

文章

AlphaFold 3 对生物分子相互作用的精准结构预测

<https://doi.org/10.1038/s41586-024-07487-w>

检查更新

乔什·阿布拉姆森^{1,7}, 乔纳斯·阿德勒^{1,7}, 杰克·邓杰^{1,7}, 理查德·埃文斯^{1,7}, 蒂姆·格林^{1,7}, 亚历山大·普里策尔^{1,7}, 奥拉夫·罗内伯格^{1,7}, 林赛·威尔莫尔^{1,7}, 安德鲁·J. 鲍德¹, 约书亚·班布里克², 塞巴斯蒂安·W·博德斯坦¹, 大卫·A·埃文斯¹, 洪嘉骏², 迈克尔·奥尼尔¹, 大卫·雷曼¹, 凯瑟琳·图尼娅苏瓦纳库尔¹, 扎卡里·吴¹, 阿克维莱·泽姆古利特¹, 埃里尼·阿瓦纳蒂³, 查尔斯·贝蒂³, 奥塔维亚·贝尔托利³, 亚历克斯·布里格兰³, 阿列克谢·切列帕诺夫⁴, 迈尔斯·康格里夫⁴, 亚历山大·L·科文-里弗斯³, 安德鲁·考伊³, 迈克尔·菲格诺夫³, 法比安·B·富克斯³, 汉娜·格拉德曼³, 里舒布·贾恩³, 优素福·A·汗^{3,5}, 卡罗琳·M.R·洛⁴, 库巴·佩尔林³, 安娜·波塔彭科³, 帕斯卡尔·萨维⁴, 苏克德夫·辛格³, 阿德里安·斯特库拉⁴, 阿肖克·蒂莱苏达姆³, 凯瑟琳·通⁴, 谢尔盖·亚克尼恩⁴, 埃伦·D·钟^{3,6}, 米哈尔·杰林斯基³, 奥古斯丁·日德克³, 维克托·巴普斯特^{1,8}, 普什米特·科利^{1,8}, 马克斯·贾德伯格^{2,8}✉, 德米斯·哈萨比斯^{1,2,8}✉ 和 约翰·M·贾珀^{1,8}✉

AlphaFold 2¹的引入在蛋白质及其相互作用结构的建模领域引发了一场革命，推动了蛋白质建模与设计^{2–6}的广泛应用。在此，我们介绍了我们的AlphaFold 3模型，该模型采用了显著更新的基于扩散的架构，能够预测包含蛋白质、核酸、小分子、离子及修饰残基的复合物的联合结构。新版AlphaFold模型在多个方面显著超越了以往的专业工具：在蛋白质-配体相互作用方面，其准确性远超最先进的对接工具；在蛋白质-核酸相互作用方面，其准确性大大高于专门的核酸预测工具；在抗体-抗原预测方面，其准确性显著优于AlphaFold-Multimer v.2.3^{7,8}。这些结果共同表明，在单一统一的深度学习框架内实现生物分子空间的高精度建模是可能的。

生物复杂体的精确模型对于我们理解细胞功能以及合理设计治疗方案^{2–4,9}至关重要。随着AlphaFold¹的发展，蛋白质结构预测取得了巨大进展，该领域也随着基于AlphaFold 2 (AF2)^{10–12}理念和技术的一系列后续方法的出现而迅速扩展。AlphaFold一经推出，便有研究表明，简单的输入修改即可实现令人惊讶的蛋白质相互作用预测精度^{13–15}，并且专门针对蛋白质相互作用预测训练的AF2系统表现出了极高的准确性⁷。

这些成功引发了一个问题，即是否有可能在一个深度学习框架内准确预测包含更广泛生物分子（包括配体、离子、核酸和修饰残基）的复合物结构。已经开发了多种针对特定相互作用类型的预测器^{16–28}，以及一种与本工作同时开发的通用方法²⁹，但这些深度学习尝试的准确性参差不齐，且往往低于基于物理启发的方法^{30,31}。几乎所有这些方法都高度专用于特定的相互作用类型，无法预测包含多种实体的一般生物分子复合物的结构。

我们在这里介绍AlphaFold 3 (AF3) ——一种能够高精度预测包含几乎所有存在于蛋白质数据库³² (PDB) 中的分子类型的复合物的模型 (图 1a,b)。在除一类以外的所有类别中，它的表现显著优于专门针对该任务的强大方法 (图 1c 和 扩展数据表 1)，包括在蛋白质结构和蛋白质-蛋白质相互作用结构方面更高的准确性。

这是通过AF2架构和训练过程的重大改进 (图 1d) 实现的，既适应了更广泛的化学结构，又提高了学习的数据效率。该系统通过用更简单的pairformer模块 (图 2a) 替换AF2的evoformer，减少了多序列比对 (MSA) 的处理量。此外，它直接利用扩散模块预测原始原子坐标，取代了在氨基酸特定框架和侧链扭转角上操作的AF2结构模块 (图 2b)。扩散过程的多尺度特性 (低噪声水平促使网络改进局部结构) 也使我们能够消除立体化学损失和网络中大部分对成键模式的特殊处理，轻松适应任意化学成分。

¹核心贡献者，Google DeepMind，伦敦，英国。²核心贡献者，Isomorphic Labs，伦敦，英国。³Google DeepMind，伦敦，英国。⁴Isomorphic Labs，伦敦，英国。⁵分子系和细胞生理学，斯坦福大学，斯坦福，加利福尼亚州，美国。⁶计算机科学系，普林斯顿大学，普林斯顿，新泽西州，美国。⁷这些作者贡献相同：乔什·亚伯拉罕森，Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore。⁸这些作者共同监督了这项工作：Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, John M. Jumper. ✉电子邮件：jaderberg@isomorphilabs.com; dhcontact@google.com; jumper@google.com

文章

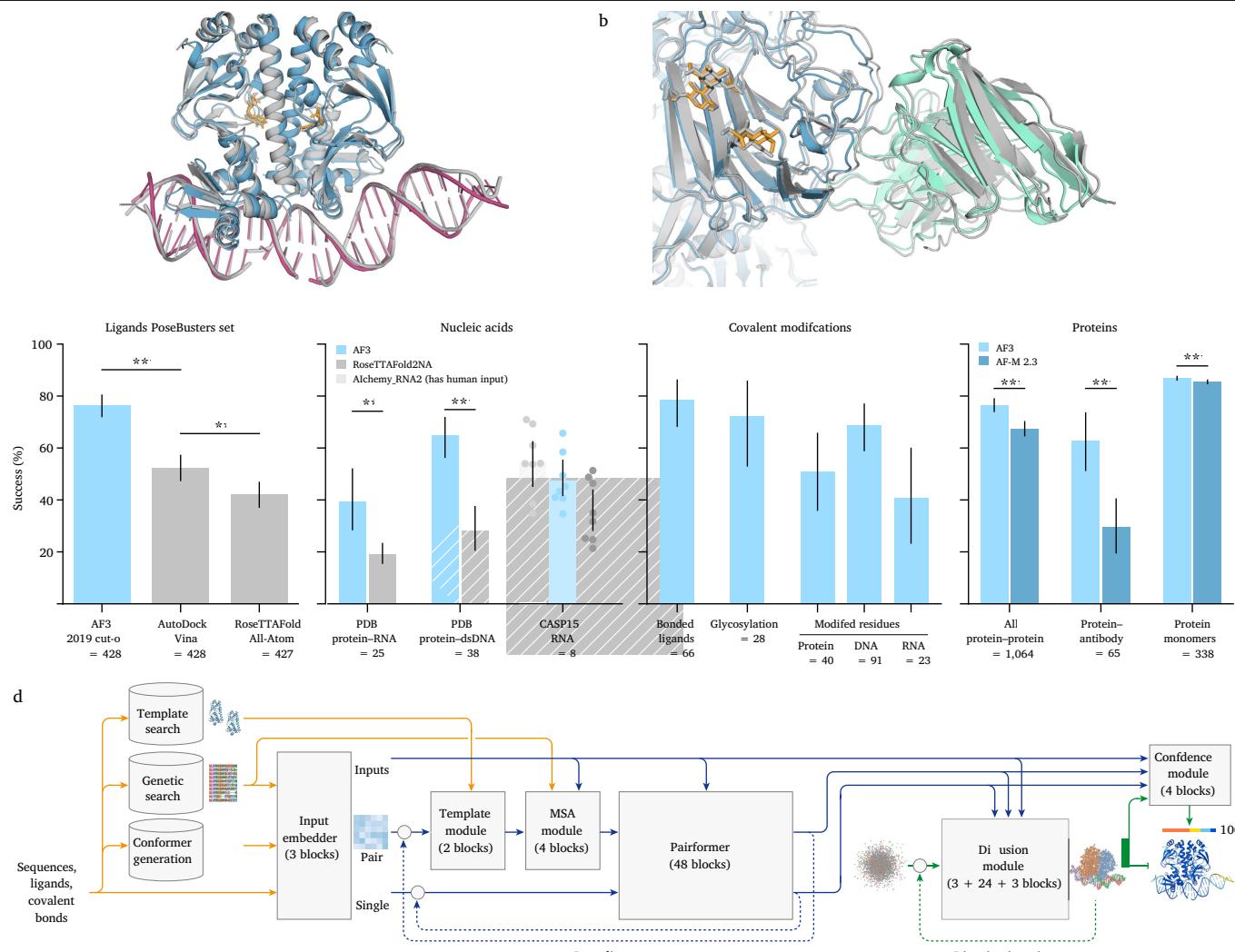


图1 | AF3 准确预测了生物分子复合物的结构。

a,b, 使用 AF3 预测的示例结构。a, 细菌 CRP/FNR 家族转录调控蛋白与 DNA 和 cGMP 结合 (PDB 7PZB; 全复合物 LDDT⁴⁷, 82.8; 局部距离测试 (GDT)⁴⁸, 90.1)。b, 人冠状病毒 OC43 刺突蛋白, 4,665 个残基, 高度糖基化并被中和抗体结合 (PDB 7PNM; 全复合物 LDDT, 83.0; GDT, 83.1)。c, AF3 在 PoseBusters (v.1, 2023 年 8 月发布)、我们最近的 PDB 评估集和 CASP15 RNA 上的表现。指标如下: 口袋对齐配体 r.m.s.d. < 2 Å 的百分比, 适用于配体和共价修饰; 蛋白质-核酸复合物的界面 LDDT; 核酸和蛋白质单体的 LDDT; 以及蛋白质-蛋白质和蛋白质-抗体界面的 DockQ > 0.23 的百分比。所有分数均来自五个模型种子 (每个种子有五个扩散样本) 中的最高置信度样本, 除了蛋白质-抗体分数, 这些分数在两个模型 (每个 AF3 种子有五个扩散样本) 的 1,000 个模型种子中排名。采样

网络架构与训练

AF3的整体结构(图1d和补充方法3)与AF2相似, 主干部分演化出化学复合物的成对表示, 随后是一个结构模块, 该模块利用成对表示来生成明确的原子位置, 但在每个主要组件中存在显著差异。这些修改既是为了适应广泛的化学实体而无需过多的特殊处理, 也是基于对AF2在不同修改下的性能观察。在主干部分, MSA处理的重要性大大降低, 采用了更小且更简单的MSA嵌入块(补充方法3.3)。

排名细节在方法中提供。对于配体, n 表示目标数量; 对于核酸, n 表示结构数量; 对于修饰, n 表示聚类数量; 对于蛋白质, n 表示聚类数量。条形高度表示平均值; 误差条表示 PoseBusters 的精确二项分布 95% 置信区间, 其他所有数据通过 10,000 次自举重采样获得。显著性水平使用 PoseBusters 的双侧 Fisher 精确检验和其他所有数据的双侧 Wilcoxon 符号秩检验计算; ***P < 0.001, **P < 0.01。精确的 P 值 (从左到右) 如下: 2.27×10^{-13} , 2.57×10^{-3} , 2.78×10^{-3} , 7.28×10^{-12} , 1.81×10^{-18} , 6.54×10^{-5} 和 1.74×10^{-34} 。AF-M 2.3, AlphaFold-Multimer v.2.3; dsDNA, 双链 DNA。d, 用于推理的 AF3 架构。矩形代表处理模块, 箭头显示数据流。黄色, 输入数据; 蓝色, 抽象网络激活; 绿色, 输出数据。彩色球体代表物理原子坐标。

与AF2的原始evoformer相比, 块的数量减少到四个, MSA表示的处理采用了廉价的成对加权平均, 并且仅使用成对表示进行后续处理步骤。'pairformer' (图2a和补充方法3.6)取代了AF2的evoformer作为主要的处理块。它仅在成对表示和单个表示上操作; MSA表示未被保留, 所有信息都通过成对表示传递。成对处理和块的数量 (48个) 与AF2基本保持不变。生成的成对表示和单个表示与输入表示一起传递给新的扩散模块 (图2b), 该模块取代了AF2的结构模块。

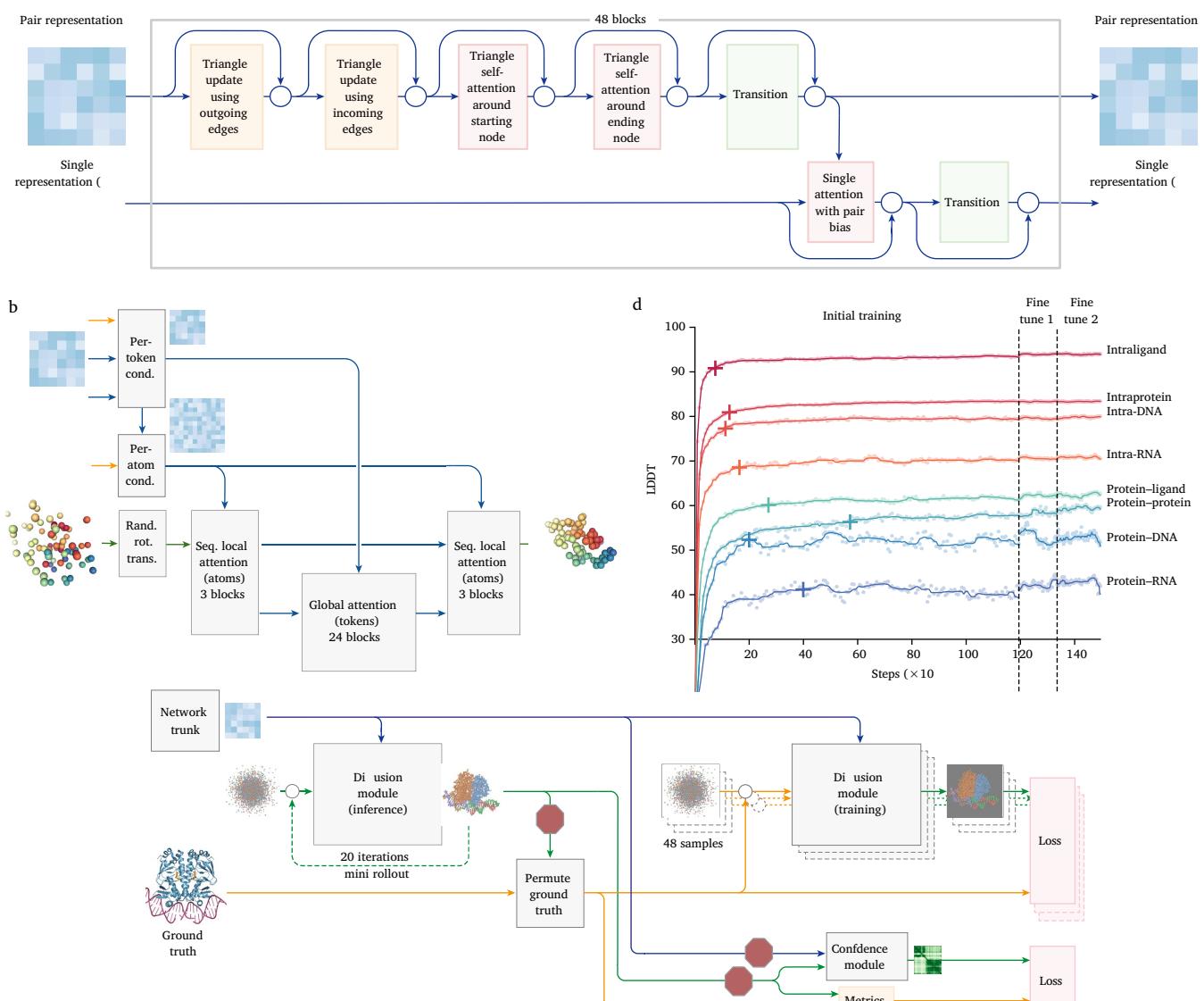


图2 | 架构和训练细节。a, Pairformer模块。

输入和输出：维度为(n, n, c)的成对表示和维度为(n, c)的单个表示。n是token的数量（聚合物残基和原子）；c是通道数（成对表示为128，单个表示为384）。每个48个块都有一组独立的可训练参数。b, 扩散模块。输入：粗略数组描绘了每个token的表示（绿色，输入；蓝色，成对；红色，单个）。精细数组描绘了每个原子的表示。彩色球体代表物理原子坐标。Cond., 条件；rand. rot. trans., 随机旋转和变换；seq., 序列。c, 训练设置（省略了距离图头）

扩散模块（图2b和补充方法3.7）直接操作于原始原子坐标和粗略抽象的标记表示，无需旋转框架或任何等变处理。我们在AF2中发现，去除结构模块的大部分复杂性仅对预测精度产生适度影响，而维持骨架框架和侧链扭转表示则对通用分子图增加了相当多的复杂性。同样，AF2在训练过程中需要精心调整立体化学违规惩罚，以确保生成结构的化学合理性。我们采用了一种相对标准的扩散方法³³，其中扩散模型被训练来接收“噪声化”的原子坐标，然后预测真实坐标。此任务要求

从网络主干末端开始。彩色数组展示了来自网络主干的激活情况（绿色，输入；蓝色，成对；红色，单个）。蓝色箭头表示抽象激活数组。黄色箭头表示真实数据。绿色箭头表示预测数据。停止标志代表梯度的停止。所描绘的两个扩散模块共享权重。d, 初始训练和微调阶段的训练曲线，显示了在我们的评估集上作为优化器步数函数的LDDT。散点图展示了原始数据点，线条则展示了使用宽度为九个数据点的中值滤波器平滑后的性能。十字标记表示平滑性能达到初始训练最大值的97%时的点。

网络通过在多种长度尺度上学习蛋白质结构，其中在小噪声情况下的去噪任务强调对非常局部立体化学的理解，而在高噪声情况下的去噪任务则强调系统的大尺度结构。在推理时，随机噪声被采样，然后递归去噪以生成最终结构。重要的是，这是一种生成训练过程，产生的是答案的分布。这意味着，对于每个答案，局部结构将被明确定义（例如，侧链键几何），即使网络对位置不确定。因此，我们能够避免基于扭转的残基参数化以及结构上的违规损失，同时处理一般配体的全部复杂性。与一些最近的³⁴工作类似，

文章

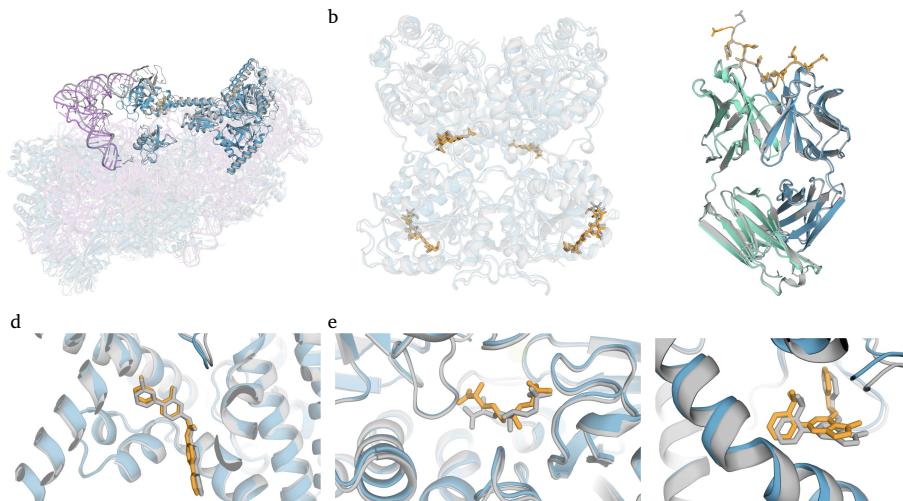


图3 | 预测复合物的示例。选定的EXT3同源二聚体 (PDB 7AU2；平均口袋对齐的均方根偏差, 1.10 Å) 的结构预测。预测的蛋白质链显示为蓝色 (预测的抗体在c中, 间皮素C端肽与单克隆抗体15B6结合, 绿色), 预测的配体和糖类为橙色, 预测的RNA为紫色, (PDB 7U8C; DockQ, 0.85)。d, 临床阶段的抑制剂LGK74与PORCN结合 (PDB 7URD; 配体均方根偏差, 1.00 Å)。e, (5S,6S)-O7-碘基DADH与AziU3/U2复合物结合, 具有新颖的折叠 (PDB 7WUX; 配体均方根偏差, 1.92 Å)。f, NIH-12848的类似物与PI5P4K 的别构位点结合 (PDB 7QIE; 配体均方根偏差, 0.37 Å)。a, 人类40S核糖体亚基 (7,663个残基), 包括18S核糖体RNA和Met-tRNaiMet (不透明紫色), 与翻译起始因子eIF1A和eIF5B (不透明蓝色; PDB 7TQL; 全复合物LDDT, 87.7; GDT, 86.9) 形成复合物。b, 糖基化的球状部分。

我们发现, 在架构中不需要对分子的全局旋转和平移保持不变性或等变性, 因此我们省略了这些特性以简化机器学习架构。

生成扩散方法的使用带来了一些技术挑战, 这些是我们需要解决的。最大的问题是生成模型容易产生幻觉³⁵, 即模型可能在无结构的区域中创造出看似合理的结构。为了对抗这种效应, 我们采用了交叉蒸馏方法, 其中我们用AlphaFold-Multimer (v.2.3) ^{7,8}预测的结构来丰富训练数据。在这些结构中, 无结构区域通常由长延伸环表示, 而不是紧凑结构, 对这些结构进行训练“教会”了AF3模仿这种行为。这种交叉蒸馏显著减少了AF3的幻觉行为 (扩展数据图1展示了在CAID 2³⁶基准集上的无序预测结果)。

我们还开发了置信度度量方法, 用于预测我们最终结构中的原子级误差和成对误差。在AF2中, 这是通过在训练过程中回归结构模块输出误差来直接实现的。然而, 这种方法不适用于扩散训练, 因为扩散训练仅训练单一步骤而非完整结构生成 (图2c)。为了弥补这一点, 我们开发了一种扩散“展开”过程, 用于在训练期间进行完整结构预测生成 (使用比正常情况更大的步长; 图2c (小型展开))。然后, 利用这一预测结构来排列对称的真实链和配体, 并计算性能指标以训练置信度头部。置信度头部利用成对表示来预测修改后的局部距离差异测试 (pLDDT) 和预测对齐误差 (PAE) 矩阵, 如同AF2中一样, 以及一个距离误差矩阵 (PDE), 该矩阵表示预测结构与真实结构之间距离矩阵的误差 (详细信息见补充方法4.3)。

图2d显示, 在初始训练阶段, 模型迅速学会预测局部结构 (所有链内指标快速上升, 并在前20,000个训练步骤内达到最大性能的97%), 而模型需要更长时间来学习全局排列 (界面指标上升缓慢, 蛋白质-蛋白质界面LDDT在60,000步后才达到97%的门槛)。在AF3开发过程中, 我们观察到某些模型能力达到上限。

相对较早地开始下降 (很可能是由于对该能力的有限训练样本过度拟合), 而其他能力仍然训练不足。我们通过增加或减少相应训练集的采样概率 (补充方法2.5.1), 并使用上述所有指标和一些附加指标的加权平均值进行早期停止, 以选择最佳模型检查点 (补充表7) 来解决这一问题。具有较大裁剪尺寸的微调阶段在所有指标上均提升了模型性能, 特别是在蛋白质-蛋白质界面上有显著提升 (扩展数据图2)。

复杂类型中的准确性

AF3能够根据输入的聚合物序列、残基修饰和配体SMILES (简化分子输入线输入系统) 预测结构。在图3中, 我们展示了一些示例, 突显了该模型在多种生物学重要和治疗相关的模式中的泛化能力。在选择这些示例时, 我们考虑了单链和界面与训练集的相似性方面的新颖性 (补充方法8.1提供了额外信息)。

我们对系统在每种复杂类型的最新界面特定基准上的性能进行了评估 (图1c和扩展数据表1)。蛋白质-配体界面的性能评估基于PoseBusters基准集, 该基准集由2021年或之后发布到PDB的428个蛋白质-配体结构组成。由于我们的标准训练截止日期是2021年, 我们训练了一个单独的AF3模型, 其训练集截止日期较早 (方法)。PoseBusters集的准确性报告为蛋白质-配体对中配体口袋对齐的均方根偏差 (r.m.s.d.) 小于2 Å的百分比。基准模型分为两类: 仅使用蛋白质序列和配体SMILES作为输入的模型, 以及额外泄露已解决的蛋白质-配体测试结构信息的模型。传统的对接方法使用后者特权信息, 尽管在实际应用中这些信息不可用。即便如此, AF3在不使用任何结构输入的情况下, 仍大大优于Vina^{37,38}等经典对接工具 (Fisher's精确检验, $P = 2.27 \times 10^{-13}$), 并且大大优于所有其他真正的盲对接方法。

类似于RoseTTAFold全原子模型 ($P = 4.45 \times 10^{-25}$)。扩展数据图3展示了三个例子，其中AF3实现了准确的预测，而对接工具Vina和Gold未能做到³⁷。PoseBusters分析使用2019年9月30日的训练截止日期进行，以确保模型未在任何PoseBusters结构上进行训练。为了与RoseTTAFold全原子结果进行比较，我们使用了PoseBusters版本1。版本2（从基准集中移除了晶体接触）的结果及其质量指标显示在扩展数据图4b-f和扩展数据表1中。我们使用多个种子来确保正确的手性和避免轻微的蛋白质-配体碰撞（与使用扩散引导等方法相反），但我们通常能够生成高质量的立体化学。此外，我们还训练了一个AF3版本，该版本接收了“口袋信息”，正如近期一些深度学习工作中所使用的^{24,26}（结果显示在扩展数据图4a中）。

AF3在预测蛋白质-核酸复合物和RNA结构方面，比

RoseTTAFold2NA¹⁵（图1c（第二个图））具有更高的准确性。由于RoseTTAFold2NA仅在低于1,000个残基的结构上得到验证，我们仅使用我们最近的PDB评估集中低于1,000个残基的结构进行此比较（方法）。AF3能够预测含有数千个残基的蛋白质-核酸结构，其中一个例子如图3a所示。需要注意的是，我们并未直接与RoseTTAFold全原子模型进行比较，但基准测试表明，在核酸预测方面，RoseTTAFold全原子模型的准确性略低于RoseTTAFold2NA²⁹。

我们还评估了AF3在十个公开的结构预测关键评估15（CASP15）RNA目标上的表现：在与RoseTTAFold2NA和AIchemy_RNA²⁷（CASP15中表现最佳的AI提交^{18,31}）的共同预测子集上，我们的平均表现更高（详细结果见扩展数据图5a）。我们未能达到CASP15中最佳人类专家辅助提交AIchemy_RNA²⁹的表现（图1c（左中））。由于数据集规模有限，我们在此不报告显著性检验统计数据。进一步分析仅预测核酸（不包括蛋白质）的准确性，结果见扩展数据图5b。

共价修饰（结合配体、糖基化以及蛋白质和核酸碱基的修饰）也能被AF3准确预测（图1c（右中））。这些修饰涉及任何聚合物残基（蛋白质、RNA或DNA）。我们将准确性报告为成功预测的百分比（口袋r.m.s.d. < 2 Å）。我们对结合配体和糖基化数据集应用了质量过滤器（与PoseBusters的做法相同）：我们仅包含具有高质量实验数据的配体（根据RCSB结构验证报告，ranking_model_fit > 0.5，即模型质量高于中位数的X射线结构）。与PoseBusters数据集一样，结合配体和糖基化数据集未根据与训练数据集的同源性进行过滤。基于结合聚合物链的同源性过滤（使用聚合物模板相似性 < 40%）仅产生了五个结合配体的簇和七个糖基化的簇。我们在这里排除了多残基糖类，因为RCSB验证报告未提供它们的ranking_model_fit值。在所有质量的实验数据上，多残基糖类的成功预测百分比（口袋r.m.s.d. < 2 Å）为42.1%（n = 131个簇），略低于所有质量实验数据上单残基糖类的成功率46.1%（n = 167个簇）。修饰残基数据集的过滤方式与我们其他聚合物测试集类似：它仅包含与训练集同源性低的聚合物链中的修饰残基（方法）。详见扩展数据表1的详细结果，以及扩展数据图6中预测的带有共价修饰的蛋白质、DNA和RNA结构示例，包括磷酸化对预测影响的分析。

在扩展建模能力的同时，AF3在蛋白质复合物准确性方面也相对于AlphaFold-Multimer（v.2.3）^{7,8}有所提高。总体而言，蛋白质-蛋白质预测成功率（DockQ > 0.23）⁴⁰有所提升（配对Wilcoxon符号秩检验， $P = 1.8 \times 10^{-18}$ ），

抗体-蛋白质相互作用预测方面，尤其是显示出显著的改进（图1c（右）；配对Wilcoxon符号秩检验， $P = 6.5 \times 10^{-5}$ ，预测从1000个而非典型的5个种子中排名最高；更多细节见图5a）。蛋白质单体LDDT的改进也具有显著性（配对Wilcoxon符号秩检验， $P = 1.7 \times 10^{-34}$ ）。AF3对MSA深度的依赖性与AlphaFold-Multimer v.2.3非常相似；具有浅层MSA的蛋白质预测精度较低（单链LDDT对MSA深度的依赖性比较见扩展数据图7a）。

预测的置信度跟踪准确性

与AF2一样，AF3的置信度指标与准确度校准良好。

我们的置信度分析基于最近的PDB评估集，未进行同源性过滤，并包括了多肽。配体类别经过筛选，仅保留高质量的实验结构，并且只考虑标准非键合配体。关于键合配体和其他界面的类似评估，请参见扩展数据图8。所有统计数据均按簇加权（方法），并仅考虑排名最高的预测（排名详情见补充方法5.9.3）。

在图4a（上排）中，我们绘制了链对界面预测的跨膜（ipTM）得分⁴¹（补充方法5.9.1）与界面准确性指标的关系：蛋白质-蛋白质DockQ、蛋白质-核酸界面LDDT（iLDDT）和蛋白质-配体成功率，其中成功率定义为在阈值口袋对齐均方根偏差值下的样本百分比。在图4a（下排）中，我们绘制了每个蛋白质、核苷酸或配体实体的平均pLDDT与我们定制的LDDT_to_polymer指标（方法中提供了指标详细信息）的关系，该指标与pLDDT预测器的训练目标密切相关。

在图4b-e中，我们突出展示了7T82的单一预测示例，其中每个原子的pLDDT着色识别出不自信的链尾、有些自信的界面以及其余自信的二级结构。在图4c中，同样的预测按链进行着色，同时图4d中显示了DockQ界面评分，并在轴上展示了每条链的着色以供参考。从图4e中我们可以看到，对于DockQ > 0.7的粉色-灰色和蓝色-橙色残基对，PAE置信度较高，而对于DockQ < 0的粉色-橙色和粉色-蓝色残基对，PAE置信度最低。扩展数据图5c,d展示了包含蛋白质和核酸链的示例的类似PAE分析。

模型局限性

我们注意到AF3在立体化学、幻觉、动态性和某些目标的准确性方面存在模型局限性。

在立体化学方面，我们注意到两类主要的违规情况。第一类是模型输出并不总是遵循手性（图5b），尽管模型接收的参考结构具有正确的手性作为输入特征。为了解决这个问题，在PoseBusters基准测试中，我们在模型预测的排名公式中加入了手性违规的惩罚。尽管如此，我们仍然在基准测试中观察到4.4%的手性违规率。第二类立体化学违规是模型偶尔会在预测中产生重叠（冲突）的原子。这种情况有时表现为同源体中的极端违规，其中整个链被观察到重叠（图5e）。在排名过程中对冲突进行惩罚（补充方法5.9.3）减少了这种失败模式的发生，但并未完全消除。几乎所有剩余的冲突都发生在蛋白质-核酸复合物中，这些复合物中核酸超过100个，总残基数超过2000个。

我们注意到，从非生成型的AF2模型切换到基于扩散的AF3模型引入了在无序区域中产生虚假结构顺序（幻觉）的挑战（图5d和扩展数据图1）。尽管幻觉区域通常被标记为非常低的置信度，但它们可能缺乏明显的带状外观。

文章

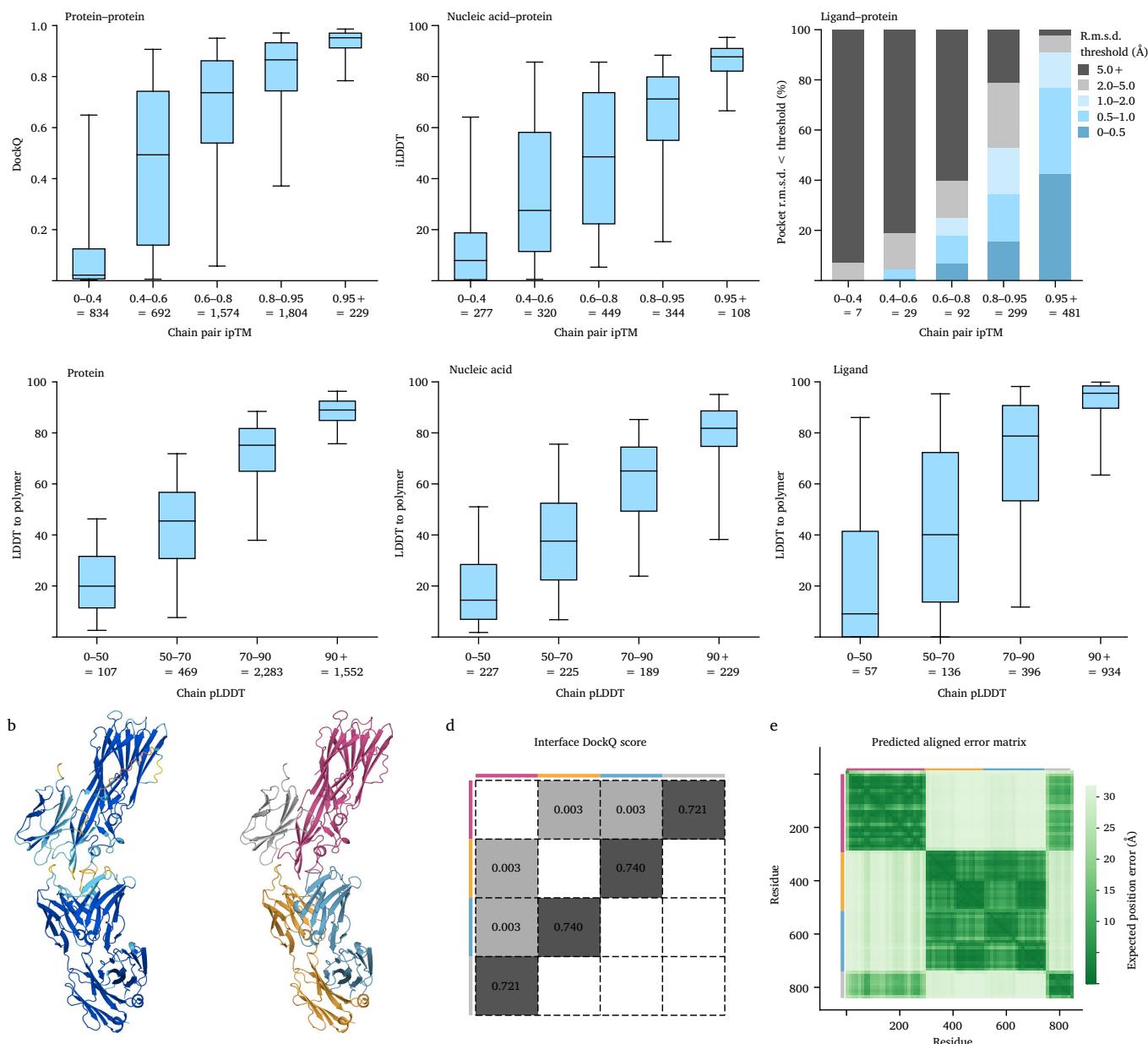


图4 | AF3 置信度追踪准确性。a. 蛋白质-蛋白质界面准确性随链对 ipTM 的变化 (顶部)。底部，LDDT 到聚合物的准确性根据链平均 pLDDT 评估不同链类型的变化。箱线图显示了 25-75% 的置信区间 (箱限)、中位数 (中心线) 和 5-95% 的置信区间 (须)。n 值报告了每个带中的簇数。b. PDB 预测结构的预测结构。c. 根据链着色的相同预测。d. 蛋白质-蛋白质界面的 DockQ 分数。e. 相同预测的 PAE 矩阵 (越暗越自信)，侧边栏有 c 的链着色。虚线黑线表示链边界。

AF2在无序区域生成的结果。为了在AF3中鼓励带状的预测，我们使用从AF2预测中提取的蒸馏训练，并添加了一个排序项，以鼓励生成更多溶剂可及表面积的结果³⁶。

蛋白质结构预测模型的一个主要局限性在于，它们通常预测的是在PDB中观察到的静态结构，而不是溶液中生物分子系统的动态行为。这一局限性在AF3中也存在，其中无论是扩散头还是整个网络的多个随机种子都无法产生溶液集合的近似。

在某些情况下，根据指定的配体和其他输入，建模的构象状态可能不正确或不全面。

例如，E3泛素连接酶在apo状态下自然地采用开放构象，并且只有在结合配体时才观察到闭合状态，但AF3专门预测apo和holo系统的闭合状态⁴²(图5c)。已经开发了许多方法，特别是在MSA重采样方面，这些方法有助于从之前的AlphaFold模型中生成多样性⁴³⁻⁴⁵，并且也可能有助于AF3的多状态预测。

尽管AF3在模型精度上取得了显著进步，但仍有许多目标的准确建模颇具挑战性。为了达到最高精度，可能需要生成大量预测并对其进行排序，这会带来额外的计算成本。我们观察到这一效应的目标类别

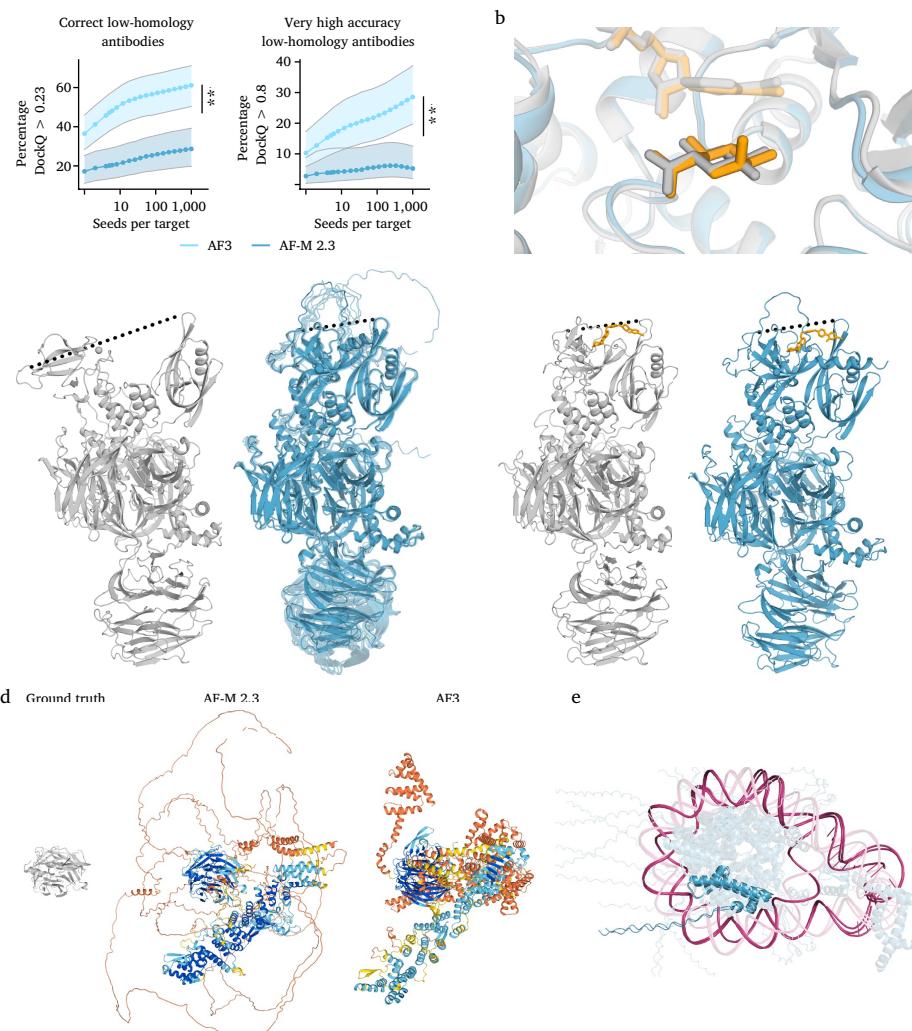


图5 | 模型局限性。a, 抗体预测质量随模型种子数量的增加而提高。排名靠前的低同源性抗体-抗原界面预测质量随种子数量的变化。每个数据点显示了在1,200个种子中，通过1,000次随机抽样（有放回）对种子进行排序的平均值。置信区间为每个数据点上对10,000次重采样集群分数的95%自举法。每个界面的样本按蛋白质-蛋白质ipTM排序。使用双侧Wilcoxon符号秩检验进行显著性检验。 $n = 65$ 个集群。确切的P值如下： 2.0×10^{-5} （正确百分比）和 $P = 0.009$ （极高准确率百分比）。b, Thermotoga maritima -d-葡萄糖苷酶和-d-葡萄糖醛酸的预测（彩色）和真实（灰色）结构——来自PoseBusters集的目标（PDB: 7CTM）。AF3预测-d-葡萄糖醛酸；不同的手性中心用星号标出。显示的预测是按配体-蛋白质ipTM排序的最高排名预测。

强烈的是抗体-抗原复合物，类似于其他近期的工作⁴⁶。

图5a显示，对于AF3，随着模型种子的增加，排名靠前的预测结果持续改进，即使在多达1,000个种子的情况下（5和1,000个种子之间的Wilcoxon符号秩检验， $P = 2.0 \times 10^{-5}$ 对于正确百分比， $P = 0.009$ 对于非常高准确率的百分比；按蛋白质-蛋白质界面ipTM排序）。这种随着多种子的显著改进在其他分子类别中并不常见（扩展数据图7b）。仅使用每个模型种子的一个扩散样本进行AF3预测，而不是五个（未展示），结果并没有显著变化，这表明运行更多模型种子对于提高抗体评分是必要的，而不仅仅是增加扩散样本。

手性（chirality）和碰撞惩罚（clash penalty）。c, 构象覆盖范围有限。cereblon在开放（apo, PDB: 8CVP；左）和闭合（holo, 与mezigdomide结合, PDB: 8D7U；右）构象中的真实结构（灰色）。预测结果（蓝色）显示，无论是apo（有10个重叠样本）还是holo结构，都处于闭合构象。虚线表示N端Lon蛋白酶样域与C端沙利度胺结合域之间的距离。d, 一个含有1,854个未解析残基的核孔复合物（PDB: 7F60）。显示了真实结构（左）以及AlphaFold-Multimer v.2.3（中）和AF3（右）的预测结果。e, 预测的含有重叠DNA（粉色）和蛋白质（蓝色）链的三核小体（PDB: 7PEU）；突出显示的是重叠的蛋白质链B和J以及自我重叠的DNA链AA。除非另有说明，预测结果均由我们的全局复合物排名指标进行排名，并考虑了手性不匹配和空间碰撞惩罚（补充方法5.9.1）。

讨论

分子生物学的核心挑战在于理解和最终调控生物系统中复杂的原子相互作用。AF3模型在这一方向上迈出了重要一步，展示了在统一框架下准确预测广泛生物分子系统结构的可能性。尽管在所有相互作用类型中实现高度准确的预测仍面临重大挑战，但我们证明了构建一个对所有这些相互作用表现出强大覆盖和泛化能力的深度学习系统是可能的。我们还证明了缺乏跨实体的进化信息并不是预测进展的重大障碍。

文章

这些相互作用，以及抗体结果的显著改善，表明AlphaFold衍生的方法能够模拟分子间相互作用的化学和物理特性，而不依赖于MSA。最后，蛋白质-配体结构预测的巨大改进表明，在一个通用的深度学习框架内处理化学空间的广泛多样性是可能的，并且无需求助于蛋白质结构预测与配体对接之间的人为分离。

自下而上构建细胞组分的模型发展是解开细胞内分子调控复杂性的关键步骤，而AF3的表现表明，开发适当的深度学习框架可以大幅减少在这些任务上获得生物学相关性能所需的数据量，并放大已经收集的数据的影响。我们预计，结构建模不仅会因深度学习的进步而持续改进，还因为实验结构测定方法的不断进步，如冷冻电子显微镜和断层扫描技术的显著改进，将为进一步提高此类模型的泛化能力提供丰富的新的训练数据。实验方法和计算方法的并行发展有望将我们进一步推向一个结构化生物学理解和治疗开发的时代。

在线内容

任何方法、附加参考文献、Nature Portfolio 报道摘要、源数据、扩展数据、补充信息、致谢、同行评审信息；作者贡献和利益冲突的详细信息；以及数据和代码可用性的声明，均可在 <https://doi.org/10.1038/s41586-024-07487-w> 获取。

1. Jumper, J. 等。AlphaFold 实现高精度蛋白质结构预测。Nature 596, 583–589 (2021)。
2. Kreitz, J. 等。利用细菌收缩注射系统进行可编程蛋白质递送。Nature 616, 357–364 (2023)。
3. Lim, Y. 等。基于计算的蛋白质相互作用筛选揭示了 DONSON 在复制起始中的作用。Science 381, eadi3448 (2023)。
4. Mosalaganti, S. 等。基于人工智能的结构预测赋能人类核孔复合物的整合结构分析。Science 376, eabm9506 (2022)。
5. Anand, N. & Achim, T. 使用等变去噪扩散概率模型生成蛋白质结构和序列。预印本在 arXiv <https://doi.org/10.48550/arXiv.2205.15019> (2022)。
6. Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 及其在生物学和医学领域的应用。Signal Transduct. Target. Ther. 8, 115 (2023)。
7. Evans, R. 等。使用 AlphaFold-Multimer 进行蛋白质复合物预测。预印本在 bioRxiv <https://doi.org/10.1101/2021.10.04.463034> (2022)。
8. Žídek, A. AlphaFold v.2.3.0 技术说明。GitHub https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md (2022)。
9. Isert, C., Atz, K. & Schneider, G. 基于几何深度学习的结构药物设计。Curr. Opin. Struct. Biol. 79, 102548 (2023)。
10. Lin, Z. 等。使用语言模型进行原子级蛋白质结构的大规模进化预测。Science 379, 1123–1130 (2023)。
11. Baek, M. 等。使用三轨神经网络准确预测蛋白质结构和相互作用。Science <https://doi.org/10.1126/science.abj8754> (2021)。
12. Wu, R. 等。从一级序列进行高分辨率从头结构预测。预印本在 bioRxiv <https://doi.org/10.1101/2022.07.21.500999> (2022)。
13. Bryant, P., Pozzati, G. & Elofsson, A. 使用 AlphaFold2 改进蛋白质-蛋白质相互作用预测。Nat. Commun. 13, 1265 (2022)。
14. Moriwaki, Y. 在 X 上的帖子。X https://x.com/Ag_smith/status/1417063635000598528?lang=en-GB (2021)。
15. Baek, M. 在 X 上的帖子。X <https://x.com/minkbaek/status/1417538291709071362?lang=en> (2021)。
16. Qiao, Z. 等。使用多尺度深度生成模型进行特定状态的蛋白质-配体复合物结构预测。Nat. Mach. Intell. 6, 195–208 (2024)。
17. Nakata, S., Mori, Y. & Tanaka, S. 使用基于扩散的生成模型进行端到端的蛋白质-配体复合物结构生成。BMC Bioinform. 24, 233 (2023)。
18. Baek, M. 等。使用 RoseTTAFoldNA 准确预测蛋白质-核酸复合物。Nat. Methods 21, 117–121 (2024)。
19. Townshend, R. J. L. 等。RNA 结构的几何深度学习。Science 373, 1047–1051 (2021)。

20. Jiang, D. 等人。InteractionGraphNet：一种新颖且高效的深度图表示学习框架，用于精确的蛋白质-配体相互作用预测。《药物化学杂志》64, 18209–18232 (2021)。

21. Jiang, H. 等人。利用图神经网络预测蛋白质-配体对接结构。《化学信息与建模杂志》<https://doi.org/10.1021/acs.jcim.2c00127> (2022)。

22. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. Di Dock：用于分子对接的扩散步骤、扭曲和转弯。预印本在 arXiv <https://doi.org/10.48550/arXiv.2210.01776> (2022)。

23. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. & Jaakkola, T. EquiBind：用于药物结合结构预测的几何深度学习。在第39届国际机器学习会议 (eds Chaudhuri, K. 等人) 20503–20521 (PMLR, 2022)。

24. Liao, Z. 等人。DeepDock：通过结合配体和结构信息增强蛋白质-配体相互作用预测。在2019年 IEEE 国际生物信息学与生物医学会议 (BIBM) 311–317 (IEEE, 2019)。

25. Lu, W. 等人。TANKBind：用于药物-蛋白质结合结构预测的三角学感知神经网络。《神经信息处理系统进展》35, 7236–7249 (2022)。

26. Zhou, G. 等人。Uni-Mol：一种通用的三维分子表示学习框架。预印本在 ChemRxiv <https://chemrxiv.org/engage/chemrxiv/article-details/6402990d37e01856dc1d1581> (2023)。

27. Shen, T. 等人。E2Efold-3D：一种端到端深度学习方法，用于准确从头预测 RNA 三维结构。预印本在 arXiv <https://arxiv.org/abs/2207.01586> (2022)。

28. van Dijk, M. & Bonvin, A. M. J. J. 推动蛋白质-DNA 对接的极限：基准测试 HADDOCK 的性能。《核酸研究》38, 5634–5647 (2010)。

29. Krishna, R. 等人。使用 RoseTTAFold All-Atom 进行广义生物分子建模和设计。《科学》384, eadl2528 (2024)。

30. Butterschoen, M., Morris, G. M. & Deane, C. M. PoseBusters：基于 AI 的对接方法未能生成物理上有效的姿势或推广到新的序列。《化学科学》15, 3130–3139 (2024)。

31. Das, R. 等人。CASP15 中三维 RNA 结构预测的评估。《蛋白质》91, 1747–1770 (2023)。

32. Berman, H. M. 等人。蛋白质数据库。《核酸研究》28, 235–242 (2000)。

33. Karras, T., Aittala, M., Aila, T. & Laine, S. 阐明基于扩散的生成模型的设计空间。《神经信息处理系统进展》35, 26565–26577 (2022)。

34. Wang, Y., Elhag, A. A., Jaitly, N., Susskind, J. M. & Bautista, M. A. 生成分子构象场。预印本在 arXiv <https://doi.org/10.48550/arXiv.2311.17932> (2023)。

35. Ji, Z. 等人。自然语言生成中的幻觉调查。《ACM 计算调查》55, 248 (2023)。

36. Del Conte, A. 等人。蛋白质固有无序预测的批判性评估 (CAID) —— 第二轮结果。《蛋白质》91, 1925–1934 (2023)。

37. Trott, O. & Olson, A. J. AutoDock Vina：通过新的评分函数、高效优化和多线程提高对接速度和准确性。《计算化学杂志》31, 455–461 (2010)。

38. Miller, E. B. 等人。解决蛋白质-配体诱导适配对接问题的可靠且准确的解决方案。《化学理论与计算杂志》<https://doi.org/10.1021/acs.jctc.1c00136> (2021)。

39. Chen, K., Zhou, Y., Wang, S. & Xiong, P. 在 CASP15 中使用 BRiQ 势能进行 RNA 三级结构建模。《蛋白质》91, 1771–1778 (2023)。

40. Basu, S. & Wallner, B. DockQ：蛋白质-蛋白质对接模型的质量度量。《PLOS ONE》11, e0161879 (2016)。

41. Zhang, Y. & Skolnick, J. 用于自动评估蛋白质结构模板质量的评分函数。《蛋白质》57, 702–710 (2004)。

42. Watson, E. R. 等人。分子胶 CELMoD 化合物是脑苷脂酶的调节剂。《科学》378, 549–553 (2022)。

43. Wayment-Steele, H. K. 等人。通过序列聚类和 AlphaFold2 预测多种构象。《自然》625, 832–839 (2024)。

44. del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. 使用 AlphaFold2 采样转运蛋白和受体的替代构象状态。《eLife》<https://doi.org/10.7554/eLife.75751> (2022)。

45. Heo, L. & Feig, M. 在实验精度下对 G 蛋白偶联受体进行多状态建模。《蛋白质》90, 1873–1885 (2022)。

46. Wallner, B. AFsample：通过大规模采样使用 AlphaFold 改进多聚体预测。《生物信息学》39, btad573 (2023)。

47. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. IDDT：使用距离差异测试比较蛋白质结构和模型的局部叠合无分数。《生物信息学》29, 2722–2728 (2013)。

48. Zemla, A. LGa：用于在蛋白质结构中寻找三维相似性的方法。《核酸研究》31, 3370–3374 (2003)。

出版者声明 Springer Nature 对出版地图和机构隶属关系中的司法权主张保持中立。

开放获取 本文根据知识共享署名 4.0 国际许可协议 (Creative Commons Attribution 4.0 International License) 授权发布，允许在任何媒介或格式中使用、分享、改编、分发和复制，前提是您需适当注明原作者及来源，提供指向知识共享许可协议的链接，并标明是否对内容进行了修改。本文中的图片或其他第三方材料包含在文章的知识共享许可协议中，除非在材料的署名行中另有说明。如果材料未包含在文章的知识共享许可协议中，且您的使用行为不符合法定规定或超出了允许的使用范围，您将需要直接从版权所有者处获取许可。如需查看此许可协议的副本，请访问 <http://creativecommons.org/licenses/by/4.0/>。

© The Author(s) 2024



方法

完整算法细节

补充方法2-5中提供了各组件的详细解释。此外，补充算法1-31中提供了伪代码，网络图见图1d和图2a-c以及补充图2，输入特征见补充表5，训练中的其他超参数见补充表3、4和7。

训练计划

在训练过程中使用的结构数据在2021年9月30日之后未被发布，并且对于用于PoseBusters评估的模型，我们过滤掉了2021年9月30日之后发布的PDB³²结构。一个优化步骤使用256个输入数据样本的小批次，在初始训练期间， $256 \times 48 = 12,288$ 个扩散样本。对于微调，扩散样本的数量减少到 $256 \times 32 = 8,192$ 。模型分三个阶段进行训练——初始训练阶段使用384个token的裁剪尺寸，以及两个连续的微调阶段，分别使用640和768个token的裁剪尺寸。更多细节在补充方法5.2中提供。

推理模式

在2021年9月30日之后，未发布任何推断时间模板或参考配体位置特征，而在PoseBusters评估中，使用了更早的截止日期，即2019年9月30日。模型可以通过使用不同的随机种子来运行，以生成不同的结果，每个种子对应一批扩散样本。除非另有说明，所有结果均通过运行同一训练模型的5个种子生成，每个模型种子生成5个扩散样本，总共从25个样本中选择置信度最高的样本。预测中排除了标准结晶辅助剂（见补充表8）。

结果显示了排名最高的样本，样本的排名取决于是否试图在全球范围内选择总体最佳输出，还是为某个链、界面或修饰残基选择最佳输出。全局排名使用pTM和ipTM的混合，并结合一些条件来减少大量碰撞的情况并提高无序率；单链排名使用链特定的pTM测量；界面排名使用相关链对的定制ipTM测量；修饰残基排名使用感兴趣残基的平均pLDDT（补充方法5.9.3）。

指标

评估将预测的结构与相应的真实结构进行比较。如果复合物包含多个相同的实体，通过最大化LDDT来确定预测单元与真实单元之间的分配。配体中原子的局部对称组中的分配问题，通过在RDKit给出的前1,000个残基对称性上进行穷举搜索来解决。

我们使用DockQ、LDDT或口袋对齐的均方根偏差（r.m.s.d.）来衡量预测的质量。对于核酸-蛋白质界面，我们通过iLDDT来测量界面精度，该指标是根据界面中不同链之间的原子距离计算得出的。DockQ和iLDDT高度相关（扩展数据图9），因此DockQ的标准阈值可以转换为等效的iLDDT阈值。由于核酸的规模较大，与通常用于蛋白质的15 Å相比，核酸的LDDT（链内和界面）计算时采用了30 Å的包含半径。为了进行置信度校准评估，我们使用了一种定制的LDDT（LDDT_to_polymer）指标，该指标考虑了给定实体的每个原子与其包含半径内的任何C 或C1 聚合物原子之间的差异。这与置信度预测的训练方式密切相关（补充方法4.3.1）。

口袋对齐的均方根偏差（r.m.s.d.）计算如下：口袋定义为距离配体任何重原子10 Å以内的所有重原子，限制在配体或被评分修饰残基的主聚合物链上，并且进一步限制在蛋白质的主链原子上。主聚合物链的定义各有不同：对于PoseBusters，

它是距离配体10 Å范围内原子数最多的蛋白质链；对于键合配体得分，它是键合的聚合物链；对于修饰的残基，它是包含该残基的链（减去该残基）。通过使用口袋进行最小二乘刚性对齐，将预测结构与真实结构对齐，然后计算配体所有重原子的均方根偏差（r.m.s.d.）。

最近PDB评估集

对最近包含8,856个PDB复合物（这些复合物在2022年5月1日至2023年1月12日期间发布）的PDB集进行了通用模型评估。该集合几乎包含了在该期间内发布且模型大小不超过5,120个模型标记的所有PDB复合物（补充方法6.1）。对每个结构中的单链和界面分别评分，而不是仅查看完整复合物的评分，然后对链和界面进行聚类，以便首先在聚类内汇总评分，然后在聚类间汇总以计算平均评分，或使用逆聚类大小的权重进行分布统计（补充方法6.2和6.4）。

评估配体时排除了标准结晶辅助剂（补充表8）、我们的配体排除列表（补充表9）以及糖类（补充表10）。结合配体和非结合配体分别进行评估。离子仅在特别提及的情况下才被包括（补充表11）。

最近的PDB数据集经过筛选，形成了一个低同源性子集（补充方法6.1），用于某些结果的分析。同源性定义为与训练集中序列的序列同一性，并通过模板搜索进行测量（补充方法2.4）。在评估复合物中，如果单个聚合物链与训练集中链的最大序列同一性大于40%，则这些链会被过滤掉，其中序列同一性是指评估集链中与训练集链相同的残基百分比。单独的肽链（少于16个残基的蛋白质链）总是被过滤掉。对于聚合物-聚合物界面，如果两个聚合物与训练集中同一复合物中的两条链的序列同一性均大于40%，则该界面被过滤掉。对于与肽的界面，如果非肽实体与训练集中任何链的序列同一性大于40%，则该界面被过滤掉。

为了比较蛋白质-蛋白质界面和蛋白质单体的预测质量与AlphaFold-Multimer (v.2.3)⁸的预测质量，并比较单蛋白质链预测质量对MSA深度的依赖性，我们将低同源性的近期PDB数据集限制在少于20个蛋白质链和少于2,560个标记的复合物上。我们与未松弛的AlphaFold-Multimer v.2.3预测结果进行比较。

为了研究抗体-抗原界面预测，我们从低同源性的近期PDB数据集中筛选出包含至少一个蛋白质-蛋白质界面的复合物，其中一条蛋白质链属于两个最大的PDB链集群之一（这些集群代表了抗体）。我们进一步筛选出最多包含2,560个标记且在PDB中没有未知氨基酸的复合物，以便与AlphaFold-Multimer v.2.3的松弛预测进行广泛比较。最终得到71个抗体-抗原复合物，包含166个抗体-抗原界面，跨越65个界面集群。

MSA深度分析（扩展数据图7a）基于计算查询序列每个位置的有效序列数（Ne）的归一化值。通过统计该位置在MSA中非空位残基的数量并使用Ne方案⁴⁹对序列进行加权，得到每个残基的Ne值，其中序列同一性阈值为80%，计算区域为两序列中非空位的部分。

核酸预测基线

在核酸结构预测的基准性能测试中，我们报告了与现有的用于蛋白质-核酸和RNA三级结构预测的机器学习系统RoseTTAFold2NA¹⁸的基线比较。我们使用与AF3预测相同的MSAs运行开源的RF2NA⁵⁰进行比较。

文章

在AF3和RF2NA之间，我们选择了一部分最近的PDB数据集的子集以满足RF2NA的标准（总残基和核苷酸数<1,000）。由于RF2NA未被训练用于预测包含DNA和RNA的系统，分析仅限于仅含一种核酸类型的目标。在撰写本文时，没有公开可用的系统用于对PDB中任意组合的生物分子类型数据进行基线比较。作为RNA三级结构预测的附加基线，

我们评估了AF3在CASP15 RNA目标上的表现，这些目标在2023年12月1日之前已公开可用（R1116/8S95, R1117/8FZA, R1126

（从CASP15网站下载https://predictioncenter.org/casp15/TARGETS_PDB/R1126.pdb），R1128/8BTZ, R1136/7ZZJ4, R1138/

[7PTK/7PTL], R1189/7YR7 和 R1190/7YR6）。我们比较了排名第一的预测结果，并且在存在多个真实结构的情况下（如R1136），预测结果会与最接近的状态进行评分。我们将比较结果展示为与RF2NA作为代表性机器学习系统的对比；

AIchemy_RNA2作为表现最佳的人工干预参赛者；

以及AIchemy_RNA作为表现最佳的机器学习系统。所有参赛者的预测结果均从CASP网站下载并在内部进行评分。

PoseBusters

虽然其他分析使用了基于截至2021年9月30日之前发布的PDB数据训练的AlphaFold模型，但我们的PoseBusters分析则是基于一个架构相同且训练计划相似的模型，唯一的区别在于使用了更早的截至2019年9月30日的数据截止点。因此，该分析未包含此日期之后发布的训练数据、推理时间模板或“ref_pos”特征。

在指定的PDB文件中，对不对称单元进行了推断，并进行了以下轻微修改。在多个PDB文件中，与感兴趣的配体发生冲突的链被移除（7O1T, 7PUV, 7SCW, 7WJB, 7ZXV, 8AIE）。另一个PDB条目（8F4J）由于系统过大（超过5,120个token），无法推断整个系统，因此我们仅包含了与感兴趣的配体在20 Å范围内的蛋白质链。每个目标生成了五个模型种子，每个种子有五个扩散样本，共产生了25个预测结果，这些结果按质量和预测准确性进行排序：排序分数根据ipTM聚合值计算（补充方法5.9.3（第3点）），如果配体存在手性错误或与蛋白质发生冲突，则进一步除以100。

对于口袋对齐的r.m.s.d.，首先通过将预测结构与真实结构的口袋骨架原子（主蛋白质链中与配体在10 Å内接触最多的链上的CA、C或N原子）对齐来进行对齐。使用PoseBusters Python包v.0.2.7⁵¹来评估口袋对齐预测的r.m.s.d.和违规情况。

尽管AlphaFold模型对蛋白质口袋是“盲”的，但对接通常是在了解蛋白质口袋残基的情况下进行的。例如，Uni-Mol将口袋定义为任何距离感兴趣的配体重原子6 Å以内的残基²⁶。为了评估在提供口袋信息的情况下AF3对接配体的能力，我们对2019年9月30日的AF3模型进行了微调，增加了一个指定口袋-配体对的额外标记特征（补充方法2.8）。具体来说，引入了一个额外的标记特征，对于感兴趣的配体实体以及任何重原子距离该配体实体6 Å以内的口袋残基，该特征被设置为真。在训练时，选择一个随机的配体实体用于此特征。需要注意的是，可能会选择具有相同实体（CCD代码）的多个配体链。在推理时，根据感兴趣配体的CCD代码选择配体实体，因此偶尔会选择多个配体链。此分析的结果显示在扩展数据图4中。

模型性能分析与可视化

数据分析使用了Python v.3.11.7 (<https://www.python.org/>)、NumPy v.1.26.3 (<https://github.com/numpy/numpy>)、SciPy v.1.9.3 (<https://www.scipy.org/>)、seaborn v.0.12.2 (<https://github.com/mwaskom/seaborn>)。

Matplotlib v.3.6.1 (<https://github.com/matplotlib/matplotlib>)，pandas v.2.0.3 (<https://github.com/pandas-dev/pandas>)，statsmodels v.0.12.2 (<https://github.com/statsmodels/statsmodels>)，RDKit v.4.3.0 (<https://github.com/rdkit/rdkit>) 和Colab (<https://research.google.com/colaboratory>)。TM-align v.20190822 (<https://zhanglab.dcmb.med.umich.edu/TM-align/>) 用于计算TM分数。结构可视化在Pymol v.2.55.5 (<https://github.com/schrodinger/pymol-open-source>) 中创建。

报告摘要

关于研究设计的更多信息，请参阅本文所附的《自然》系列期刊报告摘要。

数据可用性

所有用于创建训练和评估输入的科学数据集均从公共来源免费获取。PDB中的结构用于训练和作为模板 (<https://files.wwpdb.org/pub/pdb/data/assemblies/mmCIF/>；序列集群可从<https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt>获取；序列数据可从https://files.wwpdb.org/pub/pdb/derived_data/获取）。训练使用了2023年1月12日下载的PDB版本，而模板搜索使用了2022年9月28日下载的版本。我们还使用了2023年10月19日下载的化学成分词典 (<https://www.wwpdb.org/data/ccd>)。我们从PDB中展示了以下实验结构，访问号为7PZB（参考文献52）、7PNM（参考文献53）、7TQL（参考文献54）、7AU2（参考文献55）、7U8C（参考文献56）、7URD（参考文献57）、7WUX（参考文献58）、7QIE（参考文献59）、7T82（参考文献60）、7CTM（参考文献61）、8CVP（参考文献42）、8D7U（参考文献42）、7F60（参考文献62）、8BTI（参考文献63）、7KZ9（参考文献64）、7XFA（参考文献65）、7PEU（参考文献66）、7SDW（参考文献67）、7TNZ（参考文献68）、7R6R（参考文献69）、7USR（参考文献70）和7Z1K（参考文献71）。我们还使用了以下公开的数据库进行训练或评估。详细的使用情况描述在补充方法2.2和2.5.2中。UniRef90 v.2020_01 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/)、UniRef90 v.2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/)、UniRef90 v.2022_05 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_05/uniref/)、Uniclust30 v.2018_08 (https://wwwuser.gwdg.de/~combiol/uniclust/2018_08/)、Uniclust30 v.2021_03 (https://wwwuser.gwdg.de/~combiol/uniclust/2021_03/)、MGNify clusters v.2018_12 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/)、MGNify clusters v.2022_05 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2022_05/)、BFD (<https://bfd.mmseqs.com>)、RFam v.14.9 (<https://ftp.ebi.ac.uk/pub/databases/Rfam/14.9/>)、RNACentral v.21.0 (<https://ftp.ebi.ac.uk/pub/databases/RNACentral/releases/21.0/>)、核酸数据库（截至2023年2月23日）(<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>)、JASPAR 2022 (<https://jaspar.elixir.no/downloads/>；参见<https://jaspar.elixir.no/profile-versions>获取版本信息)、参考文献72补充表中的SELEX蛋白序列以及参考文献73补充表中的SELEX蛋白序列。

代码可用性

AlphaFold 3 将仅作为非商业用途服务器在 <https://www.alphafoldserver.com> 上提供，对允许的配体和共价修饰有相关限制。描述算法的伪代码可在补充信息中找到。代码未提供。

49. 吴涛, 侯杰, Adhikari, B. & 程杰. 影响基于深度学习的残基间接触预测的几个关键因素分析. 生物信息学 36, 1091–1098 (2020). 50. DiMaio, F. RF2NA v.0.2. GitHub <https://github.com/uw-ipd/RoseTTAFold2NA/releases/tag/v0.2> (2023). 51. Buttenschoen, M. PoseBusters v.0.2.7. GitHub <https://github.com/maabuu/posebusters/releases/tag/v0.2.7> (2023).

52. Werel, L. 等人。根瘤菌双特异性cAMP和cGMP受体蛋白Clr的结构基础。MBio 14, e0302822 (2023).
53. Wang, C. 等人。人冠状病毒OC43刺突蛋白的抗原结构揭示了暴露和隐蔽的中和表位。Nat. Commun. 13, 2921 (2022).
54. Lapointe, C. P. 等人。eIF5B和eIF1A重新定位起始tRNA以允许核糖体亚基结合。Nature 607, 185–190 (2022).
55. Wilson, L. F. L. 等人。EXTL3的结构有助于解释双域外骨素在硫酸肝素合成中的不同作用。Nat. Commun. 13, 3314 (2022).
56. Liu, X. 等人。高度活性的CAR T细胞，结合间皮素近膜区域且不被脱落间皮素阻断。Proc. Natl Acad. Sci. USA 119, e2202439119 (2022).
57. Liu, Y. 等人。Porcupine介导的Wnt酰化机制及其抑制。Nature 607, 816–822 (2022).
58. Kurosawa, S. 等人。通过硫酸消除形成酶促氮丙啶的分子基础。J. Am. Chem. Soc. 144, 16164–16170 (2022).
59. Boey, H. K. 等人。开发选择性磷脂酰肌醇5-磷酸4激酶抑制剂，具有非ATP竞争性、别构结合模式。J. Med. Chem. 65, 3359–3370 (2022).
60. Buckley, P. T. 等人。用于预防和治疗金黄色葡萄球菌感染的多价人抗体-中心蛋白融合蛋白。Cell Host Microbe 31, 751–765 (2023).
61. Mohapatra, S. B. & Manoj, N. 糖苷水解酶家族4中NAD(H)依赖的 α -D-葡萄糖苷酶的催化机制和底物识别的结构基础。Biochem. J. 478, 943–959 (2021).
62. Gao, X. 等人。Sarbecovirus ORF6介导的核质运输阻断的结构基础。Nat. Commun. 13, 4782 (2022).
63. Atkinson, B. N. 等人。设计从共价到非共价抑制剂的转换，用于羧酸酯酶Notum活性。Eur. J. Med. Chem. 251, 115132 (2023).
64. Luo, S. 等人。细菌Pip系统植物效应蛋白识别蛋白的结构基础。Proc. Natl Acad. Sci. USA 118, e2019462118 (2021).
65. Liu, C. 等人。鉴定单糖衍生物作为人及小鼠半乳糖凝集素-3的强效、选择性和口服生物利用度抑制剂。J. Med. Chem. 65, 11084–11099 (2022).
66. Dombrowski, M., Engeholm, M., Dienemann, C., Dodonova, S. & Cramer, P. 组蛋白H1与核小体阵列的结合依赖于连接DNA的长度和轨迹。Nat. Struct. Mol. Biol. 29, 493–501 (2022).
67. Vecchioni, S. 等人。金属介导的DNA纳米技术在三维空间中的结构库：模板衍射。Adv. Mater. 35, e2210938 (2023).
68. Wang, W. & Pyle, A. M. RIG-I受体采用两种不同构象以区分宿主和病毒RNA配体。Mol. Cell 82, 4131–4144 (2022).
69. McGinnis, R. J. 等人。单体分枝杆菌噬菌体免疫抑制因子利用两个域识别不对称DNA序列。Nat. Commun. 13, 4105 (2022).
70. Dietrich, M. H. 等人。针对Pf5230的纳米抗体阻断恶性疟原虫传播。Biochem. J. 479, 2529–2546 (2022).
71. Appel, L.-M. 等人。SPOC域是一个磷酸丝氨酸结合模块，连接转录机器与共转录及转录后调控因子。自然通讯 14, 166 (2023).
72. Yin, Y. 等人。胞嘧啶甲基化对人类转录因子DNA结合特异性的影响。科学 356, eaaj2239 (2017).
73. Jolma, A. 等人。依赖于DNA的转录因子成对形成改变了它们的结合特异性。自然 527, 384–388 (2015).

致谢

我们感谢G. Arena, Ž. Avsec, A. Baryshnikov, R. Bates, M. Beck, A. Bond, N. Bradley-Schmieg, J. Cavojska, B. Coppin, E. Dupont, S. Eddy, M. Fiscato, R. Green, D. Hariharan, K. Holsheimer, N. Hurley, C. Jones, K. Kavukcuoglu, J. Kelly, E. Kim, A. Koivuniemi, O. Kovalevskiy, D. Lasecki, M. Last, A. Laydon, W. McCorkindale, S. Miller, A. Morris, L. Nicolaisen, E. Palmer, A. Paterson, S. Petersen, O. Purkiss, C. Shi, G. Thomas, G. Thornton和H. Tomlinson的贡献。

作者贡献 共同贡献的作者按字母顺序排列，其余核心贡献作者（不包括共同指导作者）以及所有非指导作者同样按字母顺序排列。D.H.、M.J.和J.M.J.领导了研究工作。M.J.、J.M.J.和P.K.制定了研究策略。J. Abramson, V.B.、T.G.和C.C.H.领导了关键研究支柱。T.G.和A. Žídek负责研究的技术框架。O.B.、H.G.和S.S.协调和管理了研究项目。J. Abramson, J. Adler, E.A.、A.J.B.、J.B.、V.B.、A.I.C.-R.、J.D.、R.E.、D.A.E.、M.F.、F.B.F.、T.G.、C.-C.H.、M.J.、J.M.J.、Y.A.K.、A. Potapenko, A. Pritzel, D.R.、O.R.、A.T.、C.T.、K.T.、L.W.、Z.W.和E.D.Z.开发了神经网络架构和训练程序。J. Abramson, A.J.B.、J.B.、V.B.、C.B.、S.W.B.、A.B.、A. Cherepanov, A.I.C.-R.、A. Cowie, J.D.、T.G.、R.J.、M.O.、K.P.、D.R.、O.R.、M.Z.、A. Žemgulyt和A. Žídek开发了训练、推理、数据和评估基础设施。J. Abramson, J. Adler, A.J.B.、V.B.、A.I.C.-R.、R.E.、D.A.E.、T.G.、D.H.、M.J.、J.M.J.、P.K.、K.P.、A. Pritzel, O.R.、P.S.、S.S.、A.S.、K.T.和L.W.参与了论文的撰写。M.C.、C.M.R.L.和S.Y.为项目提供了建议。

竞争利益 与作者相关的实体已提交了美国临时专利申请，包括63/611,674、63/611,638和63/546,444，涉及使用嵌入神经网络和生成模型预测分子复合物的三维结构。除A.B.、Y.A.K.和E.D.Z.外的所有作者均对所述工作有商业利益。

附加信息

补充信息

在线版本包含可通过以下链接获取的补充材料：<https://doi.org/10.1038/s41586-024-07487-w>。

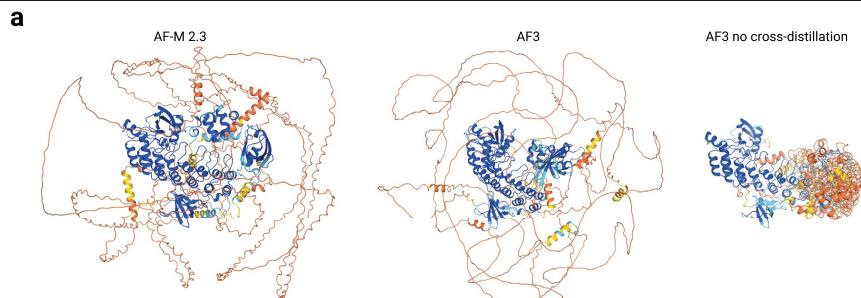
通信和材料请求应发送至Max Jaderberg, Demis Hassabis或John M. Jumper。

同行评审信息

《自然》杂志感谢Justas Dapkunas、Roland Dunbrack和Hashim Al-Hashimi对本工作的同行评审所做出的贡献。

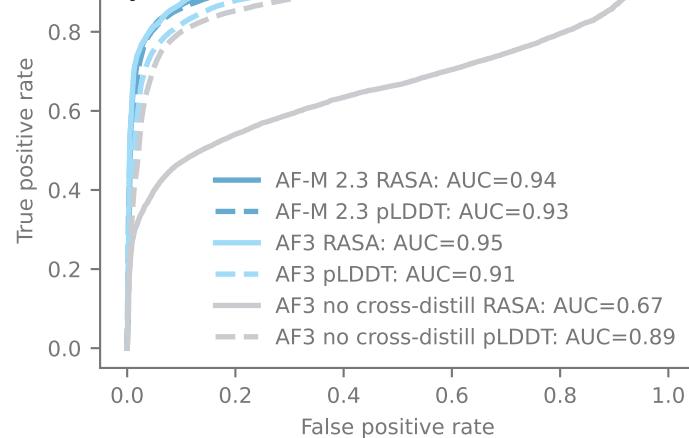
重印和权限信息可在以下网址获取：<http://www.nature.com/reprints>。

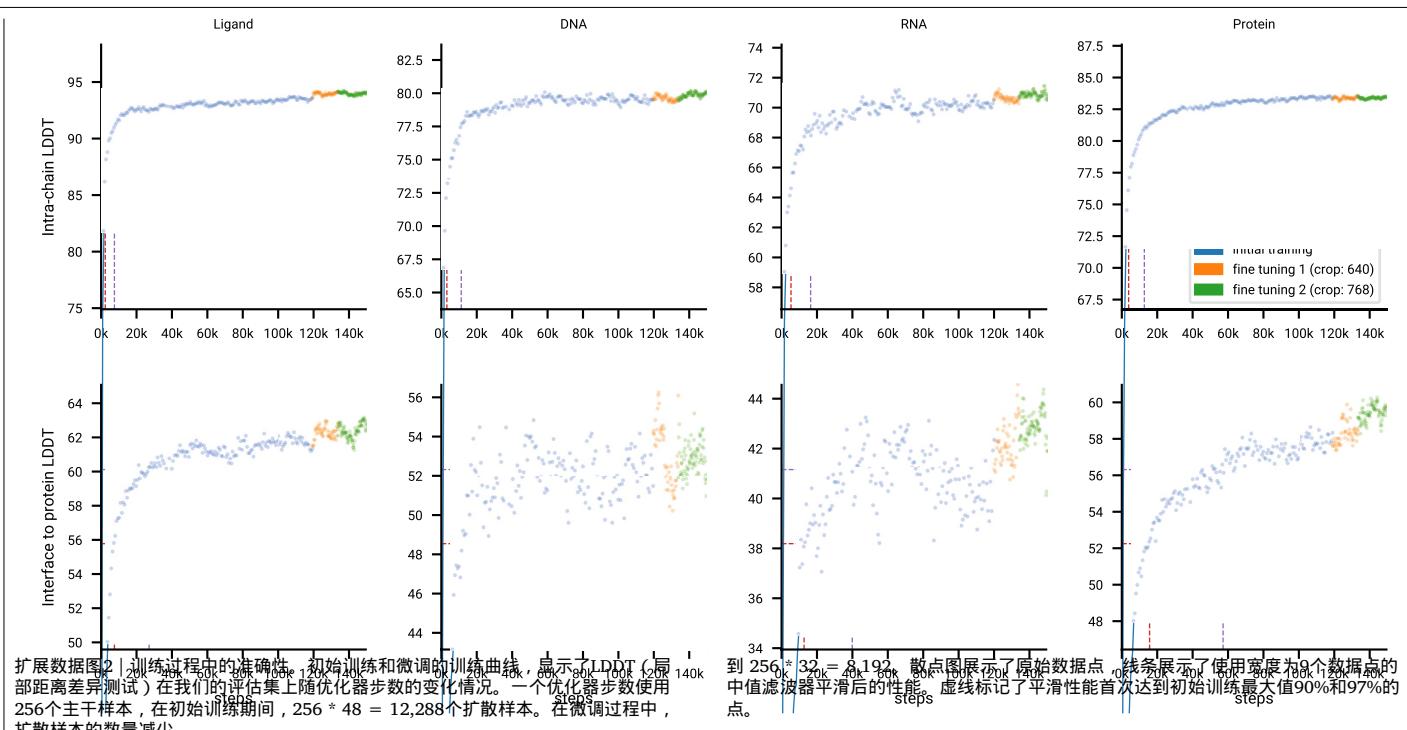
文章



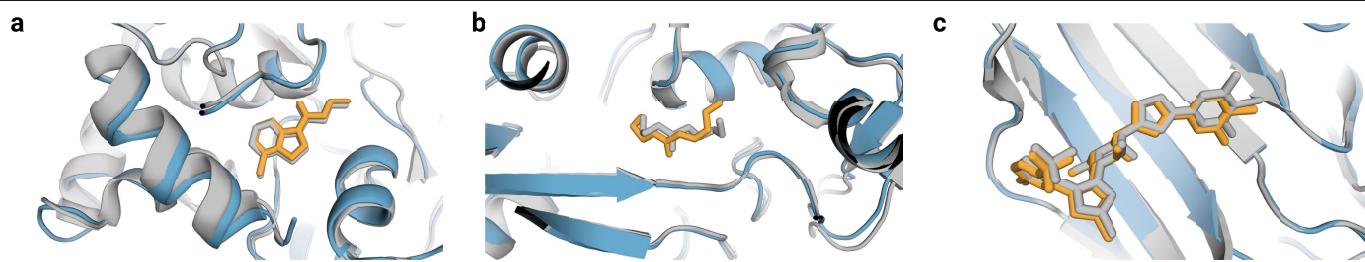
扩展数据图1 | 无序区域预测。a, 对来自AlphaFoldMultimer v2.3、AlphaFold 3以及未使用无序蛋白质PDB交叉蒸馏集训练的AlphaFold β 的无序蛋白质的预测示例。蛋白质为CAID 2 (Critical Assessment of protein Intrinsic Disorder prediction , 蛋白质内在无序性预测的关键评估) 数据集中的DP02376。预测结果按pLDDT着色 (橙色 : pLDDT < = 50 , 黄色 : 50 < pLDDT < = 70 , 浅蓝色 : 70 < pLDDT < = 90 , 深蓝色 : 90 < = pLDDT < 100)。

b, CAID 2数据集中蛋白质残基无序性的预测，这些蛋白质与AF3训练集的同源性较低。预测方法包括RASA (相对可及表面积) 和pLDDT (N = 151种蛋白 ; 46,093个残基)。

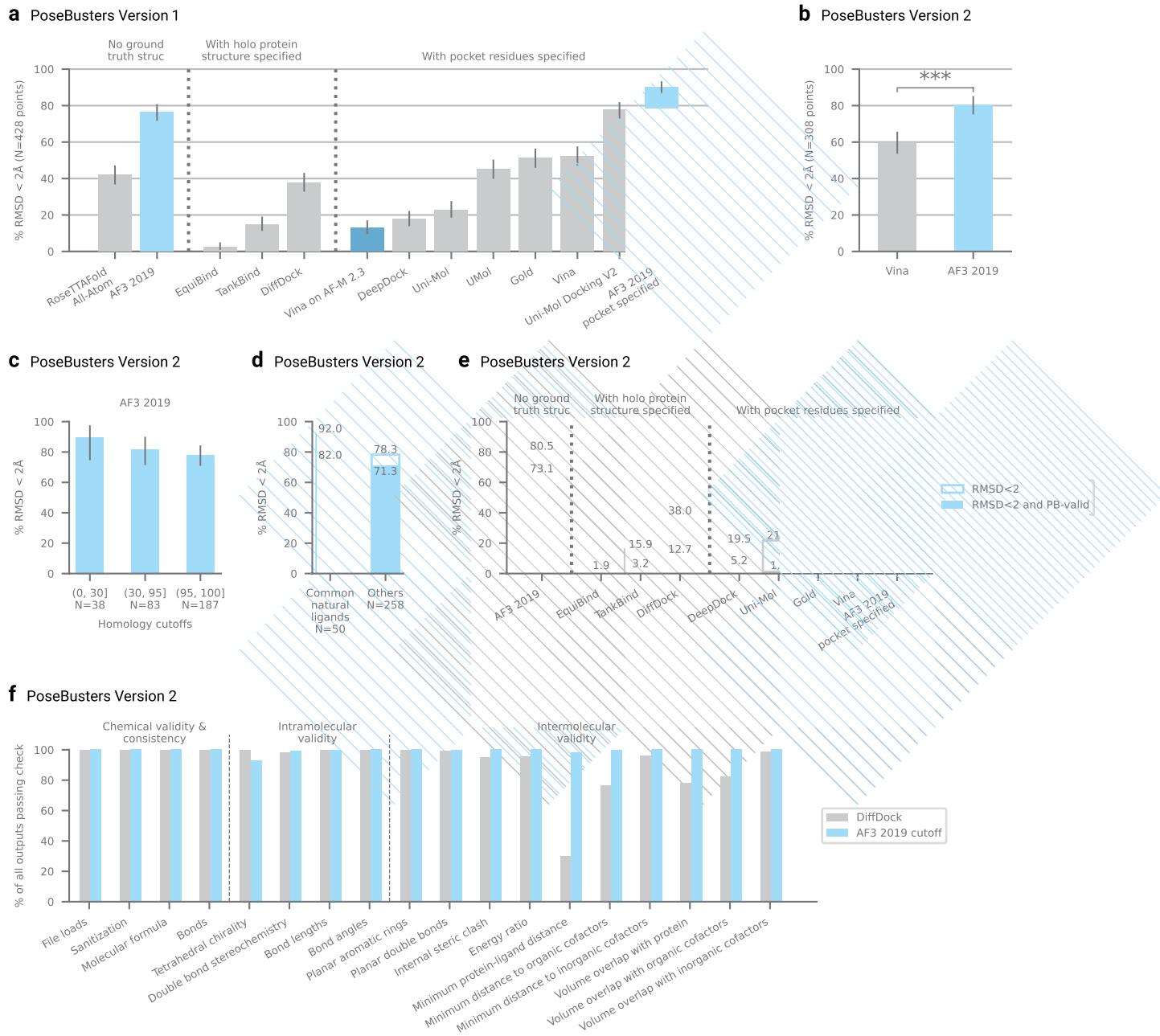




文章



扩展数据图3 | AlphaFold 3 对 PoseBusters 示例的预测，其中 Vina 和 Gold 预测不准。预测的蛋白质链以蓝色显示，预测的配体以橙色显示，真实结构以灰色显示。a, 人 Notum 与抑制剂 ARUK3004556 结合 (PDB ID 8BTI, 配体 RMSD: 0.65 Å)。
b, 假单胞菌属PDC86 Aapf结合HEHEAA (PDB ID 7KZ9 , 配体RMSD: 1.3 Å)。
c, 人半乳糖凝集素-3碳水化合物识别域与化合物22复合物 (PDB ID 7XFA , 配体 RMSD: 0.44 Å)。

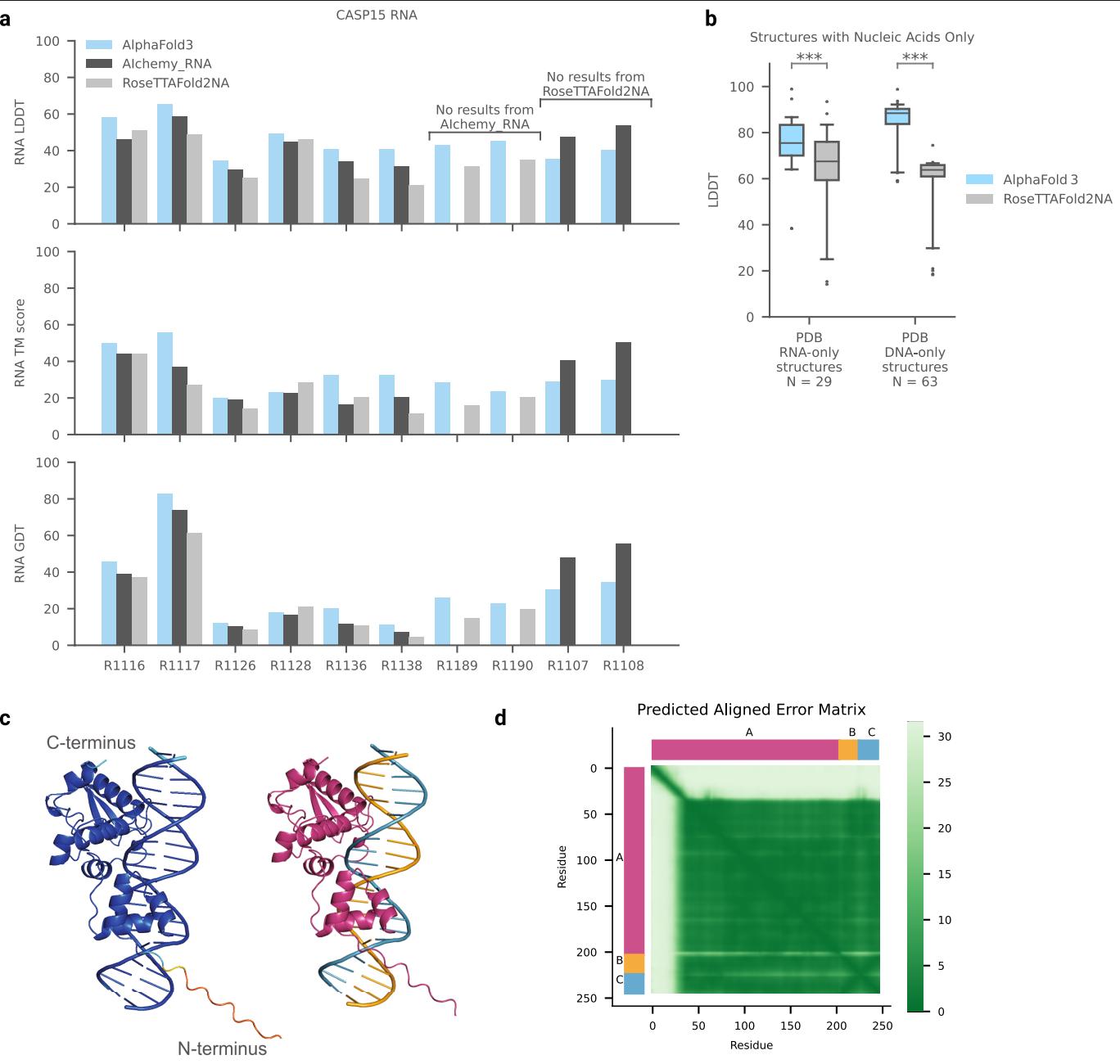


扩展数据图4 | PoseBusters分析。a, AlphaFold 3 与基线方法在PoseBusters

版本1基准集（V1，2023年8月发布）上的蛋白质-配体结合成功率比较。方法按使用的基础真值信息程度分类。注意，除UMol和AF3外，所有使用口袋残基信息的方法也使用基础真值全蛋白结构。b, PoseBusters版本2（V2，2023年11月发布）中，领先对接方法Vina与AF3 2019的比较（双侧Fisher精确检验， $N = 308$ 个目标， $p = 2.3 \times 10^{-8}$ ）。c, AF3 2019在PoseBusters V2中对低、中、高蛋白序列同源性目标的结果（整数范围表示与训练集中蛋白质的最大序列一致性）。d, AF3 2019在PoseBusters V2中按配体分类的结果。

那些被归类为“常见天然”配体和其他配体。“常见天然”配体定义为在PDB中出现超过100次且经视觉检查不属于非天然的配体。完整列表可在补充表15中找到。深色条表示RMSD < 2 Å且通过PoseBusters有效性检查（PB-valid）。e, PoseBusters V2结构精度和有效性。深色条表示RMSD < 2 Å且通过PoseBusters有效性检查（PB-valid）。浅色阴影条表示RMSD < 2 Å但未通过PB有效性检查。f, PoseBusters V2详细有效性检查比较。误差条表示精确二项分布的95%置信区间。 $N = 427$ 个目标用于RoseTTAFold全原子版本1，其他版本1中为428个目标；版本2中为308个目标。

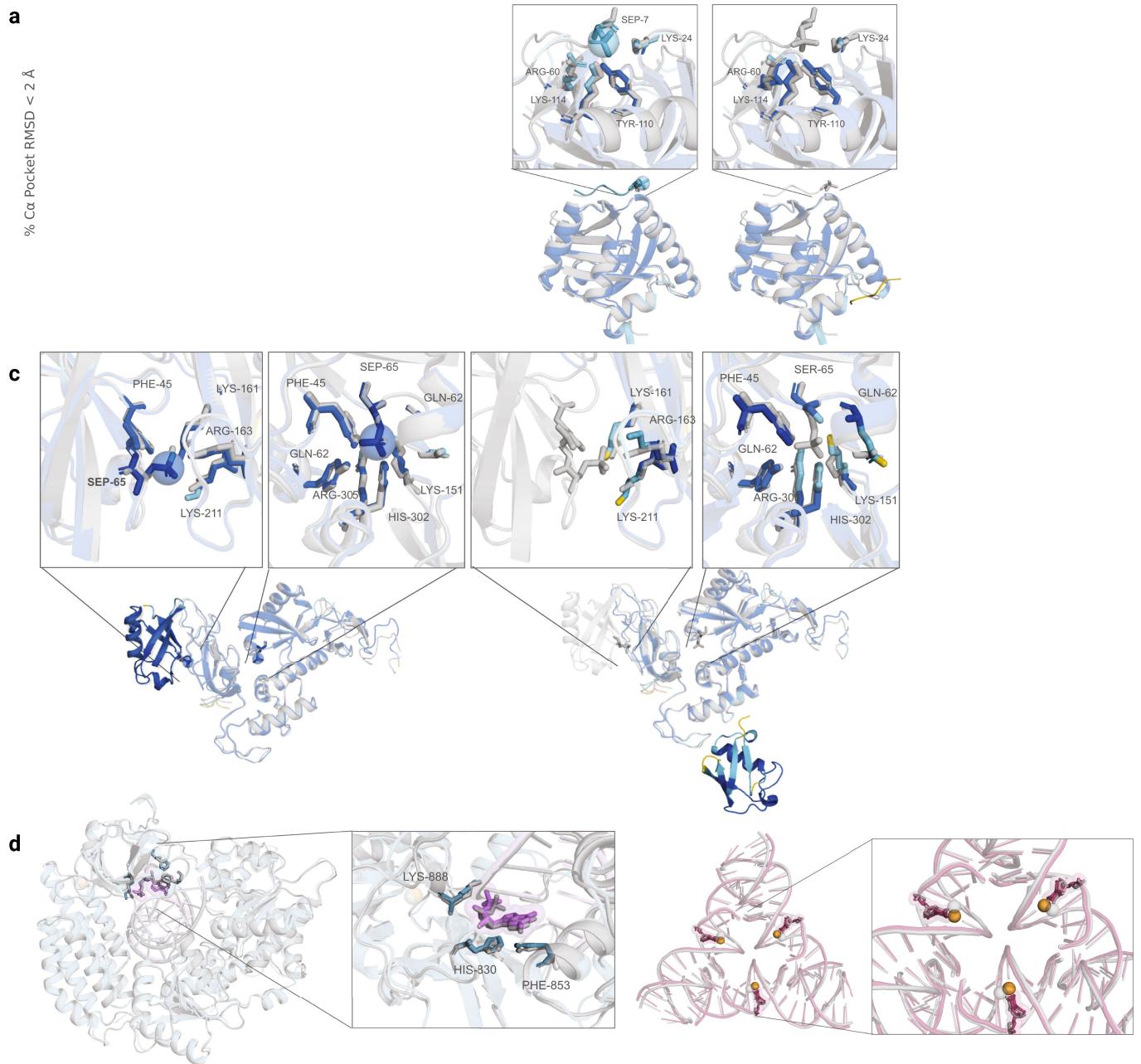
文章



扩展数据图5 | 核酸预测准确性与置信度。

a. 从AlChemey_RNA（基于AI的最高提交）、RoseTTAFold2NA（能够预测蛋白质-RNA复合物的AI方法）和AlphaFold 3中获得的CASP15 RNA预测准确性。13个目标中有10个可在PDB或通过CASP15网站进行评估。预测结果从CASP网站下载，用于外部模型。b. 最近PDB评估集中包含低同源性RNA或DNA单独复合物结构的准确性。AlphaFold 3与RoseTTAFold2NA (RF2NA) 的比较 (RNA : N = 29个结构，配对Wilcoxon符号秩检验, $p = 1.6 \times 10^{-7}$; DNA : N = 63个结构，配对双侧Wilcoxon符号秩检验)。

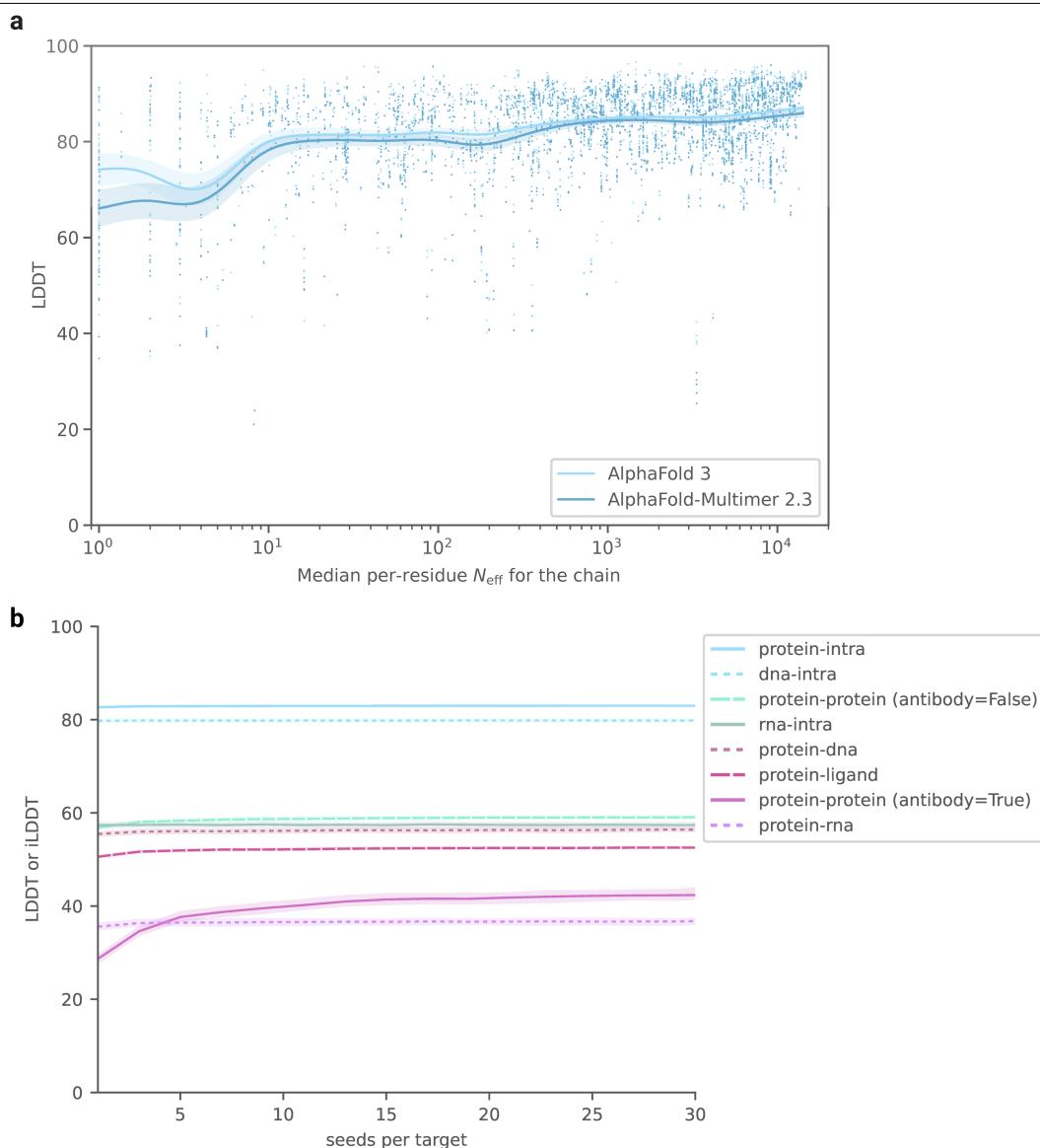
测试， $p = 5.2 \times 10^{-12}$ ）。注意，RF2NA仅在双链（至少形成10个氢键的链）上进行了训练和评估，但此数据集中的一些DNA结构可能不是双链。箱线图、中位数和须线边界分别位于(25%，75%)区间、中位数和(5%，95%)区间。c. 预测的结核杆菌噬菌体免疫抑制蛋白与双链DNA结合的结构 (PDB ID 7R6R)，按pLDDT着色 (左图；橙色：0–50，黄色：50–70，青色70–90，蓝色90–100) 和链ID着色 (右图)。注意，未完全显示的无序N端。d. 预测的每对token的排列误差 (PAE)，行和列按链ID标记，绿色渐变表示PAE。



扩展数据图6 | 修饰蛋白质和核酸的分析与示例。 a, 近期PDB评估集中包含常见磷酸化残基 (SEP, TPO, PTR, NEP, HIP) 的结构准确性分析。AlphaFold 3在模型中包含磷酸化与不包含磷酸化的比较 ($N = 76$ 个簇), 配对双侧Wilcoxon符号秩检验, $p = 1.6 \times 10^{-4}$)。注意, 为了预测不包含磷酸化的结构, 我们在修饰位置预测母体(标准)残基。AlphaFold 3在包含磷酸化模型时通常能实现更好的主链准确性。误差棒表示精确二项分布的95%置信区间。b, 人SHARP的SPOC结构域与磷酸化的RNA聚合酶II C端结构域复合物 (PDB ID 7Z1K), 预测结果按pLDDT着色 (橙色: 0–50, 黄色: 50–70, 青色: 70–90, 蓝色: 90–100)。左: 包含磷酸化模型 (平均口袋对齐RMSDC 2.104 Å)。右: 不包含磷酸化模型 (平均

口袋对齐的RMSDC 10.261 Å)。当排除磷酸化时, AlphaFold 3在磷酸肽上提供较低的pLDDT置信度。c, parkin与两个磷酸化泛素分子结合的结构 (PDB ID 7US1), 预测结果同样按pLDDT着色。左: 模型中包含磷酸化 (平均口袋对齐的RMSDC 0.424 Å)。右: 未模型化磷酸化 (平均口袋对齐的RMSDC 9.706 Å)。当排除磷酸化时, AlphaFold 3在错误预测的泛素界面残基上提供较低的pLDDT置信度。d, 含修饰核酸的示例结构。左: RNA中的鸟苷单磷酸 (PDB ID 7TNZ, 平均口袋对齐的修饰残基RMSD 0.840 Å)。右: 甲基化DNA胞嘧啶 (PDB ID 7SDW, 平均口袋对齐的修饰残基RMSD 0.502 Å)。我们标记了预测结构的残基以供参考。真实结构为灰色; 预测的蛋白质为蓝色, 预测的RNA为紫色, 预测的DNA为洋红色, 预测的离子为橙色, 预测的修饰通过球体突出显示。

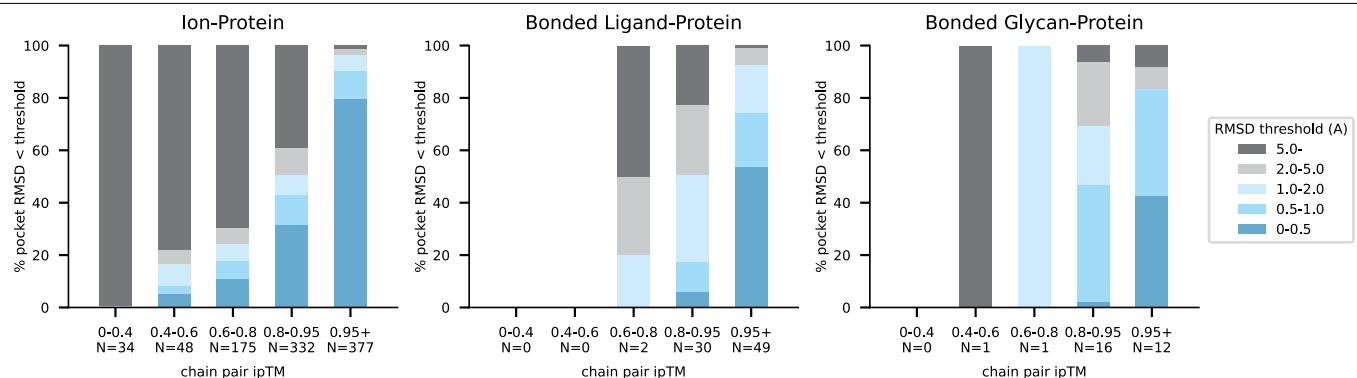
文章



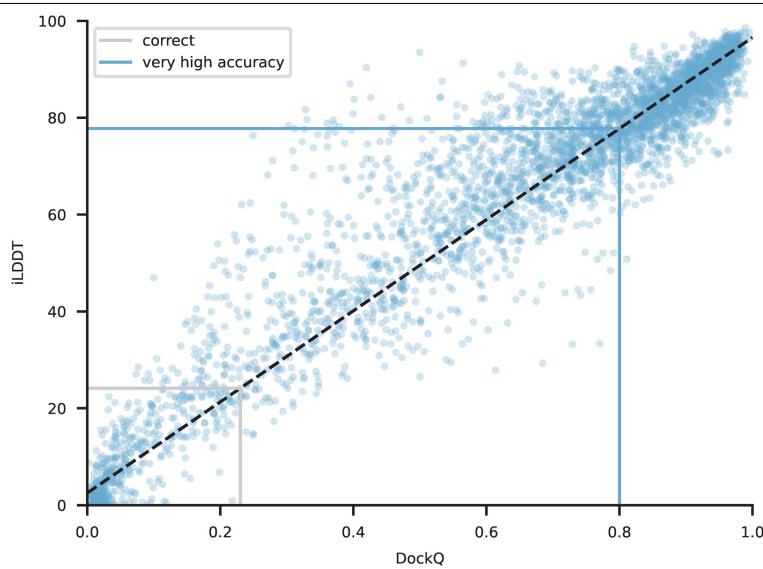
扩展数据图7 | MSA大小和种子数量对模型准确性的影响。

a. MSA深度对蛋白质预测准确性的影响。准确性以单链LDDT分数表示，MSA深度通过使用 N_{eff} 加权方案计算MSA中每个位置的非间隙残基数量并取残基间的中位数来计算（详见方法部分对 N_{eff} 的说明）。用于AF-M 2.3的MSA与AF3略有不同；为了更清晰地进行比较，数据使用了AF3的MSA深度。分析使用了低同源性Recent PDB集中的每一条蛋白质链，限制为复合物中少于20条蛋白质链且少于2,560个标记的链（详见方法部分对Recent PDB集和与AF-M 2.3比较的说明）。曲线为

通过高斯核平均平滑（窗口大小为对数 $10(N_{\text{eff}})$ 的0.2单位）获得；阴影区域为使用10,000个样本的自助法估计的95%置信区间。b，不同分子类型随着种子数量增加的排名准确度提升。预测按置信度排序，仅对每个界面的最高置信度进行评分。在低同源性的最新PDB数据集上进行评估，过滤至少于1,536个标记。评估的聚类数量：dna-intra = 386, protein-intra = 875, rna-intra = 78, protein-dna = 307, protein-rna = 102, protein-protein (antibody = False) = 697, protein-protein (antibody = True) = 58。置信区间为1,000个样本的95%自助法。



文章



扩展数据图 9 | 蛋白质-蛋白质界面中DockQ与iLDDT的相关性。每簇一个数据点，共显示4,182个簇。最佳拟合线采用 ϵ 为1的Huber回归器。DockQ分类为正确（>0.23）和极高准确度（>0.8）分别对应iLDDT值为23.6和77.6。

扩展数据表1 | 生物分子复合物预测准确性

AlphaFold 3 在 PoseBusters V1 (2023年8月发布)、PoseBusters V2 (2023年11月6日发布) 以及我们最近的PDB评估集上的表现。对于配体和核酸，N表示结构的数量；对于共价修饰和蛋白质，N表示聚类的数量。

Task	Dataset	Metric	Notes	Method	N	Mean	95% CI
Ligands	PoseBusters V1	% RMSD < 2 Å	–	RoseTTAFold All-Atom AF3 (2019 cutoff)	427	42.0	37.2 – 46.8
				Holo protein struct. given	428	76.4	72.1 – 80.3
				EquiBind TankBind DiffDock	428	2.6 15.0 37.9	1.3 – 4.6 11.7 – 18.7 33.2 – 42.6
		Pocket residues specified	Vina on AF-M 2.3 DeepDock Uni-Mol UMol Gold Vina Uni-Mol Docking V2 AF3 (2019 cutoff) pocket specified	Vina on AF-M 2.3	428	13.1	10.0 – 16.7
				DeepDock	428	17.8	14.3 – 21.7
				Uni-Mol	428	22.9	19.0 – 27.2
				UMol	428	45.0	40.3 – 49.9
				Gold	428	51.2	46.3 – 56.0
				Vina	428	52.3	47.5 – 57.2
				Uni-Mol Docking V2	428	77.6	73.3 – 81.4
Ligands	PoseBusters V2	% RMSD < 2 Å	–	AF3 (2019 cutoff)	308	80.5	75.6 – 84.8
				Holo protein struct. given	308	1.9	0.7 – 4.2
				EquiBind TankBind DiffDock	308	15.9 38.0	12.0 – 20.5 32.5 – 43.7
		Pocket residues specified	Vina on AF-M 2.3 DeepDock Uni-Mol Gold Vina AF3 (2019 cutoff) pocket specified	Vina on AF-M 2.3	308	15.3	11.4 – 19.8
				DeepDock	308	19.5	15.2 – 24.4
				Uni-Mol	308	21.8	17.3 – 26.8
				Gold	308	58.1	52.4 – 63.7
				Vina	308	59.7	54.0 – 65.3
				AF3 (2019 cutoff) pocket specified	308	93.2	89.8 – 95.7
Nucleic Acids	Protein-RNA	iLDDT		RoseTTAFold2NA AF3	25	19.0 39.4	15.6 – 23.2 28.5 – 51.9
	Protein-dsDNA	iLDDT		RoseTTAFold2NA AF3	38	28.3 64.8	20.7 – 37.5 56.4 – 71.7
	CASP 15 RNA	RNA LDDT		RoseTTAFold2NA AF3 Alchemy_RNA2 (has human input) RNAPolis (has human input) Chen (has human input) Kiharalab UltraFold	8	35.5 47.3 54.5 50.5 49.8 40.9 37.8	28.3 – 43.8 41.7 – 55.2 45.3 – 62.4 45.2 – 55.8 40.7 – 58.5 35.1 – 54.3 32.5 – 45.0
	Bonded ligands	% RMSD < 2 Å		AF3	66	78.5	68.3 – 86.2
	Glycosylation	% RMSD < 2 Å	high-quality, single-residue	AF3	28	72.1	53.1 – 85.7
			all-quality, single-residue	AF3	167	46.0	40.0 – 52.1
			all-quality, multi-residue	AF3	131	42.4	35.4 – 49.3
	Modified residues	% RMSD < 2 Å		AF3	154	59.9	52.4 – 67.0
	Modified protein residues	% RMSD < 2 Å		AF3	40	51.0	36.0 – 65.6
	Modified DNA residues	% RMSD < 2 Å		AF3	91	68.6	59.0 – 76.9
Proteins	Modified RNA residues	% RMSD < 2 Å		AF3	23	40.9	23.4 – 59.9
	All Protein-Protein	% dockq > 0.23		AF-M 2.3 AF3	1064	67.5 76.6	64.7 – 70.1 74.0 – 78.9
	Protein-Antibody	% dockq > 0.23		AF-M 2.3 AF3	65	29.6 62.9	19.6 – 40.4 51.4 – 73.5
	Monomers	LDDT		AF-M 2.3 AF3	338	85.5 86.9	84.7 – 86.1 86.2 – 87.6

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

All scientific datasets used to create training and evaluation inputs are freely available from public sources (see Data section below). No additional data was collected.

Data analysis

Data analysis used Python v3.11.7 (<https://www.python.org/>), NumPy v1.26.3 (<https://github.com/numpy/numpy>), SciPy v1.9.3 (<https://www.scipy.org/>), seaborn v0.12.2 (<https://github.com/mwaskom/seaborn>), Matplotlib v3.6.1 (<https://github.com/matplotlib/matplotlib>), pandas v2.0.3 (<https://github.com/pandas-dev/pandas>), statsmodels v0.12.2 (<https://github.com/statsmodels/statsmodels>), RDKit v4.3.0 (<https://github.com/rdkit/rdkit>), and Colab (<https://research.google.com/colaboratory>). TM-align v20190822 (<https://zhanglab.dcmbe.med.umich.edu/TM-align/>) was used for computing TM-scores. Structure visualizations were created in Pymol v2.55.5 (<https://github.com/schrodinger/pymol-open-source>). PoseBusters scoring done with PoseBusters v0.2.7 (<https://github.com/maabuu/posebusters>). RoseTTAFold2NA benchmarking done with RoseTTAFold2NA v0.2 (<https://github.com/uw-ipd/RoseTTAFold2NA>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All scientific datasets used to create training and evaluation inputs are freely available from public sources. Structures from the PDB were used for training and as templates (<https://files.wwpdb.org/pub/pdb/data/assemblies/mmCIF/>; for sequence clusters see <https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt>; for sequence data see https://files.wwpdb.org/pub/pdb/derived_data/).

Training used a version of the PDB downloaded 12 January 2023, while template search used a version downloaded 28 September 2022. We also used the Chemical Components Dictionary downloaded on 19 October 2023 (<https://www.wwpdb.org/data/ccd>).

We show experimental structures from the PDB with accession numbers 7PZB50,51, 7PNM52,53, 7TQL54,55, 7AU256,57, 7U8C58,59, 7URD60,61, 7WUX62,63, 7QIE64,65, 7T8266,67, 7CTM68,69, 8CVP43,70, 8D7U43,71, 7F6072,73, 8BTI74,75, 7KZ976,77, 7XFA78,79, 7PEU80,81, 7SDW82,83, 7TNZ84,85, 7R6R 86,87, 7USR88,89, and 7Z1K,90,91

We also used the following publicly available databases for training or evaluation. Detailed usage is described in Supplementary Methods 2.2{Genetic search} and Supplementary Methods 2.5.2{Distillation datasets}.

UniRef90 v.2020_01 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/),

UniRef90 v.2020_03 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_03/uniref/),

UniRef90 v.2022_05

https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_05/uniref/),

Uniclust30 v.2018_08

(https://wwwuser.gwdg.de/~comppbiol/uniclust/2018_08/),

Uniclust30 v.2021_03

(https://wwwuser.gwdg.de/~comppbiol/uniclust/2021_03/),

MGNify clusters v.2018_12

(https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2018_12/),

MGNify clusters v.2022_05

(https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2022_05/),

BFD

(<https://bfd.mmseqs.com>),

RFam v.14.9

(<https://ftp.ebi.ac.uk/pub/databases/Rfam/14.9/>),

RNAcentral v.21.0

(<https://ftp.ebi.ac.uk/pub/databases/RNAcentral/releases/21.0/>),

Nucleotide Database (as of 23 February 2023)

(<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>),

JASPAR 2022

(<https://jaspar.elixir.no/downloads/>; see <https://jaspar.elixir.no/profile-versions> for version information),

SELEX protein sequences from Supplementary Tables92

(<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8009048/>),

SELEX protein sequences from Supplementary Tables93

(<https://www.nature.com/articles/nature15518>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

N/A

Reporting on race, ethnicity, or other socially relevant groupings

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available data were used for each benchmark. No subsampling was performed.
Data exclusions	PDB structures were excluded on the basis of size or homology as described in the text
Replication	Code and method details were carefully checked for completeness and replicability.
Randomization	The work constitutes in-silico analysis so all treatments (software packages) were applied to all relevant data for benchmarking.
Blinding	Test sets were held back from training but researchers were not blinded. Large test sizes (all recent PDB) were used instead to avoid overfitting. Fully blind tests would be impractical over the development of the project due to the small size of recent PDB and the need for large samples size on individual new prediction modalities.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants		

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.

条款与条件 Springer Nature期刊内容，由Springer Nature客户服务中心有限公司（“Springer Nature”）提供。

施普林格·自然支持作者、订阅者和授权用户（“用户”）合理分享研究论文，前提是这种分享仅限于小规模的个人非商业用途，并且保留所有版权、商标和其他专有权利声明。通过访问、分享、接收或以其他方式使用施普林格·自然期刊内容，您同意这些使用条款（“条款”）。在此方面，施普林格·自然认为学术用途（由研究人员和学生使用）属于非商业性质。

这些条款是补充性的，并将与任何适用的网站条款和条件、相关站点许可证或个人订阅一同适用。在相关条款、站点许可证或个人订阅之间存在冲突或模糊的情况下，这些条款将优先适用（仅限于冲突或模糊的范围）。对于采用Creative Commons许可的文章，将适用所使用的Creative Commons许可的条款。

我们收集和使用个人数据，以提供访问Springer Nature期刊内容的服务。我们还可能在ResearchGate和Springer Nature内部使用这些个人数据，并按照约定以匿名方式共享，用于跟踪、分析和报告的目的。除非在隐私政策中详细说明并获得您的许可，否则我们不会将您的个人数据披露给ResearchGate或Springer Nature集团以外的公司。

尽管用户可以将Springer Nature的期刊内容用于小规模、个人的非商业用途，但需要注意的是，用户不得：

将此类内容用于向其他用户提供定期或大规模访问，或作为绕过访问控制手段的目的；

1.

在任何司法管辖区内进行此类行为将被视为刑事或法定犯罪，或引发民事责任，或

2.

属于其他非法行为；

虚假或误导性地暗示或表明得到认可、批准、赞助或关联，除非斯普林格自然明确书面同意；

3.

使用机器人或其他自动化方法访问内容或重定向消息；

4.

覆盖任何安全功能或排他性协议；或

5.

分享内容以创建斯普林格自然产品或服务的替代品，或系统化斯普林格自然期刊内容的全面数据库。

6.

根据禁止商业使用的限制，施普林格·自然不允许创建任何通过我们的内容或其作为付费服务的一部分或其他商业利益而产生收入、版税、租金或收益的产品或服务。施普林格·自然期刊内容不得用于馆际互借，图书馆员也不得大规模将施普林格·自然期刊内容上传到其或其他任何机构的存储库中。

这些使用条款会定期审查，并可能随时修订。施普林格·自然没有义务在网站上发布任何信息或内容，并可自行决定随时移除这些内容或功能，无论是否事先通知。施普林格·自然亦可随时撤销您对此许可，并移除您已保存的施普林格·自然期刊内容副本的访问权限。

在法律允许的最大范围内，施普林格·自然不对用户作出任何明示或暗示的保证、声明或担保，

关于施普林格·自然期刊内容，所有相关方均否认并放弃任何暗示的担保或法律施加的担保，

包括适销性或适用于任何特定目的。

onlineservice@springernature.co

请注意，这些权利并不自动延伸至施普林格·自然所发布的内容、数据或其他可能从第三方获得许可的材料。

如果您希望以超出本条款明确允许的范围，例如定期或以其他方式向更广泛的受众使用或分发我们的Springer Nature期刊内容，请与Springer Nature联系。