arXiv:2602.10422v1 [cs.CE] 11 Feb 2026

通过几何感知平滑驾驭异质蛋白质景观

Srinivas Anumasa[1], Barath Chandran[4], Tingting Chen[1], Nuwaisir Mohammad Rahman[1], Yingtao Zhu[1], Rushi Shah[1], Hongyu He[1], Peisong Zhang[1], Yizhen Liao[1], Yiming Tang[1], Yong Shen[6], TianfanFu[5], Rui Qing[7], Xiao Li[8],Sebastian Maurer-Stroh[3], Xinyi Su[1,3], Zhizhuo Zhang[2],Dianbo Liu[1,3]

[1]新加坡国立大学。[2]GSK.ai，美国。[3]A*STAR，新加坡。[4]鲁尔基印度理工学院。[5]南京大学。[6]西交利物浦大学。[7]上海交通大学。[8]北京大学。

摘要

生物分子的进化适应度景观极为稀疏且异质，功能序列在广阔的组合空间中形成孤立的高密度“岛屿”，而该空间中的变体大多不具备功能。蛋白质序列尤其体现了这一结构，但大多数生成式人工智能模型都隐含地假设数据分布是均匀的。我们表明，在异质生物序列空间中，这一假设从根本上不成立：固定的全局噪声水平会造成破坏性的权衡——要么过度平滑高密度功能簇，要么使稀疏区域碎片化，并产生非功能性幻觉。为解决这一局限，我们提出 *Density-Dependent Smoothing* (DDS)，一种几何感知生成框架，可根据底层序列景观的局部密度调整随机平滑。通过使扩散噪声与估计的序列密度呈反向关联，DDS 能够在高密度功能区域进行温和优化，同时推动对稀疏区域的可控探索。作为离散分子采样的插件机制，DDS 在抗体库、治疗性抗体设计、抗菌肽生成和冠状病毒抗体设计等任务中，始终优于最先进的

扩散模型和自回归模型。总而言之，这些结果表明，固定全局平滑假设从根本上限制了稀疏生物序列空间中的生成式建模，而几何感知平滑消除了这一限制，使以往固定噪声生成模型无法实现的可靠探索与设计成为可能。

关键词： 通过几何感知平滑探索异质蛋白质景观

引言

分子设计存在一个根本性的几何悖论：在极其庞大的组合空间中，存在着“稀疏性诅咒”[1]。例如，一个长度为 L的抗体序列，位于一个基数为 20^L的流形上；然而，生物功能性蛋白质仅占据其中微乎其微的一部分 [2–5]。关键在于，这些功能序列并非均匀分布[6, 7]；相反，它们形成密集且彼此隔离的适应度“岛屿”，散布在广阔的低密度盆地之间，而这些盆地中充斥着无效或不稳定的构象。要在这一异质的进化景观中探索，需要一种能够兼顾两种微妙需求的生成模型：它必须积极探索岛屿之间稀疏的“暗物质”区域，以发现新变体；同时还要精细优化密集的高适应度簇，而不扭曲其精确的功能语法。

当前最先进的方法，尤其是离散扩散框架[8,9], ，往往因根本性的几何失配而无法协调这些相互竞争的需求。这些模型通常依靠固定的全局噪声尺度（σ）来平滑数据分布，隐含地假设景观是同质的。如图1所示，这一假设迫使模型做出破坏性的折中：足以跨越稀疏区域的噪声水平会过度平滑密集的岛屿，抹去关键的结构基序（过度平滑，图1b）。反之，足以保留这些基序的较低噪声水平，会导致景观在稀疏区域碎片化，使采样过程陷入局部极小值（碎片化，图1c）。因此，现有模型往往必须牺牲生物学合理性来实现多样性，或反之。

重要的是，这种失效并非超参数调优不佳所致。在高度异质的生物序列空间中，单一的全局噪声尺度无法同时保留密集区域内细粒度的功能结构，并支持对稀疏区域的探索。因此，固定噪声生成模型不得不在多样性与生物学有效性之间面临不可避免的权衡。这一局限反映了全局平滑假设与生物适应度景观局部几何结构之间的根本性不匹配，说明在这种情况下，要实现可靠的生成建模，就需要能够根据局部数据密度进行调整的机制，而非依赖单一的全局尺度。

我们认为，要有效地在这一景观中导航，生成过程的“平滑性”必须动态匹配生物地形的“粗糙度”。我们

提出，扩散噪声不应被视为静态超参数，而应作为一种几何感知控制机制，以适应生物序列密度的局部变化[10, 11]。为将这一见解付诸实施，我们引入密度依赖平滑（DDS），该框架用直接从训练分布估计的数据依赖噪声尺度取代全局噪声调度[12]。通过利用核密度估计刻画生化特征空间的局部几何结构，DDS将扩散噪声与局部序列密度反向耦合。这使得生成动力学能够反映底层序列分布的几何结构，从而在高密度功能簇内实现温和细化，并在稀疏区域进行更广泛的探索（图1d–f）。

我们在治疗性抗体 [13] 和抗菌肽（AMP） [14] 设计任务中验证了这一方法，结果表明，DDS 无需全局统一调优的噪声尺度，并能使基于扩散的生成建模在高度异质的生物序列空间中稳定运行。除了标准指标外，我们还利用 ESM-2[15] 确认，DDS 生成的序列不仅具有多样性，而且在生物物理层面切实可行，并表现出更高的结构折叠置信度和预测结合亲和力 [15]。至关重要的是，同源性分析 [16] 显示，DDS 超越了固定噪声基线方法的局限；后者往往依赖对训练数据的简单记忆，而 DDS 能够生成真正的进化新颖性。这些结果表明，尊重生物数据的异质几何结构，是实现可靠、可扩展且以生物学为依据的从头蛋白质设计的先决条件。

结果

我们在受控合成基准和真实生物序列设计任务中评估了密度依赖平滑（DDS）。首先，我们在具有已知真实值的多模态玩具数据集上验证该方法，以量化保真度、模式覆盖率和虚假生成情况。随后，我们使用来自 Observed Antibody Space (OAS) 的配对抗体库评估 DDS 在大规模抗体序列生成中的表现[13]，，并进一步在治疗性抗体和抗菌肽基准上进行评估。对于 OAS 生成的抗体，我们还通过结构置信度、语言模型似然和进化同源性等互补指标，评估其生物学合理性。综合来看，这些结果证明了密度感知平滑在稀疏且异质的蛋白质景观中进行生成式建模的优势。DDS 无需全局统一调优的噪声参数，因此能使生成模型在高度异质的生物序列空间中稳定运行，而固定噪声模型在这一情形下会系统性地失效。

异质分子数据分布需要自适应噪声尺度。

高维蛋白质序列数据的一个显著特征是极端稀疏性。对于大小为 $V = 21$（20 种残基加上一个填充标记）的氨基酸词表以及长度为 $L = 297$ 的序列，组合空间包含 $V^L = 21^{297}$ 种可能的配置。在这个天文数字般庞大的空间中，只有极小的一部分对应于生物学上可行的抗体，这意味着观测到的序列

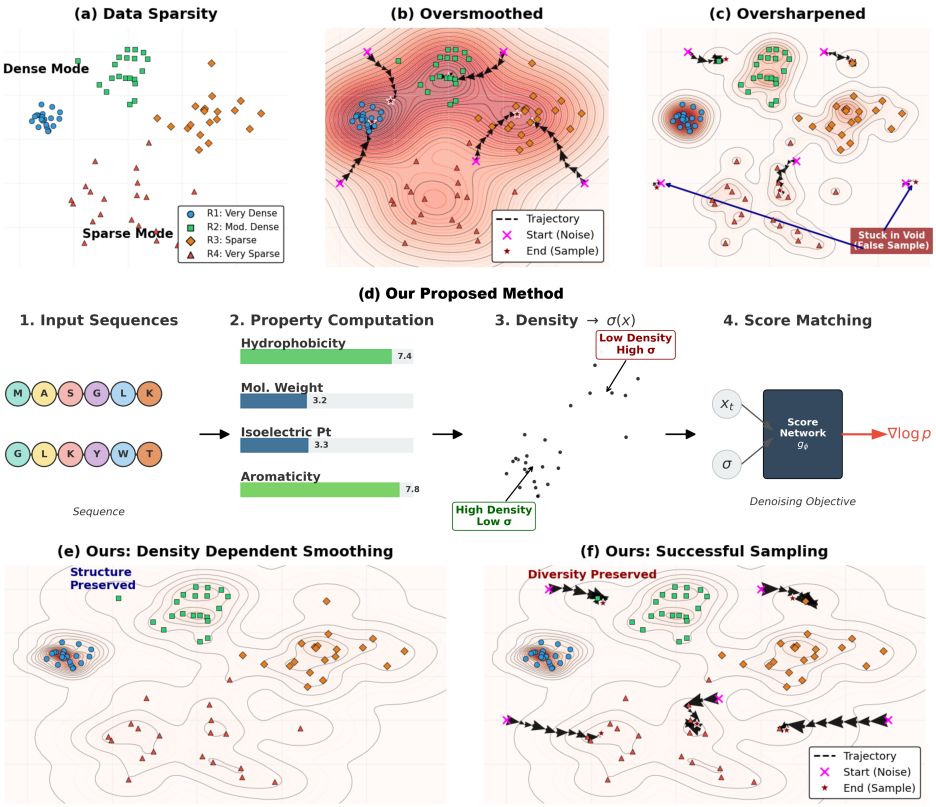


图1：**固定噪声尺度无法捕捉蛋白质景观的异质几何**。(a) 一个合成数据集，展示了生物数据中典型的密集功能簇与稀疏空洞的混合。(b)**过度平滑：** 固定的高噪声尺度会合并不同的功能模式，抹去细粒度多样性。(c) **碎片化：** 固定的低噪声尺度会在稀疏区域形成零梯度空洞，导致采样轨迹停滞。(d) **密度依赖平滑（DDS）：** 我们提出的方法通过生物物理性质估计局部密度。(e) 所得自适应景观在稀疏空洞中施加强平滑（以弥合间隙），在密集簇中施加温和平滑（以保留结构）。(f) **结果：**采样轨迹成功穿越该景观，恢复多样模式，且不会幻觉出假阳性。

占据底层分布中彼此孤立、不规则且高度稀疏的区域。这种稀疏性给生成式建模带来了根本性的挑战。

在 dWJS 等框架中，噪声水平 σ 的选择会直接影响模型如何在这一稀疏景观中导航。较大的 σ 会过度平滑数据流形，掩盖重要的结构变化；而较小的 σ 则会形成过度碎片化的景观，使采样器陷入狭窄盆地。因此，

选择适当的噪声水平对于确保训练稳定和有效生成样本至关重要。

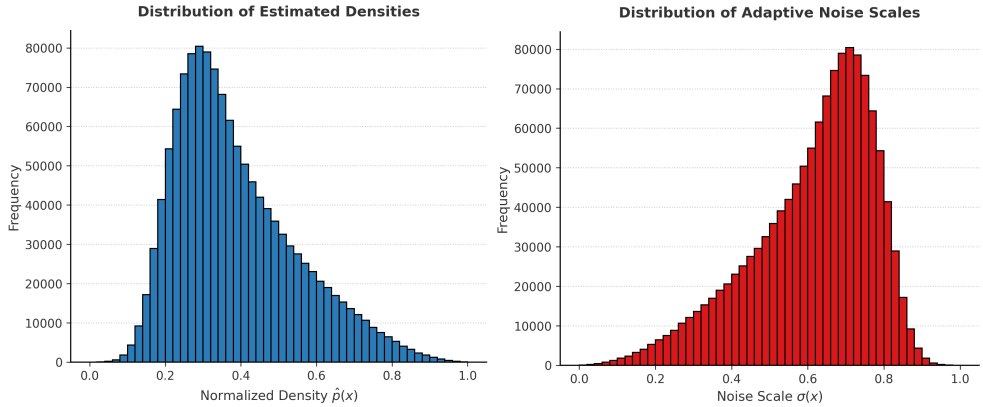
为量化抗体数据集中的稀疏性，我们使用核密度估计器（KDE）估计局部样本密度。给定具有特征表示 $f(x_i) \in \mathbb{R}^d$的训练样本 $\{x_i\}_{i=1}^N$，点 $f(x)$处的KDE估计由下式给出

$$\hat{p}\big(f(x)\big) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{f(x) - f(x_i)}{h}\right), \tag{1}$$

其中 $K(\cdot)$为核函数， $h > 0$ 表示带宽。

对于抗体数据集，我们将每条序列投影到六维生化特征空间，其组成包括(a)疏水性、(b)分子量、(c)等电点、(d)芳香性、(e)不稳定性指数以及(f) β折叠片含量——这些量与折叠、稳定性和功能可行性密切相关。由于完整数据集包含超过130万条序列，我们使用随机傅里叶特征 [17]加速KDE计算。相比之下，对于我们的合成四维玩具数据集，KDE直接在独热嵌入空间中计算，无需近似。

图 2（左）展示了估计得到的序列密度分布，揭示出数据集存在显著的异质性。少数序列集中在高密度簇中，而大多数序列则分布在生化特征空间中密度各异的稀疏区域



(a) (b)

图2：抗体序列空间的异质性要求进行自适应平滑。(a) 训练数据集估计核密度$\hat{p}(x)$的分布。数值分布范围较广，凸显了适应度景观稀疏性的差异：序列并非均匀分布，而是占据从密集簇到稀疏过渡区域的不同区域。(b) 相应自适应噪声尺度$\sigma(x)$的分布。通过将密度反向映射到噪声，我们的方法为 σ赋予动态取值范围，确保不同稀疏程度的区域获得适当水平的平滑。

。这一观察表明，稀疏性并非均匀分布，而是在序列景观的不同区域间存在显著差异。这种稀疏性的异质性意味着，在生成建模过程中使用单一的全局噪声尺度并非最优，因为它无法适应密集区域和稀疏区域各自不同的局部结构。这一观察促使我们提出密度依赖平滑（DDS），该方法显式纳入局部密度信息，以数据感知的方式调节生成动力学。

受这一观察的启发，我们将全局噪声水平替换为依赖数据的噪声水平 $\sigma(x)$，使其能够适应局部密度。具体而言，我们定义

$$\sigma(x) \propto (1 - \hat{p}\big(f(x)\big)), \tag{2}$$

使得密集区域中的样本获得较低的噪声水平，而稀疏区域中的样本则分配到较高的噪声水平。图 2 （右）展示了由此得到的 σ值分布，它与数据集的经验稀疏性分布高度吻合。

训练期间，这些依赖数据的噪声水平会针对每个样本调节去噪目标。推理期间，我们直接从训练数据中观测到的经验分布采样噪声水平 σ 。如使用以σ为条件的分数模型进行采样（方法）所示，每条轨迹都会使用其特定采样得到的 σ进行初始化和引导，从而确保生成过程反映蛋白质景观的异质稀疏性，而不是依赖单一的 σ 值。

DDS 能够准确恢复合成四维玩具数据集的底层分布。

为证明我们的 DDS 能够正确估计底层分布，我们在合成四维玩具数据集上设计了一项受控实验。该设置旨在以简化的方式模拟蛋白质序列的关键特征。

我们考虑一个大小为 $V = 21$ 的词汇表（类似于 20 种氨基酸加上一个填充标记），但将序列长度限制为 $L = 4$ ，而不是 297。因此，离散序列空间为

$$\mathcal{X} \;=\; \{0, 1, \ldots, 20\}^4, \qquad |\mathcal{X}| \;=\; V^L \;=\; 21^4.$$

类比于真实的蛋白质序列，由于生化约束，只有一小部分氨基酸序列能够产生有效的蛋白质，我们在这个玩具数据集上施加一个人工约束。具体而言，为每个词汇表元素 $v \in \{0, 1, \ldots, 20\}$ 分配一个随机的“疏水性”得分 $h\,(v) \;\sim$ 均匀$[-4, 4]$。对于候选序列 $x = (x_1,\ x_2,\ x_3,\ x_4) \in \mathcal{X}$，我们将其疏水性总和定义为

$$H(x) \;=\; \sum_{i=1}^{4} h(x_i).$$

当且仅当一个序列满足阈值判据时，它才被视为有效

$$H(x) \;\geq\; \tau,$$

其中 τ 是一个固定常数（例如，在我们的实验中为 $\tau = 20$ ）。

形式化地说，有效序列的集合为

$$\mathcal{X}_{\text{valid}} = \{ x \in \mathcal{X} : H(x) \geq \tau \}.$$

通过穷举所有可能的序列 $x \in \mathcal{X}$ 并根据上述约束进行筛选，我们得到 4D 空间中有效点的稀疏分布。这与真实蛋白质序列数据的行为相似：绝大多数可能的序列都无效，而观测数据位于组合空间中稀疏且具有结构的区域。

我们使用固定噪声尺度 $\sigma \in \{0.1, 0.2, 0.25, 0.35, 0.50, 0.75, 1.0\}$，将 DDS 与标准 dWJS 基线方法进行比较。对于 DDS，我们将噪声范围限制在 $\sigma \in [0.25, 0.35]$。这一选择依据基线模型的实证观察：中间噪声水平（$\sigma = 0.25$ 和 0.35）的表现始终优于较低或较高的噪声水平。因此，DDS 被配置为在这一经实证验证的有效噪声范围内进行插值。

该合成数据集由 120 个真实模态组成，其中 80% 用于训练，其余 20% 留作评估模态恢复效果。所有模型均采用相同的 ByteNet 主干网络，并训练 500 个训练轮次。

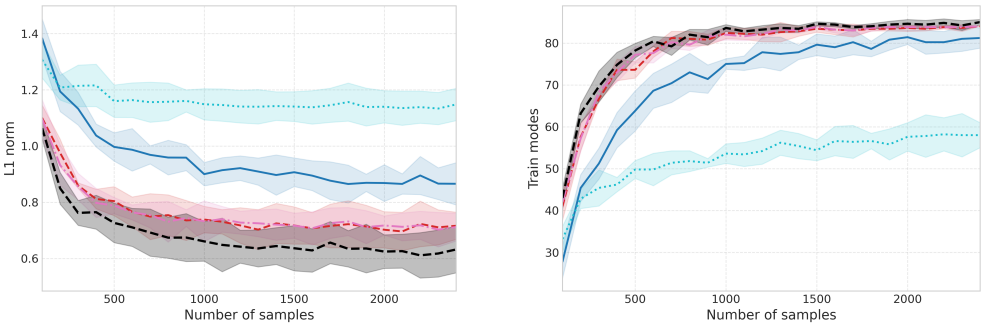
在评估过程中，我们使用数量逐渐增加的随机样本生成样本，每个推理阶段生成更多样本。对于每种噪声配置，我们使用不同的随机种子训练五个相互独立的模型。图3报告了这些运行结果在四个互补指标上的均值和标准差。为清晰起见，这里我们仅比较具有 4 个不同 σ 值的基线方法，而在 1 中，我们比较了具有全部 7 个不同 σ 值的基线方法。

自适应平滑在所有样本预算下均降低 ***L1***误差

如图 3（a）所示，随着生成样本数量增加，所有模型的 L_1 距离均有所下降，表明对真实分布的逼近效果得到改善。在所有样本预算下，DDS 始终在所有基线方法中取得最低的 L_1 误差。这表明，数据依赖平滑生成的 sigma 值能够更好地反映局部数据密度，使模型相比固定σ 基线方法更忠实地恢复真实分布。

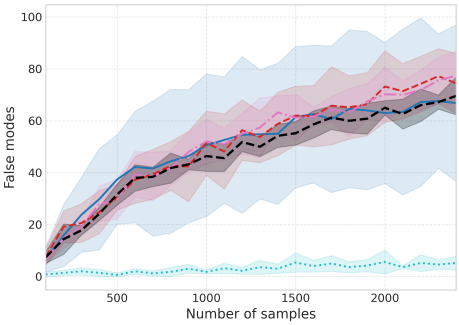
DDS 在避免过度平滑的同时，兼顾模态恢复与泛化。

图3(b) 和 (d) 展示了采样过程中恢复的真实模态 数量，以及发现的留出模态或新模态数量。随着采样预算增加，DDS 和采用 $\sigma \in \{0.25\}$ 的基线方法能够恢复大多数真实模态。相比之下，采用较大固定 σ 值的基线方法（例如 0.5 和 1.0）恢复的模态显著更少，表明过度平滑会抑制对底层模态结构的探索。在模态发现任务中，DDS 的表现与最佳固定 σ 基线方法相当，成功识别出训练过程中从未出现过的留出模态。这表明，我们的数据依赖平滑策略能够实现受控探索，并泛化到训练支持集之外。

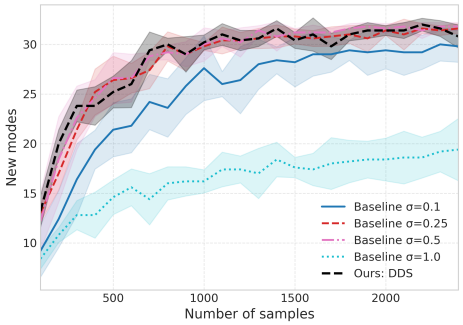


(a) L_1 范数

(b) 覆盖的真实模态



(c) 覆盖的错误模态



(d) 覆盖的新模态

图 3：数据依赖平滑改善了合成多模态景观上的保真度与探索的平衡。(a) L_1 **距离**随样本数量变化，表示生成分布与真实分布之间的差异。密度依赖平滑（DDS；黑色虚线）始终能达到更低的误差，说明其对目标分布的近似效果更好。(b) **恢复的真实训练模态数量**。DDS 的召回率与经过良好调优的低噪声基线相当，而高噪声模型覆盖的模态不足。(c) **生成的错误（虚假）模态数量**。随着采样增加，低噪声基线生成的虚假模态越来越多，而 DDS 能显著减少虚假模态的数量。(d) **恢复的新（留出）模态数量**。DDS 能够泛化到训练模态之外，在保持保真度的同时恢复未见过的模态。阴影区域表示 $\pm$ 个独立运行的均值 $n = 5$ 标准差。

DDS 在探索与正确性之间取得平衡，最大限度减少虚假模态生成。

图 3(c) 突出了探索与正确性之间的权衡。尽管激进探索可能增加多样性，但往往会产生 虚假模态，即不对应任何真实结构的样本。DDS 实现了良好的平衡：它既能有效恢复真实模态并发现新模态（图 3(b,d)），又比激进锐化的基线方法生成显著更少的虚假模态。在

相比之下，使用较小固定 σ 值的模型往往会过度探索，在真实分布之外生成大量无效样本。DDS 避免了这种失败模式，始终贴近正确的数据流形，同时仍能有意义地探索训练分布之外的区域。

DDS 生成的序列既新颖，在生物物理层面也合理。

我们在 Observed Antibody Space (OAS) 数据库中的可变长度抗体序列上评估了所提出的 DDS 模型 [13]。每条抗体序列表示为 $x = (x_1, \ldots\ldots, x_d)$，其中 $x_l \in \{1, \ldots, 20\}$ 表示位置 l上的氨基酸（AA）类型。所有序列均采用 AHo编号方案进行比对，随后进行独热编码。由于抗体序列的长度可变，我们在末尾添加间隙标记，以获得固定长度的表示：重链序列填充至 149 个位置，轻链序列填充至 148 个位置，每个位置可取 21 种符号之一（20 种 AA 加一个间隙标记）。比对和编码后，每条抗体都表示为维度为 $(149+148)\times 21$ 的二进制向量。对于 POAS，我们尤其关注新样本的发现：从随机噪声初始化，并在经过训练的 σ 条件分数模型的引导下生成新样本。

我们将模型与一组涵盖基于扩散的方法和自回归语言模型的多样化基线方法进行比较。我们的主要基线是 dWJS [8]，我们的方法正是在此基础上构建的。此外，我们还评估了以下方法：潜在序列扩散模型 SeqVDM（一种变分扩散的离散扩展形式） [11]；采用能量参数化的基于得分的模型 DEEN [18]；专为抗体设计定制的基于Transformer的语言模型 IgLM [19]；ESM2 [15]；以及一个大型预训练语言模型（ GPT-3.5 ）。

表1：基线方法在抗体序列设计上的模型性能比较。

模型 / σ	U ↑	ID ↑	ED ↑	DCS ↑	WD ↓
dWJS (σ=1.0)	1.00	95.59 ± 24.58	34.38 ± 19.60	0.45 ± 0.30	0.072 ± 0.075
dWJS (σ=0.5)	1.00	**119.29 ± 20.94**	61.71 ± 18.88	0.25 ± 0.28	0.068 ± 0.046
DDS	1.00	106.92 ± 22.33	50.71 ± 17.57	0.45 ± 0.30	**0.057 ± 0.024**
SeqVDM [11]	1.00	57.40	60.00	0.40	0.062
DEEN [18]	0.99	42.70	50.90	0.41	0.087
GPT-3.5	0.66	46.10	55.40	0.23	0.140
IgLM [19]	1.00	34.60	48.60	**0.53**	0.080
ESM2 [15]	1.00	77.56	**70.99**	0.06	0.150

[1]我们复现了 dWJS 的结果，而其余基线方法的性能数值则取自 dWJS 研究中报告的结果，该研究未提供标准差值[8]。

我们使用 Frey 等人 [8] 中所述的相同实验设置，训练了所提出的 DDS模型和基线 dWJS。两个模型均采用 ByteNet 主干网络训练 40 个训练轮次，并使用相同的架构选择、学习率和权重衰减设置，以确保比较公平。对于基线 dWJS，我们评估了两个

固定噪声尺度超参数（$\sigma \in 0.5,\ 1。0$）。相比之下，我们的 DDS模型采用数据依赖平滑机制：通过核密度估计从训练数据中估算每个样本的 σ 值，然后进行归一化和重新缩放，使其落在 $[0.4, 0.6]$ 范围内。为评估生成的抗体序列的质量，我们评估了一组涵盖唯一性、多样性、新颖性和分布相似性的指标，例如唯一性（U）、内部多样性（ID）、编辑距离（ED）、分布一致性得分（DCS）和 Wasserstein距离（WD）（方法）。

表1 总结了我们的方法与 dWJS 及其他基线方法的性能比较。优秀的生成模型应兼顾多样性和质量：生成的序列彼此之间以及与用于训练模型的样本之间都应有足够差异，同时仍接近真实的底层分布。从表 1可以看出，与基线方法相比，我们的 DDS 模型更有效地实现了这种平衡。具体而言，DDS 在保持竞争力的多样性（ID 和 ED）的同时，生成了质量更高的样本，这一点从 DCS 和 Wasserstein距离 指标中可以看出。

相比之下，采用较大固定噪声尺度（$\sigma = 0.5$）的 dWJS 虽然表现出较高的多样性，却以样本质量为代价——其 Wasserstein距离 得分明显更差——这表明不受控制的探索会将模型推向序列空间中不切实际的区域。尽管 dWJS（$\sigma = 0.5$）在 ED 指标上表现良好，说明它生成的序列与验证集不同，但其中许多序列质量较低或功能失常。

总体而言，DDS 生成的序列既新颖，又具有生物物理合理性，在探索性与真实性之间实现了更优的权衡。这一趋势与我们的合成实验结果一致：DDS 成功找回了更多真实模态，同时避免了虚假模态。所有指标均呈现一致的结果，凸显了采用数据依赖平滑而非固定噪声生成策略的优势。

先前研究 [8] 表明，数值高于 0.3 足以生成适用于湿实验环境的样本。因此，DDS 在质量与分布覆盖度之间实现了有效平衡。

DDS 在生成 CDR突变体时，能够在保持生物学合理性的同时维持较高的多样性。

接下来，我们使用 hu4D5抗体突变体数据集评估我们的方法在生成 CDR突变体这一任务上的表现 [20]。去重并移除标签冲突的样本后，数据集包含约 9k 个结合型和 25k 个非结合型 hu4D5 CDR 突变体，每个突变体最多包含 10 个突变。为评估生成序列的质量，我们训练了一个二元分类器，该分类器在独立同分布的验证集上准确率达到 85%。所有生成模型均在包含 34k 个 hu4D5 CDR H3突变体的完整数据集上训练。按照 GFlowNets [21], 中提出的评估流程，我们生成 500 条序列，并保留其中按分类器预测结合概率排名最高的 100 条用于分析。表2报告了各方法的性能对比。我们希望生成模型能够兼顾样本的多样性与质量。

GFlowNets表现出较高的内部多样性（ID），但其生成的许多序列获得的预测结合概率较低，表明它们可能落在

真实结合物的分布之外。相比之下，gg-dWJS [9] 取得了较强的结合概率（P_{bind}），但产生的多样化样本较少。我们的DDS模型实现了更理想的平衡，在保持更高多样性的同时保留了生物学合理性，这一点体现在调和平均（调和平均数）指标中，证明了其在生成既新颖又具有功能前景的序列方面的有效性。

表2：不同模型的结合概率与多样性指标比较。

模型	P_{bind}(↑)	ID(↑)	ED(↑)	调和平均数(↑)
dWJS ($\sigma = 1.0$)	0.96 ± 0.01	6.15 ± 1.37	2.66 ± 0.68	0.41 ± 0.08
dWJS ($\sigma = 0.5$)	0.95 ± 0.02	6.60 ± 1.28	2.68 ± 0.56	0.41 ± 0.07
DDS	0.93 ± 0.03	6.82 ± 1.32	2.82 ± 0.63	**0.43 ± 0.07**
gg-dWJS	**0.98 ± 0.01**	4.58 ± 1.17	2.24 ±0.58	0.36±0.08
GPT-4o	0.58 ±0.36	7.42±1.57	**3.29±0.96**	0.34±0.16
SeqVDM	0.36 ±0.26	7.48±1.21	3.25±0.64	0.24±0.16
GFlowNets	0.32	**9.16**	-	-

[1]我们复现了 dWJS 的结果，而其余基线方法的性能数据则取自 gg-dWJS 研究中报告的结果及其 GitHub 仓库[9]。

DDS 在抗菌肽（AMP）生成中实现了序列质量与多样性之间的良好平衡。

本实验旨在生成具有抗菌特性的肽。我们从 DBAASP数据库 [14]中的 AMP 和非抗菌肽构建数据集。遵循 Jain等人 [21], 的方案，我们选取长度为 12–60、靶向革兰氏阳性菌的肽，得到 6,438 条 AMP 和 9,222 条非抗菌肽。

我们按照 [22], 将数据集拆分为两个子集 D1 和 D2。D1 提供给生成模型，而 D2 仅用于训练预言分类器以进行评估。该拆分遵循与 Jain等人相同的组不相交性规则：对于任意肽 $x \in D1$，不存在属于同一组的肽 $x' \in D2$，反之亦然。因此，D1 中包含 3,219 条 AMP 和 4,611 条非抗菌肽。我们使用一个准确率达到 91% 的分类器训练预言分类器。遵循 [21], ，我们使用 3 种不同的指标以及调和平均值来评估模型，后者同时衡量质量和多样性。

调和平均（HM）[23]。 HM 指标体现了生成样本质量 与多样性 之间的权衡。对于每条生成的序列，质量得分由训练好的分类器得出，而多样性得分则通过归一化编辑距离（编辑距离除以序列长度；对于 AMP 序列，该长度为 60）计算。对这两个值取调和平均，可促使模型生成兼具高质量和多样性的样本，而不是只在其中一个维度表现出色。

表3 总结了不同模型在多项评估指标上的性能。指标$Performance$ 衡量训练后的分类器赋予

表 3：我们的方法与基线方法在 AMP数据集上的模型性能比较。

方法	调和平均值(↑)	性能 (↑)	ID (↑)	ED (↑)
DDS	**0.73 ± 0.08**	0.92 ± 0.10	45.93 ± 5.40	36.80 ± 6.41
dWJS (σ=0.5)	0.65 ±0.15	0.70 ±0.27	**48.83 ± 3.91**	**40.89 ± 4.79**
dWJS (σ=1.0)	0.36 ±0.13	**0.99 ±0.01**	21.21 ± 6.54	13.87 ± 5.97
gg-dWJS	–	0.98	25.78	15.02
GFlowNet-AL	–	0.93	22.34	28.44
DynaPPO	–	0.94	12.12	9.31
COMs	–	0.76	19.38	26.47
GFlowNet	–	0.87	11.32	15.72

[1]我们复现了 dWJS 的结果，而其余基线方法的性能数据则取自 gg-dWJS 研究报告的结果[9]。

生成的肽为 AMP 的平均概率。按照 GFlowNets [21], 中介绍的评估方案，每个模型生成 500 条序列，并根据分类器预测的 AMP 概率进行排序，保留排名前 100 条用于分析。

在 ID 和 ED 指标下，采用 $\sigma = 0.5$ 的 dWJS 表现相对较好；然而，其对应的性能得分表明，根据分类器的判断，其生成的许多样本不太可能是 AMP。相比之下，gg-DWJS 的性能值较高，但生成的样本多样性较低。

我们的模型在序列质量和多样性之间取得了良好的平衡，这体现在调和平均数（HM）指标上。值得注意的是，我们的方法生成的所有前 100 个样本都被分类为 AMP，但预测概率介于 0.85 至 0.95 之间，平均值为 0.92。这种分布体现了模型生成多样且具有生物学合理性的抗菌肽候选物的能力。

基于密度的 σ 估计对于生成样本的质量和多样性至关重要。

为了验证通过密度估计得到的偏斜 σ 分布是否确实有助于提升我们模型的性能，我们进行了一项实验：将 σ 值均匀地（$\sigma = 0.4$ 至 0.6）分配给训练数据，而不是通过基于核密度估计的密度估计得出。表4 比较了基于密度的 σ 分配方式与均匀分配 σ 值的性能。

在 ID 和 ED 指标下，均匀-σ 变体的表现略优于我们基于密度的方法；然而，在性能和调和平均值指标下，我们提出的基于密度的方法明显优于该变体。调和平均值越高，表示生成样本的多样性与质量之间的平衡越好。

还值得注意的是，就 ID 和 ED 而言，均匀-σ 模型的性能与固定 $\sigma = 0.5$ 的 dWJS 相似。这是意料之中的，因为均匀采样的 σ 值的平均值接近 0。5。然而，与固定-σ dWJS 相比，均匀-σ 模型在 性能 和调和平均值 指标上表现更佳。

这些观察结果表明，使用一系列 σ 值，即使是均匀采样的，也能起到有益的正则化作用。尽管如此，基于密度的 σ 估计还能进一步提升质量和多样性，凸显了使 σ值与底层数据几何相匹配的重要性。

表 4：基于密度的 σ 估计与均匀分配的 σ 值之间的模型性能比较

方法	调和平均值(↑)	性能 (↑)	ID (↑)	ED (↑)
DDS	**0.73 ± 0.08**	0.92 ± 0.10	45.93 ± 5.40	36.80 ± 6.41
均匀	0.68 +0.11	0.75 ± 0.22	**48.56 ± 3.86**	**40.82 ± 4.81**

DDS生成结构上合理的序列，同时避免在高平滑水平下观察到的极端记忆。

尽管前文报告的生物物理性质分布初步表明生成序列与天然蛋白质相似，但这些汇总统计可能无法充分反映单个序列是否具有生物学可行性。为解决这一问题，我们进行了三项互补的计算验证分析，分别评估结构可折叠性、序列自然性和进化合理性。

DDS生成具有高结构置信度的蛋白质

功能性蛋白质的一项基本要求是其能够折叠成稳定的三维结构。我们使用ESM-评估了生成序列的结构合理性2 [15], ，这是一种最先进的蛋白质结构预测模型。对于每条生成序列，我们计算了预测的局部距离差异测试（pLDDT）分数，该分数以0–100的尺度量化模型对其结构预测的置信度。分数高于70表示预测可信，而分数高于90则表明对预测结构具有非常高的置信度。

我们为每个模型的500条序列预测了结构，并在表5中报告了平均pLDDT分数。作为参考基线，我们还计算了验证集中留出的天然序列的pLDDT分数。

DDS生成自然外观的蛋白质序列，并避免对训练数据的简单记忆

在数百万条进化序列上训练的蛋白质语言模型捕捉了表征天然蛋白质的复杂氨基酸共现模式和位置依赖关系。我们利用ESM-2 [15], 这一拥有6.5亿个参数的大规模蛋白质语言模型，来评估我们的生成序列看起来有多“自然”。

对于每条序列，我们依次遮蔽每个位置，并测量模型预测真实氨基酸的能力，以此计算伪困惑度。形式上，

表 5：结构预测置信度分数。pLDDT 越高，表示序列能够折叠成稳定三维结构的置信度越高。pLDDT > 70 报告了结构预测结果可信的序列数量。

模型	平均 pLDDT（↑）	pLDDT > 70 （数量）（↑）
天然（参考）	75.30 ± 0.013	500
DDS（我们的方法）	75.30 ± 0.015	498
dWJS（$\sigma = 0.5$）	72.19 ± 0.036	396
dWJS（$\sigma = 1.0$）	76.52 ± 0.011	500

对于序列 $\mathbf{x} = (x_1, \ldots\ldots, x_L)$，伪困惑度定义为：

$$\mathrm{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L}\log P(x_i \mid \mathbf{x}_{\setminus i})\right), \tag{3}$$

其中，$P(x_i \mid \mathbf{x}\setminus i)$ 是 ESM-2 在给定其他所有位置后，为真实氨基酸 x_i分配的概率。伪困惑度越低，表示该序列越符合从天然蛋白质中学习到的模式。

表6：使用 ESM-2 伪困惑度（伪困惑度）和平均对数概率评估生成的蛋白质序列与真实蛋白质序列。伪困惑度越低，表示进化上的合理性越高；平均对数概率越高，则表示与语言模型的契合度越好。结果基于前 500 条序列报告。

模型 / σ	伪困惑度 ↓	平均对数概率 ↑
天然（参考）	2.51 ± 0.46	-0.90 ± 0.17
DDS（我们的方法）	2.88 ± 0.39	-1.05 ± 0.14
dWJS（$\sigma = 0.5$）	3.83 ± 1.01	-1.31 ± 0.24
dWJS（$\sigma = 1.0$）	2.42 ± 0.42	-0.87 ± 0.17

综合来看，表 5 和 6 表明，dWJS（$\sigma = 1.0$）在结构置信度和序列自然性方面均达到最高，而 dWJS（$\sigma = 0.5$）在这两个指标下的得分均最低。然而，仅凭这些指标无法区分有效的泛化与对训练数据的简单记忆，因此有必要对序列同源性进行明确分析。

DDS 在结构有效性与进化新颖性之间取得平衡

为评估生成的序列是否体现了真正的进化新颖性，而非记忆训练样本，我们使用

BLAST [16] 对 POAS训练集进行了序列相似性搜索。对于每个生成的抗体，我们在配对层面评估同源性，将序列一致性定义为重链和轻链各自最佳 BLAST 一致性中的较小值。这一定义较为保守，因为配对抗体的新颖性取决于其同源性较低的那条链。

我们报告配对级序列一致性在互不重叠区间中的分布： $< 30\%$、30–50%、50–70%、70–90%、90–95%和 $> 95\%$。70–90%区间对应于新颖但在进化上合理的序列，而高于95%的一致性表明近重复序列以及潜在的记忆。

表7：针对POAS训练集的序列同源性分析。配对级一致性计算为重链与轻链最佳BLAST一致性的最小值。数值以百分比报告，合计为100%。所有模型的样本命中率均达到100%

模型 / σ	<30	30–50	50–70	70–90	90–95	>95
DDS（我们的方法）	0.0	0.0	0.0	38.6	36.4	25.0
dWJS（$\sigma = 1.0$）	0.0	0.0	0.0	10.0	24.2	65.8
dWJS（$\sigma = 0.5$）	0.0	1.2	14.2	76.6	6.4	1.6

同源性分析表明，dWJS（$\sigma = 1.0$）在结构和语言模型指标上的优异性能，很大程度上是由记忆驱动的，因为生成序列中有65.8%与训练集的相似度超过95%。相比之下，dWJS（$\sigma = 0.5$）显著减少了记忆，并将大多数样本（76.6%）集中在理想的70–90%相似度范围内，但代价是结构置信度降低。DDS处于中间状态，在保持较高结构有效性的同时，由于数据依赖平滑而表现出更高的冗余性。

综合来看，这些结果凸显了有效性与新颖性之间的根本权衡。虽然结构预测置信度和语言模型似然倾向于 favor 接近训练分布的模型，但基于BLAST的同源性分析对于揭示原本容易被掩盖的记忆效应至关重要。DDS实现了均衡的权衡，能够生成结构上合理的序列，同时避免在高平滑水平下观察到的极端记忆现象。

这些结果表明，在较大的固定平滑强度下，表面上较高的样本质量可能由静默记忆造成。这种失效模式无法仅凭结构得分或语言模型分数检测出来，但密度感知平滑可以缓解这一问题。

DDS 生成多样化样本，与真实值的免疫原性分布相似

我们还在一项真实世界的抗体设计任务中评估了我们的方法，重点考察 DDS 能否在保留真实数据免疫原性分布特征的同时，生成多样化的 VH–VL 序列。我们构建了训练数据集

表 8：免疫原性统计与分布匹配。KS/WD 越低越好（↓）。

模型	KS ↓	WD ↓	平均分数	最高分数
真实（训练集）	–	–	0.0162 ± 0.0123	0.0756
Ours	**0.1688 ± 0.0176**	0.0043 ± 0.0006	0.0149 ± 0.0136	0.0656
dWJS($\sigma = 0.5$)	0.1804 ± 0.0180	0.0041 ± 0.0006	0.0169 ± 0.0141	0.0647
dWJS($\sigma = 1.0$)	0.2868 ± 0.0167	**0.0040 ± 0.0005**	0.0191 ± 0.0119	0.0604

表 9：生成抗体的新颖性和多样性。对于 ED/ID (↑)，数值越高越好。ED 和 ID 均已归一化。

模型	ED ↑	ID ↑
Ours	0.334 ± 0.0780	**0.507 ± 0.0040**
dWJS($\sigma = 0.5$)	**0.344 ± 0.686**	0.484 ± 0.048
dWJS($\sigma = 1.0$)	0.155 ± 0.0505	0.312 ± 0.0056

来自**冠状病毒抗体数据库 (CoV-AbDab)**[24], ——一个经过整理的资源库，收录了经实验表征、能够与冠状病毒结合的抗体。CoV-AbDab 提供抗体重链和轻链可变区序列及相关元数据。我们从官方资源库[1]下载了 CoV-AbDab 汇总表，该表为包含抗体序列注释的 CSV 文件。每个条目都包含重链可变区 (VH) 和轻链可变区 (VL) 序列。我们应用了若干预处理步骤，以获得干净的配对 VH–VL 数据集。我们通过筛除标记为 抗体 的条目，确保仅保留经典抗体，并排除纳米抗体 (纳米抗体)。从 重链可变区 或 VHH（重链）和 VL （轻链）字段中提取可变区序列。

筛选后，每条保留记录均包含一条重链和一条轻链。为避免冗余，移除了重复的 VH–VL 配对。最终数据集由用于训练的成对可变区序列组成，共计 11K 个样本。

免疫原性评估

我们使用丹麦技术大学健康科技提供的**NetMHCIIpan**网络服务器评估免疫原性风险，而非本地安装版本。对于每条抗体可变区序列，我们生成了重叠的 15-mer肽，并以 EL（洗脱配体）预测模式将其提交至服务器。我们获取预测结果，并通过汇总每条序列中所有肽的结合排名，将肽水平的预测结果归纳为序列水平的免疫原性评分。

我们使用两种分布相似性度量——柯尔莫戈洛夫–斯米尔诺夫和瓦瑟斯坦——比较生成抗体与真实抗体的免疫原性得分分布。

表 8 和表 9 总结了基于拼接后的 VH–VL 序列进行的免疫原性分布匹配以及新颖性/多样性评估。我们的模型实现了

[1] https://opig.stats.ox.ac.uk/webapps/covabdab/

与真实免疫原性分布最接近的匹配，获得最低的 KS统计量，同时保持较强的新颖性和最高的内部多样性，这表明它能够探索序列空间，且没有记忆的迹象。相比之下，dWJS (σ=1.0) 生成的序列新颖性和多样性都显著较低，这与对训练样本的部分记忆相符。dWJS (σ=0.5) 的新颖性与我们的模型相当，但内部多样性略有下降，表明其对序列空间的覆盖度稍低。除 KS/WD 外，表 8还报告了免疫原性评分的均值和最大值，分别概括了聚合肽段级结合排名后，典型情况下及最坏情况下（尾部）的序列级风险。我们的模型的均值略低于真实数据集，最大值也低于真实数据集的最大值，这表明 DDS 在匹配分布的同时避免了罕见的高风险离群值。

选择 σ 范围对 DDS 性能至关重要：

由 $\sigma_{\min}$ 和 $\sigma_{\max}$定义的平滑范围，对 DDS 的性能起着至关重要的作用。为研究模型对这些超参数的敏感性，我们在 AMP数据集上进行了消融分析，如表10所示。在本研究中，$\sigma_{\min}$ 的取值范围为 0.3 至 0.9，$\sigma_{\max}$ 的取值范围为 0.5 至 1.0，并考虑了所有满足 $\sigma_{\max} > \sigma_{\min}$的有效组合。对于每种组合，我们评估了多个 σ 带宽，范围为 0.1 至 0.6。

结果表明，与其他设置相比，0.4–0.6 和 0.4–0.7 这两个范围始终表现出色。在 AMP 实验中，我们选择了 0.4–0.6 这一范围。尽管 0.4–0.7 范围的 ED 指标略高，但从性能得分来看，0.4–0.6 的整体质量更好，为 0.92；而 0.4–0.7 范围的得分为 0.87。

对于抗体序列生成任务，我们沿用了相同的范围（0.4–0.6）。由于计算限制，我们没有对抗体进行全面的消融实验；不过，我们观察到，在 AMP 上表现良好的 sigma 范围也能有效迁移到其他数据集。

表10：针对 AMP数据集 的 (σ_{low}，σ_{high}) 消融实验。每个单元格显示 **调和平均数 / ED**。

$\sigma_{\text{low}} \backslash \sigma_{\text{high}}$	0.5	0.6	0.7	0.8	0.9	1.0
0.3	0.61 / 43.49	0.68 / 40.67	0.72 / 36.21	0.63 / 29.27	0.51 / 22.22	0.51 / 22.22
0.4	0.65 / 42.10	**0.73 / 36.80**	**0.73 / 38.64**	0.64 / 29.45	0.48 / 20.08	0.30 / 10.65
0.5	–	0.72 / 35.51	0.68 / 32.20	0.57/ 24.98	0.49 / 20.70	0.31 / 11.08
0.6	–	–	0.68 / 31.66	0.51 / 21.29	0.51 / 21.29	0.29 / 10.49
0.7	–	–	–	0.60 / 26.72	0.46 / 18.98	0.32 / 11.68
0.8	–	–	–	–	0.35 / 12.98	0.29 / 10.50
0.9	–	–	–	–	–	0.28 / 9.73

方法论背景

神经经验贝叶斯（NEB） [25] 提供了一个有理论依据的框架，将非参数平滑与经验贝叶斯估计相结合，以执行最小二乘去噪。给定一个干净样本 x，通过添加方差为 σ^2 的高斯扰动，生成对应的含噪样本：

$$y = x + \mathcal{N}(0, \sigma^2). \tag{4}$$

经验 的干净信号的贝叶斯估计为 给定

$$\bar{x} = y + \sigma^2 \nabla \log f(y), \tag{5}$$

其中 $f(y)$ 是平滑分布的密度。尽管未知 $f(y)$，但其得分 $\nabla \log f(y)$ 可以通过神经网络近似得到。

网络通过去噪目标进行训练，该目标旨在促使模型从噪声版本中准确重建干净样本：

$$L(\phi) = \mathbb{E}_{x \sim p(x),\, y \sim p(y|x)} \left\| x - \hat{x}_\phi(y) \right\|^2, \tag{6}$$

其中 $p(y|x) = \mathcal{N}(x,\ \sigma^2)$ 并且

$$\hat{x}_\phi(y) = y + \sigma^2 g_\phi(y),$$

其中 $g_\phi(y)$ 作为得分函数 $\nabla \log f(y)$ 的学习型近似。

离散行走-跳跃采样（dWJS）方法 [8] 将 NEB 应用于蛋白质和抗体序列生成。每个氨基酸序列都表示为拼接的独热编码：对于长度为 297 的序列，这会得到一个维度为 297×21 的二进制向量。向这种连续松弛表示中添加高斯噪声，使模型能够学习数据流形的局部结构。

训练完分数模型后，采样分两个阶段进行。首先将一个随机序列转换为其独热表示，并加入高斯噪声进行扰动，生成连续嵌入。随后，通过沿学习到的得分场上升来迭代更新样本，这就是“行走”步骤，使其沿估计的能量景观移动。经过 T 次这样的细化步骤后，通过在每个位置选取概率最高的氨基酸，将更新后的嵌入投影回离散域，从而生成有效的蛋白质序列。这个投影过程就是“跳跃”步骤。

使用 σ 条件化的分数模型进行采样

输入。A σ 条件化的分数网络 $g_\phi(y,\ \sigma)$；经验噪声水平分布 $\hat{p}(\sigma)$；步长 δ；迭代次数 T。

输出。从学习到的数据分布中抽取一个样本 $\hat{x}$ 。**过程**。

1. **采样噪声尺度**

$$\sigma \sim \hat{p}(\sigma).$$

2. **初始化潜变量**

$$y_0 \sim \mathcal{N}(0, \sigma^2 I_d).$$

3. 迭代朗之万更新。 对于 $t = 0,\ldots, T-1$，根据以下公式更新 y_t

$$y_{t+1} = y_t + \delta\, g_\phi(y_t, \sigma) + \sqrt{2\delta}\, \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_d).$$

4. **去噪步骤**。

$$\hat{x} = y_T + \sigma^2 g_\phi(y_T, \sigma).$$

与几何引导的关系

我们的框架在理论上可与无分类器引导（CFG）相关联 [26]，但有两个关键区别，使其适用于无监督流形学习。

标准 CFG 更新通过在无条件得分 $\epsilon_\theta(y_t)$ 与条件得分 $\epsilon_\theta(y_t, c)$ 之间进行外推，来修改得分估计 $\tilde{\epsilon}$（其中 c 是类别标签）：

$$\tilde{\epsilon}_\theta(y_t, c) = \epsilon_\theta(y_t) + w \cdot (\epsilon_\theta(y_t, c) - \epsilon_\theta(y_t)) \tag{7}$$

其中 w 是引导尺度。通常，使用 $w > 1$ 来增强条件概率，但会以牺牲多样性为代价。

我们的密度依赖平滑（DDS）用连续噪声尺度 $\sigma(x)$ 替换离散类别标签 c，该噪声尺度作为局部几何密度的代理。我们的采样过程直接使用条件得分模型 $g_\phi(y_t, \sigma)$。这在数学上等价于引导权重固定为 $w = 1$ 的 CFG 公式（标准条件采样）：

$$\tilde{g}_\phi(y_t) = g_\phi(y_t, \sigma) \tag{8}$$

这一选择确保模型忠实地遵循所学几何，而不会引入通常与高引导采样相关的模式崩塌伪影（$w > 1$）。

此外，我们将“条件类别”的概念从离散集合扩展到**连续谱**。标准 CFG 以离散标签$c \in \{1, \ldots, K\}$为条件，而我们的模型以连续噪声尺度 $\sigma \in \mathbb{R}^+$ 为条件。通过将 σ 视为局部密度的代理，我们实际上定义了**任意数量的几何类别**，范围从“高度密集”到“稀疏”。这使得得分函数能够在蛋白质序列的异质能量景观中平滑适应。

评估指标

为评估生成抗体序列的质量，我们使用一组多样化的指标进行评估，这些指标涵盖唯一性、多样性、新颖性、与留出数据的编辑距离以及分布相似性。

• **唯一性（U）**。唯一性衡量生成序列中彼此不同的比例。该值越高，表明模型越能避免生成冗余或重复的样本。形式上，若 $\mathcal{D}$ 是生成序列的集合，Unique($\mathcal{D}$) 表示不同元素的数量，则

$$U = \frac{\text{Unique}(\mathcal{D})}{|\mathcal{D}|}.$$

• **内部多样性（ID）**。内部多样性用于量化生成样本彼此之间的差异。序列集合具有多样性是理想的，因为这意味着它们探索了更广阔的序列空间。按照以往研究，我们使用平均成对莱文斯坦距离来计算多样性：

$$\text{ID}(\mathcal{D}) = \frac{\sum_{x_i \in \mathcal{D}} \sum_{x_j \in \mathcal{D} \setminus \{x_i\}} d(x_i, x_j)}{|\mathcal{D}|(|\mathcal{D}| - 1)},$$

其中，$d(\cdot, \cdot)$ 是莱文斯坦距离。数值越大，表示生成序列之间的多样性越高。

•**编辑距离（ED）**。编辑距离用于衡量生成序列与验证集 $\mathcal{D}_0$ 中样本之间的差异。该指标评估模型生成的序列是否具有新颖性，且未出现在已知数据集中。我们将 ED 计算为每个生成序列到 $\mathcal{D}_0$ 中任意序列的最小距离的平均值：

$$\text{ED}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \min_{s_j \in \mathcal{D}_0} d(x_i, s_j).$$

ED 值越高，表示新颖性越强。

• **DCS（分布一致性得分）**。DCS [8] 与图像领域的 Fr´echet Inception Distance（FID）类似。它用于评估生成序列与参考分布的一致程度，同时考量新颖性和多样性。DCS 值越高，表示与真实分布的匹配程度越好。

• **Wasserstein距离（WD）**。Wasserstein距离衡量生成序列与验证集生化性质分布的匹配程度。对于每条序列，我们使用 `BioPython` [27]计算其生化性质（例如，不稳定性指数和芳香性）。然后，我们计算生成集与验证集的性质分布之间的平均Wasserstein距离：

$$\text{WD} = W(\mathcal{P}_{\text{gen}}, \mathcal{P}_{\text{val}}).$$

较低的WD表示生成序列呈现出真实的生化特征。

• **KS统计量** 它衡量经验累积分布函数之间的最大差异：

$$\text{KS} = \max_x |F_{\text{real}}(x) - F_{\text{gen}}(x)|,$$

其中， F_{real} 和 F_{gen} 分别表示真实和生成的免疫原性评分的累积分布。

结论

更广泛地说，我们的研究结果表明，尊重生物序列空间的异质几何结构并非实现细节，而是实现可靠生成式建模的先决条件。固定全局平滑假设——当前基于扩散与语言模型的生成器普遍采用这类假设——会系统性地扭曲稀疏的生物流形，导致稠密功能簇过度平滑，或稀疏区域发生碎片化。诸如 **密度依赖平滑（DDS）** 这样的几何感知生成机制提供了一种有原则的替代方案，能够在固定噪声模型根本无法处理的情形下，实现可扩展且可靠的分子设计。

在受控的多模态玩具基准测试中，DDS 在提升保真度的同时保持了探索能力，并以少得多的虚假生成恢复了真实模式（包括留出的模式）。在真实生物设计任务中，包括大规模抗体库生成、CDR 突变体生成、抗菌肽设计和冠状病毒抗体生成，与固定-σ 基线方法相比，DDS 始终实现了更优的质量与多样性权衡。重要的是，结合结构预测和蛋白质语言模型评分的互补评估，以及基于 BLAST 的 [16]同源性分析表明，固定的大幅度平滑下较高的有效性可能源于记忆，而 DDS 能更好地平衡结构合理性与进化新颖性。最后，在 CoV-AbDab 上，DDS 在保持较强新颖性和多样性的同时，与真实免疫原性分布更加吻合，这表明密度感知噪声缩放能够提升安全相关属性的分布保真度。

我们预期，几何感知平滑将成为未来生物分子生成模型的一项基础设计原则。通过使随机平滑与稀疏且异质的适应度景观的内在几何结构相一致，DDS 等方法为实现可扩展、可靠且注重安全性的分子设计开辟了道路，使其不再受固定噪声生成框架所施加的限制。

局限性与未来工作

尽管 DDS 提供了一种简单有效的方法，我们的方法仍存在一些局限性，这些局限性也为未来研究指明了方向。我们的方法依赖于为缩放每个样本的 σ 值预先设定一个范围。上下界的选择可能会影响模型性能。如何以更有原则或自适应的方式设定这些界限，仍是一个有待解决的问题。

我们目前的研究主要是实证性的。尽管在合成数据集和生物分子序列设计任务上的结果强烈表明，自适应数据的 σ 能够提高生成样本的质量，但对于自适应噪声缩放为何以及何时能够更好地近似底层得分函数，我们仍缺乏理论理解。我们希望通过形式化分析拓展这项工作，并有望建立

将局部密度估计、 σ 选择与得分近似的准确性联系起来的误差界。

参考文献

[1] Altman, N., Krzywinski, M.: 维度的诅咒。《自然·方法》**15**(6), 399–400 (2018)[2] Romero, P.A., Arnold, F.H.: 通过定向进化探索蛋白质适应度景观。《自然综述：分子细胞生物学》**10**(12), 866–876 (2009)[3] Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., 等: 一种用于设计分支酸变位酶的基于进化的模型。《科学》**369**(6502), 440–445 (2020)[4] Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H., Ranganathan, R.: 用于确定蛋白质折叠的进化信息。《自然》**437**(7058), 512–518 (2005)[5] Bloom, J.D., Labthavikul, S.T., Otey, C.R., Arnold, F.H.: 蛋白质稳定性促进可进化性。《美国国家科学院院刊》**103**(15), 5869–5874 (2006)[6] Wittmann, B.J., Yue, Y., Arnold, F.H.: 知情的训练集设计可实现高效的机器学习辅助定向蛋白质进化。《细胞系统》**12**(11), 1026–1045 (2021)[7] Buchholz, P.C., Zeil, C., Pleiss, J.: 蛋白质序列空间的无标度特性。《公共科学图书馆·综合》**13**(8), 0200815 (2018)[8] Frey, N.C., Berenberg, D., Zadorozhny, K., Kleinhenz, J., Lafrance-Vanasse, J., Hotzel, I., Wu, Y., Ra, S., Bonneau, R., Cho, K., 等：通过离散行走-跳跃采样发现蛋白质。arXiv 预印本 arXiv:2306.12360 (2023)[9] Ikram, Z., Liu, D., Rahman, M.S.: 用于生物序列生成的梯度引导离散行走-跳跃采样。《机器学习研究汇刊》(2024)[10] Song, Y., Ermon, S.: 通过估计数据分布的梯度进行生成式建模。《神经信息处理系统进展》**32** (2019)[11]Kingma, D., Salimans, T., Poole, B., Ho, J.: 变分扩散模型。《神经信息处理系统进展》**34**, 21696–21707 (2021)[12] De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y.W., Doucet, A.: 黎曼得分生成建模。《神经信息处理系统进展》**35**, 2406–2422 (2022)

[13]Olsen, T.H., Boyles, F., Deane, C.M.: Observed Antibody Space：一个经过清理、注释和翻译的非配对及配对抗体序列多样化数据库。Protein Science 31(1), 141–146 (2022)[14] Pirtskhalava, M., Amstrong, A.A., Grigolava, M.,Chubinidze, M.,Alimbarashvili, E., Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D.E., Tartakovsky, M.: Dbaasp v3：抗菌/细胞毒活性及肽结构数据库，可作为开发新型治疗药物的资源。Nucleic Acids Research 49(D1), 288–297 (2021)[15] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., S metanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., 等：利用语言模型进行进化尺度的原子级蛋白质结构预测。Science 379(6637), 1123–1130 (2023)[16] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: 基本局部比对搜索工具。Journal of Molecular Biology 215(3), 403–410 (1990)[17] Rahimi, A., Recht, B.: 面向大规模核机器的随机特征。《神经信息处理系统进展》20 (2007)[18] Saremi, S., Mehrjou, A., Sch¨olkopf, B., Hyv¨arinen, A.: 深度能量估计网络。arXiv 预印本 arXiv:1805.08306 (2018)[19] Shuai, R.W., Ruffolo, J.A., Gray, J.J.: 用于抗体设计的生成式语言建模。bioRxiv, 2021–12 (2021)[20] Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S.M., Ehling, R.A., Bonati, L., Dahinden, J., Gainza, P., 等：通过深度学习根据抗体序列预测抗原特异性，以优化治疗性抗体。Nature Biomedical Engineering 5(6), 600–612 (2021)[21] Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B.F., Ekbote, C.A., Fu, J., Zhang, T., Kilgour, M., Zhang, D., 等：利用 GFlowNets 进行生物序列设计。载于：国际机器学习会议，第9786–9801页 (2022)。PMLR[22] Angermueller, C., Dohan, D., Belanger, D., Deshpande, R., Murphy, K., Colwell, L.: 用于生物序列设计的基于模型的强化学习。载于：国际学习表征会议 (2019)[23] Vishakh, P., Chen, Y.-H., Jane, P., Valerie, C., He, H.: 将大语言模型新颖性作为原创且高质量输出的前沿进行衡量 (2025)。arxiv.org[24] Raybould, M.I., Kovaltsuk, A., Marks, C., Deane, C.M.: CoV-AbDab：冠状病毒抗体数据库。Bioinformatics 37(5), 734–735 (2021)[25] Saremi, S., Hyv¨arinen, A.: 神经经验贝叶斯。Journal of Machine Learning

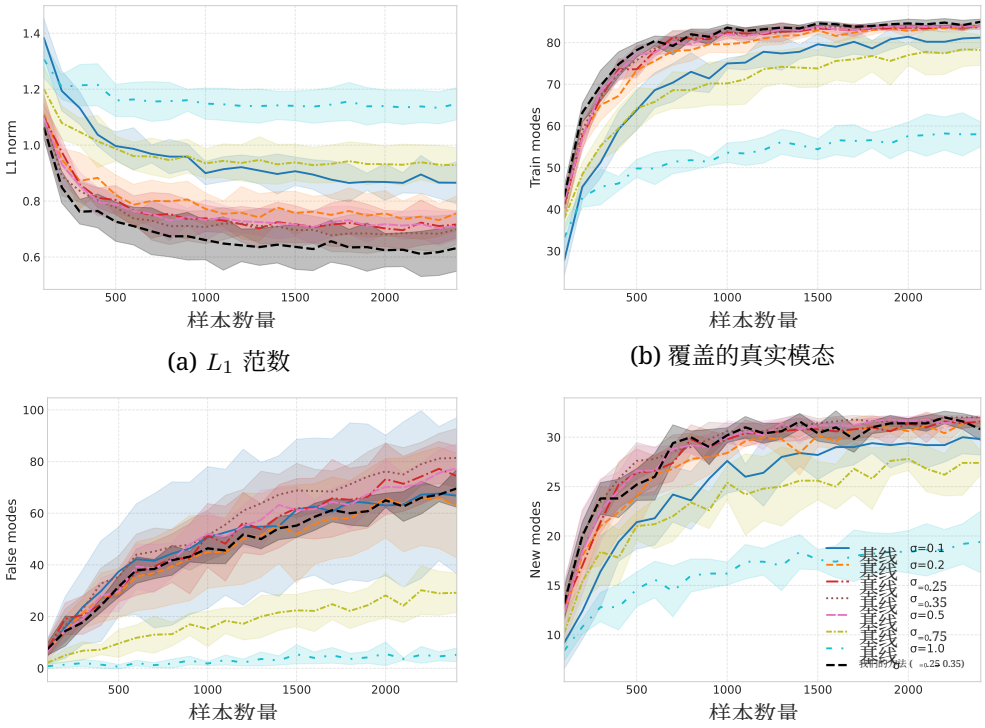
Research **20**(181), 1–23 (2019)

[26] Ho, J., Salimans, T.：无分类器扩散引导。载于：NeurIPS 2021 深度生成模型与下游应用研讨会[27] Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox,C.J., Dalke,A., Fried- berg, I., Hamelryck, T., Kauff, F., Wilczynski, B., 等：Biopython：用于计算分子生物学和 Bioinformatics 的免费 Python 工具。Bioinformatics **25**(11), 1422 (2009)

附录

玩具数据集上的扩展基线扫描。

在正文中，为了让图 σ 3易于阅读，我们比较了精简的一组数值；此处，我们在图1中列出了针对固定平滑尺度的完整基线扫描 $\sigma \in \{0.1, 0.2, 0.25, 0.35, 0.5, 0.75, 1.0\}$。这一扩展视图证实，基线性能对平滑尺度的选择很敏感：过小的 σ 会导致探索过度并引入伪模态，而过大的 σ 则会过度平滑，无法充分覆盖真实模态。DDS 在整个扫描范围内始终具有竞争力，避免了这两种失效模式，也减少了精确 σ 调优的必要性。



(a) L_1 范数

(b) 覆盖的真实模态

(c) 覆盖的错误模态

(d) 覆盖的新模态

图1：合成4D数据集上的性能