arXiv:2302.01435v1 [cs.CE] 2 Feb 2023

使用潜在空间近似轨迹收集器进行靶向特异性肽设计

Tong Lin[3,4†], Sujie Chen[1†], Tong Lin[3,4†], Dehua Pei[2*], Xiaolin Cheng[1*]和 Levent Burak Kara[3*]
[1*]俄亥俄州立大学药学院，美国俄亥俄州哥伦布市 W Lane Ave 281号，邮编 43210。[2]俄亥俄州立大学化学与生物化学系，美国俄亥俄州哥伦布市 W Lane Ave 281号，邮编 43210。[3]卡内基梅隆大学机械工程系，美国宾夕法尼亚州匹兹堡市 Forbes Ave 5000号，邮编 15213。[4]卡内基梅隆大学机器学习系，美国宾夕法尼亚州匹兹堡市 Forbes Ave 5000号，邮编 15213。[5]卡内基梅隆大学电气与计算机工程系，美国宾夕法尼亚州匹兹堡市 Forbes Ave 5000号，邮编 15213。

*通讯作者。电子邮箱：pei.3@osu.edu; cheng.1302@osu.edu; lkara@cmu.edu;
†这些作者对本研究的贡献相同。

摘要

尽管深度学习在从头分子设计中的应用十分普遍且成果显著，但针对特定蛋白质的肽生成问题仍未得到解决。主要障碍之一是高质量训练数据匮乏。为解决这一问题，我们提出了一种新颖的、基于机器学习的肽设计架构，称为潜在空间近似轨迹收集器（LSATC）。它由一系列采样器组成，这些采样器沿着高度非凸能量景观上的优化轨迹运行，从而在潜在空间中近似具有目标属性的肽的分布。该过程几乎无需人工干预，并且可以端到端地实现。我们通过设计靶向

β-连环蛋白的肽延伸展示了该模型的效果。β-连环蛋白是一种参与经典 Wnt 信号传导的关键核效应蛋白。与随机采样器相比，LSATC 能够采样得到结合评分低**36%** 的肽，同时其四分位距（IQR）小**16** 倍；此外，这些肽的疏水性低**284%** ，且四分位距小**1.4** 倍。LSATC 的表现也远胜于其他常见生成模型。最后，我们使用聚类算法，从 LSATC 设计的 100 条肽中选出 4 条进行实验验证。结果证实，LSATC 延伸的 4 条肽与 β-连环蛋白的结合均有所改善，至少提高了**20.0%**；其中两条肽的结合亲和力较基础肽提高了**3** 倍。

关键词： 自动化蛋白质特异性肽设计，机器学习，进化优化

1 引言

治疗性肽是一类药物制剂，因其独特的生化和治疗特性而有别于小分子药物。近年来，许多肽类药物被发现具有优于小分子药物的效力和安全性特征 [1]。肽可以破坏有害的蛋白质-蛋白质相互作用（PPI），而这类相互作用通常被认为与癌症的发生和进展有关。其中一个例子是许多人类癌症中常见的 Wnt/β-连环蛋白/T 细胞因子（Tcf）信号通路失调。因此，开发基于肽的蛋白质-蛋白质相互作用抑制剂已成为癌症药物研究中最热门的方向之一。

针对特定蛋白质靶标的基于结构的肽抑制剂设计长期以来一直依赖经验。传统上，肽设计主要侧重于序列扰动，包括残基突变、链间残基交换、丙氨酸扫描和化学修饰 [2, 3], ，并以从蛋白质或蛋白质-肽复合物系统中获得的结构信息为指导。这种策略的主要局限在于，它忽视了序列扰动可能引起的二级结构变化，也未考虑这些变化可能如何改变蛋白质-肽结合结构。高通量筛选（HTS）也广泛应用于肽设计。HTS 是一种蛮力方法，通过快速开展数千至数百万次生化、遗传或药理学测定来识别生物活性肽。然而，HTS 需要高度专业化的仪器、适当生物测定方法的开发与采用，以及高质量的肽库。此外，HTS 发现的活性分子往往带有偶然性 [4]。作为 HTS 的补充，虚拟筛选（如肽对接）已被广泛用于肽抑制剂设计。 [5]。肽对接通常包括一种用于探索肽构象的采样技术，以及一种用于评估所有采样所得肽结合姿势中肽结合强度的评分函数。

两步方案评估一个蛋白质-肽复合物至少需要数分钟 [6], 这限制了其探索序列的能力，进而限制了其在从头肽设计中的应用。

近年来，基于机器学习的分子设计取得了快速发展。深度生成模型中的一个重要领域是分子的高效表示。在深度学习时代之前，小分子常用摩根指纹 [7] ，蛋白质则常用原子对指纹 [8]。然而，这些表示编码的是单个分子的化学和结构信息，并非针对特定任务设计 [9]。为了解决这一局限，人们开发了深度学习模型，先在大型数据集中学习统一的表示，再针对特定任务对该表示进行微调。过去几年间，分子表示研究逐渐转向字符串转换，其中大多数方法借鉴了自然语言处理的理念。模型框架主要分为两类。第一类是循环神经网络，例如长短期记忆网络（LSTM）[10], 、门控循环单元 [11] 和循环注意力 [12]。这些模型已用于预测分子性质，例如溶解度和毒性 [13–15]。另一类是基于 Transformer [16] 的模型，它们采用多头注意力机制，更高效地处理序列数据。Transformer 架构（如 ProteinBERT[17] 和 ProteinTrans [18] ）已广泛用于蛋白质表示，并在多种蛋白质下游分类或回归任务（例如二级结构分类、荧光预测）中取得了显著成效。近年来，分子的图表示受到广泛关注，因为图能够纳入更详细的结构和空间信息 [19]。图神经网络已用于处理分子图并进行性质预测 [20–22]。

尽管学习得到的分子表示已被用于许多性质预测模型，但涉及分子生成的模型却寥寥无几。2017年，Bjerrum 使用循环神经网络生成有效分子 [23]。此后，已有多项研究发表，探讨如何生成具有优化的一般性质（如脂水分配系数（logP）、拓扑极性表面积（TPSA）和药物相似性量化评估（QED））的有效分子 [24–26],。相比之下，蛋白质序列设计方面的研究较少，且大多聚焦于针对单一蛋白质的一般性质设计，例如长度、稳定性和等电子点[27, 28]。针对蛋白质相互作用的药物设计则鲜有探索。2020年，Das 提出了一种抗菌药物设计方法，通过拒绝采样在潜在空间中搜索合适的分子SMILES[29]。2022年，Castro 提出了一种基于梯度的潜在空间搜索方法，用于设计针对雷珠单抗重链第三互补决定区的蛋白质序列 [30]。这两篇论文与我们的工作最为相关，尽管第一项研究并非肽类药物设计，第二项研究的应用范围也仅限于一种已有包含60,000个样本的数据集的蛋白质。蛋白质特异性肽设计面临的主要挑战是，准确的结合亲和力数据不足，这源于高昂的

计算或实验成本以及庞大的肽空间。据我们所知，目前尚无用于蛋白质特异性肽生成的深度学习模型。

本文提出了一种新颖的蛋白质特异性肽生成方案（图 1），称为潜在空间近似轨迹收集器（LSATC）。我们实现了一个基于门控循环单元的 Wasserstein 自编码器（WAE） [31]，以获得肽的一维连续潜在空间表示，如图1(a) 所示。为解决数据稀缺问题，我们设计了一个反馈循环，使用 CMA-ES [32]在肽潜在空间中优化生成器并收集生成器的轨迹。该过程如图 1(c) 所示。快速反馈评估器对该方案至关重要。由于当前的结合评估器（例如分子对接）计算成本高昂，我们训练了一个高效的代理模型，将评估时间从几分钟缩短至微秒（图 1(b)）。训练代理模型时，分别使用 PyRosetta 和 Biopython 对数量适中的随机采样肽的结合能和疏水性进行评估。最后，我们从已探索的轨迹中采样相关损失值低于预设阈值的肽。值得注意的是，代理模型带来的速度提升使我们能够探索肽编码器空间，以进行肽抑制剂设计和优化。此外，在深度学习中引入基于生物物理学的模型，有助于在后处理阶段筛选肽候选物。

在此，我们在一项多目标肽延伸任务中测试 LSATC 模型。具体而言，我们旨在通过延长基础肽“YPEDILD- KHLQRVIL”的 N端来增强其与β-连环蛋白的结合，并降低肽的疏水性，以尽量减少非特异性结合。为简化起见，我们仅考虑延伸 5 个氨基酸的肽，因此搜索空间限制为 320 万种肽。不过，值得注意的是，LSATC 并不局限于生成固定长度的肽。整个肽延伸过程，包括数据集生成和模型训练，只需两天即可完成。事实证明，与随机生成和其他几种常用生成模型相比，我们的 LSATC 模型能更高效地生成所需肽。体外结果表明，我们生成的肽不仅疏水性更低，而且效力也高于实验基线结果，最高提升达 3 倍。

本文的贡献如下：

•我们设计了一种肽生成器，在不存在或仅有极少结合数据的情况下，也能高效地发现更多新颖的蛋白质特异性肽序列。

• 我们阐释了基于机器学习的肽生成器在数据稀缺的条件下如何工作以及为何能够发挥作用，从而提高了模型的可解释性。

• 我们提出的设计流程已可用于体外测试，且高度自动化。我们设计了一种肽筛选策略，从采样得到的高质量肽中选出所需数量的肽用于体外测试。这免去了实验人员手动检查并最终确定测试肽的需求。

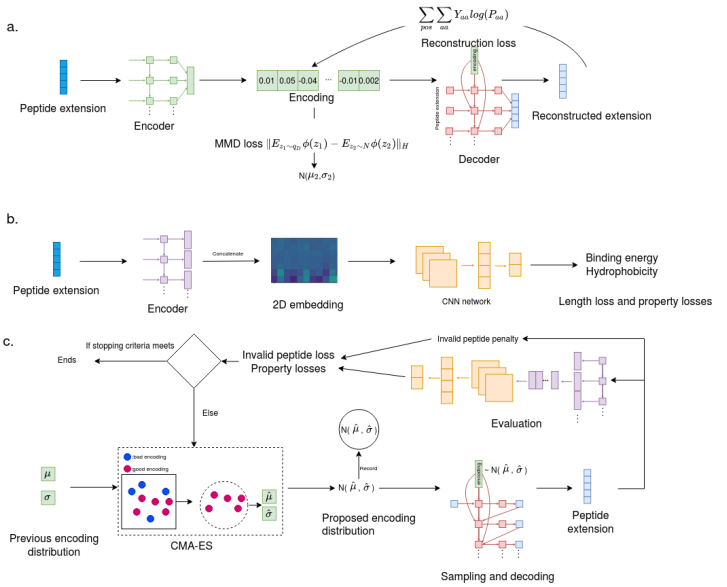

图 1 (a). 肽延伸重建网络。使用重构损失。编码的分布经正则化后符合高斯分布。(b). 性质预测网络。该模型估算肽延伸的疏水性和结合能。(c). 生成模型的优化过程。使用 CMA-ES 在潜在空间中学习高斯采样器。解码时使用重建网络。评估时使用性质预测网络作为 PyRosetta 的代理模型。每次迭代都会记录采样器。总损失包括目标肽性质，以及对采样到无效编码的惩罚。

2 结果

在计算机模拟评估中，与随机生成、高斯混合模型（GMM）和条件 Wasserstein 自编码器（CWAE）相比，LSATC 能更高效地生成结合能低且疏水性低的肽延伸。在体外测试中，所有采样得到的肽延伸均显著提高了基础肽的结合评分。本节将介绍 LSATC 各组成部分的分析和结果。

2.1 数据集准备

在 LSATC 中，序列重构模型用于将氨基酸表示从离散字母转换为连续数值。属性代理模型用于快速评估肽结合能和疏水性。一个 50万条未标注肽延伸数据集和一个 50,000 条有标注肽延伸数据集分别用于训练序列重构模型和代理模型。对于这两个数据集，所有肽延伸均不重复且为随机生成。

带标签数据集中的每条肽都有两种属性——疏水性和 β-连环蛋白结合评分。疏水性用于表示水排斥非极性分子的倾向，是一种序列依赖属性。我们使用 Biopython 计算疏水性。Kyte-Doolittle [33]标度用于测量每种氨基酸的疏水性程度。我们将窗口大小设为 3，并通过在肽序列上滑动窗口计算移动平均值。疏水性通过对所有移动平均值求和得出。请注意，我们使用固定长度的肽。因此，疏水性无需按肽长度进行归一化。结合评分的计算依据是[34]。它是范德华力能、溶剂化能、残基-残基对势、氢键能、静电能以及侧链旋转异构体内能的加权和。能量的单位为 $\frac{kcal}{mol}$ 。用于界定相互作用原子的距离阈值为 $10\mathring{A}$。计算结合评分需要肽-蛋白质复合物的三维结构信息。此类信息通过对初始肽-蛋白质复合物结构进行突变替换获得，详细说明见第 4.1 节。我们测试了 5 种不同的结合能计算方法：MM/GBSA、Rosetta FlexPepDock、flex ddG、flex ddG(gam) 和 PyRosetta。在图2中，我们绘制了 10 条经测定肽的实验结合数据与五种方法估算结果之间的相关性。结果表明，PyRosetta 在较低能量区域的线性相关性最佳。MM/GBSA 和 Rosetta FlexPepDock 可对整个区域内的肽进行评分。因此，我们选择 PyRosetta 来计算 50,000 条肽的结合能。尽管 PyRosetta 是一个轻量级软件包，生成带标签数据集仍需约 12 小时。

为了正确训练代理模型，我们对结合能进行对数变换以实现归一化，从而减小离群值的影响。数据集生成过程的详细信息见第 4.2 节。

2.2 序列重构

氨基酸所表示的肽与其性质之间的映射并不平滑，因为氨基酸空间是离散的。这往往导致优化器无法找到更优的肽。为缓解这一问题，我们创建了一个序列重构模型，用连续数值表示肽序列，并将这些数值解码回原始氨基酸空间。为实现这一目标，我们采用门控循环单元（GRU）生成式自动编码器框架进行序列重构。训练阶段和推理阶段的序列重构过程有所不同。我们将前者称为直接重构，后者称为采样重构。图3(b) 展示了二者的区别。进行解码时，需要使用已知的编码和前一个位置上的氨基酸来输出当前氨基酸。对于直接重构，由于训练期间前一个氨基酸是已知的，因此始终会正确输入；然而，对于采样重构，则需要

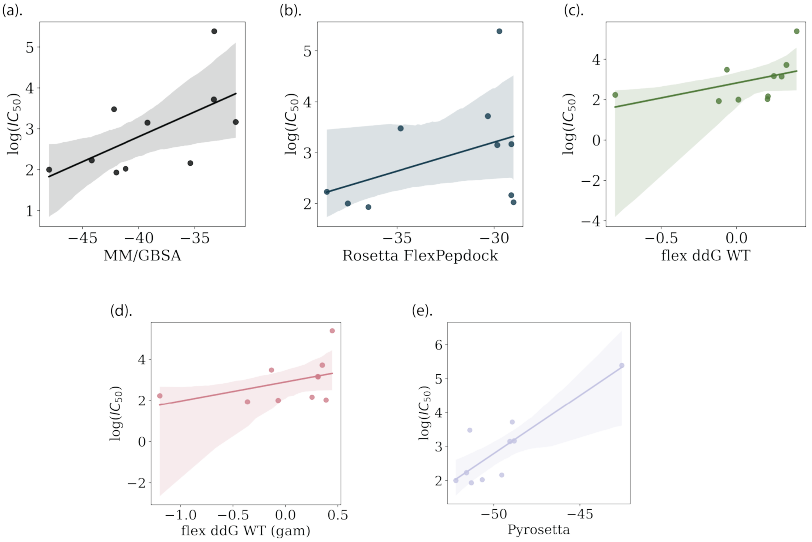

图 2 (a). 通过 MM/GBSA 计算的体外半数抑制浓度对数值与结合能的回归图。(b). 通过 Rosetta FlexPepDock 计算的体外半数抑制浓度对数值与结合能的回归图。(c). 通过 flex ddG 计算的体外半数抑制浓度对数值与结合能的回归图。(d). 通过 flex ddG(gam) 计算的体外半数抑制浓度对数值与结合能的回归图。e). 通过 PyRosetta 计算的体外半数抑制浓度对数值与结合能的回归图。

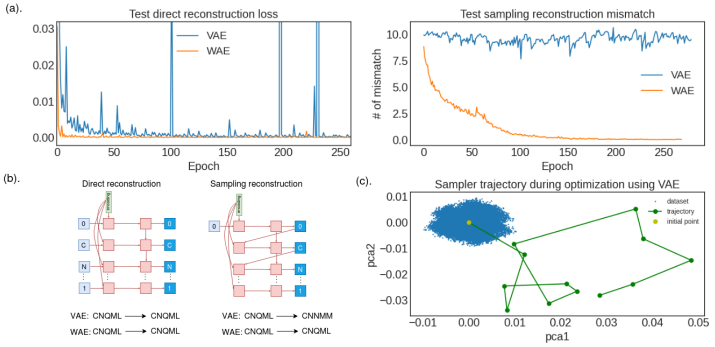

图 3 (a). 训练期间的直接重构损失（左）和采样重构损失（右）。(b). 直接重构与采样重构之间差异的示意图。图中展示了它们在变分自编码器和 Wasserstein自编码器上的两种重构结果。(c). 采样重构准确率较低的解码器的采样器优化轨迹示例。此处，采样器在与编码器所描述的潜在空间不同的潜在空间中进行优化，数据集代表了编码器描述的潜在空间。

将预测得到的前一个氨基酸作为输入，以推断当前氨基酸。这是因为在推理阶段无法获取真实的前一个氨基酸。采样重构对我们的模型至关重要。采样重构准确率较低的重构模型可能会导致我们在后续优化阶段出现编码偏移。图 3(c) 展示了使用一个有缺陷的编码采样器所得到的优化轨迹

模型。优化轨迹是指搜索良好编码过程中编码的轨迹，我们将在第 2.4 节中对此进行说明。由于编码器空间与解码器空间之间存在差异，需要从解码后的序列获取反馈的优化器不会在数据集所表示的编码器输出空间中或其附近进行搜索。

在本研究中，我们比较了常用的变分自编码器（VAE）和Wasserstein自编码器（WAE）的重构效果。VAE 使用 KL 散度将编码分布约束为目标分布，而 WAE 使用 Wasserstein 距离来实现这一目标。两种模型的详细信息见第 4.3 节。图 3(a) 右侧的图显示了训练期间直接重构的测试均方误差损失。我们发现，两种模型的测试直接重构损失都很低。右侧图中，我们将采样重构误差表示为输入与重构输出之间不匹配的数量。VAE 模型的采样重构误差一直很高，而 WAE 即使在两种模型的直接重构损失都很低的情况下，也能在 150 个训练周期后将误差降至接近 0。事实上，在推理阶段，VAE 模型倾向于将前一个输入重复作为输出。图 3(b) 给出了一个示例。根据上述实验，我们选择基于 GRU 的 Wasserstein 自编码器来编码和解码肽序列，因为该模型在直接重构和采样重构方面均表现良好。

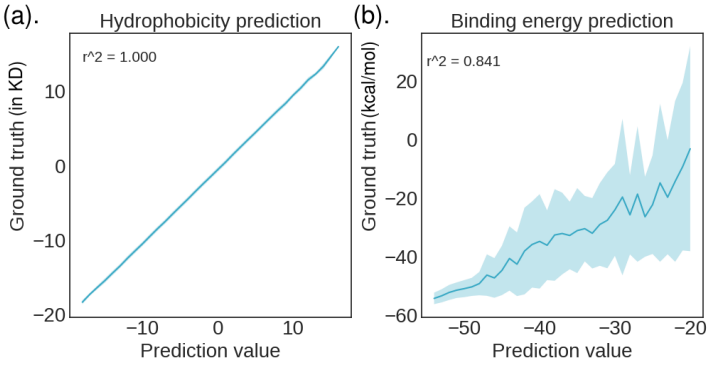

图 4 (a). 疏水性的 r^2 图。 r^2 值接近 1。(b). 结合评分预测图。 r^2 值是通过计算小区间内的预测均值与关联预测值落在该区间内的真实值均值之间的差值得出的。不确定性表示每个预测值对应的真实值范围。显然，预测值越小，不确定性也越小。

2.3 代理模型预测

准确评估肽-蛋白质结合自由能在计算上非常耗时，这在很大程度上是因为结合过程涉及显著的构象、平移和旋转变化，而这些变化难以采样。因此，要将其纳入迭代优化过程，

我们需要一种快速而准确的自由能评估方法。为此，我们在带标签数据集上训练了一个代理模型，用于预测肽与β-连环蛋白的结合评分以及肽的疏水性。

预测网络与重建网络相似，不同之处在于末端增加了一个卷积神经网络（CNN），用于预测结合能和疏水性。我们测试了几种不同的机器学习模型，其中卷积神经网络模型在预测疏水性和结合能时的测试均方误差最低。所有测试结果均见补充信息。

预测的疏水性与实验数据呈现出极佳的相关性（图4（a））。这表明，我们的代理模型能够准确捕捉肽的序列级性质。

由于代理模型仅依赖肽序列，因此结合评分的预测难度更大。它可以视为一种基于深度学习的定量构效关系（DL-QSAR）模型。为了兼顾速度和准确性，我们的目标并非获得与真实活性值高度相关的预测结果，而是对肽的结合能力进行排序，从而将重点放在结合评分较低的肽上。我们需要确保，当 PyRosetta 预测某条肽的结合评分较低时，代理模型也会预测该肽的结合评分较低。代理模型的预测值与 PyRosetta 计算值之间呈现相当好的相关性，r2 为 0.841。值得注意的是，如4（b）所示，结合评分越低，不确定性越小。这表明，如果代理模型预测某条肽的评分较低，那么它的真实评分也更有可能较低，这非常符合我们的优化需求。

2.4 CMA-ES 优化结果

CMA-ES 旨在通过无梯度方法优化高斯分布的均值和方差。我们使用高斯分布对优质编码所在区域进行参数化，并将其称为编码采样器。相对于编码的属性空间具有高度非凸性。因此，CMA-ES 是优化该采样器模型的良好优化器。具体而言，我们将采样器设计为各向同性高斯分布 $N(\mu, \sigma)$，它能够在潜在空间中采样得到优质肽，其中均值 μ的维度与编码的维度相同。 μ 和 σ 需要进行优化，以使采样得到的序列同时具有所需的结合能和疏水性。为此，我们实现了 CMA-ES。CMA-ES 的损失函数由三个部分构成：结合能、疏水性以及对无效肽的惩罚项。施加该惩罚项有助于采样器迅速逃离无效富肽区域；我们将其中超过 80% 的采样得到的肽为无效肽的区域定义为无效富肽区域。损失函数设计的细节见第 4.4 节。图5(a) 展示了优化过程中损失值的变化。黑框圈出的高损失值表明采样器进入了无效富肽区域。当进入绿框圈出的有效

富肽区域时，如粉色箭头所示，采样器表现出生成低结合能和低疏水性肽的强烈倾向。图 5(d) 的右图展示了优化过程中 μ 的轨迹，以及带标签数据集的编码。 μ 似乎在带标签数据集的外围移动。这种行为平衡了新颖性与对带标签数据集的相似性。在图5(d) 的左图中，随着优化过程的推进，学习得到的 σ 收敛至较小的值。因此，采样器探索过程的这种行为可以解释如下：当采样器到达无效富肽区域时，它会迅速逃离该区域。然而，当采样器进入有效富肽区域时，它开始通过在 μ周围的小区域内搜索，对结合评分和疏水性进行局部优化。这两个过程在整个优化过程中交替进行。如图 5(c) 所示。集中于潜在空间中的小区域是合理的，因为从图6(a) 的左图可以看出，相对于编码的总评分具有高度非凸性。如果搜索区域较大，则很容易采样得到低质量肽。

图5(b) 显示了优化过程中结合评分和疏水性的变化轨迹。最低疏水性约为 -15（KD 标度），最低结合评分约为 -50 $\frac{kcal}{mol}$，这些数值均根据生成的未见过的肽计算得出。请注意，在带标签数据集中，结合能的范围为 -59 $\frac{kcal}{mol}$ 至 $4782\frac{kcal}{mol}$，标准差为 $220\frac{kcal}{mol}$。疏水性的范围为 -19 至 17（KD 标度），标准差为 6。假设代理模型能够很好地近似这两种性质，那么这两个最小值都接近带标签数据集中的最低值。因此，收集采样器在这些局部极小值处的信息至关重要。我们在第 2.5 节进一步解释采样器信息的收集。

在优化过程中，我们发现优化器倾向于将 μ 调整到某个特定方向。图 5(e) 左图显示， μ 的轨迹倾向于在第一和第二主成分上取较大的负值。图 5(e) 右图显示了优化过程中变化最大的前 5 个编码维度。显然，这些数值在优化过程中会沿着偏好的方向变化。对于任何重构模型而言，高维编码的可解释性都是一个长期存在的研究问题，这给调整编码以生成具有特定性质的对象带来了困难。本文提出的 CMA-ES 优化器提供了一种自动调整编码的方法，以完成特定性质的生成任务。

CMA-ES 作为一种无梯度方法，非常适合肽生成，原因有三点。首先，在序列重构模型中，编码经过正则化，使其服从高斯分布。这与 CMA-ES 的假设相符，因为采样器同样采用高斯分布。其次，CMA-ES 是一种无梯度方法。无法保证生成的编码都能解码为有效序列。因此，一个重要目标是确保生成的编码有效（例如，解码后的序列长度为

5）。关于这一目标相对于编码的梯度信息是不存在的。第三，编码对应的损失函数很可能高度非凸。可能存在许多局部极小值，而我们希望收集这些局部极小值中的信息。基于梯度的方法很容易陷入局部极小值，而无梯度方法则可以越过这些极小值之间的障碍。

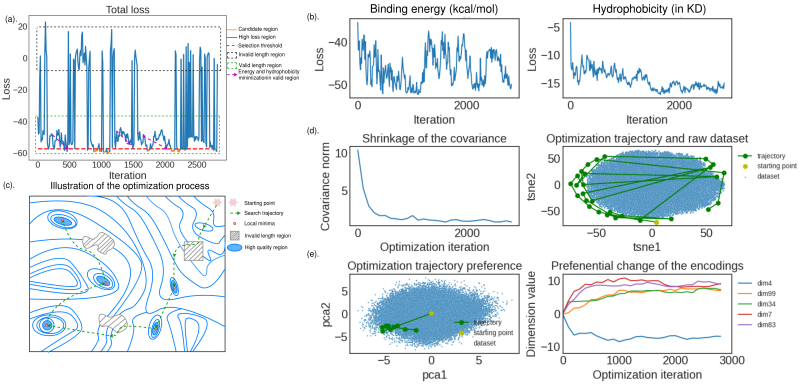

图 5 (a)。CMA-ES 训练损失。由无效肽惩罚项导致的高损失区域，以及代表有效肽丰富区域的低损失区域。在低损失区域，主要目标是降低属性损失。(b)。优化过程中的平均结合评分。（左）。优化过程中的平均疏水性（右）。(c)。优化过程中采样器行为的示意图。(d)。优化过程中的协方差幅度。该幅度通过计算每个元素绝对值之和得到（左）。用于展示采样器均值轨迹的二维 t-SNE 图。采样器正在带标签数据集的外围附近进行搜索（右）。(e)。用于展示编码变化优先方向的二维 PCA 图。显然，大多数搜索发生在带标签数据集边缘的左下角（左）。5 个编码维度的数值变化。它们分别被调整为正向或负向，表明调节器具有调节偏好（右）。

2.5 序列采样结果

如图所示 5(d)， σ 在优化过程中往往较小。在我们的实验中，我们发现，随着采样迭代次数增加，在单个 (μ， σ) 位置采样有效编码变得越来越困难。较小的 σ 可确保采样得到的肽的质量，但会导致有效编码耗尽。我们通过迭代收集有效编码，直到达到所需数量。在单个位置，随着所需肽数量增加，达到目标所需的采样迭代次数会迅速增加。这一点在图6(d) 中的最佳选择曲线上显而易见，该曲线对应于在单个位置进行采样。因此，从多个区域收集肽可以缓解肽耗竭问题，但选择合适的区域可能并不容易。例如，采样器可能连续多个迭代都停留在同一区域，并产生相似的属性损失。在这种情况下，所收集肽中损失函数最低的一些肽很可能来自同一区域，

这仍会面临编码耗尽问题。图6(e) 中的红点展示了这种情况。为了解决这一问题，我们收集损失值排名前 500 的采样器轨迹（μ，σ），并将其称为候选轨迹。这些区域在图6(e) 中以橙色显示，其对应的损失值在图5(a) 中也以橙色显示。为获得彼此距离较远的采样器，我们采用 K均值算法对 500 个采样器进行聚类，并从各个簇中选取代表性采样器。所选采样器在图6(e) 中以绿点表示，显然比分布为红点的采样器彼此间隔更远。

我们提出的 LSATC 是一组高质量采样器，每个采样器都对应于局部极小值中的一个遥远小区域。为考察这一策略是否有效，我们将其与直接对已知高质量编码建模的算法进行比较。由于疏水肽往往会以非特异性方式与蛋白质结合，我们定义了一个总评分，用于统一疏水性和结合评分。我们分别对肽的结合能和疏水性进行排序，然后将二者相加，计算总评分。评分越低，质量越好。为选择用于训练的肽，我们将带标签数据集中的肽分为 5 类。标签为 0 的肽总评分低于 10,000。这些肽的质量最高，在 50,000 条肽中占 456 条。我们在该数据集上实现了两个模型。第一个是高斯混合模型（GMM），用于对标签为 0 的肽的编码分布建模。GMM 所近似分布的粒度取决于簇的数量。簇越多，分布越精细。然而，随着簇数量增加，模型更容易出现有效编码耗尽的问题。这是因为数据集中只有 456 个编码，因此每个分布模式都可能非常集中。我们在补充信息中展示了 200 个簇时效果最佳。第二个模型是条件WAE模型，它通过将标签拼接到编码中，手动分离每个类别的编码。该模型分别对每个带标签编码的分布进行建模。将标签 0 与正态采样的编码拼接后，作为解码器的输入，即可直接采样得到高质量肽。算法的详细信息见补充信息。在采样阶段，模型只从目标类别区域采样编码。如图6(a) 所示，LSATC 采样得到的编码集中在几个偏好的区域，而 GMM 采样得到的编码则分散在低评分区域；条件WAE采样得到的编码则位于其自身的簇空间中。

我们以随机采样（即我们的带标签数据集）作为基线，比较这些模型的采样质量。对于总评分，我们将每条肽的结合评分和疏水性与带标签数据集中的相应指标进行排名。将这两个排名相加，得到总评分。评分越低，模型表现越好。随机肽的总体得分可视为背景排名。对于每个机器学习模型，我们采样了 100 条序列。在图6(b) 最右侧的图中，

从所有机器学习模型中采样得到的肽，其总评分分布相较于随机模型明显向较低分数一侧偏移。LSATC的中位数和四分位距分别为7975和12436，分别比随机模型小6.3倍和2.3倍。LSATC的总评分中位数分别比高斯混合模型和CWAE小2.8倍和4.2倍，其四分位距比高斯混合模型小1.3倍，比CWAE小2.1倍。如图6(b)所示，LSATC的中位数最低、四分位距最窄，说明它是这些模型中效率最高的采样器。与随机模型相比，LSATC采样得到的肽结合评分低36%，四分位距缩小至原来的十六分之一；疏水性低284%，四分位距缩小至原来的.4倍。此外，高斯混合模型采样得到的肽，其结合评分和疏水性的中位数也较低、四分位距也较小，这表明潜在空间中已知高质量肽的邻近区域也存在高质量肽。CWAE模型的表现是三个模型中最差的，不过仍优于随机模型。CWAE的疏水性中位数低于随机模型，但四分位距与随机模型几乎相同，这说明在图6(a)所示的标签0区域中采样，并不能保证疏水性较低。这一观察结果是合理的，因为高质量肽（标签0）非常稀少。对高质量区域的分布进行建模十分困难，因为这种分布函数应包含许多局部模态，而相较于整个区域，这些模态的数量很少。对LSATC、高斯混合模型和条件WAE的表现进行比较后发现，采样高质量肽最高效的方法是在每个高质量但范围较小的区域内进行采样。

在图6(c) 中，我们展示了 LSATC 模型采样得到的肽。延长氨基酸残基通常会因非特异性相互作用增加而提升结合能力，但我们的模型似乎会生成肽序列以最大化特异性结合。如序列标志图所示，I、R、E、Y、K 是每个位置上生成频率最高的氨基酸，其中大多数是能够形成特异性极性/离子相互作用的带电/极性残基。结果表明，我们的模型能够学习结合环境，进而生成相应的残基。这也佐证了在损失函数中惩罚疏水性有助于模型减少非特异性相互作用。完整的样本表见补充信息。

2.6 实验结果

从图6（b）可以明显看出，我们通过 LSATC 生成的肽抑制剂在结合强度和疏水性方面存在差异。因此，除了选取得分最高的肽之外，我们还倾向于根据肽的多样性对其进行聚类。该策略首先从 100 个 LSATC 采样的延伸序列中选取得分最高的 40 个延伸序列。其次，使用 GibbsCluster 从这些优选延伸序列所属的 4 个不同簇中进一步选取 4 个代表性延伸序列 [35]。具体流程见第4.6节。图7展示了四个簇各自的对数优势（LO）矩阵，这些簇分别由

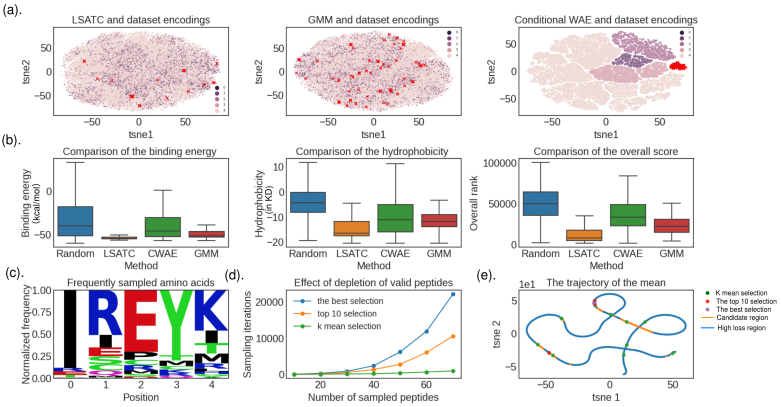

图 6 (a). 二维 t-SNE 图，展示不同生成模型生成的样本在训练用高质量肽中的分布位置。模型分别为 LSATC（左）、高斯混合模型（中）和 CWAE（右）。(b). 对比随机选择、LSATC、CWAE 和高斯混合模型采样所得的肽质量。请注意，需要使用总评分来描述肽，因为结合图和疏水性图会损失肽特异性。(c). 序列标志图，展示每个延伸位置采样最多的氨基酸。(d). 有效编码的耗减。每次迭代的采样批次固定。显然，除 K 均值选择外，所需迭代次数会随着请求肽数量的增加而迅速增加。(e). 采样器的优化轨迹，以及使用不同选择策略最终选出的采样器。

Seq2Log [36] 序列标志图。20 种天然氨基酸根据其在序列标志图相应位置上的对数优势评分以不同大小显示。该评分可以视为该簇中该位置出现相应氨基酸的概率。从簇 1 到簇 4，分别包含 16、2、11 和 11 条序列，相应的簇代表序列为 IREYK、IREFK、IRCCK、ICEYK 和 EREYK。值得注意的是，序列标志图的第一行表示该簇内最有可能出现的理论氨基酸，但这些氨基酸并不总是存在于我们的采样池中。然而，除第一个之外，其余所有氨基酸都可以在我们通过 LSATC 生成的肽抑制剂列表中找到。完整的 40 条肽收录于补充信息中。总体而言，我们整合 GibbsCluster 肽聚类方法，以更系统地选择用于结合测定的测试肽。

我们将实验结果与母肽进行比较。请注意，实验中实际使用的基础肽是”GGYPEDILDKHLQRVIL”。不过，在计算设计中去掉”GG”是一种很好的做法，因为甘氨酸通常会形成环状二级结构，这会给我们的结合能评估带来灵活性和不确定性。在补充信息 1中，基础肽是将原始基础肽中的 GG 截去后的版本。显然，LSATC 生成的所有肽都大幅提高了结合亲和力，其中表现最佳的肽“ICEYKYPEDILDKHLQRVIL”的结合亲和力提高了 3 倍以上。

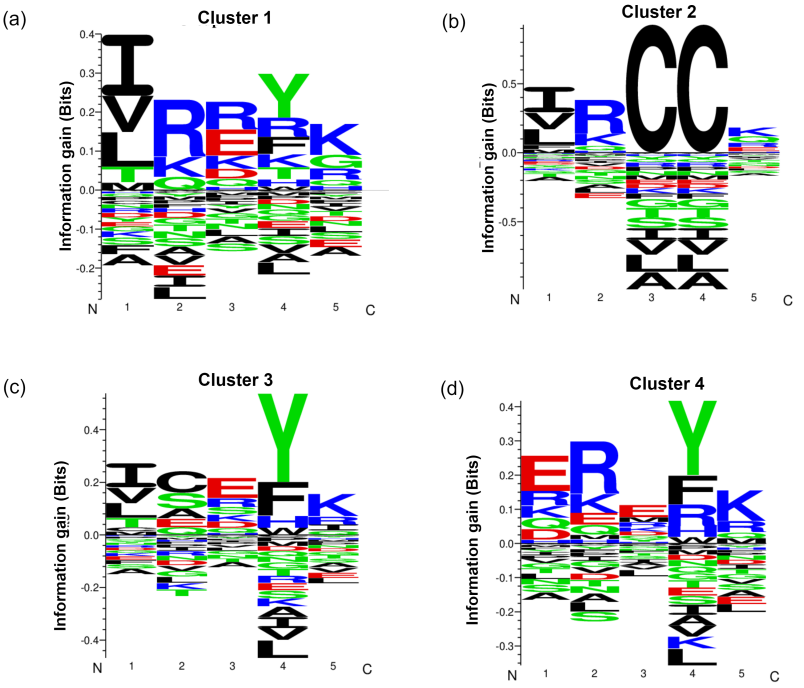

图7 四个簇的对数优势矩阵序列标志图。横轴表示氨基酸位置，在本例中总长度为 5。纵轴表示该位置氨基酸相对于其自身背景频率的信息增益。(a)、(b)、(c)、(d) 分别对应簇 1 至簇 4 的对数优势矩阵序列标志图。在序列标志图中，每列顶部的氨基酸表示该簇中该位置最有可能出现的氨基酸。

表 1 受测肽的实验结果。

来源	肽序列	体外 IC_{50} (nM)
LSATC-1	IREYKYPEDILDKHLQRVIL	51.8 ± 8.2
LSATC-2	IRCCKYPEDILDKHLQRVIL	120.3 ± 2.6
LSATC-3	ICEYKYPEDILDKHLQRVIL	47.8 ± 1.9
LSATC-4	EREYKYPEDILDKHLQRVIL	68.4 ± 10.2
基础肽	GGYPEDILDKHLQRVIL	150 ± 20

3 讨论

在本研究中，我们提出了 LSATC，这是一种自动化的蛋白质特异性肽设计方法，通过收集采样器在肽的潜在空间中的搜索轨迹来实现。该方法适用于广泛的用户，因为我们只需要肽-蛋白质复合物的初始构象。

在 LSATC 中，CMA-ES 被用作自动编码调优器，引导高斯采样器在广阔的潜在空间中迭代探索具有低疏水性和低结合能的有效肽区域。作为一种无梯度方法，CMA-ES 非常适合此类肽生成任务。原因有三点。首先，编码被正则化为高斯分布

在序列重构模型中。这与 CMA-ES 的假设一致，因为采样器同样服从高斯分布。其次，CMA-ES 是一种无梯度方法。无法保证任何生成的编码都能被解码为有效序列。因此，一个重要目标是确保生成的编码有效（例如，解码后的序列长度为 5）。关于这一目标，相对于编码不存在梯度信息。第三，相对于编码的损失函数很可能具有高度非凸性。可能存在许多局部极小值，而我们希望收集这些局部极小值中的信息。基于梯度的方法很容易陷入局部极小值，而无梯度方法则可以越过这些极小值之间的障碍。为降低计算成本，我们训练了一个代理模型来评估结合性质和疏水性质，从而使迭代搜索在计算上可行。我们已经表明，LSATC采样器能够学习一系列独立的高斯分布，以近似肽潜在空间中每个局部极小值处的一个小区域，从而有效避免编码耗竭问题。我们通过肽延长任务表明，LSATC 生成 b-catenin特异性肽的效率远高于随机选择，以及其他在 *in silico* 评估中对高质量肽的编码分布进行全局拟合的机器学习模型。

此外，我们提出了一种策略，用于从 LSATC 提出的肽中选取数量合理的代表性肽，以进行最终的体外 测试。在最终体外 测试中，所选肽的结合亲和力优于基础肽，也优于通过文库筛选方法获得的所有肽延伸，从而证明了我们的 LSATC采样器具有实际应用价值。

如第2.1节所述，LSATC 方法需要约 50,000 条针对靶蛋白的模拟结合数据。因此，肽采样的质量实质上取决于突变模型在估算肽-蛋白质复合物三维结构时的质量。对于短肽生成，突变模型准确度较高，因此可以完成从头设计任务，即生成完整的肽，而不是肽延伸。然而，对于长肽，突变模型的准确度会下降，在从头肽设计中使用 LSATC 在实践中可能会产生质量较差的肽。如今，大分子药物发现领域的深度学习技术发展十分活跃。有些研究将肽配体和蛋白质受体的序列输入深度学习模型，以进行二元分类，判断它们能否结合。未来不久，深度学习模型或许能够准确预测肽与蛋白质之间连续的结合评分。届时，LSATC 方法将不再需要突变模型，也就能够利用 LSATC 高质量地完成任意长度肽的从头设计。

4 方法

4.1 肽结合能计算

我们使用与订书肽抑制剂结合的 β-连环蛋白晶体结构（PDB:4DJS）作为肽结合能计算的起始结构。原始肽序列 YPEDILDKHLQRVIL 的两端延伸了 5 个丙氨酸，并使其形成 α螺旋结构。随后，将延伸后的肽叠合到共晶结构中的订书肽上。多聚丙氨酸延伸部分随机突变为 20 种天然氨基酸中的任意一种，并使用 Modeller 对生成的结构进行能量最小化。[37]

为了为我们的优化器寻找稳健的评分函数，我们测试了五种结合能计算工具，包括分子力学广义Born表面积（MM/GBSA）、Rosetta FlexPepDock、flexddG、flexddGgam 和 PyRosetta，以确定其中预测的结合能与体外IC50值相关性最佳的工具。 [6, 37–40]（图5）MM/GBSA 是一种常用的结合自由能估算方法，在该方法中，肽-蛋白质复合物首先进行 10 ns 的分子动力学模拟，然后使用 GMX MMPBSA 近似计算结合能。Rosetta FlexPepDock 是另一种常用的肽-蛋白质相互作用能估算方法。首先使用蒙特卡罗方法对肽-蛋白质复合物的结构进行采样，然后使用 Rosetta 评分函数计算每个采样构象的界面能。报告 50 个构象的均值。PyRosetta 是 Rosetta 的一个经过修改的 Python 版本。首先对肽-蛋白质复合物结构进行最小化，然后计算界面能。除三种最知名的结合能计算方法外，我们还测试了 Flex ddG流程。该流程用于预测肽-蛋白质界面处肽发生突变时的结合自由能变化（界面ddG）。此外，还可以通过广义加和模型（GAM）方法改进界面ddG，使其与实验测定的界面ddG相关。首先使用 backrub 对肽-蛋白质复合物结构进行构象采样，随后进行扭转角最小化和侧链重新堆积。将所得构象集合的界面ddG均值作为结合能进行报告。结果表明，PyRosetta 计算的结合能与实验IC50值之间具有最佳相关性，因此选择 PyRosetta 对机器学习模型生成的肽序列进行后续评估。

4.2 数据集准备

准备了两个数据集：一个是用于重构模型训练的无标签数据集，另一个是用于代理模型训练的带标签数据集。无标签数据集包含 500,000 个随机采样的长度为 5 的肽。请注意，总共有 3,500,000 种组合。我们使用其中 80% 来训练变分自编码器，20% 用于测试。带标签数据集包含 50,000 个随机

采样得到的肽延伸。这些延伸与基础肽拼接在一起。使用 Biopython 和 PyRosetta 评估延伸肽的疏水性和结合能。在一台配备 16 核中央处理器的计算机上使用多进程评估其结合能，大约需要 12 小时。

结合评分分布高度偏斜。由于肽与蛋白质之间存在空间位阻冲突，因此会出现许多高值。我们首先剔除离群值，将数值限制在 0.1 至 0.9 分位数范围内。然后使用 $\log(x+100)$ 将结合评分归一化到较小的范围。请注意，若不进行这一步数据预处理，代理模型就无法在任何范围内做出合理预测。对于疏水性，其数值服从正态分布。因此，我们只需将数值标准化为单位高斯分布。预处理前后的分布变化见补充信息图 xx。

4.3 序列重构模型和代理模型

序列重构模型由嵌入网络、门控循环单元和多层感知机构成。嵌入网络将单个氨基酸的离散表示转换为连续表示，并将它们串联起来以表示肽序列。随后，GRU网络和MLP网络处理这一有序表示，并输出肽编码的均值和标准差。然后，将编码的均值输入另一个GRU网络，再经过一个MLP网络来重构肽。

公式1 展示了变分自编码器的损失函数。第一项是重构损失，旨在使 x_o 由编码 z 解码得到的输出与x_i （编码得到 z 的输入）相匹配。第二项是编码正则项，它使用 KL 散度强制编码服从正态分布。在 Wasserstein自编码器中，如公式2所示，KL 散度相对于 p(z) 被最大均值差异（MMD）度量所取代。

$$\text{vae loss} = \underbrace{-E_{q_d(z|x_i)} \log p(x_o|z)}_{\text{reconstruction loss}} + \underbrace{D_{KL}(q_D(z|x_i)\|p(z))}_{\text{encoding regularizer}} \tag{1}$$

$$\text{wae loss} = \underbrace{-E_{q_d(z|x_i)} \log p(x_o|z)}_{\text{reconstruction loss}} + \underbrace{MMD(q_D(z|x_i), p(z))}_{\text{encoding regularizer}} \tag{2}$$

MMD 的定义见公式3。它是两个分布之间差异的度量 q_D 和 p 通过某个函数将其值转换到希尔伯特空间之后 ϕ。

$$\begin{aligned}\text{MMD}^2(q_D, p) &= \|E_{z_1\sim q_D}[\phi(z_1)] - E_{z_2\sim p}[\phi(z_2)]\|_H \\ &= E_{z_1\sim q_D} E_{z_1'\sim q_D}\langle[\phi(z_1)], [\phi(z_1')]\rangle - 2E_{z_1\sim q_D} E_{z_2\sim p}\langle[\phi(z_1], [\phi(z_2]\rangle \\ &+ E_{z_2\sim p} E_{z_2'\sim p}\langle[\phi(z_2)], [\phi(z_2')]\rangle\end{aligned} \tag{3}$$

在公式3中，MMD^2最终形式中的自变量由有效内积 $\langle \cdot, \cdot \rangle$ 完全定义，这种内积称为核函数。一旦知道核的形式，就无需定义变换函数 ϕ 。在本研究中，我们使用高斯核对学习得到的编码和正态采样的 z<style id='17}>进行变换，以计算两个分布之间的最大均值差异损失。此外，我们直接对编码的学习均值和方差施加较弱的 KL 惩罚项，使其服从单位高斯分布。对于重构损失，由于输入和输出均为离散值，我们采用交叉熵（CE）损失。最终损失如公式4所示。

$$loss = \underbrace{CE(x_i,\ x_o)}_{\text{重构损失}} + \underbrace{MMD(q_D(z|x)\ N(0,\ 1)) + 10^{-3} * D_{KL}(q_D(z|x)\|N(\mu_z,\ 1))}_{\text{编码正则项}} \tag{4}$$

代理模型的结构与序列重构模型相似。不同于执行重构任务，解码 GRU 网络的输出被导入 7×100 图像中，以使用 CNN 模型预测疏水性和结合能。最终损失是这两个属性的均方误差（MSE）损失之和，如公式5所示，其中星号表示真实值。在实践中，我们测试了不同的代理模型结构，并因 CNN 模型表现最佳而选用了它。在补充信息中，我们展示了所有已测试结构的训练结果。

$$\text{surrogate loss} = MSE(BE^*, BE) + MSE(Hydro^*, Hydro) \tag{5}$$

4.4 CMA-ES 优化设计

CMA-ES 是一种无梯度优化方法。它假设一个多元高斯采样器。在每一步中，该算法都会收集一组样本。采样器均值根据公式6进行更新。在该公式中，μ^i 是第 i^{th} 次迭代时的采样器均值。b 是学习率。x_n^i 是第 n^{th} 次迭代中采样的点。i^{th} 是第 w_n 的权重。x_n。$\sum w_n$ 等于 1。k 是通过适应度函数衡量的优选样本数量。我们根据选定的 k 个点和前一次均值更新采样器的均值。b 通常设为 1。在这种情况下，我们根据 k 个选定点的加权组合更新均值。

$$\mu^{i+1} = \mu^i + b \sum_{n=1}^{k} w_n (x_n^{i+1} - \mu^i) \tag{6}$$

协方差矩阵更新基于之前的协方差，以实现更准确的估计。如公式 xx 所示，它由两部分组成。第一项将当前估计的协方差与加权后的先前

协方差结合起来。第二项则是在第一项基础上的改进。

$$S_1^{i+1} = (1-b_2)S^i + b_2 \underbrace{(\sum_{n=1}^{k} w_n (\frac{x_n^{i+1}-\mu^i}{\sigma^i})(\frac{x_n^{i+1}-\mu^i}{\sigma^i})^T)}_{\text{current covariance estimation}} \tag{7}$$

正如第一项所示，$\frac{x_n^{i+1}-\mu^i}{\sigma^i}(\frac{x_n^{i+1}-\mu^i}{\sigma^i})^T$ 在更新过程中丢失了符号信息。为强化这一信息，协方差通过另一条进化路径进行更新，该路径由前几步的均值构建，而非单个样本。此路径如式8

$$y^{i+1} = (1-b_3)y^i + \sqrt{b_3(2-b_3)w_{eff}}\frac{\mu^{i+1}-\mu^i}{\sigma^i} \tag{8}$$

$$S_2^{i+1} = (1-b_4)S^i + b_4 \underbrace{y^{i+1}(y^{i+1})^T}_{\substack{\text{estimation from} \\ \text{evolution path}}} \tag{9}$$

式9 展示了基于进化路径的协方差更新规则。研究表明，这种更新能更好地耦合两个优化步骤[41]。两次协方差更新相加，构成式10所示的最终更新规则。

$$S^{i+1} = S_1^{i+1} + S_2^{i+1} \tag{10}$$

尽管上述方程中有许多超参数，但大多数参数都有最优取值 [41]。CAM-ES 中唯一需要调整的参数是初始采样规模。初始均值和协方差对结果影响不大[42]。虽然 CMA-ES 方法简单，但它在非光滑、非连续函数甚至含噪数据集上都表现良好，是一种可靠的局部优化方法[43]。

在本研究中，CMA-ES 的适应度函数由疏水性、结合能以及对采样得到的无效肽的惩罚项之和构成。优化过程中，将从当前采样器中采样 N 条肽。然而，由于编码空间很大，无法保证解码得到的肽序列包含有效词元。有时，肽长度并非 5。极少数情况下，填充词元会出现在结束词元之前，这违反了词元规则。一旦出现这些情况，生成的编码就会被视为无效。考虑到 CMA-ES 是一种基于种群的优化方法，我们希望确保在每个优化步骤中，生成的样本大多数都是有效的。因此，我们在每个步骤中设计了另一个生成循环，反复采样 1000 条肽。只有当 80% 的样本有效，或循环迭代 20 次时，循环才会退出。

达到 20 次迭代。式11 展示了每个采样编码的适应度函数 $f(z_i)$ z_i。

$$f(z_i) = w(z)(NN_b(z_i) + 0.5 * NN_h(z_i)) + w(z)I(z_i) \tag{11}$$

其中，
NN_b和 NN_h分别是结合能和疏水性的代理模型输出，

如果有效采样编码少于 80%，则 $w(z)$ 为 1；否则，该值为 0。

$I(z_i)$= 0.1*（无效编码的数量）。

在公式11中，我们将疏水性的权重设为 0.5，以优先最小化结合能。在我们的实现中，如果无效编码占样本的比例低于 20%，则会将无效编码从种群中移除，使优化器专注于有效编码。优化设计的详细示意图见补充信息第 xx 节。

4.5 采样器选择

由于编码空间中的优化具有高度非凸性，且编码会集中在单一位置，因此有必要收集优化轨迹。为避免收集相似的采样器，我们根据 CMA-ES 的损失函数对其排序，并选取排名前 500 的采样器。然后使用包含 10 个簇的 K均值算法对这 500 个采样器进行聚类，并选取距离聚类中心最近的 10 个采样器。K均值算法是一种聚类算法，旨在最小化各个簇内样本之间的距离。其目标函数如公式4.5所示。这里我们使用欧几里得距离作为距离度量。在补充信息中，我们展示了采样器簇的二维 t-SNE 图。可以明显看出，选出的采样器彼此相距较远。

$$loss_{k-mean} = \sum_{n=1}^{N}\sum_{k=1}^{K} I(c_i == k) f(z_i, \mu_k) \tag{12}$$

其中，
N 和 K 分别是采样器的总数，而 K $=10$ 是簇的数量。

I(.) 是示性函数。
z_i 是采样器均值。 c_i 是 z_i的簇。μ_k 是 k^{th}簇的均值。

4.6 实验的肽选择

我们首先使用 LSATC 采样 100 条肽，并根据总体得分对肽进行排序，选出排名前 40 的肽。接下来，我们使用 GibbsCluster[44] 对这些肽延伸序列进行聚类。GibsCluster 是一种无监督肽模式发现算法，可同时对肽数据进行采样、聚类和比对

。该算法会学习一个 $\times \times i^{th}$ 的对数优势矩阵，其中 j^{th} 行 i^{th} 列的元素表示在 j^{th} 位置出现的 i^{th}氨基酸相较于氨基酸背景信息的信息增益。m 表示 20 种天然氨基酸。n 为 5，因为我们只额外延伸 5 个残基。该信息增益与某个簇中某个位置出现某种氨基酸的概率成正比。

该簇中某条肽的代表性评分，是通过将其氨基酸在该簇的对数优势矩阵中的信息增益相加计算得出的。信息增益本质上是对数概率评分。因此，计算总和类似于计算该肽在该簇中出现的未归一化概率评分。

4.7 肽合成

使用 Fmoc 化学，在 Rink 酰胺树脂上通过固相肽合成手工合成肽。通过基于荧光偏振（FP）的竞争测定法，测量预测肽对β–连环蛋白的体外半数抑制浓度。按照此前报道的方法，将 FAM 标记的探针肽（10 nM）与 50 nM 谷胱甘肽 S-转移酶–β–连环蛋白在含有 20 mM Tris、300 mM 氯化钠、pH 8.8 和 0.01% Triton-X100 的溶液中孵育 1 h。将竞争肽进行系列稀释，稀释液为含有 20 mM Tris、300 mM 氯化钠、pH 8.8 和 0.01% Triton-X100 的溶液。1 h 后，将平衡后的探针肽 –β–连环蛋白溶液分装加入系列稀释的肽溶液中，并在室温下孵育 1 h。将样品转移至黑底黑壁的 384 孔非结合微孔板（Greiner）中，并使用 Tecan M1000 Infinite 酶标仪测量 FP。使用 GraphPad Prism v. 8.0 分析数据，并将其归一化至探针完全结合/未结合时对应的 FP 值。

参考文献

[1]Lau, J.L., Dunn, M.K.：治疗性肽：历史视角、当前发展趋势与未来方向。Bioorganic andMedic- inal Chemistry **26**（10），2700–2707（2018年）。https://doi.org/10.1016/j.bmc.2017.06.052。肽类疗法[2] Xiao, X., Wang, Y., Seroski, D.T., Wong, K.M., Liu, R., Paravastu, A.K., Hudalla, G.A., Hall, C.K.：从头设计可共同组装成 β 片层基纳米纤维的肽。Sci Adv **7**（36），7668（2021年）[3] Eustache, S., Leprince, J., Tuff´ery, P.：利用肽扫描研究构效关系的进展：对药物发现的影响。Expert Opin Drug Discov **11**（8），771–784（2016年）[4] Lyu, W., Deng, Z., Sunkara, L.T., Becker, S., Robinson, K., Matts, R., Zhang, G.：针对天然宿主防御肽诱导化合物开展高通量筛选，作为抗生素的新型替代物。Frontiers in

《细胞与感染微生物学》8（2018年）。https://doi.org/10.3389/fcimb.2018.00191[5] Valiente, P.A., Wen, H., Nim, S., Lee, J., Kim, H.J., Kim, J., Perez-Riba, A., Paudel, Y.P., Hwang, I., Kim, K.-D., Kim, S., Kim, P.M.：SARS-CoV-2 强效 D-肽抑制剂的计算设计。《药物化学杂志》64(20), 14955–14967（2021年）[6] Raveh, B., London, N., Zimmerman, L., Schueler-Furman, O.：Rosetta FlexPepDock ab-initio：肽在其受体上的同步折叠、分子对接与优化。《公共科学图书馆·综合》6(4), 18934（2011年）[7] Morgan, H.L.：为化学结构生成独特机器描述——化学文摘社开发的一项技术。《化学文献》5(2), 107–113（1965年）。https://doi.org/10. 1021/c160017a018[8] Carhart, R.E., Smith, D.H., Venkataraghavan, R.：作为构效关系研究中分子特征的原子对：定义与应用。《化学信息与计算机科学杂志》25(2), 64–73（1985年）。https://doi.org/10.1021/ci00046a002[9]Capecchi, A., Probst, D., Reymond, J.-L.：一种统摄一切的分子指纹：药物、生物分子和代谢组。《化学信息学杂志》12(1), 43（2020年）。https://doi.org/10.1186/s13321-020-00445-4[10] Hochreiter, S., Schmidhuber, J.：长短期记忆网络 9(8) （ 1997 年）。https://doi.org/10.1162/neco.1997.9.8.1735[11] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.：门控循环神经网络在序列建模中的实证评估。载于：NIPS 2014 深度学习研讨会，2014年12月（2014年）[12] Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.：视觉注意力的循环模型。NIPS'14。麻省理工学院出版社，美国马萨诸塞州剑桥（2014年）[13] Feinberg, E.N., Sur, D., Wu, Z., Husic, B.E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B.,Pande, V.S.：用于分子性质预测的 Potentialnet。《ACS 中央科学》4(11), 1520–1530（2018年）。https://doi.org/10.1021/acscentsci.8b00507[14] Li, C., Feng, J., Liu, S., Yao, J.：一种结合多种基于 SMILES 的数据增强方法、用于分子性质预测的新型分子表示学习方法。《计算智能与神经科学》2022, 8464452（2022年）[15] Paul, A., Jha, D., Al-Bahrani, R., Liao, W.-k., Choudhary, A.N., Agrawal,

A.：Chemixnet：采用多种分子表示的混合深度神经网络架构，用于预测化学性质。arXiv **摘要/1811.08283**（2018年）

[16] Vaswani, A.,Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.：注意力就是一切所需。载于：Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vish- wanathan, S., Garnett, R.（编）：神经信息处理系统进展，第30卷。Curran Associates, Inc.，???（2017年）。 https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf[17]Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M.：ProteinBERT：一种通用的蛋白质序列和功能深度学习模型。《生物信息学》**38**（8），2102–2110（2022年） https://academic.oup.com/bioinformatics/article-pdf/38/8/2102/45474534/btac020.pdf。 https://doi.org/10.1093/bioinformatics/btac020[18] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B.：ProtTrans：通过自监督学习理解生命语言。《IEEE模式分析与机器智能汇刊》**44**（10），7112–7127（2022年）[19] Guo, Z., Nan, B., Tian, Y., Wiest, O., Zhang, C., Chawla, N.V.：基于图的分子表征学习。arXiv（2022年）。 https://doi.org/10.48550/ARXIV.2207.04869。 https://arxiv.org/abs/2207.04869[20] Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., Wang, H.：用于性质预测的几何增强分子表示学习。《自然·机器智能》**4**（2），127–134（2022年）。 https://doi.org/10.1038/s42256-021-00438-4[21]Wang, Y., Wang, J., Cao, Z., Barati Farimani, A.：通过图神经网络进行分子表示的对比学习。《自然·机器智能》**4**（3），279–287（2022年）。 https://doi.org/10.1038/s42256-022-00447-x[22] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman- Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R.：分析用于性质预测的已学习分子表示。《化学信息与建模杂志》**59**（8），3370–3388（2019年）。https://doi.org/10.1021/acs.jcim.9b00237

[23] Bjerrum, E.J., Threlfall, R.：使用循环神经网络进行分子生成

网络（循环神经网络）。arXiv **abs/1705.04612** （2017年）

[24] Bagal, V., Aggarwal,R.,Vinod, P.K., Priyakumar,U.D.：MolGPT：使用Transformer 解码器模型进行分子生成。《化学信息与建模杂志》 **62(9)**, 2064–2076（2022年）。https://doi.org/10.1021/acs.jcim.1c00600

[25]Popova, M., Isayev, O., Tropsha, A.：用于从头药物设计的深度强化学习。《科学进展》**4(7)**, 7885（2018年）https://www.science.org/doi/pdf/10.1126/sciadv.aap7885。https://doi.org/10.1126/sciadv.aap7885

[26] Wang, F., Feng, X., Guo, X., Xu, L., Xie, L., Chang, S.：通过在CycleGAN 中嵌入长短期记忆网络和注意力机制，改进从头分子生成。《遗传学前沿》**12** （2021年）。https://doi.org/10.3389/fgene.2021.709500

[27]Wu, Z., Johnston, K.E., Arnold, F.H., Yang, K.K.：基于深度生成模型的蛋白质序列设计。《化学生物学当前观点》**65**，18–27（2021年）。https://doi.org/10.1016/j.cbpa.2021.04.004。《机制生物学》 * 化学生物学中的机器学习[28] Anand, N., Eguchi, R., Mathews, I.I., Perez, C.P., Derry, A., Altman, R.B., Huang, P.-S.：利用学习得到的势能进行蛋白质序列设计。《自然·通讯》**13(1)**，746（2022年）。https://doi.org/10.1038/s41467-022-28313-9[29] Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., dos Santos, C., Chen, P.-Y., Yang, Y.Y., Tan,J.P.K., Hedrick, J., Crain, J., Mojsilovic, A.：通过深度生成模型和分子动力学模拟加速抗微生物药物发现。《自然·生物医学工程》**5(6)**，613–623（2021年）。https://doi.org/10.1038/s41551-021-00689-x[30] Castro, E., Godavarthi, A., Rubinfien, J., Givechian, K., Bhaskar, D., Krishnaswamy, S.：通过正则化潜在空间优化实现基于Transformer的蛋白质生成。《自然·机器智能》（2022年）。https://doi.org/10.1038/s42256-022-00532-1[31] Tolstikhin, I., Bousquet, O.,Gelly, S.,Sch¨olkopf, B.：Wasserstein自编码器。载于：第六届国际学习表征会议（ICLR）（2018年）。https://openreview.net/forum?id=HkL7n1-0b[32] Igel, C., Hansen, N., Roth, S.：用于多目标优化的协方差矩阵自适应。《进化计算》**15(1)**，1–28（2007年）。https://doi.org/10.1162/evco.2007.15.1.1

[33] Kyte, J., Doolittle, R.F.：一种展示蛋白质疏水特性的简单方法。《分子生物学杂志》157(1), 105–132 (1982年)[34] Chaudhury, S., Berrondo, M., Weitzner, B.D., Muthu, P., Bergman, H., Gray, J.J.：Rosetta v3.2 中蛋白质对接性能的基准测试与分析。《公共科学图书馆·综合》6(8), 22477 (2011年)[35] Andreatta, M., Alvarez, B., Nielsen, M.：Gibbscluster：肽序列的无监督聚类与比对。《核酸研究》45(W1), 458–463 (2017年)[36] Thomsen, M., Nielsen, M.：Seq2logo：一种用于构建和可视化氨基酸结合基序及序列谱的方法，涵盖序列加权、伪计数，以及氨基酸富集与缺失的双向表示。《核酸研究》40, 281–7 (2012年)。https://doi.org/10.1093/nar/gks469[37] Webb, B., Sali, A.：使用 modeller 进行比较蛋白质结构建模。《生物信息学现行方案》54(1), 5–6 (2016年)[38] Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J.Z., Hou, T.：利用 MM/PBSA 和 MM/GBSA 进行端点结合自由能计算：药物设计中的策略与应用。《化学评论》119(16), 9478–9508 (2019年)[39] Barlow, K.A., O Conchuir, S., Thompson, S., Suresh, P., Lucas, J.E., Heinonen, M., Kortemme, T.：flex ddG：基于 Rosetta 集合的方法，用于估算突变引起的蛋白质-蛋白质结合亲和力变化。《物理化学杂志B》122(21), 5389–5399 (2018年)[40] Chaudhury, S., Lyskov, S., Gray, J.J.：PyRosetta：一种基于脚本的接口，用于通过 Rosetta 实现分子建模算法。《生物信息学》26(5), 689–691 (2010年)[41]Hansen, N.：CMA进化策略：教程。CoRR abs/1604.00772 (2016年) 1604.00772[42] Auger, A., Hansen, N.：一种种群规模递增的重启式CMA进化策略。载于：2005 IEEE 进化计算大会，第2卷，第1769–17762页 (2005年)。https://doi.org/10.1109/CEC.2005. 1554902[43]Hansen, N., Ostermeier, A.：进化策略中的完全去随机化自适应。《进化计算》9(2), 159–195 (2001年)。https://doi.org/10.1162/106365601750190398

[44] Andreatta, M., Lund, O., Nielsen, M.: 使用吉布斯抽样方法同时对肽数据进行比对和聚类。《生物信息学》**29**(1), 8–14 (2012年)
https://academic.oup.com/bioinformatics/article-pdf/29/1/8/17100983/bts621.pdf。
https://doi.org/10.1093/bioinformatics/bts621