

**大学生创新训练项目**

**——研究报告**

**基于深度残差交互机制的TCR-****抗原识别模型**

院 系 人工智能与自动化学院

专业班级 智能医学工程2201班

姓 名 杨子涵、薛卜元

学 号 U202215124 U202215123

指导教师 刘雪明

2025年4月6日

**目录**

1 项目概况 2

1.1 研究目的 2

1.2 研究现状 2

1.3 数据调研 2

1.4 实施方案 3

1.5 项目成员 3

2 数据分析与预处理 4

2.1 数据集分析 4

2.2 数据集预处理 6

2.2.1 BLOSUM62处理 6

2.2.2 生成负样本 7

2.2.3 转换数值表示 8

3 模型构建 9

3.1 模型结构 9

3.2 模型优势 10

4 训练过程 11

4.1 代码结构 11

4.1 训练效果 11

参考文献 12

# 1 项目概况

在当今社会，深度学习技术在预测T 细胞和抗原相互作用方面取得了显著的进展，为癌症免疫治疗等领域提供了强大的工具。然而，目前大多数研究如pMTnet-omni、DeepTCR 集中在CD8 T 细胞和MHC-I 组织相容物的结合预测上，即用于预测癌细胞抗原。相反，关于CD4 T 细胞和MHC-II 组织相容物的相互作用的预测方法却相对较少。

本项目的研究目的在于弥补这一不足，通过引入迁移学习的策略，将已经在预测CD8 T 细胞和MHC-I 组织相容物相互作用方面取得成功的模型迁移到CD4 T 细胞和MHC-II 组织相容物的相互作用的预测中。

## 研究目的

T细胞受体和抗原表位的相互作用是机体免疫系统的重要组成部分，这种相互作用可以激活T细胞，从而识别和攻击感染机体的病原体。因此，研究T细胞受体和抗原表位的结合位点是理解机体免疫系统如何工作的关键。

T细胞在适应性免疫反应中对表位的识别是至关重要的。这种识别是通过主要组织相容性复合体（MHC）分子呈递的致病抗原（epitope）与T细胞受体（TCR）相互作用来实现的，从而激活细胞介导的免疫，消除感染细胞或活化相应的免疫细胞。然而，由于这种识别机制的内在复杂性，TCR-epitope相互作用的实验检测和确定通常既费时又昂贵。因此，为了解决这一时间成本和经济成本问题，我们可以开发深度学习模型来辅助我们预测T细胞受体和表位之间的相互作用。

## 研究现状

T细胞在适应性免疫反应中对表位的识别是至关重要的。这种识别是通过主要组织相容性复合体分子呈递的致病抗原与T细胞受体相互作用来实现的，从而激活细胞介导的免疫，消除感染细胞或活化相应的免疫细胞。因此，深入了解TCR-epitope的结合机制对于癌症免疫学、自身免疫抗原发现和疫苗设计具有重要意义。然而，由于这种识别机制的内在复杂性，TCR-epitope相互作用的实验检测和确定通常既费时又昂贵。

为了解决这一问题，清华大学交叉信息研究院曾坚阳团队成功开发了能预测T细胞受体和抗原表位相互作用的深度学习模型，相关成果以“利用深度学习表征T细胞受体和表位之间的相互作用构象”（Characterizing the interaction conformation between T cell receptors and epitopes with deep learning）为题，于3月27日发表在《自然》（Nature）子刊《自然·机器智能》（Nature Machine Intelligence）上。

## 数据调研

在TCR与表位预测的研究中，多个数据库提供了丰富的数据资源。IEDB是全球最大的免疫表位数据库，涵盖了广泛的TCR与表位结合数据；VDJdb专注于TCR序列及其靶标信息，适合机器学习模型的训练；MCPAS聚焦于与疾病相关的TCR序列，特别适合特定疾病的免疫应答研究；PIRD收录了不同个体和环境下的TCR和BCR数据，支持多样性研究；IMGT则提供标准化的TCR和免疫球蛋白数据，有助于精准的生物信息学分析。这些数据库的数据质量、覆盖面和应用场景各不相同，为构建高性能的TCR与表位预测模型提供了坚实的基础。

目前关于TCR与表位的三维结构数据相对较少。这是因为获取高质量的三维结构数据需要复杂且昂贵的实验过程，如X射线晶体衍射、冷冻电子显微镜（Cryo-EM）和核磁共振（NMR）等。这些方法耗时较长，并且并非所有的TCR与表位复合物都能被成功解析。

## 实施方案

搜集一定量关于T细胞受体和抗原表位的相互作用相关的文献和数据；对文献进行整理与汇总，对文献中建立的指标体系、运用的数学模型等进行提炼并分析；使用 Python编写相关代码，并进行实验，根据相关文献选用合适的性能参数指标。优化得到的模型，获得性能良好的犯罪热点预测模型。

## 1.5 项目成员

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 角色 | 姓名 | 学号 | 专业 |
| 1 | 组长 | 杨子涵 | U202215124 | 智能医学工程 |
| 2 | 成员 | 薛卜元 | U202215123 | 智能医学工程 |

# 2 数据分析与预处理

## 2.1 数据集分析

原始数据集由IEDB（Immune Epitope Database，免疫表位数据库）下载，IEDB是全球最权威和常用的免疫学数据库之一，尤其在 T细胞、B细胞、MHC研究中应用广泛，其数据范围完美满足了我们此抗原识别模型的数据集要求，且数据集的可靠性、数据量、可获得性都能得到很高的保障。

抗原识别模型的优劣可由预测法的评价标准来衡量，优秀的抗原识别模型应在准确性和平衡性上表现优秀、能预测出符合生物规律的新组合、有良好的泛化能力、对正负样本和类不平衡情况有鲁棒性的评价标准。其中，准确性和平衡性反映在F1-score、AUC等指标上，高F1-score表明模型在精确率和召回率之间达到了较好平衡，AUC则代表了模型整体“排序能力”，AUC越高，说明模型越擅长区分正负样本。

之后，对原始数据集使用EDA分析，初步得到以下结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | CDR3 | Epitope | V | J | MHC A | MHC B |
| count | 4029 | 4029 | 4029 | 4029 | 4029 | 4029 |
| unique | 3009 | 114 | 99 | 77 | 25 | 47 |
| freq | 79 | 879 | 544 | 553 | 1135 | 879 |

表 1 EDA分析结果

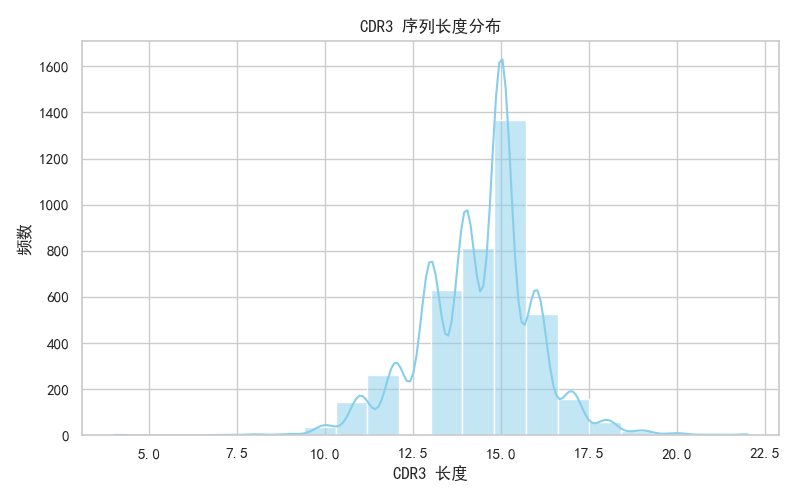


图 1 CDR3序列长度分布

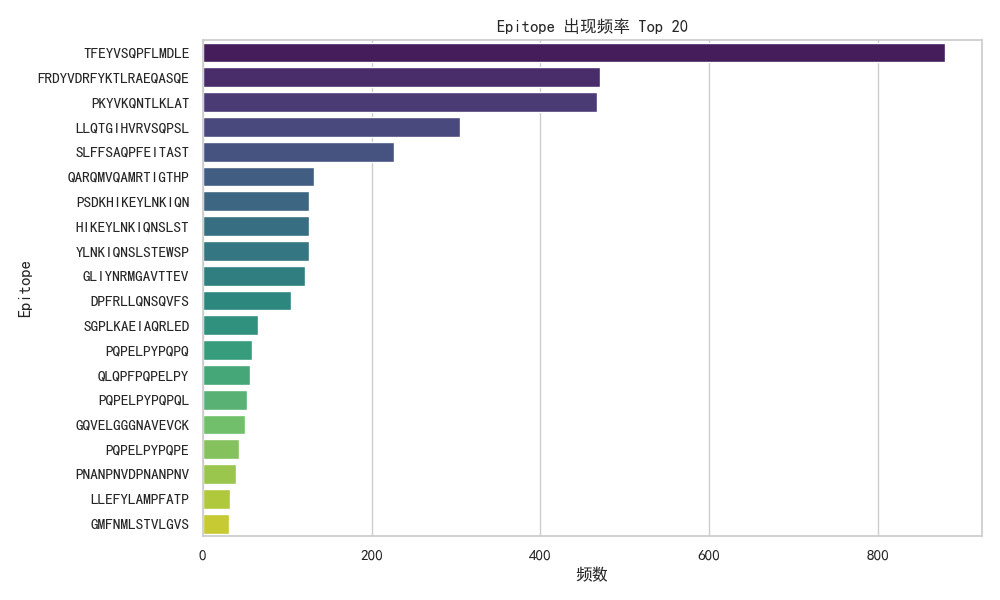


图 2 Epitope出现频率

图表, 条形图

AI 生成的内容可能不正确。

图 3 V区基因使用频率

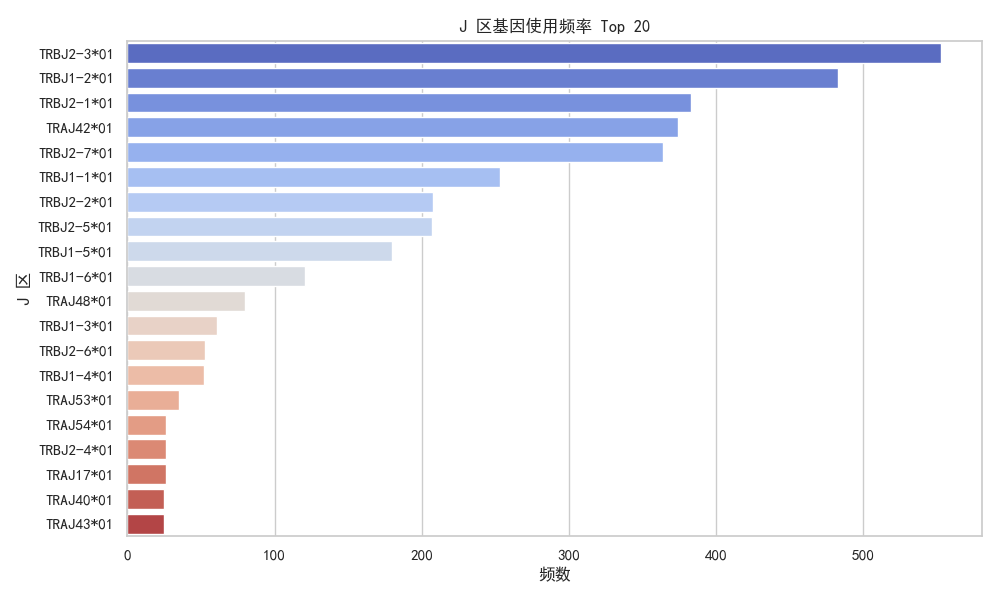


图 4 J区基因使用频率

图表, 条形图, 漏斗图

AI 生成的内容可能不正确。

图 5 MHC频率

## 2.2 数据集预处理

### 2.2.1 BLOSUM62处理

BLOSUM62（Blocks Substitution Matrix 62）是生物信息学中一种重要的替换矩阵，主要用于蛋白质序列比对和相似性评估。它为为每对氨基酸之间的替换提供了一个得分值，这些得分反映了不同氨基酸之间的相似性或可替换性，得分反映了氨基酸在进化过程中的保守性，得分越高表示这两种氨基酸在进化上越相似。通过评估序列相似性，可以推断未知蛋白质的结构和功能以及帮助识别不同物种间的同源蛋白。

在本项目中，BLOSUM62为原始数据集中的每一对氨基酸分配一个得分，得分值反映了在自然选择下这种替换发生的可能性。替换分数高（正得分）通常表示该替换在进化中较为常见，而负得分则表示这种替换不常见或可能破坏蛋白质的结构和功能。

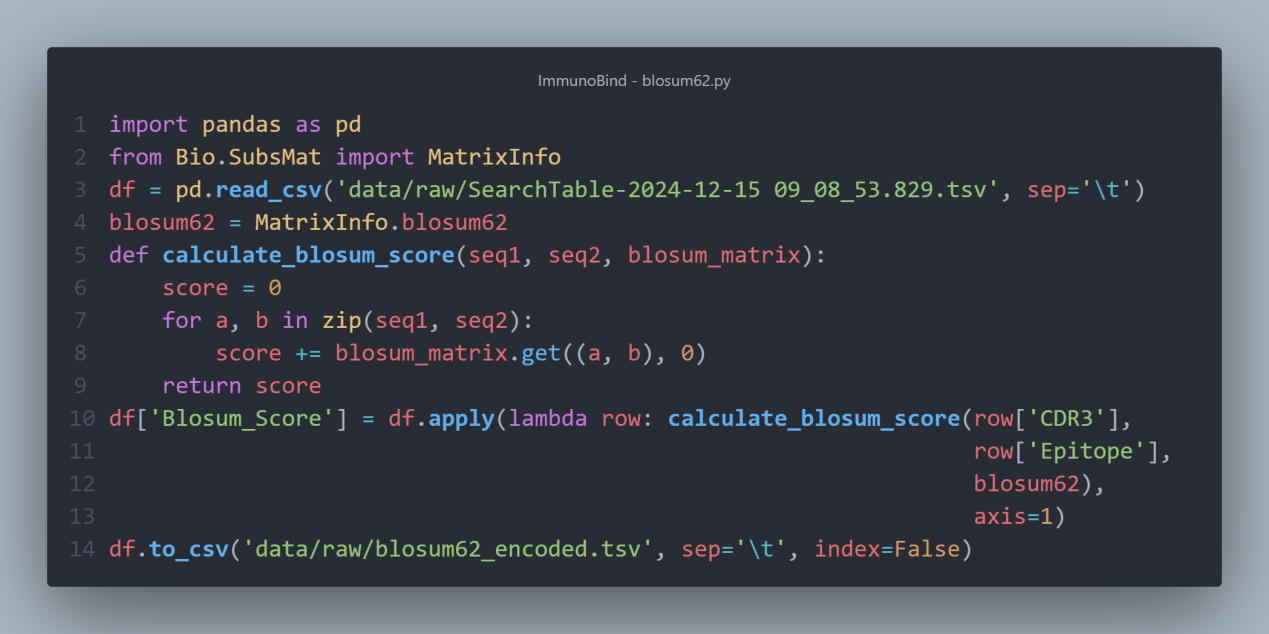


图 6 BLOSUM62处理

### 2.2.2 生成负样本

由IEDB下载的数据集均为正样本（标签1），每一对CDR3与表位信息一一匹配，以便构造训练数据集或进行其他后续分析。生成负样本的方法是将原本成对出现的两个特征中的第二个特征进行随机排列，打乱原始的正确配对关系，从而得到错误或不匹配的配对，并将这些配对统一标记为负样本（标签0）。

样本的正负比可以通过修改函数常数来实现，以便在后续训练过程中进行实时修改。



图 7 正负样本示例

### 2.2.3 转换数值表示

使用预训练的BERT模型（TCR-BERT和ProtBERT）将CDR3和Epitope序列转换为嵌入向量（embedding），并将结果保存为新的TSV文件，形成最终用于训练的文件。将免疫序列转换成深度学习模型可用的数值表示（embedding），用于后续的机器学习或可视化任务。



图 8 转换数值表示

# 3 模型构建

## 3.1 模型结构

融合了预训练模型、注意力机制、残差连接和Transformer结构，适合处理序列对（如 TCR和Epitope）的表征与交互建模。

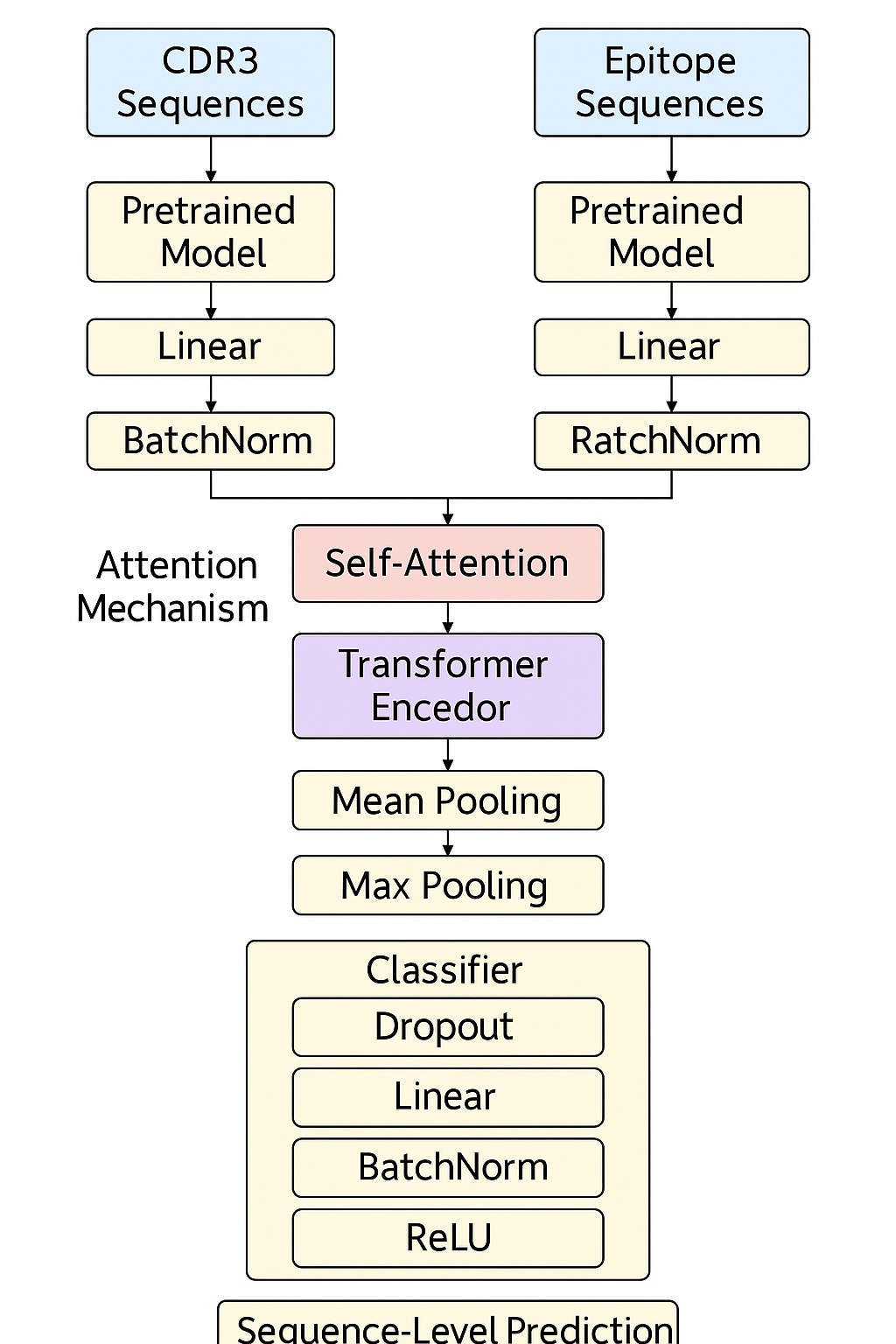


图 9 模型结构图

## 3.2 模型优势

首先，通过采用预训练模型分别对CDR3和Epitope序列进行编码，模型能够利用预先在大规模数据上学习到的丰富表征信息。这种处理方式使得模型在处理有限样本、复杂生物序列时有更好的初始化表现和泛化能力，为后续的特征融合打下坚实基础。

另外，模型在特征提取阶段使用了多层的线性映射、批归一化和ReLU激活函数，这种结构不仅加强了非线性建模能力，还能缓解梯度消失问题，提升训练的稳定性。在交互层面，模型通过点积注意力机制捕捉了不同序列间的互相关系，再结合自注意力对单模态内部进行强化，从而有效融合了两种不同来源的信息。这种设计充分利用了注意力机制在捕捉长距离依赖关系和重要特征方面的优势。

此外，模型引入了Transformer编码器层，这不仅让融合后的特征经过多层非线性变换进一步提取深层次的语义信息，同时也借助Transformer在建模序列间关系上的强大能力，对整个序列对的互动进行全局考量。在池化阶段采用了平均池化和最大池化的组合方式，更全面地捕获了序列信息的全局趋势和局部细节，进一步增强了模型对于最终特征表达的区分度和鲁棒性。

# 4 训练过程

## 4.1 代码结构

首先加载和预处理数据，包括数据增强和数据集划分；接下来构建数据加载器、初始化模型以及定义损失函数、优化器和学习率调度器；然后进入训练循环，在每个 epoch 中进行训练和验证，记录指标并依据验证结果进行早停判断；最后保存最佳模型。这样的结构清晰地展示了数据从输入、预处理、训练、评估到模型保存的完整流程。

训练循环部分采用了梯度累积策略，每累计一定批次再进行一次参数更新，以应对较大的批次需求和显存限制。循环中，使用了自动混合精度加速训练，并加入梯度裁剪以防止梯度爆炸。每个epoch内不仅完成了训练，还通过计算ROC-AUC等指标来实时评估模型在验证集上的表现。为了避免过拟合和资源浪费，代码引入了早停机制，在验证指标连续多轮未提升后自动停止训练，同时保存了性能最优的模型。训练结束后，最终加载最优模型并保存到指定路径，从而为后续模型部署或进一步调优提供了基础。

## 4.1 训练效果

**准确性与平衡性指标：**经过多轮训练，模型在测试集上的准确度达到了 0.75，表明模型在精确率和召回率之间取得了良好的平衡，能够识别出 T 细胞受体和抗原表位的相互作用关系，同时也能有效避免漏判情况。平均AUC 值为 0.69，说明模型在区分正负样本方面表现较差，这或许是因为数据集中TCR-Epitope对并非为特异性对应关系。

**模型对比分析：**选择当前研究领域内的其他先进模型，如 pTCRα-ensemble、SVM 等作为对比对象。对比在相同测试集上各模型的关键指标。

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Feature | ACC(%) | AUC |
| **pTCRα-ensemble** | **One-hot,15PCs** | **81.91** | **0.914** |
| SVM | 15PCs | 79.26 | 0.793 |
| RF | One-hot | 83.07 | 0.925 |
| **ImmunoBind** | **Bert-coding** | **75.64** | **0.695** |
| NB | 15PCs | 64.85 | 0.725 |
| LB | One-hot | 66.93 | 0.740 |
| AdaBoost | 15PCs | 64.52 | 0.700 |

表格 2 模型对比分析

# 参考文献

[1]Sidhom, J. W., Larman, H. B., Pardoll, D. M., & Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T - cell repertoires[J]. Nature Communications, 2021, 12(1): 1605.

[2]Han, Y., Yang, Y., Tian, Y., et al. Pan - MHC and cross - species prediction of T cell receptor - antigen binding[J]. bioRxiv, 2023. DOI: 10.1101/2023.12.01.569599.

[3]曾坚阳团队.Characterizing the interaction conformation between T cell receptors and epitopes with deep learning[J]. Nature Machine Intelligence, 2023.

[4]Immune Epitope Database (IEDB)[EB/OL].[n.d.].[https://www.iedb.org/].

[5]Wang, X., & Zhang, Y. Advances in T - cell receptor - antigen interaction prediction methods[J]. Journal of Immunology Research, 2022, 2022: 8292793.

[6]Baldi, P., & Longo, L. Deep learning in immunology: From sequence to function[J]. Cellular and Molecular Immunology, 2021, 18(9): 1953 - 1965.

[7]Karimi, M., & Khorshidi, A. A review of machine learning algorithms for predicting T - cell receptor - antigen interactions[J]. Briefings in Bioinformatics, 2020, 21(5): 1496 - 1509.

[8]Wu, K., & Yang, J. TCR - BERT: A pre - trained model for T - cell receptor sequence analysis[J]. Bioinformatics, 2021, 37(17): 2531 - 2538.

[9]Rost, B., & Söding, J. Protein language models: A new generation of sequence analysis tools[J]. Nature Reviews Microbiology, 2019, 17(10): 637 - 648.