



计算差异视角下MHC-I与MHC-II的 TCR-pMHC结构预测资源概述与测评*

吴晓芹¹⁾ 刘大伟²⁾ 李玢瑜²⁾ 刘 扬^{3,4)} 曹 洋^{3,4)**} 戴文韬^{1,2)**}

(¹⁾ 上海理工大学健康科学与工程学院, 上海 200093;

²⁾ 复旦大学药学院, 上海市疾病与健康基因组学重点实验室, 国家卫生健康委员会计划生育药具重点实验室,
上海市生物医药技术研究院, 上海 200237;

³⁾ 四川大学生命科学学院, 生物资源与生态环境教育部重点实验室, 成都 610065;

⁴⁾ 四川大学生命科学学院, 四川省动物疫病防控与食品安全重点实验室, 四川省微生物与代谢工程重点实验室, 成都 610065)

摘要 T细胞受体(T cell receptor, TCR)对肽-主要组织相容性复合物(peptide-major histocompatibility complex, pMHC)的特异性识别,是触发抗原特异性免疫应答的关键分子事件之一。因此,解析TCR-pMHC复合物的结构特征,不仅为阐明抗原特异性免疫识别机制奠定基础,也为疫苗设计、TCR免疫治疗靶点研究,以及TCR鉴定与优化提供重要依据。然而,由于实验测定成本高、周期长且覆盖范围有限,通过计算方法快速获得可靠的TCR-pMHC复合物结构已成为免疫学研究的重要诉求。近年来,随着深度学习的不断发展,加之结构和序列资源日益丰富,涌现出多种TCR-pMHC计算建模工具。但是,多数工具在建模时未对MHC-I类与MHC-II类复合物之间的关键差异加以区分,导致二者的建模效果存在明显差别。这些工具对I类复合物的预测通常优于II类复合物,这一现象可能源于二者在肽结合槽构型、肽段长度范围及肽段侧翼残基(peptide flanking residues, PFRs)特性等方面的不同。此外,TCR识别界面中,互补决定区(complementarity determining regions, CDRs)通常存在构象柔性,不少工具为追求整体构象稳定性或降低结构噪声,往往过度约束了本应具有柔性的区域。这种过度约束可能导致柔性CDRs被不当刚化,从而引发局部结构畸变、界面几何失真,甚至导致复合物建模失败。为此,本综述基于MHC-I和MHC-II的计算差异视角,系统整理并归纳了TCR和pMHC相关资源,包括结构数据库、序列数据库和结构预测工具。随后构建了包含25个I类和10个II类TCR-pMHC复合物数据集,对AlphaFold2、AlphaFold3、TCRmodel2、tFold-TCR及TCR-pHLA_ModellerS五种能同时预测I类和II类复合物的代表性工具开展系统评估。通过计算全原子RMSD(All-Atom Root Mean Square Deviation, All-Atom RMSD)、主链RMSD(Backbone RMSD)、模板建模得分(Template Modeling score, TM-score)和DockQ,衡量各工具在MHC-I类和MHC-II类复合物整体结构及结合界面建模的准确性。针对MHC-II类复合物,创新性地提出PFR构象偏差指数,包括PFRs-Deviation Index(PFRs-DI)、N-PFR-Deviation Index(N-PFR-DI)与C-PFR-Deviation Index(C-PFR-DI),以评估PFRs的构象准确性。同时,提出CDR构象一致性指数(CDR conformational consistency, CCC),定性评估预测工具刻画TCR CDRs构象柔性的能力。上述指标共同用于评价工具在整体构象与关键区域的建模能力,在一定程度上弥补了通用蛋白质结构模型评测指标偏重整体构象,难以准确反映关键功能区域建模质量的不足,从而构建适用于MHC-I类与MHC-II类复合物的分析框架,用于指导数据资源选择、建模策略制定与评价体系构建,进一步推动计算建模的发展,为多尺度解析TCR-pMHC识别机制及其生物学功能提供重要支撑。

关键词 免疫学, TCR-pMHC复合物, 结构预测, PFR构象偏差指数, CDR构象一致性

中图分类号 Q71, R392

DOI: 10.3724/j.pibb.2026.0015

CSTR: 32369.14.pibb.20260015

T细胞受体(T cell receptor, TCR)对肽-主要组织相容性复合物(peptide-major histocompatibility complex, pMHC)的特异性识别,是T细胞适应性免疫应答发挥功能的分子基础。免疫应答是机体识别并清除外源病原体、维持免疫稳态以及避免自身组织损伤的核心生物学过

* 上海市生物医药技术研究院省部级重点实验室创新能力专项(RC2023-01, DT2025-05), 上海科学院关键共性技术专项(SKY2022003)和国家自然科学基金(32470696)资助。

** 通讯联系人。

曹洋 Tel: 028-85410043, E-mail: cao@scu.edu.cn

戴文韬 Tel: 021-64771220, E-mail: daiwentao@sibpt.cn

收稿日期: 2026-01-09, 接受日期: 2026-04-02

程, 其精准调控对于个体生存与健康具有重要意义。在适应性免疫中, 抗原信息的有效呈递与解读是免疫应答启动的前提。早期抗原呈递机制的研究主要聚焦于肽段与MHC分子的结合特性, 旨在阐明肽-MHC相互作用在抗原加工、呈递及免疫原性筛选中的基础作用, 为理解免疫识别奠定了重要理论基础^[1-3]。在此研究背景下, Zhang等^[1]合作开发了面向MHC-I类分子的肽段结合亲和力预测方法ConvNeXt-MHC, 为相关研究提供了可用的计算工具与方法支持。

然而, 随着研究的深入, 发现仅基于肽-MHC的结合特性难以充分诠释免疫应答机制。这一过程依赖于TCR特异性识别MHC所呈递的肽段, 并在共刺激信号的协同作用下触发下游信号, 诱导T细胞发挥免疫功能, 使机体能够在复杂的抗原环境中区分自身与非自身信号, 进而对携带特定抗原肽的靶细胞产生应答^[4-5]。相较于只承载抗原信息的肽-MHC复合物, TCR-pMHC复合物在功能层级更直接地关联免疫识别与免疫效应, 是解析T细胞免疫应答调控机制、理解免疫多样性及指导免疫治疗与疫苗设计的核心研究对象之一^[6-8]。因此, 当前免疫学的研究焦点已从早期局限于肽段与MHC的二元相互作用, 逐步拓展至TCR、肽段和MHC所构成的三元复合物识别机制。

TCR-pMHC复合物作为连接抗原呈递与T细胞功能应答的关键功能单元, 其识别机制、构象柔性及分子动力学特征, 共同决定了抗原识别的特异性与敏感性, 同时直接调控T细胞激活阈值、信号传导强度、功能分化方向, 以及免疫应答强度^[9-12]。因此, 对TCR-pMHC识别机制的系统性解析, 不仅有助于阐明免疫识别的分子基础, 同时也为免疫治疗策略的优化与疫苗的精准设计提供方向。例如: 在抗感染免疫中, 对TCR-pMHC识别机制的解析能够清晰揭示T细胞对病原相关抗原的识别规律; 在肿瘤治疗领域, 基于其特异性识别原理, 可优化T细胞受体工程化T细胞(T cell receptor-engineered T cell, TCR-T)疗法和TCR融合蛋白药物以提升肿瘤靶向精准度与疗效, 同时为不同癌症类型开发个性化疫苗, 增强治疗的灵活性与靶向性; 在自身免疫疾病研究中, 对免疫识别与耐受调控失衡等异常机制的解析, 有助于阐明其发病基础, 并支持相关疾病的诊断与治疗方案的创新^[9, 11, 13-16]。

根据MHC分子的不同类别, TCR-pMHC复合

物可分为MHC-I类和MHC-II类。MHC-I类分子几乎分布于所有有核细胞表面, 主要负责向CD8+ T细胞递呈来源于细胞内的内源性抗原, 从而参与抗病毒和抗肿瘤免疫监视^[17-18]; MHC-II类分子主要表达于树突状细胞、巨噬细胞和B细胞等抗原递呈细胞, 主要向CD4+ T细胞递呈来源于胞外摄取的外源性抗原, 在适应性免疫应答的激活与调控中发挥关键作用^[19-20]。二者不仅在胞内抗原加工路径上有很大不同, 其结构也存在显著差异, 具体体现在肽结合槽、肽段长度以及肽段侧翼残基(peptide flanking residues, PFRs)等方面^[21-23]。MHC-I类复合物的肽结合槽两端相对封闭, 适合结合较短肽段(通常为8~10个氨基酸)。而MHC-II类复合物的肽结合槽在两端相对开放, 使其能容纳更长的肽段(通常为13~25个氨基酸)^[22, 24]。肽段延伸至结合槽外侧形成肽侧翼残基, 参与结合肽构象调节, 并在一定程度上影响复合物的整体结构^[25]。

此外, 在TCR识别界面中, 互补决定区(complementary-determining regions, CDRs)的构象柔性, 在不同TCR之间从相对稳定的刚性到灵活多变的柔性有着很大变动范围, 其构象柔性是TCR识别与适应不同配体的重要结构基础^[26-27]。基于这一特性, TCR在与配体结合后可能会在结合界面产生构象变化, 从而改变TCR-pMHC复合物的整体结构^[28-29]。这不仅给TCR-pMHC复合物的精确建模带来了巨大挑战, 也直接限制了计算辅助TCR靶向免疫治疗的评估与优化^[30-32]。

随着高通量免疫组库测序技术的成熟以及冷冻电镜和X射线晶体学等结构解析手段的发展, 大规模的TCR、pMHC及TCR-pMHC复合物数据资源得以持续积累^[33-35]。根据不同研究重点, 相关数据库资源可被分为4类: 复合物结构与界面注释数据库(ATLAS^[36]、The Structural T-cell Receptor Database (STCRDab)^[37]、TCR3d 2.0^[38]和TRAIT^[39]); 免疫表位与MHC结合肽数据库(EPIMHC^[40]、Immune Epitope Database (IEDB)^[41]和MHCBN^[42]); HLA分子专用数据库(Alelle Frequency Net Database (AFND)^[43]、HLA Ligand Atlas^[44]和IPD-IMGT/HLA^[45]); 以及TCR序列与抗原特异性数据库(McPAS-TCR^[46]、Pan Immune Repertoire Database (PIRD)^[47]、TCellAtlas^[48]、TCRdb 2.0^[49]和VDJdb^[50])。除数据库资源外, 还包含一系列用于解析TCR免疫组库的计算资源, 如CATT^[51]、DeRR^[52]、

STCAT^[48]、TCellSI^[53]和TCRoretta^[54]。这些资源共同为计算建模提供了丰富的训练样本^[55]，进一步推动了TCR单体与TCR-pMHC复合物建模工具的发展。

在TCR单体建模方面，现有建模方法包括同源建模与深度学习。同源建模的代表性工具包括LYRA^[56]、Repertoire Builder^[57]和TCRmodel^[58]。这些方法通过序列比对，从已解析结构数据库中筛选同源模板，随后将目标序列映射到模板结构中以预测主链骨架构象，在此基础上对CDRs高变区域进行局部构象建模，并进一步完成侧链构建与优化，从而实现TCR单体结构预测。基于深度学习建模的工具包括ImmuneFold TCR^[59]、TCRBuilder2^[60]和tFold-TCR^[61]。这类方法基于大规模蛋白质序列与结构数据进行训练，通过学习序列特征与三维结构之间的映射关系，推断残基间的空间几何距离、取向和三维坐标，实现在低同源或无模板条件下进行结构建模。

当前的TCR-pMHC复合物结构建模工具，可根据研发目标场景分为通用蛋白质结构预测模型和TCR-pMHC复合物结构建模专用工具两类。通用蛋白质结构预测模型以AlphaFold2^[62]与AlphaFold3^[63]为代表，凭借强大的结构特征提取能力在多类蛋白质复合物预测中占据重要地位。然而，考虑到TCR-pMHC结合界面具有高度的柔性和特异性，仅依赖通用模型往往难以捕捉其关键构象特征^[26, 64]，一系列针对TCR-pMHC复合物优化的工具应运而生，包括ImmuneScape^[65]、TCRdock^[66]、TCRmodel2^[67]、TCRpMHCmodels^[68]、TCR-pHLA_ModellerS^[69]和tFold-TCR^[61]。其中，TCRmodel2、TCRpMHCmodels和TCR-pHLA_ModellerS利用已解析的复合物结构作为模板，通过序列比对构建目标复合物结构。ImmuneScape在利用模板结构特征确定复合物整体构象的基础上，通过分子对接与界面精修对TCR-pMHC结合界面进行局部优化。TCRdock基于改造的AlphaFold框架，在引入TCR-pMHC复合物结构模板的基础上，对复合物进行端到端结构预测。tFold-TCR基于蛋白质语言模型(protein language model, PLM)的端到端三维结构预测方法，可从输入序列中提取链内与链间相互作用关系及进化相关信息，并结合结构预测与柔性对接模块生成TCR-pMHC复合物结构。

尽管已有研究从整体结构、局部构象及结合界

面等维度对多种TCR-pMHC复合物建模工具进行了系统评测^[70-71]，但当前的评估体系仍存在一定局限。一是评估方法未能充分反映MHC-I类与MHC-II类复合物间的关键差异；二是现有评估指标主要关注全局或核心肽段结构，难以刻画PFRs的局部构象偏差；三是未能评估不同工具在CDRs区的构象柔性区分能力。因此，现有的评测只能在一定程度上反映粗粒化结构建模质量，尚不足以全面评估当前建模工具的实用性能和优势，而对MHC-I类与MHC-II类复合物进行区别讨论、综合考虑PFRs结构与CDRs构象柔性特征，对TCR-pMHC复合物建模工具评测有着重要意义。本次工具评测，期望构建能反映TCR-pMHC复合物关键结构特征的评估体系，为建模工具开发提供参考，从而推动疫苗设计、肿瘤新抗原筛选及个体化免疫治疗靶点的深入研究。

在此背景下，本文从MHC-I与MHC-II的计算结构差异出发，总结了TCR-pMHC相关的数据资源和建模方法，并对代表性建模工具进行了综合评测。通过常用评价指标全原子RMSD (All-Atom Root Mean Square Deviation, All-Atom RMSD)、主链RMSD (Backbone RMSD)、模板建模得分(Template Modeling score, TM-score)以及DockQ，从整体结构与界面构象两个层面评估不同工具对MHC-I类和MHC-II类复合物的建模性能。对于传统评价指标难以反映MHC-II复合物中肽段侧翼残基构象特征的问题，本文进一步提出PFRs偏差指数(PFRs-Deviation Index, PFRs-DI)、N端PFR偏差指数(N-PFR-Deviation Index, N-PFR-DI)和C端PFR偏差指数(C-PFR-Deviation Index, C-PFR-DI)，以衡量不同工具对MHC-II结合肽两端构象的建模能力，从而弥补传统界面质量指标的不足。围绕TCR CDRs构象柔性刻画困难的问题，引入CDR构象一致性指数(CDR conformational consistency, CCC)，用于定性评估不同建模工具对TCR CDRs刚性与柔性构象的区分能力。通过系统比较各工具在结合界面、整体结构、PFRs复现能力和TCR CDRs构象柔性区分能力的表现以及建模过程的计算资源消耗，为MHC-I类与MHC-II类复合物的建模策略制定与评价体系构建提供统一的分析框架与可操作的科学依据，有望推动相关计算方法的持续迭代升级，并为感染性疾病与自身免疫性疾病中异常TCR识别机制的阐明提供重要支持，同时也为肿瘤免疫治疗中TCR

靶点筛选与亲和力优化策略的设计提供方法学支撑。

1 数据资源

在 TCR-pMHC 数据资源（表 1）中，根据

TCR、抗原肽与 MHC 分子的不同侧重，现有数据库可归为 4 类：复合物结构与界面注释数据库、免疫表位与 MHC 结合肽数据库、HLA 分子专用数据库，以及 TCR 序列与抗原特异性数据库。

Table 1 Overview of TCR-pMHC data resources
表1 TCR-pMHC数据资源

数据库类型	数据库名称	资源概述与特色	是否包含MHC分子与 TCR-pMHC复合物	资源链接及更新日期
复合物结构与 界面注释 数据库	ATLAS ^[36]	ATLAS数据库是由人工整理的 数据资源，系统收录了野生型 与突变型TCR及其抗原之间的 结合亲和力信息。用户可以在 数据库中检索TCR、肽段和 MHC分子，获取其亲和力数 据、结构特征以及相关实验设 计与测定条件等详细信息	是	DOI: 10.1002/prot.25260 链接: https://atlas.ibbr.umd.edu/ web/ 最新更新日期: 2017年2月3日
	The Structural T-cell Receptor Database (STCRDab) ^[37]	STCRDab数据库每周自动收集 Protein Data Bank (PDB) ^[72] 中的TCR结构数据，并提供注 释，如α/β或γ/δ链配对、结合 的MHC分子，以及抗原结合亲 和力。此外，还提供了TCR可变 区的三维空间相对取向及其夹 角关系	是	DOI: 10.1093/nar/gkx971 链接: https://opig.stats.ox.ac.uk/ webapps/stcrdab-stcrpred/ 最新更新日期: 每周自动收集和 整理
	TCR3d 2.0 ^[38]	TCR3d数据库可以查看所有经 过实验确定的TCR-pMHC复合 物的结构。此外，还提供相关 工具，可基于输入的TCR或 TCR-mimic复合物结构预测对 接构象	是	DOI: 10.1093/nar/gkae840 链接: https://tcr3d.ibbr.umd.edu/ faq 最新更新日期: 2024年9月26日
	TRAIT ^[39]	TRAIT数据库汇总了数百万个 经实验验证的TCR-抗原对，并 提供序列、结构和亲和力数据， 同时还包含部分抗原突变数据	是	DOI: 10.1093/gpbjnl/qzaf033 链接: https://pgx.zju.edu.cn/ traitdb/ 最新更新日期: 2025年7月11日
免疫表位与 MHC结合肽 数据库	EPIMHC ^[40]	EPIMHC数据库，记录了在真 实蛋白质结构中观察到的MHC 结合肽和T细胞表位。目前，数 据库包含4 867个不同来源的肽 序列，其中有84种与肿瘤相关 的抗原	主要包含MHC-I类分子	DOI: 10.1093/bioinformatics/ bti269 链接: http://imed.med.ucm.es/ epimhc/ 最新更新日期: 2016年
	Immune Epitope Database (IEDB) ^[41]	IEDB数据库收录了人类及其他 物种的抗体与T细胞表位实验数 据，涉及传染病、过敏、自身 免疫及器官移植等领域。该平 台同时集成了一系列表位预测 与分析工具，并提供配套的在 线资源	是	DOI: 10.1093/nar/gkae1092 链接: https://www.iedb.org/ 最新更新日期: 2025年1月6日

续表1

数据库类型	数据库名称	资源概述与特色	是否包含MHC分子与TCR-pMHC复合物	资源链接及更新日期
HLA分子专用数据库	MHCBN ^[42]	MHCBN是综合性数据库，收录了超过25 857条肽段，其中有1 053条已经过实验验证可以被抗原加工相关转运体（transporter associated with antigen processing, TAP）分子识别并结合的多肽	主要包含MHC-I类分子	DOI: 10.1186/1756-0500-2-61 链接: https://webs.iiitd.edu.in/raghava/mhcbn/ 最新更新日期: 2009年4月20日
	Allele Frequency Net Database (AFND) ^[43]	AFND数据库记录了人类白细胞抗原（human leukocyte antigen, HLA）、MHC-I相关基因及细胞因子的等位基因与单倍型频率数据，同时还提供HLA基因型与不良药物反应的关联信息	主要包含MHC-I类分子	DOI: 10.1093/nar/gkz1029 链接: https://www.allelefrequencies.net/ 最新更新日期: 2020年1月8日
	HLA Ligand Atlas ^[44]	HLA Ligand Atlas数据库收录了大量HLA配体数据，其中HLA-I涵盖51个等位基因及90 428个配体；HLA-II涵盖86个等位基因及142 625个配体	包含MHC-I类与MHC-II类分子，但未提供TCR-pMHC结合数据	DOI: 10.1136/jitc-2020-002071 链接: https://hla-ligand-atlas.org/welcome 最新更新日期: 2021年4月15日
	IPD-IMGT/HLA ^[45]	IPD-IMGT/HLA数据库是针对HLA序列的权威资源库。截至目前，该数据库已收录超过35 000个由WHO HLA系统命名委员会正式命名的等位基因序列	包含MHC-I类与MHC-II类分子，但未提供TCR-pMHC结合数据	DOI: 10.1093/nar/gkac1011 链接: https://www.ebi.ac.uk/ipd/imgt/hla/ 最新更新日期: 2024年6月27日
TCR序列与抗原特异性数据库	McPAS-TCR ^[46]	McPAS-TCR是通过人工审阅已发表文献构建的数据库，专门收录与特定抗原及病理状态相关的TCR序列	否	DOI: 10.1093/bioinformatics/btx286 链接: https://friedmanlab.weizmann.ac.il/McPAS-TCR/ 最新更新日期: 2022年9月10日
	Pan Immune Repertoire Database (PIRD) ^[47]	PIRD整合了人类及其他物种带注释的TCR与BCR测序数据，还提供数据可视化和交互式在线分析功能。此外，PIRD还专门构建了TBAdb数据库，用于存储已知抗原的TCR和B细胞受体（B-cell receptor, BCR）序列	否	DOI: 10.1093/bioinformatics/btz614 链接: https://db.cngb.org/pird/ 最新更新日期: 2020年2月1日
	TCellAtlas ^[48]	TCellAtlas是一个综合性T细胞数据库和在线分析平台，收录了1 677 799个高质量T细胞，覆盖37种疾病状态和16种组织类型，还包含68种T细胞亚型与状态分类	否	DOI: 10.1016/j.scib.2025.02.043 链接: https://guolab.wchscu.cn/TCellAtlas/#/ 最新更新日期: 2025年3月30日

续表1

数据库类型	数据库名称	资源概述与特色	是否包含MHC分子与TCR-pMHC复合物	资源链接及更新日期
	TCRdb 2.0 ^[49]	TCRdb 2.0收录了约7亿条TCR序列, 这些序列来源于19 701个TCR-Seq样本, 涵盖46种组织类型与147种临床状态, 是迄今为止最为全面的TCR序列数据库	否	DOI: 10.1093/nar/gkaf876 链接: https://guolab.wchscu.cn/TCRdb2/#/ 最新更新日期: 2025年9月9日
	VDJdb ^[50]	VDJdb数据库收录并整合了已发表的T细胞抗原特异性测定结果, 并提供抗原特异性与TCR序列之间的关联信息	否	DOI: 10.1093/nar/gkx760 链接: https://vdjdb.cdr3.net 最新更新日期: 2025年12月28日

1.1 复合物结构与界面注释数据库

复合物结构与界面注释数据库, 是开展 TCR-pMHC 复合物结构建模、界面特征解析及分子识别机制研究的基础。以 ATLAS^[36]、The structural T-cell receptor database (STCRDab)^[37]、TCR3d 2.0^[38] 和 TRAIT^[39] 为代表的资源库, 汇集了已解析的 TCR 单体和 TCR-pMHC 复合物三维结构, 并全面注释了 MHC-I 类与 MHC-II 类复合物构象及界面特征。

其中 ATLAS 数据库收录了野生型与突变型 TCR-pMHC 相互作用的实验结合亲和力数据, 还进一步提供了结构信息与结合功能注释。STCRDab 数据库每周从 Protein Data Bank (PDB)^[72] 数据库中自动收集 TCR 相关结构, 提供 TCR α/β 或 γ/δ 链配对、MHC 类型、抗原信息及 CDRs 环构象注释。TCR3d 2.0 数据库包含了经实验解析的 TCR 及其复合物结构, 为 TCR CDRs 环、抗原肽及 MHC 界面的空间关系提供系统化注释。TRAIT 数据库汇集了序列、结构及结合能力等多维信息, 构建了经系统整理的结合与非结合 TCR-抗原数据。

总体而言, ATLAS、STCRDab、TCR3d 2.0 与 TRAIT 四个数据库均以 TCR 相关结构与界面信息为核心, 在数据来源、注释深度与核心功能方面既具有共性, 也各有特色和侧重。共同之处在于, 它们均整合了经实验解析的 TCR 及 TCR-pMHC 三维结构, 并围绕 MHC 类型、抗原肽以及 CDRs 构象等关键信息进行了标准化整理, 为界面特征分析与计算建模提供了可靠数据支撑。以上数据库间的差异在于, STCRDab 与 TCR3d 2.0 更侧重结构本身的系统汇聚与精细结构注释, 其中 TCR3d 2.0 在结构空间关系与界面几何描述上具备更细致的注释功

能, 而 ATLAS 在结构注释基础上引入结合亲和力等实验功能数据, 实现结构与功能的直接关联。TRAIT 则对序列、结构与结合能力等多维度数据进行了大规模整合, 覆盖结合与非结合 TCR-抗原数据, 从而支持统计分析和机器学习。综上, 这 4 个数据库共同构成了支撑 TCR-pMHC 结构建模和识别结合研究与计算方法开发的关键数据基础。

1.2 免疫表位与MHC结合肽数据库

EPIMHC^[40]、Immune Epitope Database (IEDB)^[41] 和 MHCBN^[42] 数据库则从免疫表位及 MHC 结合肽层面提供数据支持, 支撑抗原来源解析、表位鉴定及免疫反应等研究。

EPIMHC 与 MHCBN 聚焦于 MHC-I 类结合肽相关数据, 提供肽段序列、结合亲和力与功能注释, 对表位预测与 MHC 结合规律的研究具有重要价值。需要指出的是, 两者均已长期未更新, 数据规模与注释深度难以反映当前领域最新进展。IEDB 是目前覆盖最广并持续更新的免疫表位数据库, 系统收录了感染、自身免疫、过敏及肿瘤背景下的 T 细胞表位实验数据, 支持 MHC-I 类分子与 MHC-II 类分子相关研究, 但其核心侧重于表位与免疫反应记录, 而非完整的 TCR-pMHC 相互作用关系。

综上, 这 3 类资源在适用场景上有所不同。IEDB 以其系统注释与持续更新为核心优势, 汇集了多疾病背景、跨物种的 T 细胞表位实验数据, 在数据广度与时效性方面具有显著优势, 因此更适用于依托最新实验数据开展的系统性免疫表位研究。EPIMHC 与 MHCBN 则可作为补充数据资源加以利用, 其方法学与数据内容仍具有参考价值, 在前沿研究中, 需结合最新结构与序列资源加以综合评估。

1.3 HLA分子专用数据库

Allele Frequency Net Database (AFND)^[43]、HLA Ligand Atlas^[44]和IPD-IMGT/HLA^[45]是HLA分子专用数据库,它们主要从序列、多态性及抗原呈递层面对HLA-I分子与HLA-II分子进行了全面注释。

AFND数据库从群体遗传学角度提供人类白细胞抗原(human leukocyte antigen, HLA)等位基因频率数据,为跨人群与疾病场景下的HLA相关研究提供了必要的背景信息。HLA Ligand Atlas数据库系统收集了51个HLA-I等位基因及90 428个配体,86个HLA-II等位基因及142 625个配体,但不包含TCR的结合信息。IPD-IMGT/HLA数据库收集了人类MHC等位基因的核酸与蛋白质序列信息,是开展HLA建模与等位基因特异性分析的权威资源,但同样也未直接提供TCR-pMHC结合数据。

总而言之,AFND、HLA Ligand Atlas与IPD-IMGT/HLA数据库在信息侧重点与应用场景上各具优势。AFND从群体遗传学视角系统整理等位基因频率数据,适用于跨种群比较分析、疾病易感性分析及疫苗设计。HLA Ligand Atlas提供经实验鉴定的HLA-I与HLA-II配体信息,更适用于抗原呈递分析、免疫肽组学研究以及新抗原筛选等应用场景,有助于理解不同HLA等位基因的肽呈递特征与功能差异。IPD-IMGT/HLA以高质量的等位基因序列与多态性注释为核心,尤其适用于开展HLA等位基因比对、系统进化分析以及结构建模前的序列准备工作。因此,将3类数据库联合使用,可从序列-配体-人群层面形成多尺度整合视角,为精准免疫学研究及转化应用提供数据支撑。

1.4 TCR序列与抗原特异性数据库

McPAS-TCR^[46]、Pan Immune Repertoire Database (PIRD)^[47]、TCellAtlas^[48]、TCRdb 2.0^[49]和VDJdb^[50]数据库集中关注TCR序列及其抗原或疾病关联信息。这类资源通常不直接包含完整的TCR-pMHC结构信息,但为大规模TCR序列建模、特异性预测及免疫分析提供了数据基础。

McPAS-TCR数据库基于已发表文献进行人工整理,重点收录了与特定疾病状态或抗原背景相关的TCR序列,并提供相应的病理或免疫学注释。PIRD和TCRdb 2.0数据库主要侧重于TCR序列的系统性收集与整理,在数据规模和样本覆盖范围上

具有显著优势,且分别汇总了大规模的TCR CDR3序列及跨物种的免疫受体序列资源。TCellAtlas数据库系统整合了人类T细胞不同亚型及功能状态的注释信息,并提供自动化注释工具,能够对单细胞RNA测序数据中的T细胞进行分类与功能状态识别。VDJdb数据库整合了已发表的T细胞抗原特异性测定结果,并提供TCR与抗原特异性之间的关联信息。

综上所述,McPAS-TCR、PIRD、TCellAtlas、TCRdb 2.0和VDJdb侧重于TCR方面的注释,虽不提供完整的三维结构信息,但在不同研究情境下具有明确的应用价值。McPAS-TCR以疾病或病原体背景为主线,尤其适用于研究特定疾病的免疫功能解析。PIRD与TCRdb 2.0包含大量与人群、组织及其他物种相关的TCR序列,适合用于构建数据驱动的TCR建模训练集。TCellAtlas整合的人类T细胞亚型与功能状态信息及其表达特征,为解析T细胞功能异质性、免疫微环境变化及疾病相关免疫状态提供重要参考。VDJdb提供的抗原注释,适用于开展特异性相关性分析以及初步的TCR-抗原匹配研究。总体而言,这5个数据库共同构成了TCR序列层面的重要基础资源,为后续TCR抗原识别特异性预测与精准免疫治疗研究提供关键数据支撑。

2 工具资源

现有TCR单体(表2)和TCR-pMHC复合物(表3)结构建模工具中,在模型输入、建模方法与流程等方面均有差异。TCR单体的建模主要采用同源建模或深度学习算法,且统一以氨基酸序列作为输入。相比之下,TCR-pMHC复合物同时涉及受体、抗原肽与MHC分子之间的相互作用,其建模任务更复杂,对方法设计也提出了更高的要求。根据研发目标场景,TCR-pMHC复合物建模方法主要分为通用蛋白质结构预测模型和TCR-pMHC复合物结构建模专用工具。在输入信息中,既包含仅需要氨基酸序列信息作为输入的模式,也包含依赖MHC等位基因、TCR基因名和CDR3序列注释信息作为输入的模式,后者在特异性建模中具有潜在优势,但受限于相关注释在公开数据资源中的可得性,其适用性在部分场景下可能受到限制。

Table 2 Overview of TCR structure modeling tools
表2 TCR结构建模工具资源

算法原理	工具名称	主要功能	输入信息	方法	工具使用链接及更新记录
同源建模	LYRA ^[56]	构建BCR和TCR结构	序列	基于多序列比对 (multiple sequence alignment, MSA) 筛选最优的骨架区与CDRs模板, 通过序列映射进行残基替换, 并利用侧链构象搜索算法重排原子空间位置, 以构建能量高度稳定的蛋白质结构模型	DOI: 10.1093/NAR/GKV535 使用链接: https://services.healthtech.dtu.dk/services/LYRA-1.0/ 最新更新日期: 2015年7月1日
	Repertoire Builder ^[57]	构建BCR或TCR模型, 其结构达到原子级分辨率	序列	基于MAFFT软件中的MSA扩展功能, 利用其预对齐MSA中的已知结构进行建模	DOI: 10.1039/c9me00020h 使用链接: https://sysimm.org/tools/ (网络服务目前无法使用) 最新更新日期: 2019年5月21日
	TCRmodel ^[58]	快速生成TCR结构模型	序列	基于同源模板构建初始骨架, 运用Rosetta对CDRs环进行构象采样与全原子能量精修, 最终生成最优TCR三维结构	DOI: 10.1093/nar/gky432 使用链接: https://tcrmodel.ibbr.umd.edu/index/ 最新更新日期: 2018年7月2日
深度学习	ImmuneFold TCR ^[59]	预测TCR结构	序列	基于蛋白质语言模型的结构预测方法, 并通过低秩适配 (low-rank adaptation, LoRA) 进行微调, 实现对TCR等免疫蛋白的结构预测	DOI: 10.1101/2024.11.13.621715 部署链接: https://github.com/CarbonMatrixLab/immune-fold 最新更新日期: 2024年11月15日
	TCRBuilder2 ^[60]	预测TCR的结构	序列	基于AlphaFold-Multimer结构模块改进的深度学习框架, 专门用于TCR的结构预测	DOI: 10.1038/s42003-023-04927-7 部署链接: https://github.com/oxpig/ImmuneBuilder 最新更新日期: 2023年5月29日
	tFold-TCR ^[61]	预测TCR单体、MHC-I类和MHC-II类复合物结构	序列	基于PLM提取序列特征, 通过几何等变结构模块, 模型可直接回归TCR (α/β 链) 三维原子坐标, 实现端到端的快速结构建模	DOI: 10.1101/2025.01.12.632367 使用链接: https://ai4s.tencent.com/tcr/homeView 最新更新日期: 2025年1月15日

Table 3 Overview of TCR-pMHC complex structure modeling tools
表3 TCR-pMHC复合物结构建模工具资源

研发目标	工具名称	主要功能	输入信息	方法	工具使用链接及更新记录
通用蛋白质结构预测模型	AlphaFold2 ^[62]	可用于MHC-I类和MHC-II类复合物的结构建模	序列	通过MSA、Evoformer演化建模模块和结构预测模块, 实现从序列到结构的端到端预测	DOI: 10.1038/s41586-021-03819-2 部署链接: https://github.com/deepmind/alphafold 最新更新日期: 2021年7月15日
	AlphaFold3 ^[63]	可用于MHC-I类和MHC-II类复合物的结构建模	序列	通过简化的MSA嵌入模块、Pairformer建模模块和扩散建模模块完成端到端建模	DOI: 10.1038/s41586-024-07487-w 使用链接: https://www.alphafoldserver.com 最新更新日期: 2024年5月8日
TCR-pMHC复合物结构建模专用工具	ImmuneScape ^[65]	预测MHC-I类和MHC-II类复合物结构	序列	利用LYRA进行TCR结构建模, 并参考已知复合物的结构作为模板, 完成TCR与pMHC的空间对接, 从而生成TCR-pMHC复合物结构	DOI: 10.1007/978-1-4939-9728-2_17 使用链接: https://sysimm.org/immune-scape/ (目前无法使用) 最新更新日期: 2019年8月9日

续表3

研发目标	工具名称	主要功能	输入信息	方法	工具使用链接及更新记录
		预测MHC-I类和MHC-II类复合物结构	MHC等位基因、TCR基因名和CDR3序列	通过引入多样化的混合模板来优化AlphaFold2的输入策略，并利用模板提供的几何约束，降低模型对MSA的依赖	DOI: 10.7554/eLife.82813 部署链接: https://github.com/phbradley/TCRdock 最新更新日期: 2023年1月20日
	TCRdock ^[66]	预测MHC-I类和MHC-II类复合物结构	序列	基于AlphaFold-Multimer架构，通过引入pMHC专用模板作为空间约束，并优化TCR-pMHC界面的预测算法，在保持预测速度的同时，提升建模精度	DOI: 10.1093/nar/gkad356 使用链接: https://tcrcmodel.ibbr.umd.edu/ 最新更新日期: 2023年7月5日
	TCRmodel2 ^[67]	针对MHC-I类复合物结构建模	序列	通过LYRA预测TCR骨架结构，并利用模板对齐技术确定TCR与pMHC的相对取向，然后调用MODELLER生成复合物结构	DOI: 10.1038/s41598-019-50932-4 使用链接: http://www.cbs.dtu.dk/services/TCRpMHCmodels/ 最新更新日期: 2019年10月10日
	TCR-pHLA_ModellerS ^[69]	预测MHC-I类和MHC-II类复合物结构	序列	基于均方误差 (mean squared error, MSE) 选择与模板库中最相似的序列作为模板，随后调用MODELLER完成建模，并通过ATLAS对输出结构打分	DOI: 暂未公开发表文献 部署链接: https://github.com/dwent/TCR-pHLA_ModellerS 最新更新日期: 2025年9月19日
	tFold-TCR ^[61]	预测TCR单体、MHC-I类和MHC-II类复合物结构	序列	基于PLM解析TCR和pMHC的序列特征，以挖掘特征中的空间约束关系，并驱动柔性对接模块完成复合物结构建模	DOI: 10.1101/2025.01.12.632367 使用链接: https://ai4s.tencent.com/tcr/homeView 最新更新日期: 2025年1月15日

2.1 模型输入

多数TCR单体及复合物建模工具以氨基酸序列作为输入；TCRdock需额外提供MHC等位基因、TCR V/J基因名和CDR3序列注释，以获得更强的结构约束，但部分复合物的相关注释信息获取困难，可能导致建模失败，一定程度上限制了该方法的使用场景。

2.2 建模方法

在建模方法上，TCR单体结构预测工具侧重于重建TCR αβ链的精确折叠，而TCR-pMHC复合物结构预测工具则需要在准确构建单体结构的基础上，进一步确定TCR与pMHC的空间结合模式及其复合物的整体构象。

2.2.1 TCR单体结构预测工具

TCR单体结构预测工具，包含同源建模工具LYRA^[56]、Repertoire Builder^[57]、TCRmodel^[58]，以及深度学习模型ImmuneFold TCR^[59]、

TCRBuilder2^[60]和tFold-TCR^[61]。LYRA基于多序列比对 (multiple sequence alignment, MSA) 选择最优的骨架区与CDRs模板，通过序列映射进行残基替换，最后使用侧链构象搜索算法对原子空间位置进行重排，从而构建TCR结构模型。Repertoire Builder利用MAFFT^[73]的MSA扩展功能，将预对齐序列与已知结构信息进行映射，实现TCR原子级分辨率的三维建模。TCRmodel采用片段组装策略，将目标序列映射至结构库以检索与局部构象匹配的短片段，然后通过对这些片段进行拼接与迭代替换，在骨架空间中逐步搜索并收敛至最优整体构象，并在此基础上结合Rosetta能量函数对骨架与侧链进行优化，以构建更稳定的TCR结构模型。ImmuneFold TCR采用ESMFold作为底层结构预测框架，通过低秩适配 (low-rank adaptation, LoRA) 进行参数高效微调，以提升对TCR等免疫蛋白的建模能力。TCRBuilder2基于AlphaFold-

Multimer^[74]进行改造, 形成适用于抗体和TCR的深度学习框架, 实现高精度结构预测。tFold-TCR利用PLM提取TCR序列信息, 并结合几何等变结构预测模块, 直接回归TCR $\alpha\beta$ 链的原子坐标, 从而实现对TCR单体结构的建模。

2.2.2 TCR-pMHC复合物结构预测工具

TCR-pMHC复合物结构预测工具不仅需要准确刻画单体折叠, 还需同时处理界面识别与分子对接等更复杂的问题。根据研发目标场景, 现有方法分为通用蛋白质结构预测模型和TCR-pMHC复合物结构建模专用工具。

AlphaFold2^[62]与AlphaFold3^[63]作为通用蛋白质结构建模工具, 可用于MHC-I类与MHC-II类复合物的结构建模, 但其缺乏对免疫识别界面的定向优化, 导致在刻画复合物关键区域的细节方面存在局限^[64]。为了更好地刻画TCR-pMHC复合物结构特征, 研究者进一步开发了一系列专用工具。以同源建模策略为核心的工具包含TCRmodel2^[67]、TCRpMHCmodels^[68]和TCR-pHLA_ModellerS^[69], 这些工具利用已解析的TCR-pMHC复合物结构为模板, 通过序列比对构建整体构象框架, 并在此基础上对CDRs区域进行局部构象优化, 从而实现复合物结构建模。在此基础上, ImmuneScape^[65]在利用模板结构特征确定复合物整体构象的前提下, 通过分子对接与界面精修对TCR-pMHC结合界面进行局部优化。以深度学习为核心的建模工具包含TCRdock^[66]和tFold-TCR^[61], 其中TCRdock以已解析的TCR和pMHC结构为模板, 通过预设多样化的对接几何参数生成多种候选结合模式, 并利用预测对齐误差等置信度指标对界面构象进行评估与筛选, 从而实现对TCR-pMHC结合方式的预测。tFold-TCR则在PLM生成的TCR和pMHC结构特征基础上, 通过深度学习模型预测分子之间的相互作用并进行结构优化, 这一过程不依赖外部模板, 体现了TCR-pMHC结构预测从模板约束转向特征学习的技术演进。

总体而言, TCR-pMHC建模工具在输入信息、建模流程及建模策略等方面存在差异, 在一定程度上影响了模型的输出质量及适用范围。因此, 本综述对多种代表性TCR-pMHC结构建模方法的结构、界面、PFRs构象偏差和TCR CDRs构象柔性区分能力进行了系统比较, 为相关研究的工具选择提供参考。

3 基准测试

在基准测试阶段, 对5种可同时用于MHC-I类与MHC-II类复合物建模的代表性工具——AlphaFold2、AlphaFold3、TCRmodel2、tFold-TCR和TCR-pHLA_ModellerS, 开展系统比较分析。为避免数据泄露, 本文仅纳入2024年7月6日之后公开的TCR-pMHC复合物结构(25个I类和10个II类)作为基准数据集, 该日期为被测工具中最晚的训练截止时间(TCR-pHLA_ModellerS)。

为系统评估不同工具在TCR-pMHC复合物建模中的性能, 基准测评从结构、界面特征、PFRs构象偏差、TCR CDRs构象柔性区分能力和计算资源消耗5个层面展开分析。对于结构、界面特征和PFRs构象偏差, 本文基于各工具打分最高的Top1预测结构, 采用All-Atom RMSD、Backbone RMSD从全原子与主链层面评估结构构象的预测精度。使用TM-score评估整体折叠相似性。在结合界面上, 使用DockQ^[54]评估TCR-pMHC界面质量。在PFRs构象偏差方面, 本文创新性地定义了PFRs-DI、N-PFR-DI和C-PFR-DI三个指标, 用于衡量工具是否能准确预测MHC-II类复合物结合肽两端的构象。

为评估工具对TCR CDRs构象柔性区分能力, 本文选用各工具(tFold-TCR因仅输出单一构象而未纳入)生成结构中排名前5的模型(Top5), 通过局部刚体叠合计算Top5结构中CDRs的成对均方根偏差(Top 5 root mean square deviation, Top5 RMSD), 量化TCR CDRs的构象变化范围, 并据此将构象变化范围划分为刚性(rigid)、模糊(ambiguous)和柔性(flexible)。此外, 由于测试数据中的复合物并未有明确的TCR CDRs构象柔性注释, 因此利用工具ITsFlexible^[75]获取的构象柔性注释作为参考。ITsFlexible基于大规模结构数据训练, 并采用E(n)-equivariant图神经网络对CDR环及其局部结构进行建模, 对TCR CDRs的柔性状态表现出较高的预测准确性, 可作为参考注释指标。在此基础上, 将预测模型的构象分类与ITsFlexible的柔性注释进行映射与整合, 构建统一的CDRs构象柔性数据集。最后, 基于该数据集, 比较模型预测结果与参考注释的一致性, 以定性评估各工具区分CDRs刚性与柔性状态的能力, 从而揭示其在局部结构建模中可能存在的构象偏好。

3.1 TCR-pMHC基准测试集

本文纳入了2024年7月6日之后(该日期是所有待评估工具中最晚的训练截止日期,确保与任何训练集没有重叠)公开的TCR-pMHC复合物结构。共收集35个复合物,包括25个I类TCR-pMHC复合物和10个II类TCR-pMHC复合物(表S1),并从PDB中下载原始结构作为参考。

为评估测试数据集与训练数据集之间可能存在的高度相似序列及潜在数据泄漏,对35个复合物进行氨基酸序列相似性分析(文档S1)。结果显示,MHC链在所有训练数据集中均表现出极高的序列相似性,与多数测试复合物的同型MHC分子相似度均达到100%(图S1)。这是全长序列和HLA分子亚型层面的分析结果,主要原因在于MHC分子具有相对保守的结构框架,特别是近膜端结构,而HLA基因的多态性体现在亚型间同抗原肽和TCR发生相互作用的区域^[76-78],所以在训练数据集中可能找到与测试数据高度相似甚至完全一致的序列。相比之下,TCR链的序列相似性明显低于MHC链,呈现较大的变化区间。这一差异反映了TCR分子序列的高度多样性,其中CDRs区表现尤为明显,而CDRs区的微小变化往往都可能会导致关键构象的改变,因此,TCR链可能是TCR-pMHC复合物结构预测的主要挑战来源^[7, 26]。

总体来看,测试数据与训练数据之间存在一定程度的序列相似性,但整体未出现与训练数据完全相同的复合物,表明所构建的基准测试数据具有相对独立性。

3.2 建模工具选择

本文选取了能够同时对MHC-I类与MHC-II类复合物进行建模的5种结构预测工具进行系统比较分析,包括通用蛋白质结构预测框架AlphaFold2和AlphaFold3,同源建模方法TCR-pHLA_ModellerS和TCRmodel2,以及基于蛋白质语言模型的深度学习模型tFold-TCR。其中,AlphaFold2、AlphaFold3、TCR-pHLA_ModellerS与tFold-TCR在本地部署运行,TCRmodel2通过其网络服务器使用。尽管TCRdock支持MHC-I类与MHC-II类复合物的建模,但其建模过程依赖于MHC等位基因、TCR V/J基因名和CDR3序列信息作为输入,部分复合物在现有数据库中难以获得可靠注释,即便在基因信息完整的情况下,仍有部分复合物未能成功建模。基于上述限制,本文未将TCRdock纳

入后续基准评测。

3.3 结构与界面的预测性能评估

为了系统评估各建模工具在TCR-pMHC复合物上的整体结构质量,本文使用PyMOL (Version 3.0)^[79]计算All-Atom RMSD和Backbone RMSD。其中,All-Atom RMSD用于从全原子层面量化预测结构与参考结构之间的整体几何偏差,Backbone RMSD基于主链原子坐标,用于衡量骨架构象的一致性。使用OpenStructure^[80]计算TM-score,从整体折叠层面评估预测结构与参考结构的全局相似性。此外,采用DockQ评估TCR与pMHC之间的界面构象是否被准确模拟。

3.4 MHC-II PFRs构象预测准确性评估

鉴于现有的RMSD、TM-score和DockQ等常用指标主要关注全局结构相似性和结合界面建模质量,难以准确衡量MHC-II类复合物中结合肽PFRs的结构质量,本文创新性地提出PFRs-DI、N-PFR-DI、C-PFR-DI三项指标,用以评估预测模型是否能正确重建肽段两端的自然构象。此外,由于PFRs相关指标的计算依赖于核心肽及其N和C端PFR边界的准确定义,所以从10个TCR-pMHC II类复合物中进一步筛选出能够从文献中明确获得核心肽及其N/C端区域信息的5个复合物作为基准数据集(表S2)。

3.4.1 PFRs偏差指数(PFRs-Deviation Index, PFRs-DI)

PFRs-DI用于评估预测模型对MHC-II复合物中抗原肽PFRs的构象准确性。该指标关注模型是否能够在空间上正确定位肽段两端残基,并同时保持其在天然结构中所呈现的外翻取向、主链折叠特征以及与MHC肽结合槽之间的几何关系^[81-83]。

在计算中,选取N端与C端PFR区域中同时存在参考结构与预测结构的残基参与计算。考虑到主链原子能够稳定反映多肽骨架的整体空间构象,而侧链原子通常具有较高的构象自由度且在不同预测模型之间存在较大不确定性^[84],本文仅比较参与计算的残基的主链原子的空间坐标,并以RMSD的平均值作为整体PFRs构象偏差。较低的PFRs-DI表示模型能够准确定位肽段的两端并保持其自然延伸趋势,较高偏差则提示模型未能正确捕捉PFRs的真实空间布局,从而可能影响TCR-pMHC的整体构象。计算公式如下:

$$PFRs - DI = \frac{1}{N_{PFR}} \sum_{i \in \{N - PFR, C - PFR\}} \sqrt{\frac{1}{K} \sum_a^K |x_a^{mov}(i) - x_a^{ref}(i)|^2} \quad (1)$$

其中: N 表示 PFR 残基数量; i 表示 N 端和 C 端的 PFRs 残基; $K=4$ (使用 backbone 原子: N、C α 、C、O); $x_a^{mov}(i)$ 表示预测结构中残基 i 的主链原子 a 的三维坐标; $x_a^{ref}(i)$ 表示参考结构中残基 i 的主链原子 a 的三维坐标。

3.4.2 N端PFR偏差指数 (N-PFR-Deviation Index, N-PFR-DI)

N-PFR 位于肽段的起始端, 其空间取向直接影响肽段的外向延伸方向以及在 MHC 肽结合槽入口处的构象稳定性, 是 pMHC 复合物与 TCR 形成初始接触界面的关键结构区域^[21]。N-PFR-DI 直接反映模型在这一关键几何区域的复现程度, N-PFR-DI 越低表示预测模型越能正确恢复 N 端真实构象。

在计算过程中, 对 N-PFR 中同时存在于参考结构与预测结构的残基, 逐一比较其主链原子的空间坐标, 并以 RMSD 的平均值作为 N-PFR 构象偏差。较低的 N-PFR-DI 值表示预测模型能够较好地恢复肽段 N 端在天然结构中的空间取向与构象特征。较高的 N-PFR-DI 值则提示模型在肽段起始处可能出现构象扭曲, 或其与 MHC 槽口入口区域的几何匹配关系与真实结构不一致。计算公式如下:

$$N - PFR - DI =$$

$$\frac{1}{N_{N - PFR}} \sum_{i \in N - PFR} \sqrt{\frac{1}{K} \sum_a^K |x_a^{mov}(i) - x_a^{ref}(i)|^2} \quad (2)$$

其中: N 表示 PFR 残基数量; i 表示 N 端残基; $K=4$ (使用 backbone 原子: N、C α 、C、O); $x_a^{mov}(i)$ 表示预测结构中残基 i 的主链原子 a 的三维坐标; $x_a^{ref}(i)$ 表示参考结构中残基 i 的主链原子 a 的三维坐标。

3.4.3 C端PFR偏差指数 (C-PFR-Deviation Index, C-PFR-DI)

C-PFR 是 MHC-II 肽段的向外延伸区域, 在维持肽段稳定性、形成 pMHC 外侧几何形态中具有重要作用^[82, 83]。C-PFR-DI 用于评估模型在此区域是否准确重现真实构象。

在计算过程中, 对 C-PFR 区域中同时存在参考结构与预测结构的残基, 逐一比较其主链原子的空间坐标, 并以 RMSD 的平均值作为 C-PFR 构象偏差。较低的偏差表示模型成功捕捉 C 端的自然弯曲和扭转方向, 较高的偏差则说明预测方法未能正确捕捉 C 端扭转角或其与 MHC-II 结构的关系。计算公式如下:

$$C - PFR - DI =$$

$$\frac{1}{N_{C - PFR}} \sum_{i \in C - PFR} \sqrt{\frac{1}{K} \sum_a^K |x_a^{mov}(i) - x_a^{ref}(i)|^2} \quad (3)$$

其中: N 表示 PFR 残基数量; i 表示 C 端残基; $K=4$ (使用 backbone 原子: N、C α 、C、O); $x_a^{mov}(i)$ 表示预测结构中残基 i 的主链原子 a 的三维坐标; $x_a^{ref}(i)$ 表示参考结构中残基 i 的主链原子 a 的三维坐标。

3.5 TCR CDRs 构象柔性区分能力评估

为定性评估不同结构预测工具在捕捉 TCR CDRs 刚性与柔性构象方面的能力, 本研究对各工具输出的 Top5 候选模型进行 CDRs 构象分析。

在保证预测结构与参考结构具有可比序列编号的前提下, 预先构建链映射表, 对所有预测模型的链号进行统一规范, 使其与参考结构序列实现一一对应。随后依据 The International ImMunoGeneTics Information System (IMGT)^[85] 数据库中查询的 CDR 编号区间, 从各工具的 Top5 结构中提取对应 CDRs 的 C α 原子坐标^[86] (表 S3)。

通过 Kabsch 算法^[87-88] 对 Top5 中任意两条预测结构的对应 CDR 进行局部最优刚体叠合, 并计算其成对 RMSD 值。对于每个 CDR, 计算 Top5 结构之间产生的 10 组成对 RMSD (Top5 RMSD) 均值, 并以其均值作为该 CDR 构象柔性的参数。

Top5 RMSD 参数被划分为 3 类构象柔性等级: 平均 RMSD 小于 1.0 Å 的 CDR 被视为强烈构象收敛, 其 Top5 模型高度一致 (rigid); 平均 RMSD 位于 1~2.0 Å 的 CDR 被认为存在中等构象收敛 (ambiguous); 平均 RMSD 大于或等于 2.0 Å 的 CDR 则被视为构象发散 (flexible), 体现为较大的构象柔性^[89]。

上述分类结果进一步与 ITsFlexible^[75] 预测的 CDRs 构象柔性注释进行对比分析 (表 S4)。ITsFlexible 注释中出现的高置信度刚性和低置信度刚性映射为 rigid, 高置信度柔性和低置信度柔性映射为 flexible, 模糊标签映射为 ambiguous。由于 ambiguous 样本不具有明确生物学意义, 因此在后续分析中被剔除。

最后, 基于上述结果进行统计分析。以 ITsFlexible 作为参考注释, 将刚性 CDR 定义为阳性类别, 柔性 CDR 定义为阴性类别。对于各结构预测工具在 MHC-I 类与 MHC-II 类复合物中的所有

CDRs, 统计其预测结果与参考注释之间对应的真阳性 (true positive)、假阳性 (false positive)、假阴性 (false negative) 和真阴性 (true negative)。在此基础上, 定义 CCC 定性评估不同工具区分 CDRs 刚性与柔性的能力 (表 S5)。CCC 的定义形式在数学表达上类似于准确性指标, 但其生物学含义并非衡量预测结果相对于实验真值的正确率, 而是用于量化模型预测的 CDRs 构象状态与 ITsFlexible 注释的构象柔性之间的一致程度。TP、FP、FN、TN 具体定义如下。

True positive (TP): ITsFlexible 注释为刚性, 模型预测为刚性;

False positive (FP): ITsFlexible 注释为柔性, 模型预测为刚性;

False negative (FN): ITsFlexible 注释为刚性, 模型预测为柔性;

True negative (TN): ITsFlexible 注释为柔性, 模型预测为柔性。

CCC 计算公式如下:

$$CCC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

4 评估结果

在该模块中, 本文基于 All-Atom RMSD、Backbone RMSD、TM-score、DockQ、PFRs-DI、N-PFR-DI、C-PFR-DI 和 CCC 8 项评估指标, 系统比较了 5 个建模工具在 MHC-I 类与 MHC-II 类复合物中的建模性能, 重点从整体结构精度、界面特征、PFRs 构象刻画能力, 以及 TCR CDRs 环构象柔性的区分能力等维度进行评估。此外, AlphaFold2 在部分样本中建模失败, 包括 MHC-I 类复合物 8RYM、8RYN、9GV6、9IKY、9J4T、9RU5 和 9RUP, 以及 MHC-II 类复合物 8TRR 和 8VD2。对于这些预测失败的复合物, 不纳入后续指标的统计。

4.1 结构与界面

4.1.1 全原子结构准确性

全原子结构准确性层面, All-Atom RMSD 分析结果显示 (图 1a~c), MHC-I 类 TCR-pMHC 复合物中 (图 1a), AlphaFold2 和 AlphaFold3 的中位 All-Atom RMSD 均约为 2.0 Å, 表明 AlphaFold2 和 AlphaFold3 在多数样本中能够保持较低 RMSD, 整体预测精度较高。tFold-TCR 的中位 RMSD 次之 (约 2.5 Å), 且箱体最窄, 说明 tFold-TCR 整体表现

较稳定。TCR-pHLA_ModellerS 的中位 RMSD 虽居于中下游 (约 4.0 Å), 但其四分位距较窄, 且极值范围最小 (大致集中在 3.0~5.0 Å), 表明其预测稳定性高。相比之下, TCRmodel2 的中位 All-Atom RMSD (约 6.0 Å) 为所有工具中最高, 箱体范围较大, 说明其建模准确性与稳定性相对较弱。

MHC-II 类复合物 (图 1b) 中, AlphaFold2、AlphaFold3 和 tFold-TCR 的中位 All-Atom RMSD 最低 (均约为 2.5 Å), 三者箱体都较窄, 多数预测在 2.0~4.0 Å, 但由于 AlphaFold2 存在少量建模失败的复合物, 提示其稳定性不如 AlphaFold3 和 tFold-TCR。其次是 TCR-pHLA_ModellerS (中位 RMSD 均约为 3.3 Å), 上四分位数较高, 最大值接近 23.0 Å, 表明其整体误差及样本间波动较大。TCRmodel2 的中位 RMSD 最高 (约 8.5 Å), 且箱体最大, 数据分布最分散, 提示其在全原子层面的整体构象偏差较大, 整体建模准确性较低。

MHC-I 类复合物和 MHC-II 类复合物的全样本 (Overall) 分析结果 (图 1c) 显示, AlphaFold2、AlphaFold3 和 tFold-TCR 显示出最低的中位 All-Atom RMSD (均约为 2.0 Å), 表明其能够较为精确地构建复合物的全原子结构。其中, tFold-TCR 四分位距较窄, 整体表现比 AlphaFold2 和 AlphaFold3 更稳定。TCR-pHLA_ModellerS 的中位 RMSD 约为 4.0 Å, 箱体较窄, 预测结果集中, 提示该工具稳定性较好, 但全原子预测结果准确性有待提高。相比之下, TCRmodel2 的全原子 RMSD 中位数最高 (约 6.5 Å), 数据分布离散, 说明其预测结果不稳定。

总体来看, 在所有成功建模的样本中, AlphaFold2 和 AlphaFold3 在全原子结构建模中表现最佳。相比之下, TCRmodel2 的中位 RMSD 最高且分布最离散, 反映出其预测精度和稳定性存在明显不足。

4.1.2 主链构象一致性

在主链构象一致性层面上, Backbone RMSD 分析结果显示 (图 1d~f), MHC-I 类复合物中 (图 1d), AlphaFold2、AlphaFold3 和 tFold-TCR 的中位值都较低 (均约为 1.5 Å)。其中 AlphaFold3 在大部分样本预测误差较小, 但存在极端离群值, 表明该工具整体建模精度较高, 偶尔出现严重错误。tFold-TCR 箱体最小, 说明其稳定性较高。TCR-pHLA_ModellerS 中位数较高 (约 3.5 Å), 但其四分位距非常窄, 分布也较为集中, 说明尽管建模精

度不如 AlphaFold 系列和 tFold-TCR, 但模型输出较为稳定、波动较小。相比之下, TCRmodel2 的中位数最高 (约 6.5 Å), 且四分位距较大, 表明其主链构象预测中不仅偏差大, 而且整体准确性较低, 稳定性不足。

MHC-II 类复合物 (图 1e) 中, tFold-TCR 的 Backbone RMSD 中位数最低 (约 2 Å), 且箱体最小, 大部分样本分布集中, 说明其预测准确性较高, 稳定性最好。其次是 AlphaFold2、AlphaFold3 和 TCR-pHLA_ModellerS, 三者中位数均约为 3.0 Å。其中 TCR-pHLA_ModellerS 箱体较大, 最大值接近 23.0 Å, 提示其对部分目标的建模构象差异较大。而 TCRmodel2 的中位数 (约 9.5 Å) 和四分位距均为所有模型中最高, 显示其主链误差显著且最不稳定。

MHC-I 类与 MHC-II 类复合物的全样本分析结果显示 (图 1f), AlphaFold2、AlphaFold3 和 tFold-TCR 的中位数相对较低 (均约为 2.0 Å)。三者中 tFold-TCR 箱体最小, 说明其整体误差较低且稳定性较好。TCR-pHLA_ModellerS (约 3.5 Å) 处于中等水平, 但四分位距较小, 表明其预测性能稳定。相较之下, TCRmodel2 表现出最高的中位值 (约 7.0 Å), 且数据分布最离散, 提示其对主链构象预测精度和稳定性均不足。

综上所述, 在所有测试的模型中, TCRmodel2 的主链构象预测表现最差 (中位值最高, 不稳定性最大), 而 AlphaFold3 和 tFold-TCR 在主链构象建模中具有更高的准确性。

4.1.3 整体构象相似性

在整体构象相似性层面上, TM-score 分析结果 (图 1g~i) 显示, MHC-I 类复合物中, AlphaFold2、AlphaFold3、TCR-pHLA_ModellerS 与 tFold-TCR 在中位数均约为 0.50 (图 1g), 提示其在部分样本中能达到正确折叠拓扑 (TM-score=0.5)^[90]。而 AlphaFold3 上四分位接近 0.90, 表明其在部分复合物上的整体结构拓扑方面具有更好的性能。TCRmodel2 (均约为 0.26) 的中位值偏低, 表明预测性能明显不足。

MHC-II 类复合物中, AlphaFold3 的 TM-score 中位数为 0.95~0.97, 分布高度集中, 显著优于其他工具 (图 1h), 表明能稳定构建整体拓扑结构。AlphaFold2、TCR-pHLA_ModellerS 与 tFold-TCR 处于中等水平 (中位数约 0.60), 提示其整体折叠大体正确但仍存在一定程度的偏差, 而

TCRmodel2 的中位数 (约 0.30) 最低, 提示其预测结果未达到正确结构折叠水平。

在 MHC-I 类与 MHC-II 类复合物的全样本分析结果 (图 1i) 中, AlphaFold3 (中位数约 0.85) 继续保持最高的 TM-score, 且分布集中, 显示其在两类复合物中的全局拓扑一致性最优。AlphaFold2、TCR-pHLA_ModellerS 和 tFold-TCR 中位数约为 0.60, 表明预测的整体几何结构能达到正确折叠阈值。TCRmodel2 的 TM-score (中位数约 0.30) 整体偏低且离散度较大, 表明其多数预测未能获得正确的全局拓扑结构, 整体建模稳定性与可靠性较低。

总体而言, AlphaFold3 在 MHC-I 类与 MHC-II 类复合物的整体构象相似性评估中表现出最为稳定且最高的整体构象建模准确性, 而 TCRmodel2 的整体构象建模能力相对较弱。

4.1.4 界面构象质量

界面构象质量层面, DockQ 分析结果 (图 1j~l) 显示, MHC-I 类 TCR-pMHC 复合物中, AlphaFold3 的 DockQ (约 0.75) 中位数最高且分布较集中 (图 1j), 表明其能够稳定生成高质量的界面构象。AlphaFold2 和 TCRmodel2 的中位数分别约为 0.65 和 0.62, 具有较高的预测上限, 但结果分布较为分散, 提示两个工具的稳定性不足。TCR-pHLA_ModellerS 的中位 DockQ 约为 0.53, 但其箱体最窄, 说明其预测结果较稳定。相比之下, tFold-TCR (约 0.47) 的整体 DockQ 偏低, 显示其在 MHC-I 界面建模中的准确性受限。

MHC-II 类复合物中, AlphaFold3 (约 0.82) 仍表现出最优的界面预测质量, 其 DockQ 中位数处于最高区间且分布集中, 下限明显高于其他工具。AlphaFold2 (约 0.61)、TCR-pHLA_ModellerS (约 0.58) 与 TCRmodel2 (约 0.59) 的 DockQ 中位数相近, 整体表现居中 (图 1k)。tFold-TCR (约 0.38) 的 DockQ 中位数最低且箱体较宽, 表明其在 MHC-II 类复合物的界面预测性能落后。

MHC-I 类复合物与 MHC-II 类复合物的全样本分析结果显示, AlphaFold3 的 DockQ 中位数 (约 0.77) 显著高于其他方法, 且分布集中 (图 1l), 表明其在两类复合物的界面构象中具有更高的准确性和稳定性。AlphaFold2 (约 0.62) 和 TCRmodel2 (约 0.61) 表现中等。TCR-pHLA_ModellerS 的中位 DockQ 约 0.54, 提示其能够在一定程度上模拟界面几何特征但存在偏差。相比之下, tFold-TCR 的

DockQ中位数约0.38, 为所有模型中最低, 说明其界面构象预测较差。

总体而言, AlphaFold3在MHC-I类与MHC-II

类复合物界面构象预测方面表现出明显优势, 而tFold-TCR在界面建模质量和稳定性方面略显不足。

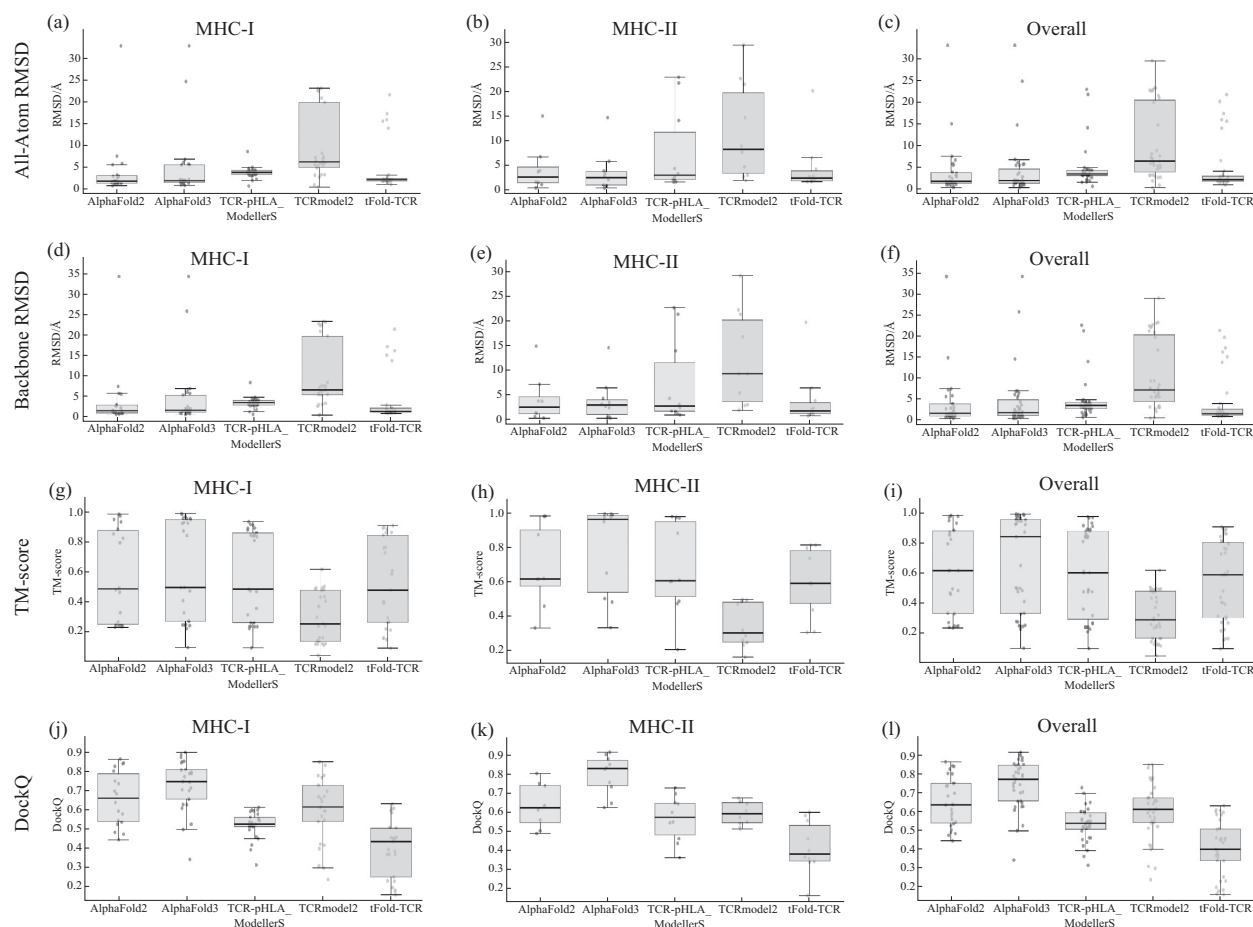


Fig. 1 Comparison of modeling quality at the structural and interfacial levels

图1 结构与界面的建模质量比较

本图分别比较了5个工具(AlphaFold2、AlphaFold3、TCR-pHLA_ModellerS、TCRmodel2和tFold-TCR)在四项评估指标上的表现。箱线图中, 箱体表示四分位距(interquartile range, IQR), 水平线表示中位数, 上下限表示极值范围, 离散点代表单个复合物的得分。AlphaFold2在MHC-I类、MHC-II类及全样本(Overall)的有效样本量分别为18、8和26, 其余工具分别为25、10和35。(a~c)全原子结构准确性(All-Atom RMSD): (a) MHC-I类复合物; (b) MHC-II类复合物; (c) 全样本。(d~f)主链构象一致性(Backbone RMSD): (d) MHC-I类复合物; (e) MHC-II类复合物; (f) 全样本。(g~i)整体构象相似性(TM-score): (g) MHC-I类复合物; (h) MHC-II类复合物; (i) 全样本。(j~l)界面构象质量(DockQ): (j) MHC-I类复合物; (k) MHC-II类复合物; (l) 全样本。

4.2 MHC-II PFRs构象偏差

对于N端PFR的构象模拟(图2a), AlphaFold2的N-PFR-DI分布最紧凑且偏左, 显示其在N端局部构象建模中的偏差最小。AlphaFold3和TCR-pHLA_ModellerS略优于AlphaFold2, 波动较大但精度尚可。TCRmodel2与tFold-TCR模型均呈现更右移的分布, 提示其预测的N端构象相对真实结构存在较大偏差。

C端PFR-DI的分布(图2b)显示,

TCRmodel2的整体偏差最小, 其结果集中分布在最左侧, 说明其C-PFR的构象模拟最接近真实结构。TCR-pHLA_ModellerS和AlphaFold3的表现略次于TCRmodel2, 分布整体右移且呈现更长的高值尾部, 提示其部分样本中仍存在局部偏差。相比之下, tFold-TCR的C-PFR-DI分布最靠右, 且呈现整体明显放大, 表明该工具在C端PFR构象预测中偏差最为突出。

两端PFRs-DI分布(图2c)显示, AlphaFold2

的表现最优, 其分布最靠左, 肽段两端的构象偏移最小。AlphaFold3与TCR-pHLA_ModellerS在整体偏差上居于中间区间, 处于中等水平; 而tFold-TCR的分布向右偏移更为明显, 反映其在长肽两端构象的整体预测偏差均显著高于前述工具, 构象模拟的准确度相对较低。

对于综合样本完成率(overall_coverage)与构象偏差两个层面, 由图2d可见, 不同工具在

MHC-II数据集上呈现出显著的性能差异。AlphaFold2在N-PFR-DI及整体PFRs-DI维度上均取得最佳表现, 但其部分长肽建模失败, 完成率略低, 表明其在成功预测时能够恢复高精度的两端构象, 但对某些结构呈现敏感性。相比之下, TCRmodel2在C端PFR预测表现突出, 因此其在MHC-II长肽中的适用性与稳健性更均衡。

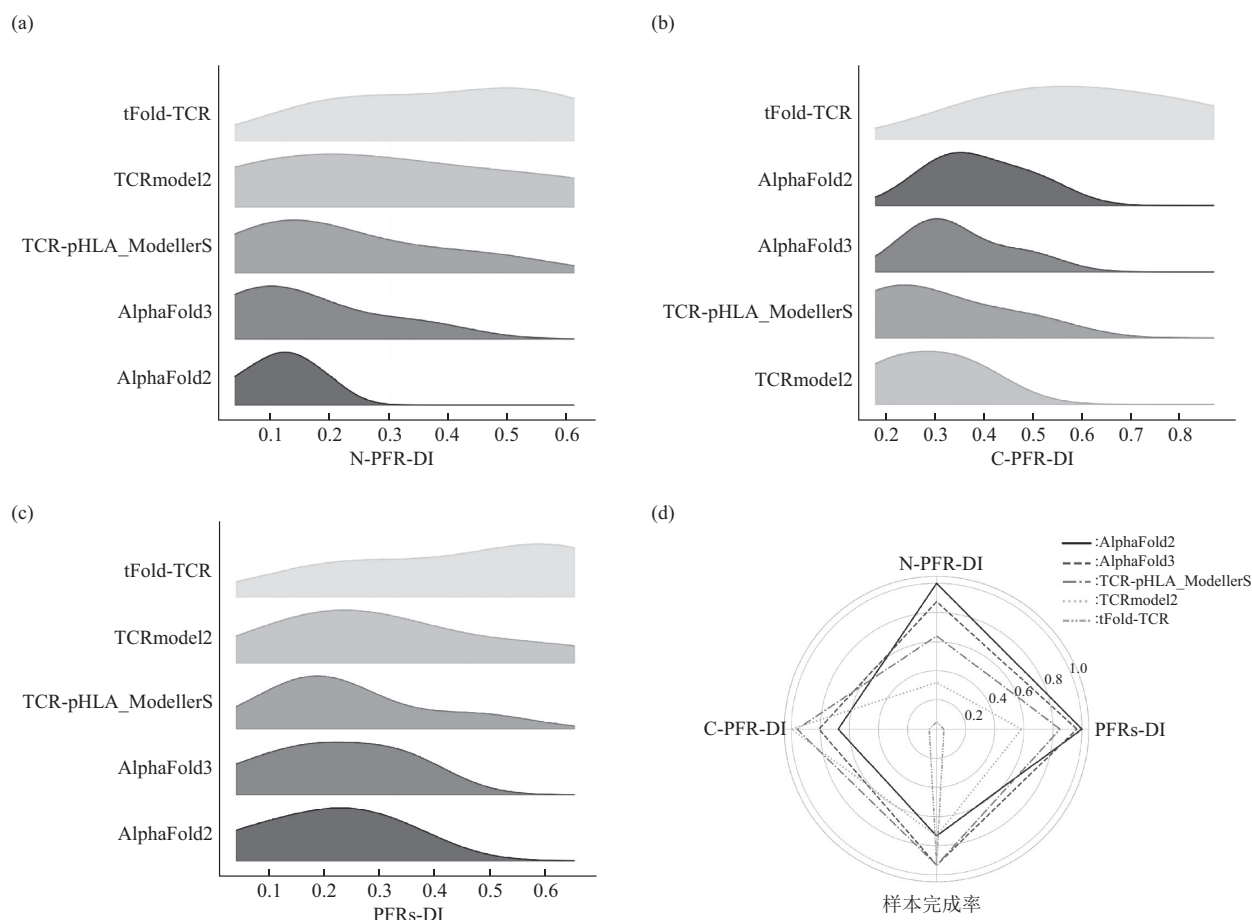


Fig. 2 Accuracy of predicted PFR conformations

图2 PFR构象准确性

本图分别比较了5个工具(AlphaFold2、AlphaFold3、TCR-pHLA_ModellerS、TCRmodel2和tFold-TCR)在MHC-II类复合物上对PFRs构象的预测表现。AlphaFold2的有效样本量为3, 其余工具为5。(a) N端PFR构象准确性; (b) C端PFR构象准确性; (c) 两端PFRs构象准确性; (d) 雷达图展示了5种建模工具在不同指标上的综合表现。每个顶点代表一个评价指标, 径向刻度表示经Min-Max归一化后的得分(数值越接近1表示性能越优, 越接近0表示性能越差)。tFold-TCR在3个指标上的归一化结果为0, 在雷达图中造成与坐标中心重合。为了增强可视化效果, 设置下限为0.05。

4.3 TCR CDRs构象柔性区分能力

在MHC-I类复合物(图3a)中, AlphaFold2与AlphaFold3在多数CDRs区表现出较高的CCC。对于TCR α 链的3个CDR区(CDR1 α 、CDR2 α 、CDR3 α), 两者整体一致性均处于较高水平, 尤其

在CDR3 α 上, 其一致性可达到0.9~1.0, 表明其在处理CDR3 α 环时能较好复现真实的构象柔性。对于TCR β 链的3个CDR区(CDR1 β 、CDR2 β 、CDR3 β), AlphaFold2和AlphaFold3整体仍保持较高CCC, 其中CDR3 β 分别为1.0和0.94, 但两者在

CDR1 β 上明显下降,提示其在部分 β 链环区的构象柔性区分相对较弱。相比之下,TCR-pHLA_ModellerS和TCRmodel2在MHC-I类复合物中的整体表现明显差于AlphaFold系列模型。TCR-pHLA_ModellerS在TCR α 链上表现中等,其中CDR1 α 约为0.6,但CDR2 α 与CDR3 α 下降至0.3左右,在TCR β 链中,其CDR1 β 约0.64,CDR2 β 约0.36,而CDR3 β 几乎接近0。TCRmodel2的表现也呈现不均衡模式,在TCR α 链中CDR1 α 约0.58,但CDR2 α 与CDR3 α 均低于0.35,在TCR β 链中CDR1 β 与CDR3 β 较高(约0.75),而CDR2 β 显著偏低(约0.22)。总体而言,在MHC-I类复合物中,AlphaFold2和AlphaFold3在多数CDRs环均展现出更高的构象区分能力。

在MHC-II类复合物(图3b)中,对于TCR α 链,AlphaFold2与AlphaFold3依然保持较高的CCC,其中CDR1 α 、CDR2 α 与CDR3 α 均在0.75以上,表明两种通用建模模型在MHC-II类复合物中仍能较稳定地预测 α 链CDR构象。然而,在TCR β 链中,其柔性构象区分能力明显下降。AlphaFold2在CDR1 β 、CDR2 β 、CDR3 β 上分别约0、0.23和0.48;AlphaFold3也呈现类似趋势,其中CDR2 β 仅约0.12,表明在MHC-II类复合物中,两种建模工具对TCR β 链CDRs环的构象柔性区分能力明显减弱。而TCR-pHLA_ModellerS在MHC-II类复合物中的表现则呈现与MHC-I类复合物相反的模式,尽管其在TCR α 链(CDR1 α 、CDR2 α 、CDR3 α)上的CCC仍较低,但该工具在CDR1 β 、CDR2 β 、CDR3 β 上均取得最高CCC,其中CDR3 β 的一致性达到1.0。这一结果反映TCR-pHLA_ModellerS在MHC-II类复合物背景下对 β 链CDRs环的构象建模更加稳定保真。TCRmodel2在MHC-II类复合物中的表现呈现中等水平,在TCR α 链CDR(CDR1 α 、CDR2 α)上CCC明显偏低(约0.2~0.3),在TCR β 链中,CDR1 β 与CDR2 β 表现出中等水平(约0.56~0.74),但在CDR3 β 上CCC下降至0.33。

综合而言,不同工具在CDRs构象柔性区分方面呈现出明显的类别依赖性与链特异性差异。AlphaFold2与AlphaFold3在MHC-I类复合物中整体表现最佳,能较稳定地区分多数CDRs环的柔性构象。但在MHC-II类复合物中,其对TCR β 链CDR的辨识能力明显减弱。相反,TCR-

pHLA_ModellerS在MHC-I类复合物中表现一般,却在MHC-II类复合物中对3条CDR β 的构象分类展现出高度一致性,提示其在 β 链上具有优势。总体来看,不同工具在MHC-I类与MHC-II类复合物中的适用性存在差异和性能互补,强调在跨体系评价与应用建模工具时需进行差异化选择。

4.4 计算资源与运行耗时

AlphaFold2、AlphaFold3、tFold-TCR和TCR-pHLA_ModellerS在本地环境下运行,仅以FASTA格式的序列作为输入。AlphaFold2和AlphaFold3分别为每个目标复合物生成10个候选结构;TCR-pHLA_ModellerS为待预测复合物生成5个结构,并同步输出ATLAS评分结果;tFold-TCR每次仅能生成1个结构。TCRmodel2通过其Web服务器访问,每个输入生成5个模型。

为评估各工具在最低硬件条件下的适用性,预测过程中均设置为模型要求的最低配置。AlphaFold系列模型在单GPU条件下运行,单个复合物的预测通常需要数小时。tFold-TCR在单GPU条件下约需15 min完成一次预测。尽管该工具提供了在线服务器,但在实际测试中,其网页端任务常出现运行数小时仍未返回结果的情况,因而难以满足高通量预测需求。TCR-pHLA_ModellerS运行无需GPU支持,单个结构预测时间5~6 min,因而更适合于大规模建模任务。TCRmodel2平均每个结构建模时间约30 min。TCRdock通过Google Colab运行,需输入MHC等位基因、TCR V/J基因名称和CDR3序列,每次运行得到1个结构。

此外,在对35个TCR-pMHC复合物的基准测试中,部分工具未能成功完成测试数据集中全部复合物的结构建模。其中,AlphaFold2预测I类复合物8RYM、8RYN、9GV6、9IKY、9J4T、9RU5和9RUP以及II类复合物8TRR和8VD2均失败,推测原因可能是目标序列的同源序列数量有限或序列相似性较低,多序列比对所提供的共进化特征不足,导致无法进入后续结构预测模块。TCRdock仅能对I类复合物中的8RJ5、8V4Z和8V51完成建模,其余复合物均由于其基因命名在建模过程中无法识别特定的MHC或V/J基因名,导致建模失败。相比之下,AlphaFold3、TCRmodel2、tFold-TCR和TCR-pHLA_ModellerS成功生成了基准测试中所有结构的预测,展现出更优越的鲁棒性。

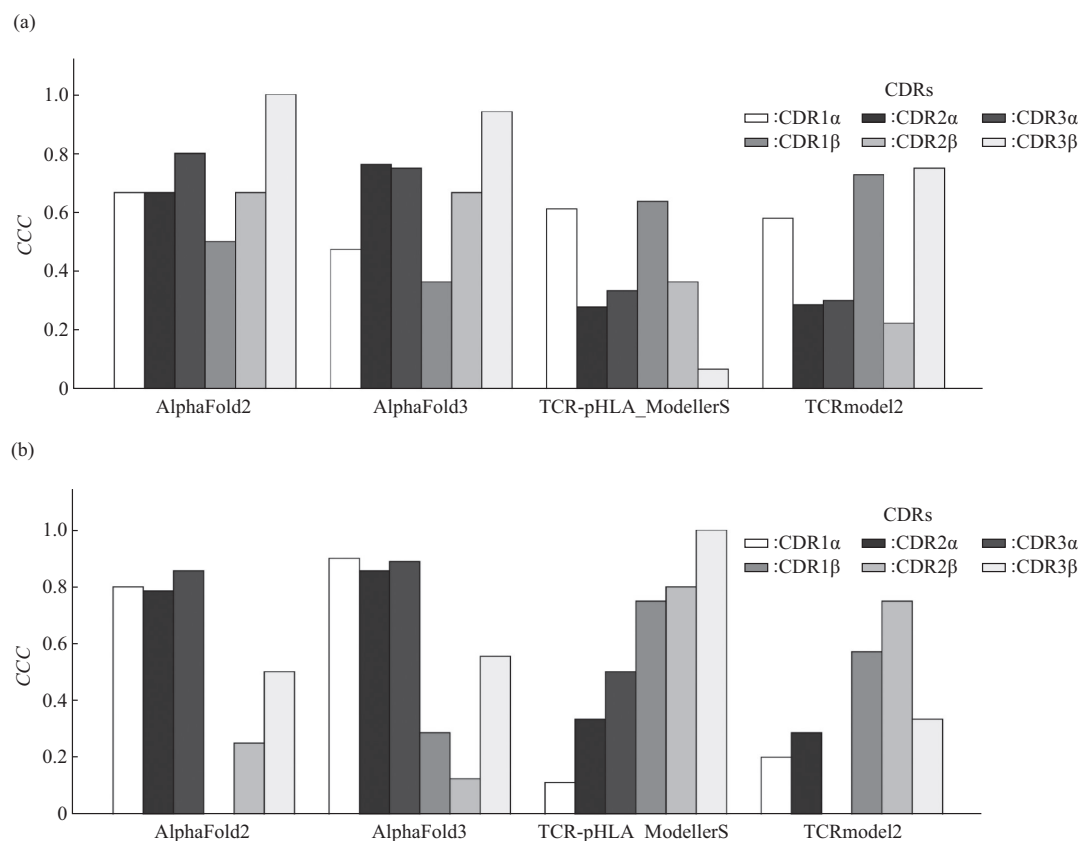


Fig. 3 Ability to distinguish rigid and flexible conformations of TCR CDRs

图3 TCR CDRs刚性及柔性构象区分能力

本图分别比较了4类工具(AlphaFold2、AlphaFold3、TCR-pHLA_ModellerS和TCRmodel2)在区分TCR CDRs构象柔性方面的表现,柱状图中的每条柱体分别代表6个CDRs环。AlphaFold2的有效样本量在MHC-I类复合物与MHC-II类复合物分别为18和8,其余工具分别为25和10。(a) MHC-I类复合物;(b) MHC-II类复合物。

5 总结与讨论

本文系统整合了TCR和pMHC相关数据资源与计算工具,并对主流工具——AlphaFold2、AlphaFold3、TCRmodel2、tFold-TCR和TCR-pHLA_ModellerS进行了全面基准测试。与传统通用结构和界面评估指标相比,本文的评测框架严格区分MHC-I类和MHC-II类复合物结构特征,并对MHC-II类复合物引入PFR构象偏差指数,精细地比较了不同工具对PFRs构象刻画能力。除通用指标和PFR构象偏差指数外,还提出CDR构象一致性指数,用于定性比较不同工具对TCR CDRs构象柔性的刻画区分能力,建立了全面且多维度的TCR-pMHC复合物结构建模工具性能评价框架。该框架可在缺乏实验复合物结构的情况下,用于快速选择合适工具获得可靠的TCR-pMHC复合物构象与界面信息,从而辅助解析T细胞免疫识别机

制、阐明TCR对特定表位pMHC的结合特异性,进一步支撑候选TCR的鉴定与优化,服务于肿瘤免疫治疗、感染性疾病疫苗研发及个体化新抗原治疗。

经过系统分析,本文发现,在MHC-I类TCR-pMHC复合物建模中,AlphaFold3在全原子和主链构象方面表现最优,同时在界面构象和全局拓扑上也表现出明显优势。在MHC-II类复合物建模中,AlphaFold3在整体界面构象和全局结构方面仍然表现最佳,但其在肽段两端构象的细节预测中仍存在偏差;AlphaFold2在PFRs构象预测上表现较好,但在建模完成率上稍显不足。在区分TCR CDRs构象柔性表现方面,尽管AlphaFold系列模型和TCR-pHLA_ModellerS分别在I类和II类复合物中表现出一定优势,但整体效果仍显不足。

从整体结果来看,现有工具对I类复合物的预测明显优于II类复合物,II类复合物的建模仍然存

在巨大挑战。其原因在于,大部分工具在建模过程中可能主要以复合物整体结构的合理性为优化目标,而缺少对肽两端PFRs构象与局部微环境的柔性特征捕获和结构优化,这不仅影响肽在结合槽中的结合方式与锚定位点选择,还会通过改变肽主链构象和侧链取向,进而调控与TCR的接触模式和识别特异性^[22]。因此,直接限制了对MHC-II类复合物的准确建模。同时,多数工具在关键接触残基以及构象柔性等层面的区分能力也存在一定局限。在TCR-pMHC预测结构推断关键结合位点的实际研究中,若模型对柔性构象进行过度刚化,可能导致对TCR识别关键位点与结合构象的误判,从而影响疫苗设计或TCR免疫治疗靶点筛选的准确性与可靠性。

依据分析结果,本文进一步从应用需求出发,总结了面向不同研究目标的TCR-pMHC复合物建模工具选择建议。由于不同方法的优势侧重不一,不主张以单一工具作为通用最优解,而是建议根据具体目的采用性能互补的组合使用策略。在需要获得高质量整体拓扑与界面几何的场景中,可优先采用通用蛋白质结构预测框架AlphaFold3。若更关注于全原子或主链构象准确性,可选用AlphaFold系列模型或基于蛋白质语言模型的建模工具tFold-TCR。在MHC-II类复合物的PFRs构象分析中,可将AlphaFold2的预测结果作为参考。在CDRs构象柔性的研究中,可引入TCR-pHLA_ModellerS这类模板驱动方法进行对照分析,以降低因构象过度收敛而掩盖关键局部特征的风险。

综合看来,每类建模工具都面临着独特的技术挑战。对通用建模工具AlphaFold来说,现有框架虽能较好地复现整体折叠与部分界面几何,但在MHC-II类复合物建模过程中,需提高TCR CDRs构象柔性区分能力,以减少对TCR CDRs的过度刚化。对于TCR-pMHC复合物改造的深度学习模型来说,需提升MHC-II类复合物在PFRs延伸构象中的细节刻画能力。对同源建模方法而言,主要瓶颈可能在于模板覆盖数量不足。当缺乏高相似模板时,模型容易受模板偏置影响而限制界面构象,因此需扩大高质量模板以降低模板缺失导致的建模误差。

不可否认的是,本文的研究也存在一定局限。在工具选取方面,本文优先选择可同时对MHC-I类和MHC-II类复合物建模、稳定运行且支持公开访问的工具。受限于工具可获得性、输入要求、运

行环境依赖与建模成功率等因素,未能纳入所有已发表或更新较快的建模方法,这可能缩小比较范围。在评测数据集方面,由于排除了训练数据,评估数据集偏少,尤其是MHC-II类复合物数量有限,且PFRs相关分析还受限于核心肽与两端PFRs可从文献中明确标注的样本,因此在统计学上难以充分检验工具间是否具有显著差异。

尽管存在这些局限,本文的基准测试在区分MHC-I和MHC-II计算建模差异的视角下,提供了对TCR-pMHC复合物结构建模工具的全面评估框架。通过提出PFRs-DI、N-PFR-DI、C-PFR-DI与CCC评估指标,有望更系统地刻画TCR-pMHC复合物在不同尺度上的结构特征,为MHC-I类与MHC-II类复合物的结构预测工具选择和建模策略制定提供统一的分析框架。该框架可为后续免疫计算方法的优化与开发提供实用指导,从而推动感染性疾病及自身免疫性疾病相关抗原识别机制的研究,并进一步支持肿瘤新抗原特异性候选TCR筛选与优化、个性化新抗原疫苗研发以及精准免疫治疗策略制定。

附件 见本文网络版 (<http://www.pibb.ac.cn>, <http://www.cnki.net>):

PIBB_20260015_Document_S1.pdf

PIBB_20260015_Figure_S1.pdf

PIBB_20260015_Table_S1.xlsx

PIBB_20260015_Table_S2.xlsx

PIBB_20260015_Table_S3.xlsx

PIBB_20260015_Table_S4.xlsx

PIBB_20260015_Table_S5.xlsx

PIBB_20260015_Table_S6.pdf

数据可用性声明 本论文的关联数据 (DOI: 10.57760/sciencedb.32190/CSTR: 31253.11.sciencedb.32190) 可在科学数据银行 (Science Data Bank, ScienceDB; <https://www.scidb.cn/s/qyU7Ff>) 中访问获取。

参 考 文 献

- [1] Zhang L, Song W, Zhu T, *et al.* ConvNeXt-MHC: improving MHC-peptide affinity prediction by structure-derived degenerate coding and the ConvNeXt model. *Brief Bioinform*, 2024, **25**(3): bbae133
- [2] Motmaen A, Jude K M, Wang N, *et al.* Targeting peptide-MHC complexes with designed T cell receptors and antibodies[DB/OL].

- New York: bioRxiv, 2025 [2026-01-02]. <https://www.biorxiv.org/content/10.1101/2025.11.19.689381v1.full>
- [3] Jian F, Cai H, Chen Q, *et al.* OnmiMHC: a machine learning solution for UCEC tumor vaccine development through enhanced peptide-MHC binding prediction. *Front Immunol*, 2025, **16**: 1550252
 - [4] Park H, Krog J, Carter B, *et al.* T-cell receptors that are *k*-binding have defined sequence features. *Front Immunol*, 2025, **16**: 1621201
 - [5] Chemla Y, Itzhaki O, Melamed S, *et al.* HLA export by melanoma cells decoys cytotoxic T cells to promote immune evasion. *Cell*, 2026, **189**(1): 233-251.e29
 - [6] Qi F, Huang Q, Xuan Y, *et al.* A roadmap for T cell receptor-peptide-bound major histocompatibility complex binding prediction by machine learning: glimpse and foresight. *Brief Bioinform*, 2025, **26**(4): bbaf327
 - [7] Gray G I, Chukwuma P C, Eldaly B, *et al.* The evolving T cell receptor recognition code: the rules are more like guidelines. *Immunol Rev*, 2025, **329**(1): e13439
 - [8] Ye F, Chen M, Huang Y, *et al.* LightCTL: lightweight contrastive TCR-pMHC specificity learning with context-aware prompt. *Brief Bioinform*, 2025, **26**(3): bbaf246
 - [9] Rodrigues Junior L, Bonorino C, Haubert A F, *et al.* Keeping it local: how CD8 TRMs regulate viral and cancer immunity. *Front Immunol*, 2025, **16**: 1618700
 - [10] Shi Y, Wang M, Dai B, *et al.* T cell-intrinsic PRR signaling in immunity and pathology. *Acta Biochim Biophys Sin*, 2026, **58**(1): 39-48
 - [11] Kim C A, Shelton N, Samudrala M, *et al.* AutoEpiCollect 2.0: a web-based machine learning tool for personalized peptide cancer vaccine design. *Molecules*, 2025, **30**(24): 4702
 - [12] Buligin A, Golikova E, Sennikov S. Development of $\alpha\beta$ and $\gamma\delta$ T cells in the thymus and methods of analysis. *Int J Mol Sci*, 2025, **26**(24): 11939
 - [13] Cheng X, Wu H, Chen P, *et al.* Progress of deep learning prediction of CD8⁺ T-cell epitopes. *Proteomics*, 2026, **26**(1): 6-21
 - [14] Gaiger N S, Coburn J, Lee M N, *et al.* Phenotypic and functional characterization of tumor-reactive T cells in malignant pleural effusions. *bioRxiv*, 2025: 2025.12.05.692662
 - [15] Spitzer F S, Camps M G M, Britten C M, *et al.* Dysfunctional CD4 T cells in an oncovirus-specific TCR-transgenic *in vivo* model. *Nat Commun*, 2025, **17**(1): 870
 - [16] Serçinoğlu O. Understanding allosteric stabilization of pMHC by peptide ligands through molecular dynamics simulations. *Turk J Immunol*, 2025, **13**(3): 132-145
 - [17] Nolan D S, Altenburg A F. Viral manipulation of the HLA class I antigen processing and presentation pathway. *Front Immunol*, 2025, **16**: 1694498
 - [18] Intonti S, Kokona D, Zinkernagel M S, *et al.* Glia modulates immune responses in the retina through distinct MHC pathways. *Glia*, 2025, **73**(4): 822-839
 - [19] Huang Z, Xu J, Wang Y, *et al.* The role of MHC molecules in cancer immunotherapy: insights into tumor immune evasion and therapeutic strategies. *Crit Rev Oncol Hematol*, 2026, **218**: 105057
 - [20] Wang M Y, Qiao Y, Wei S J, *et al.* MHC class II of different non-professional antigen-presenting cells mediate multiple effects of crosstalk with CD4⁺T cells in lung diseases. *Front Med*, 2025, **12**: 1388814
 - [21] Kusano S, Ueda S, Oryoji D, *et al.* Contributions of the N-terminal flanking residues of an antigenic peptide from the Japanese cedar pollen allergen Cr_j 1 to the T-cell activation by HLA-DP5. *Int Immunol*, 2023, **35**(9): 447-458
 - [22] Rademaker D T, Parizi F M, van Vreeswijk M, *et al.* Predicting reverse-bound peptide conformations in MHC Class II with PANDORA. *Front Immunol*, 2025, **16**: 1525576
 - [23] Saksager A B, Asmussen S R, Hede F D, *et al.* Evidence of altered antigen processing in autoimmune disease revealed by comparative immunopeptidomics[DB/OL]. New York: bioRxiv, 2025 [2026-01-01]. <https://www.biorxiv.org/content/10.64898/2025.12.12.693959v1>
 - [24] Asgary A H, Aleyasin A, Mehl J A, *et al.* PMGen: from peptide-MHC structure prediction to peptide generation[DB/OL]. New York: bioRxiv, 2025[2026-03-06]. <https://www.biorxiv.org/content/10.1101/2025.11.14.688404v2>
 - [25] Hulin-Curtis S, Geary J K, MacLachlan B J, *et al.* A targeted single mutation in influenza A virus universal epitope transforms immunogenicity and protective immunity *via* CD4⁺ T cell activation. *Cell Rep*, 2024, **43**(6): 114259
 - [26] McMaster B, Thorpe C J, Rossjohn J, *et al.* Quantifying conformational changes in the TCR: pMHC-I binding interface. *Front Immunol*, 2024, **15**: 1491656
 - [27] Jeon Y, Choi J, Roh Y, *et al.* Cytoplasmic delivery of antibodies through grafting a functional single complementarity-determining region loop. *FEBS Lett*, 2025, **599**(10): 1442-1455
 - [28] Han X, Han X, Hao Y, *et al.* Identification of novel KRAS^{G12D} neoantigen specific TCRs and a strategy to eliminate off-target recognition. *J Transl Med*, 2025, **23**(1): 78
 - [29] Mishra S, Swain P S, Pati S, *et al.* Extracellular domain of TREM2 possess two distinct ligand recognition sites: insights from machine-learning guided docking and all-atoms molecular dynamics simulations. *Heliyon*, 2025, **11**(1): e41414
 - [30] Xu X, Giulini M, Bonvin A M J J. Improved prediction of antibody and their complexes with clustered generative modelling ensembles. *Bioinform Adv*, 2025, **5**(1): vbaf161
 - [31] Gaudreault F, Sulea T, Corbeil C R. AI-augmented physics-based docking for antibody-antigen complex prediction. *Bioinformatics*, 2025, **41**(4): btaf129
 - [32] Greenshields-Watson A, Abanades B, Deane C M. Investigating the ability of deep learning-based structure prediction to extrapolate and/or enrich the set of antibody CDR canonical forms. *Front Immunol*, 2024, **15**: 1352703
 - [33] Deleuran S N, Nielsen M. NetTCR-struc, a structure driven approach for prediction of TCR-pMHC interactions. *Front Immunol*, 2025, **16**: 1616328

- [34] Akil C, Peng Y, McCarthy J, *et al.* Uncovering structural determinants of peptide recognition by public and private T-cell receptors. *bioRxiv*, 2025: 2025.11.23.690020
- [35] Slone J K, Conev A, Rigo M M, *et al.* TCR-pMHC binding specificity prediction from structure using graph neural networks. *IEEE Trans Comput Biol Bioinform*, 2025, **22**(1): 171-179
- [36] Borrmann T, Cimons J, Cosiano M, *et al.* ATLAS: a database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins*, 2017, **85**(5): 908-916
- [37] Leem J, de Oliveira S H P, Krawczyk K, *et al.* STCRDab: the structural T-cell receptor database. *Nucleic Acids Res*, 2018, **46**(D1): D406-D412
- [38] Lin V, Cheung M, Gowthaman R, *et al.* TCR3d 2.0: expanding the T cell receptor structure database with new structures, tools and interactions. *Nucleic Acids Res*, 2025, **53**(D1): D604-D608
- [39] Wei M, Wu J, Bai S, *et al.* TRAIT: a comprehensive database for T-cell receptor-antigen interactions. *Genomics Proteomics Bioinformatics*, 2025, **23**(3): qzaf033
- [40] Reche P A, Zhang H, Glutting J P, *et al.* EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics*, 2005, **21**(9): 2140-2141
- [41] Vita R, Blazeska N, Marrama D, *et al.* The immune epitope database (IEDB): 2024 update. *Nucleic Acids Res*, 2025, **53**(D1): D436-D443
- [42] Lata S, Bhasin M, Raghava G P. MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes*, 2009, **2**: 61
- [43] Gonzalez-Galarza F F, McCabe A, Santos E J M D, *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res*, 2020, **48**(D1): D783-D788
- [44] Marcu A, Bichmann L, Kuchenbecker L, *et al.* HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer*, 2021, **9**(4): e002071
- [45] Robinson J, Barker D J, Marsh S G E. 25 years of the IPD-IMGT/HLA database. *HLA*, 2024, **103**(6): e15549
- [46] Tickotsky N, Sagiv T, Prilusky J, *et al.* McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 2017, **33**(18): 2924-2929
- [47] Zhang W, Wang L, Liu K, *et al.* PIRD: pan immune repertoire database. *Bioinformatics*, 2020, **36**(3): 897-903
- [48] Shen W K, Zhang C Y, Gu Y M, *et al.* An automatic annotation tool and reference database for T cell subtypes and states at single-cell resolution. *Sci Bull*, 2025, **70**(10): 1659-1672
- [49] Yue T, Chen S Y, Shen W K, *et al.* TCRdb 2.0: an updated T-cell receptor sequence database. *Nucleic Acids Res*, 2026, **54**(D1): D504-D510
- [50] Shugay M, Bagaev D V, Zvyagin I V, *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res*, 2018, **46**(D1): D419-D427
- [51] Chen S Y, Liu C J, Zhang Q, *et al.* An ultra-sensitive T-cell receptor detection method for TCR-Seq and RNA-Seq data. *Bioinformatics*, 2020, **36**(15): 4255-4262
- [52] Chen S Y, Mao L, Fu X, *et al.* DeRR: a unique detecting method and the first landscape for T cells with dual T cell receptors from large-scale single cell data. *Genomics Proteomics Bioinformatics*, 2025. DOI: 10.1093/gpbjnl/qzaf090
- [53] Yang J M, Zhang N, Luo T, *et al.* TCellSI: a novel method for T cell state assessment and its applications in immune environment prediction. *Imeta*, 2024, **3**(5): e231
- [54] Yue T, Chen S Y, Shen W K, *et al.* TCRosetta: an integrated analysis and annotation platform for T-cell receptor sequences. *Genomics Proteomics Bioinformatics*, 2024, **22**(4): qzae013
- [55] Hudson D, Fernandes R A, Basham M, *et al.* Can we predict T cell specificity with digital biology and machine learning?. *Nat Rev Immunol*, 2023, **23**(8): 511-521
- [56] Klausen M S, Anderson M V, Jespersen M C, *et al.* LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res*, 2015, **43**(W1): W349-W355
- [57] Schmitt D, Li S, Rozewicki J, *et al.* Repertoire Builder: high-throughput structural modeling of B and T cell receptors. *Mol Syst Des Eng*, 2019, **4**(4): 761-768
- [58] Gowthaman R, Pierce B G. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res*, 2018, **46**(W1): W396-W401
- [59] Zhu T, Ren M, He Z, *et al.* Accurate structure prediction of immune proteins using parameter-efficient transfer learning. *New York: bioRxiv*, 2024[2026-01-03]. <https://www.biorxiv.org/content/10.1101/2024.11.13.621715v1>
- [60] Abanades B, Wong W K, Boyles F, *et al.* ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun Biol*, 2023, **6**(1): 575
- [61] Wu F, Zhao Y, Xiao Y, *et al.* Fast and accurate modeling of TCR-peptide-MHC complexes using tFold-TCR. *New York: bioRxiv*, 2025[2026-01-03]. <https://www.biorxiv.org/content/10.1101/2025.01.12.632367v1>
- [62] Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, **596**(7873): 583-589
- [63] Abramson J, Adler J, Dunger J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024, **630**(8016): 493-500
- [64] Quast N P, Abanades B, Guloglu B, *et al.* T-cell receptor structures and predictive models reveal comparable alpha and beta chain structural diversity despite differing genetic complexity. *Commun Biol*, 2025, **8**(1): 362
- [65] Li S, Wilamowski J, Teraguchi S, *et al.* Structural modeling of lymphocyte receptors and their antigens. *Methods Mol Biol*, 2019, **2048**: 207-229
- [66] Bradley P. Structure-based prediction of T cell receptor: peptide-MHC interactions. *Elife*, 2023, **12**: e82813
- [67] Yin R, Ribeiro-Filho H V, Lin V, *et al.* TCRmodel2: high-resolution modeling of T cell receptor recognition using deep

- learning. *Nucleic Acids Res*, 2023, **51**(W1): W569-W576
- [68] Jensen K K, Rantos V, Jappe E C, *et al.* TCRpMHCmodels: structural modelling of TCR-pMHC class I complexes. *Sci Rep*, 2019, **9**(1): 14530
- [69] Dai W, Li B, Li F, *et al.* TCR-pHLA_ModellerS. San Francisco: GitHub, 2025 [2026-01-01]. https://github.com/dwent/TCR-pHLA_ModellerS
- [70] Shi Y, Parks J M, Smith J C. Comparative analysis of TCR and TCR-pMHC complex structure prediction tools. *J Chem Inf Model*, 2025, **65**(13): 7156-7173
- [71] Ascunce-París A, Farriol-Duran R, Romero-Durana M, *et al.* A unified framework for TCR-pMHC structural model assessment. New York: bioRxiv, 2025[2026-01-04]. <https://www.biorxiv.org/content/10.1101/2025.10.09.681411v1>
- [72] Burley S K, Bhatt R, Bhikadiya C, *et al.* Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res*, 2025, **53**(D1): D564-D574
- [73] Katoh K, Standley D M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 2013, **30**(4): 772-780
- [74] Evans R, O'Neill M, Pritzel A, *et al.* Protein complex prediction with AlphaFold-Multimer. New York: bioRxiv, 2021 [2026-01-04]. <https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1>
- [75] Spoendlin F C, Fernández-Quintero M L, Raghavan S S R, *et al.* Predicting the conformational flexibility of antibody and T cell receptor complementarity-determining regions. *Nat Mach Intell*, 2025, **7**(10): 1755-1767
- [76] Wassenaar T M, Harville T, Chastain J, *et al.* DNA structural features and variability of complete MHC locus sequences. *Front Bioinform*, 2024, **4**: 1392613
- [77] Robert J, Najafi-Majd E. Evolutionary perspective of nonclassical MHC class I and innate-like T cells relevance in immune surveillance. *Cells*, 2025, **14**(20): 1592
- [78] Wu Y, Zhang N, Hashimoto K, *et al.* Structural comparison between MHC classes I and II; in evolution, a class-II-like molecule probably came first. *Front Immunol*, 2021, **12**: 621153
- [79] Schrödinger W D. Pymol. New York: Schrödinger, 2020[2026-01-02]. <http://www.pymol.org/pymol>
- [80] Biasini M, Schmidt T, Bienert S, *et al.* OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr*, 2013, **69**(Pt 5): 701-709
- [81] Yu Y, Zu L, Jiang J, *et al.* Structure-aware deep model for MHC-II peptide binding affinity prediction. *BMC Genomics*, 2024, **25**(1): 127
- [82] Lovitch S B, Pu Z, Unanue E R. Amino-terminal flanking residues determine the conformation of a peptide-class II MHC complex. *J Immunol*, 2006, **176**(5): 2958-2968
- [83] Conant S B, Swanborg R H. MHC class II peptide flanking residues of exogenous antigens influence recognition by autoreactive T cells. *Autoimmun Rev*, 2003, **2**(1): 8-12
- [84] Vanguru S, Bhattacharya D. To pack or not to pack: revisiting protein side-chain packing in the post-AlphaFold era. *Brief Bioinform*, 2025, **26**(3): bbaf297
- [85] Giudicelli V, Chaume D, Lefranc M P. IMGT/GENE-DB a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*, 2005, **33**(Database issue): D256-D261
- [86] Reinherz E L, Tan K, Tang L, *et al.* The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science*, 1999, **286**(5446): 1913-1921
- [87] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 2005, **33**(7): 2302-2309
- [88] Ribeiro R B, Cezar H M. Clusttraj: a solvent-informed clustering tool for molecular modeling. *J Chem Theory Comput*, 2025, **21**(14): 6759-6768
- [89] Liu C N, Denzler L M, Hood O E C, *et al.* Do antibody CDR loops change conformation upon binding?. *MAbs*, 2024, **16**(1): 2322533
- [90] Xu J, Zhang Y. How significant is a protein structure similarity with TM-score=0.5?. *Bioinformatics*, 2010, **26**(7): 889-895

A Computational Perspective on Differences Between MHC-I and MHC-II in TCR-pMHC Structure Prediction Resources: Review and Benchmarking*

WU Xiao-Qin¹⁾, LIU Da-Wei²⁾, LI Bin-Yu²⁾, LIU Yang^{3,4)}, CAO Yang^{3,4)**}, DAI Wen-Tao^{1,2)**}

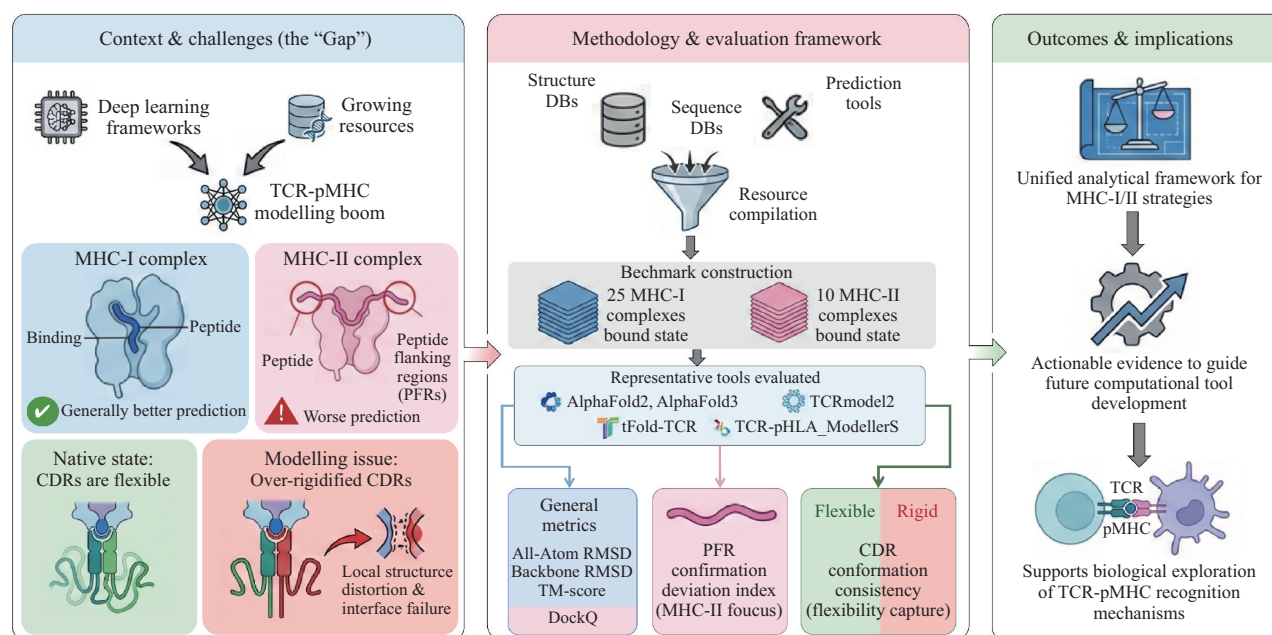
¹⁾School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;

²⁾Shanghai-MOST Key Laboratory of Health and Disease Genomics, NHC Key Lab of Reproduction Regulation, Shanghai Institute for Biomedical and Pharmaceutical Technologies, School of Pharmacy, Fudan University, Shanghai 200237, China;

³⁾Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China;

⁴⁾Animal Disease Prevention and Food Safety Key Laboratory of Sichuan Province, Microbiology and Metabolic Engineering Key Laboratory of Sichuan Province, College of Life Sciences, Sichuan University, Chengdu 610065, China)

Graphical abstract



Abstract The initiation of adaptive immune responses relies on the precise recognition and interpretation of antigenic information. In this process, the specific binding of T cell receptors (TCRs) to peptide-major histocompatibility complex (pMHC) molecules represents one of the key molecular events in the initiation of adaptive immune responses. Accordingly, the structural features of TCR-pMHC complexes provide a fundamental basis for dissecting antigen recognition mechanisms and support rational vaccine design, therapeutic target

* This work was supported by grants from the Innovation Promotion Program of NHC and the Shanghai Key Laboratory of SIBPT (RC2023-01, DT2025-05), Shanghai Academy of Science & Technology (SKY2022003), and The National Natural Science Foundation of China (32470696).

** Corresponding author.

CAO Yang. Tel: 86-28-85410043, E-mail: cao@scu.edu.cn

DAI Wen-Tao. Tel: 86-21-64771220, E-mail: daiwentao@sibpt.cn

Received: January 9, 2026 Accepted: April 2, 2026

discovery in TCR-based immunotherapy, and TCR identification and optimization. However, experimental determination of TCR-pMHC structures remains costly, time-consuming, and limited in coverage, making computational approaches essential for rapidly obtaining reliable structural information. Computational methods for predicting the structures of TCR-pMHC complexes have advanced rapidly in recent years, driven by progress in deep learning-based modeling frameworks and the increasing availability of structural and sequence resources. Despite these developments, most existing tools do not adequately distinguish the key structural and biophysical differences between MHC class I (MHC-I) and MHC class II (MHC-II) complexes during model construction. As a consequence, their predictive performance differs substantially between class I and class II complexes. In general, structural predictions for class I complexes outperform those for class II complexes. This discrepancy may be related to several fundamental differences between the two systems, including the architecture of the peptide-binding groove, the distribution of peptide lengths, and the properties of peptide flanking residues (PFRs). Compared with MHC-I molecules, MHC-II molecules usually bind longer antigenic peptides, which typically range from 13 to 25 amino acids in length. PFRs at both termini of these peptides participate in regulating the overall conformation of TCR-pMHC class II complexes and exert a pronounced effect on the geometric and physicochemical characteristics of the TCR-pMHC binding interface. Furthermore, within the TCR recognition interface, the complementarity-determining regions (CDRs) consist of segments that differ markedly in conformational behavior. They commonly include regions that are relatively rigid and structurally stable, together with highly flexible segments exhibiting substantial conformational plasticity. These rigidity-flexibility features constitute an essential structural basis enabling TCRs to recognize diverse peptide-MHC ligands and to accommodate conformational heterogeneity at the interface. However, many current modeling tools, in an effort to enforce global conformational stability or reduce structural noise, tend to over-constrain intrinsically flexible regions. Such oversimplification may lead to inappropriate rigidification of flexible CDR loops, resulting in local structural distortions, compromised interface geometry, or even complete modeling failure for specific complexes. Against this background, the review approaches the field from the perspective of computational differences between MHC-I and MHC-II complexes. We first systematically organize and summarize available resources related to TCRs and pMHCs, including structural datasets, sequence databases, prediction tools, and benchmarking studies. We then focus on five representative tools capable of predicting both class I and class II complexes—AlphaFold2, AlphaFold3, TCRmodel2, tFold-TCR, and TCR-pHLA_ModellerS. After excluding structures present in the training sets of these tools, we constructed a benchmark dataset comprising 25 class I and 10 class II TCR-pMHC complexes in the bound state and conducted a systematic evaluation using this dataset. We first employ widely used general evaluation metrics, including All-Atom Root Mean Square Deviation (All-Atom RMSD), Backbone RMSD, Template Modeling score (TM-score), and DockQ, to assess the global conformational accuracy and interface modeling quality of class I and class II complexes. For class II complexes, we propose for the first time a peptide flanking residue deviation index, including the PFRs-Deviation Index (PFRs-DI), N-PFR-Deviation Index (N-PFR-DI), and C-PFR-Deviation Index (C-PFR-DI), to quantitatively characterize conformational deviations in PFRs. In addition, we propose the CDR conformational consistency index (CCC) designed to qualitatively evaluate the ability of prediction tools to capture TCR CDR conformational flexibility. These metrics collectively assess a tool's ability to model both overall conformation and critical functional regions, thereby addressing the limitations of existing evaluation criteria that overemphasize global structure while inadequately capturing modeling quality in key functional areas. This establishes a unified analytical framework for MHC-I and MHC-II complexes to guide data resource selection, modeling strategy formulation, and evaluation system development. The framework further advances computational modeling and provides crucial support for multi-scale analysis of TCR-pMHC recognition mechanisms and their biological functions.

Key words immunology, TCR-pMHC complexes, structural prediction, PFR deviation index, CDR conformational consistency

DOI: 10.3724/j.pibb.2026.0015

CSTR: 32369.14.pibb.20260015