

# Chai-2模型发布掀起 “AI抗体设计”风暴

# CONTENT

|                          |    |
|--------------------------|----|
| 第一章 引言                   | 03 |
| 第二章 正文内容                 | 06 |
| 2.1 背景介绍                 |    |
| 2.2 抗体设计的扩散模型            |    |
| 2.3 从头设计抗体的方法            |    |
| 2.4 抗体优化设计的方法            |    |
| 2.5 蛋白质通用设计扩散模型在抗体设计中的应用 |    |
| 2.6 老子数据集在抗体抗原数据上的优势     |    |
| 第三章 全文总结                 | 21 |



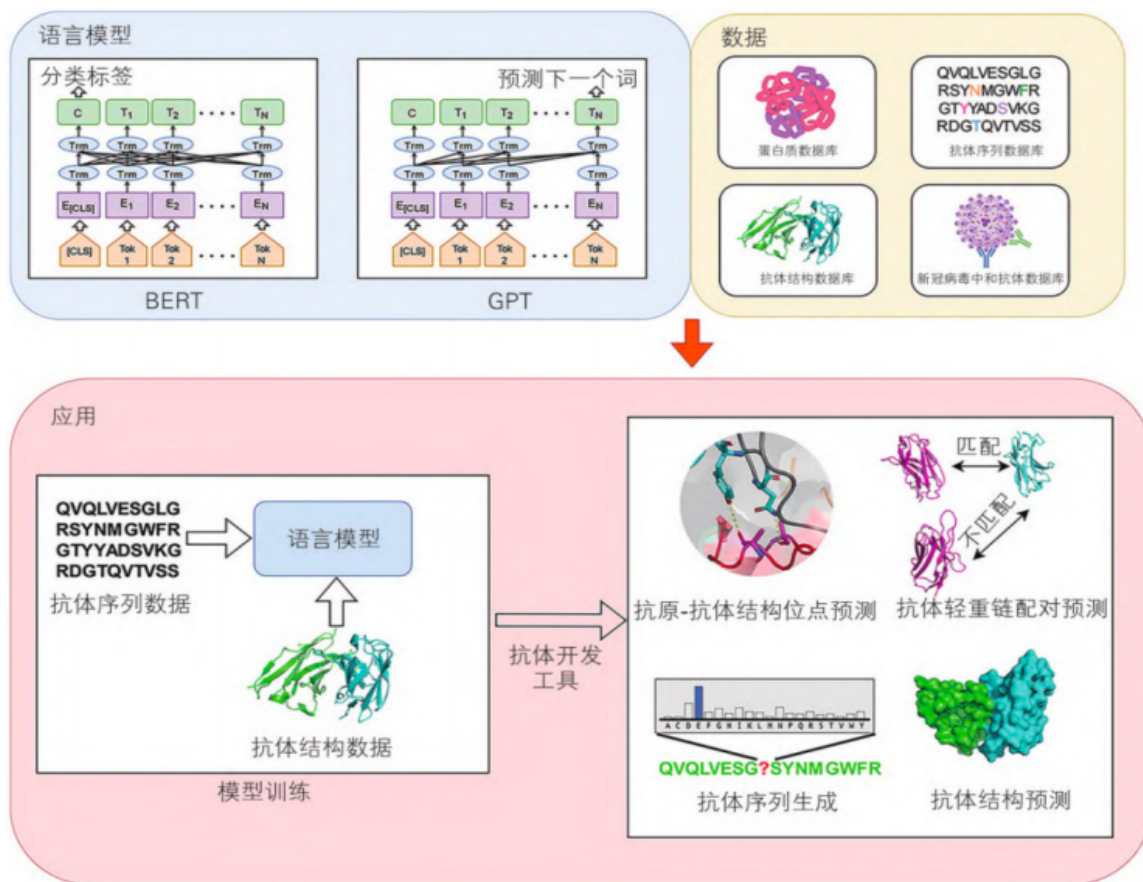
第一章



# 引言



2025年6月30日，OpenAI支持的Chai Discovery推出Chai-2，这款多模态生成模型展现出强大的抗体设计能力。Chai-2能够在不依赖现有抗体或高通量筛选的情况下，完全零样本生成抗体，实现了两位数的成功率，较传统方法提高了100倍。将原本需要几个月时间的抗体从头设计工作压缩至两周以内，显著加速药物发现的进程。



语言模型在抗体开发领域的潜在应用

随着计算机算力和深度学习技术的迅速发展，尤其是生成式语言模型（如GPT系列）和表征式语言模型（如BERT系列）的出现，为抗体设计与优化提供了更为强大和高效的工具。同时，既往的抗体文库测序工作积累了大量序列数据。这些模型通过自监督学习大量抗体序列数据，挖掘潜在的抗体进化信息和共性特征，使得利用语言模型指导实验方法进行抗体开发成为可能。

随着语言模型辅助人工智能抗体设计与优化领域相关研究进展，出现大量有价值的研究成果。Bai和Chungyoun等提出**抗体预训练语言模型构建**、**抗体序列生成**以**抗体三维结构预测**为该领域的3个重要研究任务。

在计算和AI领域，尤其是**生成式扩散模型**方面，已经开始着手解决这些难题，为抗体设计提供了新颖的方法。基于扩散的生成方法辅助抗体设计，包括**从头设计抗体**以及**互补决定区（CDR）环的优化**。Chai-2便是生成式扩散模型的一种，它采用全原子扩散模型+条件生成（抗原导向）的方式，从输入的已知靶点结构和表位残基开始，设计出特异性结合该表位的序列与全原子结构。



## 第二章



# 正文内容

## 2.1 背景介绍

抗体是免疫系统抵御外来分子和病原体的重要组成部分，它们展现出极强且特异性的与抗原结合的能力。抗体已成为领先的治疗药物，一直位列畅销药品之列。抗体因其强大的识别和生成方法而成为独特的治疗药物。其良好的药代动力学、极低的毒性以及低免疫原性，再加上易于表达和纯化的特点，进一步使其脱颖而出。

抗体分为五种不同的亚型：IgG、IgM、IgA、IgD 和 IgE。这些亚型在结构上各不相同，反映了它们在免疫反应中的特定作用。美国食品药品监督管理局（FDA）批准的单克隆抗体治疗药物中最常见的类别遵循 IgG 格式，由两条重链和两条轻链组成（图 1a）。重链包含一个可变区域（VH）和三个恒定区域（CH1、CH2、CH3），而轻链由一个可变区域（VL）和一个恒定区域（CL）组成。此外，这些区域根据功能被重新分类为片段。抗原结合片段（Fab）包含 VH、CH1、VL 和 CL 区域。其余区域构成可结晶片段（Fc）。此外，重链和轻链的可变区域共同构成了可变片段（Fv），这是大多数抗体设计工作中主要关注的对象。每个 Fv 区域由三个高度变异性区域（也称为互补决定区，CDRs）组成。CDRs 使抗体对目标抗原的表位（即与抗体相互作用的精确区域）具有特异性。因此，CDRs 具有高度的变异性（图 1b），并且在抗体设计中始终是重点关注的对象。

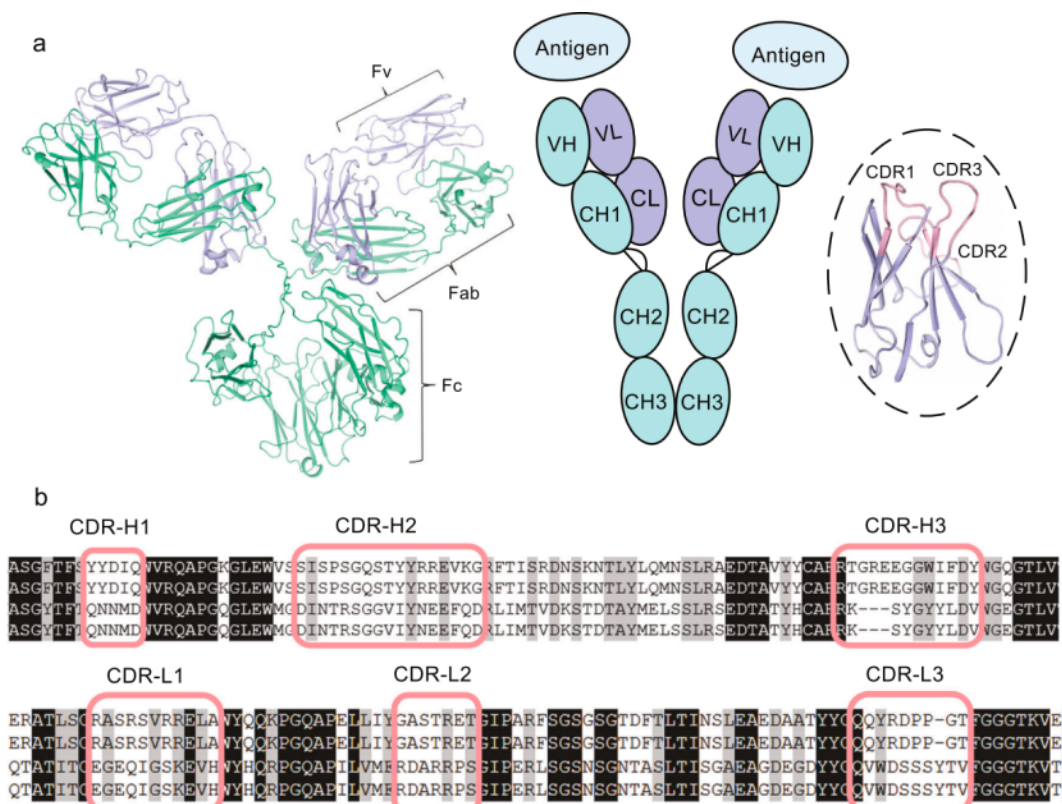
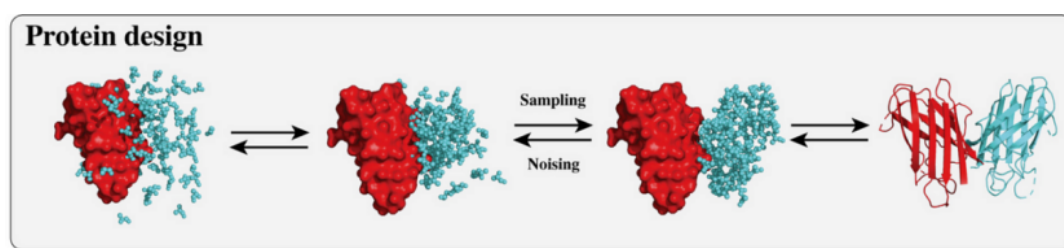


图1a 抗体典型结构；图1b 4 种人源 IgG 抗体的重链和轻链序列比对

传统的设计抗体的方法不仅成本高昂、耗时漫长，而且无法生成针对所需抗原表位的抗体。此外，通过亲和力成熟来优化现有抗体还会带来额外的问题，需要反复进行突变和筛选以获得具有更高亲和力的抗体，这一过程本质上是劳动密集型且缓慢的。计算技术可用于快速识别和优化高亲和力抗体。通常，抗体设计的计算技术包括基于自由能的传统方法以及数据驱动的方法。

随着数据量的不断增加以及模型的不断进步，用于科学研究的人工智能（AI）发展迅速。具体而言，人工智能在蛋白质设计方面做出了重大贡献。传统的诸如自回归模型、变分自编码器（VAE）和生成对抗网络（GAN）等生成模型可应用于蛋白质设计。与文本生成任务类似，语言模型在蛋白质设计中得到了广泛应用，通过遮蔽残基或更广泛地遮蔽标记等方式。



蛋白质扩散设计过程

目前，扩散模型（diffusion models）已迅速成为生成模型领域的主要研究方向。这些模型在计算机视觉、自然语言处理和计算生物学等领域展现出了卓越的能力。在几个关键领域（如从头蛋白质设计、小分子设计、分子构象生成以及盲法蛋白质-配体对接）中，它们都超越了传统方法。理论和实际应用都表明，通过给出 prompt，扩散模型在将一种分布迅速转换为另一种分布方面具有极高的有效性，这一点对于抗体生成任务尤为重要，在这种任务中，抗原起到了自然的引导作用。

扩散模型的核心思想是通过一个迭代的去噪过程，从随机噪声中逐步恢复出有意义的数据结构。将其应用于蛋白质设计，意味着模型能够从一个完全随机的初始状态，逐步“去噪”并“塑造”出目标蛋白质的三维结构和对应的氨基酸序列。这种创新性的从头设计方法，与传统的基于模板或基于物理模拟的方法截然不同，能够学习并内化蛋白质结构与序列的复杂内在规律，从而生成前所未有且功能新颖的蛋白质分子。

此外，扩散模型广泛利用大型数据集，利用大量信息来指导实验。抗体设计领域拥有丰富的数据。考虑到它们在蛋白质设计方面的成功，扩散模型代表了生成模型领域的一项前沿技术，有可能彻底改变抗体设计。



| Algorithms                    | Training Set                                          | Output Type            | Description                                                                                                                                      |
|-------------------------------|-------------------------------------------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| AbDiffuser                    | Sequences and structures from pOAS                    | Sequence and structure | An equivalent and physics-informed diffusion model for co-design of antibody sequences and structures                                            |
| AntiBARTy Diffusion           | Sequences from OAS and UniProtKB                      | Sequence               | A property-conditional diffusion model trained using the latent space of an antibody- specific language model for guided antibody de novo design |
| EAGLE                         | Structures from SAbDab                                | Sequence               | Epitope-specific based diffusion model for generation of antibody sequence using a language model.                                               |
| DiffAb/guided DiffAb/ab_opt   | Structures from SAbDab                                | Sequence and structure | Diffusion models specifically designed for joint generation of antibody sequence in CDR optimization                                             |
| Diffusion Optimized Sampling  | Sequences with affinity data                          | Sequence               | A guidance method for discrete diffusion models that follows gradients in the hidden states of the denoising network.                            |
| Antibody-specific RFdiffusion | Structures from SAbDab, TCR and PDB                   | Sequence and structure | Finetuned RFdiffusion to design <i>de novo</i> antibody structures                                                                               |
| Chroma                        | Structures from PDB, Uniprot and PFAM                 | Sequence and structure | A diffusion model developed for the design of proteins.                                                                                          |
| EvoDiff                       | Sequences from UniRef50, OpenFold, PDB and UNiClust30 | Sequence               | A general-purpose diffusion framework for controllable sequence generation.                                                                      |
| Vanilla RFdiffusion           | Structures from PDB                                   | Sequence               | A diffusion model utilizing RoseTTAFold to generate <i>de novo</i> protein structures.                                                           |

（表1）可用于抗体设计的扩散模型概述（RFdiffusion 是专门用于抗体设计的方法，而 Chroma、EvoDiff 和普通 RFdiffusion 是用于蛋白质设计的方法，但也可应用于抗体设计。pOAS 是成对可观测抗体空间的缩写）

抗体设计领域中扩散模型的最新进展见表 1。抗体设计的扩散模型的应用分为两个主要领域：

- (i) 全新抗体从头生成模式，旨在生成针对特定抗原的新型抗体；
- (ii) 抗体优化模式，侧重于基于已知序列或结构生成更优的抗体。

## 2.2 抗体设计的扩散模型

生成扩散模型最早应用于计算机视觉领域，它是一种通过逐步添加噪声并训练神经网络逆向去噪的生成模型。这种模型被成功应用于图像生成、视频生成等任务，并在蛋白质设计、分子构象生成等领域取得了显著成效。

在抗体设计中，生成扩散模型通过将噪声添加到抗体的序列或结构数据中，然后逆向去噪以生成新的抗体结构。该过程包括两个阶段：前向过程，逐步向输入抗体数据添加噪声；逆向过程，通过神经网络去除噪声，还原出新的抗体结构。

生成扩散模型中的去噪扩散概率模型（DDPM）是最常用的框架之一。它通过在训练数据中逐步添加高斯噪声，并训练网络在每一步预测出去噪后的数据，从而实现对抗体的生成。具体来说，前向过程中的每一步都通过如下公式来添加噪声：

$$X_t = \sqrt{\alpha_t} \cdot X_0 + \sqrt{1 - \alpha_t} \cdot \varepsilon$$

在逆向过程，通过训练好的神经网络对每一步的去噪结果进行预测，从而最终生成新的抗体结构。

扩散模型特别适合抗体设计任务，因为它能够基于抗原信息生成具有特定亲和力和特异性的抗体。

如图 2 所示。与用于图像生成的扩散模型如何用噪点像素来破坏图像的方式类似，向输入的抗体序列数据、结构数据或两者中添加噪声。同样，在训练神经网络的过程中也可以对其进行去噪处理。训练完成后，可以将扩散模型输入不同的噪声分布来生成新的抗体。

扩散模型要求其处理的数据需经过适当的编码。对于抗体数据（无论是序列数据、结构数据，还是两者兼而有之），都需要一种特定的编码方法，以便这些模型能够理解这些数据。从序列的角度来看，氨基酸通常使用一个 20 维向量空间中的“独热编码”来表示。

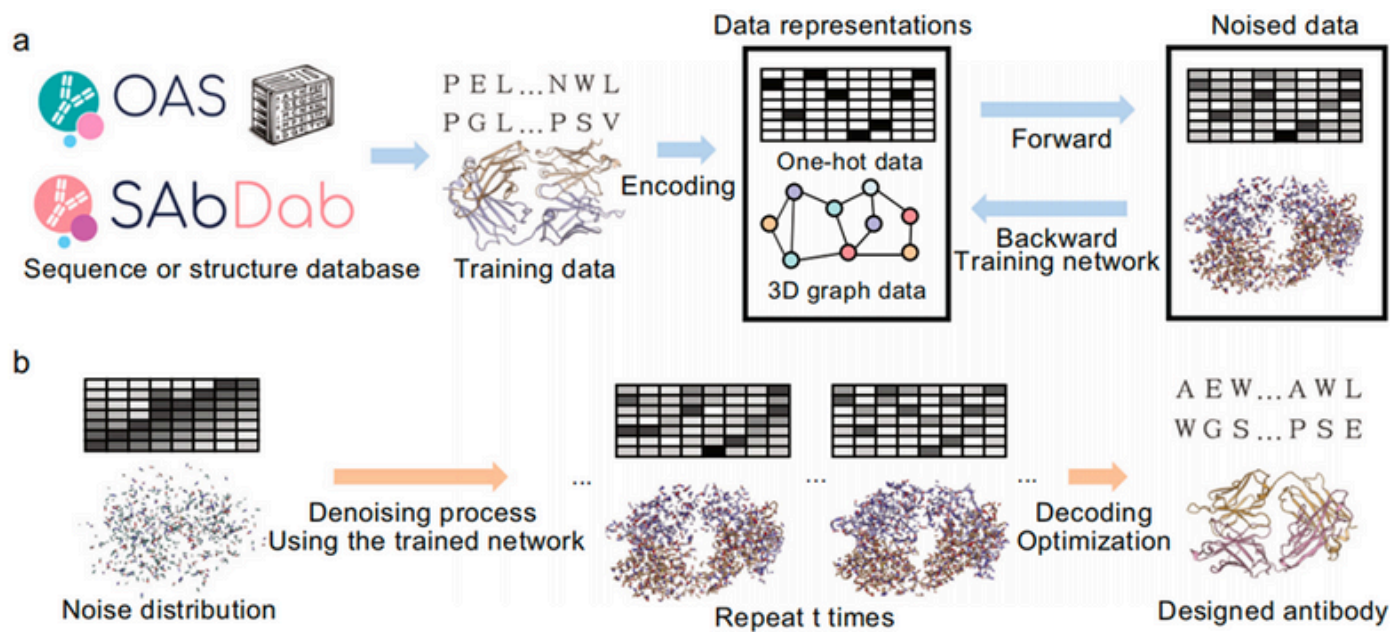


图2 基于扩散原理的抗体设计示例



## 2.3 从头设计抗体的方法

“从头设计抗体”指的是利用计算工具和方法完全从零开始生成抗体的过程，不依赖于已有的抗体。该方法旨在基于所期望的目标设计具有特定特性的新抗体，例如结合亲和力、特异性、甚至结合表位等。扩散模型由于能够学习复杂的数据分布并生成多样且高质量的样本，在新蛋白质设计方面取得了巨大成功。按照同样的思路，扩散模型还可以专门针对现有的**抗体数据**进行训练，以学习天然抗体序列和结构的潜在分布。一旦经过训练后，该模型能够生成与天然抗体相似但具有多样性和潜在优化特性的新型抗体。图3所示为生成式扩散模型在从头抗体设计中的应用。

RFdiffusion 是一种用于全新蛋白质设计的强大工具，但最初它无法解决从零开始设计抗体这一复杂的难题。然而，通过对原始的 RFdiffusion 模型进行微调，取得了重大突破。这包括将抗体框架结构作为残基对之间的二维距离和角度的二维矩阵纳入训练过程。此外，该模型继承了其前身的“热点”特性，以指定与抗体互补决定区（CDRs）的表位相互作用。针对抗体的 RFdiffusion 利用 ProteinMPNN 进行 CDR 序列设计，并采用定制的 RoseTTAFold2 进行 CDR 模型构建和筛选。这个训练过程涵盖了多样化的数据集，其中包括抗体的结构、T 细胞受体（TCR）以及界面处的环状结构，从而确保了全面的学习范围。这种模型能够生成单域纳米抗体。实验结果表明了这种方法的有效性，所设计的抗体对诸如呼吸道合胞病毒（RSV）、SARS-CoV-2、流感和 TcdB 等目标的亲和力范围从微摩尔到亚微摩尔不等，突显了其高度的特异性。尽管所设计的抗体的结合亲和力仍然较低，且总体成功率还需要进一步提高，但抗体特异性 RF 扩散标志着在针对特定抗原构象的抗体的理性从头设计方面迈出了重要的一步。

AntiBARTy Diffusion 是一种基于语言模型的生成扩散模型，能够通过对抗体序列进行条件生成，指导其生成符合特定属性（如溶解性）的抗体序列，生成的抗体在体外实验中表现出较高的溶解性。

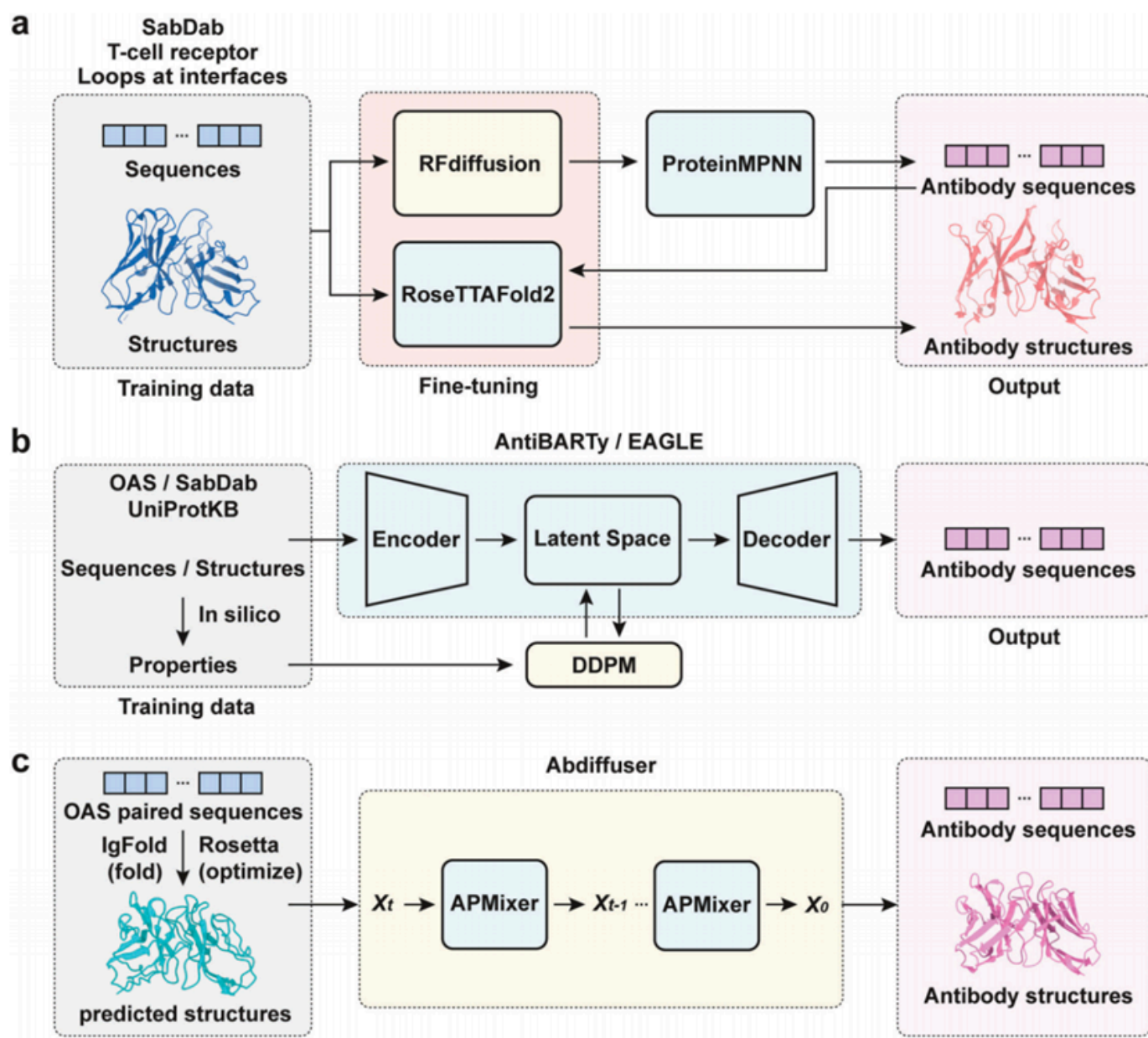


图3 生成式扩散模型在从头抗体设计中的应用

主要创新点包括：在潜空间中引入聚合嵌入（pooled embeddings）以实现条件控制与无分类器引导（classifier-free guidance），并推动了基于属性的抗体设计方法的发展。该模型架构由一个用于 AntiBARTy 的 BART 式转换器和一个用于扩散模型的 U-Net 主干网络组成。AntiBARTy 扩散模型是在一个包含超过 6 亿个抗体 Fv 序列的大型数据集上进行训练的，这些序列是从 Observed Antibody Space (OAS) 和 UniProt 知识库 (UniProtKB) 中提取的，从而确保其接触到各种抗体变异。虽然在提高免疫球蛋白的生成效率以及增强其在计算机模拟中的溶解性方面颇具前景，但该研究也承认存在一些局限性，比如缺乏实验验证以及无法设计出针对抗原的特异性抗体。借助语言模型嵌入技术，EAGLE 还能够根据抗原表位结构生成抗体序列。这些模型能够生成整个 Fv 区域（包括 VH 和 VL）的序列。

AbDiffuser 也是一种扩散模型，旨在同时生成抗体的三维结构和序列。它通过一种名为“对齐蛋白质混合器”（APMixer）的新架构融入了对称性和基于物理原理的见解。APMixer 是一种基于 MLP-Mixer 架构的神经网络，专为像抗体这样的对齐蛋白质而设计。AbDiffuser 采用了去噪扩散先验方法，其中包括位置特定的残基频率以及原子位置之间的条件依赖关系，从而能够精确生成抗体结构。该模型基于 OAS 数据集进行训练，并在曲妥珠单抗 H3 突变体上进一步进行微调，以设计针对 HER2 的结合剂。AbDiffuser 的优点包括高效利用计算资源以及生成高保真的抗体结构和序列。然而，它无法直接且灵活地基于特定靶点设计抗体。

2.4 抗体优化模式设计的方法

快速发针对特定靶点的候选抗体

示例 1

我正在研究TAU靶点。请帮我找出对这个靶点的最佳抗体。

示例 2

帮我评估这个PDL1抗体序列，找出相关的优选抗体： 链接: EVMTQSPGTLSPGERATLSCRAISGRVSSYLAWYQKPGQAPRLIYDASSRATGIPORIFSGDSGDTSTLTISRLPEDEFAVYYCQYVDSLPWTFGQGTVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKQKWKDALQSGHSGESVTEGDSKSTYSLSTLTLSKADYDKHVVACEVTHQGLSPVTKSFNRGEC 链接: EVQLVESGGGLVQPGGSLRLSCAASGFTFISRYWMSWITQAPKGLIEWANIKDQGSKEYVYDSVKGFFTISRDNAKNSLYLQMNSLRADTAYYCAREGGWFGELAFDYWGQGLTVTVSASSTKGPSPFPLAPSSKSTSGGTAALGCLVADYFPEPYTVSWNSGALTSGVHTFPAWLQSSGLYSLSVVTVPSSSLGTQTYICNVNKKPSNTKVKRVEPKGCKTHCTPCPAPEFEGGSPVFLFPKPKDTLMSRTPEVTCVAVDVSHEDPEVKRFNYYVDGVEHNKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPASEKTKAKGQPREPQVYITLPPSREEMTKNQVSLTCLVKGFYPSDAIVWEISNGQPENNYKTPPVLDSGGFFLYSKLTVDKSRWQQGNVFNCSVMHEALHNHYTQKSLSLSPGK

示例 3

请为OX40靶点分析最佳抗体选择。我需要详细的亲和力数据。

请为OX40靶点分析最佳抗体选择。我需要详细的亲和力数据。

抗体药物研发专家

用户需要OX40靶点的最佳抗体分析，特别关注亲和力数据。OX40是一个重要的免疫靶点。用于肿瘤免疫治疗。我需要通过抗体检索工具找到相关抗体，再获取最新抗体信息，最后推荐最佳抗体。

+ 新问题

你想知道什么?

内容由 AI 生成，仅供参考

推荐抗体详细信息

| 序号 | 抗体名称             | 抗原信息                                                 | KD值 (M)   | 亲和力等级     | 亲和力测定条件   | 表位信息                                                                               | 其它关键性能指标                                                                                   | reference       | 抗体序列                                                                                                         |
|----|------------------|------------------------------------------------------|-----------|-----------|-----------|------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|-----------------|--------------------------------------------------------------------------------------------------------------|
| 1  | Pogalizumab      | 抗原名称: OX40<br>物种: Homo sapiens<br>Uniprot ID: P43489 | <1.00E-12 | Very High | BLI, 37°C | X-ray晶体结构: Q114-Y119,D124,A126-P127,P129-P130,H132,A140,K142<br>Octet技术: R106-P127 | kon: 2.98E+5<br>1/Ma<br>koff: <1.00E-7<br>1/s<br>ic50: 5.98E-9<br>M<br>ec50: 8.26E-11<br>M | US20230295324A1 | 链接: 315624412<br>Seq Code: EVQLVQSGAEVKKPKASVKVSKCAE<br>链接: 315624413<br>Seq Code: DIQMTQSPSSLASVQDRVITTCRAE |
| 2  | psb1949-1        | 抗原名称: OX40<br>物种: Homo sapiens<br>Uniprot ID: P43489 | 1.10E-10  | High      | BLI, 25°C | 丙氨酸扫描: W58,N60,P62,R60,L66,P63,P69,P115                                            | -                                                                                          | US11332536B2    | 链接: 12948825<br>Seq Code: EVQLVESGGGLVQPGGSLRLSCAAS<br>链接: 12948532<br>Seq Code: DIQMTQSPFLPVTPGEPAISGRSS    |
| 3  | 1.134.9-u1-ksG1L | 抗原名称: OX40<br>物种: Homo                               | 1.79E-10  | High      | SPR       | 丙氨酸扫描: G7D,Y72,N88,G82,F94,T105,D104,V106                                          | kon: 2.07E+5<br>1/Ma<br>koff: -                                                            | US20210171647A1 | 链接: 524446850<br>Seq Code: QVQLQESGGGLVQPGGSLRLSCAAS                                                         |

抗体抗原agent根据靶点找优秀抗体

除了全新的任务之外，现有的抗体-抗原复合物结构或初始序列通常也需要进行优化，以增强抗体对抗原的亲和力以及其治疗效果。扩散模型在这一领域已展现出显著的性能。由于抗体框架保持不变，与抗原直接相互作用的互补决定区（CDR）对于优化至关重要。在已知的抗原-抗体结构基础上，优化 CDR 以提高其结合亲和力并改善抗体性能是抗体设计中可行且值得去做的事情。结构和序列信息都可以作为扩散模型的输入，以生成新的 CDR 并优化现有的抗体。例如需要从靶向OX40的抗体中选择一条候选抗体来通过扩散模型训练和改造CDR区，可以借助智慧芽的抗体抗原agent，在agent中输入：为OX40靶点分析最佳抗体选择，agent会调用工具分析智慧芽存储的抗体抗原数据库，把亲和力最强且有抗原表位信息的抗体排出来。（如上图所示）

DiffAb：开创性的序列-结构协同设计方法DiffAb是一种基于扩散模型的开创性抗体设计方法，它通过同时编码氨基酸类型、位置和方向，解决了抗体CDR区设计中序列与结构协同优化的核心难题。其过程是通过“加噪-去噪”来生成新的CDR结构。

15



DiffAb的性能优势与多样性与Rosetta等传统方法相比，DiffAb生成的抗体结构在预测的结合亲和力及结构/序列恢复方面表现更优。此外，它还能通过调整扩散步长来生成高度多样化的序列，同时保持良好的结合亲和力。

DiffAb的持续进化与增强DiffAb的功能通过引入新策略不断得到增强，例如：“属性引导采样”可以优化溶解性等物理化学性质；更先进的图神经网络模型和AlphaFold2的注意力机制也被整合进来，以提升其对-抗原抗体界面的捕捉和整体建模能力。

仅序列模型也可应用于抗体设计。例如，扩散优化采样（NOS）能够生成能够优化任意目标同时又保持对序列参考分布的高可能性的样本。

NOS使用多次去噪扩散采样步骤，其中每个步骤都应用一种干扰、梯度步骤来优化值函数。

NOS还被用于增强一种名为LaMBO的序列设计的贝叶斯优化程序，从而产生了名为LaMBO-2的新方法，该方法根据功能引导的抗体设计的获取值显著性来选择编辑位置，从而显著提高了样本目标值。

NOS生成的序列具有更好的预测亲和力。体外实验表明，由于其对更大训练数据集的广泛访问，

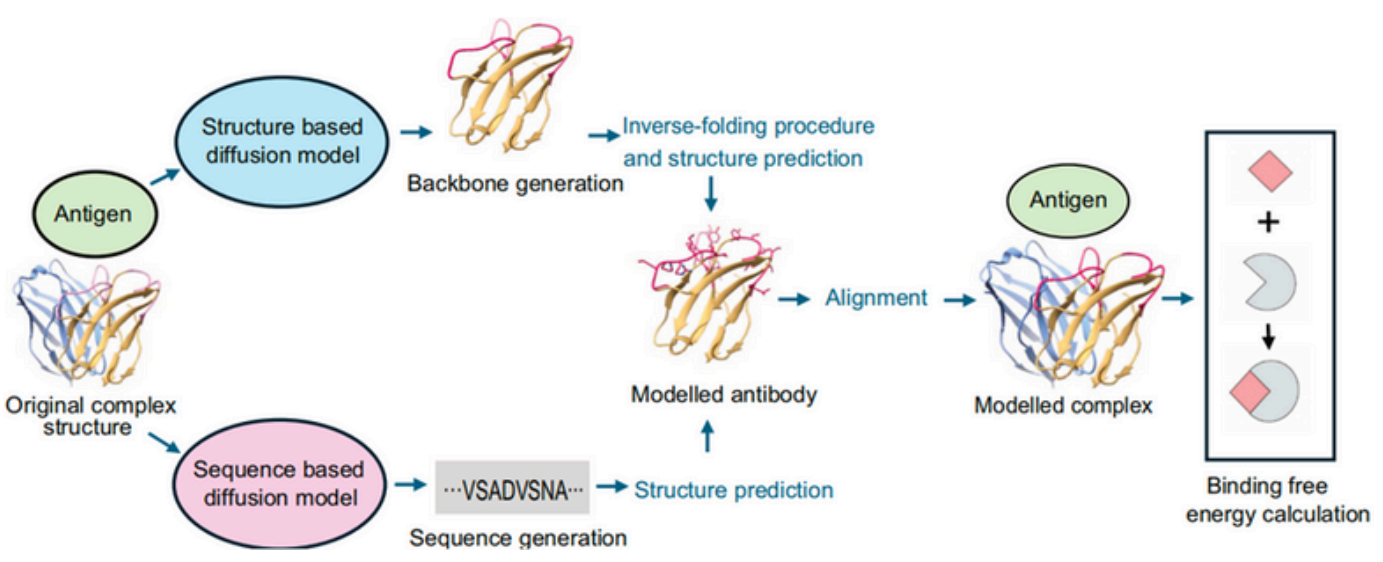
NOS在填充任务上比基于结构的方法表现出更优越的性能。

NOS还由于其引导采样策略而在优化目标和在单目标任务中保持高可能性之间提供了更好的权衡。该方法的体外实验表明，其表达率和结合率都很高。

## 2.5 蛋白质通用设计扩散模型在抗体设计中的应用

由于抗体是一种特殊的蛋白质，所以一般的蛋白质设计算法也可以应用于抗体的设计中。在扩散模型领域，此类方法可归类为基于序列和基于结构的两类。生成的结果取决于掩蔽区域：掩蔽互补决定区（CDR）会使模型生成一个新的CDR，而掩蔽整个抗体则会构建一个新的抗体。图4描述了使用蛋白质设计模型生成抗体序列或结构的通用过程以及相应的评估。

用于蛋白质设计的典型基于序列的扩散模型有 EvoDiff 和 Taxdiff 。此类算法可用于生成新的 CDR 序列或框架。这些模型是基于大量的蛋白质数据集（如 UniRef50 ）进行训练的。在训练过程中，序列会被屏蔽，而模型则会尝试重建原始序列。在 CDR 优化的应用中，可以有选择地屏蔽 CDR 区域，从而利用基于序列的模型生成新的 CDR。然后，这些序列会通过诸如 AlphaFold 或 Modeller 等结构预测工具折叠成抗体结构。



利用通用蛋白质设计扩散模型进行抗体设计的过程

此外，基于结构的扩散模型，如 RFdiffusion 和 Chroma 也是适用的。这些模型是使用 PDB 数据库中的数据进行训练的。例如，RFdiffusion 在训练过程中会扩散  $C\alpha$  坐标和主链二面角，模型随后会利用这些信息生成结构。随后，ProteinMPNN 被用于将残基分配到生成的主链上。模体支架方法（保持固定的非 CDR 区域）特别适合优化 CDR。在以 Chroma 为例，训练过程涉及对骨干结构、序列以及侧链方向进行扩散处理，同时使用贝叶斯后验 SDE 进行模型推理。Chroma 的条件生成能力（结合子结构条件器）使其非常适合 CDR 优化。无论使用基于序列的还是基于结构的扩散模型，生成的 CDR 都应被建模为复合物并进行评估。这可以通过自由能估计或亲和力估计来完成，以评估结合效力，因为模型在生成过程中不考虑抗原信息。

## 2.6 老子抗体抗原数据系统（Lao Tzu Antibody-Antigen Dataset）

智慧芽开发的老子抗体抗原数据系统，是一个经过精心筛选并全面索引的抗体-抗原配对数据集，包括表位、亲和力等信息，数据来源于经过验证的参考文献和专利。该数据系统旨在助力抗体研发团队验证抗体与靶点之间的关系，并用于训练和测试人工智能/机器学习模型——以高效且精准的方式加速模拟和全新抗体的设计过程。

如上文所述，扩散生成模型能应用于抗体的从头设计和已有抗体序列的优化设计；扩散生成模型的训练依赖于海量的抗体抗原数据，抗体抗原数据的质量也直接影响着扩散生成模型生成抗体序列或优化抗体序列的能力。

|             | Antibody-Antigen Pairings | Affinity (kD) | Epitope | IC <sub>50</sub> , EC <sub>50</sub> |
|-------------|---------------------------|---------------|---------|-------------------------------------|
| Data system | Lao Tzu                   | Lao Tzu       | Lao Tzu | Lao Tzu                             |
| Est YoY+    | ~10K+                     | ~10K+         |         | ~10K+                               |
| 2025        | 120,000+                  | 20,000+       | 2,000+  | 24,000                              |
| 2026        | 130,000+                  | 30,000+       | 3,000+  | 34,000                              |
| 2027        | 140,000+                  | 40,000+       | 4,000+  | 44,000                              |

智慧芽的老子抗体抗原数据系统是通过AI+人工审核的方式，对海量的来自于专利、文献、pdb库、NCBI、synapse、SAbDab、UniRef50、OpenFold、UniClust30等来源的抗体抗原数据进行了数据清洗、数据治理、数据标注和数据审核之后的集成系统。老子数据系统准确度达到90%以上，截止至2025年，老子数据系统涵盖了12万对+的抗体抗原配对信息，覆盖了3300+的靶点信息，超过2000+的精确表位信息，超过2万对抗体抗原亲和力的数据，24000+的抗体抗原IC50/EC50数据；

|                                          | Lao Tzu, by Patsnap                                                                                                                                                                                                                                                                                                                                                                                                                                              | Natural Antibody                                                                                                                                                                                                                                                                                                                  | PLabDab                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data source                              | Patent, PDB, literature, NCBI, Synapse                                                                                                                                                                                                                                                                                                                                                                                                                           | INN from WHO, OpenFDA                                                                                                                                                                                                                                                                                                             | Patents and academic papers                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Antibody formats                         | Mainly mAbs (including engineered), but does have some scFV, bispecifics                                                                                                                                                                                                                                                                                                                                                                                         | mAb, Fc-ab, Ab fragment, nanobodies                                                                                                                                                                                                                                                                                               | Mainly mAbs but details undisclosed, and only 0.7% therapeutic                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Target coverage                          | 3300+                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Only 438                                                                                                                                                                                                                                                                                                                          | Undisclosed                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Antibody-Antigen & Epitope Relationships | ✓                                                                                                                                                                                                                                                                                                                                                                                                                                                                | X                                                                                                                                                                                                                                                                                                                                 | X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Affinity                                 | ✓                                                                                                                                                                                                                                                                                                                                                                                                                                                                | X                                                                                                                                                                                                                                                                                                                                 | X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Therapeutic Potential                    | ✓                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Purchase specific module                                                                                                                                                                                                                                                                                                          | X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Clinical Status of Antibodies            | Require customization                                                                                                                                                                                                                                                                                                                                                                                                                                            | Purchase specific module                                                                                                                                                                                                                                                                                                          | X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Additional Comments                      | <ul style="list-style-type: none"><li>✓ Breadth of target coverage for unbiased AI/ML model training</li><li>✓ Validated datasets, with 90%+ precision</li><li>✓ Leverage Synapse's target classification tree for target standardization process</li><li>✓ Data extracted in patent demonstrate real-world reproducibility of antibodies</li><li>✓ Rich fields with standard industry ontologies, e.g. Uniprot ID, Uniprot Seq</li><li>✓ Timey update</li></ul> | <ul style="list-style-type: none"><li>❖ Lack pairing information of light &amp; heavy chain seq and corresponding relationship with antigens, i.e. cannot support in silico modelling</li><li>❖ Lack affinity data – cannot support antibody optimization/ developability study</li><li>❖ Unclear data update frequency</li></ul> | <ul style="list-style-type: none"><li>❖ Skewed and biased dataset as primarily based on literature</li><li>❖ Limited target information and lack of therapeutic antibodies</li><li>❖ Lack epitope and affinity data</li><li>❖ Only provide L&amp;HC pairing but lack antigen – difficult to conduct drug discovery &amp; target validation</li><li>❖ Less quality fields, e.g. lacking sources &amp; source ID, missing Uniprot ID and Uniprot Seq – credibility and traceability issues</li><li>❖ Unclear data update frequency</li></ul> |



智慧芽的老子抗体抗原数据系统与一些商业数据系统或开源数据集相比也具备非常多的优势，尤其体现在智慧芽在数据清洗、数据治理和人工标注与ai相结合方面；例如上图和天然抗体以及PLabDab的对比能看出，老子数据系统在靶点覆盖上远超一些常见数据集；并且在治疗性抗体的数据结构化方面有着传统的优势；利用新药情报库的目标分类树进行目标标准化处理；专利中提取的数据证明了抗体在实际应用中的可重复性；从上图的提取计划看到，2026年和2027年，老子数据系统会继续扩充抗体抗原对、亲和力、表位信息和IC50/EC5；为生成扩散等模型提供丰富、高质量的数据以训练模型和优化抗体。

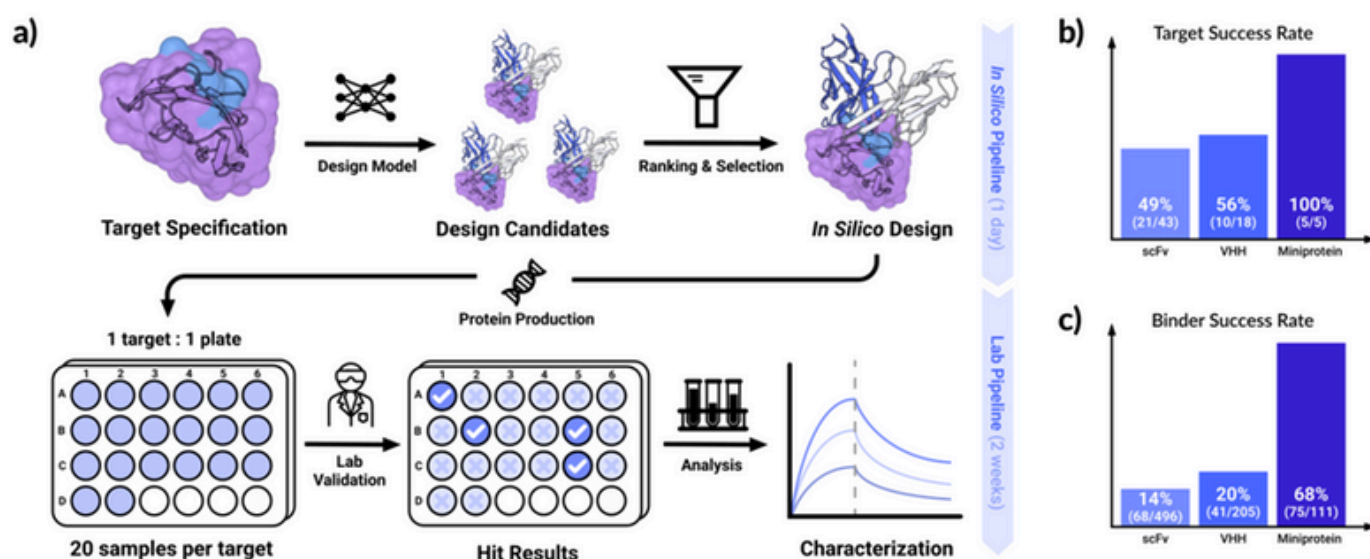
## 第三章



# 全文总结

扩散模型的运用具有诸多优势：首先，它们利用了庞大的数据库，其中包括超过10亿个抗体序列、超过8000个结构注释、以及蛋白质界面环数据，并涵盖了超过8000个结构。其次，扩散模型在捕捉抗体与抗原界面的复杂相互作用方面表现出色，从而简化了抗体的生成过程。此外，这些模型减少了对诸如小鼠杂交瘤和单个B细胞方法等耗时且成本高昂的技术的需求，这些方法既费时又费力。最后，与定向进化可能会将抗体困在不理想的状态不同，扩散模型探索了一个理论上极其广阔的序列空间L20，从而能够全面探索潜在的抗体变异。这些特性使扩散模型成为用于新抗体设计和CDR优化的有前途的工具。

尽管基于扩散的模型在抗体设计方面有着诸多令人鼓舞的应用，但仍有几个局限性需要引起关注。首先，这些模型的准确性并非绝对的；设计出的序列和结构并不总是能达到预期的效果，还需要进一步验证。目前，缺乏既快速又精确的评估这些生成抗体的方法，导致验证成本增加。其次，这些模型的训练主要依赖于阳性数据，而忽略了阴性数据。这种疏忽导致了不合理抗体的产生，这些抗体随后也无法使用。第三，尽管当前的模型主要侧重于提高结合亲和力，但它们往往忽略了其他关键属性，如溶解性、人类相容性和聚集可能性。在未来模型中解决这些属性问题对于提高设计抗体的可用药性至关重要。



Chai-2系列模型能够实现高亲和力蛋白质结合和抗体的设计，成功率领先行业

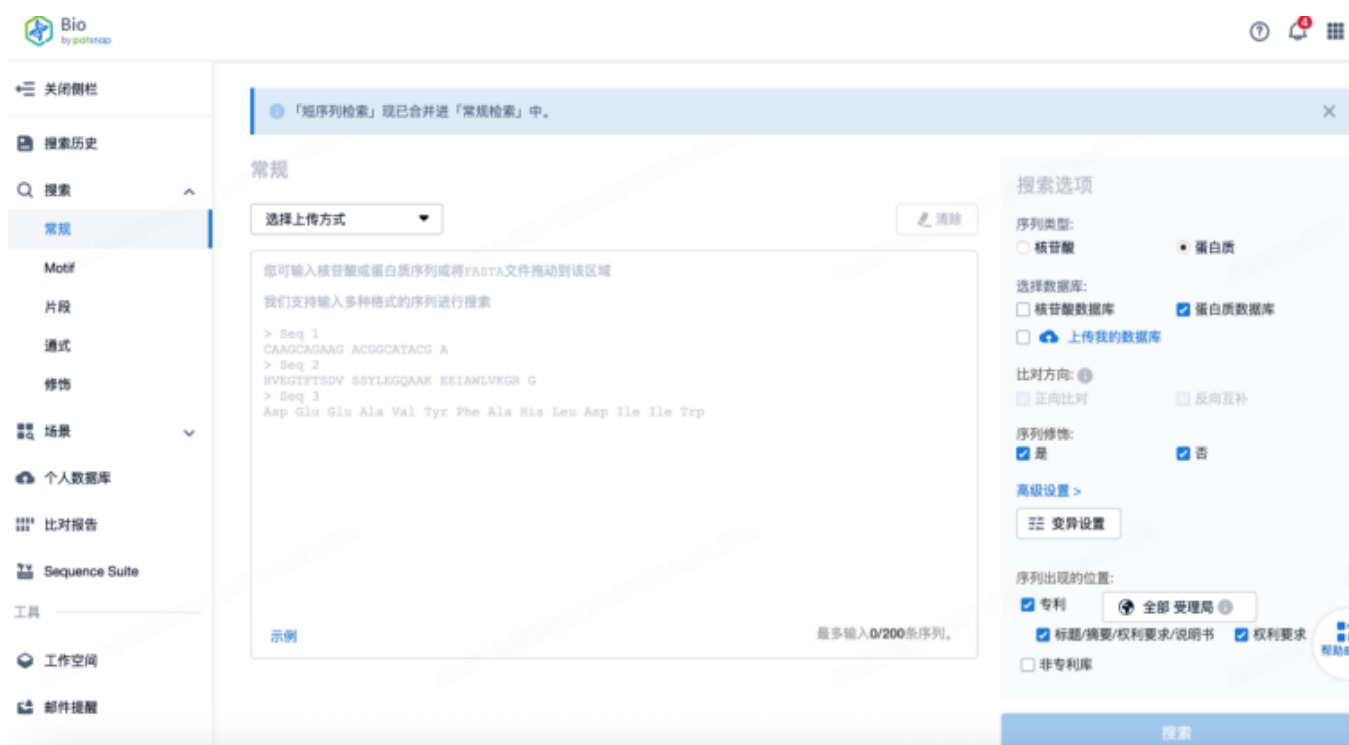
<https://www.biorxiv.org/content/10.1101/2025.07.05.663018v1.full.pdf>

基于扩散的模型的最新进展已被应用于新抗体的开发中。继 AlphaFold2 精准预测蛋白结构之后，由 Chai discovery 团队推出的Chai-2模型更进一步，实现了从零开始生成具备稳定结构和功能位点的蛋白质序列。它不仅能够“设计出抗体”，还被寄予“直接生成药物”的厚望。

然而，AI生成抗体序列 可直接用于药物开发。具备高亲和力只是第一步，真正迈入临床的抗体药物还必须经过重重考验——表达、纯化、免疫原性、体内稳定性等等，这些传统的“瓶颈”，AI暂时还无法一键解决。

因此，尽管前景看好，但基于扩散的抗体设计方法仍处于早期阶段，需要进一步改进才能在药物开发中充分发挥其潜力。

为了在抗体设计中推进基于扩散的模型，可以通过几项策略性的措施改进。首先，引入诸如流匹配或随机插值等新算法，能够优化从初始分布到生成抗体的过渡过程，从而为两个分布之间的关系提供更精确的解释。其次，利用针对广泛抗体数据集进行训练的预训练大型语言模型（LLMs）能够加深模型对抗体特性的理解，从而提升设计过程。最后，专门针对内部数据进行测试的方法能够提供既有益又关键的负向训练数据，从而提高模型的准确性和可靠性——这一点在抗生素研发中已经得到了证实。





智慧芽经过18年的发展，在专利数据和医药情报数据方面有着深厚的积累，近年来随着用户对抗体抗原数据的需求不断增加，智慧芽独创的开发出ai+人工标注的方式，高效高质量的采集了海量的抗体抗原配对信息、抗体抗原亲和力信息、抗体抗原结合的表位信息以及专利中关于此类数据的测试方法和效果数据等；这些高价值的数据对于推动ai设计抗体和优化抗体序列方面有着非常大的提效，用户如果需要多样化的搜寻抗体抗原数据，也可考虑结合使用智慧芽的SaaS产品bio数据库和生物医药相关的agent。

我正在研究TAU靶点。请帮我找出针对这个靶点的最佳抗体。

Hi, 今天你有什么问题? 芽仔可以帮你解决

快速发现针对特定靶点的候选抗体

示例 1

我正在研究TAU靶点。请帮我找出针对这个靶点的最佳抗体。

示例 2

帮我评估这个PCOL1抗体序列。找出相关的候选抗体。 链接: ENLTGSPGTLISLSPGERATLSRASCQVPSGSLAHYQZKPGQAPRLJYDASRATGPPORFSGSGSGTDFLTISRLEPEDFAVYQDQYGS LPWTFGGGQTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVQLLNNFYP REAKVQWKNVNALQSGNSQESVTEQDSKDSITYSLSSITLTKADYE KHKVVACEVTHQGLSSPVTKSFNRGEC 链接: EVQLVESGGGLVQPGGSLRLSCAASGFTFSRYMMSWVRQAPGKG LEWVANIKQDGSSEKYYVDSVNGRFTISRDNAKNSLYLQMNSRAED TATYYCARREGGWFGELAFDYWGQGLTVYSASTKGPSPFPLAPSS KSTISGGTALGCLVKDYFPEPVTVSHHSGALTSGVHTFPAALQSSG LYSLSGTVTPSSSLGTQTYICNNAHKPSNTKVDKRVKPKSQCKTHT CPICPAPEFEGGSPSLFPPKPKDTLMISRTPEVTCVAVDVSHDEPE WFNWYVDGVEVHNAKTKPREEQYNSTYRVSVLTVLDHQLNGK EYKCKVSNKALPASEKTSKAKGQPREPQVYTLPPSREMTKNQVSLTCLVKGFPYSDIAVEWESNGOPENNYKTPPVLDSDGSFFLYSKLT YKSKRWQGGNVFSCSMHEALHNHYTQKSLSPGK

示例 3

请为OX40靶点分析最佳抗体选择。我需要详细的亲和力数据。

我正在研究TAU靶点。请帮我找出针对这个靶点的最佳抗体。

抗体数据库发布

+ 新问题

你知道什么?

4. 优选抗体推荐

推荐抗体详细信息

| 排名 | 抗体名称         | 抗原信息                                          | KD值 (M)   | 亲和力等级     | 亲和力测定条件           | 表位信息                              | 其它关键性能指标                                                   | reference      | 抗体序列                                                                                                                                                    |
|----|--------------|-----------------------------------------------|-----------|-----------|-------------------|-----------------------------------|------------------------------------------------------------|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | mAb A        | 抗原名称: TAU 物种: Homo sapiens Uniprot ID: P10636 | 6.95E-19  | Very High | BLI, pH 7.4, 25°C | -                                 | -                                                          | US2020146535A1 | 链接: Seq Code: 651447996 QSVESGGRLVTPGTPLTLCTVSGLSFSGWYGVHWRQAPGKGLWQVLRAE 链接: Seq Code: 651448000 AQVLQTASPVSATVGGTVTINCQASLAVNNYLAWYQKPGQPKRLYLASS     |
| 2  | B296, PT3    | 抗原名称: TAU 物种: Homo sapiens Uniprot ID: P10636 | <1.60E-11 | High      | SPR               | R211-R221 CN110799531B            | kon: 8.56E+5 1/Ms koff: 2.30E-2 1/s ec50: (4.33±1.39E-10 M | US10766953B2   | 链接: Seq Code: 452073798 QVQLVESGGGVVQPGSRSLRLSCAASGFTFSYMSWVRQAPGKGLWVASISK 链接: Seq Code: 452073799 DIQMTQSPSSLSASVGRVITTCASQGINRYLNWFQKPGKAPKSLYRANRLI |
| 2  | PT3 Fab, PT3 | 抗原名称: TAU 物种: Homo sapiens Uniprot ID: P10636 | <1.60E-11 | High      | SPR               | R211-R221 T212, T217 CN110799531B | -                                                          | US10766953B2   | 链接: Seq Code: 181530 EVQLVESGGDLVQPGGSLKLSCAASGFTFSYAMSWVRQNPKEKLEWVASISK 链接: Seq Code: 181535 DIKMTQSPSSMYASLGERTVITTCASQGINRYLNWFQKPGKSPKTLRYRANRLI   |

此外，在制药公司内部应用隐私保护计算技术也可以进一步丰富用于训练的数据量。基于扩散的方法将继续发展，这有望彻底改变包括小分子和蛋白质设计在内的其他应用的协同进化过程。这种进步有可能极大地提高生物化学合成在各个领域的效率和精度。

## 参考文献

1. Bai and Chungyoun : DOI: 10.7644/j.issn.1674-9960.2024.07.007
2. The Chai-2 series models can achieve high-affinity protein binding and antibody design:  
<https://www.biorxiv.org/content/10.1101/2025.07.05.663018v1.full.pdf>
3. Antibody design with generative diffusion models XH He et al
4. Drug Design, Development and Therapy 2025:19

# 关于智慧芽生物医药

智慧芽生物医药作为行业的创新先锋，精心构建了由Synapse新药情报库、Bio生物序列数据库、Chemical化学结构数据库和Hiro-LS芽仔生物医药四大核心产品组成的综合SaaS产品矩阵，同时提供数据服务串联各维度数据，组成综合数据服务平台。这一平台旨在为生物医药领域的全产业链条提供全面、精确、实时且遵循统一标准的全生命周期数据服务，以支持行业内的科研、开发、生产和商业决策。

通过运用前沿的大数据和人工智能技术，结合生物医药领域专家的深入洞察和严格审核，智慧芽生物医药实现了对产业链数据的高效集成和精准处理。我们的生物医药产品系列已经建立了一个全球范围内的实时数据更新机制，能够从广泛的网络资源中提取关键的原始数据，确保为生物医药行业的各个环节提供全面、深入的数据支持和解决方案。

## 欢迎全世界朋友免费试用！



### 新药情报库

- ✓ 寻找新药研发机会
- ✓ 评估引进项目价值
- ✓ 竞争情报监控
- ✓ 优化产品管线布局

[点击此处试用](#)



### 生物序列数据库

- ✓ 序列FTO检索分析
- ✓ 序列专利动态跟踪
- ✓ 特定序列研发调研
- ✓ 序列化学修饰分析

[点击此处试用](#)



### Eureka LS

- ✓ 抗体抗原筛选
- ✓ 专利推荐PCC分子
- ✓ 生物医药百科问答
- ✓ 药物知产报告

[免费使用入口](#)



### 化学结构数据库

- ✓ 化学结构FTO检索分析
- ✓ 研发专利调研
- ✓ IP尽调及竞对监控
- ✓ 马库什结构专利检索

[点击此处试用](#)

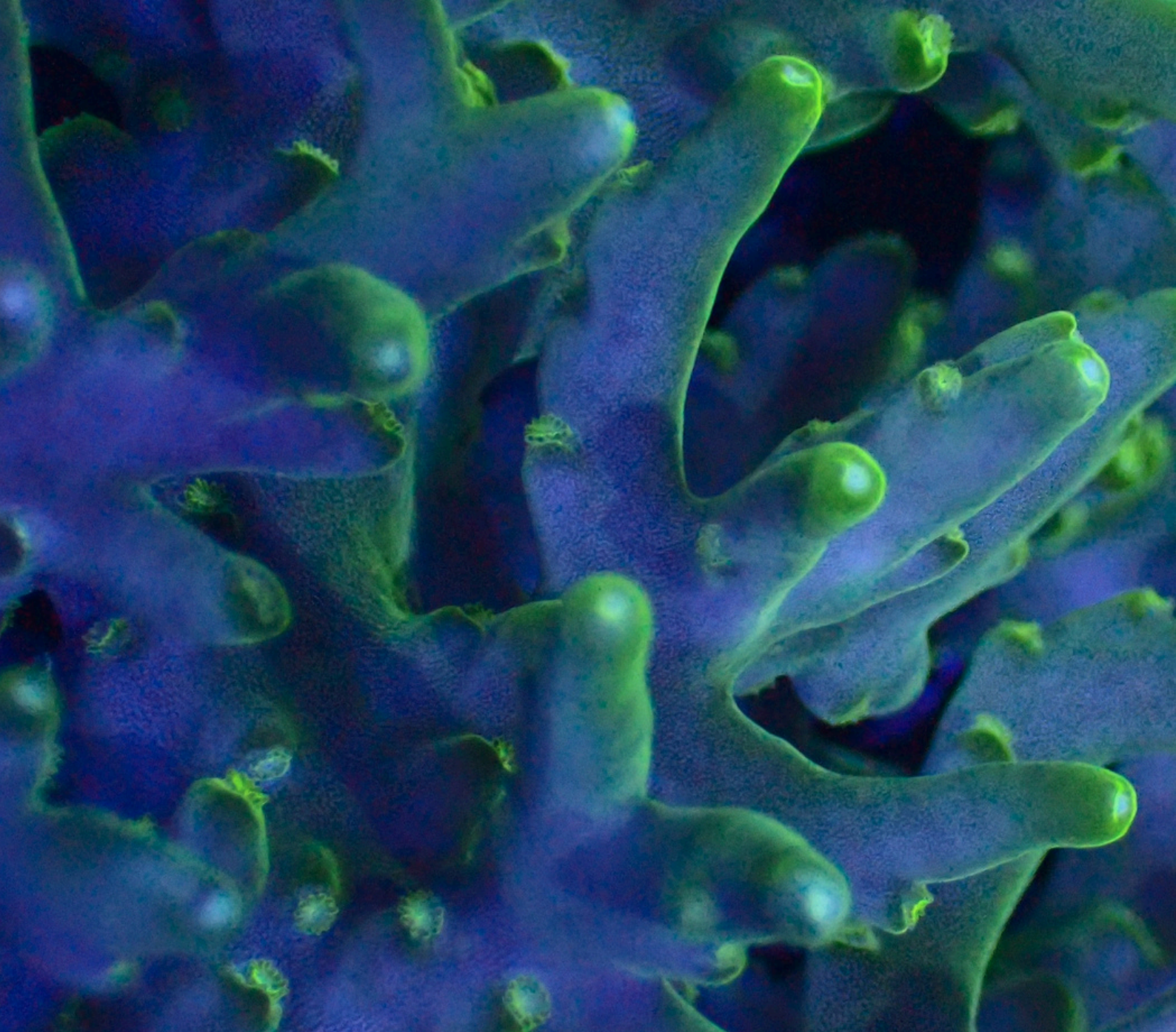


### 个性化数据集服务

- ✓ 药物管线实时监控
- ✓ AI制药模型训练数据
- ✓ 生物医药智能数据治理
- ✓ 科创情报分析平台

[点击查看详情](#)





智慧芽是AI驱动的科技创新和知识产权信息服务商，在全球范围内为知识产权、研发创新和科创信息服务等创新场景为客户提供AI智能体（Agent）、软件工具、数据服务及AI综合解决方案。

基于覆盖全球20大类、超20亿条高质量创新数据构建的行业知识图谱，结合领先的大模型、机器学习、计算机视觉和自然语言处理（NLP）技术，智慧芽打造了“更懂行业Know-How的AI中台”。依托强大的数据治理能力和算法体系，智慧芽能够输出深度、可信赖、低幻觉的专业报告与洞察，帮助客户在专利分析、技术布局、研发决策等场景实现高效创新。

公司成立于2007年，目前在全球拥有超15000家客户，覆盖生物医药、新材料、智能制造、新能源汽车、通信、半导体、高校和科研院所等领域。客户包括中国石化、海尔、美的、小米、宁德时代、小鹏汽车、大疆、药明康德、商汤科技、清华大学、北京大学、中科院、麻省理工学院、牛津大学、陶氏化学、Spotify等。

