

Pretrainable Geometric Graph Neural Network for Antibody Affinity Maturation

Huiyu Cai, Zuobai Zhang, Mingkai Wang, Bozitao Zhong, Yanling Wu, Tianlei Ying, Jian Tang

YANLINGWU, TLYING@FUDAN.EDU.CN, TANGJIAN@BIOGEOM.COM

Editor: A detailed contributor list can be found in the appendix of this paper.

Abstract

提高抗体对其靶抗原的结合亲和力是抗体治疗开发中的关键任务。本文提出了一种可预训练的几何图神经网络 GearBind，并探讨了其在计算亲和力成熟中的潜力。通过利用多关系图构建、多层次信息传递以及在大规模、无标记蛋白质结构数据上的对比预训练，GearBind 在 SKEMPI 数据集和独立测试集上表现优于现有的最先进方法。基于 GearBind 的强大集成模型成功用于增强两种具有不同形式和靶抗原的抗体的结合能力。设计的抗体突变体在 ELISA 实验中 EC_{50} 值降低了最多 17 倍， K_d 值降低了最多 6.1 倍。这些有前景的结果突显了几何深度学习的实用性以及在大分子相互作用建模任务中有效预训练的价值。

Keywords: GearBind, Geometric Deep Learning, Contrastive pretraining, Antibody Affinity Maturation, Computational Biology

Date: Aug 14, 2024

1 Introduction

抗体在人体免疫系统中起着关键作用，并且由于其能够以高亲和力选择性地靶向抗原，因此成为一种强大的诊断和治疗工具。在体内，抗体会经历亲和力成熟过程，通过体细胞超突变和克隆选择使其靶标结合亲和力逐渐提高^[1]。当遇到新的抗原表面时，从已知抗体重新利用的治疗性抗体药物，或从天然或全新设计的文库中筛选出的抗体，通常需要通过体外亲和力成熟来提升其结合亲和力，以达到通常亚纳摩尔级的水平。

体外抗体亲和力成熟的湿实验方法通常涉及多轮突变和筛选。尽管这些方法在过去几年中取得了显著改进，但总体上仍然耗时且成本高，尤其是在考虑到可能突变的组合搜索空间时^[2]。通常情况下，抗体的互补决定区（CDR）上有 50-60 个残基，这些残基在体内高度可变，并贡献了大部分的结合自由能 ΔG_{bind} ^[3]。先前的研究表明，成功的亲和力成熟通常需要多个点突变^[4, 5]。在每个抗体的 CDR 区域中，大约有 1000 个可能的点突变组合（60 个残基 × 每个残基 19 种氨基酸），要对其进行实验验证几乎是不可能的。因此，迫切需要一种快速且准确的计算方法来缩小搜索空间。

然而，计算亲和力成熟方法在平衡速度和准确性方面面临挑战。一方面，蛋白质系统过于庞大，难以通过分子动力学方法进行建模（更不用说更为准确但计算开销更高的方法了）。然而，考虑到需要对数千个突变及其组合进行建模，能够在合理的时间内完成此任务的更昂贵的量子力学方法更加难以实施。另一方面，经验性力场方法虽然速度更快，但无法充分捕捉到抗体-抗原之间的微妙相互作用，导致可靠性较低。近年来，机器学习（尤其是深度学习）展示了其解决这一复杂问题的巨大潜力。许多机器学习方法^[6-11] 将亲和力成熟问题表述为基于结构的结合自由能变化 ($\Delta \Delta G_{\text{bind}} := \Delta G_{\text{bind}}^{(\text{mt})} - \Delta G_{\text{bind}}^{(\text{wt})}$ ，其中 wt 表示野生型，mt 表示突变体) 的预测问题。然而，尽管蛋白质侧链构象对蛋白质-蛋白质相互作用至关重要，大多数现有方法要么忽略了原子级信息，要么仅间接建模。这些方法未能充分解决侧链原子之间复杂的相互作用。

另一个关键问题是，机器学习模型通常需要大量的配对数据才能保证准确性和可靠性。据我们所知，最大的公开可用的蛋白质结合自由能变化数据集——SKEMPI v2.0 数据库^[12]——仅包含 7,085 条 $\Delta \Delta G_{\text{bind}}$ 测量数据，这与 AlphaFold2^[13] 和 ESM2^[14] 等成功的深度学习模型所需的训练集相比，数据量微不足道。

为应对上述挑战，我们引入了 GearBind，这是一种深度神经网络，利用多层次几何信息传递以建模复杂的抗体-抗原相互作用。此外，通过对蛋白质结构数据进行基于对比学习的预训练，能够融合关键的结构信息，以预测 $\Delta \Delta G_{\text{bind}}$ 值（图 1）。通过一系列计算机模拟实验，我们验证了 GearBind 的卓越性能，并展示了预训练的优势。GearBind 在多种评价指标上优于现有的最先进方法。借助对比预训练，GearBind 对突变体的亲和力变化非常敏感，并反映了氨基酸替换的趋势。随后，我们使用基于 GearBind 的流程对抗体进行亲和力成熟优化，并成功提高了抗体 CR3022 对 Delta 及 Omicron SARS-CoV-2 S 蛋白受体结合域（RBD）的结合亲和力，提升了最多 17 倍。这些结果进一步强调了几何深度学习和有效预训练在抗体亲和力成熟以及更广泛的大分子相互作用建模中的重要性。

2 Results

2.1 GearBind：一个可预训练的 $\Delta \Delta G_{\text{bind}}$ 预测器

GearBind 框架旨在通过多层次和多关系消息传递，从野生型和突变体结构中提取几何表示，以预测结合自由能变化 $\Delta \Delta G_{\text{bind}}$ 。GearBind 利用蛋白质复合物中的三种不同层次的信息，这些互补的见解包括：原子级信息，捕捉精确的空间和化学特征；边级信息，捕捉角度关系；以及残基级信息，强调蛋白质复合物内更广泛的上下文。整合这些独特但相互关联的信息层次，能够更全面地了解蛋白质复合物，从而增强模型的能力。

更正式地说，当蛋白质复合物结构输入到 GearBind 时，首先构建一个多关系界面原子图，以模拟复合物内的详细相互作用。定义的关系覆盖了序列邻近性（即在同一链上的原子）和空间邻近性（包括 k -最近邻和 r 半径内的邻居关系）。原子级表示通过在界面图上应用几何关系图神经网络（GearNet²²）来获得。在此基础上，还构建了一个线图，用于在原子图中，我们将每条边视为一个线节点，连接相邻的节点，并将角度信息编码为线边特征。然后通过在线图上执行消息传递来捕捉边级相互作用，这类似于 AlphaFold 的稀疏三角注意力机制²⁰。最后，在聚合每个残基的原子和边的

表示后，应用几何图注意力层来在残基之间传递消息。这种多层次的消息传递方案将多粒度的结构信息注入到学习到的表示中，从而非常适合 $\Delta\Delta G_{\text{bind}}$ 的预测任务。

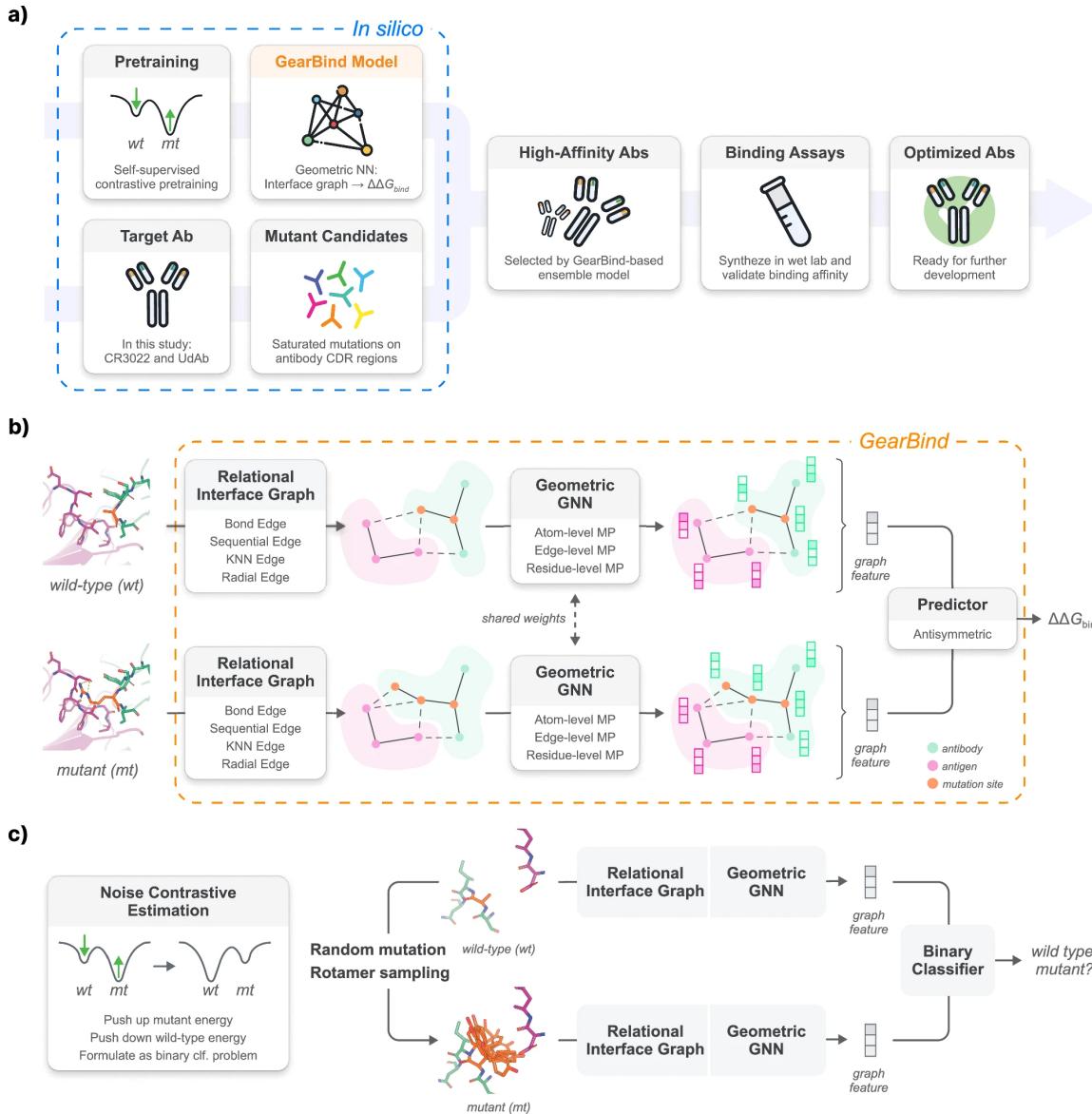


图 1: 基于 GearBind 的计算抗体亲和力成熟流程。**a. 流程概述**: 该流程采用几何神经编码器 GearBind, 首先在 CATH 数据集上进行自监督预训练, 随后在 SKEMPI v2 数据集上进行有监督学习。基于 GearBind 的集成模型用于在计算环境中对目标抗体进行亲和力成熟优化, 给定其结合结构和原生抗原。根据模型预测结果, 可找到数十种候选抗体, 其中改进的亲和力可以通过进一步的实验验证。图中使用的术语包括神经网络 (NN)、抗体 (Ab)、CDR (互补决定区)。图示资源来自 Flaticon.com。**b. GearBind 模型**: GearBind 采用共享图神经网络 GearNet, 对野生型和突变体复合物结构进行编码。对于每个结构, 构建关系界面图。接下来, GearNet 在图上执行多关系、多层次的消息传递, 以提取丰富的界面表示。基于提取到的复合物表示, 使用反对称预测器预测突变效应 $\Delta\Delta G_{\text{bind}}$ 。**c. 自监督预训练**: GearBind+P 利用大规模未标注的蛋白质结构进行自监督对比预训练。该模型旨在区分原生结构与侧链扭转角从构象库中随机采样的突变结构, 以探索蛋白质能量景观, 从而提高 $\Delta\Delta G_{\text{bind}}$ 的预测性能。

虽然 GearBind 可以在标注的 $\Delta\Delta G_{\text{bind}}$ 数据集上从头训练，但如果训练数据有限，可能会出现过拟合或泛化不佳的问题。为了解决这一问题，我们提出了一种自监督预训练任务，以利用 CATH^{22,23} 数据库中的大规模未标注蛋白质结构。在预训练阶段，编码器通过噪声对比估计²⁴ 来学习原生蛋白质结构的分布。具体而言，我们最大化原生 CATH 蛋白的概率（即降低其能量），同时最小化突变结构的概率（图 1c），这些突变结构是通过随机突变残基和从构象库中采样侧链扭转角生成的²⁵。区分原生、稳定的蛋白质结构和采样得到的突变结构，推动模型理解侧链相互作用模式，这对于蛋白质-蛋白质结合至关重要。通过这一过程，从丰富的单链蛋白质结构数据中提取的有意义知识可转移到蛋白质-蛋白质结合建模中。

2.2 在 SKEMPI 数据集上的交叉验证

我们在 SKEMPI v2.0 数据集上通过按复合物分割的五折交叉验证验证了 GearBind 的性能。我们的分割策略要求每个测试集与其对应的训练集不共享任何 PDB 复合物，这比按突变划分的策略更为现实，因为测试集中的野生型蛋白质复合物及其突变位点在训练中可能会出现。我们将 GearBind 和 GearBind+P（在 SKEMPI 上微调的预训练 GearBind）与最先进的基于物理的工具 FoldX⁹、Flex-ddG¹⁰ 以及深度学习方法 Bind-ddG⁸ 进行了比较。结果（表 1）显示，GearBind 通过其多关系图构建和多层次消息传递机制，在平均绝对误差（MAE）、均方根误差（RMSE）和 Pearson 相关系数（PearsonR）方面优于基线方法，并在 SpearmanR 上秒杀 FoldX。

预训练的 GearBind 进一步提高了性能，带来了 +5.4% 的 SpearmanR、+2.6% 的 PearsonR、-2.4% 的 MAE 和 -1.7% 的 RMSE。这突显了从大规模、未标注蛋白质结构数据中知识转移的有效性。

为了理解 GearBind 中关键架构设计的贡献，我们在 SKEMPI 数据集上对五个 GearBind 变体进行了基准测试。如图 2e 所示，这些 GearBind 变体在所有四项指标上的表现均不如完整的 GearBind。排除 GearBind 中的边级和残基级消息传递分别导致 SpearmanR 下降了 13% 和 3%，这突显了在特征提取过程中结合多层次信息的好处。

从界面图中排除侧链原子对性能的影响更为显著，导致 SpearmanR 下降 15%，这表明显式建模全原子结构的重要性。值得注意的是，将多关系界面图替换为基于 KNN 的图，导致 SpearmanR 严重下降 23%；而在多关系图上训练一个简单的 RGCG 模型，其性能与 Bind-ddG 持平（相比 GearBind 的 SpearmanR 下降 9%，但比 Bind-ddG 提高 2%）。这表明多关系图构建策略是 GearBind 的关键组成部分。

表 1：不同方法在 SKEMPI 数据集上的交叉验证性能 ($n = 5729$)

Model	MAE ↓	RMSE ↓	PearsonR ↑	SpearmanR ↑
FoldX ⁹	1.364 ± 0.134	2.027 ± 0.170	0.491 ± 0.007	0.526 ± 0.011
Flex-ddG ¹⁰	1.236 ± 0.101	1.849 ± 0.150	0.497 ± 0.034	0.484 ± 0.020
Bind-ddG ⁸	1.255 ± 0.096	1.759 ± 0.125	0.581 ± 0.037	0.443 ± 0.041
GearBind	<u>1.143 ± 0.088</u>	<u>1.639 ± 0.103</u>	<u>0.659 ± 0.030</u>	0.498 ± 0.033
GearBind+P	1.115 ± 0.072	1.611 ± 0.075	0.676 ± 0.041	0.525 ± 0.046
Ensemble	1.028 ± 0.080	1.503 ± 0.101	0.729 ± 0.016	0.643 ± 0.030

Note: MAE 表示平均绝对误差，RMSE 表示均方根误差。对于每个指标，我们报告了均值和标准误差。“+P”表示使用 CATH 数据集的几何预训练。在单个模型中，每个指标的最佳和次佳模型分别以粗体和下划线标出。

2.3 基于 GearBind 的计算机亲和力成熟集成

为了了解基准模型在 SKEMPI 上的行为，我们根据目标难度对 SKEMPI 数据集进行了分类，并绘制了 PearsonR 和 SpearmanR 评价。PDB 代码被分为“easy”（50 个以上相似数据点在训练集中），“medium”（1-50 个）和“hard”（0 个），基于数据点数量的高结构相似性（TM-score²⁶ > 0.8 ）。深度学习模型 Bind-ddG、GearBind 和 GearBind+P 在易

处理目标上表现优于基于物理的方法（如 FoldX 和 Flex-ddG），但在更高难度的目标上（图 2a, b），其泛化能力仍有改进空间。

我们还研究了不同突变对自由能变化的影响，将其分类为低($<0.5 \text{ kcal/mol}$)、中($0.5\text{--}2 \text{ kcal/mol}$)和高($>2 \text{ kcal/mol}$)绝对值区域。补充图 S9 显示，当结合自由能变化更大时，GearBind 表现更佳，表明其在高度复杂区域的优越性，尤其是 GearBind 的 PearsonR 达到 0.707，相比之下，FoldX 仅为 0.411，显示了 GearBind 在识别可能显著增强或破坏结合的突变方面的潜力。当 $|\Delta\Delta G_{\text{bind}}|$ 较小时，大多数方法与实验 $\Delta\Delta G_{\text{bind}}$ 的相关性很低，这可能暗示数据中的噪声或当前工具在建模较弱且更复杂的相互作用时的不足。

为了结合基于物理的方法和深度学习方法的优势，我们使用了所有基准方法的集成模型来执行计算亲和力成熟。集成模型的预测是 FoldX、Flex-ddG、GearBind、GearBind+P 和 Bind-ddG 的预测值的简单平均。结果表明，集成模型在所有四个评估指标上均优于单个模型（见表 1）。我们通过排除各模型并评估其在 SKEMPI 上的性能，来评估各模型对集成模型的贡献。结果（图 2f）显示，排除 GearBind 和 GearBind+P 对整体性能的影响最大。具体来说，对于 PearsonR 指标，排除 FoldX、Flex-ddG 和 Bind-ddG 单独带来的影响微乎其微（小于 0.01），但移除 GearBind 会导致显著下降（超过 0.08）。此外，虽然 FoldX 在单独使用时并非最佳模型，但将其从集成模型中移除会导致 SpearmanR 大幅下降。这表明 FoldX 在补充深度学习模型方面发挥了重要作用，从而构建了一个强大而准确的集成模型。事实上，将 GearBind、GearBind+P 和 FoldX 结合在一起的性能可与 5 模型集成的表现相媲美（参见补充图 S26）。

2.4 在 HER2 结合物测试集上的评估

我们使用在 SKEMPI 上训练的模型对 HER2 结合物测试集进行了性能测试，该测试集来自 Shahnehsaazadeh 等人²⁷。该数据集包含 419 个 HER2 结合物的高质量结合亲和力数据，使用表面等离子体共振（SPR）测得，并设计了全新 CDR 环。抗体数据集是 Trastuzumab 的变体，具有较高的编辑距离（平均为 7.6），这使得它们在低编辑距离数据上训练的 $\Delta\Delta G_{\text{bind}}$ 预测模型中具有挑战性。在基准方法中（Flex-ddG 未进行基准测试，因为其耗时过长），GearBind+P 达到了最佳的 PearsonR 和 SpearmanR（图 2c）。然后，我们对所有基准模型的预测进行了平均，以形成一个集成模型，并依次排除各模型以测量其对性能的影响。同样地，排除 GearBind+P 对 PearsonR 的影响最大，而排除 FoldX 对 SpearmanR 的影响最大，而 GearBind+P 紧随其后（图 2d）。

2.5 CR3022 和抗-5T4 UdAb 的亲和力成熟

为了验证我们方法的有效性，选择了两种抗体 CR3022 和抗-5T4 UdAb 作为亲和力成熟的研究对象。CR3022 抗体最初从一名康复的 SARS 患者中分离出来²⁸，随后被鉴定为可以结合 SARS-CoV-2^{29,30}。与此同时，针对肿瘤胎抗原 5T4 的 UdAb 以其卓越的稳定性而著称³¹。请注意，这两种抗体的形式各异，靶向不同的抗原。这些抗原在 SKEMPI 数据集中只有一个结构上相似的蛋白质链 ($\text{TM-score} > 0.8$)，并且结合位点不同，使其成为我们流程的挑战性目标（见补充表 S9 和 S10 及补充图 SI.4 和 SI.5）。

2.6 CR3022 的亲和力成熟

对于 CR3022，根据集成模型对野生型、BA.1.1 和 BA.4 SARS-CoV-2 毒株 RBD 结合亲和力变化的预测，我们在第一轮实验证中选择了 12 个突变体。我们注意到，野生型和 Delta RBD 在与 CR3022 结合界面上共享相同的氨基酸。在 ELISA 预实验中，我们测试了这些突变体在 100 nM 抗原浓度下与 SARS-CoV-2 Delta 株 RBD 的结合情况。12 个候选突变体中有 9 个相比野生型 CR3022 显示出改善的结合（补充图 Fig. S16a）。在进一步验证中，将 RBD 浓度降低到 10 nM 后，这 9 个候选突变体的 EC_{50} 值均低于野生型 CR3022（图 3a, b）。基于这些结果，我们综合了表现优异的 CR3022 突变体，并设计合成了 8 个带有双重或三重突变的候选体作为第二轮设计。8 个多点突变体中有 7 个在 Delta RBD 上表现出增强的结合力，其 ELISA EC_{50} 值相比野生型降低了 1.8 至 3.4 倍（图 3c, d）。在 Omicron 刺突蛋白上，

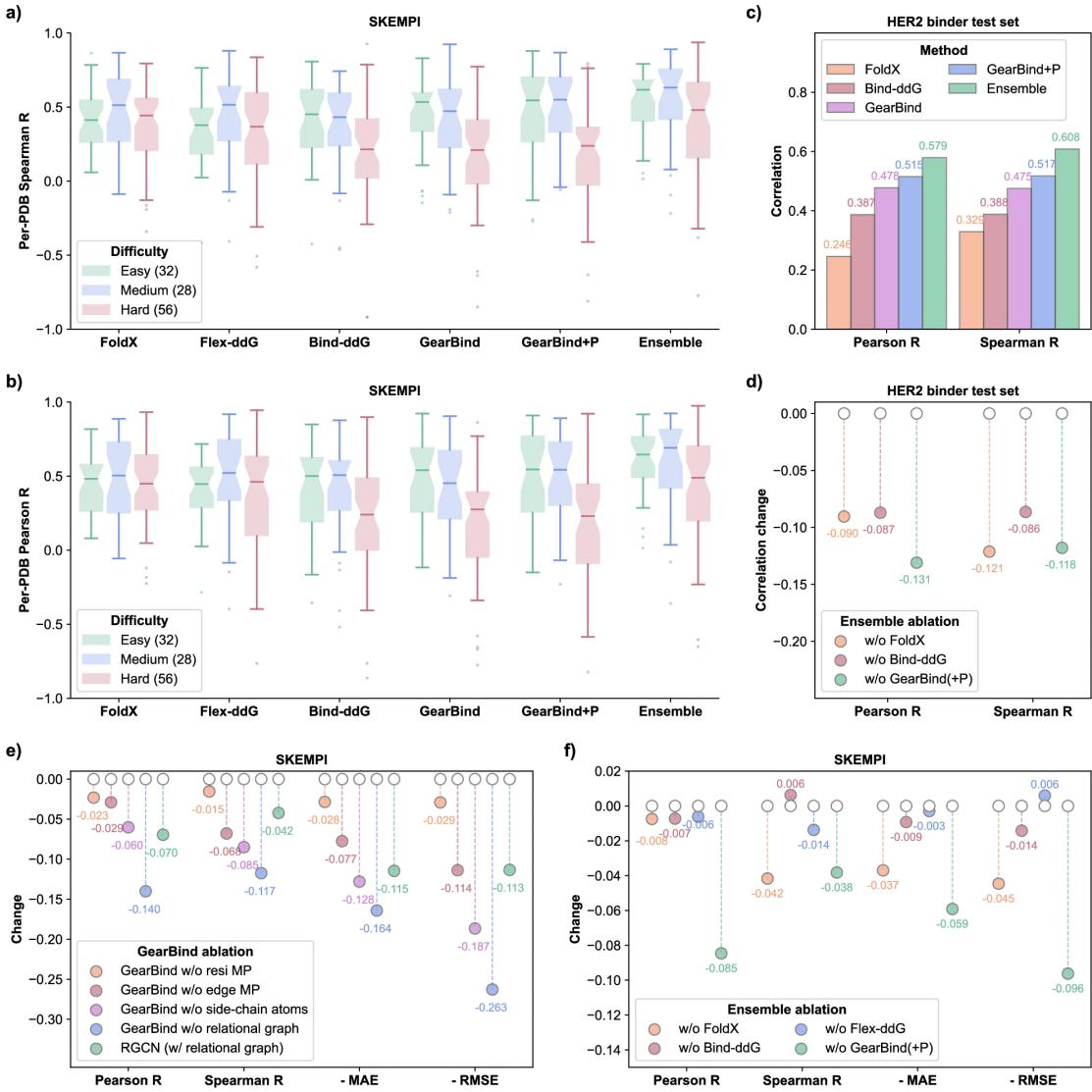


图 2: 在 SKEMPI 和 HER2 结合物测试集上的计算评估。a, b. 对 SKEMPI 子集的 Per-PDB Spearman(a)和 Pearson(b) 相关性进行比较分析, 涵盖不同难度级别的目标与各种模型预测和实验数据之间的相关性。SKEMPI 数据集中的 PDB 代码根据训练集中具有高结构相似度 (TM-score > 0.8) 的数据点数量分类为 “easy” (50+)、“medium” (1–50) 和 “hard” (0) 目标。每个难度级别的 PDB 代码数量在图例中标注。箱线图表示四分位数范围 (25th 至 75th 百分位数), 中位数用实线表示, 异常值定义为 1× 四分位数范围之外的数据点。

c. HER2 结合物测试集 ($n = 419$) 的基准测试结果, 展示了各种模型的 Pearson 和 Spearman 相关性。这些深度学习模型在 SKEMPI 上训练。d. 在 HER2 结合物测试集上排除集成模型 (FoldX + Bind-ddG + GearBind(+P)) 中各模型后的性能变化。e. 在 SKEMPI 数据集 ($n = 5729$) 上的 GearBind 架构设计变动对性能的影响。f. 在 SKEMPI 上排除集成模型 (FoldX + Flex-ddG + Bind-ddG + GearBind(+P)) 中各模型后的性能变化。

这些突变体表现出 7.6 至 17.0 倍的结合力提升, EC_{50} 达到亚纳摩尔水平, 其中 SH100D+SH103Y+SL33R 三重突变体的表现最佳, EC_{50} 最低为 0.06 nM (图 3e, f)。

我们接下来测试了新设计的突变体在 SARS-CoV RBD 上的结合情况, 以验证针对 SARS-CoV-2 RBD 的 CR3022 优化是否会导致其对原始靶标的结合能力发生显著变化。8 个突变体中有 7 个在 SARS-CoV RBD 上的 ELISA EC_{50} 值

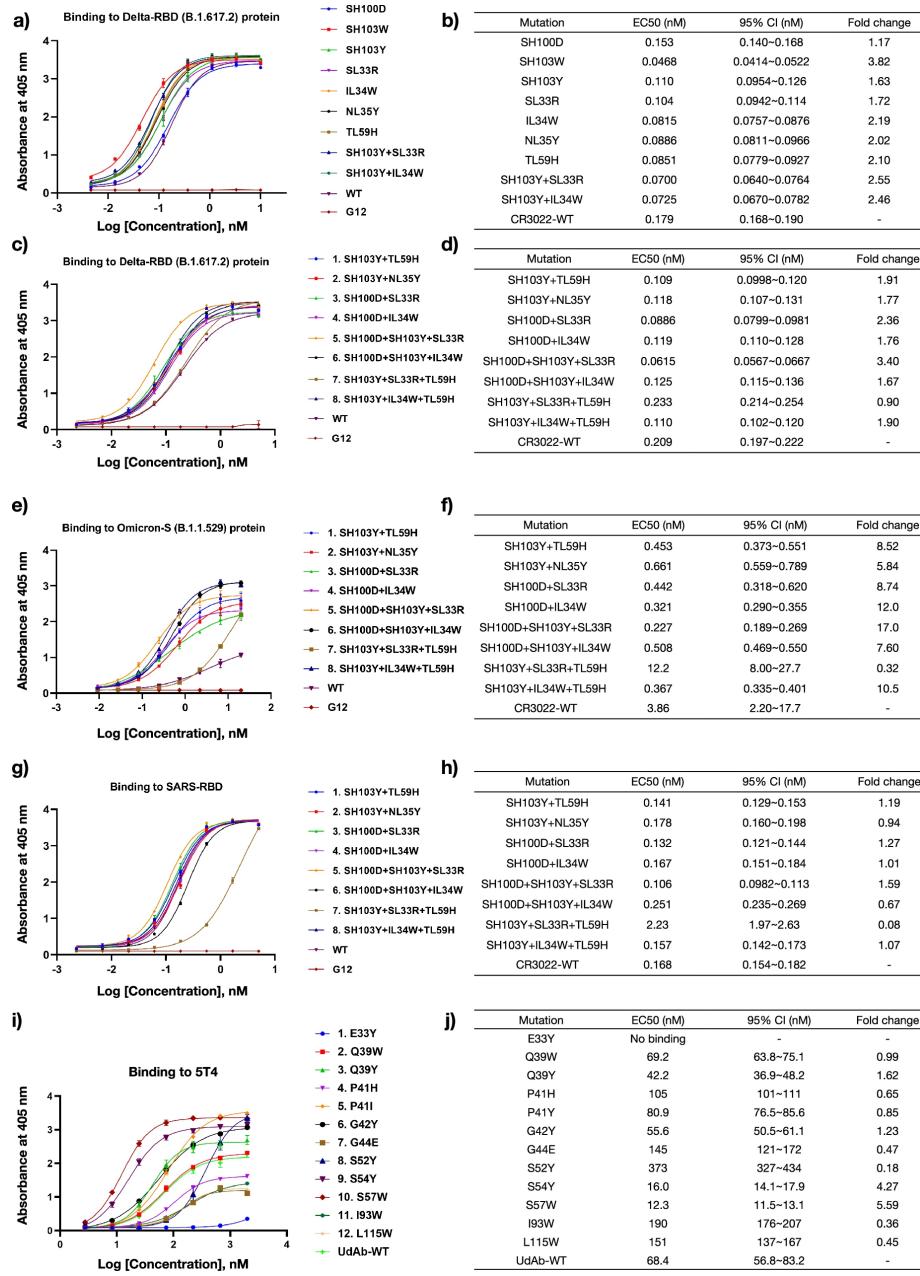


图 3: 基于 GearBind 流程设计的 CR3022 和抗-5T4 UdAb 突变体的 ELISA 结合实验结果。左侧面板 (a, c, e, g, i) 显示了浓度-响应曲线及从 ELISA 测定的 EC₅₀ 值，曲线中间点表示平均吸光度，误差条表示三个技术重复的标准偏差。右侧面板 (b, d, f, h, j) 显示了拟合的 EC₅₀ 值、其 95% 置信区间，以及通过 EC₅₀^(wt) / EC₅₀^(mt) 计算的结合倍数变化。

测试系统包括：第一轮设计的 CR3022 突变体与 Delta RBD 的结合 (a, b)；第二轮设计的 CR3022 突变体与 Delta RBD (c, d) 和 Omicron S 蛋白 (e, f) 的结合；CR3022 突变体与 SARS-RBD 的结合 (g, h)；以及抗-5T4 UdAb 的单点突变体与 5T4 的结合 (i, j)。

显示出显著变化 (图 3g, h)。综上所述，上述结果展示了我们基于 GearBind 流程在 CR3022 抗体亲和力优化中的成功应用。

2.7 抗-5T4 UdAb 的亲和力成熟

为了验证我们方法的通用性，我们将实验验证扩展至抗-5T4 UdAb。我们使用基于 GearBind 的流程开发了 12 个单点突变体，并通过 ELISA 验证其与 5T4 的结合情况。表现最佳的 UdAb 突变体 S57W 的 EC₅₀ 降低了 5.6 倍。这表明，我们的方法在增强不同形式和不同靶标抗体的亲和力方面具有潜力，使其成为一个有前景的抗体治疗开发工具。

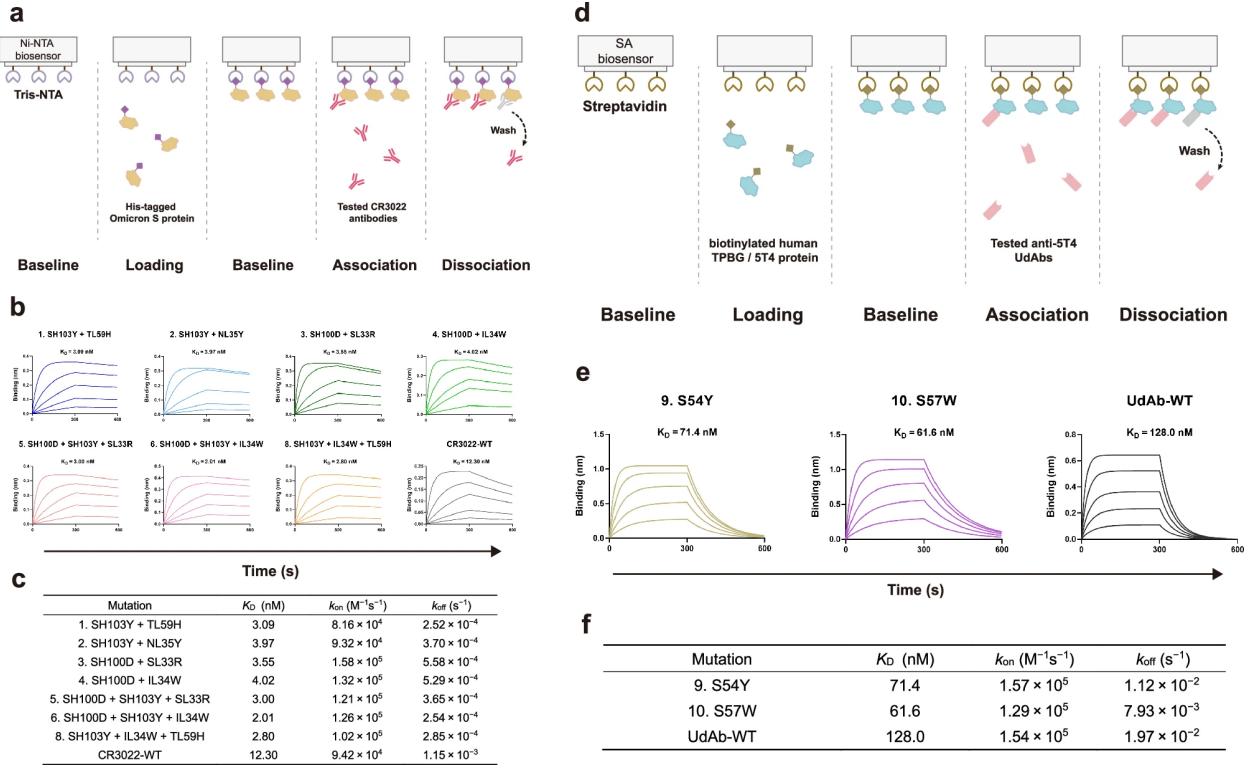


图 4: CR3022 和抗-5T4 UdAb 候选抗体的生物层干涉 (BLI) 结合分析。a. 实验方案示意图，使用 Ni-NTA 和 SA 生物传感器。b, e. CR3022 (b) 和抗-5T4 UdAb (e) 候选抗体的结合动力学。测试抗原浓度为 300 nM。数据通过全局 1:1 结合模型拟合以确定 K_D (平衡解离常数)，每个图上的注释值由 R² > 98% 的置信水平确定。c, f. 测定突变型和野生型 CR3022 抗体 (c) 以及抗-5T4 UdAb (f) 的 K_{on} (结合速率常数)、K_{off} (解离速率常数) 和 K_D 值。

我们进一步使用生物层干涉技术 (Bio-layer Interferometry, BLI) 对亲和力成熟的 CR3022 抗体和抗-5T4 UdAb 进行了验证，以更准确地评估其结合亲和力 (图 4)。测试的 7 个 CR3022 突变体在 Omicron 刺突蛋白上显示出 3.1 至 6.1 倍的结合亲和力提升，其中表现最佳的突变体为 SH100D+SH103Y+IL34W。ELISA 实验中表现最佳的 SH100D+SH103Y+SL33R 三重突变体在 BLI 测试中显示出 4.1 倍的结合亲和力提升。

对于抗-5T4 UdAb，测试的 S54Y 和 S57W 两个突变体分别表现出 1.8 倍和 2.1 倍的亲和力提升。总体而言，BLI 测量结果与 ELISA 的趋势一致，进一步确认了我们优化流程的有效性。与 ELISA 结合实验结果一致，表明基于 GearBind 流程设计的 CR3022 和 UdAb 变体的结合亲和力得到了提升。

2.8 优化抗体的结构特征

理解深度学习设计的突变体的序列-结构-功能关系不仅有助于提高我们的模型，还可以帮助解释其生物学意义。为了探索增强抗体-抗原结合的潜在机制，我们对野生型和突变抗体进行了分子动力学模拟和结构分析，重点关注 ELISA

EC₅₀ 最低的突变体，即 CR3022 的 SH100D+SH103Y+SL33R 三重突变体，以及抗-5T4 UdAb 的 S57W 突变体。我们在室温下对每个系统及其相应的野生型对照进行了 1 s 的全原子分子动力学模拟（详细方法见“Methods”部分）。

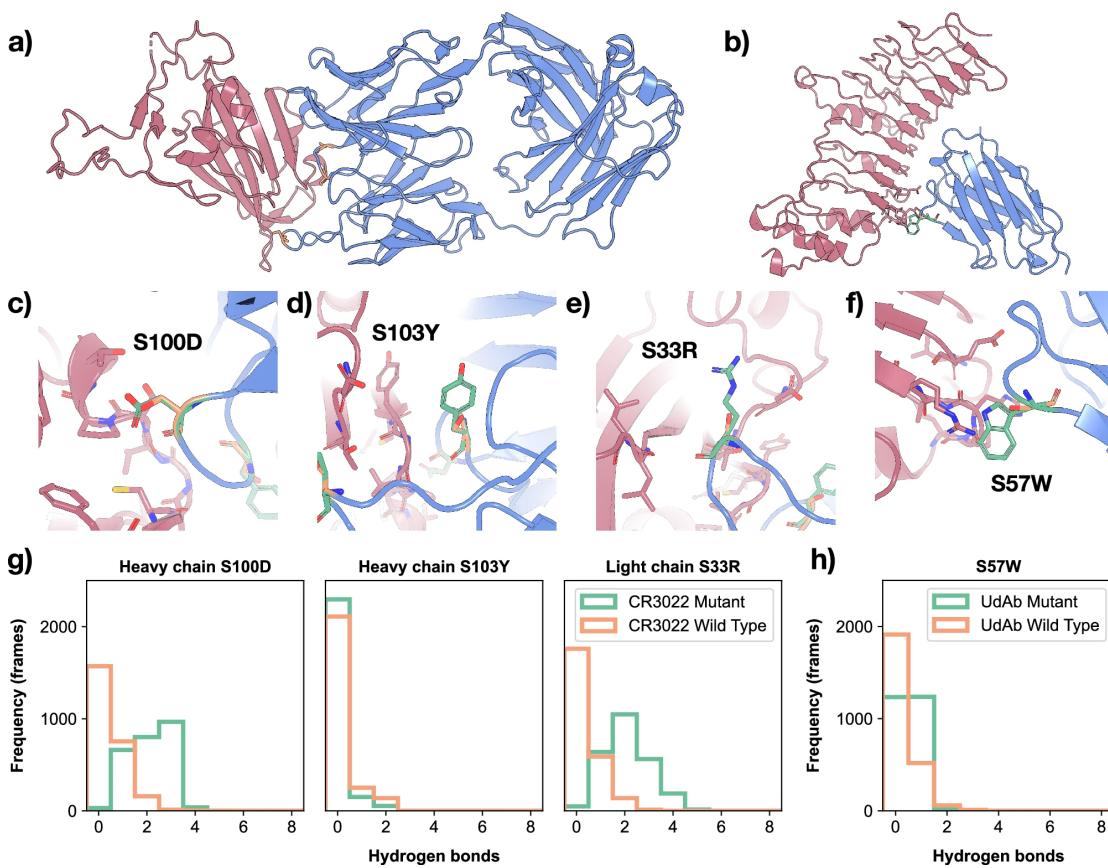


图 5 | 优化的 CR3022 和抗-5T4 UdAbs 的结构分析。**a** 用于亲和力成熟的抗体 CR3022 和 SARS-CoV-2 RBD 的复合物结构。目标抗原以红色显示，抗体以蓝色显示。突变位点 S100D、S103Y（重链）和 S33R（轻链）用橙色标出。**b** 单域抗体 UdAb 及其靶向的肿瘤抗原 5T4 的复合物结构。突变位点 S57W 用橙色标记。**c-e** CR3022 三重突变体的三个突变位点：S100D (**c**)、S103Y (**d**) 和 S33R (**e**)。**f** 单点突变体 UdAb 中的 S57W 突变位点。**g** 显示 CR3022 突变体与 RBD 复合物中突变位点与目标抗原之间的氢键数量，基于分子动力学模拟结果。**h** 显示 UdAb 与 5T4 复合物中突变位点与目标抗原之间的氢键数量，基于分子动力学模拟结果。**g, h** 氢键分布图：野生型以橙色显示，突变体以绿色显示。

基于模拟结果，在研究的四个突变体中，有三个展示了增加的氢键数量，这有助于解释其增强的结合亲和力。目标突变体包括 CR3022 的 SH100D 和 SL33R，以及抗-5T4 UdAb 的 S57W（图 5g, h）。这四个突变位点稳定了抗体结构，通过减少相应抗体中的 C_α 原子的波动（补充图 S20）。尽管 CR3022 重链中的 SI03Y 突变未增加极性接触，但通过排除更多溶剂，提高了抗体-抗原界面上的疏水相互作用，这可能是由于酪氨酸体积更大所致。总之，我们的流程设计的突变体通过形成新的相互作用，显著提高了结合亲和力，同时也观察到了稳定的结合残基和突变结构中的结构性质变化。

有趣的是，GearBind 预测的贡献图（补充图 S24）为理解这些设计突变体的潜在接触位点提供了进一步的见解。大多数贡献与分子动力学模拟的结果一致。例如，SI03Y 的潜在疏水作用也在贡献图中得到了验证，这与蛋白质结构上的推断结果高度吻合（补充图 S24c）。

3 讨论

本研究提出了一种基于可预训练几何图神经网络 GearBind 的计算抗体亲和力成熟流程，并成功将其应用于两个不同的抗体优化项目，即 CR3022 和抗-5T4 UdAb。通过大量的计算实验，我们评估了模型的性能，了解其优势和局限性。GearBind 模型的技术亮点可总结如下：1. 在图构建阶段，采用多关系图来捕捉界面上的全原子信息。定义的关系包括序列邻近性和空间邻近性。2. 在特征提取阶段，采用多层次消息传递机制，以全面捕捉蛋白质相互作用的多粒度信息。3. 提出了一种基于对比学习的预训练算法，利用 CATH 数据集中丰富的未标注单链蛋白质结构，提炼侧链扭转角的知识，从而进一步提升模型性能。

我们通过 GearBind 流程在两个真实抗体亲和力成熟项目上进行了验证。ELISA 结合实验显示，CR3022 突变体在 Delta RBD 和 Omicron 刺突蛋白上的结合力得到了显著提升。尤其是，10 个 CR3022 单点突变体中有 7 个，以及 10 个多点突变体中有 9 个，在 Delta RBD 和 Omicron S 蛋白上显示出显著的结合增强，ELISA EC₅₀ 最高降低 3.8-6.1 倍，Omicron S 蛋白上最高提升 17 倍。

此外，我们的流程对 12 个抗-5T4 UdAb 突变体中的最佳突变体 S57W 也取得了显著效果，ELISA EC₅₀ 降低了 5.6 倍，BLI 测得的 K_D 提升 2.1 倍。总之，GearBind 证明了其在设计结合亲和力增强抗体方面的有效性。基于 GearBind 流程设计的突变体通过创建新的相互作用或强化现有接触点，提高了抗体亲和力。这为将 GearBind 应用于抗体优化提供了坚实的理论基础，同时也展示了几何深度学习在抗体工程中的广泛前景。

负的 SpearmanR 值表明，大规模蛋白质语言模型的零样本预测在为蛋白质复合物的亲和力排序方面并不是可靠的方法³³。这一结果是合理的，因为语言模型所模拟的肽序列“适应度”并不一定意味着对所有其他生物分子的强结合。例如，提高 SARS-CoV-2 刺突蛋白的适应度可能会降低其对现有中和抗体的结合亲和力。另一个论点是，结构信息在构建准确且可靠的蛋白质-蛋白质相互作用算法中起着关键作用³⁴。

3.1 GearBind 的未来应用

展望未来，GearBind 的潜在应用超越了蛋白质-蛋白质结合优化。该模型可以轻松适用于蛋白质-肽和蛋白质-配体对接挑战，从而为其在小分子抑制剂和酶设计中的应用打开了可能性。

尽管前景光明，我们仍然承认当前方法中的一些局限性，并探讨未来的改进方向。首先，基于结构的 $\Delta\Delta G_{bind}$ 预测的前提是准确的复合物结构，而这在大多数抗体-抗原对中并不容易获得。为了解决这一问题，可以使用同源建模工具³⁵从模板结构中构建复合物结构。例如，我们通过同源建模构建了 CR3022 与 Omicron RBD 的复合物。更激进的方法是直接从序列预测复合物结构。随着多肽结构预测方法的不断进步³⁶，它们有望在未来成为基于结构的亲和力成熟的可靠起点。

其次，依赖于外部工具生成突变体结构增加了时间成本，并将我们的操作空间限制在突变替换上。未来的研究应聚焦于训练端到端模型，直接预测 $\Delta\Delta G_{bind}$ ，并能够处理氨基酸插入和删除。此外，我们需要更好的预训练策略和架构设计，以提升深度学习模型的泛化能力，使其对从未见过的蛋白质更加稳健。总的来说，我们的工作朝着构建可靠、强大且高效的计算亲和力成熟流程迈出了坚实的一步，这将为研究和药物开发带来巨大的应用机会。

4 Methods

4.1 数据集

SKEMPI。我们使用了 SKEMPI v2¹⁹ 数据集进行训练和验证。该数据集包含 7,085 条 $\Delta\Delta G_{bind}$ 测量数据，涉及 348 个复合物。我们按照参考文献^[6, 11] 的方法进行数据预处理，去除测量值不明确或多个测量值之间 $\Delta\Delta G_{bind}$ 变异较大的数据，最终得到 5,747 个独立突变体及其 $\Delta\Delta G_{bind}$ 测量值，覆盖 340 个复合物。对于每个突变体，我们基于野生

型 PDB 晶体结构，使用 FoldX 4^[19] 对突变体结构进行采样。使用 PDBFixer v1.8^[24] 修复 PDB 结构，以防原始结构无法被 FoldX 处理。对于无法被 torchdrug^[25] 读取的突变体结构，出于公平性考虑予以删除。

处理后的数据集被分为五个子集，每个子集包含大致相同数量的 PDB 复合物。我们进行了五折交叉验证，模型在其他四折上训练，并在特定折上的每个数据点进行推断。最终结果基于处理后的数据集中所有 4,060 个单点突变体报告。

4.2 预训练数据集

为了进行预训练，我们从蛋白质数据银行（PDB）^[26] 中检索了 123,505 个实验测定的蛋白质结构。为确保数据质量，我们仅保留了分辨率在 0.0 到 2.5 Å 之间的结构，排除了分辨率较低的结构。预训练中使用了单链和多链蛋白质。对于多链蛋白质，每次建模迭代时，我们随机选择两条链以捕捉突变时的相互作用。

4.2.1 为蛋白质复合物结构构建关系图

对于蛋白质-蛋白质复合物，我们为其界面构建多关系图，并丢弃所有其他原子。如果一个残基与结合伙伴的最近残基的欧氏距离不超过 6Å，则将其视为界面上的残基。界面上的每个原子被视为图中的节点。我们为这些原子之间添加三种类型的边，以表示不同的相互作用：1. 对于序列距离不超过 3 的两个原子，添加顺序边（sequential edge），其类型由在蛋白质序列中的相对位置决定。2. 对于具有空间距离小于 6Å 的两个原子，添加空间边（spatial edge）。3. 对于根据化学键直接连接的两个原子，添加键边（bond edge）。

4.3 CATH 预训练数据集

在预训练阶段，我们使用了 CATH v4.3.0 域的非冗余子集，其中包含 30,948 个实验蛋白质结构，这些结构的序列同一性低于 40%。此外，我们去除了长度超过 2,000 个氨基酸的蛋白质，以提高效率。在预训练过程中，我们随机截断长序列至 150 个氨基酸以内，以提高处理效率。需要注意的是，目前我们的预训练仅限于使用单链蛋白质。通过单链预训练学习到的信息可以迁移到下游的蛋白质复合物任务中，我们发现这种方法已经足够带来显著的性能提升。

4.4 HER2 结合物测试集

HER2 结合物测试集来自参考文献^[27]。原始数据包括 758 个结合物和 1,097 个非结合物的 SPR 数据。由于所有基准方法仅支持氨基酸替换，我们过滤掉了与野生型抗体（Trastuzumab）长度不同的结合物，最终保留了 419 个 Trastuzumab 突变体。 $\Delta\Delta G_{\text{bind}}$ 值基于 SPR 测得的结合亲和力，通过以下公式计算：

$$\Delta\Delta G_{\text{bind}} = -RT \ln \left(\frac{K_D^{(\text{mt})}}{K_D^{(\text{wt})}} \right)$$

需要注意的是，我们仅使用此数据集来评估基于物理的方法（FoldX、Flex-ddG）和深度学习模型（Bind-ddG、GearBind、GearBind+P），这些模型均在 SKEMPI 数据集上进行了训练。

4.5 GearBind 实现

给定一对野生型和突变体结构，GearBind 通过几何编码器和多关系图来预测结合自由能变化 $\Delta\Delta G_{\text{bind}}$ ，并通过自监督预训练进一步增强。该神经网络之所以被称为几何网络，是因为它考虑了实体（即图中的节点）之间的空间关系。在接下来的部分中，我们将讨论多关系图的构建、多层次消息传递和预训练方法。

4.5.1 为蛋白质复合物结构构建关系图

给定一个蛋白质-蛋白质复合物，我们为其界面构建多关系图，并丢弃所有非界面原子。如果一个残基与其结合伙伴中最近的残基的欧氏距离不超过 6 Å，则将其视为界面残基。界面上的每个原子被视为图中的一个节点。我们为这些原子之间添加三种类型的边，以表示不同的相互作用：

1. 对于序列距离小于 3 的两个原子，添加顺序边 (sequential edge)，其类型由它们在蛋白质序列中的相对位置决定。
2. 对于空间距离小于 5 Å 的两个原子，添加径向边 (radial edge)。
3. 每个原子还与其 10 个最近邻居连接，以确保图的连通性。

我们忽略了连接蛋白质序列中相邻原子的空间边，因为这些边对捕捉界面相互作用的意义不大。构建的关系图用 $(\mathcal{V}, \mathcal{E}, \mathcal{R})$ 表示，其中 \mathcal{V} 、 \mathcal{E} 和 \mathcal{R} 分别表示节点集、边集和关系类型集。我们使用三元组 (i, j, r) 来表示节点 i 和 j 之间的边，以及边的类型 r 。对于每个节点，我们使用残基类型和原子类型的独热向量作为其节点特征。

4.6 通过多层次消息传递构建几何编码器

在构建好的界面图之上，我们进行多层次消息传递，以建模连接的原子、边和残基之间的相互作用。我们用 $a_i^{(l)}$ 和 $e_{(j,i,r)}^{(l)}$ 分别表示第 l 层节点 i 和边 (j, i, r) 的表示，其中 $a_i^{(0)}$ 表示原子 i 的初始节点特征， $e_{(j,i,r)}^{(0)}$ 表示边 (j, i, r) 的初始边特征。随后，通过以下步骤更新表示：

$$a_i^{(l)} \leftarrow \text{AtomMP} \left(a_i^{(l-1)} \right), \quad (1)$$

$$e_{(j,i,r)}^{(l)} \leftarrow \text{EdgeMP} \left(e_{(j,i,r)}^{(l-1)} \right), \quad (2)$$

$$a_i^{(l)} \leftarrow a_i^{(l)} + \text{AGGR} \left(e_{(j,i,r)}^{(l)} \right), \quad (3)$$

$$a_{C\alpha(i)}^{(l)} \leftarrow a_{C\alpha(i)}^{(l)} + \text{ResAttn} \left(a_{C\alpha(i)}^{(l)} \right). \quad (4)$$

首先，我们在原子图上执行原子级消息传递 (AtomMP)。然后，构建一条线图 (line graph)，用于在边之间进行消息传递 (EdgeMP)，从而学习原子对之间的有效表示。边的表示通过聚合函数 (AGGR) 用于更新原子表示。最后，我们选取 Cα 原子的表示作为残基表示，并执行残基级注意力机制 (ResAttn)，这可以看作是在完全连接的图上执行的一种特殊的消息传递。

在接下来的段落中，我们将详细讨论这些组件。

4.6.1 原子级消息传递

遵循 GearNet^[22]，我们使用关系图卷积神经网络 (RGCN)^[38] 在原子之间传递消息。在消息传递步骤中，每个节点从其邻居聚合消息以更新其表示。消息是通过特定关系 (边类型) 的线性层输出，并应用于邻居的表示。形式上，消息传递步骤定义为：

$$\text{AtomMP} \left(a_i^{(l-1)} \right) = a_i^{(l-1)} + \sigma \left(\text{BN} \left(\sum_{r \in \mathcal{R}} w_r^{(a)} \sum_{(j,i,r) \in \mathcal{E}} a_j^{(l-1)} \right) \right),$$

其中， σ 是激活函数，BN 是批归一化， $w_r^{(a)}$ 是与边类型 r 相关的权重矩阵， \mathcal{R} 是所有边的关系类型集， \mathcal{E} 是边集。

其中， $\text{BN}(\cdot)$ 表示批归一化， $\sigma(\cdot)$ 是 ReLU 激活函数。

4.6.2 边级消息传递和聚合

仅建模顺序邻近性或空间距离不足以捕捉复杂的蛋白质-蛋白质相互作用 (PPI)，这些相互作用对结合有重要贡献。多项研究已表明，利用边级消息传递来整合角度信息的优势^[20, 22, 39]。在此，我们构建一条线图 (line graph)^[40]，即

在原子级图之上的关系图。只有当两个边共享一个公共节点时，它们才会连接。关系（或边类型）定义为原子级边对之间的角度，并离散化为八个区间。我们用 $(\mathcal{V}', \mathcal{E}', \mathcal{R}')$ 表示构建的线图。接下来，在该线图上使用关系消息传递：

$$\text{EdgeMP}\left(e_x^{(l-1)}\right) = \sigma \left(\text{BN} \left(\sum_{r' \in \mathcal{R}'} w_{r'}^{(e)} \sum_{(y, x, r') \in \mathcal{E}'} e_y^{(l-1)} \right) \right),$$

其中， $e_x^{(l-1)}$ 是第 $(l-1)$ 层的边表示， $w_{r'}^{(e)}$ 是与边类型 r' 相关的权重矩阵， \mathcal{R}' 是线图中的边关系类型集， \mathcal{E}' 是线图的边集。通过这种方法，我们能够捕捉到与原子对之间角度信息相关的复杂相互作用，从而改进蛋白质结合亲和力的建模。

其中， x 和 y 表示原始图中的边元组，用于简化表示。

4.6.3 边级聚合 (Aggregation)

一旦更新了边的表示，我们会将其聚合到其端点节点。这些表示被输入到线性层中，并在 AtomMP 中与特定边类型的核矩阵 $w_r^{(a)}$ 相乘：

$$\text{AGGR}\left(e_{(j,i,r)}^{(l)}\right) = \sigma \left(\text{BN} \left(\sum_{r \in \mathcal{R}} w_r^{(a)} \sum_{(j,i,r) \in \mathcal{E}} \text{Linear}\left(e_{(j,i,r)}^{(l)}\right) \right) \right),$$

这将用于更新公式 (3) 中的原子 i 的表示。

4.6.4 残基级消息传递

受限于计算复杂度，原子级和边级消息传递仅考虑稀疏的相互作用，而忽略了所有残基对之间的全局相互作用。通过在残基级别建模界面的粗粒度视图，我们能够在所有残基对之间执行消息传递。为此，我们设计了一种几何图注意力机制 (geometric graph attention mechanism)，其输入为残基的 α 碳 ($C\alpha$) 的表示，并通过公式 (4) 输出更新的残基表示。

在此，我们遵循典型的自注意力定义，使用查询 (query) 和键 (key) 向量计算注意力得分，并将该概率应用于值向量以获取残基表示 $r_i^{(l)}$ ：

$$\begin{aligned} \alpha_{ij} &= \text{Softmax}_j \left(\frac{1}{\sqrt{d}} \text{Linear}_q \left(a_{C\alpha(i)}^{(l)} \right) \cdot \text{Linear}_k \left(a_{C\alpha(j)}^{(l)} \right) \right), \\ r_i^{(l)} &= \sum_j \alpha_{ij} \cdot \text{Linear}_v \left(a_{C\alpha(j)}^{(l)} \right), \end{aligned}$$

其中 d 为表示 $a_{C\alpha(i)}^{(l)}$ 的隐藏维度，Softmax 函数在所有 j 上取值。

除了传统的自注意力，我们还在注意力机制中加入了几何信息，使其在全局复合物结构的旋转-平移变换下保持不变。因此，我们为每个残基构建一个局部坐标系，其坐标基于其氮 (N)、碳 (C) 和 α 碳 ($C\alpha$) 原子的坐标：

$$\mathbf{v}_{i1} = \mathbf{x}_{N(i)} - \mathbf{x}_{C\alpha(i)}, \quad \mathbf{v}_{i2} = \mathbf{x}_{C(i)} - \mathbf{x}_{C\alpha(i)},$$

$$\mathbf{u}_{i1}, \mathbf{u}_{i2} = \text{GramSchmidt}(\mathbf{v}_{i1}, \mathbf{v}_{i2}),$$

$$\mathbf{R}_i = \begin{bmatrix} \mathbf{u}_{i1} & \mathbf{u}_{i2} & \frac{\mathbf{u}_{i1} \times \mathbf{u}_{i2}}{\|\mathbf{u}_{i1} \times \mathbf{u}_{i2}\|} \end{bmatrix},$$

其中， \mathbf{x} 表示原子的坐标， $\text{GramSchmidt}(\cdot)$ 表示用于正交化的 Gram-Schmidt 过程。接下来，几何注意力机制被设计用来建模所有残基 j 的 β 碳 ($C\beta$) 在残基 i 的局部坐标系中的相对位置：

$$\mathbf{p}_i^{(l)} = \sum_j \alpha_{ij} \mathbf{R}_i^\top (\mathbf{x}_{C\alpha(i)} - \mathbf{x}_{C\beta(j)}),$$

其中， $\mathbf{p}_i^{(l)}$ 是残基 i 的空间表示。当复合物结构旋转时，坐标系 \mathbf{R}_i 和相对位置 $\mathbf{x}_{\text{C}\alpha(i)} - \mathbf{x}_{\text{C}\beta(j)}$ 会相应地旋转，从而抵消其影响，这确保了我们模型的旋转不变性。

最终输出是残基表示 $r_i^{(l)}$ 和空间表示 $\mathbf{p}_i^{(l)}$ 的拼接：

$$\text{ResAttn} \left(a_{\text{C}\alpha(i)}^{(l)} \right) = \text{Concat} \left(r_i^{(l)}, \mathbf{p}_i^{(l)} \right).$$

在为每个原子获取表示后，我们对所有 α 碳原子 $a_{\text{C}\alpha}$ 的表示应用均值池化层，以获得蛋白质的整体表示 \mathbf{h} 。接着，应用一个反对称预测头（antisymmetric prediction head），以确保正向和反向突变（即残基突变和其逆突变）具有相同的预测效果。预测的 $\Delta\Delta\tilde{G}_{\text{bind}}$ 值应为正向和反向突变产生相反的预测结果：

$$\Delta\Delta\tilde{G}_{\text{bind}} = \text{MLP}(h^{(\text{wt})}, h^{(\text{mt})}) - \text{MLP}(h^{(\text{mt})}, h^{(\text{wt})}), \quad (5)$$

其中， $h^{(\text{wt})}$ 和 $h^{(\text{mt})}$ 分别表示野生型和突变体复合物的表示， $\Delta\Delta\tilde{G}_{\text{bind}}$ 是通过 GearBind 模型预测的结合自由能变化。

4.7 通过噪声对比估计建模蛋白质的能量景观

由于结合自由能变化数据的样本相对较少，利用大量蛋白质结构数据对 GearBind 进行预训练将是有益的。我们的预训练方法的核心思想是建模天然蛋白质结构的分布，以识别可能导致非天然结构的有害突变。用 x 表示蛋白质结构，其分布可以通过 Boltzmann 分布建模：

$$p(x; \boldsymbol{\theta}) = \frac{\exp(-E(x; \boldsymbol{\theta}))}{A(\boldsymbol{\theta})}, \quad A(\boldsymbol{\theta}) = \int \exp(-E(x; \boldsymbol{\theta})) dx, \quad (6)$$

其中， $\boldsymbol{\theta}$ 表示编码器中的可学习参数， $E(x; \boldsymbol{\theta})$ 表示蛋白质 x 的能量函数， $A(\boldsymbol{\theta})$ 是分区函数，用于归一化分布。能量函数 $E(x; \boldsymbol{\theta})$ 通过对 GearBind 表示 $h(x)$ 应用线性层来预测：

$$E(x; \boldsymbol{\theta}) = \text{Linear}(h(x)). \quad (7)$$

给定来自 PDB 的观察数据集 $\{x_1, \dots, x_T\}$ ，我们的目标是最大化这些样本的概率：

$$\max \frac{1}{2T} \sum_t \log p(x_t; \boldsymbol{\theta}). \quad (8)$$

然而，直接优化该目标是不可行的，因为计算分区函数需要对整个蛋白质结构空间进行积分。为了解决这一问题，我们采用了一种学习基于能量模型的流行方法，称为**噪声对比估计**（noise contrastive estimation）^[24]。对于每个观察到的结构 x_t ，我们采样一个负样本结构 y_t ，然后将问题转换为一个二元分类任务，即判断一个样本是否在数据集中观察到：

$$\min \frac{1}{2T} \sum_t \log [\sigma(E(x_t; \boldsymbol{\theta}) - E(y_t; \boldsymbol{\theta}))], \quad (9)$$

其中， $\sigma(\cdot)$ 是 Sigmoid 函数。

其中， $\sigma(\cdot)$ 表示用于计算样本 x_t 属于正类概率的 Sigmoid 函数。我们可以看出，上述训练目标旨在降低正样本（即观察到的结构）的能量，同时提升负样本（即突变结构）的能量。

对于负采样，我们在对应的正样本上执行随机单点突变，并通过保持骨架不变且在突变位点从骨架依赖的侧链旋转构象库^[25] 中采样侧链扭转角来生成其构象。此外，为了进一步增强模型区分结构噪声的能力，我们随机选择 30% 的残基，在生成负样本时随机旋转其扭转角。

在 CATH 数据库上预训练之后，我们对 GearBind 编码器进行微调，以在下游任务中进行预测，从而避免过拟合。

4.8 SKEMPI 上的交叉验证

在交叉验证过程中，模型被训练和测试五次，每次使用不同的子集作为测试集，其余四个子集用于训练。在训练集中计算结果，并报告每个测试集的平均值和标准误差作为最终的交叉验证性能。在交叉验证过程中，每个数据点恰好会被纳入一次测试集。这确保了生成一个综合的“测试结果表”，其中包括每个数据点在其作为测试集时的预测值。随后，通过各种标准拆分该表，并对每个子集的性能进行评估。

在 SKEMPI 数据集上的交叉验证后，我们获得了五组模型参数。在推理阶段，我们使用这五个检查点的预测值的平均值作为模型的最终预测结果。

4.9 基线模型实现细节

FoldX. 在本工作中，我们使用 FoldX^[9] 进行突变结构生成。首先，每个 PDB 文件通过 RepairPDB 命令进行结构校正。然后，利用 BuildModel 命令生成野生型和突变体结构对。最后，通过 AnalyseComplex 命令，根据野生型和突变体结构获得 FoldX 的 $\Delta\Delta G_{\text{bind}}$ 预测结果。

Flex-ddG. 我们使用其官方实现来运行 Flex-ddG，详见 https://github.com/Kortemme-Lab/flex_ddG_tutorial。每个 PDB 文件首先通过 PDBFixer v1.8^[37] 进行处理，使用默认参数。对于每次突变，我们采样 35 个结构模型，并设置回溯路径数量为 35,000，能量函数为 fa_talaris2014。最终的 $\Delta\Delta G_{\text{bind}}$ 值通过一个加权评分项的广义加法模型预测得出。

GearBind(+P). 我们基于 TorchDrug 库^[41] 实现 GearBind。对于消息传递，我们采用了 4 层 GearBind 模型，其隐藏维度为 128。在边消息传递中，边的连接根据角度分为 8 个区间。为了从图表示中预测 $\Delta\Delta G_{\text{bind}}$ ，我们使用了 2 层 MLP。

模型使用 Adam 优化器进行训练，学习率为 1e-4，批次大小为 8。训练过程在 1 块 A100 GPU 上进行了 40 轮训练。对于预训练，我们使用相同的架构（4 层 GearBind 模型，隐藏维度为 128），使用 Adam 优化器，学习率为 5e-4，批次大小为 8，在 4 块 A100 GPUs 上进行了 10 轮预训练。**Bind-ddG.** 为了确保公平比较，我们在 SKEMPI 数据集的划分上重新实现并训练了 Bind-ddG 模型。我们遵循原始实现的配置，将隐藏层和配对表示的维度设置为 128 和 64。我们的验证结果表明，对于我们的配置，最优设置包括两层几何注意力机制和四层 MLP 预测器。我们使用 Adam 优化器进行训练，学习率为 1e-4，批次大小为 8，在 1 块 A400 GPU 上总共训练了 40 轮。

4.10 CR3022 和 anti-5T4 UdAb 的 *in silico* 亲和力成熟

选择 PDB 编号为 6XC3^[42] 的结构，其中 H 和 L 链分别组成 CR3022 抗体，而 C 链为 SARS-CoV-2 的 RBD，作为 CR3022 亲和力成熟的起始复合物。为了更好地模拟 CR3022 与 Omicron RBD 的相互作用，我们使用 SWISS-MODEL^[35] 构建了 BA.4 和 BA.1.1 突变体的复合结构。接下来，我们在 CR3022 的 CDR 区域进行饱和突变，并使用 FoldX^[9] 和 Flex-ddG^[10] 生成突变体结构。具体来说，对重链 H 的残基 26–35、50–66、99–108 和轻链 L 的残基 24–40、56–62、95–103 进行突变。总共生成了 1400 个单点突变（包括自突变）。我们使用我们的模型对突变体进行排序，并选择排名靠前的突变体进行亲和力预测和后续实验验证。

未发表的复合物结构用于优化 anti-5T4 UdAb。作为结合两个不同 5T4 表位（图 5b）的单域抗体，anti-5T4 UdAb 相较于传统抗体具有更大的界面区域。通过分析其与 5T4 的界面，我们决定在以下残基上进行饱和突变：1、3、25、27–30、31–33、39–45、52–57、59、91–93、95、99、100–102、103、105、110、112、115–117。这总共生成了 780 个单点突变（包括自突变），并按照上述相同的排名和选择策略进行筛选。

4.11 抗原制备

SARS-CoV RBD 编码基因由 Genscript (南京, 中国) 合成, 并亚克隆至带有 C 末端人 IgG1 Fc 片段和 AviTag 的 pSectag2B 载体中。重组载体被转染至 Expi 293 细胞中, 并在 37°C 下培养 5 天。培养物以 $2200 \times g$ 离心 20 分钟, 收集上清液并通过 0.22 μm 真空滤膜过滤。随后将目标蛋白通过 Genscript 的树脂柱纯化, 使用 PBS 洗涤, 并收集流出液。然后用 0.1 M 甘氨酸 (pH 3.0) 洗脱, 并用 1 M Tris-HCl (pH 9.0) 中和, 再通过磷酸盐缓冲液 (PBS) 进行透析和浓缩, 使用 Amicon 超滤离心浓缩器 (Millipore), 其分子量截断为 3 kDa。

蛋白浓度使用 NanoDrop 2000 分光光度计 (Thermo Fisher) 测量, 蛋白纯度通过 SDS-PAGE 检测。Delta RBD 蛋白由 Vazyme (南京, 中国) 采购, Omicron S 蛋白由 ACROBiosystems (北京, 中国) 采购。生物素化的人 TPBG/5T4 和人 TPBG/5T4-Fc 抗原由 ACROBiosystems (北京, 中国) 采购。

4.12 CR3022 抗体突变体与野生型的制备

不同 CR3022 抗体的重链和轻链基因被合成并亚克隆至带有 IgG1 格式的表达载体 pcDNA 3.4 中。这些构建载体被转染至 CHO 细胞, 并通过 Protein A 进行纯化。所有抗体由百奥赛图生物技术公司 (上海, 中国) 生产。

4.13 突变体和野生型 anti-5T4 UdAb 的制备

之前的工作中构建了编码野生型 anti-5T4 UdAb 的 pComb3x 载体, 并在本实验室保存。所有 anti-5T4 UdAb 的单点突变体均使用 QuickMutationTM 定点突变试剂盒 (Beyotime, 上海, 中国) 按照生产商的协议进行构建。不同突变体和野生型 anti-5T4 UdAb 的表达在 30°C 下, 于 E. coli HB2151 细菌中培养 16 小时, 同时添加 1 mM 异丙基- D-1-硫代半乳糖苷 (IPTG) 进行诱导表达。细胞在 30°C 下用多粘菌素 B 裂解 0.5 小时, 然后以 $8800 \times g$ 离心 10 分钟, 收集上清液。上清液经过 0.8 μm 聚醚砜膜过滤后, 使用 Ni-NTA (Smart Lifesciences) 纯化, 按照生产商的说明操作。

简而言之, 过滤后的上清液被装载至 Ni-NTA 树脂柱, 使用洗涤缓冲液 (10 mM Na₂HPO₄、10 mM NaH₂PO₄ [pH 7.4]、500 mM NaCl、20 mM 咪唑) 进行洗涤, 蛋白随后在洗脱缓冲液 (10 mM Na₂HPO₄、10 mM NaH₂PO₄ [pH 7.4]、500 mM NaCl、250 mM 咪唑) 中洗脱。收集的洗脱组分立即透析至 PBS, 并使用 Amicon 超滤离心浓缩器 (Millipore) 进行浓缩。蛋白浓度使用 NanoDrop 2000 分光光度计 (Thermo Fisher) 测量, 蛋白纯度通过 SDS-PAGE 检测。

4.14 酶联免疫吸附测定 (ELISA)

为了比较不同 CR3022 突变体的结合能力, 将 Delta 毒株 (B.1.617.2) RBD 和 Omicron 毒株 (B.1.1.529) S 蛋白以 100 ng/孔的浓度包被在 96 孔半面积微孔板 (Corning #3690) 上, 4°C 过夜孵育。抗原包被板用 PBST (含 0.05% Tween-20 的 PBS) 清洗三次后, 用 3% MPBS (含 3% 脱脂奶粉的 PBS) 在 37°C 阻断 1 小时。随后, 加入 1% MPBS 稀释的三倍梯度抗体 50 μL , 于 37°C 孵育 1.5 小时。使用 HRP 标记的抗-Fab 和抗-Fc 次级抗体 (Sigma-Aldrich) 检测不同的测试抗体。用 PBST 清洗五次后, 加入 ABTS 底物 (Invitrogen) 显色, 37°C 反应 15 分钟后测定酶活性。数据通过测量 405 nm 处的吸光度获取, 使用 Microplate Spectrophotometer (Biotek) 读取。通过 GraphPad Prism8.0 软件计算 EC₅₀ (达到最大效应的 50% 浓度)。

为了验证不同 UdAb 突变体, 采用上述相同的实验流程。简而言之, 人源 TPBG/5T4-Fc 抗原包被在 96 孔半面积微孔板上, 然后用 3% MPBS 阻断, 并加入不同稀释的抗体。使用 HRP 标记的抗-Flag (Sigma-Aldrich) 次级抗体进行检测, 随后加入 ABTS 底物显色, 并在 405 nm 处读取吸光度。报告的 EC₅₀ 值为三次重复实验的平均值。

4.15 生物层干涉 (BLI) 结合测定

使用 BLI 在 Octet-RED96 仪器 (ForteBio) 上测定不同抗体与 SARS-CoV-2 Omicron S 和 5T4 抗原的结合动力学。简要来说，将 8 µg/ml 的带有 His 标签的 Omicron S 蛋白和 5 µg/ml 的生物素标记 5T4 蛋白分别加载至 Ni-NTA 和链霉亲和素涂层 (SA) 生物传感器上。抗原偶联的传感器依次与三倍稀释的 CR3022 候选抗体或二倍稀释的 anti-5T4 UdAb，从 300 nM 开始，在含 0.02% PBST 中孵育 300 s 进行结合，然后浸入 0.02% PBST 中再孵育 300 s 于 37°C 进行解离。使用 Data Analysis 软件 11.1 通过 1:1 结合模型拟合曲线，所有 K_D 值通过 R^2 大于 98% 的置信区间确定。

4.16 蛋白质结构与 $\Delta\Delta G_{\text{bind}}$ 贡献分析

蛋白质结构分析是通过 Python 脚本进行的。在 Rosetta Flex-ddG 松弛处理后，获取突变抗体-抗原复合物的结构。松弛后的蛋白质结构能够提供更准确的侧链构象，这对于精确的接触和构象分析至关重要。这种分析改进了理解底层结合机制的精度，并有助于识别蛋白-蛋白相互作用中的关键特征。在 GearBind 上使用集成梯度 (Integrated Gradients, IG)⁴³，一种模型无关的归因方法，进行残基级别的解释分析。所有蛋白质结构图均使用 PyMOL v3.0 绘制。

4.17 分子动力学模拟

为了分析抗体突变的结构，我们对野生型和突变体抗体-抗原复合物进行了分子动力学模拟。初始结构取自 GearBind 使用 Rosetta Flex-ddG 松弛后的结果。在 AMBER 22 软件包中使用 LEaP 模块构建起始结构，并添加离子和溶剂以准备模拟⁴⁴。分子质子化状态按照 LEaP 在初始结构准备中的默认设置保持不变。所有系统使用 ff19SB 力场⁴⁵，并溶解于 OPC3⁴⁶ 溶剂模型的水盒中。模拟系统在 10 埃的溶剂盒中溶解。所有含有氢原子的键被 SHAKE 算法⁴⁷ 约束。粒子网格 Ewald (PME) 算法用于模拟。

分子动力学模拟用于计算长程静电相互作用⁴⁸。初始结构经过最大 10,000 步的最小化以确保收敛，然后在 NPT 系统中使用 PMEMD 加热 20 ps 并平衡 10 ps。CUDA 版本的 PMEMD 被用于加速模拟⁴⁹。模拟温度设置为室温 298K。所有系统均以 1 s 的生产 MD 模拟，其中取一个副本，并排除前 50 ns 的数据进行分析。AMBERTools 中的 CPPTRAJ 模块用于分析模拟结果，包括计算均方根偏差 (RMSD)、均方根波动 (RMSF)、氢键和二面角⁵⁰。

4.18 研究报告摘要

有关研究设计的更多信息，请参见链接至本文的 Nature Portfolio Reporting Summary。

4.19 数据可用性

原始 SKEMPI 数据库可通过 <https://life.bsc.es/pid/skempi2> 访问。CATH 数据库可通过 <https://www.cathdb.info/> 访问。HER2 绑定数据可通过 <https://github.com/AbSciBio/unlocking-de-novo-antibody-design/blob/main/spr-controls.csv> 访问。本文提供了源数据。

4.20 代码可用性

GearBind 推理代码、训练好的模型检查点和数据集预处理脚本可通过 <https://github.com/DeepGraphLearning/GearBind> 获取，遵循 Apache 2.0 许可协议。此外，还可通过 Zenodo 获取⁵¹。