct Genom*, 2019, **18**(4): 220-229
- [14] Ramazi S, Zahiri J. Posttranslational modifications in proteins: resources, tools and prediction methods. *Database*, 2021, **2021**: baab012
- [15] Chen Z, Zhao P, Li F, *et al*. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 2018, **34**(14): 2499-2502
- [16] Chen Z, Zhao P, Li F, *et al*. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform*, 2020, **21**(3): 1047-1057
- [17] Chen Z, Zhao P, Li C, *et al*. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res*, 2021, **49**(10): e60
- [18] Gnad F, Ren S, Choudhary C, *et al*. Predicting post-translational lysine acetylation using support vector machines. *Bioinformatics*, 2010, **26**(13): 1666-1668
- [19] Ning Q, Zhao X, Bao L, *et al*. Detecting Succinylation sites from protein sequences using ensemble support vector machine. *BMC Bioinformatics*, 2018, **19**(1): 237
- [20] Wang D, Zeng S, Xu C, *et al*. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 2017, **33**(24): 3909-3916
- [21] Meng L, Chan W S, Huang L, *et al*. Mini-review: recent advances in post-translational modification site prediction based on deep learning. *Comput Struct Biotechnol J*, 2022, **20**: 3522-3532
- [22] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, **9**(8): 1735-1780
- [23] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. *IEEE Trans Signal Process*, 1997, **45**(11): 2673-2681
- [24] Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 2019, **337**: 325-338
- [25] Rives A, Meier J, Sercu T, *et al*. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA*, 2021, **118**(15): e2016239118
- [26] Elnaggar A, Heinzinger M, Dallago C, *et al*. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell*, 2022, **44**(10): 7112-7127
- [27] Do D T, Le T Q T, Le N Q K. Using deep neural networks and biological subwords to detect protein S-sulfonylation sites. *Brief Bioinform*, 2021, **22**(3): bbab128
- [28] Ning W, Jiang P, Guo Y, *et al*. GPS-Palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins. *Brief Bioinform*, 2021, **22**(2): 1836-1847
- [29] Luo Z, Wang Q, Xia Y, *et al*. DLBWE-Cys: a deep-learning-based tool for identifying cysteine S-carboxyethylation sites using binary-weight encoding. *Front Genet*, 2024, **15**: 1464976
- [30] Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Brief Bioinform*, 2022, **23**: bbab376
- [31] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 2006, **22**(13): 1658-1659
- [32] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need// Guyon I, von Luxburg U, Bengio S, et al. *Advances in Neural Information Processing Systems 30*. Red Hook, NY: Curran Associates Inc., 2017: 5998-6008
- [33] Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 1975, **405**(2): 442-451
- [34] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*, 2006, **27**(8): 861-874
- [35] Davis J, Goadrich M. The relationship between precision-recall and ROC curves//Cohen W W, Moore A. *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY: Association for Computing Machinery, 2006: 233-240
- [36] Henikoff S, Henikoff J G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 1992, **89**(22): 10915-10919
- [37] Lin Z, Akin H, Rao R, *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023, **379**(6637): 1123-1130
- [38] Petrilli P. Classification of protein sequences by their dipeptide composition. *Bioinformatics*, 1993, **9**(2): 205-209
- [39] Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, **20**(3): 273-297
- [40] Breiman L. Random forests. *Mach Learn*, 2001, **45**(1): 5-32
- [41] Chen T, Guestrin C. XGBoost: a scalable tree boosting system// Krishnapuram B, Shah M, Smola A J, et al. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, 2016: 785-794
- [42] Ke G, Meng Q, Finley T, *et al*. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*, 2017, **30**: 3146-3154

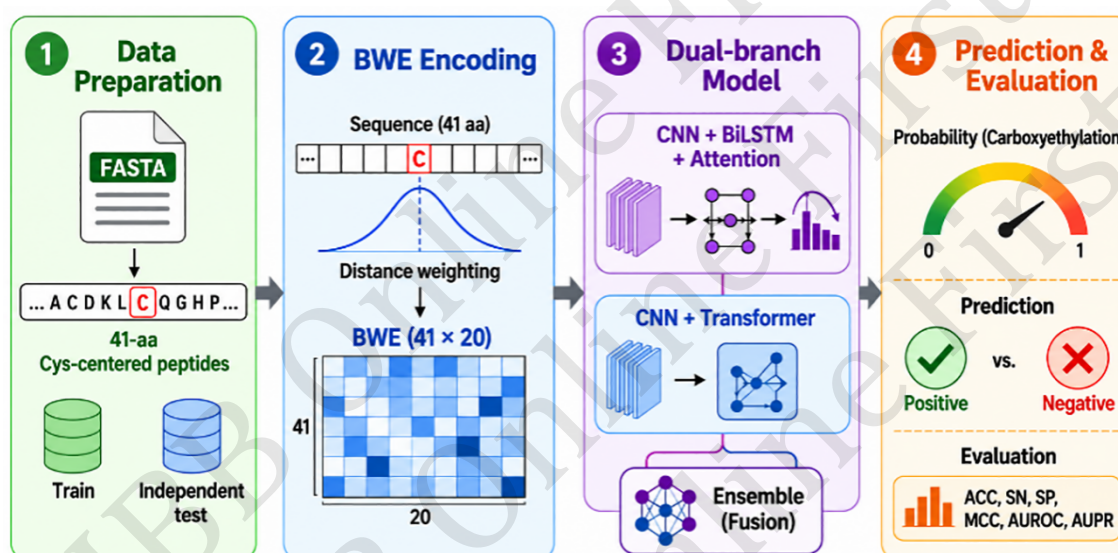
Cys-DHE: a Dual-branch Heterogeneous Deep Ensemble Model for Cysteine S-carboxyethylation Site Prediction*

WEI Xin^{1)**}, Hu Si-Qin²⁾

⁽¹⁾Business School, Jiangxi Institute of Fashion Technology, Nanchang 330201, China;

⁽²⁾School of Mega Data, Jiangxi Institute of Fashion Technology, Nanchang 330201, China)

Graphical abstract



Abstract Objective Cysteine S-carboxyethylation is a recently identified post-translational modification with potential biological relevance, but its experimental identification is costly and time-consuming. Existing computational approaches mainly rely on sequence information and may not fully exploit positional differences around the central cysteine or the complementary representation capabilities of heterogeneous neural networks. This study aimed to develop an effective sequence-based model for prioritizing potential cysteine S-carboxyethylation sites. **Methods** We developed Cys-DHE, a dual-branch heterogeneous deep ensemble model for cysteine S-carboxyethylation site prediction. The benchmark dataset contained 1 204 training samples, including 602 positive and 602 negative samples, and an independent test set containing 913 samples, including 83 positive and 830 negative samples. Each sample was represented by a 41-aa sequence window centered on the target Cys. Binary-weight encoding (BWE) was used to combine amino-acid identity with a distance-dependent weighting mechanism, thereby emphasizing residues closer to the central site. The resulting 41×20 feature matrix was simultaneously input into two complementary branches: CNN-BiLSTM-Attention for extracting local patterns and bidirectional contextual dependencies, and CNN-Transformer for modeling cross-position interactions and long-range dependencies. The prediction probabilities of the two branches were integrated using fixed weights of 0.75 and 0.25, respectively, determined according to five-fold cross-validation results on the training set. Model performance was evaluated using sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthews correlation coefficient (MCC), AUROC, and AUPR, and was further compared with individual branches and

conventional machine-learning methods. **Results** In five-fold cross-validation, Cys-DHE achieved *Sn*, *Sp*, *Acc*, *MCC*, *AUROC*, and *AUPR* values of 0.680 3, 0.842 7, 0.751 9, 0.526 2, 0.825 7, and 0.856 0, respectively. Compared with the individual branches, the fixed-weight ensemble improved *Sn*, *Acc*, *AUROC*, and *AUPR*, indicating complementary information between the two network architectures. Cys-DHE also outperformed SVM, RF, XGBoost, and LightGBM across the evaluated metrics under the same cross-validation setting. On the independent test set, Cys-DHE achieved an *Acc* of 89.60%, *Sn* of 69.88%, *Sp* of 91.57%, *MCC* of 0.508 8, *AUROC* of 0.865 4, and *AUPR* of 0.485 2. Compared with DLBWE-Cys, Cys-DHE showed higher *Sp*, *Acc*, and *MCC*, although *Sn*, *AUROC*, and *AUPR* were slightly lower. Additional local Cys analysis showed that positive sites were more frequently accompanied by additional Cys residues in their local sequence neighborhoods. Among samples containing an additional Cys within the 41-aa window, the median distance from the central Cys to the nearest additional Cys was approximately 5 residues in positive samples and 10 residues in negative samples, with a significant difference ($P=1.35\times 10^{-4}$). **Conclusion** Cys-DHE effectively integrates position-aware sequence encoding with complementary heterogeneous deep-learning architectures and provides balanced performance for cysteine S-carboxyethylation site prediction. The model is intended as a sequence-based candidate prioritization tool for narrowing potential modification sites before subsequent structural analysis and experimental validation. Sequence proximity should not be interpreted as direct evidence of three-dimensional spatial proximity or disulfide-bond formation. Future work will incorporate structural descriptors, solvent accessibility, spatial neighborhood information, and reliable disulfide-bond annotations, while expanding experimentally validated datasets to further improve model robustness, interpretability, and generalization across proteins. An online prediction server is available at <http://www.lzzzlab.top/Cys-DHE/>.

Key words cysteine S-carboxyethylation, protein post-translational modification, deep learning, heterogeneous ensemble, binary-weight encoding

DOI: 10.3724/j.pibb.2026.0446

CSTR: 32369.14.pibb.20260446

* This work was supported by a grant from The Scientific Research Plan of the Department of Education of Jiangxi Province (GJJ2402711, GJJ2402712, GJJ2402708, GJJ2502709).

** Corresponding author.

Tel: 86-791-87305677, E-mail: jci_wx@163.com

Received: August 14, 2026 Accepted: October 12, 2026