通过蛋白质-配体复合物生成器进行药物筛选分层张量变换

Jonathan P. Mailoa,[1†*] Zhaofeng Ye,[1†] Jiezhong Qiu,[1] Chang-Yu Hsieh,[1] 和 Shengyu Zhang[2*]

1) 腾讯量子实验室，腾讯，中国广东深圳 2) 腾讯量子实验室，腾讯，中国香港特别行政区

† 这些作者对本研究贡献相同

* 通讯作者：jpmailoa@alum.mit.edu，shengyzhang@tencent.com

基于分层张量变换的蛋白质-配体复合物生成器与药物筛选

摘要

生成小分子候选物（配体）在其靶蛋白口袋中的结合构象，对计算机辅助药物发现至关重要。典型的刚体对接方法忽略了蛋白质口袋的柔性，而采用分子动力学生成更准确构象的方法则受限于缓慢的蛋白质动力学。我们开发了一种分层张量变换（3T）算法，可快速生成多样化的蛋白质-配体复合物构象，用于药物筛选中的构象和亲和力估计；该方法无需机器学习训练，也无需耗时的动力学计算，同时兼顾类粗粒化协同蛋白质动力学和复合物口袋的原子级细节。与使用数百种实验蛋白质构象的传统集成对接相比，我们生成的3T构象结构在活性配体分类方面准确度显著更高。此外，我们还证明，3T可用于探索蛋白质口袋中相距较远的蛋白质-配体结合构象。3T结构变换与系统物理学相互独立，因此未来有望应用于其他计算科学领域。

引言

结构生成与优化是生命科学与物理科学领域的重要课题。生成模型算法在这些领域的应用多种多样，涵盖降水临近预报[1] 和翼型气动优化[2–4] ，以及光学纳米结构[5–7]、电极微结构[8] 、微流控器件[9,10] 和材料设计[11–13]的优化。结构生成能力在药物发现应用中更为普遍且至关重要[14–20] ，因为在确定合适的靶标特异性药物分子时，人们关注的是蛋白质和候选药物小分子（配体）的结构。深度学习领域的近期进展使生成模型算法能够更准确地预测单体蛋白质结构[21,22] ，其中一个显著的近期例子是 AlphaFold 2 在 CASP14竞赛中的成功[23] 。其他值得关注的例子也分别聚焦于生成具有理想特性的小分子[24,25] ，或对蛋白质结构进行高效采样（RiD、RIP、L-RIP、TAMD 等）[26–31]。

使用配体靶向蛋白质的特定结合口袋会带来额外挑战，因为该化合物由多个相互作用的结构组成。当柔性配体分子对接到柔性蛋白质受体口袋时，由于这一双体系复合物具有多重柔性，要找到其结合姿势是一项艰巨的任务。最常见的方法（计算成本较低）是冻结目标蛋白质受体口袋，并使用 AutoDock Vina 等分子对接软件，[32] smina，[33] 以及 Glide[34] 将配体分子对接到刚性蛋白质口袋上[35]并生成多个候选配体对接构象。遗憾的是，已知对接到刚性蛋白质口袋的配体并不能代表配体在真实蛋白质-配体复合物中的形态，因此难以用于活性配体分类，[36] 部分原因在于蛋白质结构具有柔性，能够发生内在或诱导的构象变化。[37,38] 针对这一问题，最先进的解决方案是进行集成对接，即使用同一蛋白质口袋的多个结构（通常为10^2 – 10^3 个结构），这些结构可通过实验获得或生成

长时间的分子动力学模拟（MD，模拟时长达毫秒，对应~10^{12} 个 MD 时间步）[36,39] 及其他生成方法；[40] 随后将感兴趣的配体对接到所有这些蛋白质结构上，以生成蛋白质-配体复合物的多种构象。值得注意的是，蛋白质结构生成时并未考虑特定配体是否存在于蛋白质口袋中，因为蛋白质-配体复合物具有综合性的多重柔性，同时修改蛋白质和配体结构仍然十分困难。据作者所知，对整个蛋白质-配体复合物结构进行分子动力学模拟，仍是采样并生成整个蛋白质-配体口袋多样化复合物构象的唯一可靠方法。[41,42] 遗憾的是，由于蛋白质动力学较慢，这种方法的计算成本很高，因此通常仅用于药物筛选工作流程的最后一步，例如自由能微扰（FEP）方法；在一台配备 4 个图形处理器的 GPU服务器上，每天可对 2–4 个配体进行评分，[43,44] 不过，也有人尝试将该方法用于规模更大的初筛步骤。[45] 值得注意的是，还有一些元动力学物理路径方法可用于偏置蛋白质-配体口袋的分子动力学模拟，以加速非平衡分子动力学结构采样，例如漏斗元动力学（FM）、伞形采样（US）和牵引分子动力学（SMD）。[46–50]

在这项工作中，我们提出了分层张量变换（3T）算法，这是一种通用框架，旨在为优化生成多尺度复杂系统多样的物理或生物构象结构。我们通过蛋白质-配体口袋复合物构象生成这一问题展示该算法的用法：同时生成整个蛋白质-配体复合物的多个构象。3T需要一个待优化结构的示例作为初始起点（**图1a**），随后以层次化方式将其分割为局部群组，并视情况划分成若干较小的微组和较大的宏组。在物理学或生命科学领域，基于连通性的分割是最合适的选择。我们依据可旋转键、蛋白质残基和蛋白质二级结构对蛋白质-配体复合物进行分割（**补充图1**）。然后，我们以层次化方式对这些微组和宏组应用多种结构张量变换，从而实现宏观尺度和微观尺度的

优化，从而更有效地摆脱局部能量极小值，并采样多个不同的蛋白质-配体口袋复合物构象。这些层次张量变换参数根据代价函数梯度（力场能量，见**图1b**）进行更新。我们的方法既不是机器学习（ML）方法，也不是分子动力学（MD）方法，但在结构变换部分广泛使用深度学习中常见的工具（PyTorch[51]），并在结构评估部分使用分子动力学中的原子级力场。[52–55] 只要具备可微的领域专用结构评估代价函数，这种模块化设计便可将类似3T的算法用于其他类型的复杂多尺度结构生成和优化（尤其有助于物理信息深度学习研究）。[56] 3T方法在口袋结构生成过程中允许蛋白质主链和侧链基团完全灵活地变化（**图1c**），不同于 smina、rDock、GOLD、FLAP、GRID 和 ICM 等最先进工具所采用的半柔性蛋白质-配体对接方法；这些方法仅允许选定的蛋白质侧链进行二面角旋转。，[57–61] 这些半柔性工具的复杂程度各不相同，从仅允许侧链的–OH和–NH_3 基团[60] 旋转，到穷尽尝试大多数可用的侧链旋转。[33] 虽然从理论上说，3T方法实现蛋白质完全柔性所需的计算量应远高于半柔性方法，但它通过类似粗粒化分子动力学的层次结构变换，快速排除能量上不利的蛋白质-配体构象，因此得以实现。我们利用3T为三个有代表性的蛋白质靶标生成蛋白质-配体共晶结构构象，分别为细胞周期蛋白依赖性激酶2（CDK2）、热休克蛋白90（HSP90）和凝血因子X（FXa）。与相当的分子动力学模拟相比，3T结构生成所需的计算成本可降低超过80× 与。我们进一步展示，在DUD-E和DEKOIS 2.0数据集中，与在数百个实验获得的刚性蛋白质主体构象上执行的最先进集合对接流程相比，使用这十个3T蛋白质-配体复合物构象来识别蛋白质靶标的活性配体和诱饵配体，效果更优；[62,63] 这在一定程度上是因为我们还能获得结合姿势局部能量附近的3T口袋能量景观

最低能结构。最后，我们展示了 3T 探索大型蛋白质和配体构象空间的能力，包括探索替代结合模式。

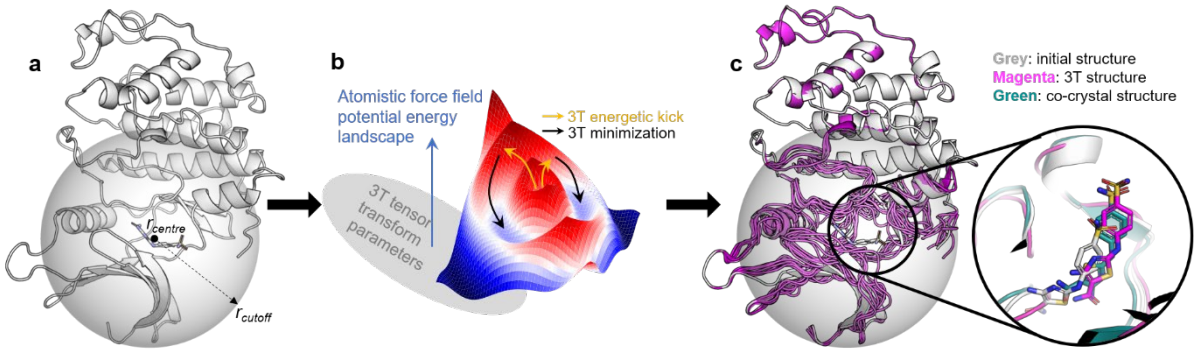


图 1 | 3T结构生成与分析流程图。a) 初始蛋白质-配体口袋体积定义。 **b)** 在分层张量变换参数超空间中进行 3T能量扰动和优化。**c)** 由 3T 生成的蛋白质-配体口袋结构（为清晰起见，仅显示一个配体构象），并与初始对接结构及相应的实验蛋白质-配体共晶结构进行比较。

结果与讨论

分层张量变换算法

原则上，3T 是一种生成式算法，通过对现有结构进行多尺度张量变换来发挥作用，并受基于物理的代价函数约束，以确保生成的多尺度变换结构符合物理现实，并具有功能相关性或经过优化。3T 包含三个主要组成部分：分层结构分割、分层张量变换分配，以及可微结构评估代价函数。

分层结构分割是必要的，因为我们希望能够对初始结构执行局部变换操作。已知蛋白质-配体复合物具有大量高维局部能量陷阱，因此在刚性蛋白质结构上进行配体对接更为实际[35]，或者至多在半柔性蛋白质结构上进行。[59,60]直接对蛋白质和配体原子的原子坐标执行结构优化，只会产生相对较小的原子移动。结构分割通过将局部相连的原子归入不同组别，并使这些原子以协调方式移动，从而消除部分这一问题。这种分组方式类似于粗粒化分子动力学中的做法，[64] 但3T保留了完整的原子级细节。分组还以层次方式进行，以实现不同层级的协调移动，包括蛋白质骨架。对于蛋白质-配体口袋复合物生成，我们将原子分为2个分割级别：微组和宏组。结合口袋（位于蛋白质-配体复合物中心的截断半径 $r_{cutoff} = 20\text{Å}$内）的原子被视为可移动原子，随后被分割成不同的微组：(1) 对于蛋白质原子，根据其氨基酸残基以及主链/侧链区分对微组进行分割；(2) 对于配体原子，根据其可旋转键对微组进行分割。随后，这些微组会进一步归入不同的宏组：(1) 每个柔性环二级结构（**方法**）中的蛋白质微组分别获得独立的宏组-

组分配，而其余微组（螺旋和片层）不分配到任何宏组中；(2) 所有配体微组均被分配到一个配体宏组中。至此，我们完成了两级3T分层分割（**补充图1**）。原则上，可以根据所研究系统的物理性质使用更多的分层分割级别。例如，如果某个配体片段没有可旋转键，但希望获得更多构象（环己烷基团就是一个例子），那么最好添加不同类型的微组分割层级，以便对感兴趣的子结构进行快速3T多尺度构象采样。

随后，我们为每个微组和宏组分别指定分层局部结构张量变换（**图2**，更直观的版本见**补充图2**）。同一微组*i*中的原子通过旋转或平移张量变换参数进行变换$\theta_{R,\ i}$分别为$\theta_{T,\ i}$。$\theta_{R,\ i}$和$\theta_{T,\ i}$每个参数各包含3个标量参数，分别对应绕微组中心所在的*x*、*y*和*z*轴的旋转角以及沿这些轴的平移。这些张量变换使可移动原子在3T优化过程中，除了各自的原子平移外，还拥有额外的协同旋转和平移自由度。我们还对仅具有一个可旋转键（锚点）的微组应用特殊的轴旋转变换$\theta_{A,\ i}$（由1个标量参数表示），使这些微组能够绕锚点自由旋转。通过对每个宏组*j*进行协同旋转和平移，还可实现更大尺度的3T协同运动，所用变换参数分别为$\theta_{R,\ j}$和$\theta_{T,\ j}$（与对应的微组参数一样，每个参数都包含3个标量参数）。

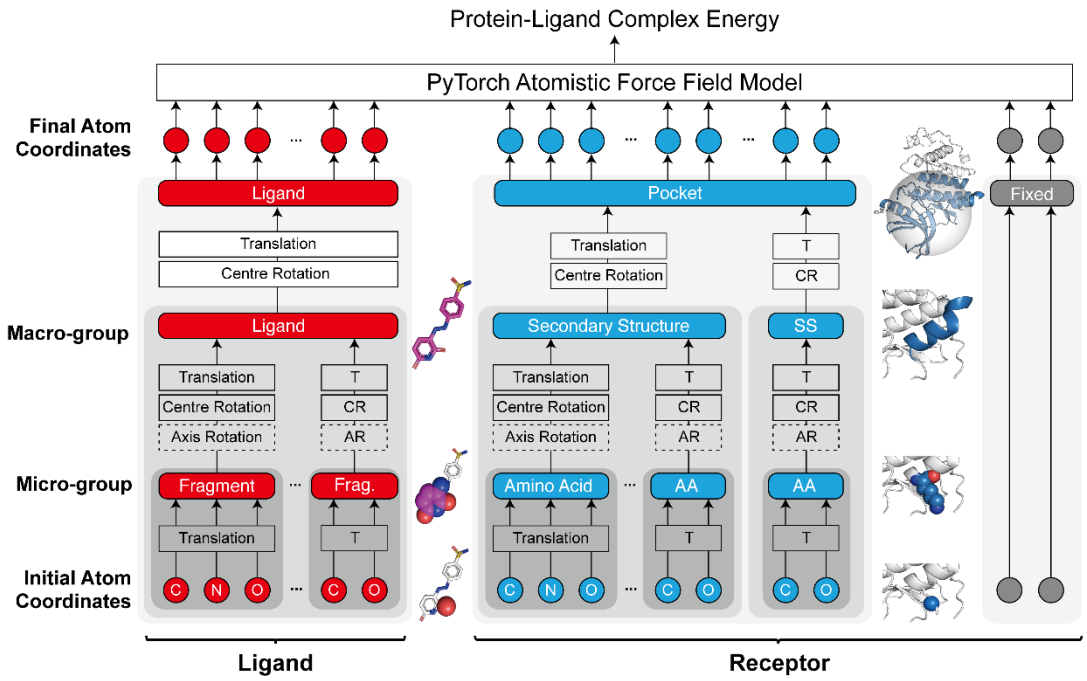


图2 | 3T方案示意图。 蛋白质-配体复合物中的原子分为固定原子（灰色圆圈）和可移动原子

原子（红色/蓝色圆圈）。可移动原子会经历多层多尺度分层张量变换（单个原子平移、侧链绕可旋转键轴进行微基团旋转、微基团笛卡尔轴旋转和平移，以及宏基团笛卡尔轴旋转和平移）。绕可旋转键进行的微基团轴旋转（虚线框）仅适用于某些微组，例如蛋白质侧链和配体边缘片段。我们并非使用原子级力场直接优化最终原子坐标，而是在 PyTorch 成本函数梯度反向传播过程中优化3T张量参数和初始可移动原子坐标。从计算角度看，这与典型的深度学习训练流程类似。有关分层结构变换的更直观示意，请参见**补充图2**。

完成这些层次分割和变换分配后，使用可微结构评估代价函数，3T优化流程本身就相对直接。我们从初始的固定原子坐标和可移动原子坐标 $\vec{r}_f$ 和 $\vec{r}_{m,\ init}$ 开始，让 $\vec{r}_{m,\ init}$ 依次经过由上述变换参数 θ's 控制的多个张量变换阶段，并计算最终的可移动原子坐标 $\vec{r}_{m,\ final}$。我们使用 PyTorch 实现张量变换和原子力场模型（**方法**），根据 $\vec{r}_f$ 和 $\vec{r}_{m,\ final}$ 计算系统总能量。我们使用 PyTorch 而非 LAMMPS[65] ，因为 PyTorch 自动微分能够灵活地执行自动微分，从而累积 3T多尺度变换参数的梯度；而 LAMMPS 不具备这种自动微分能力。以该力场能量作为主要成本函数，通过反向传播更新 3T 模型的可调参数 $\vec{r}_{m,\ init}$ 和 θ's。简而言之，我们将结构生成成本函数的计算从原始形式（**方程1**）转换为 3T 版本（**方程2)**：

$$C_{E,original} = E_{FF}\left(\vec{r}_f\ , \vec{r}_{m,init}\right) \tag{1}$$

$$C_{E,3T} = E_{FF}\left(3T\left(\vec{r}_f\ , \vec{r}_{m,init}, \theta_{R,i}, \theta_{T,i}, \theta_{A,i}, \theta_{R,j}, \theta_{T,j}\right)\right) \tag{2}$$

其中，C_E 是蛋白质-配体复合物的成本函数，E_{FF} 是原子力场能量函数，而$3T$ 是**图2**中所示的层次张量变换函数。

优化过程中，我们首先在刚性蛋白质口袋中优化配体n_{step} = 200 步。随后，我们对系统施加能量扰动，通过为微基团 θ 初始化较小的随机值，开始生成蛋白质-配体复合物口袋构象。能量扰动会使结构变形至高能态，随后通过n_{step} = 2000 步反向传播，由 3T 将其最小化。不同的随机数种子会生成不同的 3T能量扰动和最终结构（**方法**）。从计算角度看，这一最小化过程类似于标准的深度学习训练（其成本函数中没有分类标签或回归目标）。虽然应用 3T结构变换后，可优化参数的数量会增加，

但由于我们拥有额外协调的微观和宏观自由度，因此更容易摆脱局部能量极小值，也能更快达到低能平衡态。

生成的蛋白质-配体口袋结构的评估

为了评估构象结构生成的质量，我们将这些3T结构与特定蛋白质-配体复合物现有的实验共晶结构进行比较。在本研究中，我们为三种不同的蛋白质生成了3T蛋白质-配体复合物构象：CDK2、HSP90和FXa。CDK2的结合口袋是一个灵活而深的疏水腔；HSP90的结合口袋周围有两条较长的α螺旋，已知它们会在不同配体结合时发生多次显著的构象变化，[66] 而FXa的活性位点是一个灵活而浅的疏水沟槽。对于这些蛋白质，我们分别从蛋白质数据库（PDB）中提取一个蛋白质结构示例：1fin、1uyg和1ezq（见**图3a**）。

首先，我们以3T生成的CDK2构象为例，分别进行交叉对接结构分析。使用smina[33] 将4ez3 PDB共晶结构中的配体交叉对接到我们的1fin蛋白质上，[40] 以获得初始结构（见**方法**）。我们计算该交叉对接配体与原始1fin共晶配体相比的均方根偏差（RMSD），得到 $RMSD_{init}$ = 3.89 Å。该值越大，预测的交叉对接结构与真实共晶结构的差异就越大。随后，我们通过3T转换这一初始结构（见**方法**），生成新的蛋白质-配体复合物构象。从**图3b**可以看出，与初始结构相比，3T配体构象与原始1fin共晶配体构象更加吻合，均方根偏差更小（ $RMSD_{3T}$ = 1.86 Å），且蛋白质结合位点与配体的结合更加准确；而初始的smina结构使用的是刚性1fin蛋白质。相应地，均方根偏差改善了 $\Delta RMSD = RMSD_{3T} - RMSD_{init}$ = -2.03 Å。这种改善是由蛋白质骨架构象变化（红圈所示）引起的，该变化使其更接近实际的4ezq共晶蛋白质，并轻微推动配体。详情见**补充图3a**。

我们进一步分析了 3T 生成的 CDK2 蛋白构象，并将其与已知共晶结构及分子动力学模拟中发现的构象进行比较。从蛋白质数据库网站提取了 373 个 CDK2 共晶蛋白质结构。

从蛋白质数据库网站提取。每隔 1 ns 从 1fin 蛋白质-配体结构在水中的一次分子动力学模拟中提取 500 个蛋白质结构（见 **方法**）。最后，从 1fin smina 重新对接的初始结构所生成的

法），图中显示，共晶结构涵盖的蛋白质骨架构象范围远为多样，这是因为自然界中存在多

能量最小化方案，因此并未覆盖 1fin 室温分子动力学模拟构象的整个主成分空间。HSP90 和

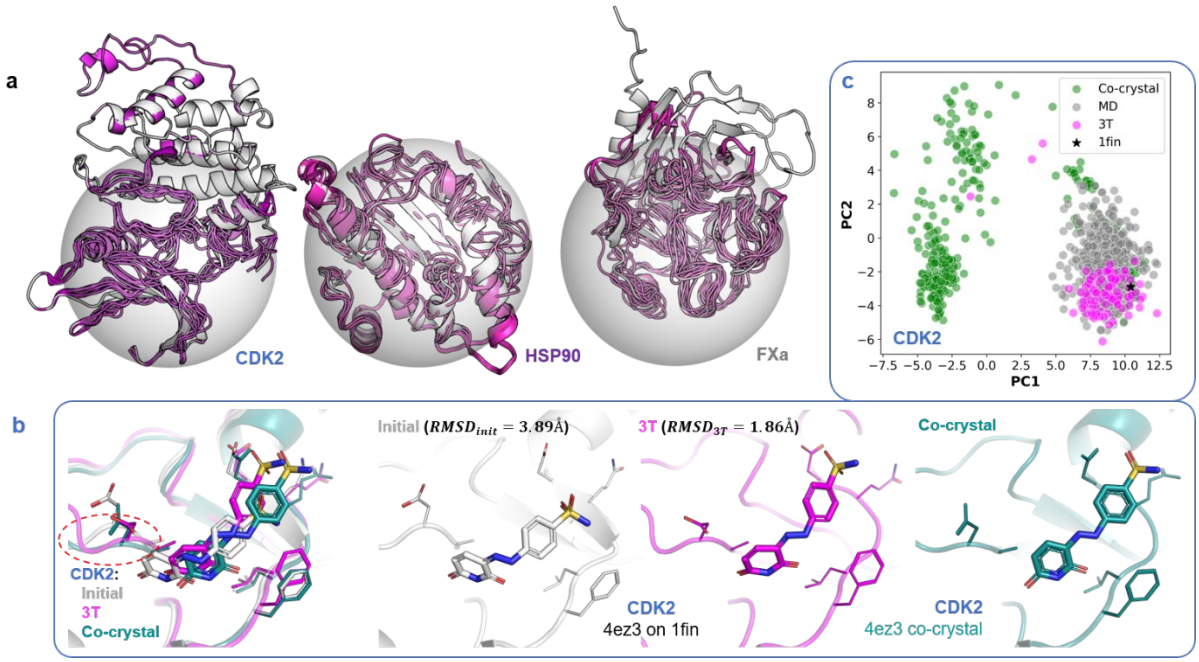


图 3 | 蛋白质-配体 使用 3T 生成蛋白质-配体复合物构象。 a) 配体依赖的结合口袋构象示例 ns

本研究生成的三种蛋白质结构，其中 CDK2 的柔性最高，HSP90 的刚性最高。为便于观察，配体几何结构已隐藏。**b)** 对 CDK2蛋白结构进行交叉对接的叠加可视化比较（将来自 4ez3 PDB 的配体结构交叉对接至来自 1fin PDB 的蛋白质结构），比较了使用 3T 和标准刚性蛋白质对接的结果，并以真实值 4ez3共晶结构为参照。红色圆圈标示出经 3T 正确转换后、与共晶结构显著更加相似的蛋白质主链结构；这也使 3T 交叉对接得到的配体比初始交叉对接结构更接近实验共晶配体构象。为便于观察，图中还展示了各个结构（灰色：初始结构；洋红色：3T；绿色：共晶结构），表明 3T 结构的均方根偏差更小，且与实验共晶结构的匹配度更高（配体构象和蛋白质结合位点）。**c)** 使用主成分分析比较 CDK2蛋白主链的微观状态：将 3T 生成的结构（对 1fin PDB 进行重新对接）与通过 500 ns 蛋白质-配体复合物分子动力学生成的 1fin 口袋结构，以及所有已知的实验 CDK2 共晶结构进行比较。由于 CDK2 存在多种主要的配体依赖性蛋白质构象，分子动力学仅覆盖了整个 CDK2 主成分子空间的一部分。3T 正确覆盖了与 1fin 分子动力学对应的子空间，而未覆盖其余无法通过 1fin 蛋白质-配体复合物实际达到的子空间。实验 1fin 共晶结构以黑色星号表示。如果使用一个初始共晶结构进行半柔性交叉对接集合，主成分在此处只会显示为一个点，因为与 3T 和分子动力学方法不同，蛋白质主链无法移动。

在活性配体分类中的适用性

我们进一步评估了 3T 复合物构象在活性配体分类中的实际效用，并将其与使用传统方法获得的复合物口袋结构进行比较。近期研究表明，将候选药物配体分子对接至单一刚性蛋白质口袋，不足以实现活性配体分类。[36,39] 实际上，需要将配体对接至数百种不同的靶蛋白口袋刚性构象。[36,39] 仅评估简单的对接评分并不足够，还需要基于集成对接评分构建机器学习模型，才能获得性能良好的活性配体分类器。[39] 这些不同的蛋白质构象结构难以获取或获取成本高昂，因此，配体（*A*）通常会被对接至不匹配的刚性蛋白质结构（*B*）（取自不同的实验蛋白质-配体（*B*-*C*）复合物），或对接至通过在溶剂中对蛋白质口袋进行长时间分子动力学模拟而生成的刚性蛋白质结构（*D*）。另一方面，3T 结构生成使我们能够明确生成依赖于配体、并针对每一对蛋白质-配体量身定制的蛋白质-配体复合物口袋构象。

我们通过进行与 Ricci-Lopez等人类似的机器学习辅助“集成对接”，展示了这种多功能性。[39] 在先前的研究中，根据数据集的不同，配体对接采用了数量不等的刚性蛋白质结构（CDK2：402，HSP90：64，FXa：136）。对于每个蛋白质宿主数据集中的每种配体，我们使用 smina[33] 将其与各自的蛋白质靶标结构进行分子交叉对接，以获得用于生成 3T 的一个初始结构（**方法**）。对于每个初始蛋白质-配体复合物，我们通过 3T 能量最小化生成 10 种构象。在本研究中，我们仅为对接到 CDK2、HSP90 和 FXa 各自一种刚性蛋白质上的每种配体生成 10 个 3T 构象结构。我们采用与先前研究相同的 30×4 折交叉验证（30×4cv）和梯度提升树（GBT）分类器算法，以确保比较的仅是构象特征质量，而非所用的分类方法。[39] 我们还指出，3T 不仅能生成蛋白质-配体复合物口袋构象，还能生成结构优化过程中局部能量极小值周围的势能景观

过程，可将其用作额外特征。我们假设，决定配体与靶蛋白口袋结合可能性的因素，并不只是蛋白质-配体口袋结构的形状（例如对接评分），还包括蛋白质-配体口袋能量极小值的可及性（局部能量极小值周围的能垒景观），如**补充图4**所示。该特征提取过程及后续的 30×4cv 分类过程见**图 4a**。对于我们为每个蛋白质-配体复合物生成的 10 种 3T 构象，我们不仅提取对接评分，还提取蛋白质-配体结合形成能 $\Delta E = E_{complex,\ 3T} - E_{ligand,\ 3T} - E_{protein,\ init}$ 并贯穿整个 3T 优化过程（**方法**）。由于蛋白质尺寸较大，在本分类研究中，CDK2 蛋白质-配体口袋构象使用 $r_{cutoff} = 25$Å重新生成，而 HSP90 和 FXa 均保持为 $r_{cutoff} = 20$Å（**补充表1**）。

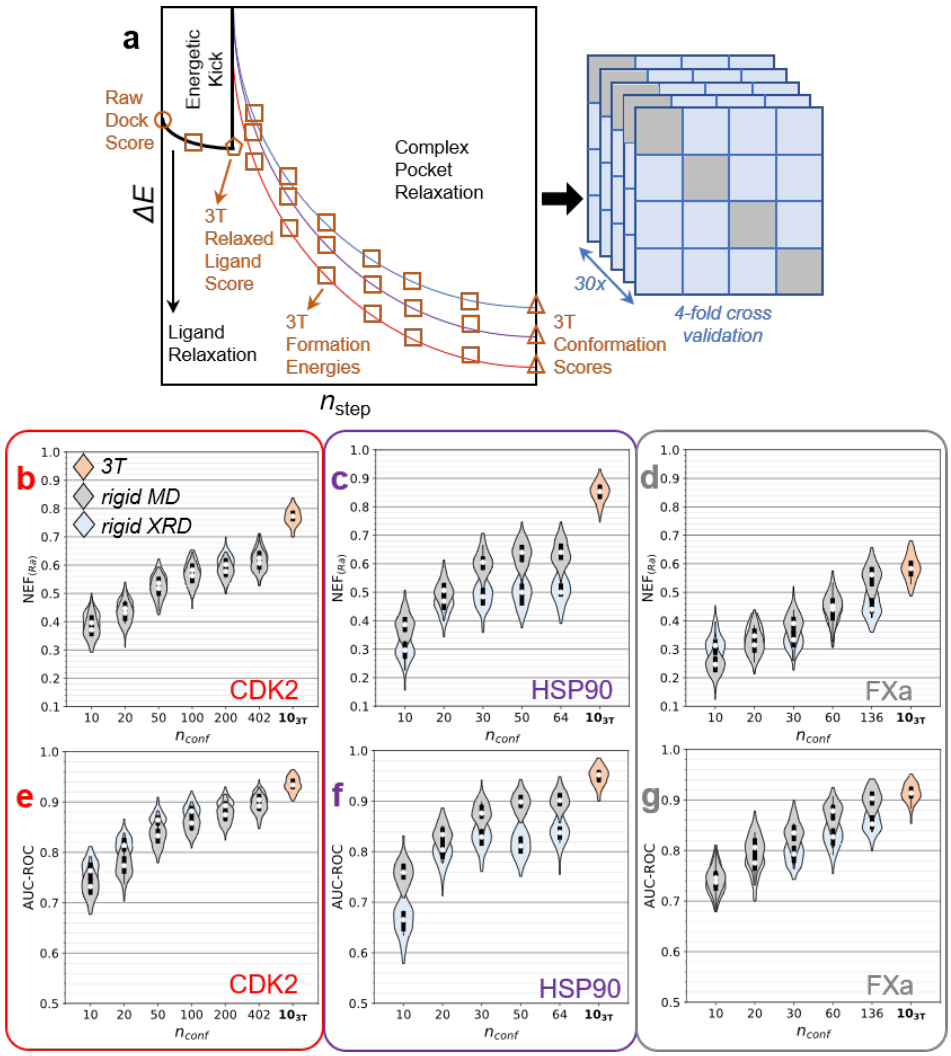


图 4 | 使用3T生成的构象进行活性配体分类。a) 3T构象特征提取过程（结合口袋交叉对接评分和形成能 ΔE），随后进行30× 4折交叉验证（使用梯度提升树分类器）。**b-d)** 图中分别展示了不同数量的CDK2、HSP90和FXa结合口袋构象对应的NEF_{R_a} 分类指标。在**e-g)**中同样展示了ROC曲线下面积指标。这三种蛋白质在结构柔性方面存在差异：CDK2和FXa主要由柔性环构成，而HSP90则主要由α螺旋和β折叠构成。我们发现，从一个实验蛋白质构象生成的10个3T配体依赖性口袋构象中提取的特征，其效果等同于或优于从数量显著更多的刚性实验X射线衍射蛋白质构象（刚性X射线衍射）或模拟MD构象（刚性分子动力学）中提取的特征。

蛋白质	活性配体	指标	3T分类器	刚性分子动力学分类器	刚性X射线衍射分类器	$n_{conf,\ 3T}$ 对比 $n_{conf,\ MD-XRD}$
CDK2	442/3764 ($R_a = 0.117$)	NEF_{R_a}	**0.771 ± 0.030**	0.608 ± 0.033	0.624 ± 0.039	**(10 + 2)** *vs* 402
		ROC曲线下面积	**0.935 ± 0.014**	0.892 ± 0.017	0.904 ± 0.015	
HSP90	298/2452 ($R_a = 0.122$)	NEF_{R_a}	**0.851 ± 0.035**	0.640 ± 0.042	0.505 ± 0.046	**(10 + 2)** *vs* 64
		ROC曲线下面积	**0.949 ± 0.018**	0.903 ± 0.019	0.836 ± 0.024	
FXa	298/7191 ($R_a = 0.040$)	NEF_{R_a}	**0.584 ± 0.043**	0.554 ± 0.046	0.452 ± 0.044	**(10 + 2)** *vs* 136
		ROC曲线下面积	**0.913 ± 0.018**	0.902 ± 0.021	0.855 ± 0.021	

表格1 | 3在 3 种不同蛋白质宿主上使用梯度提升树分类器进行活性配体分类的指标。 通常，用于标准集成交叉对接分类器的n_{conf} 越多，分类器的表现就越好。我们展示了，尽管标准构象分类器使用的刚性实验蛋白质构象结构显著更多，3T构象分类器（10 + 2 = 1 初始+ 1 加上松弛配体+ 10 能量踢动构象）在 3 种不同蛋白质宿主上的表现始终优于刚性蛋白质构象分类器。

为实现与现有研究的公平比较，并直接比较不同的数据集样本分布，我们采用曲线下面积——受试者工作特征（ROC曲线下面积）指标，以及归一化富集因子 $NEF_{\chi} = a_s \ min($ $/ \quad \chi m \ a \ ,)$，其中 m 是数据集中配体的总数，χ 是选取的排名靠前配体所占的比例（设为 $\chi = R_a = a/m$），a 是真正具有活性的配体总数，a_s 是选出的 χm 配体中真正具有活性的配体总数。[39,68] 为了研究与实验获得的蛋白质构象相比，我们的蛋白质-配体复合物口袋构象有何用处，我们针对三种蛋白质，基于不同数量的刚性实验 X 射线衍射蛋白质构象宿主（刚性X射线衍射），重新计算了先前研究[39] 的 NEF_{R_a} 。[39] 此外，我们还对水中的蛋白质进行长时间分子动力学模拟，提取其结构，作为所有配体对接共用的刚性蛋白质宿主，并按照另一项先前研究的方法，计算这一基于分子动力学的参考（刚性分子动力学）对应的分类指标。[36] 需要指出的是，这些分子动力学结构不依赖于配体，因为在本研究中针对每个明确的蛋白质-配体对分别进行分子动力学模拟（Holo-MD）在计算上将难以承受。支持我们假设的ROC曲线下面积结果见图4b：使用十个3T蛋白质-配体复合物构象构建的“集合对接”CDK2活性配体分类器，显著优于使用402个刚性蛋白质构象宿主（包括刚性分子动力学和刚性X射线衍射）构建的集合对接CDK2活性配体分类器。使用各自的十个3T构象，我们也分别优于基于64个HSP90和136个FXa刚性蛋白质构象构建的分类器（图4c-d、表1）。这些分类器的NEF_{R_a} 指标与ROC曲线下面积指标一致（图4e-g），表明3T构象显著优于相应的刚性构象，进一步证明了3T构象在分类方面的效用。在三种蛋白质中，HSP90的性能提升尤为显著；它的蛋白质口袋结构是其中柔性最低的。我们还表明，在生成3T构象时限制或消除蛋白质结构的柔性，会显著降低3T分类器的性能（补充图5、补充表1）。

虽然乍看之下，使用数量为 $n_{conf,3T}$ = 10 、源自 n_{conf} = 1 个蛋白质宿主构象结构的构象结构构建的分类器，其分类结果竟能与使用大量实验蛋白质宿主构象结构（n_{conf} = 64–402）构建的分类器相近甚至更好，似乎不太合乎直觉，但需要指出的是，3T结构依赖于配体，其特征包含更多信息（**方法**）。还需指出的是，尽管每个配体对应的3T构象各不相同且具有随机性（这使得分类器特征的使用更难以解释），我们通过确保不同配体使用相同的3T随机种子来缓解这一问题。这意味着，在结构经过弛豫、形成最终的蛋白质-配体复合物构象几何结构之前，各配体经历的初始蛋白质结构畸变几乎完全相同。最后需要指出的是，使用 $n_{conf,3T}$ = 4 的3T分类器实际上已足以优于刚性分子动力学和基于刚性X射线衍射的分类器（**补充图4**）。

除了准确性显著高于传统方法且所需实验资源显著更少之外，与 Holo-MD 或穷举式smina半柔性对接方法相比，3T 所需的计算资源也显著更少[33]（参见**补充表2**），不过这个 3T 原型版本的速度仍慢于轻量级半柔性对接方法（如 rDock）；后者仅允许侧链上的羟基和氨基$_3$ 基团旋转。[60] 如果使用 Holo-MD 生成蛋白质-配体复合物，3T结构生成所需的计算资源比 Holo-MD 少逾80×（基于激进 MD 假设，参见**方法**）。对每种配体采用这种 Holo-MD 方法在计算上难以实现。降低成本的一种方法是在刚性蛋白质对接（刚性分子动力学）之前进行蛋白质口袋MD，并在所有目标配体之间分摊计算成本。在这种刚性分子动力学对接情景下，与使用实验结构进行刚性蛋白质对接相比，很难取得更好的活性配体分类性能，尤其是在存在多个主要蛋白质构象时，例如 CDK2（**图3c**）。有关这些方法差异的更多详情，请参见**补充流程图1**。

远距离构象与口袋采样

我们还展示了 3T 生成与初始结构相距甚远、难度显著更高的构象的能力。这一能力有助于探索不同且相距较远的蛋白质口袋中的中间态亚稳态配体构象（相距~10Å）。这些任务需要采用激进得多的 3T能量扰动和优化工作流程（详情参见**补充信息**），而非本文前文用于大规模虚拟筛选的温和设置。在**图5**中，我们利用 3T 研究配体（阿普洛尔）和 G蛋白偶联受体蛋白（β2-肾上腺素能受体，ADRB2）的构象变化。正如先前报道的那样，配体会结合到位于细胞外前庭、通向正构结合位点入口的不同蛋白质口袋中（**图5a**）。[69] 我们展示了更激进的 3T 设置如何使原本位于 ADRB2 正构结合位点的阿普洛尔分子跃迁至细胞外前庭结合位点，探索暂时性结合模式。Phe193 和 Tyr308 是构象变化的关键能量屏障。在初始阶段，阿普洛尔结合于表面，此时 Phe193 和 Tyr308 封闭了隧道（**图5b**）。随着阿普洛尔深入隧道，Phe193 和 Tyr308 会发生显著的构象变化，打开入口（**图5c- d**）。阿普洛尔到达正构结合位点后，Phe193 和 Tyr308 会再次关闭隧道，从而稳定配体结合（**图5e**）。在第二个示例中，3T结构生成过程中的蛋白质柔性尤为重要，因为蛋白质结构中的 Phe193 和 Tyr308 侧链充当门户，控制细胞外前庭结合位点与正构结合位点之间的通行。

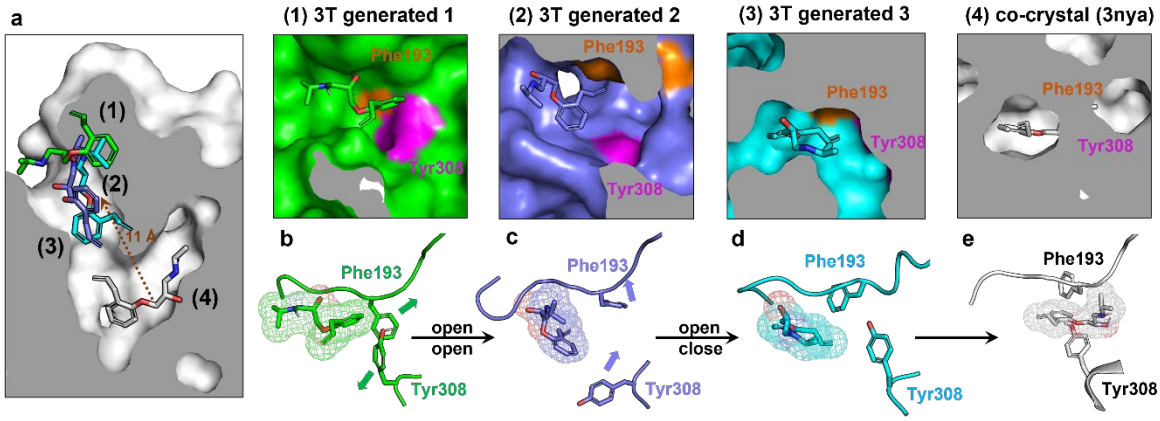


图 5 | E　探索远距离姿势和结合口袋位置的蛋白质-配体构象。 a)阿普洛尔的姿势　　ol

叠加在G蛋白偶联受体3nya共晶结构上的ADRB2前庭结合位点（1–3）和正构受体结合位点（4）中的分子。前庭结合位点的3T构象（**b-d**）是通过对原始正构位点共晶构象（**e**）施加强烈的能量扰动生成的。在3T构象中，我们注意到，这些3T蛋白质-配体构象也对应于Phe193和Tyr308侧链基团依次打开和关闭通往正构结合口袋的通道。在此前长达5 μs的分子动力学模拟中，曾观察到配体处于前庭结合位点的构象。[69]

结论

总之，我们展示了一种名为分层张量变换（3T）的新算法，可用于生成逼真的复杂多尺度结构，例如蛋白质-配体复合物构象。该结构生成方法结合了一个示例初始结构、一个可微结构评估代价函数、一系列层次化多尺度张量变换，以及用于初始结构畸变的随机能量扰动。利用此方法，我们可以针对给定的蛋白质靶标和候选药物配体分子，生成独特的蛋白质-配体复合物构象。我们证明，这些3T构象有助于活性配体分类。来自十种3T构象的特征显著优于来自数百种刚性蛋白质构象的特征，而且生成这些特征所需的计算成本比在图形处理器上进行可比的分子动力学模拟低80× 以上 相较于。我们还进一步证明了3T探索远端蛋白质口袋结合位点及蛋白质-配体口袋中间过渡态的能力。得益于3T的模块化特性，只要存在相对低成本的结构评估代价函数，将其应用于物理科学的其他领域（例如光学纳米结构或微流控结构的生成与优化）应当十分直接。

数据可用性

本文提供了源数据。本研究中使用的所有蛋白质和配体输入结构，以及用于生成**图3**、**图4**和**图5**的原始输出数据，均可在腾讯量子实验室的 GitHub 上公开获取：https://www.github.com/tencent-quantum-lab/3T。由于数据量庞大（本研究总共生成了超过 20 万个蛋白质-配体结构），同一页面仅提供少量生成的蛋白质-配体输出结构示例。其余部分输出结构可在合理请求后向作者索取。

代码可用性

用于构建**图3**、**图4**和**图5**所需的3T结构生成代码，以及用于RMSD分析和活性配体分类任务的预处理与后处理代码，均已公开发布在腾讯量子实验室的GitHub上：https://www.github.com/tencent-quantum-lab/3T。

参考文献

1. Ravuri, S. 等 使用雷达深度生成模型进行高精度降水临近预报。《自然》 **597**, 672–677（2021年）。2. Li, J., Zhang, M., Martins, J. R. R. A. & Shu, C. 利用基于深度学习的几何筛选，高效优化气动外形。*AIAA J.* **58**, 4243–4259（2020）。3. Chen, W. & Ramamurthy, A. 用于高效三维翼型参数化与生成的深度生成模型。载于*AIAA Scitech Forum*论文集1–16（2021年）。doi:10.2514/6.2021-1690。4. Chen, W., Chiu, K. & Fuge, M. D. 利用生成对抗网络进行气动设计优化与形状探索。载于*AIAA Scitech Forum*论文集（2019年）。doi:10.2514/6.2019-2351。5. Peurifoy, J. 等 利用人工神经网络进行纳米光子粒子模拟与逆向设计。*Sci. Adv.* **4**, eaar4206（2018年）。6. Kiarashinejad, Y., Abdollahramezani, S. & Adibi, A. 基于降维设计电磁纳米结构的深度学习方法。*npj Comput. Mater.***6**, 12（2020年）。7. Getman, F., Makarenko, M., Burguete-Lopez, A. & Fratalocchi, A. 通过通用逼近器实现宽带矢量超薄光学，在可见光区域的实验效率高达99%。*Light Sci. Appl.* **10**, 47（2021年）。8. Gayon-Lombardo, A., Mosser, L., Brandon, N. P. & Cooper, S. J. 深入探究孔隙：利用生成对抗网络，在周期性边界条件下随机重建三维多相电极微结构。*npj Comput. Mater.* **6**, 82（2020年）。9. Lee, X. Y., Balu, A., Stoecklein, D., Ganapathysubramanian, B. & Sarkar, S. 利用深度强化学习为微流控器件设计流动形状。载于神经信息处理系统进展第32卷 arXiv:1811.12444（2018年）。10. Lee, X. Y., Balu, A., Stoecklein, D., Ganapathysubramanian, B. & Sarkar, S. 深度强化学习在工程设计中的案例研究：在用于流动塑形的微流控器件中的应用。*J. Mech. Des.* **141**, 111401（2019年）。11. Long, T. 等 用于晶体结构逆向设计的受约束晶体深度卷积生成对抗网络。*npj Comput. Mater.* **7**, 66（2021年）。12. Kang, S. 等 利用机器学习势加速识别多组分无机晶体的平衡结构。*npj Comput. Mater.* **8**, 108（2022年）。13. Yao, Z. 等 利用深度生成模型对纳米多孔晶态网状材料进行逆向设计。*Nat. Mach. Intell.* **3**, 76–86（2021年）。14. Jørgensen, P. B., Schmidt, M. N. & Winther, O. 分子科学的深度生成模型。*Mol. Inform.* **37**, 1700133（2018年）。15. Li, Y., Pei, J. & Lai, L. 利用三维深度生成模型进行基于结构的从头药物设计。*Chem. Sci.* **12**, 13664–13675（2021年）。16. Shin, J. E. 等 利用自回归生成模型进行蛋白质设计与变体预测。《自然·通讯》 **12**, 2403（2021年）。17. Trinquier, J., Uguzzoni, G., Pagnani, A., Zamponi, F. & Weigt, M. 利用简单的自回归模型高效生成蛋白质序列。《自然·通讯》 **12**, 5800（2021年）。

18. Gebauer, N. W. A., Gastegger, M., Hessmann, S. S. P., Müller, K. R. & Schütt, K. T. 利用条件生成神经网络逆向设计三维分子结构。《自然·通讯》 **13**, 973 (2022). 19. Thygesen, C. B. 等。使用深度马尔可夫模型高效生成蛋白质结构片段。载于第*38*届国际机器学习大会论文集，第 PMLR139 卷，10258–10267 (2021). 20. Ding, X. & Zhang, B. DeepBAR：一种快速且精确的结合自由能计算方法。《物理化学快报》 **12**, 2509–2515 (2021). 21. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. 用于基于图的蛋白质设计的生成模型。载于《神经信息处理系统进展》，第33卷 (2019). 22. Anand, N. & Huang, P. 蛋白质结构生成建模。载于《神经信息处理系统进展》，第32卷 (2018). 23. Jumper, J. 等。利用AlphaFold实现高度准确的蛋白质结构预测。《自然》 **596**, 583–589 (2021). 24. Chen, Z., Min, M. R., Parthasarathy, S. & Ning, X. 通过单个片段修饰进行分子优化的深度生成模型。*Nat. Mach. Intell.* **3**, 1040–1049 (2021). 25. Skinnider, M. A. 等。深度生成模型实现了新型精神活性物质的自动化结构解析。*Nat. Mach. Intell.* **3**, 973–984 (2021). 26. Wang, D. 等。利用自适应强化动力学高效采样高维自由能景观。*Nat. Comput. Sci.* **2**, 20–29 (2022). 27. Noé, F., Olsson, S., Köhler, J. & Wu, H. 玻尔兹曼生成器：利用深度学习采样多体系统的平衡态。《科学》*(80-.)* **365**, eaaw1147 (2019). 28. Ho, B. K. & Agard, D. A. 通过局部扰动探究蛋白质结构中大幅构象变化的柔性。《公共科学图书馆·计算生物学》 **5**, e1000343 (2009). 29. Kokh, D. B., Czodrowski, P., Rippmann, F. & Wade, R. C. 用于探索蛋白质结合位点柔性并预测瞬时结合口袋的扰动方法。*J. Chem. TheoryComput.* **12**, 4100–4113 (2016). 30. Stank, A. 等。TRAPP网络服务器：预测蛋白质结合位点柔性并检测瞬时结合口袋。*Nucleic Acids Res.* **45**, W325–W330 (2017). 31. Stein, E. G., Rice, L. M. & Brünger, A. T. 扭转角分子动力学：一种用于核磁共振结构计算的高效新工具。*J. Magn. Reson.* **124**, 154–164 (1997). 32. 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与准确性。*J. Comput. Chem.* **31**, 445– 461 (2010). 33. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. 从CSAR 2011基准测试中总结的smina经验评分经验。《化学信息与建模杂志》 **53**, 1893–1904 (2013). 34. Halgren, T. A. 等。Glide：一种快速、准确的对接与评分新方法。2. 数据库筛选中的富集因子。*J. Med. Chem.* **47**, 1750–1759 (2004). 35. Forli, S. 等。使用AutoDock套件进行计算蛋白质-配体对接和虚拟药物筛选。*Nat. Protoc.* **11**, 905–919 (2016).

36. Evangelista Falcon, W., Ellingson, S. R., Smith, J. C. & Baudry, J. 药物发现中的集成对接：需要从分子动力学模拟中获取多少种蛋白质构象，才能重现已知的配体结合？*J. Phys. Chem. B* **123**, 5189–5195 (2019). 37. Gallicchio, E. & Levy, R. M. 用于模拟蛋白质-配体结合亲和力的全原子采样方法进展。*Curr. Opin. Struct. Biol.* **21**, 161–166 (2011). 38. Koshland Jr., D. E. 钥匙-锁理论与诱导契合理论。*Angew. Chemie Int. Ed.English* **33**, 2375–2378 (1994). 39. Ricci-Lopez, J., Aguila, S. A., Gilson, M. K. & Brizuela, C. A. 利用集成对接和机器学习改进基于结构的虚拟筛选。《化学信息与建模杂志》 **61**, 5362–5376 (2021). 40. Wang, A. 等。基于弹性网络模型的集成对接可提高蛋白质-配体结合姿势预测的准确性。《化学信息与建模杂志》 **60**, 2939–2950 (2020). 41. Rodriguez-Bussey, I. G., Doshi, U. & Hamelberg, D. 增强分子动力学采样 o

f

药物靶标构象。《生物聚合物》 **105**, 35–42 (2016年). 42. Guterres, H. & Im, W. 利用高通量分子动力学模拟改进蛋白质-配体对接结果。《化学信息与建模杂志》 **60**, 2189–2198 (2020). 43. Wang, L. 等 通过现代自由能计算方案和力场，在前瞻性药物发现中准确、可靠地预测相对配体结合效力。《美国化学会志》 **137**, 2695–2703 (2015). 44. Fratev, F. & Sirimulla, S. 一种改进的、适用于柔性配体结合结构域的自由能微扰（FEP）+ 采样方案。《科学报告》 **9**, 16829 (2019年). 45. Heinzelmann, G. & Gilson, M. K. 自动化执行蛋白质-配体绝对结合自由能计算，用于对接精修和化合物评估。《科学报告》 **11**, 1116 (2021年). 46. Limongelli, V. 2020年的配体结合自由能与动力学计算。《*Wiley*计算与分子科学跨学科评论》 **10**, 1–32 (2020). 47. Limongelli, V., Bonomi, M. & Parrinello, M. 漏斗元动力学：一种精确的结合自由能计算方法。《美国国家科学院院刊》 **110**, 6358–6363 (2013年). 48. Souaille, M. & Roux, B. 加权直方图分析方法的扩展：结合伞形采样与自由能计算。《计算物理通讯》 **135**, 40–57 (2001)

.

49. Torrie, G. M. & Valleau, J. P. 蒙特卡罗自由能估计中的非物理采样分布：伞形采样。*J. Comput. Phys.* **23**, 187–199（1977年）。50. Izrailev, S., Stepaniants, S., Balsera, M., Oono, Y. & Schulten, K. 亲和素-生物素复合物解离的分子动力学研究。*Biophys. J.* **72**, 1568–1581（1997年）。51. Paszke, A. 等 PyTorch：一种命令式风格的高性能深度学习库。载于神经信息处理系统进展第33卷（2019年）。52. Mailoa, J. P. 等 一种用于预测复杂多元素扩展系统中直接协变力的快速神经网络方法。*Nat. Mach. Intell.* **1**, 471–479（2019年）。53. Batzner, S. 等用于数据高效、精准原子间势的E(3)等变图神经网络。《自然·通讯》 **13**, 2453（2022年）。54. Park, C. W. 等 采用直接力架构的精准且可扩展图神经网络力场与分子动力学。*npj Comput. Mater.* **7**, 73（2021年）。

55. Xie, T., France-Lanord, A., Wang, Y., Shao-Horn, Y. & Grossman, J. C. 用于材料原子尺度动力学无监督学习的图动力网络。《自然·通讯》 **10**, 2667（2019年）。56. Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. PIGNet：一种面向通用药物-靶标相互作用预测的物理信息深度学习模型。*Chem. Sci.* **13**, 3661–3673（2022年）。57. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. 使用GOLD改进蛋白质-配体对接。*Proteins* **52**, 609–623（2003年）。58. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. & Mason, J. S. 用于分析和比较蛋白质与配体的通用参考框架。配体和蛋白质指纹（FLAP）：理论与应用。《化学信息与建模杂志》 **47**, 279–294（2007年）。59. Cross, S., Baroni, M., Carosati, E., Benedetti, P. & Clementi, S. FLAP：虚拟筛选中的GRID分子相互作用场。使用DUD数据集进行验证。《化学信息与建模杂志》 **50**, 1442–1450（2010年）。60. Ruiz-Carmona, S. 等 rDock：一种用于将配体对接至蛋白质和核酸的快速、灵活开源程序。《公共科学图书馆·计算生物学》 **10**, e1003571（2014年）。61. Fernández-Recio, J., Totrov, M. & Abagyan, R. 内坐标中的软性蛋白质-蛋白质对接。《蛋白质科学》 **11**, 280–291（2002年）。62. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. 增强型实用诱饵目录（DUD-E）：更优配体和诱饵分子，助力更优基准测试。《药物化学杂志》第55卷，6582–6594（2012年）。63. Bauer, M. R., Ibrahim, T. M., Vogel, S. M. & Boeckler, F. M. 使用DEKOIS 2.0评估和优化虚拟筛选工作流程——一个包含高难度对接基准数据集的公共库。《化学信息与建模杂志》 **53**, 1447–1462（2013年）。64. Dane, E. L. 等肿瘤穿透型PEG-脂质纳米盘递送STING激动剂，可引发强效抗癌免疫。《自然·材料》 **21**, 710–720（2022年）。65. Thompson, A. P. 等 LAMMPS——一种灵活的模拟工具，用于原子、介观和连续尺度的基于粒子的材料建模。《计算物理通讯》 **271**, 108171（2022年）。66. Amaral, M. 等 蛋白质构象柔性调节动力学和热力学的

f

药物结合。《自然·通讯》 **8**, 2276（2017年）。67. David, C. C. & Jacobs, D. J. 主成分分析：一种确定蛋白质本征动力学的方法。第1084卷（Humana出版社，2014年）。68. Liu, S. 等人。用于前瞻性虚拟筛选的实用模型选择。《化学信息与建模杂志》 **59**, 282–293（2019年）。69. Dror, R. O. 等人。药物与G蛋白偶联受体结合的途径和机制。《美国国家科学院院刊》 **108**, 13118–13123（2011年）。

方法

配体结构收集

“COCRY”数据集的配体三维结构来自蛋白质数据库[1]。对于下载的这些*pdb*文件，我们移除了所有水和溶剂分子。随后，将所有这些共晶结构与参考蛋白质结构对齐（CDK2 中为 1fin，HSP90 中为 1uyg，FXA 中为 1ezq）。之后提取天然配体，构建“COCRY”数据集。除“COCRY”数据集外的小分子来自多个来源（DUD-E、DEKOIS 2.0 和 CSAR），具体如 Ricci-Lopez 等人[1–4] 的研究所述。对于 DUD-E 和 DEKOIS 2.0 数据集，三维结构已预先生成，可直接用于对接。对于 CSAR 数据集，则使用 OpenBabel 根据 SMILES 生成三维结构。[5]

输入结构准备

从蛋白质数据库中获取单个蛋白质（*pdb*）构象结构（CDK2 为 1fin，HSP90 为 1uyg，FXa 为 1ezq）。[1] 随后，使用 smina[6] ，并采用参数“—scoring=vinardo –factor=100 –num_modes=5 – exhaustiveness=16”，将配体分子对接到刚性蛋白质上。然后，将对接配体提取到独立的 *mol2* 文件中。使用 PyMol（v2.5）、OpenBabel（v3.1.1）和 RDKit（v2020.09.1.0），从 *pdb* 和 *mol2* 文件中提取二级结构、残基名称和可旋转轮信息。[5,7,8] 使用 GROMACS 软件[10] 为蛋白质 *pdb* 指定 CHARMM 力场[9]，并将其转换为蛋白质 *gro* 文件。同样，使用 SwissParam 网络服务器[11] 为配体 *mol2* 文件指定 CHARMM 风格的力场，随后利用 charmm2gromacs-pvm 功能将其转换为 GROMACS格式。[10] 然后，将它们合并为一个蛋白质-配体复合物 *gro* 文件（包含完整的 CHARMM 力场），再使用经过一些错误修复的 InterMol 软件定制版本[13]，将其进一步转换为 LAMMPS 输入数据格式[12] 。这种 LAMMPS 输入数据格式可直接载入我们的 3T PyTorch模型，以计算原子级力场能量（结构评估代价函数）。使用 PyMol 为 *pdb* 文件指定二级结构。这里，我们只使用三种

标记，即螺旋、折叠片和环。使用 OpenBabel 和 RDKit 从 *mol2* 文件中提取可旋转轮信息。r_{centre} 根据一批对齐的共晶配体的质心进行计算。

3T PyTorch 模型开发与结构生成

3T 结构生成算法通过 PyTorch 的自动微分功能实现，[14] 在计算上与标准 PyTorch 深度学习模型看起来完全相同，不同之处在于该过程不涉及机器学习或训练数据。3T 模型分为两部分：第一部分是进行结构变换的分层张量变换模块，第二部分是结构与力场能量计算模块。InterMol 生成的 LAMMPS 力场类型在 3T PyTorch模型中重新实现，以支持原生的、类似"训练"的 PyTorch 结构生成。对于给定 3T 模型步骤中的系统快照，3T 系统（无论是 $E_{complex,\ 3T}$， $E_{ligand,\ 3T}$，还是 $E_{protein,\ init}$）的能量代价函数只是系统快照中 Lennard-Jones、库仑、键、键角以及正常和非正常二面角能量的总和，其计算方式与 LAMMPS 中通常采用的方式相同。[12]Adam 优化器[15] 和多步学习率调度器用于优化。开始时，先执行 200 个优化步骤，使结合口袋中的配体结构松弛。随后，整个可移动的蛋白质-配体口袋（范围为 r_{cutoff}）受到 3T能量扰动，之后执行 2000 个 3T 优化步骤。微组平移扰动采用均匀随机分布 $[-1.5, +1.5]$ Å θ_i ；$[-0.15, +0.15]$ 弧度用于 θ_i 旋转扰动。苯丙氨酸、组氨酸和色氨酸等蛋白质微组的侧链可能非常刚性，从而形成很深的局部结构能量极小值。因此，当我们检测到这些结构有足够空间（1 Å 范围内没有原子冲突）时，会在 3T能量扰动步骤中以 50% 的概率对 $\theta_{A,\ i}$ 额外进行 180 度旋转，以生成更多样的蛋白质-配体复合物口袋构象。生成的构象将以 *xyz* 或 *cif* 文件的形式输出，并记录优化过程中的成本函数（能量景观）。该过程

在能量扰动过程中使用 10 个不同的随机数种子（但这些种子在不同配体间保持一致）重复 10 次，以生成十个 3T 构象。3T 的 PyToch 组件在腾讯云平台的单个 NVIDIA T4 GPU 上运行。

蛋白质结构的主成分分析

分析处理了三组结构：（1）共晶蛋白质结构；（2）从长时间分子动力学模拟中提取的蛋白质构象；（3）由 3T 生成的蛋白质构象。首先，将所有结构与参考结构对齐（例如，CDK2 的 1fin）。共晶蛋白质构象取自该蛋白质所有可用的 PDB；分子动力学构象是从蛋白质-配体结构的全原子复合物分子动力学模拟中每隔 1 ns 采样得到的 500 个结构；而 3T 构象则是由 1fin smina 重新对接的初始结构生成的 90 个结构。随后，提取结合口袋内蛋白质骨架 α碳原子的 (x,y,z) 坐标作为特征。接着，使用共晶数据拟合 scikit-learn 主成分分析模型（n_component=2），再用这些模型转换分子动力学和 3T 数据。最后，绘制主成分 PC1 和 PC2，以展示蛋白质构象的分布。

$\Delta RMSD$ 计算

使用 smina 评分函数对这十种 3T 构象进行评分。选取得分最低的三个结合口袋结构作为最佳候选结构，随后根据蛋白质的结合口袋原子，使用 PyMol 将它们与对应的实验共晶结构进行对齐，并使用 Spyrmsd 软件包计算配体均方根偏差。[16] 接着，将这三个对齐后的结合口袋结构中最佳的均方根偏差与初始 smina 姿势的均方根偏差进行比较（后者同样与实验共晶结构对齐）。我们还检验假设$\langle RMSD_{init} - RMSD_{3T} \rangle = \langle \Delta RMSD \rangle > 0$在统计上是否显著，并使用 SciPy 的 stats 软件包计算p值。

活性配体分类

记录的 3T 形成能（每种构象 2200 个步骤）按 100 倍进行下采样，并缩小 1000 倍（使其单位更接近对接评分），从而为每种构象生成 22 个能量特征。初始结构、配体弛豫结构和 10 种构象对应的对接评分也作为特征纳入（共 12 个特征），因此每个配体总计有 232 个特征。这些特征直接用作基于梯度提升树的 30×4cv 活性配体分类器的输入；AUC-ROC 和 NEF_{R_a} 计算则使用先前工作中提供的同一 Jupyter 笔记本。[1] 为确保公平比较构象特征，而非比较分类算法，分类算法的设置保持不变。

分子动力学设置与半柔性对接计算资源估算

本比较使用两种类型的机器。GPU 机器使用一张 NVIDIA T4 显卡。CPU 机器使用一台 48 核 AMD EPYC 7K62 处理器中的 16 个核心。半柔性 Smina 对接（刚性蛋白质骨架、可旋转侧链）使用参数“--scoring=vinardo -- factor=100 --num_modes=3 -exhaustiveness=16 --flexdist=4 --flexdist_ligand=ref_ligand.sdf”完成。对于 rDock，在 PRMFILE 中使用参数“RECEPTOR_FLEX=4”，对接时使用参数“-n=64”。进行分子动力学计算时，将上述 GROMACS 蛋白质-配体结构置于水中溶剂化，并对体系进行电荷中和和能量最小化，随后进行 NVT 和 NPT 平衡。之后测量 GROMACS 生产阶段分子动力学的速度（NPT，时间步长为 2.0 fs，温度为 300 K）。据估计，要获得足够的蛋白质结构多样性，需要 1 μs 至 1 ms 的分子动力学模拟时间；我们采用了较为激进的假设，认为 1 μs 的分子动力学模拟已足够。据此计算所需的 GPU 计算成本。[17]

方法参考文献

1. Ricci-Lopez, J., Aguila, S. A., Gilson, M. K. & Brizuela, C. A. 通过集成对接和机器学习改进基于结构的虚拟筛选。《化学信息与建模杂志》 **61**, 5362–5376 (2021)。2. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. 增强型实用诱饵目录（DUD-E）：更优的配体和诱饵分子，实现更好的基准测试。《药物化学杂志》第 55 卷，6582–6594 (2012)。3. Bauer, M. R., Ibrahim, T. M., Vogel, S. M. & Boeckler, F. M. 使用 DEKOIS 2.0 评估和优化虚拟筛选工作流程——一个具有挑战性的对接基准数据集公共库。《化学信息与建模杂志》 **53**, 1447–1462 (2013)。4. Dunbar, J. B. 等 CSAR 数据集 2012 年发布：配体、亲和力、复合物和对接诱饵分子。《化学信息与建模杂志》 **53**, 1842–1852 (2013)。5. O'Boyle, N. M. 等 Open Babel：开放式化学工具箱——1758-2946-3-33.pdf。《化学信息学杂志》 **3**, 33 (2011)。6. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. 从 CSAR 2011 基准测试中汲取的 smina 经验评分教训。《化学信息与建模杂志》 **53**, 1893–1904 (2013)。7. PyMOL分子图形系统，Schrödinger, LLC。https://pymol.org/2/ (2021)。8. RDKit：开源化学信息学工具。http://www.rdkit.org/ (2020)。9. Vanommeslaeghe, K. 等 CHARMM通用力场（CGenFF）：一种适用于类药分子、与 CHARMM全原子加和生物力场兼容的力场。《计算化学杂志》 **31**, 671–690 (2010)。10. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS：一种消息传递并行分子动力学实现。《计算物理通讯》 **91**, 43–56 (1995)。11. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam：一种用于小分子有机化合物的快速力场生成工具。《计算化学杂志》 **32**, 2359–2368 (2012)。12. Thompson, A. P. 等 LAMMPS——一种灵活的模拟工具，可用于原子、介观和连续尺度上的基于粒子的材料建模。《计算物理通讯》 **271**, 108171 (2022)。13. Shirts, M. R. 等 从 SAMPL5 数据集分子动力学引擎比较中汲取的经验。《计算机辅助分子设计杂志》 **31**, 147–161 (2017)。14. Paszke, A. 等 PyTorch：一种命令式风格的高性能深度学习库。载于神经信息处理系统进展第 33 卷 (2019)。15. Kingma, D. P. & Ba, J. L. Adam：一种随机优化方法。载于第三届国际学习表征会议论文集 1–15 (2015)。16. Meli, R. & Biggin, P. C. Spyrmsd：Python 中的对称性校正 RMSD 计算。《化学信息学杂志》 **12**, 49 (2020)。17. Evangelista Falcon, W., Ellingson, S. R., Smith, J. C. & Baudry, J. 药物发现中的集成对接：需要多少个来自分子动力学模拟的蛋白质构象，才能重现已知的配体结合？《物理化学杂志 *B*》 **123**, 5189–5195 (2019)。

致谢

作者感谢腾讯量子实验室的 M. Shao 提供技术支持，并感谢其在腾讯云平台上的协助。本研究完全在腾讯量子实验室内，使用腾讯云平台开展。

作者贡献

J.P.M. 负责 3T算法开发、结构生成和特征提取。Z.Y. 和 J.P.M. 负责蛋白质-配体复合物结构的初步制备及输入数据预处理。Z.Y. 负责刚性蛋白质对接、均方根偏差计算和梯度提升树分类。J.P.M. 和 Z.Y. 负责分子动力学结构生成和微观状态分析。Z.Y. 和 J.Q. 负责基于无配体分子动力学结构的大规模刚性蛋白质对接。J.P.M. 和 Z.Y. 负责数据和计算成本分析。C.-Y.H. 和 S.Z. 提供反馈并指导研究。所有作者均参与论文撰写。

竞争性利益

作者声明不存在任何竞争性利益。

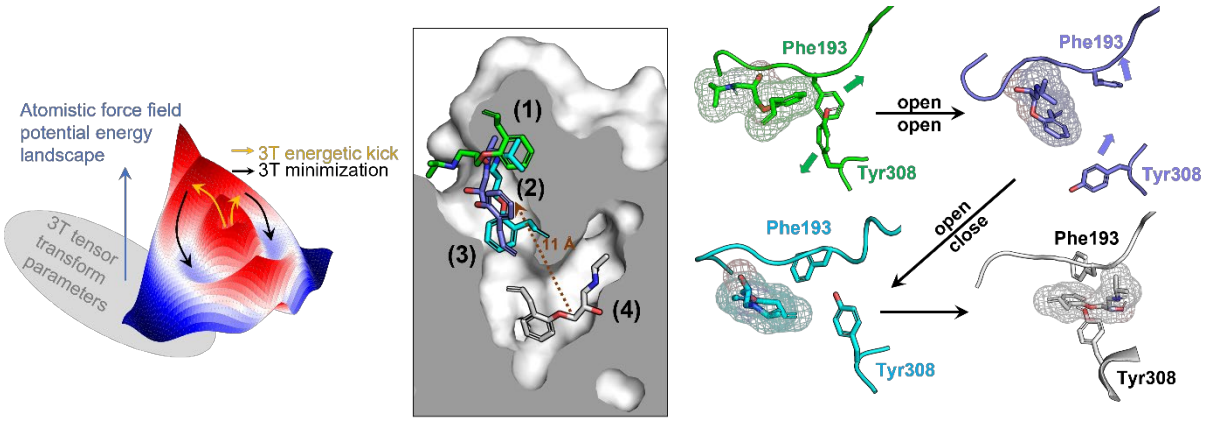
补充信息

在线版本包含补充材料，可通过……获取。

材料与通信

有关本手稿的通信及材料索取事宜，请联系 Jonathan Mailoa：jpmailoa@alum.mit.edu；或联系 Shengyu Zhang：shengyzhang@tencent.com。

仅供目录使用



概要

我们采用多尺度结构能量最小化生成多个完全柔性的蛋白质-配体口袋结构，从而实现准确的活性配体分类，并探索远距离配体结合构象。

补充信息 –

基于分层张量变换的蛋白质-配体复合物生成器与药物筛选

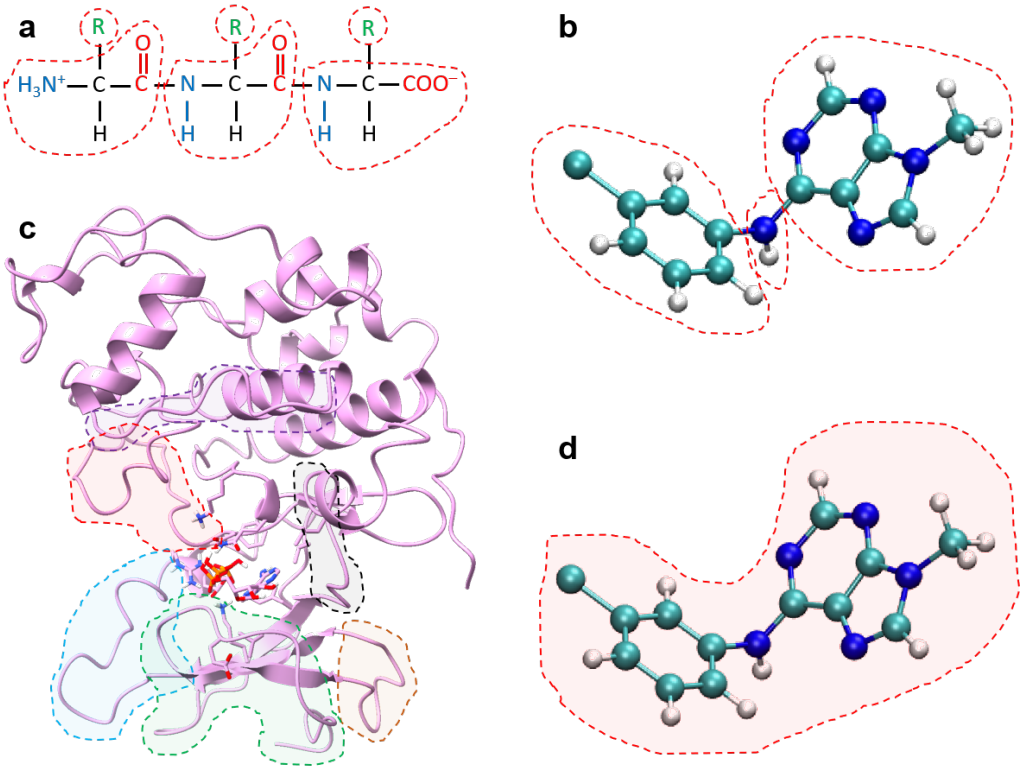
分层张量变换

Jonathan P. Mailoa，[1†*] Zhaofeng Ye，[1†] Jiezhong Qiu，[1] Chang-Yu Hsieh，[1] 和 Shengyu Zhang[2*]

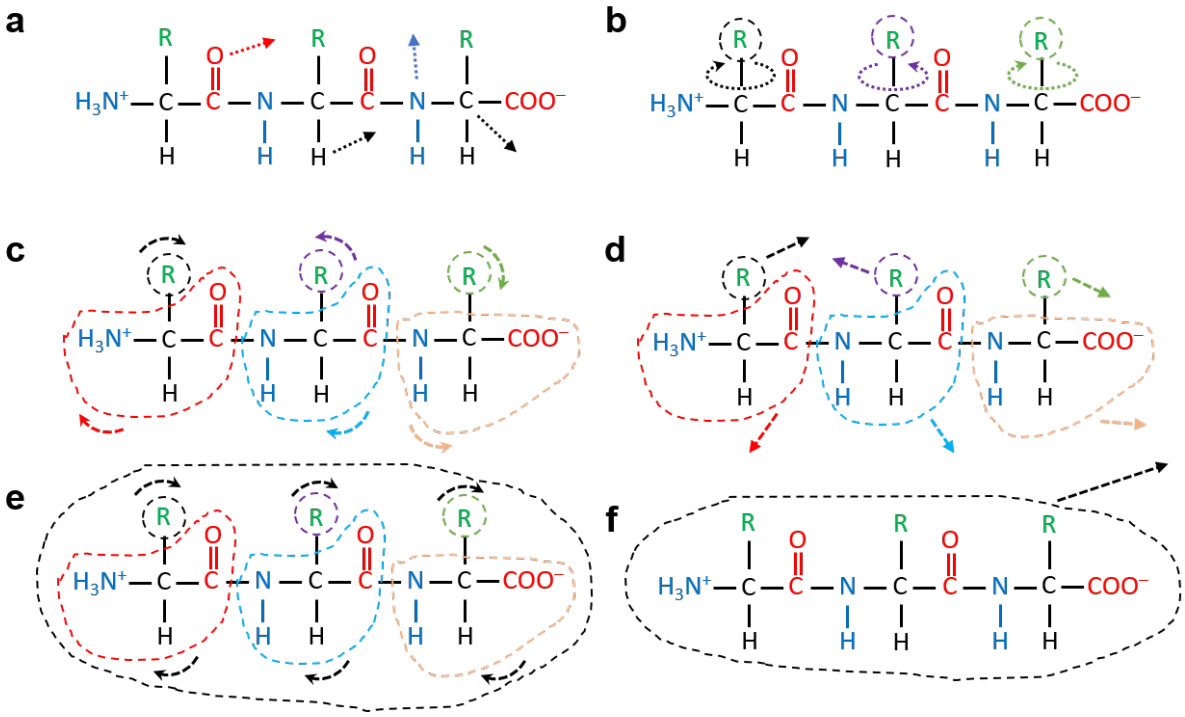
1) 腾讯量子实验室，腾讯，中国广东深圳 2) 腾讯量子实验室，腾讯，中国香港特别行政区

† 这些作者对本研究贡献相同

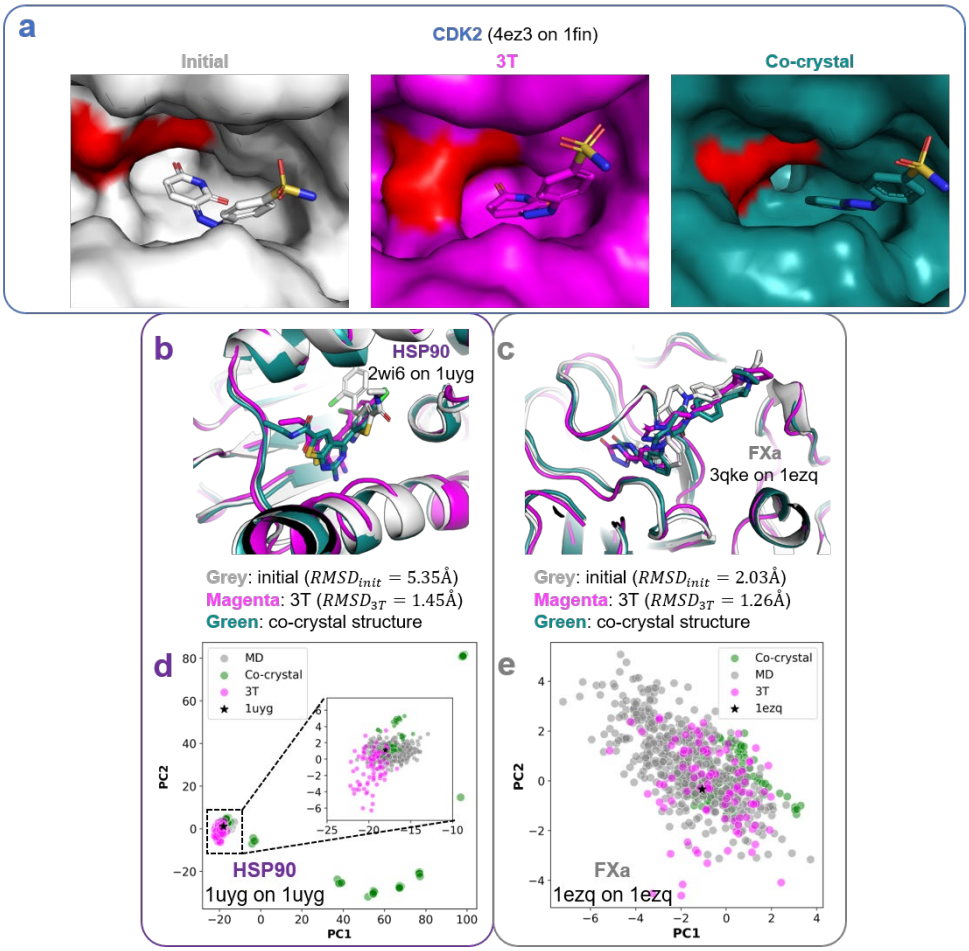
* 通讯作者：jpmailoa@alum.mit.edu，shengyzhang@tencent.com



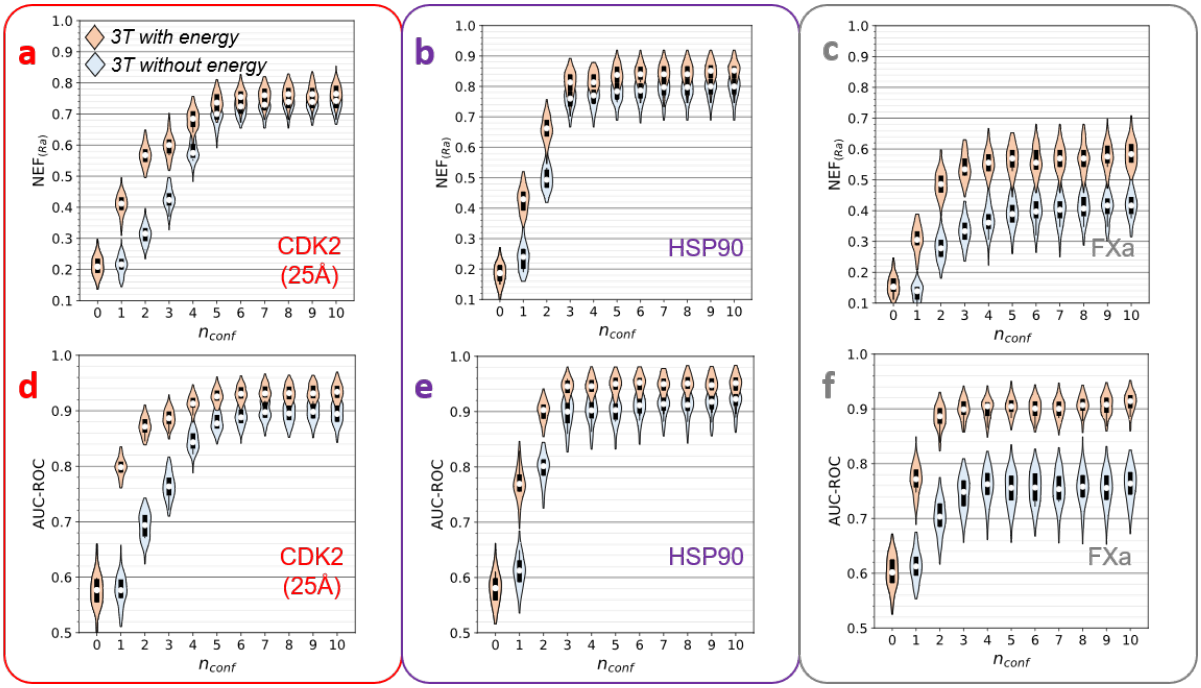
补充图 1 | 3T 蛋白质-配体口袋分层结构分割。a) 基于氨基酸主链和侧链划分的蛋白质原子微组，**b)** 基于配体可旋转键划分的配体原子微组，**c)** 基于残基二级结构划分的蛋白质原子宏组，以及**d)** 包含所有配体原子的配体原子宏组。



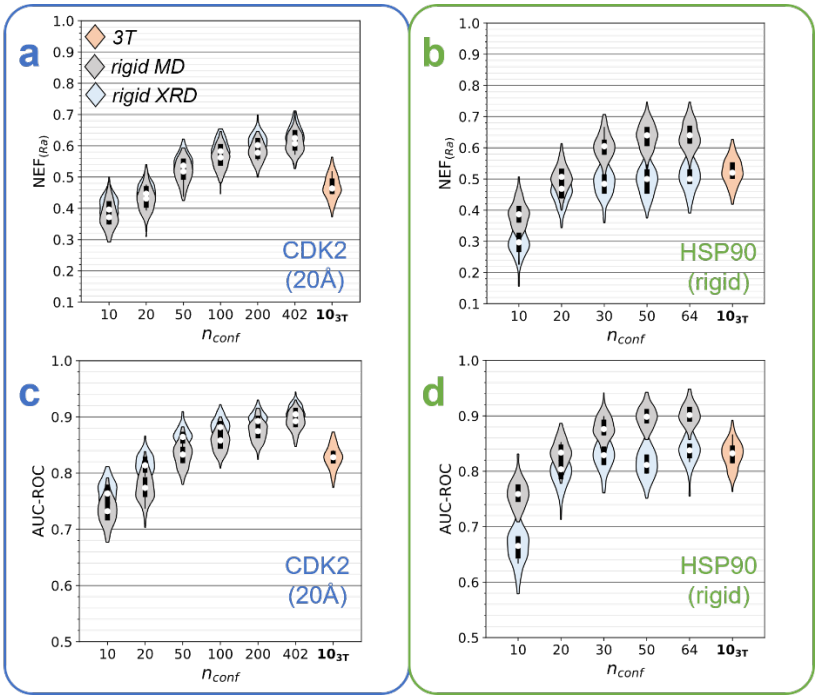
补充图 2 | 3T层级结构变换的可视化说明。 **a)**单个原子平移，**b)**单个微基团侧链旋转，**c)**单个微组中心旋转，**d)**单个微基团平移，**e)**单个宏组中心旋转，以及**f)**单个宏基团平移。这些变换的参数通过 PyTorch 基于力场代价函数的优化器进行优化。对于宏组中心旋转，通过协调旋转宏组内的各个微基团，使其绕各自的微组中心旋转来实现。



补充资料　ary Figure 3 | 3T结构的物理验证。 a) CDK2配体构象与蛋白质表面网格的比较4ez3 PDB中的配体与来自1fin的蛋白质进行交叉对接，结果表明，3T变换能够使蛋白质口袋结构收缩，并将配体推向右侧，这与实际的4ez3共晶结构相似。原始的4ez3-1fin交叉对接结构具有较为开放的刚性蛋白质口袋，因此产生了错误的配体构象。**b-c)** HSP90（2wi6 PDB中的配体与1uyg PDB中的蛋白质交叉对接）和FXa蛋白（3qke PDB中的配体与1ezq PDB中的蛋白质交叉对接）的单个示例：对比由3T生成的交叉对接蛋白质-配体结构（洋红色）、由smina生成的初始交叉对接结构（灰色）以及实验共晶结构参考（绿色）的叠加结构。**d-e)** 对HSP90和FXa蛋白所有可用的实验共晶PDB结构（绿色）、使用单次全原子复合物分子动力学模拟生成的结构（灰色）以及使用3T生成的重新对接结构（洋红色）进行蛋白质主链结构主成分分析比较。全原子复合物分子动力学模拟和3T重新对接分别仅使用HSP90的1uyg PDB和FXa的1ezq PDB。对应实验共晶结构的主成分分析结果以深灰色突出显示。对于HSP90，自然界中存在数种不同的蛋白质-配体口袋构象（绿色）。全原子复合物分子动力学模拟和3T仅占据属于实际1uyg PDB实验共晶子空间的主成分子空间，不涉及其他区域。对于FXa，共晶结构、全原子复合物分子动力学模拟和3T似乎共享同一个构象空间。需要指出的是，与CDK2和HSP90蛋白相比，可用的实验FXa蛋白质数据库共晶结构要少得多。



补充图 4 | 在不同数量的3T生成构象下，分类时纳入或排除3T能量景观特征的影响。 a–c) CDK2结合口袋（$r_{cutoff} = 25$Å）、HSP90结合口袋（$r_{cutoff} = 20$Å，非刚性）和FXa结合口袋（$r_{cutoff} = 20$Å）的NEF_{R_a}活性配体分类指标。**d–f)** 与**a–c)**中相同蛋白质的ROC曲线下面积活性配体分类指标。分类统计采用30× 4折交叉验证（使用梯度提升树分类器）。显而易见，纳入3T能量景观特征对于从诱饵分子中识别活性配体非常有帮助。$n_{conf} = 0$ 表示我们不通过3T能量扰动生成任何完全柔性的蛋白质-配体口袋结构，而只使用初始smina交叉对接结构，以及在保持刚性蛋白质口袋的同时弛豫配体后得到的结构（见**图4**）。



补充图 5 | 减少或消除 3T蛋白质柔性对活性配体分类的影响

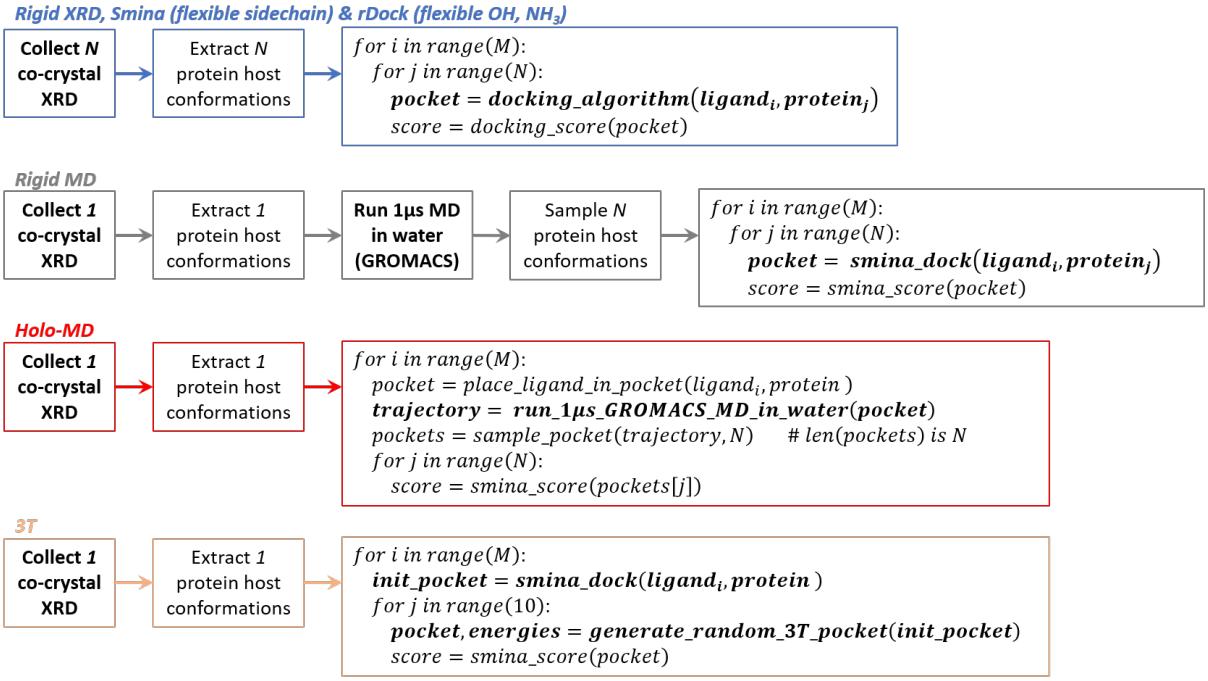
性能。 a–b) CDK2结合口袋（NEF_{R_a} 的活性配体分类指标$r_{cutoff} = 20$Å）和 HSP90结合口袋 通过 3T 特意设为完全刚性且不可变换的结合口袋），以及 **c-d)**相同蛋白质的 ROC曲线下面积活性配体分类指标。 30× 4折交叉验证（采用梯度提升树分类器）用于计算分类统计数据。可以看出，在较小的 CDK2结合口袋上进行 3T 不足以构建良好的分类器，因此需要允许结合口袋柔性具有更大的半径。同样可以看出，仅对配体进行 3T（HSP90蛋白质口袋结构完全刚性）会显著降低活性配体分类器的性能。

蛋白质	活性 配体		3T 分类器	刚性分子动力学 分类器	刚性X射线衍射 分类器	$n_{conf,3T}$ *vs* $n_{conf,MD-XRD}$
CDK2 (25	442/3764 ($R_a = 0.117$)	NEF_{R_a}	**0.771** **± 0.030**	0.608 ± 0.033	0.624 ± 0.039	**(10 + 2)** *vs* 402
		ROC曲线下面积	**0.935** **± 0.014**	0.892 ± 0.017	0.904 ± 0.015	
CDK2 (20	442/3764 ($R_a = 0.117$)	NEF_{R_a}	0.469 ± 0.037	0.608 ± 0.033	**0.624** **± 0.039**	(10 + 2) *vs* **402**
		ROC曲线下面积	0.828 ± 0.020	0.892 ± 0.017	**0.904** **± 0.015**	
HSP90	298/2452 ($R_a = 0.122$)	NEF_{R_a}	**0.851** **± 0.035**	0.640 ± 0.042	0.505 ± 0.046	**(10 + 2)** *vs* 64
		ROC曲线下面积	**0.949** **± 0.018**	0.903 ± 0.019	0.836 ± 0.024	
HSP90 (刚性)	298/2452 ($R_a = 0.122$)	NEF_{R_a}	0.524 ± 0.042	**0.640** **± 0.042**	0.505 ± 0.046	(10 + 2) *vs* **64**
		ROC曲线下面积	0.830 ± 0.024	**0.903** **± 0.019**	0.836 ± 0.024	
FXa	298/7191 ($R_a = 0.040$)	NEF_{R_a}	**0.584** **± 0.043**	0.554 ± 0.046	0.452 ± 0.044	**(10 + 2)** *vs* 136
		ROC曲线下面积	**0.913** **± 0.018**	0.902 ± 0.021	0.855 ± 0.021	

补充表 1 | 3T：采用梯度提升树分类器，在 3 种不同蛋白质宿主中进行活性配体分类的指标。 在除正文**表1**中的信息外，我们还纳入了以下情况下的分类统计：$r_{cutoff} = 20$Å的结合口袋用于CDK2，以及使用完全刚性的HSP90蛋白结合口袋进行3T蛋白质-配体口袋构象生成。可以看出，当3T的柔性蛋白质口袋生成能力降低或被移除时，其活性配体分类效用会降低。

蛋白质	蛋白质刚性	方法	XRD 实验	每个配体的分子动力学计算成本	每个配体的对接成本
CDK2	刚性	刚性X射线衍射	406×	0	12.9 CPU 小时
		刚性分子动力学*	1×	0.058 GPU 小时	12.9 CPU 小时
	半-柔性的	smina（柔性侧链）	406×	0	710.1 CPU 小时
		rDock（柔性羟基、NH_3）	406×	0	6.3 CPU 小时
	完全灵活	Holo-MD*	1×	217.2 GPU 小时	0
		3T	1×	0	2.5 GPU 小时 / 22.7 CPU 小时
	* CDK2蛋白含 69k 个原子 --> GROMACS 图形处理器速度 = 110.5 纳秒/天				
HSP90	刚性	刚性X射线衍射	64×	0	4.1 CPU 小时
		刚性分子动力学*	1×	0.051 GPU 小时	4.1 CPU 小时
	半柔性柔性	Smina（柔性侧链）	64×	0	178.3 CPU 小时
		rDock（羟基、NH_3 均可灵活旋转）	64×	0	1.7 CPU 小时
	完全灵活	Holo-MD*	1×	124.3 GPU 小时	0
		3T	1×	0	2.1 图形处理器 小时 / 10.1 中央处理器 小时
	* HSP90蛋白含 37k 个原子 --> GROMACS 图形处理器速度 = 193.1 纳秒/天				
FXa	刚性	刚性X射线衍射	136×	0	6.5 中央处理器小时
		刚性分子动力学*	1×	0.028 GPU 小时	6.5 CPU 小时
	半柔性	Smina（柔性侧链）	136×	0	170.5 CPU 小时
		rDock（柔性羟基、NH_3）	136×	0	2.5 CPU 小时
	完全柔性	Holo-MD*	1×	201.7 GPU 小时	0
		3T	1×	0	1.8 图形处理器 小时 / 10.7 CPU 小时
	* FXa 蛋白质含 60k 个原子 --> GROMACS 图形处理器速度 = 119.0 纳秒/天				

补充表 2 | 基于所用蛋白质-配体复合物结构生成方法，对 CDK2、HSP90 和 FXa 数据集上的各类基于对接的活性配体分类任务进行实验与计算资源估算。 活性配体 基于所用蛋白质-配体复合物结构生成方法，对 CDK2、HSP90 和 FXa 数据集上的各类基于对接的活性配体分类任务进行实验与计算资源估算。这六种方法根据生成的蛋白质结构的刚性分类（刚性：配体交叉对接期间蛋白质构象不发生变化；半柔性：配体交叉对接期间允许部分蛋白质侧链旋转；全柔性：配体交叉对接期间所有蛋白质骨架和侧链原子均可自由移动）。粗体标出的三种方法是正文图4中用于比较分类性能的方法**图4**。XRD 实验列指的是启用基于这些方法的活性配体分类任务所需的 X 射线衍射共晶结构数量（与正文**图4**中使用的实验结构数量相同）。对于刚性分子动力学、Holo-MD 和 3T，仅需一个此类实验共晶蛋白质结构作为初始结构。不过，Holo-MD 和刚性分子动力学方法还需要额外的分子动力学模拟成本，用于后续蛋白质-配体复合物结构生成。对于 3T，所有结构生成成本（单次刚性蛋白质对接加 10 构象生成）均归为"对接成本"。对接计算成本估算值取随机选取的三个配体的平均值，但 3T 的 CPU 小时估算值取随机选取的 16 个配体的平均值。



补充流程图 1 | 补充表2中所述各类基于对接的活性配体分类任务的伪代码和流程图。对于 CDK2、HSP90 和 FXa 蛋白质，*N* 的值分别为 406、64 和 136。刚性 XRD 所用的对接算法是标准的 smina 对接算法（蛋白质宿主结构固定）。实验或计算成本较高的子任务以粗体标出。**补充表2**中的"时间/配体"指标表示每个 $ligand_i$ 的平均资源需求。与这些特征收集任务相比，对接评分计算（不含构象搜索）以及分类模型训练本身所需时间可以忽略不计。

更为激进的 3T 能量扰动

在正文中，我们采用相对温和的能量扰动设置来筛选 10–20k 个蛋白质-配体复合物，同时确保 3T 不会生成大量不具物理意义的蛋白质-配体复合物构象。本节介绍另一种方案：处理数量少得多的目标蛋白质-配体样本时，我们可以承担更大的风险，采用激进得多的能量扰动设置。我们将这一新设置用于一个棘手的问题：探索蛋白质口袋中的远距离亚稳态结合构象。例如，阿普洛尔配体对接至 β2 G蛋白偶联受体后，其前庭结合位点与受体口袋相距较远（约 10–15 Å）。我们选择三个初始结构，即排名第 1st 至 3rd 的 smina 对接姿势。我们希望观察 3T 能否从初始正构结合位点出发，在前庭结合位点内探索不同构象（目前没有可用于实验比较的实验共晶结构，因为该位点只是一个暂时的过渡位点；Dror 博士通过长达 5 μs 的分子动力学研究发现了这一点 等人，doi:10.1073/pnas.1104614108）。

为实现这些任务，我们进行了以下大幅度修改：

a. 在第一个配体弛豫步骤之前（正文图 4a），我们对配体的宏组变换参数施加 3T能量扰动。我们将 $\theta_{R,j}$ 均匀设定在 $[-\pi, \pi]$ 之间，用于绕 x、 y和 z 轴旋转；而初始 $\theta_{T,j}$ 值设为 $[-3, -9, -9]$ ，大致指向前庭结合位点的通常方向。必须这样做，因为我们目前尚无自动口袋识别功能，并且希望避免 3T 将配体扰动到随机的致密蛋白质区域中（如果初始结构完全不合理，例如配体与多个蛋白质组分相互缠绕，那么让 3T 恢复结构的效率会很低）。在这种激进情况下，让 3T 进行随机且大幅度的宏组平移能量扰动，在计算上并不明智。

b. 配体宏组不再遵循蛋白质宏组的旋转模式（**补充图 2e**），即微组围绕各自微组中心协调旋转。配体宏组将改为围绕整个配体宏组的中心旋转。这样既能提高配体的活动性，又能使蛋白质结构变换更加保守。c. 配体弛豫步骤的 PyTorch 优化器设置提高了 100×。d. 配体的 3T 参数（$\vec{r}_{m,init}$ 和 θ's）的梯度乘以 3×，而蛋白质 3T 参数保持不变。这能确保配体得到更快的调整，同时使蛋白质结构变换更加保守。e. 初始配体弛豫步骤增加到 10,000 步（原为 200 步）。第二次能量扰动后的完整口袋弛豫步骤仍为 2,000 步。f. 每个蛋白质-配体口袋复合物对生成 30 个构象，而不是正文中设定的每对 10 个构象。这意味着我们最初会生成 90 个构象。g. 进行 2,000 步口袋弛豫后，我们从 90 个构象中选出 10 个结构，并将完整口袋弛豫步骤从 2,000 步延长至 5,000 步。h. 10 个结构中有 1 个在经过 5,000 步后仍未完全弛豫，因此需要额外进行 5,000 步完整口袋弛豫，使蛋白质结构进一步弛豫至最终构象。

总的来说，我们探索的所有构象在一台配备 Nvidia T4 GPU 的机器上耗用了约 55 个 GPU 小时的计算时间。