基于 RetESP 的酶–底物配对预测研究

摘要

蛋白质-小分子相互作用是药物发现和酶工程的核心任务。传统方法鉴于酶种类和底物结构的多样性，实验筛选耗时，成本高且泛化能力有限而受限。本研究提出 RetESP，一种基于 Retentive Network（RetNet）的深度学习框架，结合MolFormer或图神经网络（GCN+EGNNC）以及投影头，高效预测酶-底物配对。即使我们不得不压缩隐藏层的大小，但RetESP仍旧在不同的数据集上表现出色，同时计算成本较低。本研究为多模态生物信息学任务提供了灵活的解决方案，并为酶-底物配对预测提供了新思路。

关键词：蛋白质-小分子相互作用；深度学习；高维嵌入

目录

表格与插图清单

表索引

图索引

基于 RetESP 的酶–底物配对预测研究

一、介绍

酶作为生物体系中催化化学反应的核心分子，其高效、特异的催化功能对于细胞代谢、信号传导以及能量转换均至关重要。蛋白质-分子相互作用预测是药物发现、靶点识别和生物催化研究的核心任务，在加速药物筛选、优化生物工程过程及揭示分子功能机制等方面具有重要应用价值[1]。传统实验方法在揭示酶－底物相互作用方面发挥了不可替代的作用。如分子对接和基于规则的模型依赖复杂的物理模拟和手工设计的特征，不仅计算开销大，且对复杂分子系统或未见蛋白质的泛化能力有限[2]，这大大限制了底物设计的效率和泛化能力。但多数酶具有一定的底物混杂性和独特的激活/识别机制，且不同底物的催化效率各不相同。此外，实验验证的蛋白质-分子相互作用数据（如 BindingDB 数据库[3][1]）覆盖范围有限，仅包含少量已知相互作用对，难以满足高通量预测的需求。如半胱氨酸蛋白酶家族要求肽底物应该具有一定规则的识别氨基酸序列：如 C60 的 Sortase A 亚族，能够识别多肽底物的 C 端 LPXTG 序列[22]，经常被用于体外和细胞内蛋白质修饰以及肽合成的连接。但体外分离的 LPXTG 五肽是否可作为有效的最短识别底物仍有待考察。且在肽底物的选用上，考虑氨基酸所处环境的电子结构，是否需要封端？使用什么封端？也是计算模拟或实验不可避免的问题。对这些问题都实验人员提出了较高的要求。面对日益庞大的酶种类和多样化的底物结构，实验筛选往往耗时且成本高昂。通过准确预测蛋白质与小分子的相互作用强度，研究人员能够显著降低高成本的实验验证需求。因此，开发高通量、精准的计算方法成为当前研究的热点方向。

近年来，深度学习技术的快速发展为蛋白质-分子相互作用预测提供了新的解决方案。蛋白质语言模型将整个蛋白质序列解释为一个句子，并将其组成部分氨基酸解释为单个单词。蛋白质序列被限制采用特定的三维结构，这些结构经过优化，以完成特定功能。这些约束反映了自然语言处理中的语法和语义规则，采用自然语言处理的策略完成对自然科学的理解似乎是可行的。同时 ProtTrans 也

通过一系列主动学习，向我们证明蛋白质语言模型无需标签即可学习，且能够捕捉氨基酸的生物物理特征[16]。

基于序列的 Transformer 模型（如 ESM-2、MolFormer）通过在大规模蛋白质和分子数据集上预训练，学习到通用的序列表示，显著提升了特征提取的效率和预测性能[4][5]。同时，图神经网络通过显式建模分子的拓扑结构和化学键特征，在捕捉局部化学模式方面表现出色[6][7]。然而，这些方法仍面临若干挑战：1）Transformer 模型的计算复杂度随序列长度呈平方增长，处理长蛋白质序列时效率较低；2）图神经网络依赖手工设计的分子图特征，可能丢失复杂化学模式，且预处理复杂；3）多模态数据（蛋白质序列与分子表示）的有效融合仍需优化，以平衡模态特异性和交互信息捕捉；4）现有模型通常需要微调大型预训练模型或依赖额外数据集，导致计算资源需求高，限制了实际部署能力。

Retentive Network (RetNet) 作为一种高效序列建模架构被提出[8]。它能够通过多尺度保留机制替代 Transformer 的自注意力，显著降低了长序列处理的计算复杂度，同时保留了强大的序列建模能力。通过整合预训练的分子表示模型（如 MolFormer）或图神经网络（如 GCN+EGNNC）表示模块，似乎能够以较低的推理成本为蛋白质-分子相互作用预测提供了新的可能性。此外，利用投影头来对高维嵌入做进一步的精炼也能够提高多模态任务的预测精度[9]。

在此背景下，本研究提出 RetESP（Retentive Enhanced Sequence-Pair model），一个结合 RetNet、MolFormer 或 GCN+EGNNC 以及投影头的深度学习框架，用于高效预测蛋白质-分子相互作用。通过设计四种模型变体（GCN+EGNNC+RetNet、GCN+EGNNC+RetNet+ProjectHead、MolFormer+RetNet、MolFormer+RetNet+ProjectHead），我们探索了图神经网络、预训练 Transformer 和投影头嵌入精炼策略在多模态任务中的性能。RetESP 利用 RetNet 高效建模蛋白质序列，MolFormer 或 GCN+EGNNC 提取分子特征，投影头优化嵌入质量，旨在以较低的计算成本实现高精度的相互作用预测。本研究不仅为蛋白质-分子相互作用预测提供了灵活高效的解决方案，为酶－底物配对任务提供了新的解决思路和技术支撑。还为多模态生物信息学任务的模型设计提供了新的参考。

二、相关工作

早期研究主要基于分子对接模拟和手工特征工程。AutoDock Vina[12]通过基于物理知识计算结合自由能从而预测相互作用，但其计算复杂度随系统规模呈指数增长，且评分效果极其依赖于蛋白质或配体的其实评估构象。随着科技的不断发展，深度学习在蛋白质-配体相互作用预测中取得显著进展。FEATURE[13]框架通过定义空间化学特征描述符，构建随机森林分类器，完成对序列内容特征的整合，但经典的统计学习方法及其依赖与认为设置的领域知识特征。PUResNet 将稀疏表示技术应用于蛋白质结构，在以稀疏方式表示高维数据的领域中可以找到相似之处，从而执行更有效的潜在药物结合位点识别[19]。ACP-CLB 使用基于图形和统计特征的特征提取的 CNN 通道和采用 ESM-2 预训练模型进行更深层特征提取，完成对抗癌肽的高准确度预测[14][13]。ESP 通过改进的 ESM-1b 预训练模型来表示酶，D-MPNN 被用与为 GNN 提供有关小分子的信息[15]。

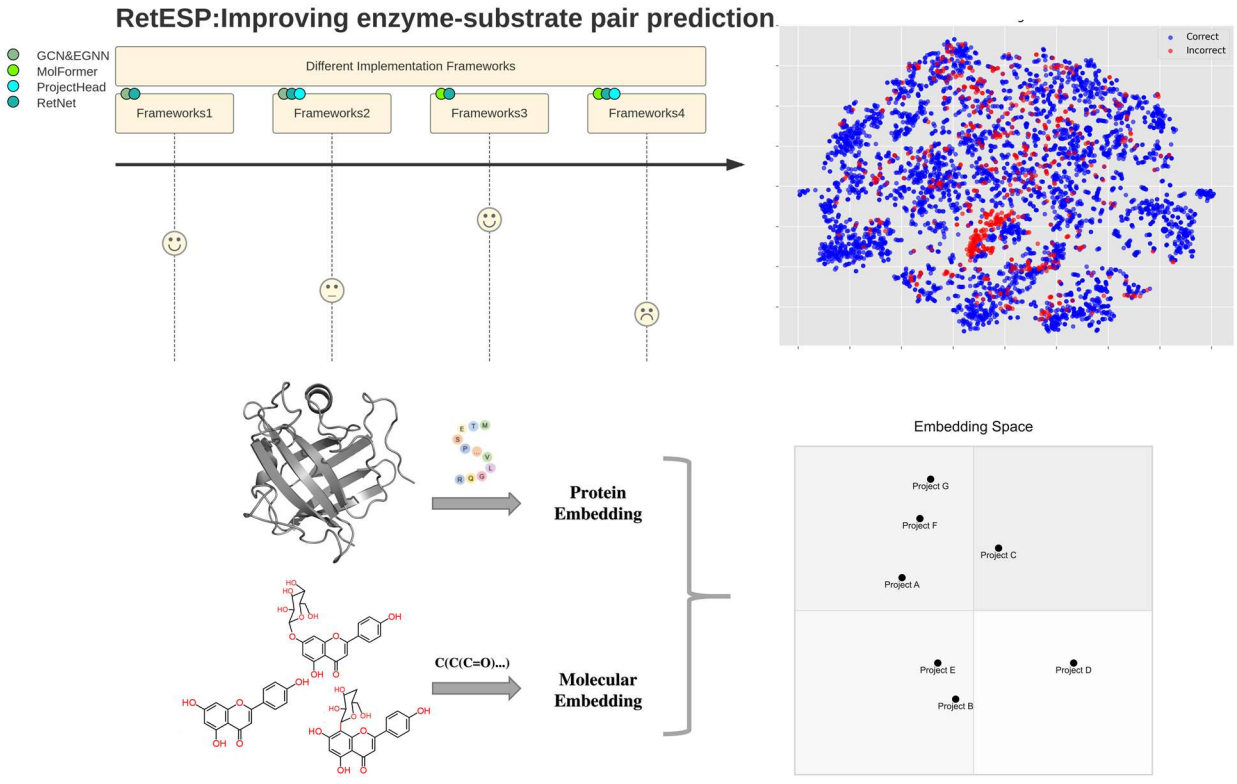

图 1 RetESP 项目缩略图示

FusionESP[18]提出分层融合策略：使用 ESM-2 提取蛋白质嵌入，MolFormer 提取小分子表示，跨模态注意力机制对齐特征空间，但其三阶段训练流程复杂

（预训练→微调→融合）。 DeepLPI，利用 1D CNN 和 biLSTM 提取蛋白质特征，但对长序列的处理效率较低[17][16]。

相较于 ESP，FusionESP 这些模型，在生成蛋白质嵌入时通过使用参数量庞大的预训练蛋白质语言模型，这对使用者的硬件设备提出了较高的要求，并且导致系统的超参数优化变得更加困难。我们预期使用尽可能简单的自然语言理解，实现低成本且高效高准确度的预测推理。

并且，在面对更复杂的问题时，BioStructNet 已经向我们证明，利用迁移学习迁移从较大数据集中学到的知识，并且通过收集小型数据集做微调可以进一步提高模型的泛化能力[20]。

三、方法

我们的目标是开发一种能够预测蛋白质酶-小分子/肽底物之间配对关系判断的机器学习模型，具体来说需要分别为酶和底物组织一个良好的蛋白质/化学高维嵌入，通过一定的组合策略将其映射到同一维低维空间以完成酶-底物之间的配对，余弦相似度被用来评估配对与否与置信度。

以拥有较低的推理成本，良好的性能以及可并行训练而著称的 Transformer 继任者 RetNet 被我们用于读取作为输入的经过特别编码的蛋白质序列列表，以提取蛋白质表示。为了确定合适的小分子表示表示。我们选择了两条路线：一是将小分子表示为非欧几里得数据表示--图：其中小分子的原子可以很自然地看作是个节点，与其相成化学键的原子被边连接，然后使用 GCN 和 EGNNC 分别提取具有不同侧重点的特征，EGNNC 是基于卷积的边特征增强图神经网络，卷积被用于扩展以处理多维边特征。二是选择 MoLFormer 用于分子的数值表示生成。我们使用了 MolFormer 在来自 ZINC/PubChem 中获取的约 12 亿训练数据的 10%进行训练的版本”MoLFormer-XL-both-10pct”，且不进行微调。该版本在”https://huggingface.co/ibm/MoLFormer-XL-both-10pct”上公布。

此外，我们还有两种映射的组合策略。正如图 2 中 A，B 中 BaseUnit 单元所示，对于 ProjP/M 映射出来的嵌入，直接进行组合并用余弦相似度评估二者之间的可配对性。ProjectHead 部分则是对 ProjP/M 映射出来的嵌入进行了进一步的处理，处理后的表示同样使用余弦相似度进行评估。至此，RetESP 拥有四个实验架构，分别是 RetNet+GCN&EGNNC，RetNet+GCN&EGNNC+ProjectHead，RetNet+MolFormer，RetNet+MolFormer+ProjectHead。下文中我们统一分别按序用 A,B,C,D 来简单标识这四种不同的实验架构。

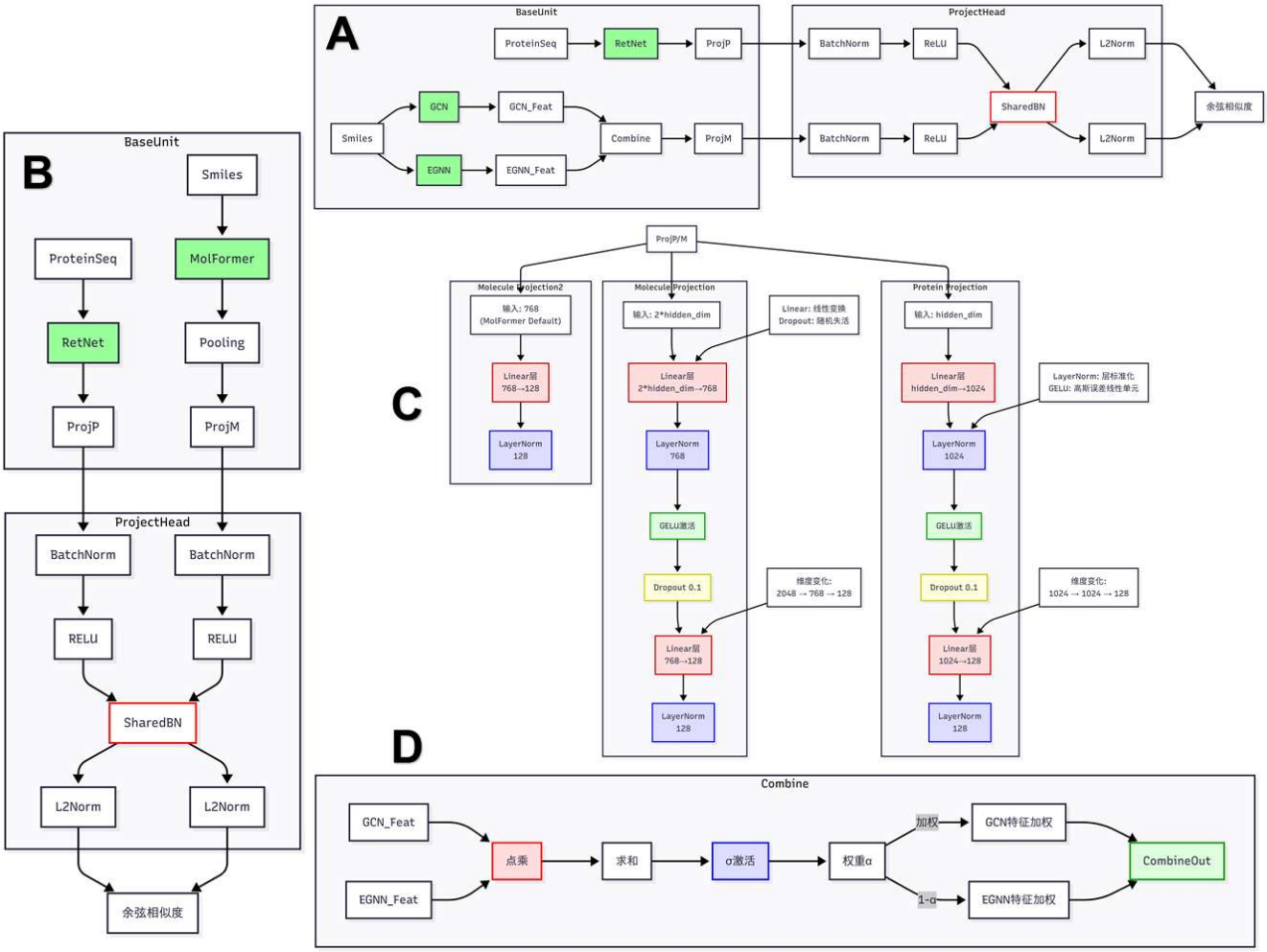

图 2 RetESP 模型框架

A 为对应的融合了几何表示的路线一。B 对应学习了大规模小分子表示的预训练模型路线二。C 对应 RetESP 中对 ProjP/ProjM 模块的具体实现，其中小分子的两种 Projection Layer 对应两种路线的 ProjM。D 对应路线一框架中的 combine 模块实现细节，用于融合 GCN 和 EGNNC 从小分子中的提取的特征。

四、实验

（一）数据集准备

为了确保与先前研究进行公平比较，本研究中使用的数据集均来自现有研究与个人扩展。我们选用了 Feiran 等人从 BRENDA 和 SABIO-RK 数据库提取的 17010 个配对数据集,考虑到部分数据存在 Kcat 值极小的情况，其酶催化反应生物学上不显著，所以我们将其值小于 0.0025 的条目的反应判断为酶与底物不配对。整理后我们将其命名为 DLkcat_ds[21]；Zhenjiao 等人提出的 FusionESP 中使用了来自 Zine15 和 PubChem 的已整合并公开的酶底物配对数据集，我们选用并命名为 FusionESP_ds。DLKcat_ds 与 FusionESP_ds 之间的配对数据存在一定的重叠。HumanDataset 是从其他研究者提供的实验配对情况整合版本。

为了减少数据集之间的重叠与影响，对数据集进行了清洗，以确保分析的准确性和可靠性。我们以后者为模板，将其内容整合到归并数据集中。并为其维护一个酶序列到底物 smiles 表示的字典，值为其可能的底物 smiles 集合。逐个读取另一个数据集中的每一个数据条目，通过动态规划计算条目的酶序列与字典各个键序列间的最长公共子序列（LCS），倘若 LCS 占两序列间较短序列的 80%长度，那么认为它们大部分重叠。对于大部分重叠或完全重复的酶序列，判断其 smiles 是否出现在所维护的 smiles 列表中，若不存在则将该数据条目完整纳入归并数据集里，存在则舍弃该序列并进入下一数据条目的判断；对于未检查出重叠或重复的酶序列，其数据条目完整纳入归并数据集里。所有数据集之间的合并都采用这种方法。

表 1 不同来源的数据集构成

	阳性数	阴性数	总数
DLKcat_ds	16577	433	17010
FusionESP_ds	18146	50575	68721
HumanDataset_ds	2836	2917	5753

（二）数据预处理

我们设计了一套预处理管道。对于第一条路线，分别对蛋白质氨基酸序列和化学分子 smiles 表示进行清洗、编码与图结构构建，以产生模型可直接消费的离散特征与图数据。氨基酸序列部分包括无效字符剔除，这一步去除了包含未知

或非标准氨基酸符号（如 X、B、U、O）的序列，以保证输入数据的一致性与可解释性。类似于 ProteinBERT 中只保留标准氨基酸及少量特殊符号的做法，有助于提升模型训练质量[9]；为了避免过长序列导致的计算开销过大，我们仅保留序列长度在 5 到 2000 之间的样本，这样能在一定程度上过滤极短序列带来的信息稀疏问题；并使用字典 SEQ_MAP 将 20 种标准氨基酸映射到 1 – 20 的整数空间，未在表中的字符映射为 0，以保持编码的稀疏性和可逆性整数映射；通过对每条通过过滤的序列，逐位替换为对应整数 ID，最终得到形状为 [L] 的一维整数向量，L 为序列长度；正如[11]中建议的那样。

对于小分子的表征，先确定 smiles 字符串的表示是否能够被解析为分子对象，若无法解析则丢弃该样本，以确保后续特征计算的有效性；同时检测芳香环中的原子，并将其符号转换为小写字母，以在原子类型编码时区分芳香与非芳香原子；通过获取每个原子的元素符号，并通过预定义的 MOL_MAP 字典映射为整数 ID，支持 C、H、O、N、P、S 等常见元素编码;为了构建原子间二值邻接矩阵 A ∈ {0,1}^{N×N}，表示原子间是否存在化学键；初始化与 A 同尺寸的浮点边特征矩阵 E，遍历分子图提取每条化学键的键级（单键、双键、芳香键等）并赋值给 E[i,j] 和 E[j,i]，以提供键类型信息给后续神经网络。最终将所有样本的最大序列长度与最大原子数记录下来，便于后续的批次填充；

相比之下，第二条路线由于成熟的小分子嵌入的预训练模型能够直接提供对小分子的表示，我们只对氨基酸序列进行了如上文的处理。所有成功预处理的样本按字典形式汇总，并连同全数据集中最大氨基酸序列长度 max_seq_len 与最大原子数 max_atoms 一并序列化至磁盘（pickle 格式），以便在训练时统一进行批次填充与掩码处理。具体来说，以 pickel 数据文件格式存储。在后续的训练过程中，针对预先保存的数据集，我们采用自定义的 LazyDataset 类按需加载样本，动态构造以下输入张量：

1）序列张量：将长度不等的氨基酸 ID 序列零填充到批内最大序列长度（max_seq_len）；

2）原子张量：将小分子中原子 ID 序列零填充到批内最大原子数（max_atoms）；倘若使用路径二则无原子张量产生。

3）邻接矩阵与键特征：分别填充到［max_atoms×max_atoms］的稠密矩阵。

（三）模型设计

为了节约训练成本，隐藏维度被设置成512。正如数据预处理所言，序列嵌入层利用将氨基酸ID映射到隐藏维度下的连续向量空间，并采用 Xavier 均匀初始化以保证初始权重的方差平衡，从而稳定深层网络的训练过程。在嵌入后，使用2层 RetNet 模块堆叠而成，每层包含4头多头自注意力；沿序列长度求平均嵌入，并在其后加上 LayerNorm 与 GELU 以加速收敛并捕捉序列全局上下文。分子嵌入层同样将原子类型ID转换为隐藏维度下的向量表示，并同样使用Xavier均匀初始化。在节点向量上堆叠2层 GCN，每一子层与邻接矩阵的点积被LeakyReLU激活。使用2层 EGNNC——基于与边特征矩阵的点积且Sigmoid进行门控激活得到其表示；对GCN和EGNN得到的表示，通过节点级点积与Sigmoid生成注意力门控，对两种输出分别加权求和，实现两种图编码方式的自适应融合。接着BatchNorm归一化和LeakyReLU激活以轻量方式融入边权信息，提升对图结构的表达能力。这也正是图2所介绍的。

（四）模型训练

对由DLKcat_ds和FusionESP_ds组合成的非冗余数据集，以0.8：0.1：0.1的比例进行训练集：验证集：测试集的划分。优化器选用 AdamW，结合权重衰减抑制过拟合，初始学习率为0.001，权重衰减系数由为0.0005。使用 MSE 损失函数进行损失计算，MSE 损失适用于回归任务，鼓励预测的余弦相似度接近真实标签（1 或 0），便于优化嵌入空间的分布，具体公式如下：

$$Loss = \frac{1}{n}\sum_{i=1}^{N}\left(Cos\,Sim_i - Label_i\right)^2 \qquad \text{公式(1)}$$

其中n是数据点的数量;Cos_Sim 是单对中酶和底物的 128 维向量的余弦相似性度量;Label是真正的标签，阳性酶-底物对的值为1，阴性对为0。在每个训练迭代中，我们使用 Cosine Annealing 调度（周期 T_{max}=10 轮）动态调整学习率，以平滑衰减至较小值，促进模型收敛。同时在反向传播前对所有可学习参数执行梯度裁剪（clip norm ≤ 1.0），防止梯度爆炸并提高训练稳定性。在每个 epoch 结束后，使用验证集对当前模型进行评估，计算以下指标：平均损

失（MSE）、精度（Accuracy）、加权敏感性/特异性与二分类平衡准确率（Balanced Accuracy）、以及 AUC（ROC 曲线下面积）。当验证集 AUC 超越历史最优时，即刻保存检查点（best_model.pth），保存内容包括当前 epoch、模型权重、优化器状态和当前最优 AUC；否则，早停计数器加一，当连续 patience 个 epoch 验证 AUC 无提升时触发早停，终止训练。在这里，我们的 patience 被设置为 100。四种不同架构的模型在训练集上以 64 的总批量大小训练了 100 个 epoch。

（五）训练结果

表 2 四种实验架构 HumanDataset_ds 上的性能比较

	ACC	BACC	AUC	F1	MCC	TN	FP	FN	TP
A	0.7858	0.7704	0.8498	0.7256	0.5559	4387	677	1192	2471
B	0.7796	0.7621	0.8362	0.7131	0.5429	4414	650	1273	2390
C	0.8109	0.7894	0.8704	0.7443	0.6092	4797	442	1251	2464
D	0.7954	0.7893	0.8653	0.7535	0.5786	4322	917	915	2800

表 3 四中实验架构在测试集上的性能比较

	ACC	BACC	AUC	F1	MCC	TN	FP	FN	TP
A	0.8871	0.8880	0.9559	0.8856	0.7768	2442	200	413	2373
B	0.8681	0.8696	0.9433	0.8633	0.7425	2451	191	525	2261
C	0.8793	0.8796	0.9482	0.8821	0.7604	2415	432	249	2547
D	0.8767	0.8765	0.9422	0.8732	0.7537	2550	297	399	2397

RetESP 能够返回蛋白质/小分子高维度嵌入后计算得到的余弦相似性分数，不仅能够根据判断阈值确定酶-小分子对是否配对，还能解释模型对其预测的置信度。我们将 0.5 设置为判断阈值，其中，如果分数高于 0.5，则酶和小分子之间的余弦相似性分数被预测为阳性，否则为阴性。为了提供更详细的预测准确性评估，图 3 和图 4 分别显示了我们在测试集和 HumanDataset_ds 中酶-底物对分数的正确预测（蓝色）和错误（红色）预测的分布。堆叠直方图条形显示真实预测（蓝色）和错误预测（红色）的预测分数分布。

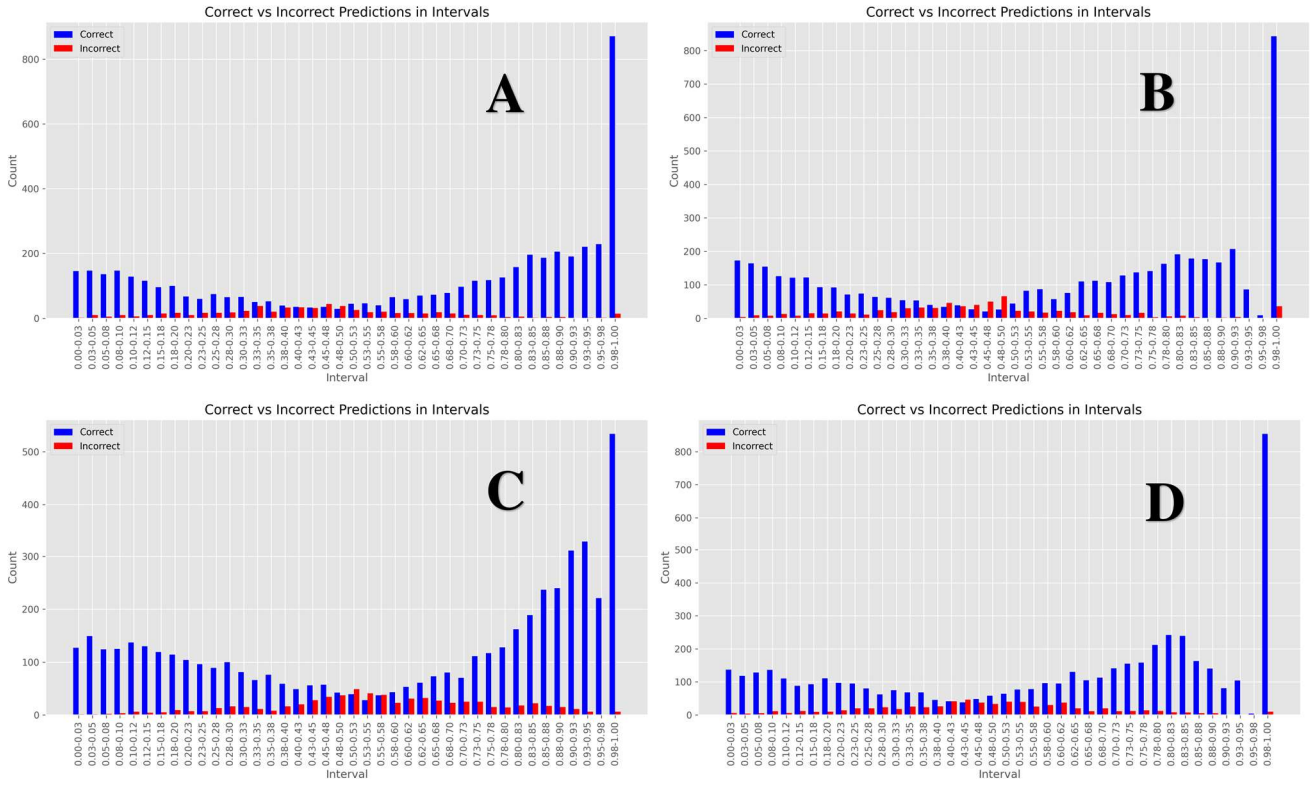

图 3 四种不同实验架构的 RetESP 在测试集上的表现

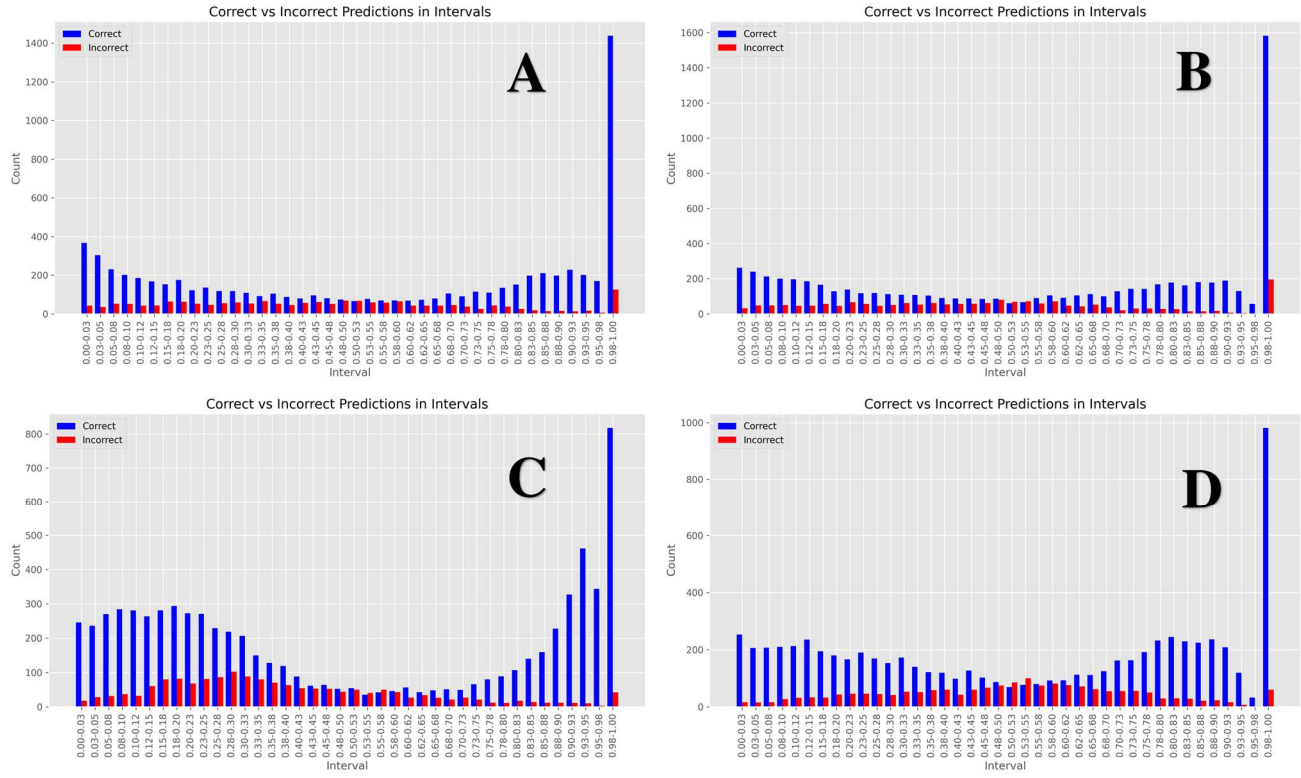

图 4 四种不同实验架构的 RetESP 在 HumanDataset_ds 上的表现

具有耦合 ProjectHead 的模型预测分数更加均匀，而不带 ProjectHead 的模型的正确预测分数更接近于 1，表明带有耦合 ProjectHead 的 RetNet 预测至信度相对较高，且具有更多的普适性。t-SNE 可视化显示带有 ProjectHead 的架构的阳性阴性样本嵌入在高维空间中分离更明显，表明投影头在一定程度上改善了特征表示。

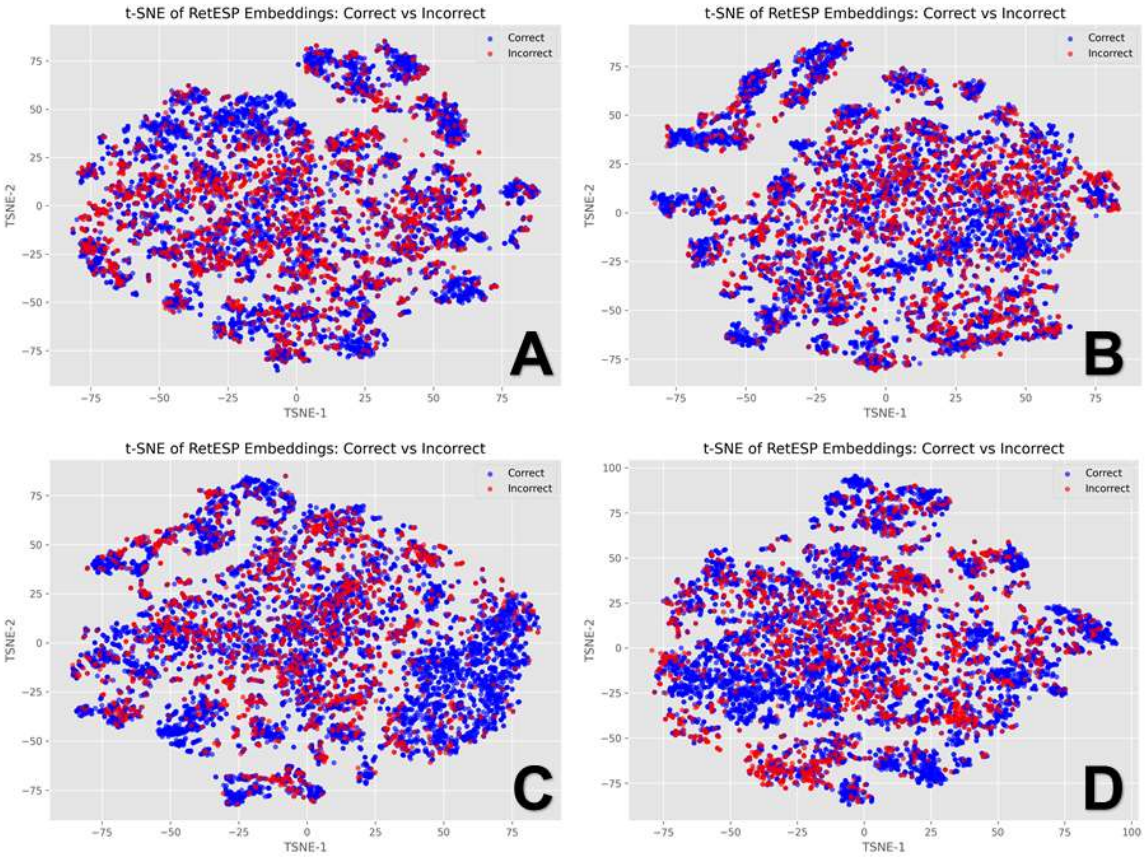

图 5 t-SNE 可视化四种不同实验架构的 RetESP 在测试集的性能

校准曲线将连续的预测值和真实值数据进行离散化，常用于评估模型预测概率与实际发生频率的一致性。在这里，我们对这四个实验架构都分别绘制的校准曲线，如图 7 所示。

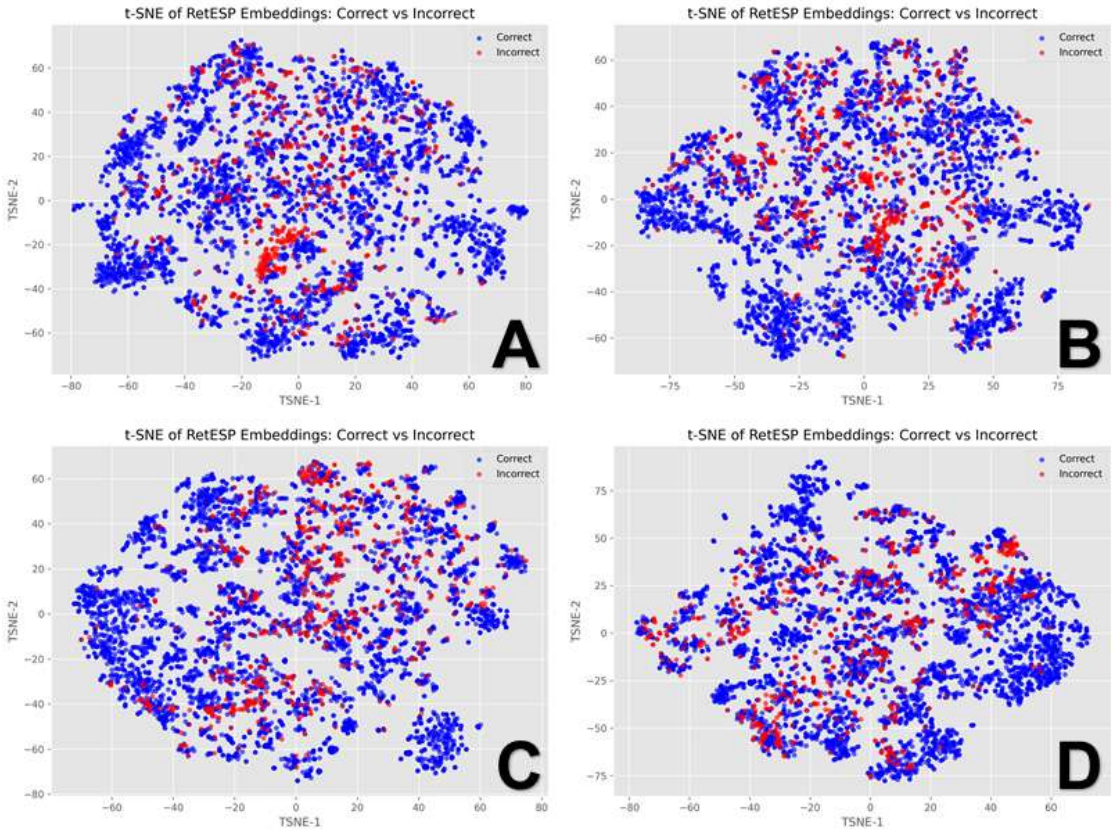

图 6 t-SNE 可视化四种不同实验架构的 RetESP 在 HumanDataset_ds 的性能

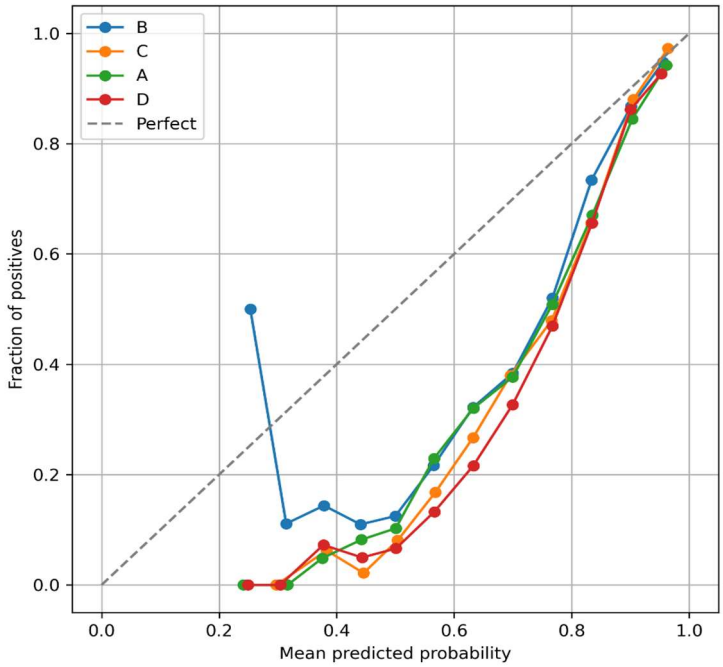

图 7 四种不同实验架构的 RetESP 在 HumanDataset_ds 中的校准曲线

（六）非耦合投影头有助于改善特征空间下的映射

至今为止，我们在设计 ProjectHead 时，我们直接将其和 BaseUnit 融合在一起，BaseUnit 给出的蛋白质/小分子编码被传入到新的单一全连接层进行融合，我们称之为单 Refine 层。ProjectHead 跟随着优化器一同训练。为了进一步讨论投影头的作用，我们设计了非耦合投影头，具体来说，就是将不带有 ProjectHead 的两个实验架构 A 和 C，分别加载了其约在 70 个 epoch 左右时保存的最佳检查点，这一步是为了读取成熟的蛋白质/小分子高维嵌入。并且冻结原有的层。为他们加上了 ProjectHead，与 B,D 架构不同的是，这次优化器只会优化这 ProjectHead 单元的参数。除此之外，对 A,C 这两个实验架构，添加的 ProjectHead 还有无额外映射或额外添加了一层同纬度下的全连接映射的差异，成为双 Refine 层；然后分别额外训练了 100 个 epoch。训练曲线见附录图 S2。投影头组合情况见图 9，至此对于每一条路径，我们分别得到了单 Refine 层/双 Refine 层的架构，按次序命名为 E,F,G,H。

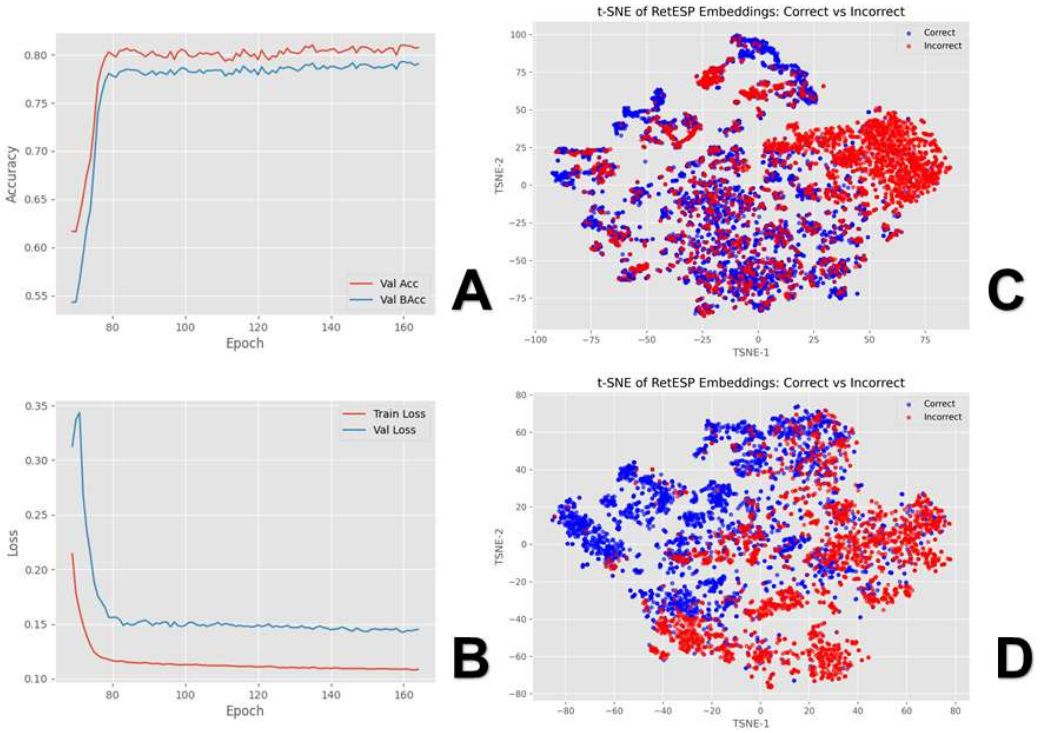

图 8 带有非耦合投影头架构训练结果

A 为该架构训练时在验证集上的 Acc，Bcc 指标；B 为损失的变化曲线；C 为该架构在测试集上预测结果的 t-SNE 降维下的可视化的性能比较。D 为该架构在 HumanDataset_ds 上预测结果的 t-SNE 降维下的可视化的性能比较。

实际上，只有 C 的额外添加全连接映射的 ProjectHead 的架构取得了在验证集上更优秀的结果。如图 8 所示。通过观察图 8（c 和 d），我们可以清楚看到，相比于图 X 和图 Y，带有非耦合投影头的架构明显地将阳性和阴性的预测划分成了两个区域，非耦合投影头实验表明，单独训练投影头可进一步优化嵌入空间，尤其在变体 C 的添加全连接映射的双 Refine-ProjectHead 的架构，添加全连接层的投影头使正负样本分离更明显。但是从训练结果我们可以看出，不论耦合还是非耦合的 ProjectHead，给模型整体带来的提升及其有限，模型的性能取决于 BaseUnit 提供的蛋白质/小分子表示的质量。

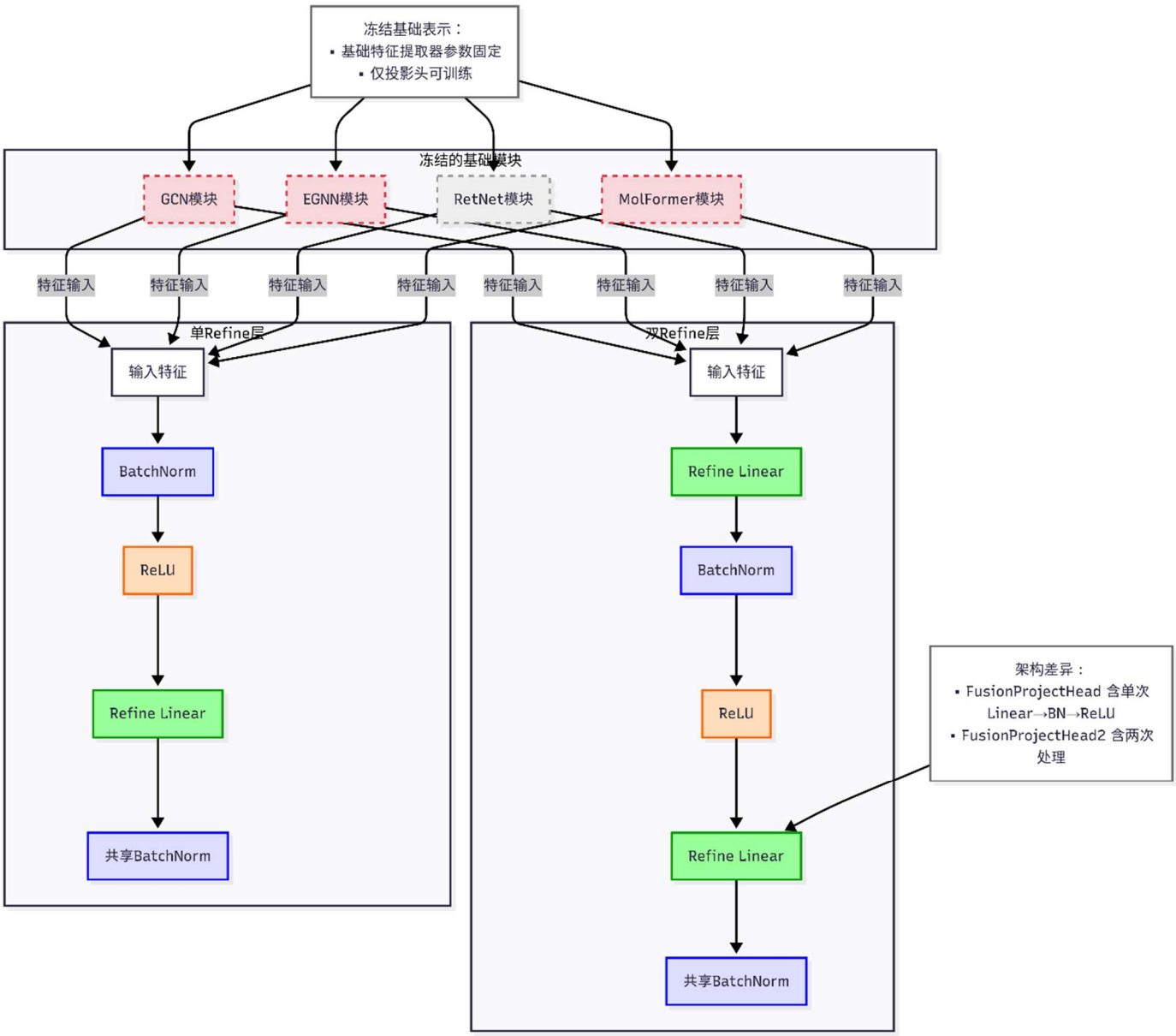

图 9 具有不同投影头的实验框架设计

五、结论

本研究提出 RetESP，一种高效的酶-底物配对预测模型，支持 MolFormer 与几何图网络双路径，适应不同场景。成功融合了蛋白质序列的自然语言处理与小分子的几何表示，兼顾了计算效率和预测精度。RetESP 利用 Retentive Network（RetNet）的高效序列建模能力，结合 MolFormer 或图神经网络（GCN+EGNNC）提取的分子特征，通过投影头进一步优化嵌入空间，实现了对酶-底物相互作用的高精度预测。实验结果表明，RetESP 在测试集以及 HumanDataset_ds 上均取得了不错的性能，尤其是在投影头的辅助下，模型的预测置信度和泛化能力得到一定的提升。尽管由于计算资源限制，隐藏维度被设定为 512，可能限制了模型对复杂相互作用的建模能力，RetESP 仍能捕获良好的高维嵌入表示，并实现不错的预测效果。未来研究可通过增加隐藏维度、整合更多数据集或探索对比损失等其他损失函数，进一步提升模型性能。RetESP 的设计理念和实验结果为多模态生物信息学任务提供了新的参考，尤其在药物发现和酶工程等领域具有广泛的应用潜力。

项目源代码请见：https://github.com/AmiHaruka/RetESP。

参考文献

[1] D. Ashwin, M. Cole, T. John J, and C. Jianlin, “Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions,” 2021, doi: https://doi.org/10.1093/bib/bbab476.

[2] T. Oleg and O. Arthur J., “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading,” 2009, doi:https://doi.org/10.1002/jcc.21334.

[3] L. T., L. Y., W. X., J. R. N., and G. M. K., “BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities,” 2006, doi: https://doi.org/10.1093/nar/gkl999.

[4] R. Jerret, B. Brian, C. Vijil, P. Inkit, M. Youssef, and D. Payel, “Large-scale chemical language representations capture molecular structure and properties,” 2022, doi: https://doi.org/10.1038/s42256-022-00580-7.

[5] L. Zeming, A. Halil, R. Roshan, H. Brian, Z. Zhongkai, and L. Wenting, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” 2023, doi: https://doi.org/10.1126/science.ade2574.

[6] K. Thomas N. and W. Max, “Semi-Supervised Classification with Graph Convolutional Networks.” 2016. doi: https://doi.org/10.48550/arxiv.1609.02907.

[7] G. Satorras, Victor, E. Hoogeboom, and M. Welling, “E(n) Equivariant Graph Neural Networks.” 2021. doi: https://doi.org/10.48550/arxiv.2102.09844.

[8] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, and J. Xue, “Retentive Network: A Successor to Transformer for Large Language Models.” 2023. doi: https://doi.org/10.48550/arxiv.2307.08621.

[9] C. Ting, K. Simon, N. Mohammad, and H. Geoffrey, “A Simple Framework for Contrastive Learning of Visual Representations.” 2020. doi: https://doi.org/10.48550/arxiv.2002.05709.

[10] B. Nadav, O. Dan, P. Yam, R. Nadav, and L. Michal, “ProteinBERT: a universal deep-learning model of protein sequence and function,” 2022, doi: https://doi.org/10.1093/bioinformatics/btac020.

[11] T. Farzana, H. Sultana Umme, M. Tanjim, N. Lutfun, H. Mohammad Shahadat, and A. Karl, “Protein Sequence Classification Through Deep Learning and Encoding Strategies,” 2024, doi: https://doi.org/10.1016/j.procs.2024.06.106.

[12] C. Sandro, F. Stefano, P. Alex L, H. Rodney, G. David S, and O. Arthur J, “Virtual screening with AutoDock: theory and practice,” 2010, doi: https://doi.org/10.1517/17460441.2010.484460.

[13] W. Yaoxin, X. Yingjie, Y. Zhenyu, L. Xiaoqing, and D. Qi, “Using Recursive Feature Selection with Random Forest to Improve Protein Structural Class Prediction for Low-Similarity Sequences,” 2021, doi: https://doi.org/10.1155/2021/5529389.

[14] G. Aoyun, L. Zhenjie, L. Aohan, Z. Zilong, Z. Quan, and W. Leyi, “ACP-CLB: An Anticancer Peptide Prediction Model Based on Multichannel Discriminative Processing and Integration of Large Pretrained Protein Language Models,” 2025, doi: https://doi.org/10.1021/acs.jcim.4c02072.

[15] K. Alexander, R. Sahasra, E. Martin K. M., and L. Martin J., “A general model to predict small molecule substrates of enzymes based on machine and deep learning,” 2023, doi: https://doi.org/10.1038/s41467-023-38347-2.

[16] E. Ahmed, H. Michael, D. Christian, R. Ghalia, W. Yu, and J. Llion, “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning,” 2021, doi: https://doi.org/10.1109/tpami.2021.3095381.

[17] Z. Wang, “A new paradigm for applying deep learning to protein–ligand interaction prediction,” Briefings in Bioinformatics, Apr. 2024.

[18] D. Zhenjiao, F. Weimin, G. Xiaolong, C. Doina, and L. Yonghui, “FusionESP: Improved Enzyme–Substrate Pair Prediction by Fusing Protein and Chemical Knowledge,” 2025, doi: https://doi.org/10.1021/acs.jcim.4c02357.

[19] K. Jeevan, T. Hilal, and C. Kil To, “PUResNet: prediction of protein-ligand binding sites using deep residual neural network,” 2021, doi: https://doi.org/10.1186/s13321-021-00547-7.

[20] A. Ali M., A. Lina, and H. Mats, “Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants?,” 2009, doi: https://doi.org/10.3368/le.86.1.79.

[21] L. Feiran, Y. Le, L. Hongzhong, L. Gang, C. Yu, and E. Martin K. M., “Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction,” 2022, doi: https://doi.org/10.1038/s41929-022-00798-z.

[22] B. Tora, P. Vijaykumar S., C. Devapriya, and R. Rajendra P., “Sorting of LPXTG Peptides by Archetypal Sortase A: Role of Invariant Substrate Residues in Modulating the Enzyme Dynamics and

Conformational Signature of a Productive Substrate," 2014, doi: https://doi.org/10.1021/bi4016023.

附录

1. RetESP 四种实验架构的训练情况

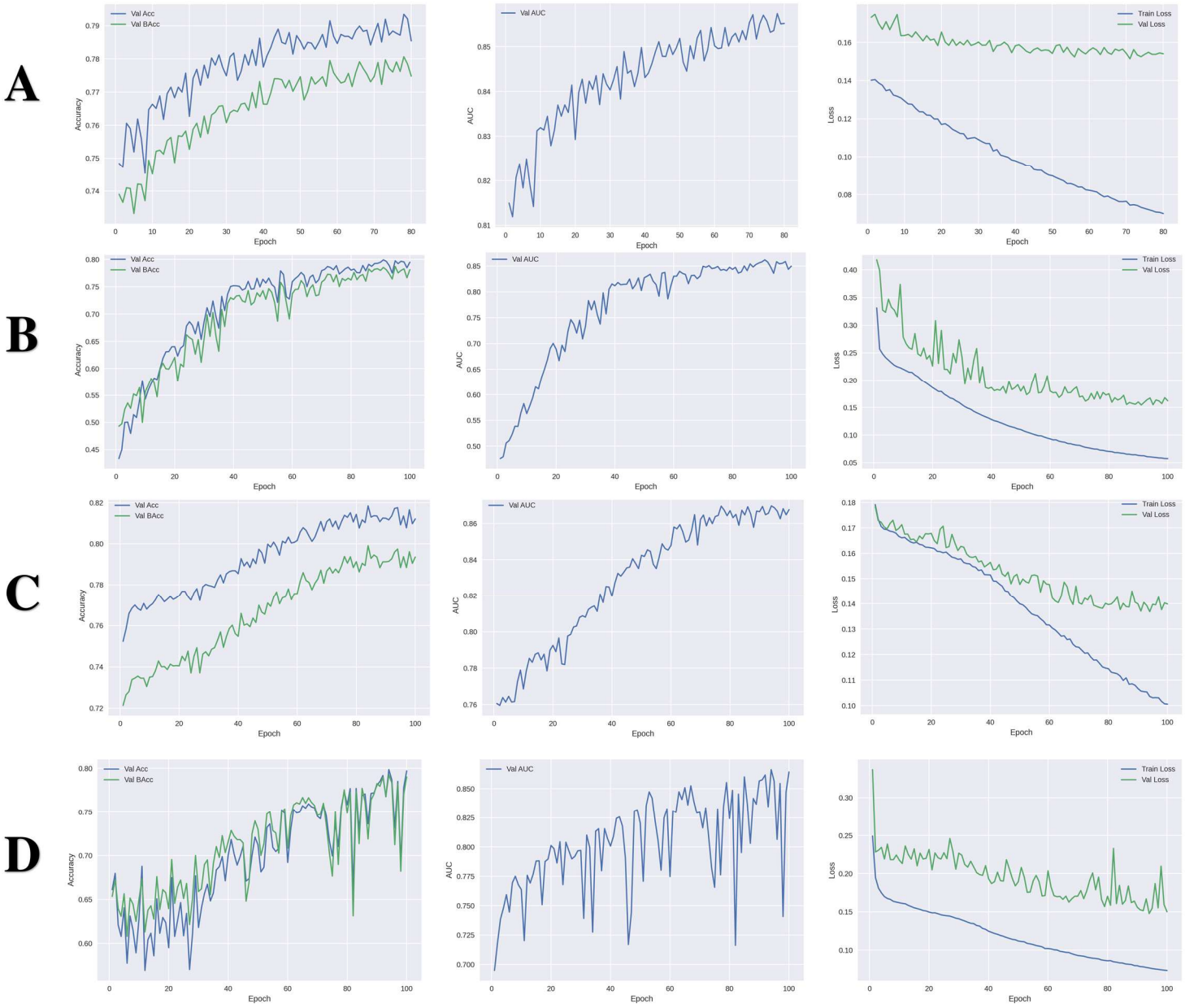

图 S1 具有不同投影头的实验框架训练曲线

A 为 GCN&EGNNC+RetNet；B 为 GCN&EGNNC+RetNet+单 Refine-ProjectHead；C 为 MolFormer+RetNet；D 为 MolFormer+RetNet+单 Refine-ProjectHead；指标分别为验证集 Acc&验证集 BAcc，验证集 AUC，训练集 Loss&验证集 Loss。

2. RetESP 四种不同投影头实验架构的训练情况

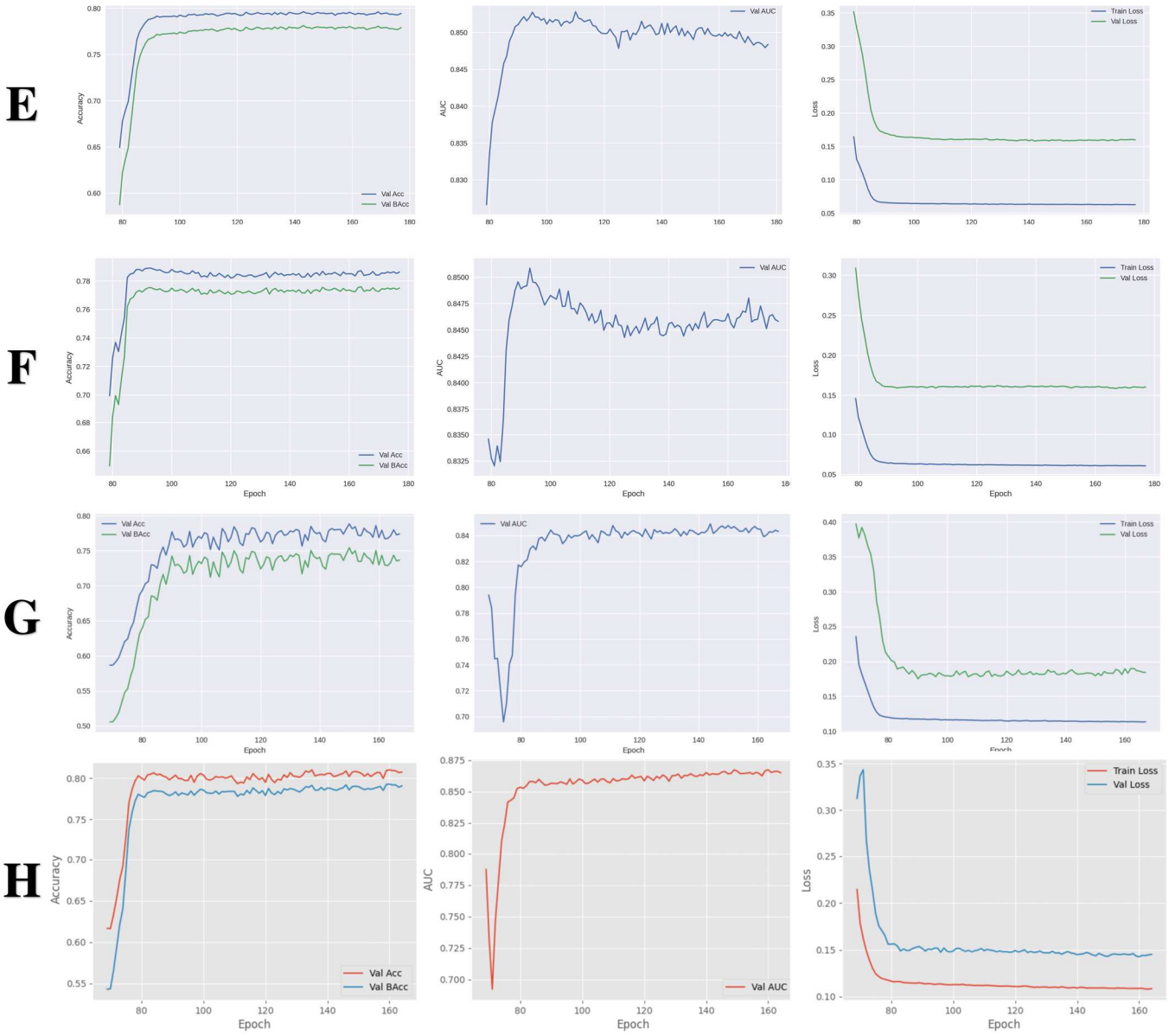

图 S2 具有不同投影头的实验框架训练曲线

E 为 GCN&EGNNC+RetNet+单 Refine-ProjectHead；F 为 GCN&EGNNC+RetNet+双 Refine-ProjectHead；G 为 MolFormer+RetNet+单 Refine-ProjectHead；H 为 MolFormer+RetNet+双 Refine-ProjectHead；在这里，BaseUnit 的参数都被冻结。指标分别为验证集 Acc&验证集 BAcc，验证集 AUC，训练集 Loss&验证集 Loss。