

## Abstract

深度生成模型在从头药物设计领域越来越受到关注。然而，针对新靶标的配体分子的合理设计仍然具有挑战性，特别是在控制生成分子的性质方面。在这里，受 DNA 编码化合物库技术的启发，我们介绍了 DeepBlock，这是一种基于块的配体生成深度学习方法，可针对目标蛋白质序列进行定制，同时实现精确的属性控制。DeepBlock 将生成过程巧妙地分为两个步骤：构建块生成和分子重建，分别由我们提出的神经网络和基于规则的重建算法完成。此外，DeepBlock 协同优化算法和深度学习来调节生成分子的属性。实验表明，DeepBlock 在生成具有亲和力、合成可及性和药物相似性的配体方面优于现有方法。此外，当与以毒性为优化目标的模拟退火或贝叶斯优化相结合时，DeepBlock 成功生成了低毒性配体，同时保留了与靶标的亲和力。

## 1 Introduction

寻找能够与特定蛋白质结合的配体（ligands）是药物发现中的一个重要目标。虚拟筛选作为一种重要的方法学已经得到了广泛应用，它利用计算机程序从小分子库中识别生物活性化合物??。然而，虚拟筛选的有效性受限于庞大的化学空间和所使用的化合物库。相比之下，de novo 药物设计策略则通过从零生成分子结构，提供了探索现有库之外更广泛化学空间的有前景途径??。

近年来，深度生成模型在 de novo 分子生成方面取得了重大进展。各种形式的自回归模型??、变分自编码器（VAE）??、生成对抗网络（GAN）??、归一化流（normalizing flow）??和扩散模型（diffusion model）??被提出用于生成简化分子输入线条系统（SMILES）字符串、分子图或三维（3D）构象。这些基于配体的药物发现模型旨在学习化学空间的底层分布，并随后从该分布中采样新的、有效的分子。然而，这些模型并不能直接生成针对特定蛋白靶点的分子，因此需要额外的基于对接筛选或与强化学习技术相结合??。

最近，配体生成的一个趋势是考虑分子如何与蛋白质靶点相互作用。诸如 LiGAN??和 3D-SBDD??等方法将来自蛋白质口袋的结构信息整合到分子生成过程中，从而为特定结合靶点定制分子。然而，值得注意的是，这些方法依赖于结构信息，这在针对具有未知三维结构的新靶蛋白时可能并不总是可用的。

此外，生成分子的可合成性??在实际药物发现项目中是一个至关重要的问题，但在许多基于深度学习的配体生成模型中，这一方面往往被忽视。尽管生成的分子在理论上可能显示出较强的亲和力，但其实际合成可能面临重大挑战，从而限制了生成模型的应用。此外，大多数现有方法仅关注生成分子与靶标之间的亲和力，而忽略了毒性、代谢以及其他对药物成功上市批准至关重要的属性。

DNA 编码化合物库??技术已成为一种被广泛接受的湿实验室方法，用于初步筛选。该方法利用组合化学??通过分子构建块的反应来快速生成大量候选化合物。受 DNA 编码化合物库技术的启发，我们引入了一种基于深度学习的名为 DeepBlock 的 de novo 药物设计框架，利用分子构建块。这里的构建块是指能够相互化学反应的分子片段。DeepBlock 的核心概念是将分子生成过程分解为两个顺序步骤：首先，根据作为输入的蛋白质序列生成构建块，其次将其组装成完整的分子。通过利用这些块的内在性质及其化学相互作用，DeepBlock 能够设计出高质量的理性分子。在此基础上，我们在 DeepBlock 中设计了有效机制，以解决两个关键任务：基于蛋白质序列的分子生成以及在生成过程中的属性控制。

在 DeepBlock 中，我们引入了块生成网络（Block Generative Network，BGNet），这是一种条件深度生成模型，旨在基于给定的蛋白质序列生成块序列。BGNet 包含两个显著增强其性能的关键特性。首先，它通过在大规模分子数据集上预训练的分子块自动编码器构建。这种预训练扩展了化学空间，并缓解了蛋白-配体对数据集规模有限可能导致的过拟合问题。其次，我们在 DeepBlock 中引入了一个关键组件，即靶标贡献感知模块。该模块增强了模型自主识别配体与残基相互作用的能力，从而弥补了蛋白质序列中缺乏三维结构信息的不足。BGNet 将这两个特性结合在一起，展示了其生成多样化且具有生物活性分子片段的能力，有效解决了蛋白质序列数据所带来的挑战。此外，我们将 BGNet 与模拟退火（simulated annealing，SA）算法??或贝叶斯优化（Bayesian optimization，BO）结合使用，

以控制生成过程，旨在在保持与靶蛋白的强结合亲和力的同时优化其他属性。我们进行了以药物毒性为优化目标的实验，结果表明我们的框架能够生成毒性水平较低的分子。

## 2 Results

### 2.1 DeepBlock 框架

在本研究中，我们提出了一种基于深度学习的框架，命名为 DeepBlock，用于生成和优化配体分子（图 1a），其通过将生成过程分解为两个步骤：构建块的生成和分子的重构。我们首先设计了一种分子碎片化和重构算法（图 1e），旨在将数据集中的配体分子转换为块序列，以训练块生成器 BGNet（图 1b），并将生成的块重构为有效的分子。为了丰富模型对分子的理解，BGNet 采用了一种配体和蛋白质的双重编码方案，并结合了一个结合贡献感知网络，以预测每个蛋白质残基的重要性在配体结合的预测。该方法确保结合位点的残基在相互作用中被视为更重要，从而直接影响靶标的表示。通过在大规模分子数据集上进行自监督预训练，并结合进化尺度建模 (ESM-2) ?? (一种蛋白质语言模型) 对蛋白质序列进行深入特征提取，DeepBlock 的预测能力得到了增强（图 1c）。在生成 BGNet 时，蛋白质序列作为输入条件，以靶标配体的块序列变量为目标（图 1d）。因此，我们将 BGNet 与优化算法相结合，利用 BGNet 生成相邻候选分子，并使用 SA 或 BO 算法进行探索，从而在生成过程中实现对分子属性的受控调节（图 1d）。

### 2.2 分子碎片化与重构

回溯合成有趣化学子结构 (BRICS) 算法?? 基于回溯合成化学定义了一系列可切割的化学键，并广泛用作分子碎片化的指导规则。然而，结构内在的组合爆炸使得仅通过 BRICS 切割的块来重构初始分子几乎无法实现。在此，我们提出了一种基于图的碎片化与重构算法，以解决这一挑战。整个转换过程遵循“分子到块序列再到分子”的路径，同时严格保证生成的分子在起点和终点保持相同。更多细节请参见方法部分中的“分子碎片化与重构算法”。为了验证其可靠性，我们将该转换应用于 ChEMBL 数据集中的每个分子??。在 2,205,345 个分子中，仅有 70 个分子未能成功转换（更详细的分析见补充图 1–5），成功率达到约 99.9963%。这一高成功率表明了我们方法的可靠性和实用性，有效支持下游模型学习任务。

**评估与对比。**如图 2a 所示，由 DeepBlock 生成的分子在对接亲和力上与 TargetDiff 和 Pocket2Mol 相当，且优于其他基线模型。此外，Vina 评分的分布（图 2d）显示，Pocket2Mol 和 TargetDiff 具有更大的方差，且存在更多的离群值，而 DeepBlock 的分布更为集中，离群值更少。这表明 DeepBlock 生成的候选分子更为一致且可靠。值得注意的是，高对接评分的分子可能包含罕见或不可行的子结构，这在药物开发中是不可接受的。因此，我们进一步分析了具有高对接亲和力的分子的定量药物相似性 (QED) 和回溯评分??，筛选了 Vina 评分超过 7.445 的分子，并在图 2c 中以散点图的形式展示结果。Pocket2Mol 倾向于生成具有较差 QED 或可合成性的高亲和力分子，而 TargetDiff 的高亲和力分子在合成可行性方面也存在问题，这表明这些方法生成的分子中可能存在不现实的结构。而另一方面，DeepBlock 生成的分子不仅实现了高结合亲和力，还表现出优越的药物相似性和易于合成的特性。此外，如图 2a 所示，DeepBlock 在 Retro?? 预测的成功率方面相较基线模型取得了令人满意的结果，这确保了后续药物合成的实用性。此外，DeepBlock 生成的分子与已知参考配体具有更高的结构相似性，但仍保持较大的多样性。此外，如图 2b 所示，DeepBlock 生成的分子与其他管道相比，更好地与已知生物靶点对齐。DeepBlock 生成的分子的更高结合评分表明其比其他方法更具可行性。

生成的分子在效率上显著快于基线模型，确保了质量和效率之间的平衡。如图 2e 所示，DeepBlock 中使用的预训练方案提高了生成分子的有效性、新颖性和唯一性。此外，我们还在替代数据集 PDBbind ?? 上对 DeepBlock 进行了评估。结果（补充表 1）显示，DeepBlock 继续表现出色，表明其在不同数据集上的稳健性和泛化能力。

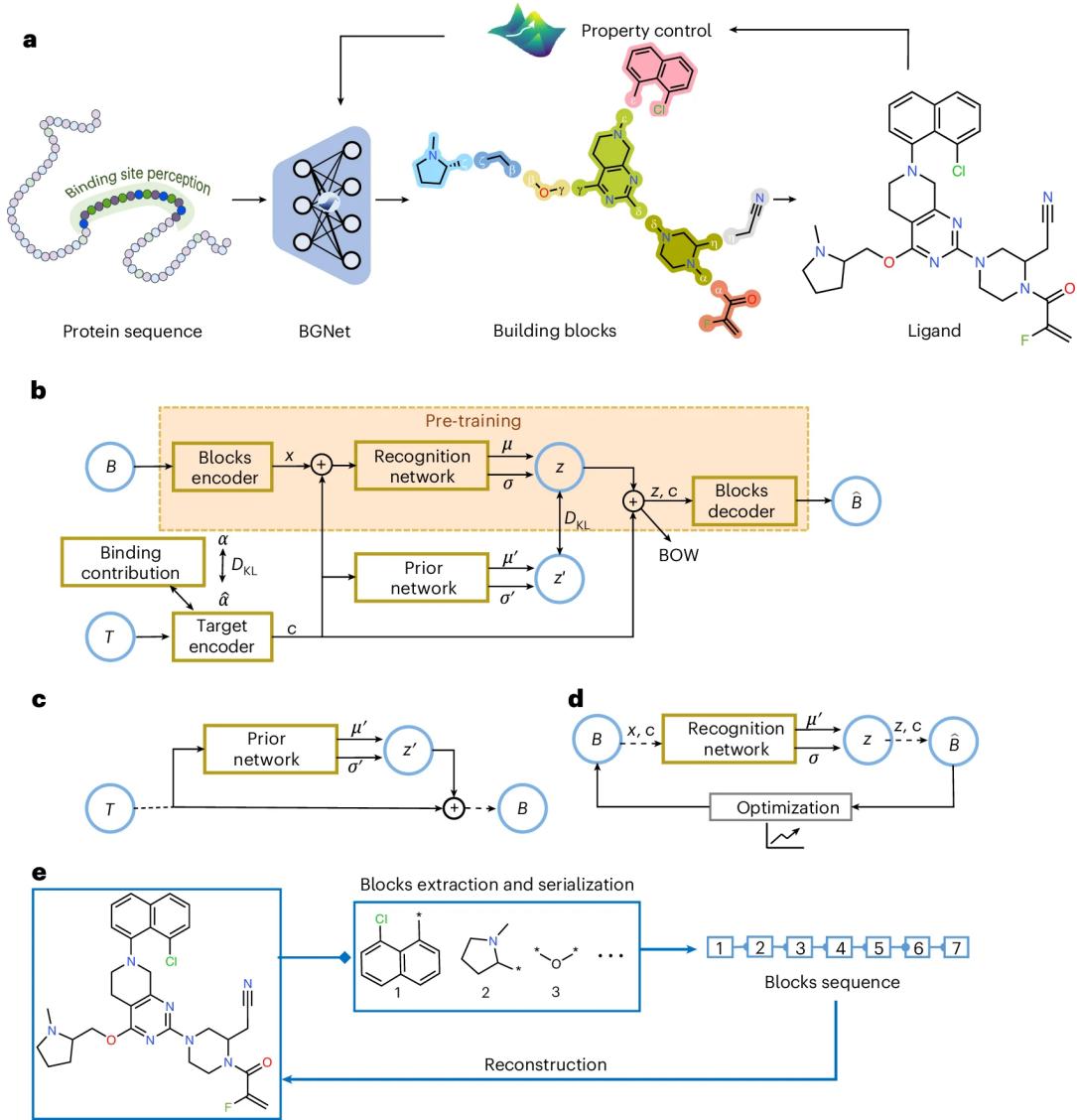


图 1: DeepBlock 框架概述。a, DeepBlock 的工作流程。使用不同颜色表示构建块以区分各个块。希腊字母用于表示虚拟原子，也称为断点，因为它们表示化学键被断开的地方。b, DeepBlock 的整体架构。 $x$  和  $c$  分别表示学习到的块序列变量  $B$  和靶标蛋白变量  $T$  的表示。 $z$  和  $z'$  是从均值  $\mu$  和  $\mu'$ 、方差  $\sigma^2$  和  $\sigma'^2$  的分布中采样的  $B$  和  $T$  的潜在向量。 $\hat{\alpha}$  和  $\alpha$  分别是预测的贡献得分和真实得分。 $D_{KL}$  表示 Kullback–Leibler 散度，BOW 表示词袋损失。橙色背景突出显示的是模型的预训练组件。浅蓝色圆圈表示输入和输出向量，深黄色方框表示神经网络。c, DeepBlock 的生成过程。d, 生成过程中的属性控制。Optimization 表示优化算法，包括 SA 和 BO。e, 分子碎片化与重构。数字 1 到 7 表示不同的构建块。

## 2.3 分子生成

## 2.4 结合贡献

在 DeepBlock 中，我们定义了残基的结合贡献系数，其与残基和蛋白质口袋中心之间的距离呈负相关，并开发了一个神经网络来自动预测该系数（详情参见方法部分“残基的结合贡献”）。在此，我们分析了模型预测的贡献值与真实值之间的相关性。计算得到的 Pearson 相关系数为 0.68，具有统计显著性  $p < 0.001$ ，表明预测值与实际值之间存在正相关性。以 ABL2 蛋白为例，我们在图 3a 中展示了预测值与实际值的对比分布（其他测试蛋白见补充图 6–11）。其中两条曲线的紧密对齐表明模型在准确捕捉贡献系数相对大小方面的能力。如图 3b 所示，结合位点的整体预测贡献值高于其他蛋白质残基。然而，对于某些特定的蛋白质和残基，其预测的贡献值并未严格遵循基于距离的负相关规则。出现这种现象的主要原因有两个：首先，贡献值预测作为辅助任务，模型可能无法完全学习到，从而存在固有的误差；其次，模型可能检测到多个结合位点，而真实值仅反映在缺乏已知靶标结构的情况下，我们还基于自动预测的贡献评分（而非从结构信息中获得的真实值）评估生成分子的质量。如图 3c 所示，使用预测贡献评分生成的分子亲和力略低于使用结构信息生成的分子，但在药物相似性、合成可行性以及与已知配体的相似性上表现相当。亲和力的这一细微差异进一步验证了我们贡献预测模块的准确性和可靠性。

此外，我们还对贡献预测模块进行了消融实验。如图 3c 所示，docking 配体结合亲和力。DeepBlock 在缺乏贡献预测模块的情况下，其亲和力略低于包含该模块的 DeepBlock。此外，药物相似性、合成可行性以及与已知配体的相似性也显示出轻微的下降。这种对接亲和力的差异并不显著，这可以归因于贡献预测模块并未引入额外信息，而是辅助学习蛋白质序列的表示。因此，虽然贡献预测模块可以提高生成分子的质量，但其重要作用之一是作为分析蛋白结合位点的辅助工具。

## 2.5 新靶标的配体生成案例研究

为展示 DeepBlock 在缺乏结构信息的情况下针对新蛋白生成配体的能力，我们选择 KIAA1363 作为案例进行分析。KIAA1363（基因名 NCEH1）是一种丝氨酸水解酶，已被观察到与肿瘤细胞的侵袭性呈正相关<sup>??</sup>。识别针对 KIAA1363 的活性抑制剂有望成为一种有效的抗癌治疗策略<sup>??</sup>。在此，我们基于 KIAA1363 的氨基酸序列使用 DeepBlock 生成了 100 个配体，并随后选择了 5 个具有最高对接亲和力的分子。如图 4a 所示，生成的五个样本分子具有与已知抑制剂 JW480<sup>??</sup>相似的子结构和对接口袋，同时表现出高对接亲和力、良好的药物相似性和可合成性，这表明 DeepBlock 在基于靶标序列的配体生成中的潜力。

此外，我们进行了长期动力学模拟以验证结合稳定性。从配体的均方根偏差（RMSD）曲线（图 4c）可以看出，JW480 和生成的分子均与配体保持稳定结合。进一步的轨迹分析显示，JW480 主要与 KIAA1363 的 Gly114 和 Leu36 形成不稳定的范德华相互作用，而我们生成的分子与 Gly114 形成了更稳定的范德华相互作用，同时还观察到了氢键相互作用。生成的分子能够在结合口袋中形成类似于已知抑制剂的关键相互作用，但更加稳定。此外，我们还选择了一个经过良好表征的靶标——严重急性呼吸综合征冠状病毒 2 (SARS-CoV-2) 主蛋白酶 (Mpro) 进行验证。对接模拟进一步确认了 DeepBlock 生成的分子与 Mpro 靶标的结合稳定性（补充图 12）。

为了研究模型的块重构机制与回溯合成路径之间的关系，我们可视化了模型预测的块序列的重构过程及 Retro\* 预测的回溯合成路径（图 4a 左上样本）。在图 4d 的左侧，DeepBlock 的块解码器生成了长度为 6 的块序列，标记为 1 到 6。在图 4d 的右侧，Retro\* 预测了三种反应物的两步反应路径。值得注意的是，反应设计中的三个反应物可以直接对应于活性块中的某些块。这表明早期阶段使用的化学键切割方法可以准确地作用于反应性化学键，同时保留子结构的化学意义。此外，反应物的相应块在原始块序列中也保持连续关系，表明模型能够有效组合块并表达结构信息。

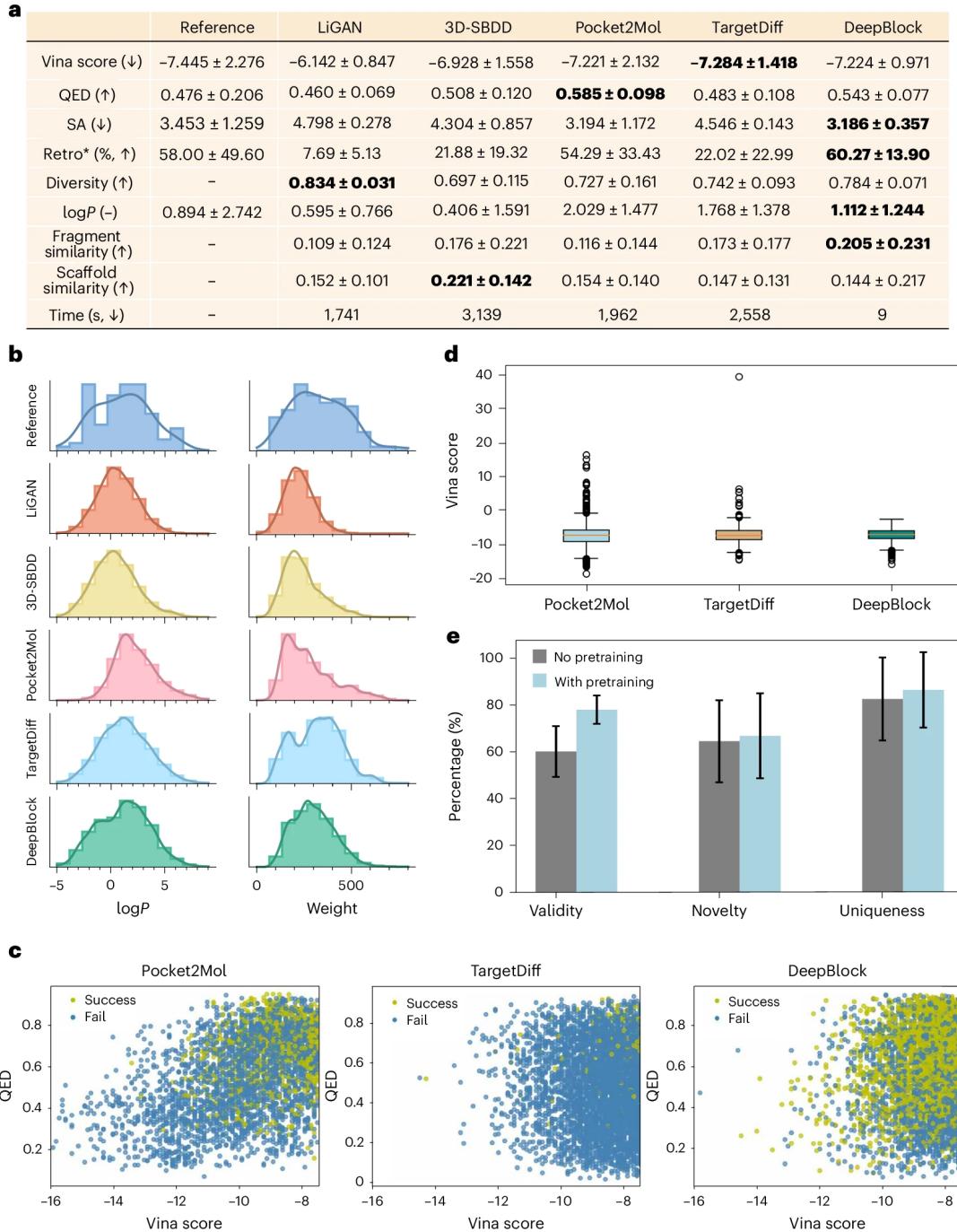


图 2: 提出的方法与基线的对比。**a**, 参考分子以及基线方法和我们方法生成的分子的对比指标, 其中粗体表示我们结果中与基线相比最优的值。我们首先分别对每个蛋白的样本组计算均值, 然后计算整体均值  $\pm$  标准差 (无偏)。粗体格式表示最佳值。**b**, 所有生成分子和参考分子的  $\log P$  与分子量的分布。**c**, 生成分子的散点图。每个点代表一个分子, 并根据其 Retro\* 得分进行着色。红色表示通过 Retro\* 无法预测合成路径的分子, 表明它们难以合成; 绿色表示能够成功预测合成路径的分子, 表明它们易于合成。**d**, 生成分子的 Vina 评分的箱线图。中心线表示中位数; 需表示 1.5 倍四分位距; 箱线图的上下限分别表示最大值和最小值。**e**, 预训练与未预训练的性能对比。所有值均以均值  $\pm$  标准差表示 ( $N = 20,000$  个分子)。

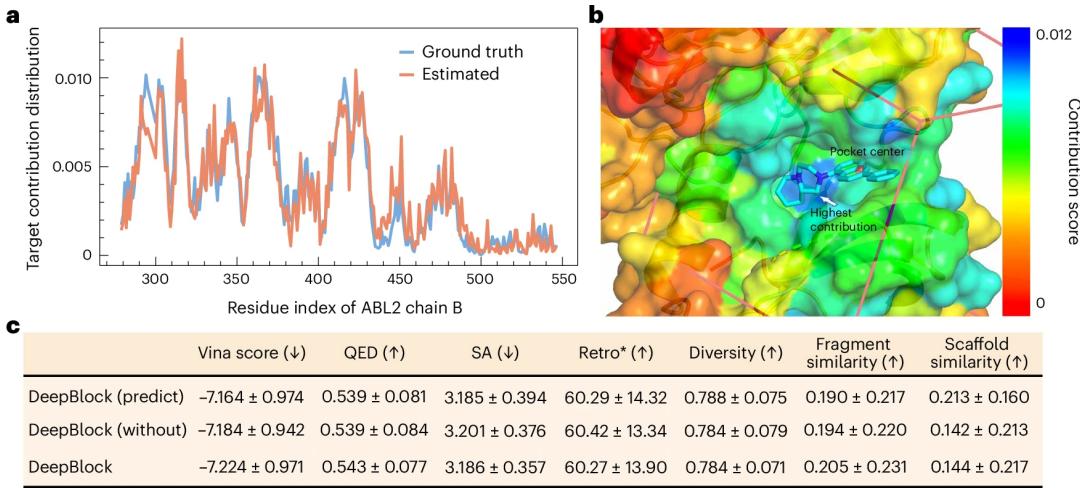


图 3: 残基的结合贡献分析。a, 靶标贡献的真实值与估计分布的对比, 以 ABL2 (Protein Data Bank ID 4XLI) 为例。b, 靶标贡献的可视化。颜色条表示贡献评分, 其中蓝色表示较高的评分, 红色表示较低的评分, 黄色表示中间值。评分范围从 0.0 (最低) 到 0.012 (最高)。红色线框表示模拟的对接口袋, 尺寸为 25 Å (长度、宽度和高度)。靶标中心区域显示出较高的预测贡献值。在 3D 分子结构中, 碳原子 (C) 用青色表示, 氮原子 (N) 用蓝色表示。c, DeepBlock 在具有预测贡献系数 (DeepBlock (predict))、无贡献预测模块 (DeepBlock (without)) 和完整模型 (DeepBlock) 下的性能比较。( $\uparrow$ ) 表示越高越好, ( $\downarrow$ ) 表示越低越好。

## 2.6 分子属性控制

**结合亲和力优化。**在此, 我们从 CrossDocked 2020 (参考文献??) 的测试集中选择了 “F16P1 (3kc1)” 作为目标受体, 并从 ChEMBL??数据集中随机选择了 5,000 个小分子作为优化的初始分子 (详细信息请参见方法部分的 “结合亲和力优化”)。图 5a 中的每个数据点表示特定结合亲和力范围内分子的优化结果 (优化前亲和力  $\pm 0.5$ )。例如, 在优化前亲和力范围为  $-7 \pm 0.5$  的分子中, 有 57.53% 的分子成功优化, 优化前后的平均相似性值为 0.307。可以观察到, 初始结合亲和力较低的分子更能体现出模型的优化效果。尽管这些分子经过了对接能力的改进, 其相似性曲线的变化仍相对平滑, 确保了分子相似性的保留。在图 5b 中, 我们更直观地展示了低亲和力分子优化后的数值变化。可以看到, 模型有效优化了初始亲和力大于  $-7.2$  的分子, 将其对接评分保持在约  $-8$  左右。

使用预优化结合亲和力的中值  $-7.2$  作为阈值, 我们进一步分析了 2,346 组预优化亲和力大于  $-7.2$  的低亲和力分子在优化前后的分子属性变化。如图 5c 所示, 76.04% 的分子显示出结合亲和力的改善, 平均提升超过  $0.5 \text{ kcal mol}^{-1}$ 。尽管与原始的 5,000 个小分子 (已经是类药物化合物) 相比, 预测结果的药物相似性和可合成性略有下降, 但 Lipinski “五规则” 的平均满足度略有提高, 表明其具有较高的类药物潜力。对于这一子集的分子, 优化前后的平均相似性为  $0.288 \pm 0.125$ 。作为参考, 我们还测试了从 ChEMBL ??中随机选择的 5,000 对分子的平均相似性, 结果为  $0.268 \pm 0.086$ 。考虑到所做的修改

对于低亲和力分子, 模型在中低亲和力分子上的结构特征保留仍在理想范围内。最后, 我们随机选择了一部分优化后的分子进行可视化 (补充图 13)。这些分子在结合亲和力方面表现出不同程度的改善, 但其他属性的波动并不一致。值得注意的是, 优化前后分子的结构高度相似, 关键子结构得到了保留。

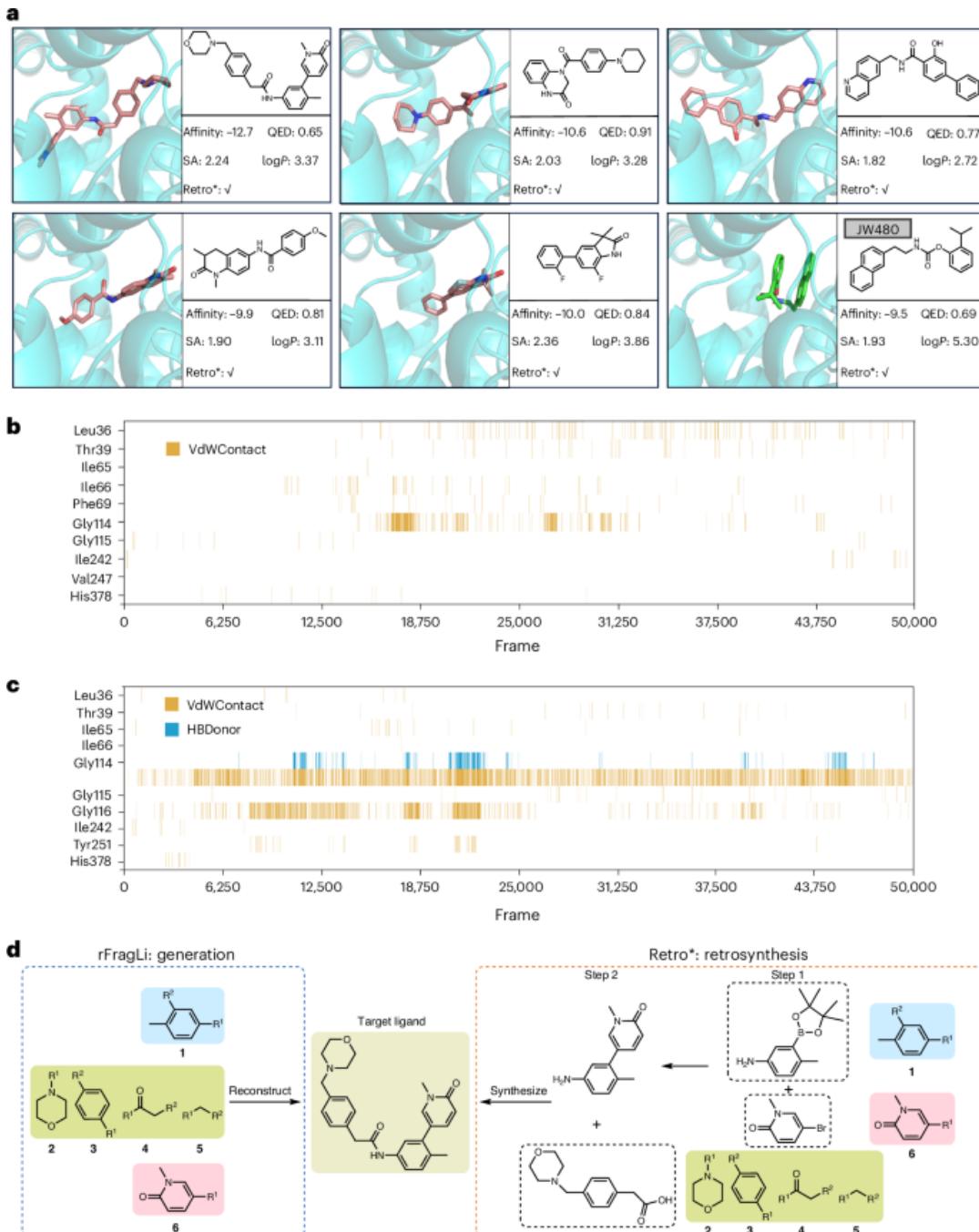


图 4: KIAA1363 靶标的配体生成案例研究。**a**, 针对 KIAA1363 靶标生成的配体样本, 显示其对接后的 3D 构象 (碳原子为粉色, 氮原子为蓝色, 氧原子为红色)。与抑制剂 JW480 对比 (碳原子为绿色, 氮原子为蓝色, 氧原子为红色)。我们还计算了它们的亲和力及其他指标, 其中 Retro\* 的勾选标记表示成功预测出回溯合成路径。**b**, JW480 的残基水平蛋白-配体相互作用。**c**, DeepBlock 生成的具有最高亲和力的分子 (a 中左上角的分子) 的残基水平蛋白-配体相互作用。VdWContact 和 HBDonor 分别表示范德华接触和氢键供体。**d**, 生成配体的合成可行性。在左侧为我们生成的分子块, 中间为重构的分子, 右侧为 Retro\* 预测的回溯合成路径。

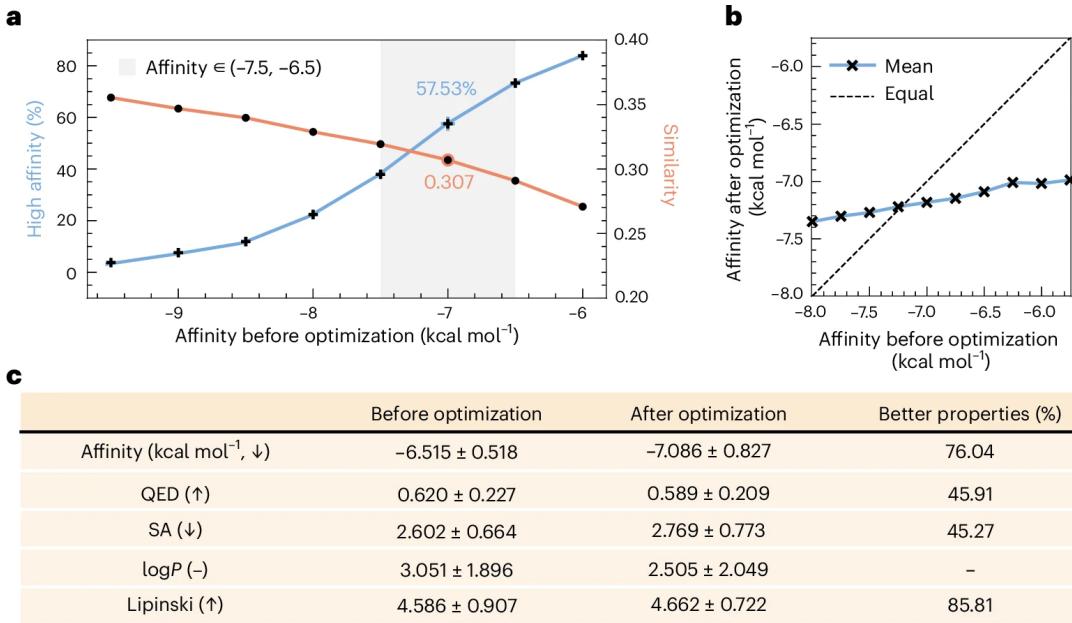


图 5: 优化前后亲和力的对比。a, 每个样本点表示特定亲和力范围内（优化前  $\pm 0.5$ ）的分子优化结果。初始亲和力较低的分子更能体现出模型的优化效果。b, 更直观地展示低亲和力分子优化后的数值变化。c, 对低结合亲和力分子的优化性能。大多数分子的对接能力得到了改善。

## 2.7 靶标感知的分子优化

我们提出了两种靶标感知的分子优化方法——基于模拟退火 (SA) 的方法 (SATMO) 和基于贝叶斯优化 (BO) 的方法 (BOTMO)，分别在离散分子空间和分子潜在空间中进行优化。与现有的分子优化方法不同，我们的方法引入了靶标感知的分子优化，并结合靶标约束，确保生成的分子在优化过程中保持与靶标的亲和力。如图 6 所示，这两种优化方法在提高毒性和 QED 的同时，保持或略微提升了对接亲和力和合成可行性。BOTMO 方法的性能受到迭代过程中生成的新候选分子数量  $N$  的影响。如图 6a 所示，较大的  $N$  通常会带来更好的优化结果，但同时也延长了优化时间 (图 6d)。当  $N = 10$  时，优化结果与 SATMO 相当，但优化速度较慢。此外，如图 6g 所示，我们展示了 SATMO 优化前后分子能量和各项属性的分布。可以观察到，优化后能量和毒性显著降低。然而，亲和力在优化后显示出双峰分布，这可能是毒性和亲和力之间的权衡结果。在药物相似性方面，与优化前相比，优化后的分布在 0.5 以上出现峰值，且右侧具有更高比例的高类药物性分子。

## 3 Discussion

DeepBlock 利用反应性构建块的优势，促进具有高类药物性、易合成性和高对接亲和力分子的理性设计。反应性块的质量是决定重构分子整体质量的关键因素。在我们的方法中，我们构建了一个包含 10,701 个块的广泛字典，涵盖了大量常用的片段。通过预训练步骤，我们显著扩展了模型的化学空间。这一过程减少了蛋白-配体对数据集规模有限所带来的过拟合风险，从而提升了模型的整体性能。然而，我们的方法也存在一定的局限性。DeepBlock 目前只能从现有的块字典中生成块，这限制了其生成分子的多样性。未来研究的方向之一是探索块的 de novo 生成方法，从而摆脱现有块字典的限制，解锁生成更具多样性和新颖性的分子的潜力。此外，DeepBlock 以 SMILES 字符串生成二维 (2D) 分子结构，实现了可控的属性和对新靶标的适用性。尽管 SMILES 字符串在多种药物开发场景中提供了

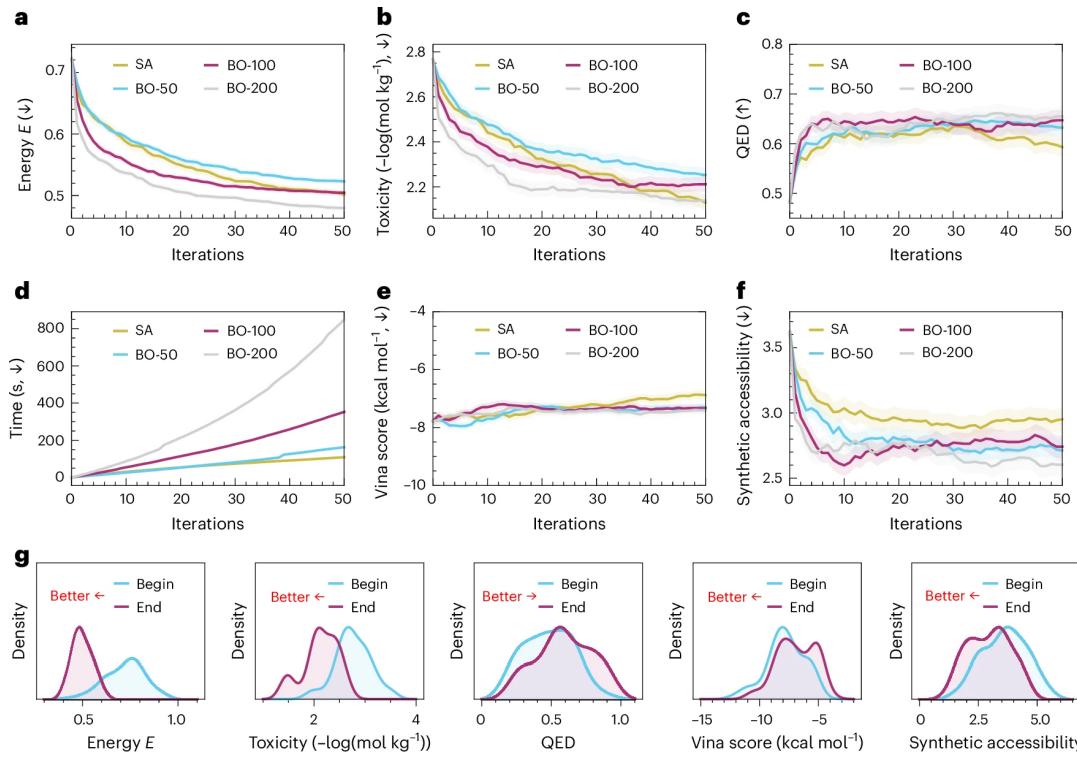


图 6: 优化过程与结果。a–c, 优化迭代过程中能量函数 (a)、毒性 (b) 和 QED (c) 的变化情况, 从 100 次独立优化过程中收集数据。BO-50、BO-100 和 BO-200 分别表示 BO 方法在迭代过程中生成  $N = 50, 100$  和 200 个新候选分子的情况。阴影部分的宽度表示标准差 ( $N = 100$  次独立优化过程), 图中对标准差应用了 0.1 倍缩放以便更好地可视化。d, 从实验开始到每次迭代结束所花费的时间。e, f, 每次迭代的 Vina 评分 (e) 和合成可行性 (f)。显示了平均模型评分及其标准差 ( $N = 100$  次独立优化过程)。g, 通过 SA 算法优化前后分子属性的分布。

足够的结构信息, 但缺乏三维 (3D) 结构细节。未来研究将重点将我们的方法与 LiGAN 等方法相结合, 基于分子构建块开发可控的 3D 分子生成方法。这种混合方法可能会结合 2D 和 3D 药物设计方法的优势, 进一步提高药物发现的效率和有效性。

## 4 Methodology

### 4.1 数据集

**蛋白-配体对训练数据。**为了训练整个模型, 我们使用了 CrossDocked 2020 数据集 ??, 最初包含 2,250 万蛋白-配体对。根据文献 16 中描述的过滤和拆分方法, 我们筛选了结合构象 RMSD 大于 1 Å 的数据点, 最终得到 184,057 个精炼子集。为了确保训练集和验证集之间没有重叠, 我们使用 mmseqs2 ?? 在 30% 序列同一性下对数据进行聚类。从这些聚类中, 我们随机选择了 10,000 个蛋白-配体对用于训练, 并从剩余聚类中选择了 100 个蛋白用于测试。此外, 我们使用 20% 的训练数据用于超参数调整和模型选择。最终结果报告于本文中

**蛋白质** 蛋白质序列通常使用 20 种氨基酸进行编码, 因此构建字典相对容易。我们直接利用 ESM-2 的内置转换工具, 实现了端到端的嵌入向量 (氨基酸特征) 生成。

**预训练的类药物分子数据** 在预训练阶段，我们使用了一个大规模的类药物分子数据集。ChEMBL 是由欧洲生物信息学研究所维护的一个开放数据库，专用于药物发现 ??。它包含大量的药物和生物活性数据，包括化学结构、物理化学性质以及药理活性数据。我们使用的数据集版本为 ChEMBL31，其中包括约 230 万个以 SMILES 字符串表示的药物分子。

**分词和嵌入** 分子或配体。我们按照“分子碎片化与重构算法”中的描述处理每个类药物分子或配体分子，以获得块序列。对于 ChEMBL 数据集，我们仅保留出现次数超过 100 次的块（见补充图 14），最终得到 5,109 个独特块。对于 CrossDocked 数据集，我们保留了所有块，总计 7,483 个。在过滤、去重、合并和排序之后（见补充表 2），我们得到一个包含 10,701 个词汇（包括 4 个特殊词）的块字典。有关详细的处理过程，请参见补充图 14 和补充表 2。我们随机初始化了配体编码器模型的嵌入层参数，并采用该词汇表的维度大小。

## 4.2 分子碎片化与重构算法

我们开发了一种可靠的方法，将分子转换为块序列，并确保其可逆性和与各种化学键断裂规则的正交性。

1. **通过利用 BRICS 规则 ??**，我们追踪可断裂的化学键，并在单步操作中提取所有不可再分的最小块。这些块随后被抽象为化学独立的图结构。接下来，我们设计了一种图搜索方法，将图转换为序列。节点遍历的顺序严格按照块的相对质量和断裂点的原子索引定义。
2. **在重构分子的过程中**，我们从序列中顺序读取块和标识符，以构建图。我们提取断裂的化学键，并连接这些块以形成完整的分子。

为了确保我们分解方法的唯一性，我们采用以下方法：

- (1) **BRICS 规则**: 我们应用 BRICS 规则来识别可断裂的化学键，并将其切断，生成最小的不可分子结构或块。这确保了不同分子之间的一致且可重复的分解。
- (2) **断裂点排序**: 在将块序列化为序列时，我们基于 RDKit 原子索引对断裂点进行升序排序，该索引来源于 SMILES 任意目标规范 (SMARTS) 表达式中的原子位置。
- (3) **图遍历起点**: 我们从连接度最高的块开始图遍历，遍历顺序由先前确定的断裂点排序决定。

**从分子到块序列** 为了便于说明，我们以 KRAS ??抑制剂药物 Adagrasib ??为例演示碎片化方法，如补充图 15 所示。在补充图 15b 中，BRICS 规则指出 Adagrasib 中存在七个可断裂的化学键，以可断裂字母表示。我们还突出显示了可断裂键两端的原子，图中的数字表示每个原子的索引。

通过断开所有七个化学键并在键的每一端连接一个虚拟原子，我们得到八个块，如补充图 15c 所示。这些块被称为节点，用数字  $N$  表示。虚拟原子  $x$  的符号用希腊字母（例如， $\alpha$ 、 $\beta$  等）表示，在此上下文中，我们将这些虚拟原子称为断点。总结来说，一个节点包含多个不等价的断点，这些断点必须连接到相邻节点的断点，从而在补充图 15d 中形成一个独特的图结构。

为了区分节点内的不同断点，我们基于 RDKit ??原子索引对断点进行升序排序，该索引由 SMARTS 表达式中原子的顺序确定。节点内的每个断点依次重新分配一个断点标签，其中  $x = \alpha, \beta, \dots, x_p$ ,  $P_x$  表示块  $N$  中的断点总数。

**简化块序列** 在本文的 Adagrasib 示例中，我们获得了长度为 15 的序列，但其中包含 7 个标识符。从补充图 15g 可以识别出节点 1、8 和 6 为理论叶节点，这些节点仅通过 1 个标识符在断点处与父节点连接。在没有该符号的情况下，如果父节点连接到断点  $x$ ，仍然可以确定该断点。此外，更短的序列更有利於模型学习并降低计算成本。

### 4.3 DeepBlock 的架构

在 DeepBlock 中，基于靶标蛋白  $T$  生成块  $B$  的生成过程可以表示为条件分布  $p(B|T) = p(B|z, T)p(z|T)dz$ ，其中  $z$  表示潜在向量。我们的目标是使用参数化为  $\theta$  的深度神经网络来近似  $p(B, z|T)$ ，记为  $p_\theta(B, z|T)$ 。在我们的模型中，如图 1b 所示， $p_\theta(B|z, T)$  对应于先验网络，而  $p_\theta(z|T)$  对应于块解码器。在分子生成阶段（图 1c），我们首先从先验网络  $p_\theta(z|T)$  中采样一个潜在向量，然后使用  $p_\theta(B|z, T)$  生成  $B$ 。为了训练先验网络  $p_\theta(z|T)$ ，我们引入了由  $\phi$  参数化的识别网络，记为  $q_\phi(z|B, T)$ ，用于近似真实的后验分布  $p(z|B, T)$ 。我们通过最小化  $q_\phi(z|B, T)$  与  $p_\theta(z|T)$  之间的 KL 散度 ??，并使用  $p_\theta(B|z, T)$  从潜在空间  $N(\mu, \sigma^2)$  或  $N(\mu', \sigma'^2)$  中随机采样潜在向量  $z$ ，以提高模型的泛化能力。

BGNet 的输出生成的块用于编码配体和蛋白结合口袋的序列。此外，结合贡献感知网络用于识别残基与靶标口袋中心之间的相互作用。残基与靶标位点中心之间的贡献得分记为  $\alpha$ 。靶标口袋中心附近的更高得分通常表明更高的贡献得分，表明在配体结合中更相关。这些靶标贡献用于加权编码器学习的蛋白质残基特征的求和，最终有助于靶标特定的表示。

**块编码器与解码器** 块编码器使用门控循环单元 (GRU) ??门控机制。它学习配体分子块序列  $B$ ，并从 GRU 的每一层获取隐藏状态，作为下游任务的输入。块解码器使用可微分的重构方法将隐藏表示转换回分子结构。生成损失函数被定义为优化分子序列的重构准确性。

隐藏状态被连接起来以形成配体表示  $x = \text{GRU}(B)$ 。配体解码器  $\text{Decoder}_L$  同样基于 GRU。给定潜在向量  $z$  和靶标表示  $T$ ，它重构序列  $B = \text{GRU}(W(z)c)$ 。其中  $c$  表示  $z$  和  $w$  的连接，并且表示在将  $z|c$  传递给配体 GRU 解码器之前用于调整维度的权重矩阵。

此外，我们在配体解码器中加入了 BOW 损失，以解决条件变分自编码器 (CVAE) 中潜在变量消失的问题。BOW 损失将目标序列表示  $x$  分解为两个变量：一个保留词序的变量，记为  $x_{\text{bow}}$ ，另一个忽略词序的变量，记为  $x_{\text{sum}}$ 。

假设在给定  $z$  和  $c$  的条件下， $x_{\text{bow}}$  和  $x_{\text{sum}}$  是条件独立的，我们有：

$$p(x, z|c) = p(x_{\text{bow}}|z, c)p(x_{\text{sum}}|z, c)p(z|c)$$

为了近似  $p(x|z, c)$ ，我们使用神经网络并最小化负对数似然  $\log p(x_{\text{bow}}|z, c)$ 。通过这种方式，潜在变量捕捉到目标序列的全局信息。近似公式如下，其中 MLP 表示多层感知器：

$$P = p(x_{\text{bow}}|z, c) = \text{Softmax}(\text{MLP}_{\text{bow}}(zc))$$

**残基的结合贡献** 在蛋白-配体相互作用研究中，配体与蛋白质残基之间的距离通常会影响它们的相互作用和结合亲和力。通常，当残基与配体之间的距离显著减少时，其相互作用减弱，结合亲和力降低。因此，我们设计了一个系数  $a_i$ ，称为残基的结合贡献，其与靶标口袋中心到各残基的距离呈负相关：

$$a_i = \text{Softmax}\left(-\left(\frac{d_i}{H}\right)\right)$$

这里  $k = 20$ ， $r = 2.5$ ， $d_i$  表示从残基质心到结合位点中心的距离。数据集中参考配体分子的质心用作口袋中心，并计算所有残基到口袋中心的距离（单位为埃；参见补充图 17）。

我们使用多层感知器网络预测每个残基的结合贡献。分子仅在蛋白结构的有利结合位点处结合，因此我们可以更多依赖靠近结合位点的残基来生成配体分子。因此，在靶标编码器中，残基特征通过方程 (3) 加权平均以获得靶标特征。模型使用方程 (7) 的损失函数最小化预测值与实际值之间的误差。

残基的结合贡献作为本研究中的辅助训练机制，具有多重优势。事实上，仅通过学习预测值和实际值的设计，模型能够持续感知配体残基之间的相互作用。即使没有实际值，理论上模型也可在训练数据中缺乏三维结构的情况下使用预测值进行贡献分析。通过在空间中采样，模型不需要额外的结构信息，仅需蛋白质序列作为输入。

#### 4.4 BGNet 训练

**重构损失。**我们使用交叉熵来衡量输入配体块序列与模型重构序列  $B$  之间的重构损失。假设序列长度为  $K$ ,  $y_i$  表示序列  $B$  中第  $i$  个块的预测概率分布向量,  $C$  表示类别数:

$$\mathcal{L}_{\text{recon}}(y, y') = -\frac{1}{K} \sum_{i=1}^K y_i \log y'_i$$

**BOW 损失。**我们实际上应用了一个多分类器，并使用交叉熵来计算其损失。在此情况下， $p$  表示实际序列  $B$  中块类别的多热编码向量，而  $\hat{p}$  表示  $p$  的预测概率分布向量。 $C$  表示类别数:

$$\mathcal{L}_{\text{bow}}(p, \hat{p}) = -\frac{1}{C} \sum_{i=1}^C p_i \log \hat{p}_i$$

**潜在空间分布损失。**为了使先验网络学习配体分子信息，我们需要最小化预测概率分布  $q_\phi(z|B, c)$  为  $N(\mu', \sigma'^2)$  与先验网络预测的概率分布  $p_\theta(z) = N(\mu, \sigma^2)$  之间的差异:

$$\mathcal{L}_{\text{latent}}(q, p) = D_{\text{KL}}(q_\phi(z|c, B) \| p_\theta(z|c))$$

**贡献分布损失。**我们利用离散 KL 散度最小化预测的靶标贡献值与实际值之间的差异，从而引导模型提高结合位点信息的准确性:

$$\mathcal{L}_{\text{contrib}}(a, \hat{a}) = \frac{1}{C} \sum_{i=1}^C a_i \log \left( \frac{a_i}{\hat{a}_i} \right)$$

在此阶段，我们设置  $c = 0$ ，并仅优化配体编码器、识别网络和配体解码器（包括 BOW 多分类器）。损失函数表示如下:

$$\mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{bow}} + \lambda_2 \mathcal{L}_{\text{latent}} + \lambda_3 \mathcal{L}_{\text{contrib}}$$

**潜在空间分布损失** 此处潜在空间分布损失  $\mathcal{L}_{\text{latent}}$  被修改为最小化由推理网络预测的概率分布  $q_\phi(z, c|x, y)$  与标准正态分布  $N(0, I)$  之间的差异:

$$\mathcal{L}_{\text{latent}}(q_\phi, p) = D_{\text{KL}}(q_\phi(z|c, B) \| N(0, I))$$

其中  $D_{\text{KL}}(P \| Q)$  表示分布  $P$  和  $Q$  之间的 KL 散度。此阶段的生成模型本质上是一个 VAE 结构。我们使用 Adam 优化器 ?? 来优化  $\mathcal{L}_{\text{latent}}$ 。学习率随着训练轮次递减，并使用 KL 退火 ?? 来调整  $\omega_t$  的值，从第一轮的 0 逐渐增加到后期的 1，如第 5.1 节所述。经过 200 轮训练后，我们选择在验证集上整体损失  $\mathcal{L}_{\text{recon}}$  最小的模型参数作为后续训练阶段的初始参数。

**训练** 正式训练的流程如图 1b 所示，使用蛋白-配体对数据集 CrossDocked 2020。我们使用 Adam 优化器 ?? 最小化以下损失函数:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{bow}} + \lambda_2 \mathcal{L}_{\text{latent}} + \lambda_3 \mathcal{L}_{\text{contrib}}$$

在预训练阶段使用的退火方法与正式训练阶段相同，即  $\omega_t = 1$ 。在训练 200 轮后选择模型参数作为最终测试基准。在这一点上，靶标结合编码器能够捕捉结合位点信息并编码靶标表示。先验网络可以从靶标信息预测配体分子特征，以进行后续采样和重构操作。

#### 4.5 分子生成

分子生成的流程如图 1c 所示。其仅需输入靶标蛋白质氨基酸序列。通过靶标编码器（包括靶标结合贡献感知），得到基于靶标的条件分子潜在空间  $N(\mu', \sigma'^2)$ 。从潜在空间中进行多次随机采样，可生成多样化的靶标对齐的分子序列，并通过配体解码器解码。然后按照“从片段序列到分子”中描述的方法，模型生成一组过渡配体分子。重构的分子将进行合法性、唯一性和生成理论初始化合物的能力检查 ??。此过程重复进行，直到收集到足够数量的分子。

## 4.6 分子属性控制

在此，我们使用 DeepBlock 进行结合亲和力优化（“结合亲和力优化”）和靶标感知的分子优化（“靶标感知分子优化”）。原则上，当输入一组初始配体、蛋白质靶标及其氨基酸序列并重新采样时，模型可有效生成新的分子。例如，我们随机从 ChEMBL 数据集中选择 5,000 个小分子作为初始分子进行优化。每个分子输入 BGNet，如补充图 19 所示。我们观察到高亲和力的分子具有更多与理论结合位点相对齐的分子片段，证明了我们方法的可行性。

**结合亲和力优化** 通过将靶标蛋白和具有较差亲和力的分子输入经过训练的 DeepBlock，模型可以生成对原靶标具有更强亲和力的新分子。我们评估了亲和力评分的变化。为此，我们从 CrossDocked 2020 测试集中选择了靶标 “F16P1 (3kc1)”。从 ChEMBL 数据集中采样 5,000 个分子并输入 BGNet，如补充图 19 所示。改进后的分子包含的块随后被重构为优化后的分子。我们使用 Tanimoto 指纹 ?? 评估优化前后分子的亲和力评分，并比较对接亲和力的变化。

**靶标感知的分子优化** 补充算法 3 实现了基于模拟退火 (SA) 的靶标感知分子优化方法。该方法使用采样策略迭代优化分子集。如果新分子集的能量低于当前分子，则新分子将替代前者。如果新分子的能量低于当前分子，则以更高的概率  $P$  生成新的分子集。分子能量函数表示为：

$$E(x) = \omega_1(\text{QED}(x) - 0.5) + \omega_2(\log P + 0.3)$$

在 BO 的实现中，我们同样使用能量函数作为目标函数。通过识别网络对初始分子进行编码，将其作为潜在向量  $z$ ，并使用高斯过程来拟合潜在向量与训练数据，如补充图 20 所示。我们执行 BO 循环以期望改进，在每次迭代中生成新的候选分子并添加到高斯过程中，从而提高亲和力评分。最终的计算机优化包括  $\mathcal{L}_D$  (半数致死剂量) 子数据集，来源于 TOXRIC (急性毒性) ??，以及无量纲  $\log P$  用于训练毒性预测器 TOXCV。我们使用 MACCS (分子访问系统) 指纹 ?? 作为特征表示，并利用自动化机器学习工具 TPOT ?? 训练和选择用于定量估算分子毒性的最佳回归模型和超参数。在测试集上的回归性能如下：均方根误差 (RMSE) = 0.4326,  $R^2$  = 0.6029，优于 TOXRIC 项目提供的基准 ??。

为了评估 DeepBlock 的分子优化能力，我们基于靶标受体 CCPR (1a2g) ?? 生成了 100 个不同的初始分子。方程 (11) 中的能量函数权重设置为  $\omega_1 = 0.6$  和  $\omega_2 = 0.4$ 。每个分子经过  $N = 50$  次由补充算法 3 进行的优化迭代。我们还在 BO 的期望改进过程中测试了不同候选数量 (50、100 和 200) 的分子优化。由于 BO 无法对每个分子进行独立优化，因此我们在每次迭代中选择前 100 个分子进行分析，并研究了优化过程中各种分子属性的变化和优化结果。

## 4.7 指标

1. 我们使用 QuickVina2 ?? 计算 Vina 评分 ??，它是对分子与其靶标之间结合亲和力的基于物理的估计。对于 DeepBlock 生成的 SMILES，我们使用 ETKDG 算法 ?? 生成分子的初始 3D 构象。
2. QED 用于定量评估类药物性，衡量分子成为药物候选物的可能性 ??。得分大于 0.5 的化合物通常被认为具有成为药物分子的潜力。
3. 合成可行性 (SA) 用于评估化合物的合成难度。本文使用 SAscore ??，得分越低表示化合物越容易合成。
4.  $\log P$  反映了物质在细胞-水相中的分布 ??。
5. 我们使用回溯合成工具 Retro\* ?? 预测生成分子的合成路径，并计算成功预测的比例。
6. 计算从相同蛋白生成的分子对之间的 Tanimoto 指纹的平均相似性 ??，以衡量模型生成分子的多样性，这反映了模型的化学空间。

7. 片段相似性比较生成集和参考集中的 BRICS 片段分布。
8. 骨架相似性比较生成集和参考集中的 Bemis–Murcko 骨架频率 ??。
9. 我们计算每种策略生成的 100 个分子的分子指纹。

#### 4.8 分子动力学模拟

在获得蛋白–配体复合物结构后，我们首先使用 Maestro 中的 Protein Preparation Workflow 模块对蛋白质进行预处理，以防止不正确的结构导致问题。对于 AMBER 力场，我们分别对配体和蛋白质采用 AMBER14SB ?? 和 AMBERGSB7。水分子参数化采用 TIP3P。在进行 5,000 次能量最小化和 10 ps 的约束力学模拟后，进行了 500 ps 的非约束分子力学模拟。生成复合物的稳定性通过配体相对于结合位点的均方根偏差（RMSD）进行评估。

#### 4.9 统计与可重复性

数据分析使用 Python 3.10、pytorch 1.12.1、rdkit 2020.09.5 和 QuickVina2 进行。未排除任何数据分析。 $P < 0.05$  被认为具有统计学显著性。未使用统计方法预先确定样本量。实验和数据分析均为随机化进行，实验人员在实验和结果评估过程中保持盲测。为了确保结果的可重复性，我们在每个实验的图例和文本中报告了统计数据。

#### 4.10 报告摘要

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

### 5 数据及代码可用性

图 2–6 的数据源随本文提供。本研究中使用的所有数据集均为公开可用。CrossDocked 2020 数据集的原始数据来自<https://github.com/gmnn/models/tree/master/data/CrossDocked2020>。用于预训练 BGNet 的数据集来自 ChEMBL 数据库 ([https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl\\_31](https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_31))。用于训练模型的处理数据集可通过 figshare 获取：[https://figshare.com/articles/dataset/crossdocked\\_pocket10\\_with\\_protein.tar.gz/25878781](https://figshare.com/articles/dataset/crossdocked_pocket10_with_protein.tar.gz/25878781)。小鼠腹腔注射  $\mathcal{L}_{D50}$  子数据集来自 TOXRIC (<https://toxric.bioinfor.cn/tech/home>)。

训练模型的源代码和权重可在 GitHub 上获取，链接为<https://github.com/BioChemAI/DeepBlock>，并已上传至 Zenodo，DOI 为<https://doi.org/10.5281/zenodo.13852436>（参考文献 ??）。