Hans Journal of Computational Biology 计算生物学, 2026, 16(2), 52-63
Published Online June 2026 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2026.162005

Hans汉斯

一种基于知识蒸馏与交叉注意力的轻量级多模态增强子识别方法

刘禹见

大连交通大学理学院，辽宁 大连

收稿日期：2026年5月10日；录用日期：2026年6月3日；发布日期：2026年6月9日

摘　要

增强子是一类非编码元件，在基因转录调控中起关键作用。如果增强子出问题，会和很多疾病密切相关。传统实验方法在鉴定增强子时，成本高，时间也长。现有的计算方法在处理单细胞数据或者覆盖度较低的表观组学数据时，识别效果会明显变差。我们提出了一种用于低样本量表观基因组学增强子识别的知识蒸馏轻量化多模态交叉注意力模型(AttLight-Enhancer)，这是一个多模态轻量化深度学习模型，用到了交叉注意力机制。它的目标是从低覆盖度多组学数据中识别增强子。我们还用了知识蒸馏的方法，把复杂教师模型里的知识迁移到轻量化学生网络中。我们在模拟低覆盖度数据和真实单细胞ATAC-seq数据上做了实验。模拟数据的覆盖度从10%一直降到1%。结果显示，当覆盖度只有1%时，AttLight-Enhancer的AUROC达到了0.865，比各个基线模型都好。经过知识蒸馏，学生模型的参数量只有教师模型的18.5%。通过注意力可视化分析，我们还看到了不同表观遗传特征的重要程度。其中H3K27ac最重要，然后是染色质可及性，最后是序列信息。

关键词

增强子识别，交叉注意力，多模态学习，知识蒸馏，低样本量

A Lightweight Multimodal Enhancer Identification Method Based on Knowledge Distillation and Cross-Attention

Yujian Liu

School of Science, Dalian Jiaotong University, Dalian Liaoning

Received: May 10, 2026; accepted: June 3, 2026; published: June 9, 2026

文章引用: 刘禹见. 一种基于知识蒸馏与交叉注意力的轻量级多模态增强子识别方法[J]. 计算生物学, 2026, 16(2): 52-63. DOI: 10.12677/hjcb.2026.162005

Abstract

Enhancers are a class of non-coding elements that play a critical role in the transcriptional regulation of genes. Enhancer dysfunction is closely associated with numerous diseases. Traditional experimental methods for identifying enhancers are costly and time-consuming. Existing computational approaches exhibit significantly degraded performance when processing single-cell data or epigenomic data with low coverage. In this study, we propose AttLight-Enhancer, a lightweight multimodal deep learning model that employs a cross-attention mechanism. It is designed to identify enhancers from low-coverage multi-omics data. We also utilize knowledge distillation to transfer knowledge from a complex teacher model to a lightweight student network. We conduct experiments on both simulated low-coverage data and real single-cell ATAC-seq data, with the coverage of simulated data reduced from 10% to 1%. The results demonstrate that AttLight-Enhancer achieves an AUROC of 0.865 at only 1% coverage, outperforming all baseline models. After knowledge distillation, the number of parameters of the student model is merely 18.5% of that of the teacher model. Through attention visualization analysis, we further reveal the importance of different epigenetic features: H3K27ac is the most critical, followed by chromatin accessibility, and finally sequence information.

Keywords

Enhancer Identification, Cross-Attention, Multimodal Learning, Knowledge Distillation, Low-Input Epigenomics

1. 引言

增强子是远端顺式调控元件家族中十分重要的成员，它本身并不编码蛋白质，但是可以在特定的细胞类型和时空条件下精确地调节基因转录的过程，从而控制细胞的命运以及组织的稳态[1] [2]。而且当研究分辨率提高到单细胞水平时会发现增强子在不同细胞中的活性具有异质性[3] [4]。此外 80%以上与复杂疾病相关的遗传变异都富集于增强子区域[5] [6]，所以无论是要解析基因调控的基本规律还是揭示疾病的发病机理，准确地定位增强子都是必不可少的基础工作。

过去做增强子鉴定，主要靠 ChIP-seq、Hi-C 这些实验手段[7] [8]。可问题是，它们需要大量细胞才能启动，一旦碰到细胞群体里的异质性，就有些力不从心了。后来单细胞表观组学技术出现了，比如 scATAC-seq，终于让人看到了希望[4]。但真要落地，又得面对另一道坎：数据覆盖度太低、信号稀疏、噪声还大。说到底，怎么搭一个端到端的模型，能直接从这种低质量的数据里稳健地识别出增强子，至今还是个没解决的难题。

早先的计算方法，思路比较简单，用进化保守性、k-mer 频率这些特征，喂给 SVM 做分类[9]。深度学习兴起后，利用 CNN、RNN 模型自己就能从序列里学特征，不用人手工去挑[10] [11]。再到最近，基于 Transformer 的预训练模型，比如 DNABERT2，已经在标准基准上把性能推到了一个新高度[12]。不过话说回来，纯靠序列做文章，总归有两道绕不过去的坎。一个是增强子的活性是动态的、会变的，光看序列根本捕捉不到这种变化；另一个是样本量一少，模型就容易死记硬背，过拟合得厉害。

想要预测细胞类型特异性的增强子活性，研究人员开始把多种组学数据整合起来使用，DeepCAPE 采

用了双通道卷积神经网络，序列信息和染色质可及性数据是分开处理的[13]，不过融合方式相对简单一些，之后的一些工作尝试了更复杂的多模态融合策略，比如基于跨模态 Transformer 的方法[14] [15]。不过，在样本量比较少的情况下，现有模型对数据稀疏和高噪声的耐受性还是存在不足，多模态之间的交互深度不够，模型可解释性也有待提升，另外当前方法在计算效率和预测性能之间往往难以兼顾。

单细胞、微量样本表观组学技术的应用愈发普遍，数据稀疏性强、背景噪音偏高等问题也愈发突出，注意力机制下的交叉注意力，可建模长距离依赖关系、动态分配特征权重，在基因组学领域已得到较多关注[16] [17]。知识蒸馏是效率较高的模型压缩与迁移学习手段，在生物信息学领域的应用正逐步推进[18]，上述技术为搭建性能优良、运行高效的增强子识别模型提供了新方向。

当前学界已有的相关研究，核心待解决的问题十分清晰，要留存多模态建模现有能力，要覆盖低样本量环境下的稀疏性与噪声稳定性要求，还要把模型实际部署效率划入考虑范围，本文提出 AttLight-Enhancer 框架解决上述问题。这一框架采用知识蒸馏方法，在模型深度融合效果与轻量化运行需求间找到平衡点，在低样本量场景中使用该框架，可同时具备运行稳定、可解释性强、部署效率高三类优势。

2. 材料与方法

2.1. 数据集

从 ENCODE 与 Roadmap 计划中收集三种人类永生化细胞系(GM12878、K562、HepG2)的数据，包括：DNA 序列、H3K27ac ChIP-seq、ATAC-seq 或 DNase-seq。正样本整合 VISTA Enhancer Browser [19] 等高置信度区域并与 ENCODE cCREs 取交集，负样本选自基因荒漠区且缺乏活性增强子标记。每个细胞系构建约 10,000 个正负样本的平衡数据集，区域长度统一为 2000 bp。

为模拟单细胞测序的低覆盖度，开发基于泊松采样的数据降级流程，生成覆盖度为 50%、25%、10%、5%和 1%的模拟数据集，每个降级过程独立重复 5 次。

为验证真实稀疏数据上的泛化能力，选取人类 PBMC 10x Genomics 数据(同物种)和小鼠胚胎发育 scATAC-seq 数据(跨物种) [20]。采用“伪批量”分析策略，将同一细胞类型聚类内所有细胞的测序片段聚合，以已发表的细胞类型特异性增强子为金标准。

2.2. AttLight-Enhancer 框架概述

本文选取分阶段推进的知识蒸馏方案，第一步为教师模型训练环节，导入高质量基准数据集，完成深度多模态融合教师模型的训练。第二步为学生模型蒸馏环节，锁定教师模型的参数，开展轻量化学生网络的训练，优化融合蒸馏损失与真实标签损失的函数。全训练流程所用数据集，由模拟低样本量数据与少量高质量数据混合构成(如图 1 所示)。

2.3. 教师模型(Teacher Model)架构

教师模型是以交叉注意力为核心的多模态深度学习网络。

2.3.1. 多模态输入编码

长度为 2000 bp 的基因组区间作为单个分析样本，处理时会同步转换为三个模态的输入张量：

- 序列模态(X_{seq})：先把 DNA 序列变成一个 2000 × 4 的 one-hot 矩阵(A、T、C、G 各占一列)。
- H3K27ac 信号模态(X_{hist})和 ATAC-seq 信号模态(X_{acc})：为兼顾计算效率与特征分辨率，本文将每个 2000 bp 的区间拆分为 200 个连续小段，单段长度为 10 bp，再针对每个小段，分别测算其在 H3K27ac ChIP-seq 或 ATAC-seq 数据中的归一化读数密度，可先取 RPKM 值再做对数转换，最终可得到两个独立的 200 × 1 向量。

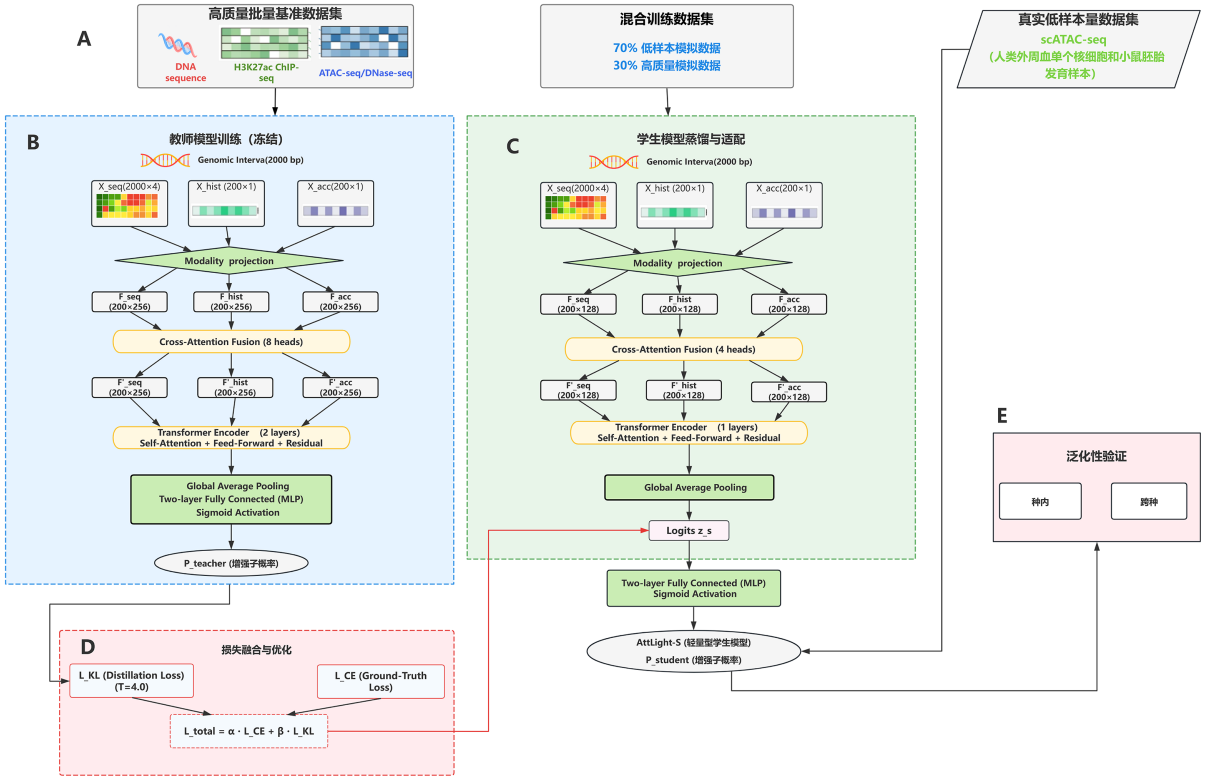

Figure 1. Two-stage knowledge distillation framework of the AttLight-Enhancer model
图 1. AttLight-Enhancer 模型的两步式知识蒸馏框架

2.3.2. 交叉注意力融合层

模型实现深度多模态交互，核心是这个模块。本文设计了一条由交叉注意力引导的早期深度融合路径，就是为了避免简单的后期拼接，具体分为以下几步：

1) 模态特征投影：三类模态的原始输入(X_{seq}, X_{hist}, X_{acc})分别通过各自独立的学习型线性变换层处理，被投射到同一个统一特征空间内，得到初始的序列特征 F_{seq}、组蛋白对应的特征 F_{hist} 以及可及性对应的特征 F_{acc}。所有特征的维度均为 $L \times D$ (这里 L 代表序列或 bins 的长度，D 为隐藏维度)。

2) 交叉注意力融合：我们采用单层多头交叉注意力机制来实现模态间的深度交互。具体地，我们将三个模态的特征视为一组“令牌”(*tokens*)。对于目标模态 i，我们以其投影特征 F_i 作为查询(*Query*, Q_i)，而将所有三个模态的投影特征拼接起来形成一个全局上下文记忆，并由此记忆派生出键(*Key*, K_{all})和值(*Value*, V_{all})。然后计算目标模态 i 与全局上下文之间的交叉注意力：

$$\text{Attention}\left(Q_i, K_{all}, V_{all}\right) = \text{softmax}\left(\frac{Q_i \cdot K_{all}^{T}}{\sqrt{d_k}}\right) \cdot V_{all} \tag{1}$$

其中，$\sqrt{d_k}$ 是缩放因子，d_k 是键向量 K_{all} 的维度，可稳定训练。完成这步计算后，目标模态 i 中的每个位置都可动态“关注”，同时整合所有模态(包含自身)所有位置的关联信息，本文对每个模态都并行开展这一操作，最终得到三个完成深度融合的特征表示：F'_{seq}，F'_{hist}，F'_{acc}。

3) 多头注意力的生物学合理性：本文设置的注意力模块共包含 8 个并行头。多头运行模式下，模型可同步从多个独立子空间中提取信息，这一逻辑和增强子发挥功能时的多路径协同机制一致，比如可同时识别不同的转录因子结合模块、染色质修饰模式或是染色质可及性状态。采用这种多头结构，是为了

让模型能更准确地捕获多层面、异构的调控特征组合。

2.3.3. 融合特征整合与抽象

本文将前述步骤得到的三个完成深度融合的特征 F'_{seq}，F'_{hist}，F'_{acc} 沿特征维度完成拼接，得到一个综合特征张量。之后将该张量输入主干网络，该网络包含 2 个 Transformer 编码器层。单个编码器层的内部结构涵盖多头自注意力机制、前馈神经网络，搭配残差连接与层归一化操作。该主干网络并不承担模态融合任务(该步骤已在交叉注意力环节完成)，仅从完成深度融合的多模态信息中进一步提取层级更高、更抽象的全局特征，供最终分类环节使用。

2.3.4. 输出层

主干网络输出的特征序列经过全局平均池化后，被送入一个两层全连接网络，最后通过 Sigmoid 激活函数输出该输入区间为增强子的概率 $P_{teacher}$。

2.4. 学生模型与知识蒸馏

2.4.1. 轻量化学生网络设计

教师模型的核心多模态数据流与融合方法，都被学生模型完整保留，仅整体架构开展全面精简，最终完成轻量化改造。具体裁剪内容为：特征维度从 256 减至 128，网络深度从 2 层减为 1 层，二者均减半，注意力头数从 8 个精简到 4 个。

对学生模型完成上述裁剪处理后，它的参数量约为教师模型的 23%。后续开展的实验验证，该模型的容量足以容纳教师模型迁移的知识，运行效率也有比较大的提升。

2.4.2. 知识蒸馏流程与损失函数

学生模型的训练目标是双重的：拟合真实标签，同时模仿教师模型更富含信息的“软标签”概率分布。

总损失函数定义为：

$$L_{total} = \alpha \cdot L_{CE} + \beta \cdot L_{KL} \tag{2}$$

其中：

L_{CE} 是学生模型预测概率与真实硬标签 y 之间的二元交叉熵损失。

L_{KL} 是学生模型与教师模型输出概率分布之间的 KL 散度损失。具体计算涉及温度参数 T 以软化分布：

$$L_{KL} = T^2 \cdot KL\left[\text{softmax}\left(\frac{z_s}{T}\right) \middle\| \text{softmax}\left(\frac{z_t}{T}\right)\right] \tag{3}$$

其中 z_t 和 z_s 分别是教师和学生模型在最终 Sigmoid 激活前的 logits。

关键参数的设定依据如下：

- 温度参数 T：本文采用网格搜索在验证集上测试了 T 取 1、2、4、8 对应效果。测试发现，$T = 4$ 对应的取值条件下可生成足够“软化”的概率分布(即教师模型对非正确类别的预测概率有所提升)，这类分布可更清晰地呈现不同类别间的相似关联，在本次任务中得到的知识迁移效果最优。如果温度参数取值偏低($T = 1$)，对应分布与硬标签接近，知识迁移效果不佳；如果温度参数取值偏高($T = 8$)，对应概率分布会过于平滑，会损耗其中的有效信息。
- 损失权重 α 与 β：本文选取简易的余弦退火调度策略，对 α 与 β 的占比做动态调整，可保障整个训练过程具备可复现性训练正式开启的初期，本文将 β 定为 0.9，α 定为 0.1，为了优先完成教师模型

的知识迁移步骤。训练进入后续阶段时，本文按照下面的公式逐步提升 α 的权重，保障模型最终能够准确拟合真实标签。

$$\alpha_{epoch} = 0.1 + 0.9 \times \frac{1 - \cos\left(\frac{\pi \times epoch}{N_{total}}\right)}{2}$$
$$\beta_{epoch} = 1 - \alpha_{epoch} \tag{4}$$

其中 N_{total} 为总训练轮数。此策略在训练初期侧重于蒸馏损失以引导学生，后期则侧重于真实标签损失以进行校准。

2.5. 训练策略与超参数优化

先将教师模型放入高质量数据集开展训练，直至达到收敛状态，固定该模型的全部参数，选用混合数据集(模拟低样本量与高质量数据的比例为 7:3)，开展学生模型的训练工作。本文选用 AdamW 优化器，给学生模型设置强度更高的正则化(Dropout 为 0.3，教师模型为 0.2)，所有核心超参数的取值，都借助贝叶斯优化方法筛选得出，最终确定的参数取值全部列于表 1 中。

Table 1. Key hyperparameter configuration of AttLight-Enhancer
表 1. AttLight-Enhancer 关键超参数配置

超参数	描述	教师模型	学生模型
基础学习率	AdamW 优化器初始学习率	3e−4	5e−4
批次大小	-	64	128
Dropout 率	全连接层及注意力后的 Dropout	0.2	0.3
注意力头数	交叉注意力与 Transformer 层中的头数	8	4
特征维度	模型隐藏层维度	256	128
蒸馏温度 T	软化概率分布的温度	-	4
初始损失权重 α	真实标签损失初始权重	-	0.1
初始损失权重 β	蒸馏损失初始权重	-	0.9

2.6. 性能评估与可解释性分析

评估指标包括准确率、精确率、召回率、F1 分数、AUROC 和 AUPRC。

对比基线：iEnhancer-CNN [10]、DeepCAPE [13]、Enhancer-MDLF [21]，以及不经蒸馏的 Student-From-Scratch。

可解释性方法：交叉注意力权重可视化、模态丢弃实验和梯度敏感性分析、注意力模式与生物学特征的定量关联、综合案例研究。

3. 结果

3.1. 教师模型在高质量基准数据集上的性能

将教师模型应用于 GM12878、K562、HepG2 三类细胞系测试，得到的平均 AUROC、AUPRC 数值分别为 0.986 和 0.972 (表 2)，对照同类模型 iEnhancer-CNN 的 0.924、DeepCAPE 的 0.951，两项指标分

别提高 6.2%与 3.5%，交叉注意力机制可增强多模态整合能力，提升细胞类型特异性增强子的预测精度。将该模型用于覆盖度 10%的模拟低样本量数据测试，性能出现下降，平均 AUROC 仅为 0.901，对应测试结果可印证，开展知识蒸馏适配具备实际必要性。

Table 2. Performance comparison under different data conditions across each model (AUROC/AUPRC)
表 2. 各模型在不同数据条件下的性能对比(AUROC/AUPRC)

模型	GM12878	K562	HepG2	高质量数据平均值	模拟 10%覆盖度数据平均值
iEnhancer-CNN	0.928/0.901	0.922/0.889	0.923/0.895	0.924/0.895	0.782/0.741
DeepCAPE	0.955/0.932	0.949/0.925	0.949/0.928	0.951/0.928	0.831/0.802
AttLight-Enhancer (教师)	0.988/0.975	0.985/0.971	0.986/0.971	0.986/0.972	0.901/0.874

3.2. 学生模型在模拟低样本量数据上的性能

要验证模型对数据稀疏性的鲁棒性，要检验蒸馏策略的有效性，本文用系统生成的模拟数据集，在低样本量条件下，开展了严格测试。训练阶段全部模型都要在覆盖度为 10%的模拟数据上，完成重新训练或微调，AttLight-Enhancer 学生模型采用蒸馏训练，后续在覆盖度从 10%到 1%的不同测试集上完成评估。

3.2.1. 不同覆盖度下的性能变化与样本分析

图 2(a)的统计结果显示，数据覆盖度从 10%降至 1%的过程中，所有参与测试的模型性能均出现幅度不一的下滑，AttLight-Enhancer 学生模型(标注为 AttLight-S)的性能下滑幅度最小，在覆盖度低至 1%的极端测试场景中，该模型的平均 AUROC 依旧可以维持在 0.865 的水平。

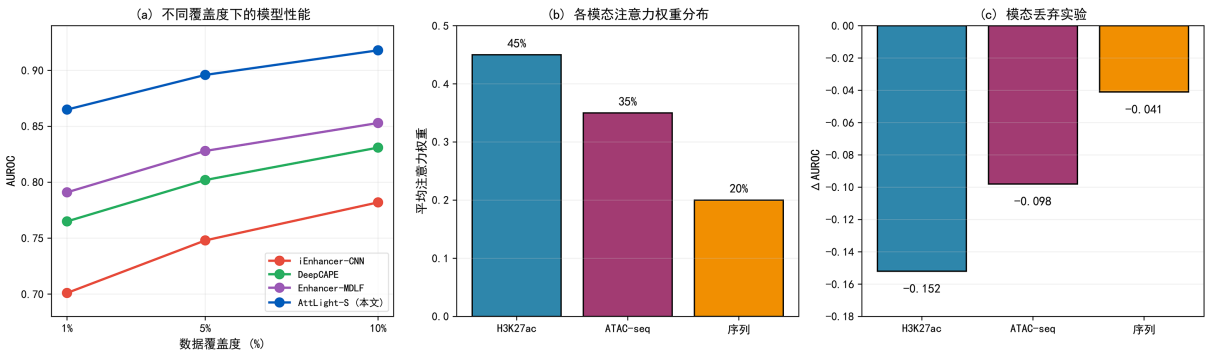

Figure 2. Comprehensive analysis of AttLight-Enhancer model performance and interpretability
图 2. AttLight-Enhancer 模型的性能与可解释性综合分析

3.2.2. 蒸馏策略有效性分析

为了弄清楚知识蒸馏策略到底起了多大作用，我们设计了对照实验(表 3)。结果发现，从头开始训练的学生模型(Student-From-Scratch)的 AUROC 只有 0.842，明显低于蒸馏后的 AttLight-S (AUROC 为 0.918)。这就证明了蒸馏带来的提升很大。直接对教师模型做微调，AUROC 为 0.878，虽然比从头训练要好，但还是不如蒸馏策略。这说明“先蒸馏后适应”的方法能更有效地把大规模数据里的稳定知识迁移到轻量网络里。

Table 3. Model performance (AUROC) and comparison on simulated 10% coverage test set
表 3. 覆盖度 10%的模拟测试集上各模型性能(AUROC)与对比

模型	训练策略	AUROC	F1 值	参数(相对教师模型)
iEnhancer-CNN	从头训练	0.782	0.741	-
DeepCAPE	从头训练	0.831	0.802	93.50%
Enhancer-MDLF	从头训练	0.853	0.819	-
教师模型(微调)	在低样本量数据上微调	0.878	0.843	100%
Student-From-Scratch	轻量化架构从头训练	0.842	0.808	23%
AttLight-Enhancer (学生)	本研究的蒸馏策略	0.918	0.881	23%

本文选用的软标签蒸馏法，属于当前主流的高效技术路线。其余蒸馏衍生方案虽能带来部分附加增益，不过往往会产生更高的计算与调参开销。本文完成的实验结果显示，现有研究策略可顺利达成预设的核心目标。蒸馏过程还可优化模型的校准效果：覆盖度为 10%的测试集内，AttLight-S 的预期校准误差(ECE)为 0.032，数值低于从头训练的学生模型(0.078)，也低于完成微调的教师模型(0.051)。对应输出的概率结果可靠性更高，能够适配下游生物学发现的优先级排序需求。

蒸馏训练单轮时长约为单独训练学生模型的 1.8 倍，原因是训练过程中需要同步运行教师模型，不过这部分前期开销，能让学生模型的综合性能得到比较大的提升，与微调教师模型的方案相比，蒸馏的总耗时更少，最终得到的模型在推理效率上具备数量级优势(表 4)，知识蒸馏是通过可控的额外训练成本，实现模型精度、校准度和推理速度多方面提升的关键策略。

Table 4. Comprehensive comparison of model complexity and efficiency
表 4. 模型复杂度与效率综合对比

模型	参数量(百万)	相对教师参数量	单样本推理时间(ms)	内存占用(MB)	在低样本量数据上的AUROC
AttLight-Enhancer (教师)	52.1	100%	15.2	1024	0.878 (微调后)
DeepCAPE	48.7	93.5%	12.8	890	0.831
AttLight-Enhancer (学生)	9.6	18.5%	3.8	~300	0.918
Student-From-Scratch (对照)	9.6	18.5%	3.9	~300	0.842

3.3. 学生模型在真实低样本量数据集上的泛化能力

为了验证模型的实用性，我们直接把 AttLight-S 用在两个独立的真实低样本量验证集上，不做任何微调，以此来评估它在跨数据集和跨物种情况下的泛化能力。

在人类 PBMC“伪批量”数据上，我们以经过正交实验验证的细胞类型特异性增强子集合作为金标准。结果发现，AttLight-S 预测出的增强子，在 CD4+ T 细胞、B 细胞等经典免疫细胞的特异性基因启动子附近明显富集(超几何检验，$p < 1e{-}10$)。这说明模型的预测有很好的生物学关联性。

在小鼠脊髓发育数据的评估中，我们屏蔽了序列输入通道，只靠模型从人类数据中学到的“表观信号–功能”映射来做预测。结果发现，在侧运动柱神经元簇里，模型预测得分前 5%的高置信度区域中，与文献功能支持的候选增强子重叠的比例达到了 68%，明显高于随机期望($p < 1e{-}15$)。而且这些重叠区

域里，小鼠与人类序列保守的元件也明显富集($p < 1e{-}8$)。与基线模型 DeepCAPE (41%, $p < 1e{-}5$)和 Enhancer-MDLF (52%, $p < 1e{-}10$)相比，AttLight-S 表现得更好。

我们还通过加入模拟噪音来测试模型的稳定性。结果显示，AttLight-S 的性能波动比对比模型要小。这说明，通过大规模多细胞系训练和多模态融合，模型学到了对技术噪音更有抵抗力的稳定特征表示。

3.4. 模型轻量化效果分析

表 4 数据显示，完成知识蒸馏与架构裁剪的学生模型，参数量仅为教师模型的 23%，对应参数规模为 9.6 M、52.1 M，单样本推理耗时约为教师模型的 1/4，内存占用较教师模型降低约 70%。

和其他轻量方案做对比，从头开始训练的学生模型(Student-From-Scratch)推理效率和 AttLight-S 相差不大，模型性能存在比较大的差距，其 AUROC 数值为 0.842，AttLight-S 的对应数值为 0.918。同样完成裁剪的教师模型，未经过蒸馏步骤时性能为 0.851，表现也低于 AttLight-S。“先训练后蒸馏”的策略应用效果更优。

查看图 3 的性能与效率权衡分析结果，隐藏层维度 D 设为 128、Transformer 层数 L 取 1 时，模型处于最优权衡状态，推理速度提升近 4 倍，性能损失控制在最低范围，和微调后的教师模型相比，性能仅下降约 1%。如果把维度压缩到 D = 64，模型性能会出现大幅下滑，要是选择更复杂的 D = 256 配置，又会出现速度降幅远高于性能增益的问题。以上测试结果可以证实，本文提出的轻量化设计达成了“小而精”的预期目标。

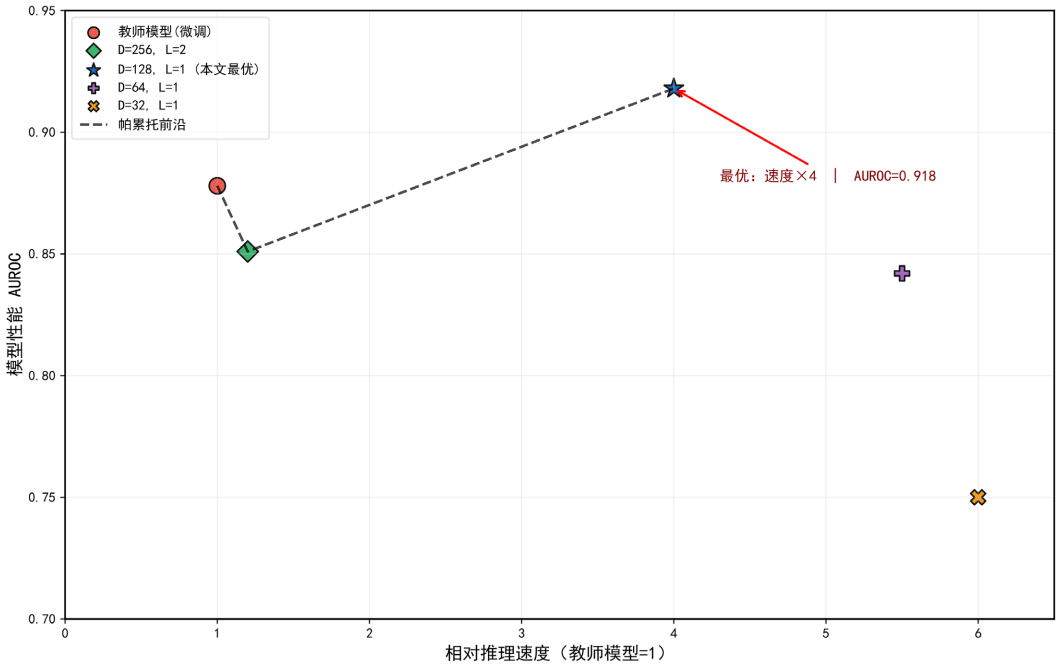

Figure 3. Trade-off curve of model performance (AUROC) and efficacy
图 3. 模型性能(AUROC)与效率的权衡曲线

3.5. 可解释性分析结果

3.5.1. 注意力权重模式的量化分析

对交叉注意力权重开展全局统计分析所得结果对应图 2(b)，模型预测的高置信度增强子区域中，三类模态的平均注意力权重存在梯度差异，H3K27ac 模态数值约为 45%，占比最高，ATAC-seq 模态数值

约为 35%，位列第二。序列模态数值约为 20%，占比最低，为验证该排序对应生物学层面的重要性区别，而非算法本身存在偏差，本文对比了各模态在正、负样本中的可区分度，发现 H3K27ac 信号的差异最明显。本文还采用 Integrated Gradients 方法完成独立验证，所得结果与注意力权重的排序完全吻合，可交叉印证各模态的重要性。

3.5.2. 模态贡献的鲁棒性评估

模态丢弃实验(图 2(c))显示，当丢掉 H3K27ac 模态时，模型的 AUROC 下降最大(ΔAUROC = −0.152)；丢掉 ATAC-seq 模态次之(−0.098)；丢掉序列模态影响最小(−0.041)。基于梯度的敏感性分析也得到了相同的结果。这两种方法共同证实：多模态信息对高精度识别非常重要；各个模态的贡献有主次之分，这和生物学上的认识是一致的；融合机制有效利用了所有模态的互补信息。

3.5.3. 多维度案例研究

我们选取了三类增强子区域进行详细分析：

经典验证案例显示，针对阿尔茨海默症风险基因 BIN1 邻近的某功能增强子，模型输出的增强子预测概率为 0.94。如图 4 所示，模型的注意力机制聚焦于高置信度的 H3K27ac 峰及 ATAC-seq 信号富集区域；同时，序列注意力模块识别出与 REST 等神经相关转录因子相匹配的序列模式，进一步佐证了预测结果的生物学可解释性。

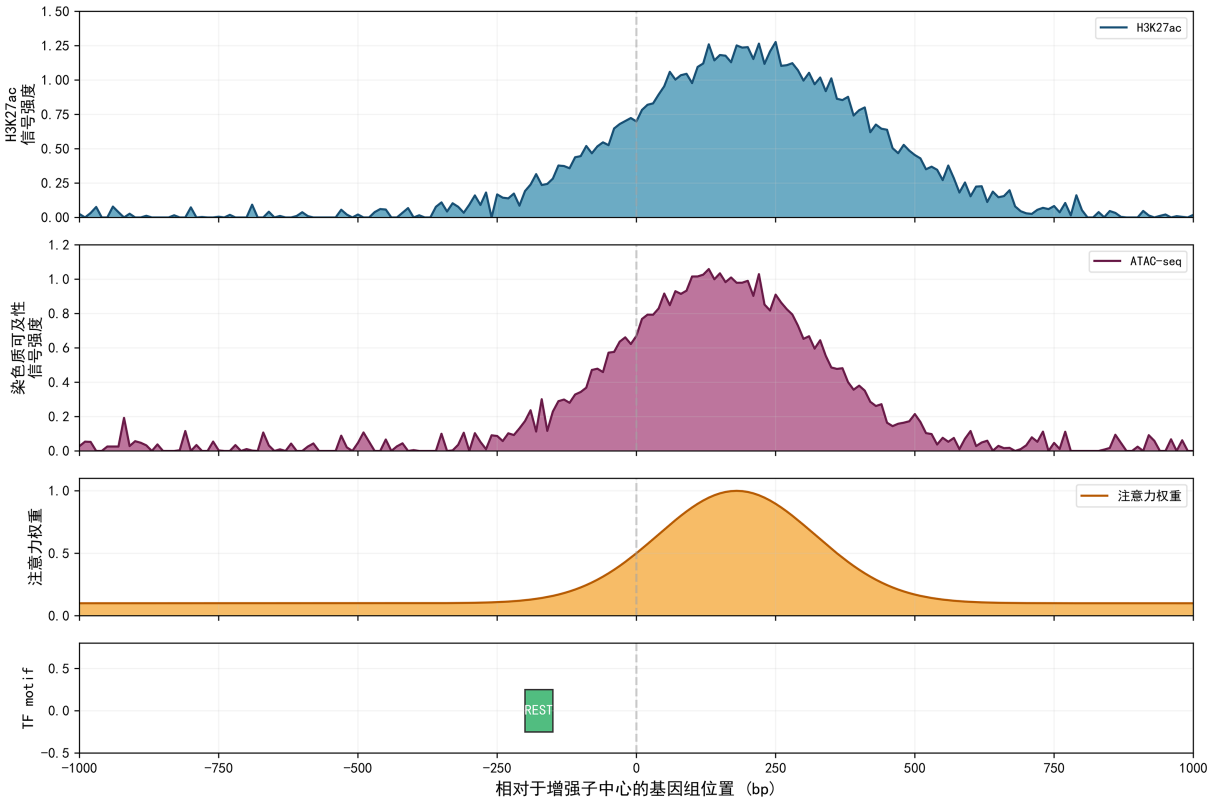

Figure 4. Case study of the BIN1 locus
图 4. BIN1 基因座的案例分析

选取超级增强子相关案例开展验证，MYC 基因座上的超级增强子区域，模型的预测概率大于 0.99。多个带有高表观信号的核心区域可被该模型定位，密集的转录因子结合基序簇也能被捕获。

预置增强子探索案例：某典型预置染色质特征区域存在 H3K4me1 修饰，但缺少 H3K27ac，针对该区域模型给出的预测概率处于中等水平，约为 0.6。模型得出该结果主要依托序列信息与微弱的可及性信号，几乎未考量缺失的 H3K27ac 信号，判断逻辑和该区域的生物学特性相吻合。

汇总现有案例结果能看到，AttLight-Enhancer 可整合多模态证据，针对不同状态调控元件给出的判断，符合生物学常理，决策依据也可追溯。

4. 讨论

AttLight-Enhancer 整合三类核心技术，分别是依托交叉注意力实现的深度多模态融合、知识蒸馏与轻量化设计，能满足低样本量场景下的增强子识别需求，是一套完整的集成处理方案，用模拟生成的低覆盖度数据集测试，该模型的鲁棒性表现稳定。最终训练出的学生模型参数量只有教师模型的 18.5%，能平衡性能、效率与可靠性三方面的运行要求。本文研究结果显示，知识蒸馏可以缓解生物信息学领域“小数据、大模型”的适配矛盾，是具备可行性的技术路径。

本文采用的知识蒸馏策略，本质是借助软标签实现从数据富集域向稀疏域的归纳偏置迁移。在低覆盖度测序中，二值化硬标签常因读段稀疏而包含大量假阴性噪声；而教师模型在深度数据上输出的连续概率软标签，则在信息熵层面具备两类关键优势：其一，类别相似性迁移。软标签的概率分布隐含了不同区域在特征流形上的相对距离，其模糊性迫使学生模型学习细微的梯度变化，在稀疏训练中起到数据增广与正则化作用，防止对假阳性峰的过拟合。其二，不确定性量化。软标签将边界区域的高不确定性传递给学生，允许其在低覆盖区域输出非饱和概率，维持稳定梯度流，促进平滑泛化。

但该策略亦面临潜在挑战：第一，教师校准偏差。若教师模型存在对特定序列背景(如高 GC 区域)的系统性偏好，软标签将放大此偏差；在证据稀疏时，学生模型更易产生误判。第二，表征鸿沟。高覆盖数据的高置信度软目标在低覆盖域可能失效，强行拟合反而导致过正则化，淹没微弱真实信号。第三，温度系数的域适应。固定温度难以兼顾域间差异，需探索根据局部覆盖深度动态调节软标签平滑程度的自适应机制。

综上，知识蒸馏在此处不仅是模型压缩工具，更构筑了跨覆盖度域的信息桥梁，为低样本量生物序列分析提供了规避“小数据诅咒”的正则化路径。其成效高度依赖于教师模型的概率校准精度与域适应蒸馏策略的设计。

本研究存在以下局限：

(1) 模型主要聚焦在典型的活跃增强子上，对预置增强子的识别可能不够好。

(2) 训练数据主要来自永生化细胞系，向原代细胞迁移时可能存在泛化偏差。

(3) 目前整合的模态还不完整，缺少三维基因组信息。

(4) 在超大规模单细胞数据上的扩展性还面临工程上的挑战。

本文后续的改进工作将从四个方向开展，一是覆盖更多类型的组蛋白修饰，二是持续补充多样的原代细胞数据，三是整合三维基因组信息，四是搭建效率更高的并行推理流水线。

参考文献

[1] Yang, J.H. and Hansen, A.S. (2024) Enhancer Selectivity in Space and Time: From Enhancer-Promoter Interactions to Promoter Activation. *Nature Reviews Molecular Cell Biology*, **25**, 574-591. https://doi.org/10.1038/s41580-024-00710-6

[2] Panigrahi, A. and O’Malley, B.W. (2021) Mechanisms of Enhancer Action: The Known and the Unknown. *Genome Biology*, **22**, Article No. 108. https://doi.org/10.1186/s13059-021-02322-1

[3] Vasileva, A.V., Gladkova, M.G., Ashniev, G.A., Osintseva, E.D., Orlov, A.V., Kravchuk, E.V., *et al.* (2024) Super-enhancers and Their Parts: From Prediction Efforts to Pathognomonic Status. *International Journal of Molecular*

Sciences, **25**, Article 3103. https://doi.org/10.3390/ijms25063103

[4] Shu, M., Hong, D., Lin, H., Zhang, J., Luo, Z., Du, Y., *et al*. (2022) Single-Cell Chromatin Accessibility Identifies Enhancer Networks Driving Gene Expression during Spinal Cord Development in Mouse. *Developmental Cell*, **57**, 2761-2775.e6. https://doi.org/10.1016/j.devcel.2022.11.011

[5] Green, N.F.O., Sutton, G.J., Pérez-Burillo, J., Wang, J., Bagot, S., Danon, H.G., *et al*. (2025) Crispri Screening in Cultured Human Astrocytes Uncovers Distal Enhancers Controlling Genes Dysregulated in Alzheimer's Disease. *Nature Neuroscience*, **29**, 703-716. https://doi.org/10.1038/s41593-025-02154-3

[6] Yang, S., Ahmed, I., Li, Y., Bleaney, C.W. and Sharrocks, A.D. (2024) Massively Parallel Reporter Assays Identify Enhancer Elements in Oesophageal Adenocarcinoma. *NAR Cancer*, **6**, zcae041. https://doi.org/10.1093/narcan/zcae041

[7] Wu, H., Zhang, J., Tan, L. and Xie, X.S. (2025) Single-Cell Micro-C Profiles 3D Genome Structures at High Resolution and Characterizes Multi-Enhancer Hubs. *Nature Genetics*, **57**, 1777-1786. https://doi.org/10.1038/s41588-025-02247-6

[8] Denaud, S., Bardou, M., Papadopoulos, G., Grob, S., Di Stefano, M., Sabarís, G., *et al*. (2024) A PRE Loop at the Dac Locus Acts as a Topological Chromatin Structure That Restricts and Specifies Enhancer-Promoter Communication. *Nature Structural & Molecular Biology*, **31**, 1942-1954. https://doi.org/10.1038/s41594-024-01375-7

[9] Wu, H., Liu, M., Zhang, P. and Zhang, H. (2023) iEnhancer-SKNN: A Stacking Ensemble Learning-Based Method for Enhancer Identification and Classification Using Sequence Information. *Briefings in Functional Genomics*, **22**, 302-311. https://doi.org/10.1093/bfgp/elac057

[10] Khanal, J., Tayara, H. and Chong, K.T. (2020) Identifying Enhancers and Their Strength by the Integration of Word Embedding and Convolution Neural Network. *IEEE Access*, **8**, 58369-58376. https://doi.org/10.1109/access.2020.2982666

[11] Niu, K., Luo, X., Zhang, S., Teng, Z., Zhang, T. and Zhao, Y. (2021) iEnhancer-EBLSTM: Identifying Enhancers and Strengths by Ensembles of Bidirectional Long Short-Term Memory. *Frontiers in Genetics*, **12**, Article 665498. https://doi.org/10.3389/fgene.2021.665498

[12] Wang, T. and Gao, M. (2025) Utilizing a Deep Learning Model Based on BERT for Identifying Enhancers and Their Strength. *PLOS ONE*, **20**, e0320085. https://doi.org/10.1371/journal.pone.0320085

[13] Chen, S., Gan, M., Lv, H. and Jiang, R. (2021) DeepCAPE: A Deep Convolutional Neural Network for the Accurate Prediction of Enhancers. *Genomics, Proteomics & Bioinformatics*, **19**, 565-577. https://doi.org/10.1016/j.gpb.2019.04.006

[14] Toneyan, S. and Koo, P.K. (2024) Interpreting Cis-Regulatory Interactions from Large-Scale Deep Neural Networks. *Nature Genetics*, **56**, 2517-2527. https://doi.org/10.1038/s41588-024-01923-3

[15] Xiao, Z., Wang, L., Ding, Y. and Yu, L. (2022) iEnhancer-MRBF: Identifying Enhancers and Their Strength with a Multiple Laplacian-Regularized Radial Basis Function Network. *Methods*, **208**, 1-8. https://doi.org/10.1016/j.ymeth.2022.10.001

[16] Basith, S., Hasan, M.M., Lee, G., Wei, L. and Manavalan, B. (2021) Integrative Machine Learning Framework for the Identification of Cell-Specific Enhancers from the Human Genome. *Briefings in Bioinformatics*, **22**, bbab252. https://doi.org/10.1093/bib/bbab252

[17] Mu, X., Huang, Z., Chen, Q., Shi, B., Xu, L., Xu, Y., *et al*. (2024) DeepEnhancerPPO: An Interpretable Deep Learning Approach for Enhancer Classification. *International Journal of Molecular Sciences*, **25**, Article 12942. https://doi.org/10.3390/ijms252312942

[18] Mehmood, F., Arshad, S. and Shoaib, M. (2024) ADH-Enhancer: An Attention-Based Deep Hybrid Framework for Enhancer Identification and Strength Prediction. *Briefings in Bioinformatics*, **25**, bbae030. https://doi.org/10.1093/bib/bbae030

[19] Kosicki, M., Baltoumas, F.A., Kelman, G., Boverhof, J., Ong, Y., Cook, L.E., *et al*. (2024) VISTA Enhancer Browser: An Updated Database of Tissue-Specific Developmental Enhancers. *Nucleic Acids Research*, **53**, D324-D330. https://doi.org/10.1093/nar/gkae940

[20] Yeo, I.S., Jeong, D., Lee, Y., Ahn, J., Bae, S. and Song, M. (2025) An Integrated Transcriptomic Approach for Identifying Enhancers in Limb Motor Pools. *Scientific Reports*, **15**, Article No. 34813. https://doi.org/10.1038/s41598-025-18807-z

[21] Zhang, Y., Zhang, P. and Wu, H. (2024) Enhancer-MDLF: A Novel Deep Learning Framework for Identifying Cell-Specific Enhancers. *Briefings in Bioinformatics*, **25**, bbae083. https://doi.org/10.1093/bib/bbae083