第 62 卷　第 1 期
2026 年 1 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 1
Jan., 2026

DOI:10.13232/j.cnki.jnju.2026.01.004

SMFF:基于结构的多模态抗原抗体结合亲和力预测方法

施佳豪,姜　舒*,鞠恒荣,丁卫平
(南通大学人工智能与计算机学院,南通,226019)

摘　要:目前抗原-抗体结合亲和力预测的人工智能方法都基于序列或结构的单一建模,难以捕获更全面的信息.因此,提出了一种基于结构的多模态特征融合方法.该方法主要包括三个模块:多模态抗体信息挖掘模块、多模态抗原信息挖掘模块和融合预测模块.在多模态抗体信息挖掘模块中,引入Roformer网络分别提取抗体的重链和轻链信息,同时采用GearNet挖掘抗体的结构信息.随后,借助交叉力注意机制自适应地融合序列和结构信息.在多模态抗原信息挖掘模块中,使用蛋白质大语言模型ESM2(Evolutionary Scale Modeling v2)来实现抗原序列信息的抽取.在融合预测模块中,为了实现全面且高效的亲和力预测协同,基于卷积神经网络(Convolutional Neural Networks,CNN)构建了一个多尺度特征协同提取网络,进一步提升抗体和抗原表示的可判别性.最后,将多尺度的抗体和抗原表示输入融合层,生成稳健的亲和力预测结果.实验结果表明,提出的模型在基准数据集上的表现超越了所有的基线模型,并在独立测试集上表现优异,证明该方法具有强大的预测与泛化能力.

关键词:抗原-抗体结合亲和力预测,多模态融合,结构信息,注意力机制
中图分类号:Q31　　**文献标志码**:A

SMFF: Structure-based multi-modal framework for antigen-antibody binding affinity prediction

Shi Jiahao, Jiang Shu*, Ju Hengrong, Ding Weiping
(School of Artificial Intelligence and Computer Science, Nantong University, Nantong, 226019, China)

Abstract: Currently, artificial intelligence methods for predicting antigen-antibody binding affinity predominantly rely on either sequence-based or structure-based unimodal modeling approaches, which limit their ability to capture comprehensive interaction information. Therefore, we present a structure-based multi-modal feature fusion framework for antigen-antibody affinity prediction. The framework mainly consists of three components: a multi-modal antibody information mining module, a multi-modal antigen information mining module, and a fusion prediction module. In the antibody information mining module, we employ the Roformer network to separately extract heavy chain and light chain information while utilizing the GearNet to capture structural information of antibodies. The sequence and structural information is then adaptively fused through a cross-attention mechanism. For antigen information extraction, we implement the ESM2 (Evolutionary Scale Modeling v2) protein language model to process antigen sequence data. The fusion prediction module incorporates a multi-scale feature extraction network based on CNN (Convolutional Neural Networks) to enhance the discriminative power of antibody and antigen representations, enabling comprehensive and efficient affinity prediction. Finally, the multi-scale representations of antibodies and antigens are fed into a fusion layer to generate robust affinity prediction results. Experimental results demonstrate that our proposed model surpasses all baseline methods on benchmark datasets and shows

基金项目:国家自然科学基金(62406153),江苏省高等学校自然科学研究(23KJB520031),江苏省高校自然科学研究项目(24KJB520032)
收稿日期:2025－09－03
*通信联系人,E-mail:jshmjs45@ntu.edu.cn

excellent performance on independent test sets, demonstrating the method's robust predictive power and generalization capability.

Keywords: antigen-antibody binding affinity prediction, multi-modal fusion, structural information, attention mechanism

近年来，人工智能的快速发展带动了生物领域的进步，其中抗体开发是生物领域中一项极具挑战性的工作．抗体在人类生命活动中发挥着重要作用，如图1所示，其Y形结构的两端，称为互补决定区（Complementarity Determining Region，CDR），可与高亲和力靶抗原上的表位结合[1−3]，从而特异性识别入侵的抗原．生物制药行业利用这种特异性开发了单克隆抗体（Monoclonal Antibody，MAb）作为治疗药物，不仅对疾病治疗的成功率高、疗效显著，而且副作用较小[4−7]．随着抗体-药物偶联物（Antibody-Drug Conjugates，ADC）等生物技术的进步，即使是传统意义上“不可成药”的疾病靶点也能实现靶向治疗．抗体药物可用于治疗各类癌症及类风湿性关节炎等自身免疫性疾病，且在人体免疫系统中扮演着关键角色[8−12]．

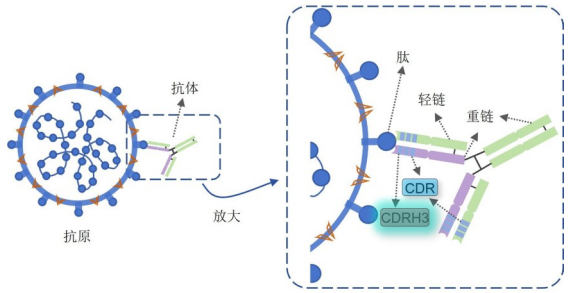

图1　抗原-抗体结合示意图

Fig. 1　Schematic of antigen-antibody binding

确定抗原-抗体相互作用的亲和力是抗体开发的重要环节．目前，亲和力测定的实验方法包括放射免疫测定（Radioimmunoassay，RIA）[13]、酶联免疫吸附测定（Enzyme Linked Immunosorbent Assay，ELISA）[14]、表面等离子体共振（Surface Plasmon Resonance，SPR）[15]和生物层干涉测定法（Bio-Layer Interferometry，BLI）[16]等，但这些方法往往耗费大量资源且十分耗时．由于人工智能技术（Artificial Intelligence，AI）迅速发展，学者们开始致力于开发基于AI的预测模型以预测抗原-抗体结合亲和力．早期方法多利用抗体结构信息进行预测，例如，Myung et al[17]提取游离溶剂可及表面积、残基深度和二级结构信息等特征，有效捕捉了抗原-抗体相互作用的关键特征；Cai et al[18]通过多关系图构建、多层次几何消息传递及对大规模未标记蛋白质结构数据的对比预训练等技术，有效提取了抗原-抗体复合物界面的几何特征，并构建了GearBind模型预测抗原-抗体亲和力的变化．还有一些科学家关注序列信息并创建亲和力预测模型，例如，Yuan et al[19]提出的一种新颖的模型DG-Affinity，通过蛋白编码和抗体编码的方式融合特征后输入神经网络，实现抗原-抗体之间的亲和力预测；Kang et al[20]基于Hag-Net网络结构，探索全序列、仅接触和仅抗体三种不同的图表示策略，并比较其在结合亲和力变化预测任务上的性能．

尽管上述工作取得了值得称赞的进展，但这些模型只考虑到单一模态的建模．当前一些蛋白质相关工作表明[21−24]，序列信息和结构信息对于生物信息挖掘与建模都是有益的．并且在抗原-抗体结合亲和力预测的任务中，稀缺的数据给模型带来了过拟合的风险和泛化能力差的问题，挖掘单一模态的信息导致这些工作的预测能力只体现在经过训练的抗原-抗体对，在遇到没见过的抗原-抗体对时，其预测能力和精度会大大降低．此外，这些工作仅采用简单的特征级联融合策略，未能充分建模抗原-抗体结合过程中多模态特征间的动态交互机制，特别是无法自适应地识别不同模态特征对亲和力预测的差异化贡献权重．

因此，本文提出了一种基于结构的新型多模态特征融合方法（Structure-Based Multi-Modal Feature Fusion Framework，SMFF）．该方法首先构建多模态抗体信息挖掘模块，将抗体序列分为重链和轻链，使用预训练的重链和轻链编码器分别获取嵌入；通过抗体结构预测工具得到抗体3D结构，构建多关系图并输入图编码器以获得特征

嵌入;再基于抗体结构信息和序列信息,使用交叉注意力机制实现模态交互.在多模态抗原信息挖掘模块中,引入蛋白质大语言模型实现抗原序列特征抽取.最后,将抗体交互特征嵌入和抗原交互特征嵌入输入融合预测模块,通过多尺度特征协同提取网络整合抗体与抗原之间的信息,最终通过多层感知器预测亲和力数值.

1 背景

1.1 蛋白质编码器 抗体是一种特殊的蛋白质,近年来蛋白质编码器的快速发展为抗原-抗体结合亲和力预测提供了便捷的信息编码工具.例如,Lin et al[25]提出仅在蛋白质序列上训练的大型语言模型ESM2(Evolutionary Scale Modeling v2),用于学习蛋白质序列语义.Wang et al[26]提出了轻量级语言模型ProtFlash,通过引入相对旋转角度的概念改进传统的位置编码机制,使其能更好地捕捉蛋白质序列中的长程依赖关系;Zhang et al[27]提出基于蛋白质结构的预训练模型GearNet,该模型引入稀疏边缘消息传递机制和多视图对比学习,显著改善了蛋白质结构的表征.受这类工作启发,Hie et al[28]将蛋白质语言应用于人类抗体进化研究,得到了结合亲和力和活性更高的抗体.本文选用在大量抗体序列上预训练的Roformer[29]编码器来提取抗体序列信息以及选用大语言模型ESM2[25]生成抗原的高质量嵌入.

1.2 抗体结构预测 深度学习方法为蛋白质结构预测领域带来了革命性突破.AlphaFold[30]的问世使蛋白质的准确结构预测在很大程度上得到普及.同时,一些专门针对抗体的深度学习方法(例如,DeepAb[31]和ABlooper[32])显著提高了CDR回路的建模精度,尤其是对挑战性较大的重链可变互补决定区3(Variable Heavy Chain Complementarity Determining Region 3,CDRH3)回路的预测.DeepAb通过预测残差之间的几何约束,并将这些约束输入到Rosetta生成抗体完整的结构;ABlooper采用端到端的方式进行CDR回路的结构预测,虽然需要一些后处理步骤提升预测准确性,但能提供回路质量评估.此外,NanoNet[33]作为专门训练的工具,可预测单链抗体(纳米抗体)并提供快速的预测结果.而想要将抗体结构信息运用到亲和力预测中,预测的时间和准确度都需要得到保证.因此,本文采用Ruffolo et al[34]提出的IgFold预测抗体的3D结构(PDB格式),该工具能在25 s内预测出质量与AlphaFold等其他方法相当或更优的抗体结构,使以往难以实现的研究方向成为可能.

2 方法

SMFF主要包括三个模块:多模态抗体信息挖掘模块、多模态抗原信息挖掘模块和融合预测模块,如图2所示.图2a为多模态抗体信息挖掘模块,引入在大量抗体序列上预训练的Roformer网络分别提取抗体的重链和轻链信息,同时构建抗体图,并采用预训练的GearNet-Edge挖掘抗体的结构信息,随后通过交叉注意力机制自适应地融合序列和结构信息.图2b为多模态抗原信息挖掘模块,通过蛋白质大语言模型ESM2来实现抗原信息抽取.图2c为融合预测模块,为了实现全面且高效的亲和力协同预测,基于CNN(Convolutional Neural Networks)构建了多尺度特征协同提取网络MF_CNN,进一步提升抗体和抗原表示的可判别性,最终将多尺度的抗体和抗原表示输入融合层,生成稳健的亲和力预测结果.

2.1 图构建 本文构建了抗体的多关系图,首先通过IgFold预测抗体结构,再根据预测的PDB结构文件构建抗体残基水平图.考虑到蛋白图的规模问题,仅保留α-碳坐标作为图节点以大幅减小图的大小,随后采用Zhang et al[27]提出的方法将抗体结构表示为关系图$\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{R})$,其中,$\mathcal{V}$和$\mathcal{E}$表示节点和边缘的集合,$\mathcal{R}$表示边缘类型的集合.如图2d所示,本研究为残基水平图构建了三种类型的边缘:序列边、半径边和k最近邻边.

2.1.1 序列边 如果节点i和j之间的序列距离低于预定义的阈值d_{seq}(例如,$|j-i|<d_{\text{seq}}$),则添加序列边.本文将d_{seq}设为2,即仅当连续距离不超过2时添加边,并且根据相对序列距离$d\in\{-2,-1,0,1,2\}$,将序列边进一步划分为五种类型.

2.1.2 半径边 采用基于欧氏距离的阈值判定

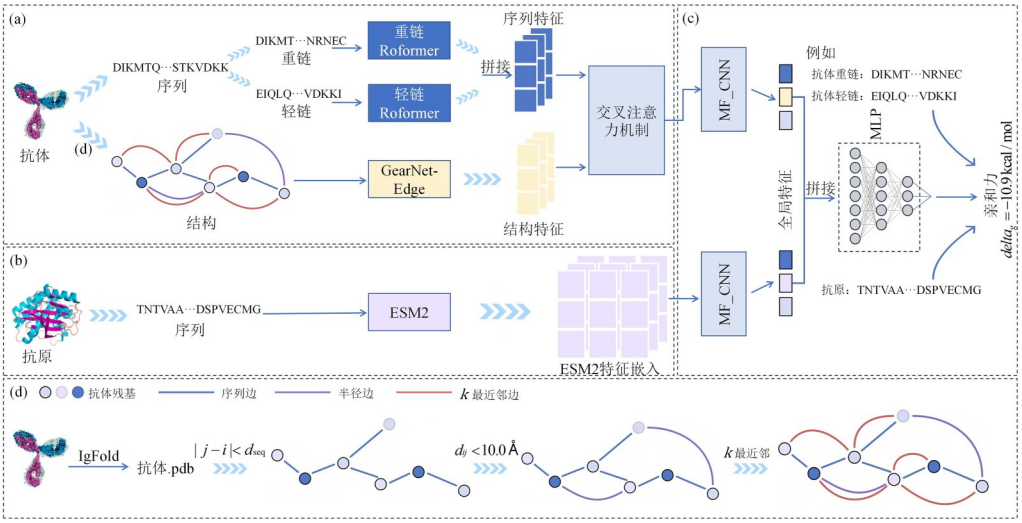

图2　SMFF流程图

Fig. 2　The flow chart of proposed SMFF

方法，即当节点 i 与 j 之间的空间距离小于阈值（默认设置为10.0 Å，可根据需求调整）时，在对应节点间建立连接边.

2.1.3　k 最近邻边　为了体现蛋白质之间不同的空间尺度，每个节点都根据欧几里得距离连接到其 k 最近邻.

将半径边和 k 最近邻边视为不同的边类型，并进一步将序列边细化为子类型，这种构造方法能反映不同的几何属性，提供抗体结构特征的全面表示.

2.2　多模态抗体信息挖掘模块

2.2.1　抗体序列编码　由图2a可见，本文将抗体重链与轻链分开，使用两个分别在大规模抗体重链序列和轻链序列上进行预训练的编码器进行编码，以体现抗体两条链在抗原-抗体结合亲和力预测中各自的作用. 模型采用基于Roformer架构的预训练模型对抗体两条链进行编码，Roformer通过引入旋转位置编码（Rotary Position Embedding，RoPE），能有效捕获序列中的长距离依赖关系和相对位置关系，这对抗体序列中关键氨基酸的位置与功能关系的建模至关重要. 在各种长文本分类基准数据集上的实验结果表明[29]，具有旋转位置嵌入的增强模型（即Roformer）性能更优. 设抗体重链序列表示 $H=\{h_1,h_2,\cdots,h_{N_H}\}$，轻链序列表示 $L=\{l_1,l_2,\cdots,l_{N_L}\}$，其中，$N_H$ 和 N_L 分别为重链和轻链的序列长度. 编码后的重链和轻链表示如式（1）所示：

$$E_H=Roformer_H(H),E_L=Roformer_L(L) \tag{1}$$

其中，$Roformer_H$ 和 $Roformer_L$ 分别表示在重链和轻链序列上预训练的Roformer编码器.

完成独立编码后，将重链和轻链的表示进行拼接，如式（2）所示：

$$E=Concat(E_H,E_L) \tag{2}$$

其中，$E\in R^{l_s\times d_s}$，l_s 是序列长度，d_s 是嵌入维度. 这种方式既保留了重链和轻链各自的独立信息，又通过联合表示充分利用了两条链在结合特异性和亲和力预测中的协同作用.

2.2.2　抗体结构编码和交叉注意力机制　与其他图神经网络（Graph Neural Network，GNN）模型相比，GearNet-Edge通过在抗体图上构建边图和边消息传递，能模拟残基与其他序列或空间相邻残基的不同相互作用之间的依赖关系，这对抗体的空间结构建模尤为重要，这主要是由于抗体的结合亲和力预测与其3D结构密切相关. 因此，选用GearNet-Edge作为结构编码器.

通过one-hot编码构建了抗体图节点特征，表示为 $f\in\{0,1\}^{n\times 21}$. 边的特征 $f(i,j,r)$ 通过两个边节点的节点特征、边类型的one-hot编码及它们之间的序列和空间距离串联而成.

然后，GearNet-Edge编码器通过构造边图

$\mathcal{G}'=(\mathcal{V},\mathcal{E}',\mathcal{R}')$来捕捉边与边之间的依赖关系，其中，$\mathcal{V}'$表示抗体图$\mathcal{G}$中边的集合，每一条边视为$\mathcal{G}'$中的节点，$\mathcal{E}'$表示原始图中两条边之间的连接关系，$\mathcal{R}'$表示边图中的边类型. 具体而言，在抗体图$\mathcal{G}$中，若边$(i,j)$和边$(k,l)$共享一个节点(即$j=k$或$i=l$)，则在$\mathcal{G}'$中添加从$(i,j)$到$(k,l)$的边. 为了反映两条边的相对几何关系，使用$\angle(i,j,k)$表示边$(i,j)$与边$(k,l)$之间的夹角并将角度区间$[0,\pi]$离散为八个子区间，不同的角度子区间对应不同的边类型$r\in\mathcal{R}'$. 此外，为了提高计算效率，本文在构建边图时仅保留抗体图中最显著的边交互关系，以减少计算负担.

对于构建的边图$\mathcal{G}'$中边(i,j)的特征更新，其消息传递式如式(3)所示：

$$m_{(i,j)}^{(l)}=\sigma\left(BN\left(\sum_{r\in\mathcal{R}'}W_r\sum_{(k,l)\in\mathcal{N}'(i,j)}m_{(k,l)}^{(l-1)}\right)\right) \tag{3}$$

其中，$m_{(i,j)}^{(l)}$为边(i,j)在第l层的特征；$\mathcal{N}'(i,j)$表示边(i,j)的邻居集合；W_r是边类型r的可学习权重矩阵；σ为非线性激活函数$ReLU$；BN(Batch Norm)表示批正则化. 通过多层消息传递，边的特征逐层被更新，捕捉了抗体结构中边与边之间的多尺度关系. 在完成边消息传递后，更新的边特征被重新引入抗体图$\mathcal{G}$，以进一步增强节点的表示能力. 结合边消息，节点i的特征更新式被重新定义，如式(4)所示：

$$u_i^{(l)}=\sigma\left(BN\left(\sum_{r\in\mathcal{R}}W_r\sum_{j\in\mathcal{N}_r(i)}h_j^{(l-1)}+FC\left(m_{(i,j)}^{(l)}\right)\right)\right) \tag{4}$$

其中，$u_i^{(l)}$为节点i在第l层的特征，$\mathcal{N}_r(i)$表示节点i的邻居集合，$h_j^{(l-1)}$为邻居节点j的特征，FC(Fully Connected Layer)是全连接层，用于将边特征$m_{(i,j)}^{(l)}$映射到与节点特征匹配的维度. 然后，将所有抗体节点特征$u_i^{(l)}$组合起来形成结构嵌入表示，表示为$U\in R^{l_t\times d_t}$，其中，l_t是结构嵌入中的节点数，d_t是结构特征的维度.

接着，开发了一个交叉注意力模块，以交互方式将基于Roformer的序列表示与基于GearNet-Edge的结构表示交互式集成，从两种模态中提取关键的相关特征. 多模态输出X_a的计算式如式(5)所示：

$$\boldsymbol{Q}=EW_Q,\boldsymbol{K}=UW_K,\boldsymbol{V}=UW_V$$
$$X_a=Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

交叉注意力模块使本模型能通过两种抗体嵌入的相互作用学习独立抗体形态(1D序列和3D结构)之间的关系. 因此，本模型可以从相关的抗体模态中捕获更全面的信息，增强抗原-抗体结合亲和力预测的准确性.

2.3 多模态抗原信息挖掘模块 本文主要关注最常见的抗原类型，即蛋白质抗原. 本研究利用蛋白质语言模型ESM2从抗原序列中提取特征，如图2b所示. 在训练过程中，ESM2利用了来自多个著名数据库的大规模未标记的蛋白质序列数据，包括UniRef50，UniRef90[35]，Pfam[36]和TrEMBL[37]. 这种对不同数据集的广泛训练使ESM2能学习蛋白质序列的丰富特征，包括潜在的结构特性，这也是选择ESM2作为抗原序列编码器的原因.

考虑到计算约束和模型容量限制，本文选择ESM2-t33_650M_UR50D作为基础模型，其包含了6.5亿个参数. 设抗原序列表示为$S=\{a_1,a_2,\cdots,a_L\}$，其中，L是序列的长度，a_i表示第i个氨基酸残基. 首先将每个氨基酸a_i转换为其one-hot编码表示来预处理序列，在序列的开头和结尾分别添加特殊标记"$\langle cls\rangle$"和"$\langle eos\rangle$"，并将序列填充到固定长度以符合模型输入要求，得到处理后的序列S'，其计算式如式(6)所示：

$$S'=\{\langle cls\rangle,a_1,a_2,\cdots,a_L,\langle eos\rangle,\langle pad\rangle,\cdots\} \tag{6}$$

然后将处理后的序列输入到ESM2模型，该模型采用33层Transformer编码器，通过多头注意力机制来捕获氨基酸残基之间的依赖关系，以学习抗原序列的丰富特征. 通过编码层后，最终输出$H^{(33)}$是一个大小为$(L+2,1280)$的张量，其中每个位置的向量都包含该残基的深度特征信息. 为了提取抗原的整体特征，本文对所有氨基酸残基特征进行池化，以获得最终的抗原特征载体X_p. 池化操作涉及平均每个氨基酸位置的特征向量，X_p的计算式如式(7)所示：

$$X_p=\frac{1}{L}\sum_{i=1}^{L}H_i^{(33)} \tag{7}$$

其中，$H_i^{(33)}$代表第i个氨基酸残基的特征向量，得

到X_p是整个抗原特征的1280维向量.

2.4　融合预测模块　本文将2.2和2.3中通过编码器学习到的抗体特征嵌入X_a和抗原特征嵌入X_p输入到预测网络中以预测亲和力数值.与以往直接拼接两种特征嵌入再送入多层感知器这类简单的融合方法不同,本文通过一种多尺度特征提取卷积神经网络MF_CNN整合抗体与抗原的信息,再输送到多层感知器中,最终得到亲和力数值,MF_CNN利用包含卷积、池化和*Relu*的三层CNN骨架进行多尺度特征提取.随后使用多层FC和残差操作进一步融合这些特征以获得最终输出,如式(8)~(9),最后使用多层感知器预测亲和力数值,计算式如式(10)所示:

$$F_a = Residual\left(FC\left(X_a\right)\right) = FC\left(X_a\right) + X_a \tag{8}$$

$$F_p = Residual\left(FC\left(X_p\right)\right) = FC\left(X_p\right) + X_p \tag{9}$$

$$Affinity = MLP\left(\left[F_a, F_p\right]\right) \tag{10}$$

3　实验与结果

3.1　数据集　本文分别在三个具有亲和力数值的抗原-抗体对基准数据集BioMap,14H,14L和独立数据集上评估了SMFF,并与其他基准模型进行比较.

BioMap数据集来源于BioMap组织的抗原-抗体结合亲和力预测竞赛[38],包含1706对抗原-抗体配对数据,并以$delta_g$(吉布斯自由能变化,单位为kcal•mol^{-1})作为标注.该数据集涵盖638种独特抗原和1277种独特抗体,其中大多数抗体来自人和小鼠,另有少量抗体来自仓鼠、黑猩猩、恒河猴、兔子、大鼠和美洲驼,不包括纳米抗体.

14H和14L数据集来源于LL-SARS-CoV-2数据库[39],每个数据集包含了CDR区域中1~3个突变.本文过滤了缺乏14H和14L数据集亲和力信息的原始数据,并对具有相同抗体序列的所有序列的亲和力测量值进行平均,最终得到14H中有13922个唯一的重链数据条目,14L中有18708个唯一的轻链数据条目.在14H数据集中,仅重链存在变化,每个条目的轻链和抗原保持不变;同理,14L数据集中仅轻链不同,重链和抗原在所有条目中保持一致.

本文使用两个不同的独立数据集评估模型的泛化能力,第一个来自公开的独立测试集Antibody-Benchmark[40],其中包含67个抗原-抗体复合物的结构文件,抗原与抗体自身的结构文件以及结合亲和力数值$delta_g$.为了避免数据泄露,筛选了两个独立数据集,保证独立数据集上的这些复合物与训练数据没有重复.另外,去除了没有亲和力数值的条目,且只选择抗体两条链的条目以方便测试,最终筛选出42个条目.根据数据集中抗体的PDB标识与抗原的PDB标识下载对应的蛋白质序列,最终形成优化后的数据集.第二个独立数据集来自PPB-Affinity[41],这是一个整理蛋白-蛋白结合亲和力的数据库,其中包含了一部分抗原-抗体对的结合亲和力数据,首先删除了没有明确亲和力值的数据并保留了1206个条目,其次从这1206个条目中选出抗体两条链且受体一条链的条目,最后根据复合物结构文件下载fasta序列并映射出重链轻链以及抗原的序列信息,最终形成922个条目的优化数据集.需要注意的是,由于PPB-Affinity不直接包含结合亲和力数值$delta_g$,只是给出了用KD(解离常数,单位为M)表示的结合亲和力数值,因此,通过式(11)做一个转换:

$$delta_g = RT\ln\left(KD\right) \tag{11}$$

其中,R是气体常数,T是温度,KD是解离常数,将结合亲和力转换成为$delta_g$.

3.2　实验设置和评价指标　本文以8∶1∶1的比例拆分训练集、验证集和测试集,并在三个基准数据集上进行了十次独立实验.SMFF基于PyTorch实现,并在A6000GPU上进行训练,所有可训练参数均由Adam算法优化,学习率为0.001.本模型中,Roformer抗体编码器输出特征向量的维度设置为768,ESM2抗原编码器输出特征向量的维度设置为1280,抗体序列(含重链和轻链)的最大长度都设置为170,抗原序列最大长度设置为512,并裁剪掉多余的部分,GearNet结构特征编码器输出特征向量的维度设置为512,MF_CNN中卷积神经网络的通道数设置为140,隐藏层大小设置为76,嵌入维度设置为256.本文在数据集上训练了50个epoch,批量大小(Batch Size)设为8.为了防止模型过拟合,如果15个

epoch内指标没有增加,则中断训练并保存最佳模型.

训练模型的损失值变化如图3所示,由图3可见,模型在训练过程中快速收敛,此后仅在小范围内轻微波动,最终保存验证集表现最佳阶段的模型,说明SMFF即便在小数据集上训练,也能有稳定的训练过程和效果.

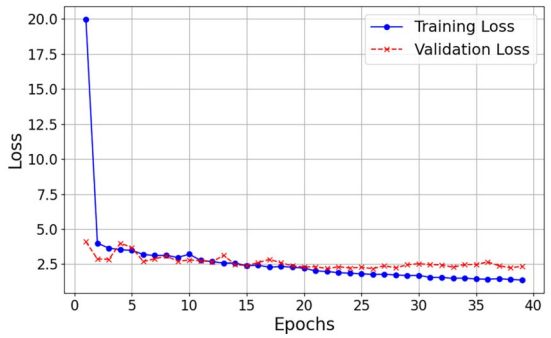

图3 每个epoch上SMFF在BioMap上的训练和验证损失值

Fig. 3 Training and validation loss of SMFF on BioMap per epoch

在抗原-抗体结合亲和力预测任务中,对模型性能的全面评估需要兼顾预测值与真实值之间的数值一致性以及排序一致性.因此,本文采用该领域通用指标Pearson相关系数和Spearman相关系数作为核心评估标准[17,19],二者分别从线性关联与单调关联两个维度提供互补的评估视角.

Pearson相关系数量化了预测值与真实值之间的线性相关程度,其取值为[−1,1].该指标通过计算协方差与标准差的比值,反映预测值随真实值变化的线性响应特性.如式(12)所示,其数学定义基于原始数据的偏差乘积:

$$r_{\text{Pearson}}=\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}} \tag{12}$$

较高的Pearson相关系数表明模型能准确地捕捉结合亲和力的数值变化趋势,Spearman相关系数通过衡量预测排名与真实排名之间的单调关联来评估模型对样本相对强弱关系的识别能力.该指标基于秩次差异计算,对原始数值分布不敏感,更能反映非线性的关联模式,如式(13)所示:

$$\rho_{\text{Spearman}}=1-\frac{6\sum d_i^2}{n\left(n^2-1\right)} \tag{13}$$

其中,d_i表示第i个样本的预测值与真实值的排名差.Spearman系数在抗原-抗体结合亲和力预测中具有重要价值,一方面,药物筛选过程中通常更关注候选分子的相对优劣而非绝对数值;另一方面,该指标对异常值及数据的非线性变换具备较强的鲁棒性.

3.3 结果 本文将提出的模型与以下基线模型进行比较.

AbMAP[42]是一种抗体高变区的蛋白质语言模型.

AntiBERTa2[43]是一种预先训练的抗体特异性序列编码器模型.

ESM-F是一种基于ESM2[25]的抗原-抗体亲和预测模型.

Ens-Grad[44]是Liu et al[44]提出的CNN架构,用于抗体CDR设计.

Vanilla BERT是原始BERT模型,具有随机初始化的权重,在多个亲和力数据集上进行抗原-抗体亲和力预测训练.

3.3.1 在基准数据集上的性能比较 表1展示了SMFF在BioMap,14H和14L三个数据集上的性能比较,其中,表中黑体字表示结果最优,"±"后面的值表示Pearson和Spearman相关系数的标准误差.

由表1可见,在所有三个数据集的亲和力预测任务中,SMFF在Pearson相关性和Spearman相关性指标方面的表现优于其他基线模型.具体而言,SMFF在14H数据集上实现了0.640的Pearson相关性,在14L数据集上实现了0.677的Pearson相关性.值得注意的是,SMFF在14H数据集上提升得更多,这可能是由于14H的重链是变化的,而重链在抗体与抗原结合的时候发挥着更重要的作用.

为了验证该模型预测LL-SARS-CoV-2以外抗原-抗体亲和力的能力,本文使用BioMap数据集来评估模型的预测性能.由表1可见,本文提出的模型达到了0.697的Pearson相关性和0.726的Spearman相关性.与之前的结果一样,同样优于其他基线模型.

为了进一步分析实验结果,绘制了SMFF和其他基线模型在BioMap数据集上预测亲和力与真实值的散点图,如图4所示.由图可见,本文提出的模型的散点在图中更密集地分布在$y=x$线附近,这表明本模型预测的亲和力值更接近真实值.

表1　本模型与基线方法在三个数据集上的性能比较

Table 1　Performance comparison between our model and baseline methods on three datasets

	14H		14L		BioMap	
	r_{Pearson}	ρ_{Spearman}	r_{Pearson}	ρ_{Spearman}	r_{Pearson}	ρ_{Spearman}
AbMAP	0.606±0.015	0.510±0.015	0.674±0.012	0.685±0.016	0.637±0.027	0.673±0.029
AntiBERTa2	0.623±0.011	0.545±0.008	0.673±0.013	0.684±0.012	0.633±0.022	0.670±0.031
Vanilla BERT	0.594±0.021	0.480±0.025	0.607±0.024	0.611±0.027	0.492±0.036	0.498±0.037
Ens-Grad	0.601±0.016	0.476±0.023	0.637±0.019	0.645±0.023	0.645±0.031	0.664±0.033
ESM-F	0.634±0.007	0.516±0.010	0.674±0.011	0.681±0.014	0.628±0.028	0.644±0.024
SMFF	**0.640±0.013**	**0.552±0.016**	**0.677±0.015**	**0.686±0.015**	**0.697±0.026**	**0.726±0.024**

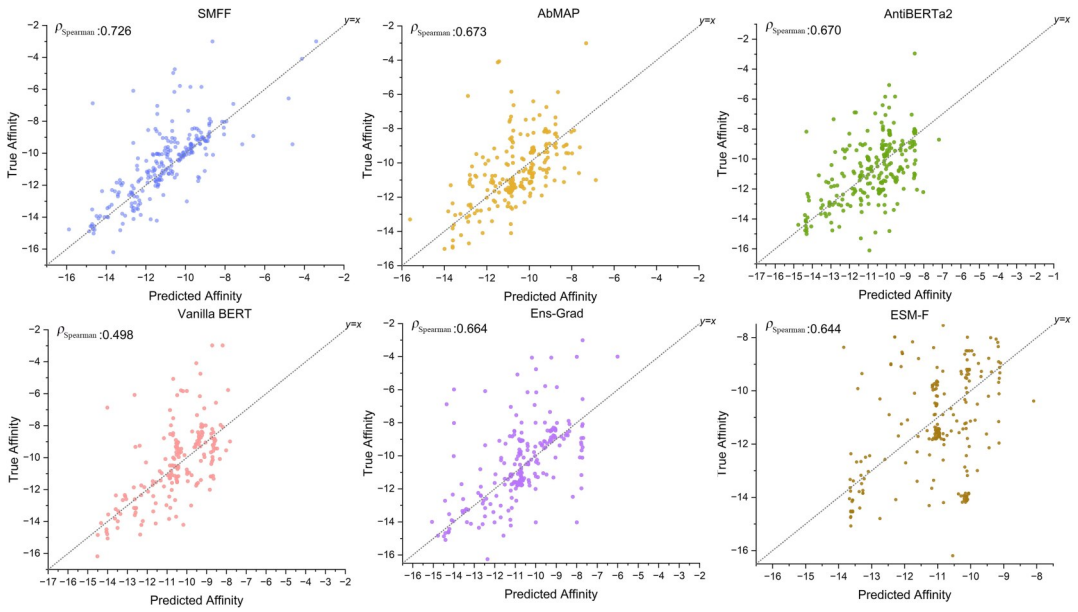

图4　针对BioMap数据集,SMFF与基线方法预测抗原-抗体结合亲和力的性能比较

Fig. 4　Performance comparison of SMFF versus baseline methods for predicting antibody-antigen binding affinity on BioMap dataset

3.3.2　在独立数据集上的性能比较　泛化性和稳健性是抗原-抗体结合亲和力预测中的关键问题,尤其是当模型遇到没见过的抗原与抗体时.在该任务中,相似或相同的抗体或抗原导致的数据冗余可能导致预测任务简化,这可能混淆方法的性能评估.从实际应用的角度来看,训练集中的大多数抗原和抗体不会出现在测试集中,因此,为了验证模型在面对未见过的数据时能否保持良好表现,本文使用了两个独立测试集分别评估SMFF与其他基线方法的表现.结果分别如图5和图6所示,SMFF在两个独立数据集上的表现远超过其他基线方法,在预测未知的抗原-抗体对结合亲和力时达到0.8左右的Pearson相关系数和Spearman相关系数,比其他基线方法高出20%以上,表明提出的模型即便在小数据集上进行训练,也不会因为过拟合导致泛化能力降低.

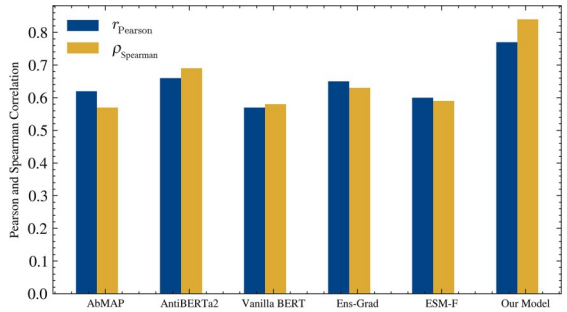

图 5 Antibody-Benchmark 独立集上的性能比较

Fig. 5 Performance comparison on Antibody-Benchmark independent set

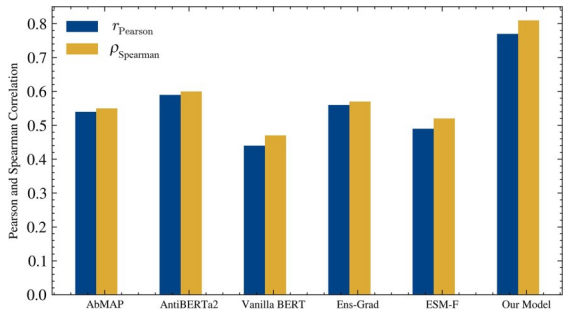

图 6 PPB-Affinity 独立集上的性能比较

Fig. 6 Performance comparison on PPB-Affinity independent set

4 讨论

4.1 消融实验 为了验证 SMFF 中每个模型的贡献和有效性,本文以三个基准数据集为基础进行消融研究,将完整 SMFF 模型的性能与其变体(分别移除预训练模型编码、交叉注意力机制和结构数据)进行比较,如表 2 所示,其中,表中黑体字表示结果最优,"±"后面的值表示 Pearson 和 Spearman 相关系数的标准误差. 需要注意的是,当结构数据被移除时,交叉注意力机制也变得无效. 当去除结构信息时,模型的性能显著下降,凸显了多模态信息,尤其是结构信息在抗原-抗体结合亲和力预测中的重要性. 同样,没有交叉注意力机制的模型表现也很差,这反映了多模态交互的重要性. 而对预训练编码向量的消融研究结果表明,预训练模型编码的序列信息嵌入和结构信息嵌入提供了有效且全面的特征信息,显著提升了模型性能. 总体而言,消融实验表明,模型中的多模态信息、模态交互和独特的多模态特征编码都是有效和必要的.

4.2 在突变限制数据集中的挑战及池化方法的分析 实验结果表明,SMFF 在所有数据集上都优于其他基线模型,其性能提升部分源于预训练模型,部分得益于结构信息与多模态交互的结合. 然而,如表 1 所示,SMFF 在 14H 和 14L 数据集上的表现略低于 BioMap 数据集,这可能是因为 LL-SARS-CoV-2 数据库仅包括抗体序列中的 1~3 个氨基酸突变,限制了模型对少数突变驱动的高亲和力变体的预测能力,这也是未来需要解决的挑战. 在 MF_CNN 结构中,本文采用了最大池化策略,并测试了平均池化和无池化的结果,如表 3 所示,其中,表中黑体字表示结果最优,"±"后面的值表示 Pearson 和 Spearman 相关系数的标准误差. 这些结果表明,在 MF_CNN 中实施的最大池化策略是有效的.

5 结论

本文提出一种名为 SMFF 的多模态深度学习模型,用于抗原-抗体结合亲和力的高精度预测. 该模型首次将抗体结构信息引入亲和力预测任务,并借助注意力机制有效捕捉不同模态间的

表 2 针对 SMFF 模型及其移除不同组件的变体在 14H, 14L 和 BioMap 数据集上的消融研究结果

Table 2 Ablation study results on 14H, 14L and BioMap datasets for the SMFF model and its variants with different components removed

	14H		14L		BioMap	
	$r_{Pearson}$	$\rho_{Spearman}$	$r_{Pearson}$	$\rho_{Spearman}$	$r_{Pearson}$	$\rho_{Spearman}$
SMFF	**0.640±0.013**	**0.552±0.016**	**0.677±0.015**	**0.686±0.015**	**0.697±0.026**	**0.726±0.024**
w/o Pre-trained	0.604±0.024	0.483±0.022	0.618±0.025	0.619±0.027	0.498±0.033	0.519±0.039
w/o Cross-Attention	0.626±0.014	0.544±0.011	0.663±0.017	0.674±0.011	0.677±0.023	0.691±0.026
w/o Structural-Data	0.612±0.019	0.520±0.013	0.648±0.018	0.655±0.014	0.640±0.025	0.686±0.021

表3　SMFF在14H, 14L和BioMap数据集上采用不同池化策略的表现

Table 3　The performance of SMFF with different pooling strategies on 14H, 14L and BioMap datasets

	14H		14L		BioMap	
	r_{Pearson}	ρ_{Spearman}	r_{Pearson}	ρ_{Spearman}	r_{Pearson}	ρ_{Spearman}
Maxpooling	**0.640±0.013**	**0.552±0.016**	**0.677±0.015**	**0.686±0.015**	**0.697±0.026**	**0.726±0.024**
Avgpooling	0.633±0.015	0.547±0.013	0.671±0.018	0.674±0.011	0.688±0.021	0.710±0.023
no pooling	0.612±0.021	0.526±0.026	0.644±0.023	0.653±0.025	0.663±0.033	0.684±0.029

内在关联.实验结果表明,SMFF在抗原-抗体结合亲和力预测中表现优异,性能显著优于现有主流方法.尽管该模型展现出良好的预测能力,仍存在若干可改进的方向.例如,当前模型仅集成抗体结构信息,未来可进一步引入抗原结构特征以增强表达能力,该方面也将作为后续研究的重点内容.

参考文献

[1] Hviid L, Lopez-Perez M, Larsen M D, et al. No sweet deal: the antibody-mediated immune response to malaria. Trends in Parasitology, 2022, 38(6): 428－434.

[2] Oostindie S C, Lazar G A, Schuurman J, et al. Avidity in antibody effector functions and biotherapeutic drug design. Nature Reviews. Drug Discovery, 2022, 21(10): 715－735.

[3] Rascio F, Pontrelli P, Netti G S, et al. IgE-mediated immune response and antibody-mediated rejection. Clinical Journal of the American Society of Nephrology, 2020, 15(10): 1474－1483.

[4] Posner J, Barrington P, Brier T, et al. Monoclonal antibodies: Past, present and future. Concepts and Principles of Pharmacology, 2019, 260: 81－141.

[5] 丁莉坤,樊婷婷,刘美佑,等.单克隆抗体药物临床监测研究进展.药物不良反应杂志,2022,24(9):484－489.

[6] 王志明,高健,李耿.治疗性单克隆抗体药物的现状及发展趋势.中国生物工程杂志,2013,33(6):117－124.

[7] 张天民.单克隆抗体药物的研究进展.医药导报,2005,24(6):494－498.

[8] Jin Y M, Schladetsch M A, Huang X T, et al. Stepping forward in antibody-drug conjugate development. Pharmacology & Therapeutics, 2022, 229: 107917.

[9] 康殷楠,陈顺,解有成,等.抗体药物偶联物在HER2阳性晚期胃癌中的应用进展和展望.中国癌症杂志,2023,33(8):790－800.

[10] 李博乐,冯红蕾,魏枫,等.肿瘤抗体药物偶联物的研发进展和挑战.中国肿瘤临床,2022,49(16):850－857.

[11] 唐思洁,糜坚青.抗体药物偶联物在血液肿瘤中的临床应用研究进展.上海交通大学学报(医学版),2024,44(12):1607－1614.

[12] 徐兵河,马飞,王佳玉,等.抗体药物偶联物治疗恶性肿瘤临床应用专家共识(2020版).中国医学前沿杂志(电子版),2021,13(1):1－15.

[13] Alhabbab R Y. Radioimmunoassay (RIA)//Basic Serological Testing. Cham: Springer International Publishing, 2018: 77－81.

[14] Tabatabaei M S, Ahmed M. Enzyme-linked immunosorbent assay (ELISA)//Christian S L. Cancer cell biology: Methods and Protocols. New York: Humana Press, 2022: 115－134.

[15] Sparks R P, Jenkins J L, Fratti R. Use of surface plasmon resonance (SPR) to determine binding affinities and kinetic parameters between components important in fusion machinery//Fratti R, Methods in Molecular Biology. New York: Humana Press, 2019: 199－210.

[16] Rhea K. Determining the binding kinetics of peptide macrocycles using bio-layer interferometry (BLI)// Coppock M B, Winton A J. Peptide Macrocycles. New York: Humana Press, 2022, 2371: 355－372.

[17] Myung Y, Pires D E V, Ascher D B. CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function. Bioinformatics, 2022, 38(4): 1141－1143.

[18] Cai H Y, Zhang Z B, Wang M K, et al. Pretrainable geometric graph neural network for antibody affinity maturation. Nature Communications, 2024, 15(1): 7785.

[19] Yuan Y, Chen Q S, Mao J, et al. DG-Affinity: predicting antigen-antibody affinity with language models from sequences. BMC Bioinformatics, 2023, 24(1):430.

[20] Kang Y, Leng D W, Guo J J, et al. Sequence-based deep learning antibody design for *In silico* antibody affinity maturation. (2022—8—15) [2025—9—2]. https://arxiv.org/abs/2103.03724.

[21] Abdine H, Chatzianastasis M, Bouyioukos C, et al. Prot2text: Multimodal protein's function generation with gnns and transformers. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(10):10757—10765.

[22] Hu F, Hu Y S, Zhang W H, et al. A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. Advanced Science, 2023, 10(22):e2301223.

[23] Su J, Han C C, Zhou Y Y, et al. Saprot: Protein language modeling with structure-aware vocabulary. (2024—04—01) [2025—9—2]. https://openreview.net/forum?id=6MRm3G4NiU.

[24] Wang X Y, Zheng Z X, Ye F, et al. Dplm-2: A multimodal diffusion protein language model. (2024—10—17) [2025—9—2]. https://arxiv. org/abs/2410.13782.

[25] Lin Z M, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379(6637): 1123—1130.

[26] Wang L, Zhang H, Xu W, et al. Deciphering the protein landscape with ProtFlash, a lightweight language model. Cell Reports Physical Science, 2023, 4(10):101600.

[27] Zhang Z B, Xu M H, Jamasb A, et al. Protein representation learning by geometric structure pretraining. 2022, arXiv:2203.06125.

[28] Hie B L, Shanker V R, Xu D, et al. Efficient evolution of human antibodies from general protein language models. Nature Biotechnology, 2024, 42(2): 275—283.

[29] Su J, Ahmed M, Lu Y, et al. Roformer: enhanced transformer with rotary position embedding. Neurocomputing, 2024, 568:127063.

[30] Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: Making protein folding accessible to all. Nature Methods, 2022, 19(6):679—682.

[31] Ruffolo J A, Sulam J, Gray J J. Antibody structure prediction using interpretable deep learning. Patterns, 2022, 3(2):100406.

[32] Abanades B, Georges G, Bujotzek A, et al. ABlooper: Fast accurate antibody CDR loop structure prediction with accuracy estimation. Bioinformatics, 2022, 38(7):1877—1880.

[33] Cohen T, Halfon M, Schneidman-Duhovny D. NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning. Frontiers in Immunology, 2022, 13:958584.

[34] Ruffolo J A, Chu L S, Mahajan S P, et al. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nature Communications, 2023, 14(1):2389.

[35] Suzek B E, Huang H Z, McGarvey P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007, 23(10):1282—1288.

[36] Punta M, Coggill P C, Eberhardt R Y, et al. The Pfam protein families database. Nucleic Acids Research, 2012, 40(Database issue):D290—D301.

[37] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Research, 1996, 24(1): 21—25.

[38] Chen B, Cheng X Y, Li P, et al. xTrimoPGLM: Unified 100b-scale pre-trained transformer for deciphering the language of protein. (2024—12—09) [2025—9—2]. https://arxiv.org/abs/2401.06199.

[39] Engelhart E, Emerson R, Shing L, et al. A dataset comprised of binding interactions for 104, 972 antibodies against a SARS-CoV-2 peptide. Scientific Data, 2022, 9(1):653.

[40] Guest J D, Vreven T, Zhou J, et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. Structure, 2021, 29(6):606—621, e5.

[41] Liu H Q, Chen P Y, Zhai X C, et al. PPB-Affinity: Protein-protein binding affinity dataset for AI-based protein drug discovery. Scientific Data, 2024, 11(1): 1316.

[42] Singh R, Im C, Qiu Y, et al. Learning the language of antibody hypervariability. Proceedings of the

National Academy of Sciences of the United States of America,2025,122(1):e2418918121.

[43] Barton J, Galson J D, Leem J. Enhancing antibody language models with structural information. BioRxiv,(2024－01－04)[2025－9－2]. https://doi.org/10.1101/2023.12.12.569610.

[44] Liu G, Zeng H Y, Mueller J, et al. Antibody complementarity determining region design using high‐capacity machine learning. Bioinformatics, 2020,36(7):2126－2133.

(责任编辑　高善露)