人工智能在抗体发现中的应用进展

闵 雪[1,2], 霍子瑞[2], 陶维红[1,3*]

(1. 河南大学, 抗体药物开发技术国家地方联合工程实验室, 河南 开封 475004; 2. 上海药物研究所分子影像中心, 上海 201203; 3. 上海溪长生物技术有限公司, 上海 201712)

摘要: 传统的抗体开发主要依赖动物实验、杂交瘤等技术的反复筛选和验证, 这导致抗体开发的生产周期长、成本高、成功率低。而最近, 人工智能 (artificial intelligence, AI) 基于蛋白质数据处理、计算等方面的巨大优势正不断推动抗体药物研发技术的进步, 在预测抗体结构、输出和优化靶向抗体序列等方面发挥着重要作用, 显著提高了抗体药物的命中率。因此, 本文将概述AI用于抗体发现与设计的最新发展与挑战, 并简要讨论传统抗体发现方法的局限性, 希望为AI驱动的抗体药物研发提供实用指导。

关键词: 抗体设计; 人工智能; 结构预测; 序列设计与优化

中图分类号: R914 **文献标识码**: A **文章编号**: 0513-4870(2026)04-1180-10

Advances of artificial intelligence in the discovery of antibody

MIN Xue[1,2], HUO Zi-rui[2], TAO Wei-hong[1,3*]

(*1. Joint National Laboratory for Antibody Drug Engineering, Henan University, Kaifeng 475004, China; 2. Molecular Imaging Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; 3. Shanghai Constream Biotech Co., Ltd., Shanghai 201712, China*)

Abstract: Conventional antibody discovery paradigms mainly rely on animal immunization and hybridoma screening, which results in long development lead times, high costs, and a low success rate for antibody therapeutics. Recently, artificial intelligence (AI) is constantly driving the progress of antibody drug research and development technology with advantages in large-scale proteomic data processing. It plays a great important role in the prediction of antibody structure and *de novo* generation of target-specific sequences, which also substantially enhances the hit rate of drug candidates. Therefore, this article aims to summarize recent developments and challenges in the application of AI technology to antibody discovery and design, while also considering the limitations to traditional methodologies. Hope to provide practical guidance for AI-driven antibody drug development.

Key words: antibody discovery; artificial intelligence; structure prediction; sequence design and optimization

自21世纪以来, 治疗性抗体药物在全球药品市场中的比重持续上升, 受到了广泛关注。主流抗体开发方法包括杂交瘤技术、单B细胞抗体技术、转基因小鼠技术以及抗体文库展示技术等。其中, 杂交瘤技术因其特异性单抗的无限量、标准化生产而被视为生物技术领域的重要突破[1]。单B细胞抗体技术通过从免疫动物体内直接提取B细胞的抗体基因, 并构建表达载体, 以进行抗体筛选[2]。转基因小鼠技术则可模拟人体免疫反应, 其生成的抗体无需额外的人源化或亲和力成熟过程。上述方法虽优势各异, 却普遍依赖动物免疫, 且生产抗体的周期漫长, 显著制约了特定抗体的制备与优化。抗体文库展示技术虽摆脱了动物依赖, 但其筛选效果仍受限于抗体库的容量与多样性[3]。因此, 突破传统抗体发现技术的局限, 构建高效且普适的

收稿日期: 2025-08-16; 修回日期: 2026-01-10.
*通讯作者 E-mail: taoweihong123@163.com
DOI: 10.16438/j.0513-4870.2025-0967

新一代抗体筛选平台, 已成为抗体药物研发领域的重要发展方向。

如图 1 所示, 人工智能 (artificial intelligence, AI) 技术已贯通抗体发现与设计的各环节。首先, AI 辅助的抗体设计摆脱了依赖动物免疫的局限, 弥补了抗体展示文库容量不足的缺陷。一方面, AI 算法理论上能实现自主设计并筛选所有可能存在的抗体序列, 从中得到最优候选序列。这极大扩展了基于传统手段发现抗体序列的边界。例如, 最近报道的基于 RFdiffusion 的抗体设计, 从靶点特定表位出发, 生成的原子精度 VHH 在结构上与 PDB (protein data bank) 数据库中已知的抗体结构具有极大差异, 验证了全新序列计算与设计的可能性[4]。另一方面, 针对 GPCR、离子通道、病毒膜蛋白等“难成药”靶点, AI 设计可把“表位–构象”同时纳入优化目标, 绕过天然免疫耐受而设计并优化抗体序列。Nabla Bio 团队[5]的模型在 8 周内设计并验证了三百余种靶向 G-蛋白偶联受体的激活剂。其中 34 条候选肽的 EC_{50} 较文献基准分子提高了 1～1.5 个数量级, 且无需动物免疫或文库筛选。这体现出 AI 在复杂膜蛋白靶点预测方面的速度与精度优势, 帮助突破了传统免疫耐受与文库覆盖度不足的双重瓶颈。鉴于此, 本文系统总结了 AI 在蛋白质结构预测、抗体设计与优化过程中的最新进展, 主要关注语言模型和扩散模型的应用。同时讨论了 AI 在抗体开发中面临的核心挑战, 希望为 AI 推动抗体药物设计提供指导。

1　AI 应用于抗体发现

1.1　结构预测

1.1.1　传统蛋白质结构预测　靶点蛋白的三维结构是基于结构的抗体药物开发的关键起点。传统上, 蛋白质三维结构主要通过结构生物学实验方法 (如 X 射线晶体学、冷冻电镜) 获取, 这依赖大量实验数据, 过程耗时且成本高昂。AI 算法的引入打破了“先有靶点结构再进行抗体设计”的传统流程, 促使蛋白质预测向几何优化建模和端对端的结构预测发展。其以近原子精度、低成本、高通量的优势, 为抗体设计带来了革命性的改变。

几何优化建模与端对端的模型预测在结构输出时各具优势, 互为补充。前者是通过预测残基间的空间约束, 并结合基于知识或物理的力场来优化算法以生成蛋白质的三维结构。这使其在模型构建时方法可控, 过程透明。然而, 缺点是其在处理大分子量蛋白质时计算成本较高, 且对蛋白质的实验数据依赖性较

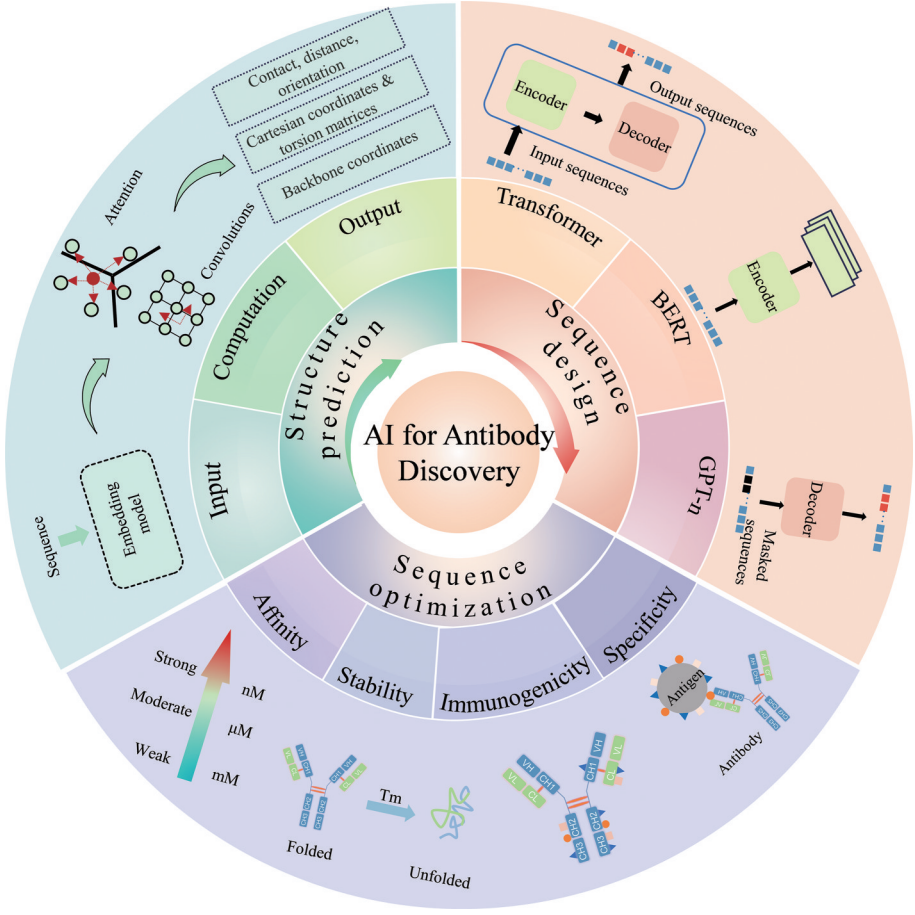


Figure 1　Artificial intelligence for antibody discovery

强[6]。相比之下, 端到端的结构预测方法直接从序列到三维结构, 这使其在预测速度和泛化能力方面具有显著优势。以RoseTTAFold和AlphaFold2为代表的深度学习网络, 使用注意力机制与先进的神经网络模块学习蛋白质的共进化–几何约束, 从而能在短时间内输出预测结构。二者进行骨架预测的中位精度最高可达0.96 Å, 与冷冻电镜实验密度的吻合度处于实验误差范围内[7,8]。这充分验证了模型的可靠性。但端对端的模型在训练时往往使用多序列比对 (multiple sequence alignment, MSA) 数据集, 对于快速进化的蛋白质序列结构, 尤其是具有高变异性的抗体互补决定区 (complementarity-determining region, CDR) 序列结构, 局部序列特征的建模能力会显著下降, 导致预测可靠性降低[9,10]。这时几何优化模型可以依靠物理知识与经验给出合理构象推测, 一定程度上弥补了端对端模型的不足。

1.1.2 无MSA的蛋白质结构预测 近年来, 无MSA的蛋白结构预测模型在生物制药领域也展现出了巨大的潜力。蛋白质语言模型 (protein language model, PLM) 能够将传统依赖MSA的结构预测流程简化为"序列到结构"的单步映射。其原理是大规模自监督模型可以学习序列的统计学规律并捕捉氨基酸残基上下文之间的嵌入特征, 将这些共进化信息编码于模型参数中, 使其在推理预测时不需再次检索同源序列数据集[11,12]。这极大提升了蛋白质结构预测的速度。OmegaFold和ESMFold模型作为PLM的代表性应用, 在不依赖MSA或已知结构模板的情况下, 生成原子级预测结构时的速度较AlphaFold和RoseTTAFold提升了一到两个数量级[13]。PLM的规模越大, 模型预测的精准度越高, 在信息检索上也更高效。创新模型HelixFold-Single的蛋白质序列训练规模达到3亿个, 即使在预测超过800个氨基酸的大质量蛋白质时也能展现出较MSA搜索方法和AlphaFold2推理更低的时间成本[14]。

无MSA方法的另一个重要优势是擅长处理孤儿蛋白或新型蛋白质。基于单序列蛋白质结构预测的自动化算法trRosettaX-Single在孤儿蛋白质的预测上明显优于源自MSA的AlphaFold2。对25种孤儿蛋白预测时, trRosettaX-Single的预测精度比AlphaFold2高出约14%, 预测的蛋白质结构与已知晶体结构的匹配度更高[15]。RGN2网络使用基于PLM的AminoBERT模型从单一序列预测蛋白质结构。相较AlphaFold2和RoseTTAFold, RGN2在孤儿蛋白及多种设计蛋白上的预测精度更接近实验误差, 均方根偏差 (root-mean-square deviation, RMSD) 数值平均低1.4 Å。在预测速度方面, RGN2的计算时间最多可缩短100万倍[16]。这使宏基因组孤儿蛋白结构注释成为可能, 显著拓宽了抗体与疫苗靶点的发现空间。

无MSA方法在预测速度上具有显著优势, 但对于具有大量同源序列的靶标蛋白, MSA依赖方法可能提供更高的精度[17]。当前较先进的PLM应用——MSA Transformer和AlphaFold的EvoFormer模块, 便是以进化相关蛋白质的MSA作为输入, 将MSA集合中包含的协同进化信息纳入注意力机制的蛋白质语言神经网络中[18]。MSA Transformer模型的这一特点使其可以高准确率地预测蛋白质二级结构以及氨基酸残基之间的长距离接触, 并在蛋白质结构预测中展现了较其他算法更优的表现[19-21]。鉴于短肽和抗体CDR的高度变异性, 将PLM和MSA结合更有望提高预测的准确性和适用性, 帮助治疗性抗体的设计与开发[22,23]。

1.1.3 抗体结构预测 相较于普通蛋白质结构, 抗体CDR H3区域的结构预测因其复杂多变的构象成为各研究机构的一大挑战。当前的深度学习方法在CDR环建模方面取得了显著进展。具体表现在: 首先, 使用抗体专业数据库进行预训练, 显著提升了深度学习模型预测抗体结构的精度。这是因为数据集的序列–结构分布与真实多样性出现显著偏差时, 模型便会产生系统性误差。ABodyBuilder平台的实验数据印证了这一规律: 在VHH、长CDR H3及新物种抗体等数据稀缺区域, 模型平均RMSD升至2.1 Å; 而在数据充足的保守结构域, RMSD仍能保持在0.5 Å的亚埃级水平[24], 二者差距直观地反映了数据偏差对预测可靠性的决定性影响。因此, 当前的抗体预测模型研究策略更多集中在构建专用数据库以进行模型训练。NanoNet模型使用具有更长CDR H3环结构的纳米抗体数据集进行训练, 能显著提升预测抗体V_H结构域时的准确度。模型在CDR H3环预测RMSD可达2.3 Å, 而常规抗体数据集预测模型RosettaAntibody为6.5 Å[25]。这同时也体现了专用的抗体数据库对特异性结构抗体预测方法的重要性。其次, 无MSA为无序区域的预测提供了可能。不同抗体在空间折叠上保持进化独立, 这使得CDR H3环的MSA结果不可靠。Ruffolo等[26-28]开发的DeepH3、DeepAb、DeepSCAb系列深度学习算法在无MSA的CDR H3结构预测上表现良好, 结构识别率RMSD分别达到2.2 ± 1.1、2.3 ± 1.3、3.2 ± 1.3 Å。这主要是因为残差神经网络提高了模型对数据集的表征能力, 使其可以从抗体链序列中学习残基间几何结构特征。预测过程不依赖MSA, 而是以高度变异序列的单个信号直接主导模型的学习。这使模型更专注学习CDR H3构象的关键残基对信息, 从而有效提高对CDR H3环特征结构的预测准确性。

此外, 深度学习模型针对CDR H3引入了特殊的几何约束。将物理几何规则嵌入模型参数中, 模型能够更好地捕捉抗体结构的本质特征, 减少了对大规模标记数据的依赖。有研究证实了在抗体CDR环结构设计过程中使用约束能量模型来表示CDR环结构的有效性。相较无约束能量模型, 使用约束进行预测时RMSD降低了33.4%, 而所需样本量减少了三分之一[29]。Abx模型在结构空间中引入物理几何约束, 从而生成具有特定结构要求的抗体。消融实验验证了移除约束后亲和力下降8.3%, 同时结构偏差增加了0.5 Å, 体现了几何约束对于确保抗体序列正确折叠的重要性[30]。

1.2　抗体设计

1.2.1　基于结构的抗体设计　蛋白质设计的经典方法遵循"结构优先"原则, 即预先设定目标三维结构骨架, 再搜寻与之匹配的氨基酸序列。抗体设计同样沿用此范式, 基于结构的抗体设计核心内容主要是抗体和抗原结构建模、表位–抗原结合位点预测。

如此前所述, 多种用于抗体结构预测的模型已取得显著进展。但归因于抗体–抗原相互作用的复杂机制, 表位–抗原结合位点预测尤其是构象表位的预测仍面临极大挑战。传统计算方法是用现有的抗体–抗原结合界面物理数据集进行训练, 以学习用于区分表位残基的关键结构特征[31]。然而, 抗原表面的任何氨基酸残基簇都可能与抗体结合, 表位残基的区分标准仍需要人为界定。随着深度学习的引入, 构象表位预测不再依赖局部结构特征, 而是把序列、结构信息和局部、全局特征整合起来。

图神经网络 (graph neural network, GNN) 擅长将三维坐标转化为图形式, 研究局部与整体的关联性特征, 尤其擅长构象表位预测。Paragraph、EPMP模型利用GNN的这一特点实现表位–抗原结合位点相关残基的快速预测[32,33]。PECAN模型同时考虑抗体和抗原的整体结构, 将其转化为图表示, 并通过图卷积网络、注意力机制和转移学习机制学习整体结构的表示, 最终预测抗体和抗原的结合界面。该模型在抗体结合位点结构预测中精确率–召回率曲线下的面积达到0.70, 而抗原表位预测中为0.21[34]。SEMA方法通过迁移学习的方法微调ESM-1v和ESM-IF1模型, 基于主要抗原的序列和三级结构来进行表位预测。在独立测试中ROC 曲线下面积分别达到0.756与0.759, 显著优于仅依赖单一信息源的既有工具[35]。

实际上, 设计抗原结合位点除了进行表位预测, 还需要考虑抗体–抗原对接。许多抗体设计往往因为对接表现不佳而被淘汰。即使是在蛋白质–蛋白质连接预测中表现优异的AlphaFold2, 被用于抗体–抗原复合物的准确建模仍具有许多挑战[36]。因此, 提升抗体–抗原复合物预测仍是需要被进一步解决的问题。

1.2.2　基于序列的抗体设计　基于序列的抗体设计突出的特点是序列数据库容量非常大, 更有利于挖掘稀有蛋白质及抗体的特性, 帮助结构未知蛋白质和抗体的设计。事实上, 近两年开发的PLM在推理蛋白质序列方面表现优异。ProGen2、proseLM、xTrimoPGLM-100B等语言模型无需结构模板即可在特定拓扑、分子互作以及零样本功能预测方面展现出独特优势[37-39]。而早期利用ProtGPT2语言模型生成的序列集具有92.9%的同源命中率, 其规避了天然序列数据集中出现的高度保守长结构域 (长度大于15个氨基酸)[40]。这提示了语言模型并非简单地重复训练数据集, 而是内化了序列中隐含的共进化规律。这一特点也为特异性抗体和靶向蛋白的按需设计奠定了算法依据。

将PLM从蛋白质序列扩展到抗体序列, 理论上也能学习抗体重轻链序列中隐含的"免疫语言"。专门针对抗体序列训练的语言模型能够精准识别并捕捉抗体的独特特性, 在生成抗体序列时能有效降低免疫原性、优化亲和力等关键特性, 提升了抗体序列的整体性能和应用潜力。此类模型的训练与推理过程如图2A所示: 海量抗体序列通过语言模型 (如掩码模型或自回归模型) 进行训练, 学习序列分布与保守特征。随后模型利用抗原–抗体配对数据或功能约束进行微调 (监督学习或强化学习等), 优化模型目标生成的能力。推理时, 只需给定抗原信息和设计约束即可输出满足要求的候选抗体。

预训练抗体生成语言模型PALM-H3基于未配对的抗体序列预训练, 经微调后结合抗原–抗体配对数据从头设计抗体的CDR H3[41]。对SARS-CoV-2新变体的高亲和力抗体预测进行结果验证, 预测抗体与XBB刺突蛋白上的关键轻链残基形成了高浓度的相互作用。最短的氢键为1.8 Å, 最长的可达2.6 Å, 显示出极强相互作用。这彰显了该模型在针对新型病原体开发高亲和力抗体的巨大应用潜力。仿生抗体语言模型BALMFold使用包含3.36亿条40%非冗余的未标记抗体序列数据集训练。该模型可以捕捉抗体特异性和保守性质, 在抗原结合预测中表现优于通用方法, 为治疗开发工作提供了有价值的指导[42]。免疫球蛋白语言模型IgLM采用自然语言的文本填充策略替代了单向上下文生成序列的方法而进行抗体设计。其在抗体V_H和V_L序列上进行了训练, 能够生成链类型不同以及物种来源不同的全长抗体序列, 进一步拓展了抗体设计的应用范围[43]。

1.2.3　抗体协同设计　抗体序列决定结构, 结构决定

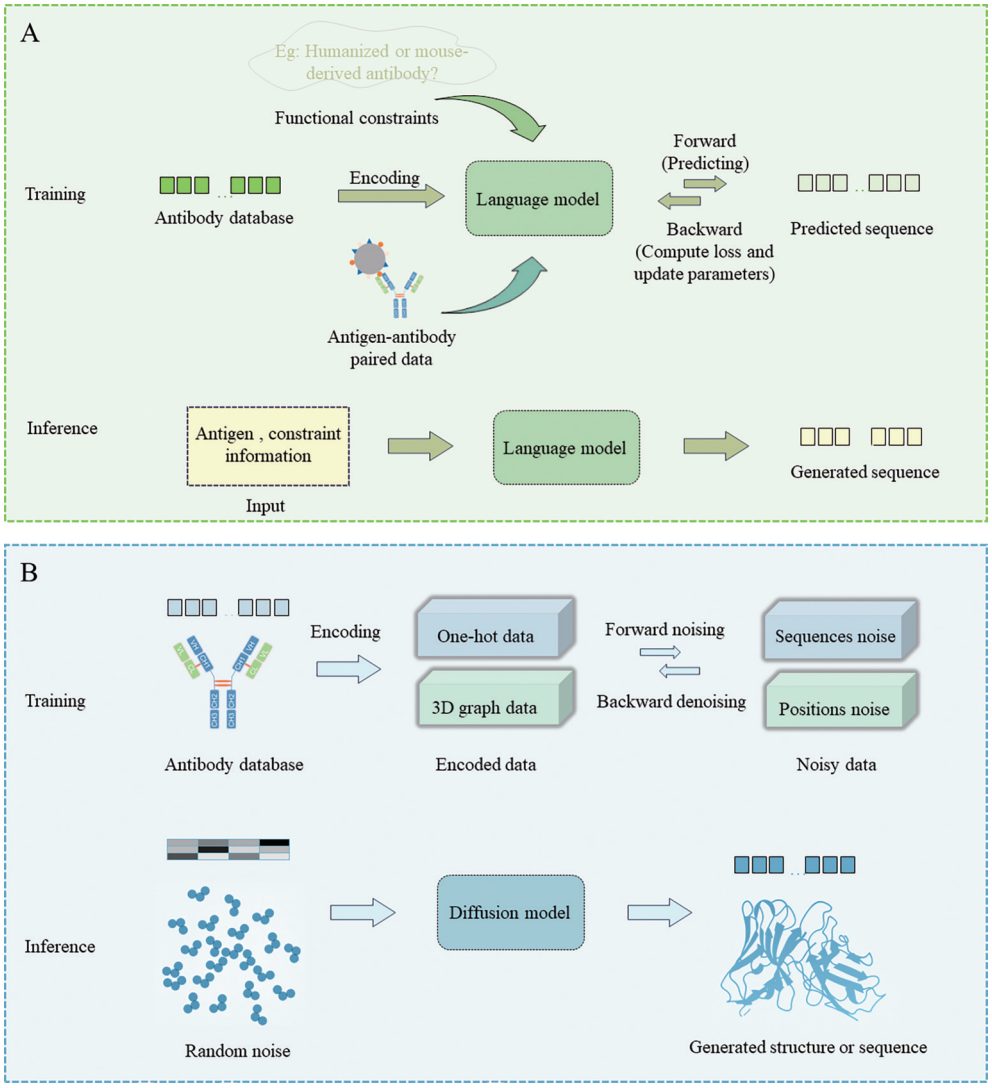


Figure 2 Common AI-based models for antibody design. A: The training and inference process of antibody language models; B: The training and inference process of diffusion model

功能。序列与结构的协同优化可同步提升抗体的亲和力、稳定性等特性, 并避免单一变量引发的折叠错误或空间冲突。扩散模型通过迭代去噪可同时更新原子坐标与氨基酸类型, 完成序列与结构协同采样。该类模型在抗体的从头设计中展现出超越传统方法的优势。如图2B所示, 语言模型与扩散模型在数据的处理方式上存在很大不同。语言模型常利用自回归和掩码模型以序列文本数据进行训练。而扩散模型在训练时, 会将抗体数据 (如三维结构坐标或氨基酸序列) 编码为图数据或独热向量, 通过前向过程逐步添加噪声破坏原始信息, 反向过程则学习从噪声中重建原始数据, 并通过损失函数优化重建能力。推理时, 从随机噪声出发, 通过训练好的降噪模型直接生成符合目标分布的抗体结构或序列。

代表模型之一RFdiffusion, 通过微调的RoseTTAFold结构预测网络进行迭代去噪而逐步生成蛋白质结构, 结合ProteinMPNN设计以预测匹配靶标结构的氨基酸序列[44]。ProteinGenerator同样基于RoseTTAfold网络, 但该模型是一个序列空间模型, 它从随机氨基酸序列开始迭代去噪, 生成序列和结构对[45]。DiffAb是首个抗体去噪扩散模型, 利用等变网络互补决定区的序列和结构进行联合建模, 能够同时优化抗体的序列和结构, 生成高度多样性的序列[46]。基于属性的从头抗体设计工具AntiBARTy Diffusion利用语言模型的潜在空间来训练属性条件扩散模型, 实现了具有特定特性的新抗体的定向设计[47]。近期开发的联合序列−结构扩散模型Antibody-SGM是一种创新的基于分数的生成模型[48]。它专注于抗体重链设计, 可以联合生成抗体的序列和结构, 突破骨架与序列分离设计的局限, 支持全原子抗体生成、抗原特异性CDR设计及序

列–结构同步优化, 为抗体设计提供了更高效的工具。总之, 扩散模型在生成可控氨基酸序列上展现出极大潜力, 被广泛用于特定功能或结构的蛋白质从头设计。其输出的多样性也显著促进了抗体新药的开发。

1.3　抗体序列优化

1.3.1　序列基础的优化　对天然抗体或计算机生成抗体的定向进化传统上依赖随机突变, 或者在重组后通过实验进行筛选目标序列。然而, 当前AI模型可以从序列和结构中学习到定向进化的规律, 这一优势在抗体的特性改造中发挥重要作用。多种语言模型集合的ESM-1v模型能够从蛋白质非冗余序列数据集学习序列的进化规则并对输入的序列提出合理的氨基酸突变, 显著提升了抗体的适应性。在亲和力优化方面, 该方法能将4种临床相关的高度成熟抗体的结合亲和力再次提高7倍, 而3种未成熟抗体的结合亲和力则提高了160倍[49]。Kang等[50]构建的亲和力预测模型利用了GNN和自然语言模型等深度学习技术来捕捉序列信息中对亲和力有利的氨基酸特征。通过保留特异性突变位点, 模型能将靶向SARS-COV2蛋白的抗体模板至高提升269.65倍, 这体现了计算机算法基于蛋白质序列信息简化抗体优化的巨大价值。

1.3.2　结构基础的优化　在不同的机器学习任务和嵌入中, 序列基础的可开发性参数 (developability parameters, DP) 相关性会显著高于结构基础的DPs。这强调了序列DP比基于结构的DP更具可预测性[51]。但这同时也凸显了基于结构的DP比基于序列的DP表现更高的独立性, 体现了结构参数在治疗性抗体设计中的重要性。AI抗体开发模型在联合抗体结构信息的情况下能设计出成药性更优的抗体序列。逆向折叠模型通过给定的蛋白质三维结构推导氨基酸序列, 在基于结构的抗体设计和蛋白质特性优化方面中发挥重要作用。例如, 专门针对抗体结构优化的微调反折叠模型ABMPNN通过结合GNN的局部结构感知能力和自回归解码器的序列生成能力, 能够捕捉氨基酸残基的局部空间和化学环境信息[52]。这使其可以结合抗体的框架信息来约束CDR环的预测, 在CDR序列恢复方面显著优于通用蛋白质模型ProteinMPNN。另一反向折叠模型ESM-IF1结合抗体–抗原复合物信息用以指导抗体的进化, 使两个靶向SARS-CoV-2的治疗性抗体的中和效果提高了25倍[53]。AntiFold方法通过微调ESM-IF1显著改进了CDR序列恢复和零样本亲和力预测。该模型用于预测抗体–抗原结合亲和力时较其他现有反向折叠工具展现更优越的性能[54]。

1.3.3　多特性协同优化　抗体优化本质上是对抗体变体特性进行筛选的过程, 而大多数抗体筛选方法仅能实现单一特性的筛选。例如, 噬菌体展示技术一次仅能优先选择单个属性进行筛选, 这限制了抗体的多种特性同时优化。为解决抗体多种特性协同进化困难的问题, Makowski等[55]尝试对抗体亲和力和特异性进行二元标记, 通过深度学习模型预测与亲和力及非特异性结合相关的连续指标, 从而识别兼具高亲和力与低非特异性结合的协同优化变体。结合上下文语言模型的高通量实验方法则通过预测序列变体的结合亲和力, 并依靠遗传算法筛选出与天然抗体高度相似的序列, 从而实现亲和力与免疫原性的协同优化[56]。这些方法可以进一步推广到抗体和蛋白质的其他特性筛选中。除此之外, 多种数据驱动的机器学习方法也专注于提升抗体的可开发性。针对抗体稳定性、溶解度和聚集性等生物物理特性的预测工具在基于语言模型和神经网络模型的算法加持下进一步提高了抗体药物可开发性和成功率[57-59]。这些研究结果表明了AI促进设计与优化抗体候选药物的重要作用, 加速了数据驱动的抗体工程的发展。

2　AI在抗体开发中面临的挑战

AI在抗体发现中的前景广阔, 但由于数据与技术限制, 当前应用也存在许多挑战。首先是缺乏合适数据, 抗体语言模型相对PLM仍缺乏大量合理数据支撑。大部分抗体专用语言模型基于OAS (observed antibody space) 和SAbDab (structural antibody database) 数据库, 如表1所示[38,41-43,60-74]。与蛋白质数据库相比, 抗体库容量小得多, 这一问题尤其表现在抗体的结构预测中。目前有效的抗体–抗原复合物结构的动态数据仍然缺乏, 这不利于针对抗体识别位设计的语言模型发展[75]。其次, 由于抗体CDR-H3环的高变结构, 针对抗体抗原复合物的建模依旧是十分困难的。抗体CDR环具有多个不稳定键, 在与抗原结合时发挥作用, 这使其不能由单个静态结构捕获或表示[76]。因此, 基于不同数据集的多种抗体预测方法在揭示CDR-H3环上的结果产生较大差异, 甚至部分模型预测结果具有显著的非天然缺陷[10]。另一方面, CDR-H3环的构象数量随长度呈指数增长, 导致AI模型在抗体的H3环预测的平均RMSD随环的长度增加而增加。这提示了具有长序列H3环的复杂抗体较难通过建模进行设计与优化[26]。多目标优化对当前的抗体模型来说也是一个重要的挑战。抗体的一系列特性参数并非一个线性依赖关系, 改善一种特性可能会影响一系列基本参数的变化[77]。抗体优化需要平衡功效、安全性和可制造性, 因此, 需要开发更复杂的算法来平衡抗体的亲和力、稳定性和免疫原性。

Table 1 Summary of antibody-specific language model. OAS: Observed antibody space; CDR: Complementarity-determining region; SAbDab: Structural antibody database; PlaAbDab: Patent and literature antibody database; CoV-AbDab: Coronavirus antibody database; MSA: Multiple sequence alignment; ADB: Antibody database; INDI: Integrated nanobody database for immunoinformatics; PDB: Protein data bank; KNN: K-nearest neighbor

Application	Model	Database	Character
Antibody design	IgLM	OAS	Leverages bidirectional context for designing; conditioned on chain type and species
	PALM-H3	OAS	Design CDR H3 sequence based on antigen-antibody pairing data
	p-IgGen	OAS	Generate low immunogenic sequences and matching antibody sequences
	MAGE	SAbDab, PlaAbDab, CoV-AbDab	Generate paired heavy and light chain sequences without template antibody or protein structural information
	IgHuAb	OAS, PlaAbDab	*De novo* generation of specific paired antibody sequences
Structure prediction	tFold-Ab	SAbDab	End-to-end prediction of backbone and side-chain conformations without MSA
	IgFold	OAS	End-to-end prediction of paired antibody atomic coordinates
	xTrimoABFold	ADB	Template search without MSA based on modalities of sequence and structure
	BALMFold	OAS, SAbDab	Prediction of full atomic structures, antibody functionality and the evolutionary trajectory without MSA or templates
Sequence recovery	AbLang	OAS	Sequence recovery without antibody germline information; generate three different types of antibody sequence representations
	IgBlend	OAS, SAbDab, PlaAbDab, INDI	Strong generalization ability; multi-modality encoder integrate structural and sequential data
	IgBert, IgT5	OAS	Learn inter-chain features from paired data; sequence recovery and binding affinity prediction
	proseLM	SAbDab, PDB, CATH 4.2	Incorporate structural and functional context
Sequence optimization	AntiBERTy	OAS	Study the process of antibody affinity maturation
	AbBERT	OAS	Assess sequence immunogenicity and optimizing antibody CDR sequences
	AntiBERTa	OAS, SAbDab	Provide self-attention heatmaps to enhance model interpretation; prediction of antibody binding epitopes and antibodies humanization
	BERT2DAb	OAS	Incorporate antibody secondary structure information; screen binding sequences and predict binding affinity
	PLAN	OAS	KNN model retrieves similar human antibody amino acids; D_E value guides the reverse mutation of key residues
	REALM	OAS	Embed co-evolutionary information of antibody regions and structures; detect antibody hydrophobicity and thermal stability

3 总结与展望

AI的不同机制和算法在处理庞大而复杂的抗体数据时具有不可替代的优势。不同模型通过捕捉抗体序列和结构中的复杂信息、解析抗体–抗原复合物, 从而高效生成多样化的抗体序列。但由于抗体结构复杂的作用机制, AI模型仍存在诸多挑战如预测精度不高、CDR H3环建模困难、多目标优化效率低等。随着越来越多优化模型的出现, AI驱动抗体药物设计也在不断趋于完善。AI算法也正与传统实验技术形成互补趋势。一方面传统实验为AI提供训练的数据并验证算法输出结果, 另一方面AI算法也为传统实验提供指导方向, 减少不必要的实验并对当前数据库进行验证与淘选, 从而提高数据集的整体质量并用以训练更高精度的模型。未来, AI将集成抗体研发的各个方面, 从靶点识别和先导化合物优化到临床试验设计与验证, 真正实现药物研发一体化, 从根本上重塑药物发现的格局。

作者贡献: 闵雪完成了文献收集、撰稿及修改和定稿; 霍子瑞完成了图表制作及论文修改; 陶维红提供了指导性意见。

利益冲突: 所有作者均声明没有利益冲突。

References

[1] Parray HA, Shukla S, Samal S, et al. Hybridoma technology a versatile method for isolation of monoclonal antibodies, its applicability across species, limitations, advancement and future perspectives [J]. Int Immunopharmacol, 2020, 85: 106639.

[2] Pedrioli A, Oxenius A. Single B cell technologies for monoclonal antibody discovery [J]. Trends Immunol, 2021, 42: 1143-1158.

[3] Fan Y, Feng R, Zhang X, et al. Encoding and display technologies for combinatorial libraries in drug discovery: The coming of age from biology to therapy [J]. Acta Pharm Sin B, 2024, 14: 3362-3384.

[4] Bennett NR, Watson JL, Ragotte RJ, et al. Atomically accurate *de novo* design of antibodies with RFdiffusion [J]. Nature, 2026, 649: 183-193.

[5] Nabla Bio, Biswas S. *De novo* design of hundreds of functional GPCR-targeting antibodies enabled by scaling test-time compute [J]. BioRxiv, 2025. https://doi.org/10.1101/2025.05.28.656709.

[6] Peng CX, Liang F, Xia YH, et al. Recent advances and challenges

in protein structure prediction [J]. J Chem Inf Model, 2024, 64: 76-95.

[7] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network [J]. Science, 2021, 373: 871-876.

[8] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596: 583-589.

[9] Fernández-Quintero ML, Math BA, Loeffler JR, et al. Transitions of CDR-L3 loop canonical cluster conformations on the micro-to-millisecond timescale [J]. Front Immunol, 2019, 10: 2652.

[10] Fernández-Quintero ML, Kokot J, Waibl F, et al. Challenges in antibody structure prediction [J]. MAbs, 2023, 15: 2175319.

[11] Ruffolo JA, Madani A. Designing proteins with language models [J]. Nat Biotechnol, 2024, 42: 200-202.

[12] Alanazi W, Meng D, Pollastri G. Porter 6: Protein secondary structure prediction by leveraging pre-trained language models (PLMs) [J]. Int J Mol Sci, 2024, 26: 130.

[13] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model [J]. Science, 2023, 379: 1123-1130.

[14] Fang X, Wang F, Liu L, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model [J]. Nat Mach Intell, 2023, 5: 1087-1096.

[15] Wang W, Peng Z, Yang J. Single-sequence protein structure prediction using supervised transformer protein language models [J]. Nat Comput Sci, 2022, 2: 804-814.

[16] Chowdhury R, Bouatta N, Biswas S, et al. Single-sequence protein structure prediction using a language model and deep learning [J]. Nat Biotechnol, 2022, 40: 1617-1623.

[17] Raisinghani N, Alshahrani M, Gupta G, et al. Probing functional allosteric states and conformational ensembles of the allosteric protein kinase states and mutants: Atomistic modeling and comparative analysis of AlphaFold2, OmegaFold, and AlphaFlow approaches and adaptations [J]. J Phys Chem B, 2024, 128: 11088-11107.

[18] Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships [J]. Nat Commun, 2022, 13: 6298.

[19] Chen Y, Chen G, Chen CYC. MFTrans: A multi-feature transformer network for protein secondary structure prediction [J]. Int J Biol Macromol, 2024, 267: 131311.

[20] Hong Y, Song J, Ko J, et al. S-Pred: Protein structural property prediction using MSA transformer [J]. Sci Rep, 2022, 12: 13891.

[21] Hong Y, Lee J, Ko J. A-Prot: Protein structure modeling using MSA transformer [J]. BMC Bioinform, 2022, 23: 93.

[22] Moussad B, Roche R, Bhattacharya D. The transformative power of transformers in protein structure prediction [J]. Proc Natl Acad Sci U S A, 2023, 120: e2303499120.

[23] Meng Q, Guo F, Tang J. Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model [J]. Brief Bioinform, 2023, 24: bbad217.

[24] Leem J, Dunbar J, Georges G, et al. ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation [J]. MAbs, 2016, 8: 1259-1268.

[25] Cohen T, Halfon M, Schneidman-Duhovny D. NanoNet: Rapid and accurate end-to-end nanobody modeling by deep learning [J]. Front Immunol, 2022, 13: 958584.

[26] Ruffolo JA, Guerra C, Mahajan SP, et al. Geometric potentials from deep learning improve prediction of CDR H3 loop structures [J]. Bioinformatics, 2020, 36: i268-i275.

[27] Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning [J]. Patterns, 2022, 3: 100406.

[28] Akpinaroglu D, Ruffolo JA, Mahajan SP, et al. Simultaneous prediction of antibody backbone and side-chain conformations with deep learning [J]. PLoS One, 2022, 17: e0258173.

[29] Fu T, Sun J. Antibody complementarity determining regions (CDRs) design using constrained energy model [C] // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. Washington DC, USA: Association for Computing Machinery, 2022: 389-399.

[30] Zhu T, Ren M, Zhang H. Antibody design using a score-based diffusion model guided by evolutionary, physical and geometric constraints [C] // Proceedings of the 41st International Conference on Machine Learning. Cambridge (MA): JMLR, 2024: 62531-62548.

[31] Sun J, Wu D, Xu T, et al. SEPPA: A computational server for spatial epitope prediction of protein antigens [J]. Nucleic Acids Res, 2009, 37: W612-W616.

[32] Chinery L, Wahome N, Moal I, et al. Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors [J]. Bioinformatics, 2023, 39: btac732.

[33] Del Vecchio A, Deac A, Liò P, et al. Neural message passing for joint paratope-epitope prediction [J]. ArXiv, 2021. https://doi.org/10.48550/arXiv.2106.00757.

[34] Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces [J]. Bioinformatics, 2020, 36: 3996-4003.

[35] Shashkova TI, Umerenkov D, Salnikov M, et al. SEMA: Antigen B-cell conformational epitope prediction using deep transfer learning [J]. Front Immunol, 2022, 13: 960985.

[36] Gaudreault F, Corbeil CR, Sulea T. Enhanced antibody-antigen structure prediction from molecular docking using AlphaFold2 [J]. Sci Rep, 2023, 13: 15107.

[37] Nijkamp E, Ruffolo JA, Weinstein EN, et al. ProGen2: Exploring the boundaries of protein language models [J]. Cell Syst, 2023, 14: 968-978.e3.

[38] Ruffolo JA, Bhatnagar A, Beazer J, et al. Adapting protein

language models for structure-conditioned design [J]. BioRxiv, 2024. https://doi.org/10.1101/2024.08.03.606485.

[39] Chen B, Cheng X, Li P, et al. xTrimoPGLM: Unified 100-billion-parameter pretrained transformer for deciphering the language of proteins [J]. Nat Methods, 2025, 22: 1028-1039.

[40] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design [J]. Nat Commun, 2022, 13: 4348.

[41] He H, He B, Guan L, et al. *De novo* generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model [J]. Nat Commun, 2024, 15: 6867.

[42] Jing H, Gao Z, Xu S, et al. Accurate prediction of antibody function and structure using bio-inspired antibody language model [J].Brief Bioinform, 2024, 25: bbae245.

[43] Shuai RW, Ruffolo JA, Gray JJ. IgLM: Infilling language modeling for antibody sequence design [J]. Cell Syst, 2023, 14: 979-989.e4.

[44] Watson JL, Juergens D, Bennett NR, et al. *De novo* design of protein structure and function with RFdiffusion [J]. Nature, 2023, 620: 1089-1100.

[45] Lisanza SL, Gershon JM, Tipps S, et al. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion [J]. BioRxiv, 2023. https://doi. org/10.1101/2023.05.08.539766.

[46] Luo S, Su Y, Peng X, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures [J]. Adv Neural Inf Process Syst, 2022, 35: 9754-9767.

[47] Venderley J. AntiBARTy diffusion for property guided antibody design [J]. ArXiv, 2023. https://doi. org/10.48550/arXiv. 2309.13129.

[48] Xie X, Valiente PA, Lee JS, et al. Antibody-SGM, a score-based generative model for antibody heavy-chain design [J]. J Chem Inf Model, 2024, 64: 6745-6757.

[49] Hie BL, Shanker VR, Xu D, et al. Efficient evolution of human antibodies from general protein language models [J]. Nat Biotechnol, 2024, 42: 275-283.

[50] Kang Y, Jin K, Pan L. AI designed, mutation resistant broad neutralizing antibodies against multiple SARS-CoV-2 strains [J]. Sci Rep, 2025, 15: 15533.

[51] Bashour H, Smorodina E, Pariset M, et al. Biophysical cartography of the native and human-engineered antibody landscapes quantifies the plasticity of antibody developability [J]. Commun Biol, 2024, 7: 922.

[52] Dreyer FA, Cutting D, Schneider C, et al. Inverse folding for antibody sequence design using deep learning [J]. ArXiv, 2023. https://doi.org/10.48550/arXiv.2310.19513.

[53] Hsu C, Verkuil R, Liu J, et al. Learning inverse folding from millions of predicted structures [C] // Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022: 8946-8970.

[54] Høie MH, Hummer AM, Olsen TH, et al. AntiFold: Improved structure-based antibody design using inverse folding [J]. Bioinform Adv, 2025, 5: vbae202.

[55] Makowski EK, Kinnunen PC, Huang J, et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space [J]. Nat Commun, 2022, 13: 3788.

[56] Bachas S, Rakocevic G, Spencer D, et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness [J]. BioRxiv, 2022. https://doi. org/10.1101/2022.08.16.504181.

[57] Harmalkar A, Rao R, Richard XY, et al. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features [J]. MAbs, 2023, 15: 2163584.

[58] Rosace A, Bennett A, Oeller M, et al. Automated optimisation of solubility and conformational stability of antibodies and proteins [J]. Nat Commun, 2023, 14: 1937.

[59] Makowski EK, Wang T, Zupancic JM, et al. Optimization of therapeutic antibodies for reduced self-association and non-specific binding *via* interpretable machine learning [J]. Nat Biomed Eng, 2024, 8: 45-56.

[60] Turnbull OM, Oglic D, Croasdale-Wood R, et al. p-IgGen: A paired antibody generative language model [J]. Bioinformatics, 2024, 40: btae659.

[61] Wasdin PT, Johnson NV, Janke AK, et al. Generation of antigen-specific paired chain antibody sequences using large language models [J]. BioRxiv, 2025. https://doi. org/10.1101/2024.12.20.629482.

[62] Marinov TM, Wasdin PT, Jordaan G, et al. An expandable synthetic library of human paired antibody sequences [J]. PLoS Comput Biol, 2025, 21: e1012932.

[63] Wu J, Wu F, Jiang B, et al. tFold-Ab: Fast and accurate antibody structure prediction without sequence homologs [J]. Biorxiv, 2022. https://doi.org/10.1101/2022.11.10.515918.

[64] Ruffolo JA, Chu LS, Mahajan SP, et al. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies [J]. Nat Commun, 2023, 14: 2389.

[65] Wang Y, Gong X, Li S, et al. xtrimoabfold: *De novo* antibody structure prediction without MSA [J]. ArXiv, 2022. https://doi.org/10.48550/arXiv.2212.00735.

[66] Olsen TH, Moal IH, Deane CM. AbLang: An antibody language model for completing antibody sequences [J]. Bioinform Adv, 2022, 2: vbac046.

[67] Malherbe C, Uçar T. Igblend: Unifying 3D structures and sequences in antibody language models [J]. BioRxiv, 2024. https://doi.org/10.1101/2024.10.01.615796.

[68] Kenlay H, Dreyer FA, Kovaltsuk A, et al. Large scale paired antibody language models [J]. PLoS Comput Biol, 2024, 20: e1012646.

[69] Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity

maturation with language models and weakly supervised learning [J]. ArXiv, 2021. https://doi.org/10.48550/arXiv.2112.07782.

[70] Vashchenko D, Nguyen S, Goncalves A, et al. AbBERT: Learning antibody humanness *via* masked language modeling [J]. BioRxiv, 2022. https://doi.org/10.1101/2022.08.02.502236.

[71] Leem J, Mitchell LS, Farmery JH, et al. Deciphering the language of antibodies using self-supervised learning [J]. Patterns, 2022, 3: 100513.

[72] Luo X, Tong F, Zhao W, et al. BERT2DAb: A pre-trained model for antibody representation based on amino acid sequences and 2D-structure [J]. mAbs, 2023, 15: 2285904.

[73] Zou H, Yuan R, Lai B, et al. Antibody humanization *via* protein language model and neighbor retrieval [J]. BioRxiv, 2023. https://doi.org/10.1101/2023.09.04.556278.

[74] Nishino T, Kato N, Tsutaoka T, et al. REALM: Region-empowered antibody language model for antibody property prediction [C] // 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Lisbon, Portugal: IEEE, 2024: 7104-7106.

[75] Dunbar J, Krawczyk K, Leem J, et al. SAbDab: The structural antibody database [J]. Nucleic Acids Res, 2014, 42: D1140-D1146.

[76] Regep C, Georges G, Shi J, et al. The H3 loop of antibodies shows unique structural characteristics [J]. Proteins, 2017, 85: 1311-1318.

[77] Warszawski S, Borenstein Katz A, Lipsh R, et al. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces [J]. PLoS Comput Biol, 2019, 15: e1007207.