天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于多模态信息融合和图卷积网络的酶催化效率预测
作者：窦壮壮，王林
DOI：10.13364/j.issn.1672-6510.20240179
收稿日期：2024-09-08
网络首发日期：2025-06-23
引用格式：窦壮壮，王林. 基于多模态信息融合和图卷积网络的酶催化效率预测[J/OL]. 天津科技大学学报. https://doi.org/10.13364/j.issn.1672-6510.20240179

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2025-06-23 10:09:01　网络首发地址：https://link.cnki.net/urlid/12.1355.n.20250619.1436.005

2025 年　天津科技大学学报 Journal of Tianjin University of Science and Technology　2025

DOI：10.13364/j.issn.1672-6510.20240179

基于多模态信息融合和图卷积网络的酶催化效率预测

窦壮壮，王　林

(天津科技大学人工智能学院，天津 300457)

摘　要：酶周转数k_{cat}是反映酶催化效率的重要指标，在蛋白质工程以及合成生物学等多个领域中都发挥着重要作用。然而，传统实验测量方法耗时长且成本高昂。此外，即使在底物和酶完全一致的情况下，微小的实验条件差异也能导致测量结果出现波动。本文提出酶催化效率预测模型，通过整合化合物分子图、SMILES 序列、蛋白质序列、蛋白质分子 3D 结构以及反应温度进行深度跨模态特征学习，实现对k_{cat}值的精确预测。该模型使用了图卷积网络，从化合物分子图和蛋白质分子 3D 结构中提取特征。利用 Prot_Bert 和 SMILES Transformer 两个预训练模型，对蛋白质和化合物进行深层特征提取，进一步提升模型的多模态数据处理能力和k_{cat}值预测精度。

关键词：生物分子多模态信息；酶周转数；蛋白质分子 3D 图；Prot_Bert；SMILES Transformer；图卷积网络

中图分类号：TP399　　**文献标志码：**A　　**文章编号：**1672-6510 (0000)00-0000-00

Prediction of Enzyme Catalytic Efficiency Based on Multimodal Information Fusion and Graph Convolutional Networks

DOU Zhuangzhuang，WANG Lin

(College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China)

Abstract: Enzyme turnover number k_{cat} is an important indicator of enzyme catalytic efficiency and plays an important role in many fields such as protein engineering and synthetic biology. However, traditional experimental measurement methods are time-consuming and costly. In addition, even when the substrate and enzyme are exactly the same, slight differences in experimental conditions can cause fluctuations in measurement results. This paper proposes an enzyme catalytic efficiency prediction model that integrates compound molecular graphs, SMILES sequences, protein sequences, protein molecular 3D structures, and reaction temperatures for deep cross-modal feature learning to achieve accurate prediction of k_{cat} values. The model uses a graph convolutional network to extract features from compound molecular graphs and protein molecular 3D structures. Using two pre-trained models, Prot_Bert and SMILES Transformer, deep feature extraction of proteins and compounds is performed, further improving the model's multimodal data processing capabilities and k_{cat} value prediction accuracy.

Key words: biomolecular multimodal information; enzyme turnover number; protein 3D structure; Prot_Bert; SMILES Transformer；graph convolutional network

酶周转数(k_{cat})是衡量酶催化效率的重要指标，通　　常表示酶每秒钟转化底物的数量。它是酶催化速率常

收稿日期：2024-09-08；**修回日期：**2025-05-10

基金项目：天津市企业科技特派员项目(20YDTPJC00560)

作者简介：窦壮壮（2000—），男，山东人，硕士研究生；**通信作者：**王　林，副教授，linwang@tust.edu.cn

数的常见表达形式。酶周转数反映酶分子在特定条件下对底物的处理速度，是评估酶活性和效率的关键参数之一。在生物学研究、医药领域和工业生产中，准确预测酶周转数对于理解酶的功能、优化酶工程和设计新的生物催化剂具有重要意义。然而，传统的化学实验方法测定 k_{cat} 不仅耗时而且成本昂贵，极大地阻碍了其在大规模应用中的效率和可行性。因此，需要开发一种高效准确的计算模型预测酶的 k_{cat}，以加快酶功能研究和工业应用的进展。

已有研究主要集中在从底物结构和蛋白质序列出发进行高通量 k_{cat} 预测。Li 等[1]提出基于深度学习的 k_{cat} 预测模型(DLKcat)，通过图卷积网络(GNN)和卷积神经网络(CNN)分别提取化合物分子图以及蛋白质序列特征。虽然基本实现了高通量 k_{cat} 预测，但该方法未充分考虑温度、pH 等实验条件对 k_{cat} 值的影响，且方法相对单一，导致模型预测结果的准确性不足。Qiu 等[2]提出基于深度学习的温度依赖性酶周转率预测模型(DLTKcat)，利用图注意力网络(GAN)和卷积神经网络(CNN)分别提取化合物和蛋白质序列特征，并且考虑温度对 k_{cat} 值的影响。然而，与 DLKcat 类似，该方法也局限于仅通过底物结构和蛋白质序列进行高通量预测，且目前实验测得的具有 k_{cat} 的底物-酶的条目较少，经过筛选后所可用于训练的数据更少，因此在训练数据中不足以支撑模型去充分学习底物以及蛋白质特征。

为了解决上述问题，本文提出基于多模态信息融合的 k_{cat} 值预测模型 Multi-Modal-Temperature Kcat (MMTKcat)。该模型结合了化合物分子图、SMILES 序列、蛋白质序列、蛋白质分子 3D 结构信息以及温度，利用图卷积网络(GCN)[3]以及 Prot_Bert[4]和 SMILES Transformer[5]两个预训练模型，能够有效地从多角度提取化合物以及蛋白质特征，缓解训练数据量不足的问题并提高预测的准确性。

1 数据集

数据集均来源于公开数据库 BRENDA[6]和 SABIO-RK[7]。通过数据库提供的应用程序接口(API)批量获取数据，其中 SABIO-RK 提供酶学委员会号码(EC 号)、底物名称、生物名称、蛋白质标识符(UniProt ID)、酶类型、温度和 k_{cat}。通过 BRENDApyrser 工具从 BRENDA 数据库批量提取相关数据。之后，通过查询 PubChem 化合物数据库，批量获取底物的 SMILES 字符串。所有蛋白质的氨基酸序列基于数据库中提取的 UniProt ID，通过 UniProt 数据库[8]进行查询，所有的蛋白质序列均可以直接通过 UniProt 映射批量获取。

在获取 SMILES 序列和氨基酸序列后，将两个数据集中的数据合并筛选，去除所有具有相同 SMILES 序列、氨基酸序列、温度和 k_{cat} 的重复数据。对于具有相同 SMILES 序列、氨基酸序列和温度但 k_{cat} 不同的条目，仅保留 k_{cat} 最大的条目。这种方法参考了 Li 等[1]和 Qiu 等[2]的研究方法。最终保留 16249 条底物-酶-温度条目。根据氨基酸序列是否在自然状态下存在，又可将酶分为野生型酶(WT)和突变型酶(MUT)，数量如图 1 所示。

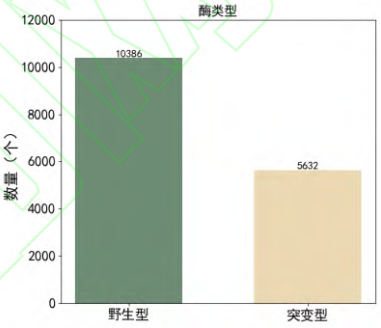


图 1 野生型酶、突变型酶数量

Fig. 1 Number of Wild-type Enzymes and Mutant Enzymes

为进一步提升数据集的完整性，利用由 Google DeepMind 开发的 AlphaFold 人工智能系统[9]，该系统可以根据蛋白质的氨基酸序列预测该蛋白质的三维结构。AlphaFold 通常能够达到与实验方法测得结果相媲美的精度。通过 AlphaFold 的 API 接口批量获取 4012 份蛋白质的三维结构文件，剔除极少部分无法获取结构的条目，最终本实验数据集中总计保留了 16018 条有效条目，且每条数据均包含 EC 号、底物名称、UniProt ID、k_{cat}、底物的 SMILES 字符串、蛋白质的氨基酸序列、蛋白质分子 3D 结构和温度。

2 构建多模态信息融合的学习模型

2.1 化合物表示

2.1.1 分子指纹和邻接矩阵生成

利用 RDKit 工具库对化合物的 SMILES 序列进行处理并生成分子图。首先，读取每个化合物的 SMILES 字符串，利用 RDKit 库[10]将 SMILES 字符串转换为分子对象并提取原子的符号和芳香性信息。随后，创建关键

字典用于存储每个原子的关键类型信息。基于此信息以指定的半径（r-radius）提取分子指纹并生成表示分子结构的向量。分子指纹提取过程根据指定的半径迭代更新节点的表示，在每次迭代中，节点的表示会结合来自其相邻节点的信息，从而得到包含指定半径范围内结构信息的指纹向量。最后，生成分子的邻接矩阵表示原子之间的连接关系，为后续利用图卷积网络进行化合物特征提取提供便利。

2.1.2　SMILES Transformer 模型

SMILES Transformer 模型被用于化合物 SMILES 字符串特征提取部分，与 GCN 提取的化合物分子图特征共同组成底物的特征表示，用于后续的深度学习模型。SMILES Transformer 是一种专门用来处理化学分子 SMILES 字符串的预训练模型。该模型在从 ChEMBL24 数据库中随机抽取的 861,000 个未标记 SMILES 字符串上进行无监督预训练，学习 SMILES 序列的上下文相关表示，并生成高维的分子指纹。这些分子指纹不仅可以捕捉分子之间的结构信息，还可以解决本实验在训练过程中训练数据不足的问题。

2.2　蛋白质表示

2.2.1　Prot_Bert 模型

蛋白质是生物体内复杂的分子结构，由 20 多种不同的氨基酸组成，且这些氨基酸可以以几乎无限的方式排列组合，这就导致了蛋白质种类的多样性和复杂性。因为本实验选取的 4012 种蛋白质序列不足以支撑模型充分学习和泛化蛋白质的多样性和复杂性，所以为了能够精准高效的提取蛋白序列特征，引入 Prot_Bert 模型。该模型是基于 BERT(bidirectional encoder representations from Transformers)架构的预训练模型，专门用于蛋白质序列分析。Prot_Bert 模型通过在蛋白质数据库 UniRef100[11]上进行自监督学习，能够精准捕捉蛋白质序列中的复杂生物物理特性和进化信息。其中，UniRef100 是由 2.17 亿个蛋白质序列组成的数据集，为模型的预训练提供了丰富的数据基础。Prot_Bert 模型采用自监督学习方法在大规模未标记蛋白质序列训练，使其能够自动从丰富的序列数据中学习，无需手动注释训练数据。利用 BERT 架构，Prot_Bert 可以同时从蛋白质序列的上下文中捕捉前后信息。这种双向表示对于识别蛋白质序列内的依赖关系和结构特征至关重要。通过大量的预训练，Prot_Bert 模型能够生成高维蛋白质序列嵌入，有效表示蛋白质的生物物理特性。这些表示可直接用于下游任务，例如蛋白质二级结构预测、功能预测和亚细胞定位以及本实验中的预测任务。

2.2.2　利用蛋白质分子 3D 图生成接触图

蛋白质的功能很大程度上是由其结构决定的，其 3D 结构数据为关键功能残基的空间排布以及酶与底物间的相互作用提供重要线索，其中氨基酸组成及其在蛋白质结构中的独特折叠模式会显著影响活性位点的可及性、底物结合位点以及酶的整体稳定性。这些因素共同决定了酶的功能特性和催化效率。因此，为了从蛋白质 3D 结构中提取信息，分别读取每个蛋白质 3D 结构的 PDB 文件，并提取蛋白质中每个氨基酸内的 C_α 坐标。读取指定 PDB 文件中的原子信息，提取属于链 A 的 C_α 原子的坐标。然后，利用提取的 C_α 坐标计算距离矩阵，进而基于距离矩阵生成接触图[12]。

将蛋白质三维结构建模为无向图 $G=(V,E)$，其中每个顶点 $v_i \in V$ 表示蛋白质主链上的第 i 个残基，边集 E 定义为残基间的接触关系。对于任意两个残基 i 和 j，计算其欧氏距离，为

$$d_{ij} = \left\| r_i - r_j \right\|_2 \tag{1}$$

其中 r_i 表示第 i 个 C_α 原子的三维坐标。当 d_{ij}<10Å 时，设定 A_{ij}=1，反之 A_{ij}=0，从而得到接触图的邻接矩阵 $A \in \{0,1\}^{n \times n}$，其中 n 表示残基个数。

在本实验中所使用的蛋白质三维结构文件均来自野生型酶(WT)，对于同一个 UniProt ID 下，由于蛋白质突变或缺失，突变型酶(MUT)与野生型酶的蛋白质序列长度不一致，通过接触矩阵进行必要的补全操作。如果接触矩阵小于序列长度，则在接触矩阵的列和行上添加 0 进行填充，接触矩阵的对角线上均全部填充 1。

2.2.3　*n*-gram 在蛋白质序列中的应用

通过 *n*-gram[13]表征蛋白质序列中的上下文语义特征，为了确保表征后的蛋白质序列长度保持不变将原始氨基酸序列 $S=[s_1,s_2,...,s_L]$ 填充为 $S'=['+',s_1,s_2,...,s_L,'-']$，经 *n*-gram 表征后得到的仍为长度为 L 的 *n*-gram 片段：$G_t=[g_1,g_2,...,g_L]$ 其中每个片段 g 会被映射为对应的整数索引或嵌入向量，最终形成序列特征矩阵 $\boldsymbol{X} \in R^{L \times d}$，$d$ 为嵌入维度。这些特征作为 GCN 中节点特征，与接触矩阵 $\boldsymbol{A}$ 一起用于图卷积学习。通过这样的方式，不仅可以聚合更多的上下文信息，还能够极大地捕获突变酶发生突变区域的信息。聚合后的特征向量将作为图卷积网络(GCN)中的节点特征，使 GCN 能够将节点特征与接触图中的结构信息结

合，进行特征提取。这种处理方法有效地编码了蛋白质序列中的信息，为后续的模型训练和预测提供了丰富的数据支持。

2.3　温度嵌入

为了将反应温度信息嵌入模型，将摄氏温度(Temp)和开尔文温度(Temp_K)倒数作为特征进行拼接。设摄氏温度为T，开尔文温度倒数为$1/T_k$模型将它们作为特征向量进行拼接，将拼接后的特征向量与底物以及蛋白质特征向量$\boldsymbol{x}$通过全连接操作进行合并以实现特征融合，得到最终的输入特征向量为

$$y=\sigma(\boldsymbol{W}[x,T,\frac{1}{T_k}]) \tag{2}$$

其中：σ表示激活函数，$\boldsymbol{W}$是全连接层的权重矩阵。

MMTKcat模型的流程图如图2所示。MMTKcat模型集成了GCN以及SMILES Transformer和Prot_Bert两个预训练模型，用于从酶和底物中提取特征预测k_{cat}值。使用GCN分别对蛋白质3D结构和化合物分子图提取结构特征，而SMILES Transformer模型和Prot_Bert模型分别用于从SMILES序列和蛋白质序列中提取序列特征。

SMILES Transformer模型和Prot_Bert模型分别将SMILES序列和蛋白质序列特征提取池化后变为1024维度的向量，这些池化后的向量经过两次线性变换和ReLU激活函数，最终得到2条均为64维度的SMILES序列特征向量和蛋白质序列特征向量。

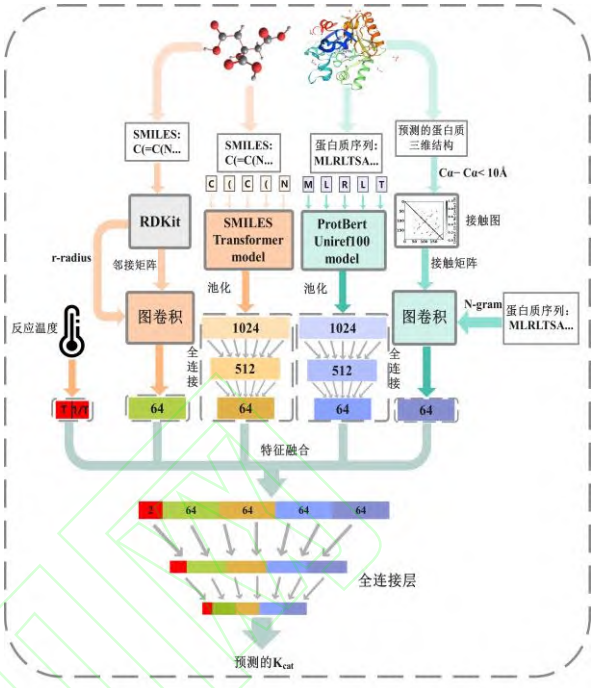


图2　MMTKcat模型的流程图

Fig. 2　Process diagram of MMTKcat model

3　本文方法

3.1　图卷积网络

图卷积网络(GCN)是用来提取图结构数据特征的深度学习模型。它可以直接在不规则的图数据上进行卷积操作。图数据由节点和边组成，每个节点均具有各自的特征，节点之间的边表示两个节点间的连接关系。GCN通过在图的邻域上进行卷积操作，逐层聚合节点及其邻居节点的信息，用来提取图结构中的高阶特征。GCN的核心操作是基于节点的特征向量和邻接矩阵，通过线性变换和非线性激活函数更新每个节点的表示。给定一个具有邻接矩阵$\boldsymbol{A}$，图卷积网络层与层之间的传播方式为

$$H^{l+1}=\sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{l}W^{l}) \tag{3}$$

其中：H^{l}表示第l层的激活矩阵。A表示图的邻接矩阵，D表示度矩阵，W^{l}表示第l层的可学习权重矩阵，σ表示非线性激活函数。

在MMTKcat模型中，图卷积网络被应用于从蛋白质3D结构和底物邻接矩阵中提取关键的结构特征[14]。利用蛋白质三维结构数据，构建基于C_{α}原子位置的接触图，这些接触图表征蛋白质分子中残基之间的空间

关系，通过对这些接触图进行卷积操作，图卷积网络(GCN)能够逐层聚合并提取蛋白质结构的高阶特征信息，使模型能够更充分地利用蛋白质的三维构象。

在提取化合物分子图的特征方面，使用 GCN 在分子的邻接矩阵上进行卷积操作，逐步提取分子结构的高阶特征。通过处理邻接矩阵，GCN 可以有效捕捉底物的分子结构信息，并生成丰富的特征表示。

3.2 Transformer 模型

Transformer 模型是由编码器和解码器组成的深度学习框架，主要依靠自注意力机制处理数据，在机器翻译、文本生成、语音识别等任务上表现出色。其核心组件包括多头自注意力机制、前馈神经网络、残差连接和层归一化。

多头自注意力机制表示为

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{h}_1,\text{h}_2,...,\text{h}_\text{h})W^O \tag{5}$$

$$\text{h}_\text{i}=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{6}$$

其中：Q 为查询矩阵，K 为键矩阵，V 为值矩阵，d_k 为键的维度。

前馈神经网络表示为

$$\text{FFN}(x)=\text{RELU}(xW_1+b_1)W_2+b_2 \tag{7}$$

其中：x 为输入特征向量，来自多头注意力机制或前一层前馈神经网络的输出；W_1 和 W_2 为的权重矩阵；b_1 和 b_2 为偏置向量。

残差连接和层归一化表示为

$$\text{LayerNorm}(x+\text{SubLayer}(x)) \tag{8}$$

其中：x 为输入特征向量；$\text{SubLayer}(x)$ 为子层操作；LayerNorm 为层归一化操作，用于对每个批次的输入特征进行归一化处理，以减少内部协变量偏移。

SMILES Transformer 是基于 Transformer 模型设计，用于处理 SMILES 字符串。SMILES Transformer 使用编码器-解码器网络，每个网络分别包含 4 个 Transformer 块，每个 Transformer 块中有 4 个头的注意力机制，嵌入维度为 256，并包含 2 个线性层，表示为

$$\text{Z=Encoder(X),Y=Decoder(Z)} \tag{9}$$

其中：$\boldsymbol{X}$ 是输入的 SMILES 序列，$\boldsymbol{Z}$ 是编码器输出的中间表示，$\boldsymbol{Y}$ 是解码器生成的输出序列。SMILES Transformer 使用 ChEMBL24 数据集中的 861000 个未标记的 SMILES 序列进行预训练。预训练通过最小化输入 SMILES 和输出概率之间的交叉熵完成。

$$\mathcal{L}=-\sum_i \text{X}_i \log \text{Y}_i \tag{10}$$

从 Transformer 输出的符号级别表示中提取分子级别表示。将最后一层的平均池化和最大池化输出、最后一层和倒数第二层的第一个输出连接起来，最终得到 1024 维的指纹。

$$\text{f}=[\text{mean}(\text{H}^{(L)}),\max(\text{H}^{L}),\text{H}_1^{(L)},\text{H}_1^{(L-1)}] \tag{11}$$

其中：$\text{H}^{(L)}$ 表示最后一层的隐藏状态，$\text{H}^{(L-1)}$ 表示倒数第二层的隐藏状态。

BERT 是基于 Transformer 架构的预训练语言模型，它在自然语言处理任务中取得了巨大成功，能够生成深度上下文表示。Prot_Bert 的架构基于 BERT，已针对蛋白质序列的特点进行调整。Prot_Bert 增加了层数，并在更长的序列上进行训练，表示为

$$H^{(l+1)}=\text{Transformer}(\text{H}^{(l)}),\ \ l=1,...,L \tag{12}$$

Prot_Bert 首先在长度为 512 的序列上训练 300000 步，然后在长度为2000 的序列上训练100000步。这种分步训练策略使模型能够在较短序列上提取有用的序列特征，同时在较长序列上进行更有效地训练。

3.3 性能评价指标

平均绝对误差(MAE，用符号 E_{MAE} 表示)表示预测值与真实值之间的平均绝对差，衡量的是误差的平均绝对值，是一种直接且易于理解的误差度量。MAE 的计算公式为

$$E_{\text{MAE}}=\frac{1}{n}\sum_{i=1}^{n}|\ y_i-\hat{y}_i\ | \tag{13}$$

均方根误差(RMSE，用符号 E_{RMSE} 表示)是预测值与真实值之间误差的平方的平均值的平方根。它对大的误差更加敏感，因此在存在较大误差时，RMSE 会比 MAE 更大。RMSE 的计算公式为

$$E_{\text{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{14}$$

决定系数(R^2)是衡量模型解释方差比例的指标，取值范围为 0~1。R^2 越接近 1，说明模型解释变量的变异越多，模型的预测能力越强。R^2 的计算公式为

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

4 实验结果与分析

4.1 五折交叉验证

采用五折交叉验证的方法对 MMTKcat 模型进行评估，确保模型的稳定性和泛化能力。将整个数据集分成 5 个互斥的子集，每次使用其中 4 个子集进行训练，剩下的 1 个子集进行测试。之后轮流将每个子集作为测试集。五折交叉验证测试集 RMSE 随训练轮数(Epoch)的变化趋势如图 3 所示。

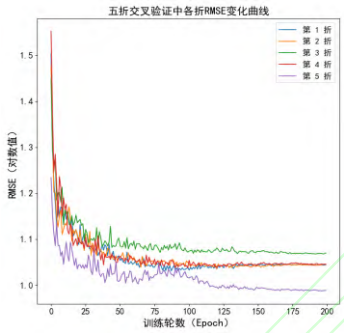


图 3 五折交叉验证测试集 RMSE 随训练轮数的变化趋势

Fig. 3 Changing trend of RMSE of the five-fold cross-validation test set with training epochs

为了进一步评估模型的性能，分别采用平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)这 3 个指标评价 MMTKcat 模型的性能，每一折的预测效果如图 4 所示。利用这些指标有助于全面了解模型的预测能力和稳定性。

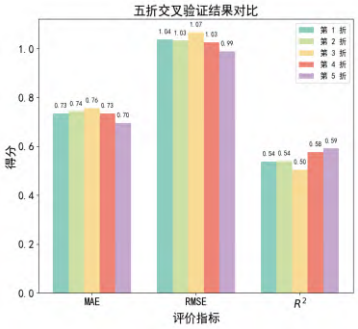


图 4 五折交叉验证结果比较图

Fig. 4 Comparison of five-fold cross-validation results

为了更加直观地观察模型预测性能，将五折交叉验证中每一折测试集的预测值与真实值放入带有热力图的散点图中如图 5 所示。因为 5 个子数据集都会在五折交叉验证中充当测试集，所以该散点图中共有 16018 个点。

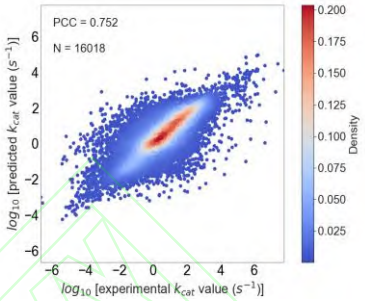


图 5 带有热力图的散点图

Fig. 5 Scatter plot with heatmap

4.2 与其他模型比较

为了验证 MMTKcat 模型预测效果的有效性，在给定的数据集中引入模型 DLKcat[1]、DLTKcat[2]、DeepEnzyme[15]，通过固定随机种子的方式保证训练集和测试集一致，并根据原文提供的代码进行对比实验。

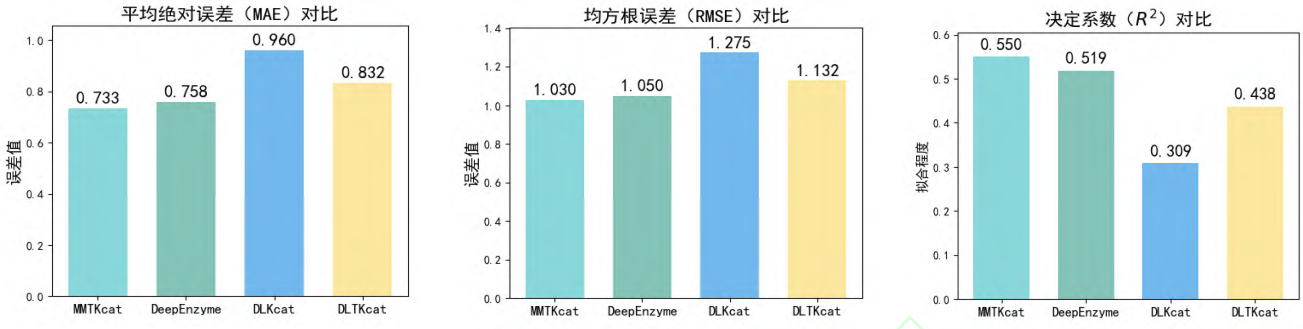
DLKcat 模型利用图卷积网络(GNN)和卷积神经网络(CNN)分别提取化合物的分子结构特征和蛋白质的序列特征，用以预测酶的催化效率常数。

DLTKcat 模型在 DLKcat 的基础上进一步引入温度变量，通过图注意力网络(GAN)和卷积神经网络(CNN)分别提取化合物和蛋白质的特征，考虑温度对 k_{cat} 的影响，使预测结果相对而言更加精确。

DeepEnzyme 模型使用图卷积网络捕获蛋白质序列三维结构的特征与底物特征，通过 Transformer 模型提取蛋白质序列特征，最后进行特征拼接从而进行 k_{cat} 的预测。

所有方法的整体性能通过五折交叉验证评估，具体结果如图 6 所示。MMTKcat 在 3 个评价指标中的表现均优于所对比模型，在 MAE 上，MMTKcat 的误差为 0.73，分别比 DeepEnzyme、DLKcat 和 DLTKcat 降低 3.95%、23.96%和 12.05%。在 RMSE 上，MMTKcat 的误差为 1.03，分别比 DeepEnzyme 的、DLKcat 和 DLTKcat 降低 1.90%、19.53%和 8.85%。在 R^2 上，MMTKcat 的值为 0.55，分别比 DeepEnzyme、DLKcat 和 DLTKcat 提高 5.77%、77.42%和 25.00%。

综上所述，MMTKcat 模型在误差控制和拟合能力方面均表现出更优的预测效果。



(a) MAE　　(b) RMSE　　(c) R^2

图 6　不同模型 MAE、RMSE、R^2 比较

Fig. 6　Comparison of MAE, RMSE, and R2 under different models

4.3　不同参数对模型预测性能的影响

4.3.1　GCN 层数

在 MMTKcat 模型中，图卷积网络(GCN)用于从蛋白质的 3D 结构和底物的邻接矩阵中提取结构特征。GCN 的层数对模型的预测性能有着重要影响。随着层数的增加，每个节点的特征将包含更高阶邻居的信息，增加 GCN 的层数能够使模型捕获更广泛的结构信息。然而，层数过多也会带来过平滑、计算复杂度增加等问题。因此，本实验分别使用 1、2、3 层图卷积网络测试不同层数对于 MMTKcat 模型预测性能的影响，结果见表 1。经过对结果的分析发现 1 层 GCN 在本研究中表现最佳，可能是因为它能够在适当的 r-radius 和 *n*-gram 参数设置下，捕捉到局部重要特征，避免模型复杂度过高和过拟合的风险。同时，这样的设置也有助于提高计算效率，使模型在较短时间内达到较好的预测性能。

表 1　GCN 层对模型预测性能的影响

Tab. 1　Impact of GCN layers on the model's predictive performance

GCN 层数	MAE	RMSE	R^2
1	**0.73**	**1.03**	**0.55**
2	0.74	1.04	0.54
3	0.78	1.1	0.49

4.3.2　r-radius 和 *n*-gram

较大的 r-radius 虽然可以捕捉更加丰富的结构信息，但是也可能引入更大的噪声和更多的冗余信息，从而影响模型预测的准确性。适当的 *n*-gram 数可以更好地捕捉序列中相邻元素的局部上下文信息，有助于提高模型的预测能力。但是，过大的 *n*-gram 数可能导致序列信息的过度分割，增加计算复杂度，导致特征稀疏性。在模型设计中需要仔细调节 r-radius 和 *n*-gram 数，以平衡信息提取的全面性和特征表示的有效性，使模型达到最佳的预测效果。r-radius 和 *n*-gram 数量对模型预测性能的影响结果见表 2，当 r-radius 为 2 且 *n*-gram 为 3 时模型的性能最佳。

表 2　r-radius 和 *n*-gram 数量对模型预测性能的影响

Tab. 2　Impact of r-radius and n-gram numbers on model prediction performance

r-radius	*n*-gram	MAE	RMSE	R^2
1	1	0.96	1.27	0.32
1	2	0.85	1.13	0.46
2	1	0.78	1.07	0.51
2	2	0.78	1.07	0.52
2	**3**	**0.73**	**1.03**	**0.55**
3	2	0.79	1.10	0.48
3	3	0.74	1.04	0.54

4.4　消融实验

为了评估模型中不同部分对最终预测性能的具体贡献，进行消融实验。通过逐步移除模型的特定组件或特性，能够深入分析每个因素对模型整体性能的影响，从而确定哪些因素对于优化预测结果至关重要。为了系统性地评估 MMTKcat 模型中引入预训练模型、整合蛋白质三维结构信息以及考虑温度拼接策略是否对预测性能产生了积极的贡献，并进一步解析这三者中哪一项因素对实验结果的影响更为显著，本实验分别对仅使用 GCN 提取底物以及蛋白质特征模型(Case 1)、去除 GCN 提取蛋白质 3D 结构部分其余部分保留(Case 2)、去除温度拼接其余部分保留(Case 3)在相同训

练集、测试集上分别进行消融实验，结果见表 3。由表 3 可知，使用蛋白质 3D 结构提取蛋白质特征可以极大增强模型预测能力。与此同时，MMTKcat 模型可以通过本文中的 2 个预训练模型得到有用的底物与蛋白质特征向量进行 k_{cat} 预测。引入温度指标，也同样会在一定程度上使模型的预测准确性得到提高。

表 3　消融实验

Tab. 3　Sensitivity analysis

实验	MAE	RMSE	R^2
Case 1	0.76	1.07	0.51
Case 2	1.03	1.32	0.25
Case 3	0.75	1.03	0.54
MMTKcat	**0.73**	**1.03**	**0.55**

4.5 案例分析

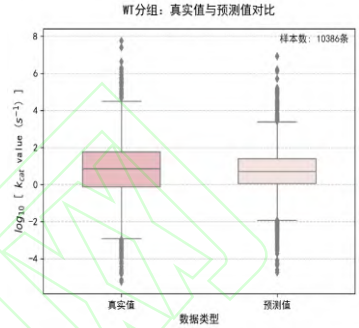
为了评估 MMTKcat 模型在实际应用中的效果，随机选取 DLKcat 数据集中的部分条目进行案例分析，所选取的 7 条酶-底物-K_{cat} 数据均未出现在本模型的训练集中。因为这 7 条数据不包括温度信息，所以移除了 MMTKcat 模型中的温度模块，其余部分保持不变，预测结果见表 4。由表 4 可知，模型在 DLKcat 数据集中同样能够实现较好的预测。

表 4　案例分析

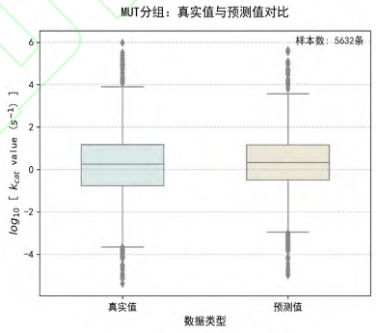
Tab. 4　Case analysis

底物名称	酶名称	真实值(s^{-1})	预测值(s^{-1})
Benzaldehyde	Benzaldehyde dehydrogenase (NAD^+)	26.4	23.96
3α,7α-Dihydroxy-5β-cholestane	Cholesterol 7α-hydroxylase	0.217	0.363
gamma-Nonalactone	Hormone-sensitive lipase	0.38	0.634
Adenosine	Adenosine kinase	0.5	0.295
2-hydroxy-5-methylnaphthalene-1-carboxylic acid	Naphthoate synthase	0.135	0.144
trans,trans-Farnesyl diphosphate	Amorpha-4,11-diene synthase	0.186	0.278
D-Alanyl-D-alanine	D-Ala-D-Ala dipeptidase	156	138.5

在前期数据集整理阶段，将酶划分为突变型酶（MUT，约占数据总量的 35%）和野生型酶（WT，约占数据总量的 65%），为了直观评估模型在不同酶类型上的预测效果，绘制了真实值与预测值的箱线图，如图 7(b)与图 7(a)所示。箱线图能够有效反映数据的分布特征，包括中位数、上下四分位数、异常值分布等信息，适用于对比真实分布与模型预测分布的一致性。



（a）野生型酶



（b）突变型酶

图 7　真实值与预测值箱线图对比

Fig. 7　Comparison of box plots of true and predicted values

在野生型酶组中，真实值与预测值的整体分布区间高度一致，中位数位置接近，四分位范围（IQR）也基本重合，且两者在极端值区域（上、下须）展现出相似的延展性。这说明模型能够较好地捕捉到真实数据的主要分布趋势。在突变型酶组中，真实值与预测值的分布同样保持了较好的对应关系。两者的中位数接近，且上下四分位数的范围相似，表明模型在突变型酶样本上也能较为准确地回归 k_{cat}。

综上所述，真实值与预测值在 WT 与 MUT 两组中均表现出明显的分布一致性，说明 MMTKcat 模型不仅能准确预测 k_{cat} 值，还具备较强的泛化能力，能够适应不同类型酶分子的预测任务。

5　结　语

本文提出基于多模态信息融合和图卷积网络的酶催化效率预测模型(MMTKcat)，用于预测酶-底物的催化常数。通过整合蛋白质序列、3D 结构特征以及底物 SMILES 序列、分子图，利用 Prot_Bert 以及 SMILES Transformer 两个预训练模型和 GCN 分别提取序列信息和结构信息，结合反应温度，构建多模态特征融合的预测框架。实验结果表明，该模型在预测酶催化常数上具有显著优势，相较于传统方法具备更高的准确性和鲁棒性。尽管如此，模型仍有不足之处，在预测酶周转数时仅考虑了反应物而没有将生成物纳入考虑的范围内，除此之外现在基于蛋白质 3D 结构进行预训练的模型也趋于成熟，如 SAPROT[16]在提取蛋白质 3D 结构方面也已经展现出了极其出色的能力，若在将来研究中使用该方法或许能够进一步提升 MMTKcat 预测能力。随着 AlphaFold 等相关蛋白质结构预测软件的不断发展和改进，本实验中数据集中的蛋白质分子 3D 结构的精确度有可能会得到了进一步提升，模型预测能力也随之增强。

参考文献：

[1] LI F, YUAN L, LU H, et al. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction[J]. Nature catalysis, 2022, 5(8): 662-672.

[2] QIU S, ZHAO S, YANG A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates[J]. Briefings in bioinformatics, 2024, 25(1): bbad506.

[3] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2024-06-10]. https://doi.org/10.48550/arXiv.1609.02907.

[4] ELNAGGAR A, HEINZINGER M, DALLAGO C, et al. ProtTrans: toward understanding the language of life through self-supervised learning[J]. IEEE Transactions on pattern analysis and machine intelligence, 2021, 44(10): 7112-7127.

[5] HONDA S, SHI S, UEDA H R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery[EB/OL]. [2024-06-10]. https://doi.org/10.48550/arXiv.1911.04738.

[6] SCHOMBURG I, JESKE L, ULBRICH M, et al. The BRENDA enzyme information system: from a database to an expert system[J]. Journal of biotechnology, 2017, 261: 194-206.

[7] WITTIG U, REY M, WEIDEMANN A, et al. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics[J]. Nucleic acids research, 2018, 46(D1): 656-660.

[8] UNIPROT CONSORTIUM. UniProt: the universal protein knowledge base in 2023[J]. Nucleic acids research, 2023, 51(D1): 523-531.

[9] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structures prediction with AlphaFold[J]. Nature, 2021，596(7873): 583–589.

[10] LANDRUM G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling[J]. Greg landrum, 2013, 8: 5281.

[11] SUZEK B E, YUQI W, HONGZHAN H, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches[J]. Bioinformatics, 2015, 31(6): 926-932.

[12] PIETRO D L, KEN N, PIERRE B. Deep architectures for protein contact map prediction[J]. Bioinformatics, 2012, 28(19): 2449–2457.

[13] BEPLER T, BERGER B. Learning protein sequence embeddings using information from structure[EB/OL]. [2024-06-10]. https://doi.org/10.48550/arXiv.1902.08661.

[14] GLIGORIJEVI V, RENFREW P D, KOSCIOLEK T, et al. Structure-based protein function prediction using graph convolutional networks[J]. Nature communications, 2021, 12(1): 3168.

[15] WANG T, XIANG G, HE S, et al. DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D-structures[J]. Briefings in Bioinformatics, 2024, 25(5): bbae409.

[16] SU J, HAN C, ZHOU Y, et al. SaProt: protein langua ge modeling with structure-aware vocabulary[EB/OL]. [2024-06-10]. https://www.biorxiv.org/content/biorxiv/early/2023/10/02/2023.10.01.560349.full.pdf.