基于表面的多模态流匹配分子设计

Fang Wu
fangwu97@stanford.edu 斯坦福大学 美国加利福尼亚州斯坦福

Zhengyuan Zhou
zzhou@ucsd.edu 加利福尼亚大学圣迭戈分校 美国加利福尼亚州拉霍亚

Shuting Jin
shutingjin@wust.edu.cn 武汉科技大学 中国湖北省武汉市

Xiangxiang Zeng
xzeng@foxmail.com 湖南大学 中国湖南省长沙市

Jure Leskovec
jure@cs.stanford.edu 斯坦福大学 美国加利福尼亚州斯坦福

Jinbo Xu†
jinboxu@moleculemind.com MoleculeMind 中国北京

arXiv:2601.04506v1 [cs.LG] 8 Jan 2026

摘要

治疗性肽在靶向以往难以成药的结合位点方面展现出巨大潜力，近年来深度生成模型的发展使针对特定蛋白质受体进行全原子肽协同设计成为可能。然而，分子表面在蛋白质相互作用（PPIs）中的关键作用尚未得到充分探索。为弥补这一空白，我们提出了一种全模态设计肽生成范式，称为SurfFlow。这是一种新颖的基于表面的生成算法，能够对肽的序列、结构和表面进行全面协同设计。SurfFlow 采用多模态条件流匹配（CFM）架构，学习表面几何形状和生化性质的分布，从而提升肽的结合准确性。在综合性 PepMerge 基准测试中，SurfFlow 在所有指标上均持续优于全原子基线方法。这些结果凸显了在从头肽发现中考虑分子表面的优势，并展示了整合多种蛋白质模态以更有效地发现治疗性肽的潜力。

CCS概念分类

• **应用计算** → 分子结构生物学； • **计算方法** →神经网络。

关键词

deep生成模型，几何表示学习 习

ACM参考文献格式：

Fang Wu、Zhengyuan Zhou、Shuting Jin、Xiangxiang Zeng、Jure Leskovec 和 Jinbo Xu†。2025。基于表面的多模态流匹配分子设计。载于第*31*届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集 *V.2*（*KDD '25*），*2025*年*8*月*3*日至*7*日，加拿大安大略省多伦多。 ACM，纽约州纽约市，12页。https://doi.org/10.1145/ 3711896.3737139

允许为个人或课堂使用而免费制作本作品的全部或部分数字副本或纸质副本，前提是不得为营利或商业利益而制作或分发副本，且副本首页须附有本声明及完整引文。作者以外的其他权利人所拥有的本作品组成部分，其版权必须得到尊重。注明出处后，允许摘录。若要以其他方式复制、再版、发布到服务器或向列表重新分发，则须事先获得具体许可和/或支付费用。请通过permissions@acm.org 申请许可。*KDD '25*, 加拿大安大略省多伦多© 2025 版权所有归所有者/作者所有。出版权授权予 ACM。ACM ISBN 979-8-4007-1454-2/2025/08 https://doi.org/10.1145/3711896.3737139

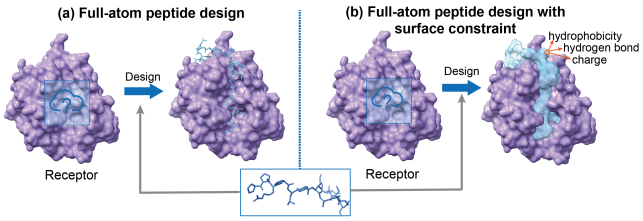


图 1：有无表面约束时全原子肽设计的比较。

1 引言

肽是由大约2至50个氨基酸组成的短链蛋白质，在细胞信号传导、酶催化和免疫反应等多种生物过程中发挥关键作用 [60]。它们能够以高亲和力和特异性结合细胞表面受体，因此是药理学中的重要介质，可产生细胞内效应，同时具有低毒性、低免疫原性且易于递送等特点 [44]。传统的肽发现模拟或搜索方法依赖于频繁计算物理能量函数，但庞大的设计空间给这一过程带来了阻碍 [5]。这促使人们日益需要计算方法来促进计算机辅助 肽设计与分析。

近年来，Diffusion [21] 和流 [36] 模型展现出在分子设计 [20], 、抗体工程 [40,66], 、蛋白质设计 [68, 69] 以及肽发现 [47, 65]领域的巨大潜力。游离肽通常处于高能量、高熵状态，构象不稳定，只有与目标受体结合后才会发挥功能。因此，肽设计必须明确以结合口袋 [59]为条件。此外，由于残基通过侧链基团形成的非共价力彼此相互作用，研究者正日益致力于利用全原子几何结构 [26, 34],捕捉蛋白质-肽相互作用，从而推动序列-结构协同设计超越单纯的骨架设计。

与此同时，研究蛋白质-蛋白质相互作用（PPIs）时，人们日益关注分子表面，因为 PPIs 在很大程度上取决于相互作用蛋白质的表面如何彼此互补 [25, 52]。表面的静电势和疏水性是决定 PPIs 强度与特异性的关键因素 [23, 28, 31], , 其几何形态（如凸起、沟槽和裂隙）则使锁钥机制 [43] 或诱导契合机制得以发挥作用，这些机制对于特异性结合至关重要。这些表面构成了一个基本的

界面，决定蛋白质如何识别并结合彼此。基于这些原因，肽生成过程中同时考虑所有分子模态——序列、结构和表面——至关重要（见图 1），从而提升各个方面的一致性，这就是我们所称的全模态设计。

为实现这一目标，我们提出了一种新颖的全模态设计生成算法 SurfFlow。该算法将多模态流匹配（FM） [1, 2, 36] 应用于内部结构和分子表面，这些结构和表面由表面点位置及单位法向量表示为 SE(3) 中的刚性坐标系。仅有互补的表面几何形状并不能保证成功结合；要实现精确结合，还必须在结合界面正确布置电荷、极性和疏水性 [16],，因此 SurfFlow 纳入了这些生化性质约束。具体而言，它利用离散流匹配（DFM） [9]，通过连续时间马尔可夫链（CTMC）在离散数据空间中处理一些表面分类特征。此外，针对如何以可扩展的方式捕捉不规则表面几何形状、多尺度特征和蛋白质间相互作用这一挑战，我们提出了等变表面几何网络（ESGN），该网络可动态建模异构表面图，并独特地同时纳入表面内部和表面之间的相互作用。最后，鉴于环化性和二硫键等肽的关键属性会影响其稳定性和结合亲和力 [8],，我们将这些因素作为额外条件，以增强 SurfFlow 的能力和泛化性。我们在综合性肽设计基准测试 PepMerge [33],上评估了 SurfFlow，实验表明，它在所有指标上均持续优于全原子基线方法，凸显了在从头肽设计中考虑表面的优势。

2 预备知识与背景

蛋白质由多个氨基酸组成，每个氨基酸由其类型、主链坐标系和侧链扭转角 [14] 定义。第 i 个残基的类型记为 $a_i \in \{1 \dots 20\}$，由其侧链 R基团决定。每个残基的刚性坐标系由 4 个主链重原子 N、Cα、C 和 O 的坐标构建，其中 Cα 位于原点。该坐标系由位置向量 $\boldsymbol{x}_i \in \mathbb{R}^3$ 和旋转矩阵 $O_i \in \mathrm{SO}(3)$ [24]表示。与骨架不同，侧链具有更大的柔性，其侧链原子之间最多包含 4 个可旋转扭转角，记为 $\boldsymbol{\chi}_i \in [0, 2\pi)^4$。此外，主链扭转角 $\varphi_i \in [0, 2\pi)$ 会影响氧原子的位置。

我们进一步考虑分子表面：沿蛋白质移动一个特定半径（约 1 Å）的探针，以计算溶剂可及表面（SAS）和溶剂排斥表面（SES）。探针的坐标将表面定义为定向点云 $Q = \{q_i : 1 \le i \le m\}$。每个点 q_i 都具有相关属性（$\boldsymbol{x}_i^s$，$\boldsymbol{n}_i^s$，$\boldsymbol{\tau}_i^s$，Υ_i^s），其中 $\boldsymbol{x}_i^s \in \mathbb{R}^3$ 是 3D 坐标，$\boldsymbol{n}_i^s \in \mathbb{R}^3$ 是单位法向量。$\boldsymbol{\tau}_i^s \in \mathbb{R}^{\psi_\tau}$ 和 $\Upsilon_i^s \in \mathbb{R}^{\psi_\Upsilon}$ 表示其连续和分类的物理化学性质，例如疏水性、氢键和电荷 [15, 52]。

本研究基于目标蛋白质设计肽。形式上，给定一个含有 n_{pep} 个残基的肽 C^{pep} 和一个含有 n_{rec} 个残基的目标蛋白质 C^{rec}，我们的目标是对条件联合分布 $p(C^{\text{pep}} \mid C^{\text{rec}})$进行建模。受体可以用 $C^{\text{rec}} = \{(a_i, O_i, \boldsymbol{x}_i, \boldsymbol{\chi}_i)\}_{i=1}^{n_{\text{rec}}}$充分且简洁地参数化，其中 $\boldsymbol{\chi}_i[0] = \varphi_i$ 以及 $\boldsymbol{\chi}_i \in [0, 2\pi)^5$。同时也描绘了配体肽表面，得到 $C^{\text{pep}} = \left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}} \cup \left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^{m}$（含有$m \gg n_{\text{pep}}$）。实际上，可以使用 PyMol [12]或 MSMS [49]等软件计算蛋白质的原始分子表面。

概率路径与流。设 $\mathbb{P}(\mathrm{M})$ 为流形 M 上、具有黎曼度量 g 的概率分布空间。概率路径 p_t： $[0,1] \times \mathrm{M} \to \mathbb{P}(\mathrm{M})$ 是两个分布 p_0，$p_1 \in \mathbb{P}(\mathrm{M})$ 之间由连续参数 t 索引的插值。

M 上的流由一个单参数微分同胚Φ： $\mathrm{M} \to \mathrm{M}$ 定义，它是对由时变向量场 $u_t \in \mathrm{U}$ 描述的瞬时变形进行积分得到的结果。$u_t(x) \in \mathrm{T}_x\mathrm{M}$ 是路径 p_t在 x时刻 t 的梯度向量。通过在 M 上、在 $t \in [0,1]$ 范围内求解以下常微分方程（ODE），并以 $\phi_0(x) = x$ 为初始条件：$\mathrm{d}\phi_t\,\mathrm{d}t\,(x) = u_t(\phi_t(x))$。我们得到时变流 ϕ_t： $\mathrm{M} \to \mathrm{M}$，并通过设定 $\Phi(x) = \phi_1(x)$ 得到最终的微分同胚。给定源密度 p_0，$\phi_t(x)$ 会诱导一个推前运算 $p_t = [\phi_t]_{\#}p_0$。它沿着 $u_t(x)$ 将点密度 $x \sim p_0$ 重塑为更复杂的密度，而变量变换算子 # 定义为 $[\phi_t]_{\#}p_0(x) = p_0\left(\phi_t^{-1}(x)\right)\det\left[\frac{\partial \phi_t^{-1}}{\partial x}(x)\right]$。时变密度 p_t 由福克-普朗克方程刻画：$\mathrm{d}p_t\,\mathrm{d}t = -\mathrm{div}(u_t p_t)$。在这些条件下，$u_t$是 p_t的概率流，而 p_t是由 u_t生成的边缘概率路径。流匹配（FM）通过拟合带参数 θ 的向量场$v_\theta \in \mathrm{U}$，使其匹配能够产生概率路径 p_t 的目标向量场 u_t，从而训练连续归一化流（CNF）。其目标位于切空间中，形式如下：

$$\mathcal{L}_{\text{RFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1], x\sim p_t(x)} \|v_\theta(x,t) - u_t(x)\|_g^2, \tag{1}$$

由于 u_t难以求解，我们使用条件向量场 $u_t(x|x_1)$构造 $p_t(x|x_1)$。目标函数变为：$\mathrm{L}_{\text{CRFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\ x_1\sim p_1(x_1)\ x\sim p_t(x|x_1)} \|v_\theta(x,\ t) - u_t(x|x_1)\|_g^2$，它与 L_{RFM} [36] 具有相同的梯度。在推断过程中，可以求解与神经向量场 v_θ 相关的常微分方程，在一段时间内将 $x_0 \in \mathrm{M}$从 p_0 推至 p_1。

连续时间马尔可夫链。连续时间马尔可夫链 [45] 是一类离散的连续时间随机过程，与概率流密切相关。假设分类变量 x 有 S 个状态，其轨迹 x_t 随时间 $t \in [0,1]$ 遵循连续时间马尔可夫链，x_t 在停留于当前状态与定期跳转至另一个随机选取的状态之间交替。跳转的频率和目的地由速率矩阵 $R_t \in \mathbb{R}^{S\times S}$ 决定，其非对角元素均为非负数。接下来一个无穷小时间步 $\mathrm{d}t$内，跳转至不同状态 j的概率 x_t为 $R_t(x_t,\ j)\mathrm{d}t$。转移概率写作：

$$\begin{aligned} p_{t+\mathrm{d}t|t}\left(j \mid x_t\right) &= \begin{cases} R_t\left(x_t, j\right)\mathrm{d}t & \text{for } j \neq x_t \\ 1 + R_t\left(x_t, x_t\right)\mathrm{d}t & \text{for } j = x_t \end{cases} \\ &= \delta\left\{x_t, j\right\} + R_t\left(x_t, j\right)\mathrm{d}t, \end{aligned} \tag{2}$$

其中 $\delta\{i,j\}$ 是克罗内克δ。$\delta\{i,j\}$ 在 $i = j$ 时为 1，否则为 0。$R_t(x_t, x_t) := -\sum_{k\neq x} R_t(x_t, k)$，以使 $p_{t+\mathrm{d}t|t}(\cdot \mid i)$ 的总和为 1。使用紧凑记法，$p_{t+\mathrm{d}t|t}$ 因此是一个分类分布，其概率为 $\delta\{x_t,\cdot\} + R_t(x_t,\cdot)\,\mathrm{d}t$，记作 $\mathrm{Cat}\left(\delta\{x_t, j\} + R_t(x_t, j)\,\mathrm{d}t\right)$。也就是说，$j \sim p_{t+\mathrm{d}t|t}(j \mid x_t) \Longleftrightarrow j \sim$

$\mathrm{Cat}\left(\delta\{x_t, j\} + R_t(x_t, j)\,\mathrm{d}t\right)$。实际上，我们需要在有限的时间间隔 Δt 内模拟序列轨迹。可以使用欧拉步进 [56] 来模拟轨迹。

$$x_{t+\Delta t} \sim \mathrm{Cat}\left(\delta\{x_t, x_{t+\Delta t}\} + R_t\left(x_t, x_{t+\Delta t}\right)\Delta t\right), \tag{3}$$

其中，变量 x 在时间 $t = 0$ 从初始样本 $x_0 \sim p_0$ 开始。速率矩阵 R_t 与初始分布 p_0 共同定义了连续时间马尔可夫链。令时间 t 的边缘分布为 $p_t(x_t)$，柯尔莫哥洛夫方程使我们能够将速率矩阵 R_t 与 $p_t(x_t)$ 的变化联系起来：

$$\partial_t p_t\left(x_t\right) = \underbrace{\sum_{j \neq x_t} R_t\left(j, x_t\right) p_t(j)}_{\text{incoming}} - \underbrace{\sum_{j \neq x_t} R_t\left(x_t, j\right) p_t\left(x_t\right)}_{\text{outgoing}}. \tag{4}$$

流入与流出的概率质量之差，就是边缘分布 $\partial_t p_t(x_t)$ 的时间导数。随后，我们得到 $\partial_t p_t = R_t^\top p_t$，其中边缘分布被视为概率质量向量：$p_t \in [0,1]^S$，并在向量空间中定义一个常微分方程。若概率路径 p_t 满足 $\partial_t p_t = R_t^\top p_t$（对 $\forall t \in [0,1]$ [9]），则称其由 R_t 生成。

3 方法

分子表面同时 encapsulates 蛋白质在欧几里得空间中的三维几何结构及其生化属性，例如疏水性和电荷 [15]。表面形状与这些生化性质之间的相互作用对于定义蛋白质的功能至关重要。给定具有特定几何和生化约束的目标受体，SurfFlow 可同时生成肽的内部结构和外部表面。此外，它还可以考虑环化性和二硫键等关键因素（见图 2）。

3.1 用于表面生成的流匹配

采用条件流匹配框架，根据其目标蛋白质 $p\left(C^{\mathrm{pep}} \mid C^{\mathrm{rec}}\right)$ 学习条件肽分布。该联合概率在经验上被分解为内部结构元素概率与外部表面元素概率的乘积：

$$\begin{aligned} p\left(C^{\mathrm{pep}} \mid C^{\mathrm{rec}}\right) \propto\; & p\left(\left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\mathrm{pep}}} \middle| C^{\mathrm{rec}}\right) \\ & p\left(\left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right), \end{aligned} \tag{5}$$

其中，$p\left(\left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right)$ 进一步被拆分为四个基本元素 $p\left(\left\{\boldsymbol{x}_i^s\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right)$、$p\left(\left\{\boldsymbol{n}_i^s\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right)$ 的乘积，$p\left(\left\{\boldsymbol{\tau}_i^s\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right)$，以及 $p\left(\left\{\Upsilon_i^s\right\}_{i=1}^m \middle| C^{\mathrm{rec}}\right)$。表面点的位置 $p\left(\boldsymbol{x}_i^s \mid C^{\mathrm{rec}}\right)$、方向 $p\left(\boldsymbol{n}_i^s \mid C^{\mathrm{rec}}\right)$、连续属性 $p\left(\boldsymbol{\tau}_i^s \mid C^{\mathrm{rec}}\right)$和类别属性 $p\left(\Upsilon_i^s \mid C^{\mathrm{rec}}\right)$ 上的不同概率流的构造详述如下。

理化特征。流匹配通常应用于连续空间。然而，某些生化特征具有离散的分类取值。例如，可以根据每个点的氢键能力将其分类为供体、受体或中性。这一挑战也出现在蛋白质生成任务中，这些任务仅关注结构 [39],，其中残基类型遵循分类分布。为解决这一问题，先前的研究 [33, 34] 采用软独热编码，将分类分布映射到概率单纯形，或直接将流匹配应用于多项分布。然而，这种直接的方法在蛋白质协同设计中可能导致性能欠佳。近期，研究者提出了更先进的、专为离散空间设计的流匹配方法 [9, 19, 53]，以克服这些局限。

离散流模型（DFM）是为在离散空间中运行而设计的生成算法，它通过模拟从噪声过渡到数据的概率流来实现这一点。DFM 沿着与噪声到数据流一致的离散变量轨迹运行，从而生成新的样本。基于 Campbell 等人 [9], 和 Gat 等人 [19], 的工作，我们使用连续时间马尔可夫链为生化性质 $\Upsilon^s \in \mathbb{R}^{m \times \psi_\Upsilon}$ 实现了一个 DFM。具体而言，我们训练一个神经网络 $v^{\mathrm{cat}}(\cdot)$ 来近似真实的去噪分布 $p_{1|t}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)$。训练采用交叉熵损失，使模型在给定带噪输入 $\Upsilon_t^s \sim p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)$ 时，学习预测干净的数据点 Υ_1^s：

$$\mathcal{L}_{\mathrm{cat}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,1), p\left(\Upsilon_1^s\right), p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)}\left[\log v^{\mathrm{cat}}\left(\Upsilon_t^s, t, C^{\mathrm{rec}}\right)\right]. \tag{6}$$

这里，我们不采用从均匀先验分布 $p_{t|1}^{\mathrm{unif}}\left(\Upsilon_t^s \mid \Upsilon_1^s\right) = \text{分类分布}\left(t\delta\{\Upsilon_1^s, \Upsilon_t^s\} + (1-t)\frac{1}{S}\right)$向 Υ_1^s 进行线性插值的方式，而是引入一个人工掩码状态 M，此时条件路径变为 [9]:

$$p_{t|1}^{\mathrm{mask}}\left(\Upsilon_t^s \mid \Upsilon_1^s\right) = \mathrm{Cat}(t\delta\{\Upsilon_1^s, \Upsilon_t^s\} + (1-t)\delta\{\mathrm{M}, \Upsilon_t^s\}). \tag{7}$$

值得注意的是，$\mathrm{L}^{\mathrm{cat}}(\cdot)$ 与用于训练扩散模型 [9] 的模型对数似然和证据下界（ELBO）密切相关。它也不依赖于 $R_t\left(\Upsilon_t^s, j \mid \Upsilon_1^s\right)$。有许多种 $R_t\left(\Upsilon_t^s, j \mid \Upsilon_1^s\right)$ 的选择，它们都能生成相同的 $p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)$。在推理时，我们为 $\Upsilon_t^s \neq j$ 选择的速率矩阵如下：

$$R_t^*\left(\Upsilon_t^s, j \mid \Upsilon_1^s\right) := \frac{\mathrm{ReLU}\left(\partial_t p_{t|1}\left(j \mid \Upsilon_1^s\right) - \partial_t p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)\right)}{S \cdot p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right)}, \tag{8}$$

其中 $\mathrm{ReLU}(a) = \max(a, 0)$，$\partial_t p_{t|1}$可通过对方程 7 中 $p_{t|1}$的显式形式求导得到。这种对 R_t^*的选择假设 $p_{t|1}\left(\Upsilon_t^s \mid \Upsilon_1^s\right) > 0$。

位置。我们采用欧几里得条件流匹配来生成表面点位置 $\boldsymbol{x}^s \in \mathbb{R}^{m \times 3}$。我们使用标准的各向同性高斯分布 $\mathrm{N}(0, I_3)$ 作为先验分布，其中目标分布为 $p(\boldsymbol{x}^s \mid C^{\mathrm{rec}})$。条件流定义为采样噪声 $\boldsymbol{x}_0^s \sim \mathrm{N}(0, I_3)$ 与数据点 $\boldsymbol{x}_1^s \sim p(\boldsymbol{x}^s \mid C^{\mathrm{rec}})$ 之间的线性插值。这种线性插值确保了直线路径，通过沿着欧几里得空间 [38]中两点之间的最短路径移动，提高了训练与采样效率。通过使用独立耦合技术，对线性流 ϕ_t^{pos} 取时间导数可得到条件向量场 u_t^{pos}：

$$\phi_t^{\mathrm{pos}}\left(\boldsymbol{x}_0^s, \boldsymbol{x}_1^s\right) = t\boldsymbol{x}_1^s + (1-t)\boldsymbol{x}_0^s, \tag{9}$$

$$u_t^{\mathrm{pos}}\left(\boldsymbol{x}_t^s \mid \boldsymbol{x}_1^s, \boldsymbol{x}_0^s\right) = \boldsymbol{x}_1^s - \boldsymbol{x}_0^s = \frac{\boldsymbol{x}_1^s - \boldsymbol{x}_t^s}{1-t}. \tag{10}$$

我们使用时间依赖的平移不变表面网络 $v^{\mathrm{pos}}(\cdot)$，根据当前插值态 $\boldsymbol{x}_t$ 和时间步 t预测条件向量场。表面点云的 CFM 目标函数

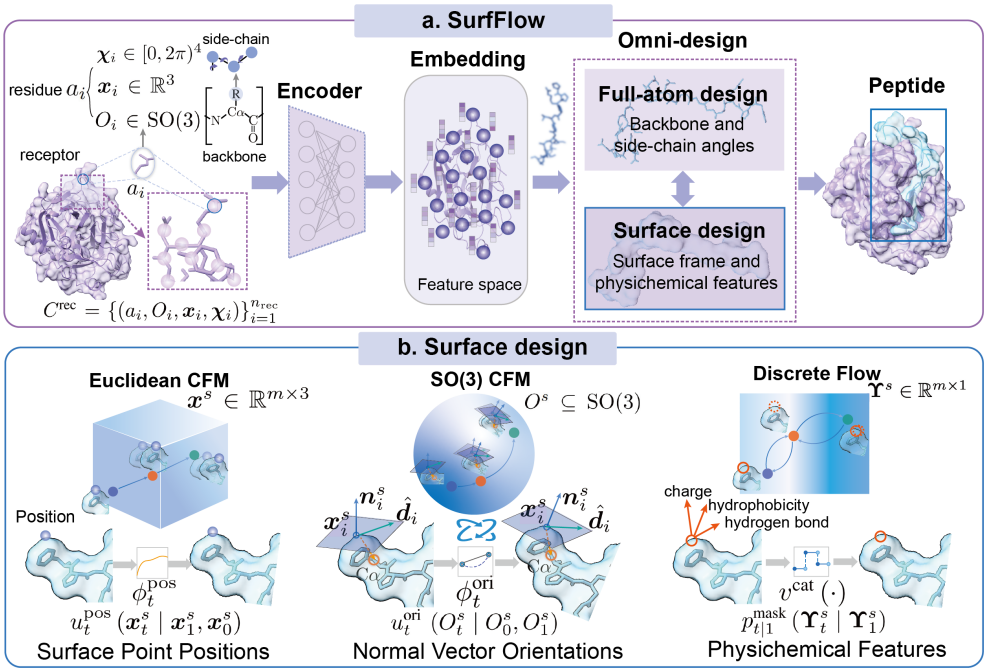


图 2：SurfFlow 的肽全方位设计工作流程，该流程在生成过程中考虑了序列、结构和分子表面之间的多模态一致性。

位置表示为：

$$\mathcal{L}_{\text{pos}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),p(\boldsymbol{x}_1^s),p(\boldsymbol{x}_0^s),p(\boldsymbol{x}_t^s|\boldsymbol{x}_0^s,\boldsymbol{x}_1^s)} \left\| v^{\text{pos}}\left(\boldsymbol{x}_t^s, t, C^{\text{rec}}\right) - \left(\boldsymbol{x}_1^s - \boldsymbol{x}_0^s\right)\right\|_2^2, \tag{11}$$

其中，U(0，1) 是 $[0,1]$ 上的均匀分布。在生成过程中，我们首先从先验分布 $\boldsymbol{x}_0^s \sim \mathrm{N}(0,\ \boldsymbol{I}_3)$ 中采样，并使用学习得到的预测器 $v^{\text{pos}}(\cdot)$ 通过 N 步前向欧拉法求解概率流，从而获得具有 $t = \left\{0, \ldots, \frac{N-1}{N}\right\}$ 的残基 j 的位置：

$$\boldsymbol{x}_{t+\frac{1}{N}}^s = \boldsymbol{x}_t^s + \frac{1}{N} v^{\text{pos}}\left(\boldsymbol{x}_t^s, t, C^{\text{rec}}\right). \tag{12}$$

法向量取向。 法向量 $\boldsymbol{n}_i^s$ 是一个单位向量，垂直于表面在 $\boldsymbol{x}_i^s$ 处的切平面。它揭示了表面取向和曲率的重要几何信息，并与蛋白质的功能性 [52] 直接相关。例如，凸区域可能更易于结合，而凹区域则可能更适合作为参与底物结合的口袋或裂隙 [29]。为捕捉其取向，我们构造了一组相对于全局坐标系的旋转矩阵 $O^s \subseteq \mathrm{SO}(3)$，其元素定义为 $O_i^s = \left(\boldsymbol{n}_i^s, \hat{\boldsymbol{d}}_i,\ \boldsymbol{n}_i^s \times \hat{\boldsymbol{d}}_i\right) \in \mathrm{SO}(3)$。其中，$\hat{\boldsymbol{d}}_i$ 是一个与 $\boldsymbol{n}_i^s$ 正交的单位向量，通过将 $\boldsymbol{n}_i^s$ 与从表面点 $\boldsymbol{x}_i^s$ 指向其最近的 Cα 坐标的方向之间的叉积归一化得到。该坐标系 O_i^s 有效描述了蛋白质表面与其下方骨架结构之间的几何关系。

三维旋转群 SO(3) 是一个光滑的黎曼流形，其切空间 $\mathfrak{so}(3)$构成一个由反对称矩阵组成的李代数。 $\mathfrak{so}(3)$中的元素可以视为绕特定轴的无穷小旋转，并表示为 $\mathbb{R}^3$ [6]中的旋转向量。我们采用 SO(3) 上的均匀分布作为先验分布 $p\left(O_s^0\right)$。正如欧几里得空间中的流匹配基于两点之间的最短路径，我们通过使用测地线将这一思想扩展到 SO(3)；测地线定义了两个取向 [30]之间的最小旋转距离。这些测地线为旋转的插值和演化提供了自然的框架，同时保留了流形的几何结构 [7, 69]。条件流 ϕ^{ori}和向量场 u_t^{ori}通过在 $O_0^s \subseteq \mathrm{U}(\mathrm{SO}(3))$与 $O_1^s \in p\left(O^s \mid C^{\text{rec}}\right)$之间进行测地线插值构建，其中测地距离随时间线性递减：

$$\phi_t^{\text{ori}}\left(O_0^s, O_1^s\right) = \exp_{O_0^s}\left(t \log_{O_0^s}\left(O_1^s\right)\right), \tag{13}$$

$$u_t^{\text{ori}}\left(O_t^s \mid O_0^s, O_1^s\right) = \frac{\log_{O_t^s}\left(O_1^s\right)}{1-t}, \tag{14}$$

其中，exp( ) 和 log(·) 分别是 SO(3) 上的指数映射和对数映射，可使用罗德里格斯公式高效计算。我们使用旋转等变表面网络 $v^{\text{ori}}(\cdot)$ 来预测以旋转向量表示的向量场 u_t^{ori}。SO(3) 上的 CFM 目标函数定义如下：

$$\mathcal{L}_{\text{ori}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),p(O_1^s),p(O_0^s),p(O_t^s|O_0^s,O_1^s)} \left\| v^{\text{ori}}\left(O_t^s, t, C^{\text{rec}}\right) - \frac{\log_{O_t^s}\left(O_1^s\right)}{1-t}\right\|_{\mathrm{SO}(3)}^2, \tag{15}$$

其中，向量场 $v^{\text{ori}}(\cdot)$ 定义在 SO(3) 的切空间 $\mathfrak{so}(3)$ 中，其范数 $|\cdot|^2$ 源自 SO(3) 上的正则度量。在推理阶段，该过程从 $O_0^s \sim$

U(SO(3)) 处初始化，并沿着 SO(3) 中的测地线前进，随时间 t逐步迈出小步：

$$O_{t+\frac{1}{N}} = \exp_{O_t^s}\left(\frac{1}{N} v^{\text{ori}}\left(O_t^s, t, C^{\text{rec}}\right)\right). \tag{16}$$

总体训练损失。结合所有模态，最终的流匹配目标函数用于条件肽生成，由公式 11、15 和 6 中的三个损失函数以及若干额外损失加权求和得到。可表示为：

$$\mathcal{L}_{\text{CFM}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{cat}}\mathcal{L}_{\text{cat}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}}, \tag{17}$$

其中，λ_* 是用于控制不同损失分量影响的超参数。$\mathrm{L_{con}}$ 是连续生化性质的损失函数，而 $\mathrm{L_{str}}$ 是流匹配目标函数，用于对残基的位置 $\{a_j\}_{j=1}^{n_{\text{pep}}}$、方向 $\{O_j\}_{j=1}^{n_{\text{pep}}}$、氨基酸类型 $\{\boldsymbol{x}_j\}_{j=1}^{n_{\text{pep}}}$和侧链扭转角 $\{\boldsymbol{\chi}_j\}_{j=1}^{n_{\text{pep}}}$的因子化分布进行建模，如公式 5 和附录 B.4 所述。用于参数化生成过程的网络细节见附录 A。

3.2 等变表面几何网络

SurfFlow 的成功与为 $v^{\text{cat}}(\cdot)$、$v^{\text{pos}}(\cdot)$ 和 $v^{\text{ori}}(\cdot)$ 精心设计骨干架构密切相关，然而，这些架构面临着若干根本性挑战。首先，分子表面的几何形态不规则，具有连续、极其复杂的内在特性，并呈现出复杂的特征。因此，使用传统的离散网格或基于图的表示难以准确捕捉其结构。其次，可扩展性是一项重大挑战，因为分子表面可能包含数百万个顶点，导致处理此类大规模数据时出现计算瓶颈。近期，等变几何编码器 [15, 51, 57] 在分子表面建模方面展现出良好前景，并在结合亲和力预测 [67],、配体结合位点预测 [71], 和逆折叠 [58]等预测任务中表现出色。然而，它们在生成任务中的适用性，尤其是需要进行蛋白质协同设计的任务，仍基本未得到探索。更重要的是，这些方法没有考虑蛋白质间复杂的相互作用，而这类相互作用对于蛋白质-肽设计等任务至关重要，因此在靶标感知生成建模方面存在明显不足。

为弥合这一差距，我们提出了一种专为蛋白质设计挑战量身打造的等变表面几何网络（ESGN），并引入了若干关键创新。值得注意的是，对于受体口袋，我们也可以生成其表面点云 $\mathcal{Q}^{\text{rec}} = \{q_j^{\text{rec}}\}_{j=1}^{m_{\text{rec}}}$，其中每个实体都具有相关的静态属性 $\left(\boldsymbol{x}_j^{s,\text{rec}}, \boldsymbol{n}_j^{s,\text{rec}}, \boldsymbol{\tau}_j^{s,\text{rec}}, \Upsilon_j^{s,\text{rec}}\right)$。因此，我们可以在不同时间步 t构建动态异质表面图 $\mathrm{G} = (\mathrm{G_{rec}}, \mathrm{G_{pep}})$，其中内部 和间点连接关系（$\mathrm{E_{rec}}$，$\mathrm{E_{pep}}$，$\mathrm{E_{inter}}$）由空间距离阈值 r确定。我们通过对各图的 SE(3)不变理化性质进行简单映射，初始化每个图的节点特征，得到

$$\boldsymbol{h}_i^{(0)} = f_e\left(\Upsilon_i^s\right) \qquad \boldsymbol{h}_j^{(0)\,\text{rec}} = (\,) \text{ 和,}$$

$f_e\left(\Upsilon_j^{s\,\text{rec}}\right) \qquad f_e : \mathbb{R}^{\psi_\Upsilon} \to \mathbb{R}^{\psi_h}$

$(,)$，其中 : 表示多层感知器（MLP）。

接着，在将 $\boldsymbol{h}_i^{(0)}$ 和 $\boldsymbol{h}^{(0)}{}_{,\,j}^{\text{rec}}$ 传递到 ESGN 的每一层 $l \in [L]$之前，我们会计算用于描述局部几何结构的关键标量：其中一个是点间距离 $d_{ij}^s = \|\boldsymbol{x}_i^s - \boldsymbol{x}_j^s\|^2$，另一个是 $d^s{}_{,ij}^{\text{rec}} = \|\boldsymbol{x}^s{}_{,i}^{\text{rec}} - \boldsymbol{x}^s{}_{,i}^{\text{rec}}\|^2$，即法向量与两点间连接有向直线之间的夹角（$\boldsymbol{x}_{ij}$、$\boldsymbol{x}_{ji}$），通过 $\varphi_{\mathbf{n}_i ij} = \angle \boldsymbol{n}_i \boldsymbol{x}_{ij}$和 $\varphi_{\mathbf{n}_j ji} = \angle \boldsymbol{n}_j \boldsymbol{x}_{ji}$计算得出，其中两个相交平面可由两个表面点的法向量构成。之后，任意相连节点（即，$\forall e_{j\to i} \in \mathrm{E_{rec}} \cup \mathrm{E_{pep}}$）之间的内部图消息更新如下：

$$\boldsymbol{m}_{j\to i} = f_m\left(\boldsymbol{h}_i^{(l)}, \boldsymbol{h}_j^{(l)}, \boldsymbol{e}_{\text{RBF}}^{(ij)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_i ij)}, \boldsymbol{a}_{\text{SBF}}^{(\mathbf{n}_j ji)}\right), \tag{18}$$

$$\boldsymbol{m}'_{j\to i} = \varpi_{ij} \cdot \boldsymbol{m}_{j\to i}, \quad \varpi_{ij} = \frac{\exp(\boldsymbol{W}_{\text{m}} \boldsymbol{m}_{j\to i} + b_{\text{m}})}{\sum_{k\in\mathcal{N}_{(i)}} \exp(\boldsymbol{W}_{\text{m}} \boldsymbol{m}_{k\to i} + b_{\text{m}})}, \tag{19}$$

其中，$f_m(\cdot)$ 是另一个附加的多层感知器，用于将每对节点特征转换为消息 $\boldsymbol{m}_{j\to i} \in \mathbb{R}^{\psi_m}$。$\boldsymbol{e}^{(ij)}$ RBF 是 d_{ij}^s 和 $d^s{}_{,ij}^{\text{rec}}$ 的径向基函数表示。$\boldsymbol{a}^{(\mathbf{n}_i ij)}{}_{\text{SBF}}$ 和 $\boldsymbol{a}^{(\mathbf{n}_j ji)}{}_{\text{SBF}}$ 是 [18] 的三维球面傅里叶-贝塞尔表示。随后采用 Softmax 对消息进行重新加权，其中权重矩阵 $\boldsymbol{W}_{\text{m}} \in \mathbb{R}^{\psi_m \times 1}$ 和向量 $b_{\text{m}} \in \mathbb{R}$ 均为可学习参数，并且 $\mathrm{N}^{\text{intra}}{}_{(i)} = \{j | e_{j\to i} \in \mathrm{E_{rec}} \cup \mathrm{E_{pep}}\}$ 是同一蛋白质内节点 i 的邻域集合。

除表面内相互作用外，我们还利用距离加权注意力技术 [17, 64]，获取 $\forall e_{j\to i} \in \mathrm{E_{inter}}$ 的互连跨图消息：

$$a_{j\to i} = \frac{\exp\left(\left\langle f_q\left(\mathbf{h}_i^{(t),l}\right), f_k\left(\mathbf{h}_j^{(l)}\right)\right\rangle\right)}{\sum_{j'} \exp\left(\left\langle f_q\left(\mathbf{h}_i^{(t),l}\right), f_k\left(\mathbf{h}_{j'}^{(t),l}\right)\right\rangle\right)}, \tag{20}$$

$$\boldsymbol{\mu}_{j\to i} = a_{j\to i} \mathbf{h}_j^{(l)} \cdot \phi_d\left(d_{ij}^{s\,(l)}\right), \forall e_{j\to i} \in \mathcal{E}_{\text{inter}}, \tag{21}$$

其中，$a_{j\to i}$ 是由可训练多层感知机 $f_q(\cdot)$ 和 $f_k(\cdot)$ 计算得到的注意力分数。$\phi_d(\cdot)$ 作用于点间距离 $d_{ij(l)}^s$，以获取 $\boldsymbol{\mu}_{j\to i} \in \mathbb{R}^{\psi_\mu}$。随后，将*intra*-消息和*inter*-消息聚合，并从每个点 q_i 的邻近区域传播，以更新其节点特征和坐标，具体如下：

$$\boldsymbol{h}_i^{(l+1)} = f_h\left(\boldsymbol{h}_i^{(l)}, \sum_{j\in\mathcal{N}_{(i)}^{\text{intra}}} \boldsymbol{m}_{j\to i}, \sum_{j'\in\mathcal{N}_{(i)}^{\text{inter}}} \boldsymbol{\mu}_{j'\to i}\right), \tag{22}$$

$$\boldsymbol{x}_i^{(l+1)} = \boldsymbol{x}_i^{(l)} + \frac{1}{\|\mathcal{N}_{(i)}^{\text{intra}}\|} \sum_{j\in\mathcal{N}_{(i)}^{\text{intra}}} f_x^{\text{intra}}\left(\boldsymbol{m}_{j\to i}\right) \boldsymbol{x}_{ij}^{(l)} \tag{23}$$

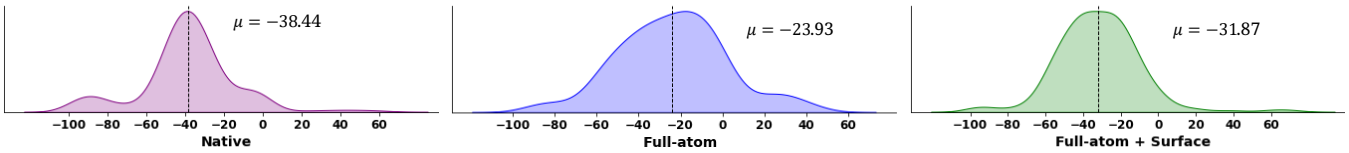
$$+ \frac{1}{\|\mathcal{N}_{(i)}^{\text{inter}}\|} \sum_{j'\in\mathcal{N}_{(i)}^{\text{inter}}} f_x^{\text{inter}}\left(\boldsymbol{\mu}_{j'\to i}\right) \boldsymbol{x}_{ij'}^{(l)}, \tag{24}$$

其中，$f_h(\cdot)$ 是节点操作，通过多层感知器实现，用于聚合图内消息和跨图消息。$f\mathrm{intra}_x$：$\mathbb{R}^{\psi_m} \to \mathbb{R}$ 和 $f\mathrm{inter}_x$：$\mathbb{R}^{\psi_\mu} \to \mathbb{R}$ 分别将 $\boldsymbol{m}_{j\to i}$ 和 $\boldsymbol{\mu}_{j'\to i}$ 转换为标量分数，以控制方向向量 $\boldsymbol{x}_{ij}^{(l)} = \boldsymbol{x}(l)_i - \boldsymbol{x}(l)_j$ 和 $\boldsymbol{x}(l)_{ij'} = \boldsymbol{x}_i^{(l)} - \boldsymbol{x}(l)_{j'}$ 的影响。$\mathrm{N}^{\text{inter}}(i) = \{j | e_{j\to i} \in \mathrm{E_{inter}}\}$ 是对应蛋白质中节点 i 的邻域集合。值得注意的是，随着流在不同时间步 $t \in [0, 1]$ 对表面点云 $\boldsymbol{x}_t^s$ 的坐标进行插值，图间和图内连接关系（$\mathrm{E_{inter}}$、$\mathrm{E_{pep}}$）会动态变化，因此必须根据距离阈值 r 重新构建。由此可轻易得出，ESGN 保持了等变性。

[1]为简洁起见，省略上标递归 。

表 1：不同方法在序列-结构协同设计任务中的评估，以及对 SurfFlow 关键组件的消融研究。最佳结果和次优结果分别以粗体和下划线标示。

	几何			能量			设计	
	氨基酸恢复率 % ↑	均方根偏差 Å ↓	二级结构恢复率 % ↑	结合位点恢复率 % ↑	稳定性 % ↑	亲和力 % ↑	设计百分比 ↑	多样性 ↑
RFdiffusion [61]	40.14	4.17	63.86	26.71	**26.82**	16.53	**78.52**	0.38
ProteinGen [37]	45.82	4.35	29.15	24.62	23.48	13.47	71.82	0.54
Diffusion [39]	47.04	3.28	74.89	49.83	15.34	17.13	48.54	0.57
PepGLAD [26]	50.43	3.83	80.24	19.34	20.39	10.47	75.07	0.32
PPIFlow [34]	48.35	3.59	68.13	25.94	15.77	12.08	46.53	0.51
PepFlow [33]	51.25	2.07	83.46	86.89	18.15	21.37	65.22	0.42
SurfFlow（不含 ESGN）	52.59	2.05	83.77	86.91	19.42	19.82	68.41	0.60
SurfFlow（不含位置）	53.26	1.99	84.79	87.15	21.30	22.38	72.09	0.60
SurfFlow（不含方向）	53.04	2.00	84.60	87.04	20.79	22.46	72.36	0.60
SurfFlow（不含生物物理属性）	52.31	2.03	83.96	86.98	19.55	20.25	70.83	0.58
SurfFlow	**54.07**	**1.96**	**85.11**	**87.38**	22.46	**22.51**	73.60	**0.61**



Fig图 3：设计肽和天然肽的结合能分布，其中数值越低越好

3.3 条件蛋白质设计

受可控图像生成成功经验的启发 [70], ，我们提出了由关键因素 c 条件化的肽设计，例如序列长度 n_{pep}、环化性以及是否存在二硫键。我们的目标变为 $p\left(C^{\text{pep}} \mid C^{\text{rec}},\ c\right)$，使流模型能够纳入额外条件。因此，向量场网络调整为 $v^{\text{pos}}\left(\boldsymbol{x}_t^s,\ t,\ C^{\text{rec}},\ c\right)$、$v^{\text{ori}}\left(O_t^s,\ t,\ C^{\text{rec}},\ c\right)$，以及 $v^{\text{cat}}\left(\Upsilon_t^s,\ t,\ C^{\text{rec}},\ c\right)$。在实践中，我们聚焦于两个主要目标：(1)环肽通过限制骨架来提高稳定性，从而降低构象柔性并增强对酶降解的抵抗力。这种结构形式带来更明确、更稳定的构象，因此能够提高结合亲和力。(2)二硫键是半胱氨酸残基之间的共价相互作用，有助于正确折叠和结构稳定化。这些键可保护肽免受氧化损伤和蛋白水解酶的影响，增强其对严苛工业环境的抵抗力。此外，二硫键还可通过形成构象限制来提升生物活性，进一步提高肽类药物的治疗潜力。

4 实验

我们全面评估 SurfFlow 在无条件和有条件的序列-结构协同设计任务以及侧链堆积问题上的表现。基准测试使用源自 PepBDB [63] 和 Q-BioLip [62]的数据集 PepMerge。遵循 Li 等人[33], 的方法，我们在去除重复项并应用经验标准（例如，分辨率 < 4Å，肽长度介于 3 至 25 之间）后，使用 mmseqs2 [54], 根据 40% 的序列一致性对复合物进行聚类。由此得到 8,365 个非冗余复合物，分属 292 个簇。测试集由 10 个簇中的 158 个复合物组成。更多详情和其他结果见附录 B。

4.1 无条件序列-结构协同设计

基线方法。 我们选取两类最先进水平的蛋白质设计方法作为基线方法。第一类忽略侧链构象，包括 RFDiffusion [61] 和 ProteinGen [37]。RFDiffusion 生成蛋白质主链，随后由 ProteinMPNN [11] 预测序列。ProteinGen 则通过联合采样主链及其对应序列，改进了 RFDiffusion。另一类采用全原子风格的蛋白质生成方法，包括 Diffusion [39],PepGLAD [26], 和 PepFlow [33]。

评估指标。 生成的肽从三个关键方面进行评估。(1) **几何**：设计的肽应与天然序列和结构高度相似。我们使用氨基酸恢复率（**氨基酸恢复率**）量化生成肽与真实值之间的序列一致性。复合物对齐后，通过 C_α 原子的均方根偏差（**均方根偏差**）评估结构相似性。二级结构相似率（**二级结构恢复率**）衡量共有二级结构所占的比例，而结合位点比率（**结合位点恢复率**）则比较生成肽与天然肽在目标蛋白质上的结合位点重叠情况。(2)**能量**：我们的目标是设计高亲和力的肽结合剂，以提高蛋白质-肽复合物的稳定性。**亲和力**定义为结合亲和力高于天然肽（结合能更低）的生成肽所占百分比，而

稳定性 表示总能量低于天然状态的复合物所占比例。能量计算使用 Rosetta [3]完成。**(3) 设计**：**可设计性** 反映设计序列与结构之间的一致性。其衡量方式为：能够折叠成与相应生成结构相似结构的序列所占比例，其中 C_α RMSD < 2 Å为阈值。我们使用 ESMFold [35] 对序列进行重新折叠。**多样性**以一减去两两 TM评分 [73], 的平均值来衡量，表示设计肽之间的结构差异。

结果。 表 1 显示，与基线方法相比，我们的 SurfFlow 能生成多样性和一致性显著更高、结合能更佳且相似度更高的肽。具体而言，SurfFlow 达到了最先进水平的氨基酸恢复率（AAR）54.07% 和均方根偏差（RMSD）1.96 Å，较全原子 PepFlow 分别提升了 5.51% 和 5.31%。此外，它在设计肽方面也表现出更强的能力，结合比例更准确，达到 87.38%，亲和力比例也更高，达到 22.51%（见图 3）。这些统计数据证实了显式建模表面几何和生化约束的优势。值得一提的是，解耦方法 RFDiffusion 在稳定性（26.82% 对比 22.46%）和可设计性（78.52% 对比 73.60%）方面表现更优，这是因为它在整个蛋白质数据库（PDB）上进行训练，因此倾向于生成具有更稳定基序的结构。

图 4 展示了全原子 PepFlow 和基于表面的 SurfFlow 设计的两个肽示例。附录 C.1 展示了整个过程中表面点云位置的变化。无论天然长度如何，SurfFlow 都能生成拓扑结构相似的肽。值得注意的是，生成的肽具有相近的侧链组成和构象，因此能够在适当的结合位点与目标蛋白质高效相互作用。

我们还在表 1 中研究了 SurfFlow 各组件的作用，其中使用朴素的 PointNet [46] 替代 ESGN。结果表明，ESGN 对表面相互作用的有效建模是 SurfFlow 成功的关键。移除生物物理特征会显著降低性能，使可设计性下降 3.76%，多样性下降 4.91%。此外，结果还表明，纳入表面取向有益，可使 AAR 提升 1.94%。我们还在表 2 中报告了表面一致性，并在附录 5 中报告了推理计算成本；在复杂度相当的情况下，SurfFlow 显著提升了生成肽与真实值之间的表面相似性。

4.2 条件化序列-结构协同设计

生物学界普遍认可，有效肽具有若干关键特征。例如，环化性是指多肽链通过环状键序列形成环结构。许多环肽天然存在，并表现出抗菌或毒性，而另一些则是在实验室中合成的 [22]。这些环肽通常具有很强的抗消化能力，因此对开发新型口服药物（如抗生素和免疫抑制剂）的研究人员颇具吸引力 [10]。此外，有证据表明，二硫键模式对肽的折叠和结构稳定至关重要。将二硫键有意引入天然或工程化肽中，往往能够增强其生物

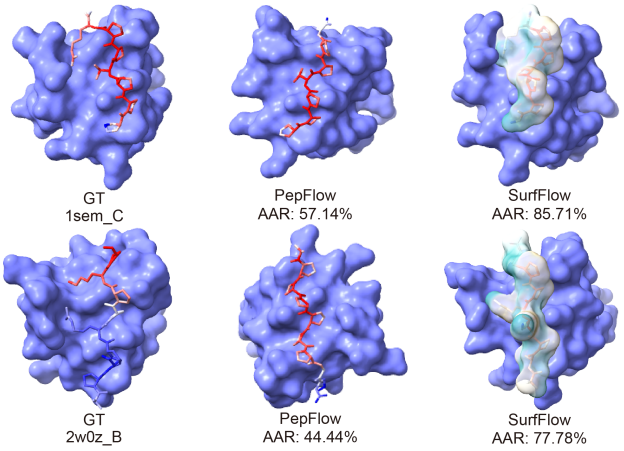


图 4：由深度学习算法设计的肽及其参考对象。

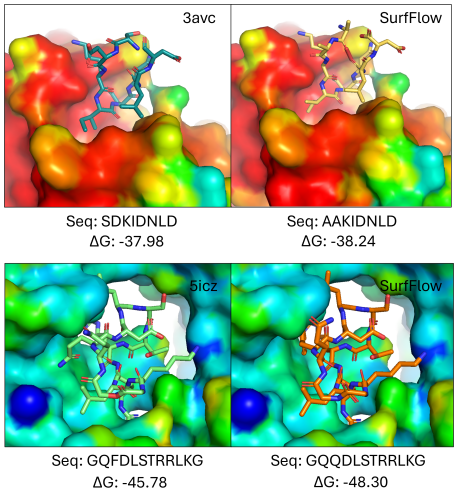


图 5：带有环状条件的肽设计。

活性、特异性和稳定性 [4]。基于这些认识，我们提出一项条件协同设计挑战。该挑战旨在计算生成的肽集合中环肽及含有二硫键的肽所占比例，以供评估。

结果。 表 3 汇总了结果。在没有条件约束的情况下，无论是全原子深度生成模型还是我们的全能设计方法，都未能满足环化性或二硫键的要求。即使引入分子表面，测试样本中 189 个设计肽里也只有 2% 至 4% 表现出环化性或含有二硫键。相比之下，在加入额外条件进行训练，并提示模型生成具有特定属性的肽后，SurfFlow 显著提高了具有目标特征的肽所占比例：具有环化性的比例从 2.67% 提高到 8.02%，含有二硫键的比例则从 4.27% 提高到 9.10%。图 5 展示了两个带有环化条件的示例。有趣的是，在天然肽为非环肽的 3AVC 案例中，SurfFlow 生成了一种新型环肽，其结合能更低（$\Delta G = -37.98$ 与

表 2：PepMerge 上的表面一致性评估。

模型	交并比(↑)	倒角距离(↓)	法向量一致性(↑)
PepFlow	0.7252	9.437	0.3088
SurfFlow	**0.8104**	**6.347**	**0.3859**

表 3：不同机制设计的环肽和含二硫键的肽，以及原始 PepMerge 数据集中的相应比例。

指标	PepFlow	SurfFlow（不含 c）	SurfFlow	PepMerge
环化性	2.13%	2.67%	**8.02%**	15.50%
二硫键%	3.21%	4.27%	**9.10%**	18.18%

$\Delta G = -38.24$）。对于原本含有环肽的 5ICZ，SurfFlow 设计出了一种新型环肽，且结合能更优（$\Delta G = -45.78$ 与 $\Delta G = -48.30$）。

4.3 侧链堆积

侧链堆积是蛋白质结构建模中的一项关键任务，重点在于预测肽侧链角度。我们采用多个模型，为每个肽生成 64 种不同的侧链构象，并通过部分采样策略高效地恢复最可能的侧链角度。该方法使我们能够有效探索构象空间，同时降低计算开销并保持较高精度。

基线方法。 我们将 SurfFlow 与多种成熟方法进行比较。RosettaPacker [32] 和 SCWRL4 [27] 等基于能量的方法依赖物理能量最小化技术来确定侧链位置。此外，DLPacker [42],、AttnPacker [41], 和 DiffPack [72]等基于学习的模型利用注意力机制和扩散等多种深度学习策略，根据学得的蛋白质结构表征预测侧链构型。

指标。 我们采用两项指标进行评估。首先，计算 4 个关键扭转角的平均绝对误差（MAE）：$\{\chi_i\}_{i=1}^{4}$。鉴于侧链固有的柔性以及微小偏差在结构生物学中的重要性，我们还报告预测结果与真实值之间偏差在 $20°$ 以内的比例。这项补充指标反映了侧链预测的实际准确性，重点衡量模型在生物学相关误差范围内的表现。

结果。 表 4 报告了结果；可以看出，与全原子模型和其他基线方法相比，引入分子表面有助于更准确地预测四个侧链角度。SurfFlow 的正确率最高，达到 63.02%。准确率的提升表明，表面信息为预测侧链构象提供了关键背景，可能是因为它能更好地表征局部环境以及影响侧链定位的潜在相互作用。

5 结论

本文介绍了 SurfFlow，这是一种新型模型，可同时生成蛋白质的所有模态——序列、结构和分子表面。我们将 SurfFlow 应用于解决一个特定的肽设计挑战，并

表 4：侧链堆积任务的评估结果。

	均方误差 ° ↓				
	χ_1	χ_2	χ_3	χ_4	正确率 % ↑
Rosseta	38.31	43.23	53.61	71.67	57.03
SCWRL4	30.06	40.40	49.71	53.79	60.54
DLPacker	22.44	35.65	58.53	61.70	60.91
AttnPacker	19.04	28.49	40.16	60.04	61.46
DiffPack	17.92	26.08	36.20	67.82	62.58
PepFlow	17.38	24.71	33.63	58.49	62.79
SurfFlow	**17.13**	**23.86**	**31.97**	**55.08**	**63.02**

将环化性和二硫键等关键特征融入生成过程。实证结果证明，在蛋白质发现中考虑分子表面是合理且有前景的。局限性和未来工作详见附录 D。

致谢

本研究得到北京市科学技术委员会和中关村科学城管理委员会项目（项目编号：Z221100003522019）的支持。

参考文献

[1] Michael S Albergo、Nicholas M Boffi 和 Eric Vanden-Eijnden。2023。随机插值：统一流与扩散的框架。*arXiv* 预印本 *arXiv:2303.08797* (2023)。[2] Michael S Albergo 和 Eric Vanden-Eijnden。2022。利用随机插值构建归一化流。*arXiv* 预印本 *arXiv:2209.15571* (2022)。[3] Rebecca F Alford 等人。2017。用于大分子建模与设计的 Rosetta 全原子能量函数。《化学理论与计算杂志》13, 6 2017 3031 3048 ( ), – 。[4] Ioana Annis、Balazs Hargittai 和 George Barany。1997。肽中的二硫键形成。《酶学方法》289 (1997), 198–221。[5] Gaurav Bhardwaj、Vikram Khipple Mulligan、Christopher D Bahl、Jason M Gilmore、Peta J Harvey、Olivier Cheneval、Garry W Buchko、Surya VSRK Pulavarti、Quentin Kaas、Alexander Eletsky 等人。2016。超稳定受限肽的精准从头设计。《自然》538, 7625 (2016), 329–335。[6] José Luis Blanco-Claraco。2021。SE(3)变换参数化与流形上优化教程。*arXiv* 预印本 *arXiv:2103.15980* (2021)。[7] Avishek Joey Bose 等人。2023。用于蛋白质主链生成的 SE(3) 随机流匹配。*arXiv* 预印本 *arXiv:2310.02391* (2023)。[8] Laura K Buckton、Marwa N Rahimi 和 Shelli R McAlpine。2021。作为细胞内靶标药物的环肽：肽类药物开发的下一个前沿。《欧洲化学杂志》27, 5 (2021), 1487–1513。[9] Andrew Campbell 等人。2024。离散状态空间上的生成流：实现多模态流及其在蛋白质协同设计中的应用。*arXiv* 预印本 *arXiv:2402.04997* (2024)。[10] David J Craik。2006。无缝蛋白质将自身松散的末端系在一起。《科学》311, 5767 (2006), 1563–1564。[11] Dauparas 等人。2022。使用 ProteinMPNN 进行基于深度学习的稳健蛋白质序列设计。《科学》378, 6615 (2022), 49–56。[12] Warren L DeLano 等人。2002。PyMol：一种开源分子图形工具。《*CCP4* 蛋白质晶体学通讯》40, 1 (2002), 82–92。[13] Prafulla Dhariwal 和 Alexander Nichol。2021。扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展34 (2021), 8780–8794。[14] Matthew Fisher。2001。《莱宁格生物化学原理》，作者 David L. Nelson 和 Michael M. Cox。《化学教育家》 6 (2001), 69–70。[15] Pablo Gainza 等人。2020。利用几何深度学习从蛋白质分子表面解读相互作用指纹。《自然·方法》17, 2 (2020), 184–192。[16] Pablo Gainza 等人。2023。利用学习得到的表面指纹从头设计蛋白质相互作用。《自然》617, 7959 (2023), 176–184。[17] Octavian-Eugen Ganea 等人。2021。用于端到端刚性蛋白质对接的独立 SE(3) 等变模型。*arXiv* 预印本 *arXiv:2111.07786* (2021)。[18] Johannes Gasteiger、Shankari Giri、Johannes T Margraf 和 Stephan Günnemann。2020。用于非平衡分子的快速且具不确定性感知能力的定向消息传递。*arXiv* 预印本 *arXiv:2011.14115* (2020)。

[19] Itai Gat 等人。2024。离散流匹配。arXiv 预印本 arXiv:2407.15595 (2024)。[20] Jiaqi Guan、Wesley Wei Qian、Xingang Peng、Yufeng Su、Jian Peng 和 Jianzhu Ma。2023。用于目标感知分子生成与亲和力预测的三维等变扩散。arXiv 预印本 arXiv:2303.03543 (2023)。[21] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020。去噪扩散概率模型。神经信息处理系统进展 33 (2020), 6840–6851。[22] Knud J Jensen。2009。生物制药应用中的肽与蛋白质设计。(2009)。[23] Susan Jones 和 Janet M Thornton。1996。蛋白质-蛋白质相互作用原理。《美国国家科学院院刊》93, 1 (1996), 13–20。[24] John Jumper 等人。2021。利用 AlphaFold 实现高精度蛋白质结构预测。《自然》596, 7873 (2021), 583–589。[25] Panagiotis L Kastritis 和 Alexandre MJJ Bonvin。2013。大分子相互作用的结合亲和力：大胆探究蛋白质为何相互作用。《英国皇家学会界面杂志》10, 79 (2013), 20120835。[26] Xiangzhe Kong、Wenbing Huang 和 Yang Liu。2024。基于几何潜扩散的全原子肽设计。arXiv 预印本 arXiv:2402.13555 (2024)。[27] Georgii G Krivov、Maxim V Shapovalov 和 Roland L Dunbrack Jr. 2009。利用 SCWRL4 改进蛋白质侧链构象预测。《蛋白质：结构、功能与生物信息学》77, 4 (2009), 778–795。[28] Houtim Lai 等人。2024。Interformer：一种用于蛋白质-配体对接和亲和力预测的相互作用感知模型。(2024)。[29] Roman A Laskowski 等人。1996。蛋白质裂隙在分子识别与功能中的作用。《蛋白质科学：蛋白质学会刊物》5, 12 (1996), 2438。[30] John M Lee。2018。黎曼流形导论。第2卷。Springer。[31] Youhan Lee 等人。2023。预训练序列、结构和表面特征，实现全面的蛋白质表示学习。发表于第十二届国际学习表征会议。[32] Julia Koehler Leman 等人。2020。Rosetta 中的大分子建模与设计：近期方法和框架。《自然·方法》17, 7 (2020), 665–680。[33] Jiahan Li 等人。2024。基于多模态流匹配的全原子肽设计。arXiv 预印本 arXiv:2406.00735 (2024)。[34] Haitao Lin 等人。2024。PPFLOW：利用扭转流匹配进行目标感知肽设计。bioRxiv (2024), 2024–03。[35] Zeming Lin、Halil Akin、Roshan Rao、Brian Hie、Zhongkai Zhu、Wenting Lu、Allan dos Santos Costa、Maryam Fazel-Zarandi、Tom Sercu、Sal Candido 等人。2022。进化尺度的蛋白质序列语言模型可实现精准的结构预测。bioRxiv (2022)。[36] Yaron Lipman 等人。2022。用于生成建模的流匹配。arXiv 预印本 arXiv:2210.02747 (2022)。[37] Sidney Lyayuga Lisanza 等人。2023。利用 RoseTTAFold 序列空间扩散联合生成蛋白质序列与结构。bioRxiv (2023), 2023–05。[38] Xingchao Liu、Chengyue Gong 和 Qiang Liu。2022。流动更直、速度更快：学习利用整流流生成和迁移数据。arXiv 预印本 arXiv:2209.03003 (2022)。[39] Shitong Luo、Yufeng Su、Xingang Peng、Sheng Wang、Jian Peng 和 Jianzhu Ma。2022。利用基于扩散的生成模型进行抗原特异性抗体设计与优化。bioRxiv (2022)。[40] Karolis Martinkus 等人。2024。AbDiffuser：全原子生成具有体外功能的抗体。神经信息处理系统进展 36 (2024)。[41] Matthew McPartlon 和 Jinbo Xu。2023。一种用于蛋白质侧链堆积和逆折叠的端到端深度学习方法。《美国国家科学院院刊》120, 23 (2023), e2216438120。[42] Mikita Misiura 等人。2022。DLPacker：利用深度学习预测蛋白质中的氨基酸侧链构象。《蛋白质：结构、功能与生物信息学》90, 6 (2022), 1278–1290。[43] Julie L Morrison 等人。2006。蛋白质-蛋白质相互作用的锁钥模型。《生物信息学》22, 16 (2006), 2012–2019。[44] Markus Muttenthaler、Glenn F King、David J Adams 和 Paul F Alewood。2021。肽类药物发现的趋势。《自然综述：药物发现》20, 4 (2021), 309–325。[45] James R Norris。1998。马尔可夫链。第2号。剑桥大学出版社。[46] Charles R Qi、Hao Su、Kaichun Mo 和 Leonidas J Guibas。2017。PointNet：在点集上进行三维分类与分割的深度学习。发表于 IEEE 计算机视觉与模式识别会议论文集。652–660。[47] Vishva Saravanan Ramasubramanian、Soham Choudhuri 和 Bhaswar Ghosh。2024。一种用于稳定、亲和力驱动、受体感知肽生成的混合扩散模型。bioRxiv (2024), 2024–03。[48] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。利用 CLIP 潜变量进行分层文本条件图像生成。arXiv 预印本 arXiv:2204.06125 1, 2 (2022), 3。[49] Robinson 等人。2014。MSM：一种用于多模态表面匹配的新型灵活框架。《神经影像》100 (2014), 414–426。[50] Vıctor Garcia Satorras、Emiel Hoogeboom 和 Max Welling。2021。E(n) 等变图神经网络。发表于国际机器学习会议。PMLR, 9323–9332。[51] Vignesh Ram Somnath、Charlotte Bunne 和 Andreas Krause。2021。蛋白质的多尺度表征学习。神经信息处理系统进展 34（2021），25244–25255。[52] Zhenqiao Song、Tinglin Huang、Lei Li 和 Wengong Jin。2024。SurfPro：基于连续表面的功能性蛋白质设计。arXiv 预印本 arXiv:2405.06693（2024）。[53] Hannes Stark 等。2024。狄利克雷流匹配及其在DNA序列设计中的应用。arXiv 预印本 arXiv:2402.05841（2024）。[54] Martin Steinegger 和 Johannes Söding。2017。MMseqs2 支持对海量数据集进行灵敏的蛋白质序列搜索与分析。《自然-生物技术》35，11（2017），1026–1028。[55] Daiwen Sun、He Huang、Yao Li、Xinqi Gong 和 Qiwei Ye。2024。DSR：作为蛋白质隐式神经网络的动态表面表示。神经信息处理系统进展 36（2024）。[56] Haoran Sun 等。2022。基于分数的连续时间离散扩散模型。arXiv 预印本 arXiv:2211.16750（2022）。[57] Freyr Sverrisson、Jean Feydy、Bruno E Correia 和 Michael M Bronstein。2021。蛋白质表面的快速端到端学习。载于 IEEE/CVF计算机视觉与模式识别会议论文集。15272–15281。[58] Xiangru Tang 等。2024。BC-Design：一种用于高精度逆向蛋白质折叠的生物化学感知框架。bioRxiv（2024），2024–10。[59] Peter Vanhee、Almer M van der Sloot、Erik Verschueren、Luis Serrano、Frederic Rousseau 和 Joost Schymkowitz。2011。肽配体的计算设计。《生物技术趋势》29，5（2011），231–239。[60] Lei Wang 等。2022。治疗性肽：当前应用与未来方向。《信号转导与靶向治疗》7，1（2022），48。[61] Joseph L Watson 等。2023。利用 RFdiffusion 从头设计蛋白质结构与功能。《自然》620，7976（2023），1089–1100。[62] Hong Wei、Wenkai Wang、Zhenling Peng 和 Jianyi Yang。2024。Q-biolip：一个全面的、基于四级结构的蛋白质-配体相互作用资源。《基因组学、蛋白质组学与生物信息学》22，1（2024）。[63] Zeyu Wen、Jiahua He、Huanyu Tao 和 Sheng-You Huang。2019。PepBDB：一个全面的生物肽-蛋白质相互作用结构数据库。《生物信息学》35，1（2019），175–177。[64] Fang Wu 等。2022。具有构象柔性的等变图匹配网络预训练及其在药物结合中的应用。《先进科学》9，33（2022），2203796。[65] Fang Wu 等。2024。D-Flow：用于D-肽设计的多模态流匹配。arXiv 预印本 arXiv:2411.10618（2024）。[66] Fang Wu 和 Stan Z Li。2024。抗体结构-序列协同设计的分层训练范式。神经信息处理系统进展 36（2024）。[67] Fang Wu 和 Stan Z Li。2024。Surface-VQMAE：分子表面上的向量量化掩码自编码器。载于国际机器学习会议。PMLR，53619–53634。[68] Kevin E Wu 等。2024。通过折叠扩散生成蛋白质结构。《自然-通讯》15，1（2024），1059。[69] Jason Yim 等。2023。利用 SE(3)流匹配快速生成蛋白质主链。arXiv 预印本 arXiv:2310.05297（2023）。[70] Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。2023。为文本到图像扩散模型添加条件控制。载于 IEEE/CVF国际计算机视觉会议论文集。3836–3847。[71] Yang Zhang 等。2023。Equipocket：用于配体结合位点预测的 E(3) 等变几何图神经网络。arXiv 预印本 arXiv:2302.12177（2023）。[72] Yangtian Zhang 等。2024。Diffpack：一种用于自回归蛋白质侧链堆积的扭转扩散模型。神经信息处理系统进展 36（2024）。[73] Yang Zhang 和 Jeffrey Skolnick。2004。用于自动评估蛋白质结构模板质量的评分函数。《蛋白质：结构、功能与生物信息学》57，4（2004），702–710。[74] Qinqing Zheng 等。2023。用于生成建模与决策的引导流。arXiv 预印本 arXiv:2311.13443（2023）。

网络参数化

为了对肽基于其目标蛋白质 $p\left(C^{\text{pep}} \mid C^{\text{rec}}\right)$ 的联合分布进行建模，我们采用编码器-解码器框架来生成肽。编码器提取受体 C^{rec} 的几何和生化特征，作为生成过程的条件；解码器则回归我们多模态流匹配架构的向量场。

编码器。 我们首先使用一个与时间无关的等变几何编码器来捕获受体的上下文信息。具体而言，它以目标蛋白质 C_{rec} 的序列和结构为输入，并计算隐藏的残基表示 $\boldsymbol{h}_{\text{rec}} \in \mathbb{R}^{n_{\text{rec}} \times \psi_{\text{递归}}}$ 以及残基对嵌入 $\boldsymbol{z}_{\text{rec}} \in \mathbb{R}^{n_{\text{rec}} \times \psi_{\text{对}}}$。

解码器。解码器接收 $(\boldsymbol{h}_{\text{rec}}, \boldsymbol{z}_{\text{rec}})$，且依赖于时间。它由两个几何网络组成：一个是 6 层的不变点注意力 (IPA) 模块 [24]，用于回归内部结构 $\left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}}$ 的向量场；另一个是等变图神经网络 (EGNN) [50] 的变体，用于回归表面几何 $\left\{\left(\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^{m}$ 的向量场。具体来说，我们首先采样一个随机时间步 $t \sim \mathrm{U}(0,1)$，为肽 C_{pep}的每种模态构建依赖于时间的向量场，包括序列、结构和表面。IPA 和 EGNN 均将时间戳 t、肽内部结构的插值状态 $(a_t, O_t, \boldsymbol{x}_t, \boldsymbol{\chi}_t)$ 以及受体信息 $(\boldsymbol{h}_{\text{rec}}, \boldsymbol{z}_{\text{rec}})$ 作为输入；肽表面的插值状态 $\left(\boldsymbol{x}_t^s, \boldsymbol{n}_t^s, \boldsymbol{\tau}_t^s, \Upsilon_t^s\right)$ 也会输入 EGNN。随后，IPA 恢复原始肽的内部结构 $\left(\hat{}a_1, \hat{}O_1, \hat{}\boldsymbol{x}_1, \hat{\boldsymbol{\chi}}_1\right)$，而 3 层 EGNN 恢复表面$\left(\hat{}\boldsymbol{x}_1^s, \hat{}\boldsymbol{n}_1^s, \hat{}\boldsymbol{\tau}_1^s, \hat{}\Upsilon_1^s\right)$。此外，还施加了两项额外损失作为约束：主链位置损失 $\mathrm{L}_{\text{bb}}(\theta)$ 和扭转角损失 $\mathrm{L}_{\text{tor}}(\theta)$。

等变性。联合分布 $p\left(C^{\text{pep}} \mid C^{\text{rec}}\right)$必须满足旋转平移等变性，以确保泛化。也就是说，对于任意平移向量 $\boldsymbol{\epsilon} \in \mathbb{R}^3$ 和任意正交矩阵 $O \in \mathbb{R}^{3\times3}$，它都应满足：

$$p\left(OC^{\text{pep}} + \boldsymbol{\epsilon} \mid OC^{\text{rec}} + \boldsymbol{\epsilon}\right), \tag{25}$$

其中 $OC^{\text{pep}} + \boldsymbol{\epsilon} = \left\{\left(a_j, O_j, O\boldsymbol{x}_j + \boldsymbol{\epsilon}, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\text{pep}}} \cup \left\{\left(O\boldsymbol{x}_i^s + \boldsymbol{\epsilon}, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\right)\right\}_{i=1}^{m}$。通过采用零质心处理，我们从所有输入坐标中减去受体的质心，从而实现平移不变性，同时也提高了训练稳定性。此外，可以证明，当先验分布 $p(a_0)$、$p(O_0)$、$p(\boldsymbol{x}_0)$、$p(\boldsymbol{\chi}_0)$、$p(\boldsymbol{x}_0^s)$、$p(\boldsymbol{n}_0^s)$、$p(\boldsymbol{\tau}_0^s)$和 $p(\Upsilon_0^s)$均为 SE(3)不变，向量场 $v^{\text{cat}}(\cdot)$和 $v^{\text{con}}(\cdot)$为 SE(3)不变，向量场 $v^{\text{ori}}(\cdot)$为 SO(3)等变且 T(3)不变，并且向量场 $v^{\text{pos}}(\cdot)$为 SE(3)等变时，由 ODE采样过程生成的密度 $p(C^{\text{pep}} \mid C^{\text{rec}})$具有 SE(3)等变性。值得注意的是，采用 IPA 和 EGNN 可确保这些向量场满足等变性和不变性要求。

条件设计。 近年来，控制 Diffusion 或流等深度生成模型的输出已成为研究热点。除了受体 C_{rec}之外，我们通常还希望创建满足特定条件的肽，例如环化性和二硫键。传统上，条件控制包括分类器引导 [13] 和无分类器引导 [21]。证据表明，只要有足够的高质量训练数据，无分类器引导往往能取得更好的结果，而且无需重新训练分类器架构，就能生成几乎无限种类别的样本。因此，无分类器引导在现代人工智能系统 [48, 74]中得到了广泛应用。鉴于这一优势，我们采用无分类器引导来实现肽条件的插入 c 并将 $p(C_{\text{pep}}|C_{\text{rec}})$ 迁移至 $p(C_{\text{pep}}|C_{\text{rec}},\ c)$。具体而言，按照惯例，我们用 $\emptyset$ 表示空条件，并设 $p(C_{\text{pep}}|C_{\text{rec}},\ \emptyset) := p(C_{\text{pep}}|C_{\text{rec}})$ 和 $u_t(\cdot|\cdot\ \emptyset) := u_t(\cdot|\cdot)$。以表面点位置为例，我们的训练损失变为 [74]:

$$\mathcal{L}_{\text{pos}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),b,p\left(\boldsymbol{x}_1^s\right),p\left(\boldsymbol{x}_0^s\right),p\left(\boldsymbol{x}_t^s|\boldsymbol{x}_0^s,\boldsymbol{x}_1^s\right)} \left\| v^{\text{pos}}\left(\boldsymbol{x}_t^s, t, C^{\text{rec}}|(1-b)\cdot c + b\cdot\emptyset\right) - \left(\boldsymbol{x}_1^s - \boldsymbol{x}_0^s\right)\right\|_2^2, \tag{26}$$

其中 $b \sim$ 伯努利分布(p_{uncond})表示使用空条件的概率。

B 实验细节

B.1 训练与采样

所有实验均在 4 块 NVIDIA A100 GPU 上实现。具体而言，我们训练 SurfFlow 320K 次迭代，并将初始学习率设为 5e-4。我们使用平台期调度器，因子设为 0.8，耐心值设为 10。最小学习率为 5e-6。每个分布式节点的批大小为 32。我们使用 Adam优化器，并进行梯度裁剪。EGNN 解码器采用 0.15 的丢弃率。不同损失分量的权重分别设为 $\lambda_{\text{pos}} = 0.2$、$\lambda_{\text{ori}} = 0.2$、$\lambda_{\text{cat}} = 1.0$、$\lambda_{\text{con}} = 1.0$ 和 $\lambda_{\text{str}} = 1.0$。对于编码器部分，残基嵌入维度设为 $\psi_{\text{rec}} = \psi_{\text{pair}} = 128$。对于解码器部分，IPA 的节点和边嵌入维度分别设为 128 和 64，而 EGNN 的节点和边嵌入维度分别设为 16 和 8，因为表面点 $m + m'$ 的数量比复合物残基 $n_{\text{pep}} + n_{\text{rec}}$的数量高出几个数量级。我们将生成肽的长度设为与对应天然肽的长度相同。我们从 PepMerge 官方代码仓库下载数据 [33]：
https://drive.google.com/drive/folders/1bHaKDF3uCDPtfsihjZs0zmjwF6UU1uVl。

B.2 物理化学表面特征

实验中使用了三种表面生化性质。具体而言，自由电子和潜在氢键供体（FEPH）是分类变量，而静电性质和疏水性是连续变量。因此，$\psi_\tau = 1$ 和 $\psi_\Upsilon = 2$。我们借助 MaSIF [15]的脚本获取这些表面特征。

自由电子和质子供体。 以氢键势为参照，计算 FEPH 在分子表面上的位置。对于分子表面上最近原子为极性氢、氮或氧的顶点，将其视为氢键中的潜在供体或受体。随后，根据重原子之间的方向，为每个顶点赋予一个取自高斯分布的值。这些初始值的范围为 -1（氢键受体的最佳位置）至 +1（氢键供体的最佳位置）。然后，根据 FEPH 为负还是为正，将该点确定为受体或供体（二元标签）。

疏水性。 根据顶点最近原子所属氨基酸的 Kyte-Doolittle 标度，为每个顶点赋予一个疏水性标量值。这些值在原始标度中的范围为 -4.5（亲水性）至 +4.5（疏水性最强），随后被归一化至 -1 到 1 之间。

***Poisson-Boltzmann* 连续介质静电学。** 使用 PDB2PQR 准备蛋白质文件，以进行静电计算；然后使用 APBS（v.1.5）计算每种蛋白质的 Poisson-Boltzmann 静电性质。使用 APBS 套件中提供的 Multivalue，为网格化表面上的每个顶点赋予相应的电荷。高于 +30 和低于 -30 的电荷值分别截断至这些数值，随后将所有数值归一化至 -1 到 1 之间。

连续表面特征损失函数。 与离散物理化学特征相比，连续特征的损失更容易分析。具体而言，由于连续特征值被归一化至 -1 到 1 之间，我们采用归一化分布 $\mathrm{N}(0,1)$ 作为先验分布。相应地，对于 $\Upsilon_i^s \sim \mathrm{N}(0,1)$，流和向量场为：

$$\phi_t^{\mathrm{con}}\left(\Upsilon_0^s, \Upsilon_1^s\right) = t\Upsilon_1^s + (1-t)\Upsilon_0^s, \tag{27}$$

$$u_t^{\mathrm{con}}\left(\Upsilon_t^s \mid \Upsilon_1^s, \Upsilon_0^s\right) = \Upsilon_1^s - \Upsilon_0^s = \frac{\Upsilon_1^s - \Upsilon_t^s}{1-t}. \tag{28}$$

随后，使用时间依赖的平移不变表面网络 $v^{\mathrm{con}}(\cdot)$，根据当前插值态 Υ_t 和时间步 t预测条件向量场。表面连续物理化学特征的条件流匹配目标函数表述如下：

$$\mathcal{L}_{\mathrm{con}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),p(\Upsilon_1^s),p(\Upsilon_0^s),p(\Upsilon_t^s|\Upsilon_0^s,\Upsilon_1^s)} \left\| v^{\mathrm{con}}\left(\Upsilon_t^s, t, C^{\mathrm{rec}}\right) - \left(\Upsilon_1^s - \Upsilon_0^s\right)\right\|_2, \tag{29}$$

B.3 表面一致性的评估指标

虽然指标中考虑了结构一致性，但尚不清楚生成的蛋白质-肽结构的生成表面与计算得到的蛋白质-肽结构表面的相似程度。受近期分子表面方法的启发 [55],，我们采用三个评估指标：交并比、倒角距离和法向量一致性，用于衡量设计的分子表面与靶蛋白的分子表面之间的相似性。为简单起见，三个指标均归一化至 0 – 1 范围。这些指标从多个角度对模型性能进行了全面评估，具体如下所述。

交并比。 交并比（IoU）用于评估重建体积与真实形状之间的重叠程度（越高越好）。对于任意两个形状 A，$B \subseteq \mathbb{S} \in \mathbb{R}^n$，交并比定义为 $\mathrm{IoU} = \frac{|A\cap B|}{|A\cup B|}$。

倒角距离。 倒角距离（CD）用于量化两个点集 $\mathrm{X}_1, \mathrm{X}_2 \subset \mathbb{R}^n$ 之间的差异（越低越好）。其计算方式为：$\mathrm{d_C}(\mathrm{X}_1, \mathrm{X}_2) = \frac{1}{2}\left(\mathrm{d}\vec{c}(\mathrm{X}_1, \mathrm{X}_2) + \mathrm{d}\vec{c}(\mathrm{X}_2, \mathrm{X}_1)\right)$，其中 $\mathrm{d}\vec{c}(\mathrm{X}_1, \mathrm{X}_2) = \frac{1}{|\mathcal{X}_1|}\sum_{\boldsymbol{x}_1\in\mathcal{X}_1} \min_{\boldsymbol{x}_2\in\mathcal{X}_2} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$。

法向量一致性。 法向一致性（NC）用于评估表面法向的对齐程度（越高越好）。对于两个归一化的单位向量 $\vec{n}_i$ 和 $\vec{m}_j$，法向一致性由它们的点积给出。为了评估预测表面点及其法向 $X_{\mathrm{pred}} = \{(\boldsymbol{x}_i, \vec{n}_i)\}$ 与真实表面点及其法向 $X_{\mathrm{gt}} = \{(\boldsymbol{y}_j, \vec{m}_j)\}$ 之间的表面法向，表面法向一致性 Γ 定义为：$\Gamma\left(X_{\mathrm{gt}}, X_{\mathrm{pred}}\right) = \frac{1}{|X_{\mathrm{gt}}|}\sum_{j\in|X_{\mathrm{gt}}|}\left|\vec{n}_j \cdot \vec{m}_{\theta(\boldsymbol{y}_j,\ X_{\mathrm{pred}})}\right|$，其中 $\theta(\boldsymbol{y}_j, X_{\mathrm{pred}}) = \arg\min_{i\in|X_{\mathrm{pred}}|} \left\|\boldsymbol{y}_j - \boldsymbol{x}_i\right\|_2^2$。

B.4 内部结构的流匹配

公式 17 中的 $\mathrm{L_{str}}(\theta)$ 计算肽结构 $p\left(\left\{\left(a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\right)\right\}_{j=1}^{n_{\mathrm{pep}}} \middle| C^{\mathrm{rec}}\right)$ 的损失。这里，我们简要介绍如何对残基的位置 $\boldsymbol{x}$、类型 a、主链扭转角 O 和侧链角 $\boldsymbol{\chi}$ 进行全原子条件流匹配。具体而言，目标函数为$\phi_t(\boldsymbol{x}_0, \boldsymbol{x}_1) = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$，用于残基的位置 $\boldsymbol{x}$；$\exp_{O_0}\left(t \log_{O_0}(O_1)\right)$，用于主链取向 O；$\phi_t(\boldsymbol{\chi}_0, \boldsymbol{\chi}_1) = [t\boldsymbol{\chi}_1 + (1-t)\boldsymbol{\chi}_0] \bmod 2\pi$，用于侧链角 $\boldsymbol{\chi}$；以及 $\phi_t(\tilde{a}_0, \tilde{a}_1) = t\tilde{a}_1 + (1-t)\,\tilde{a}_0$，用于残基类型 a。其中，$\tilde{a}$ 是通过软独热编码操作对 a 进行的表示，并满足 $\mathrm{logit}(a_i) = \tilde{a}_i \in \mathbb{R}^{20}$。

C 附加结果与可视化 C.1 表面点云演化

我们展示了一些表面点云从时间 $t = 0$ 演化至终止时间 $t = 1$ 的示例。可以看到，经过时间 $t = 0.5$ 后，点云开始形成可辨识的形状。

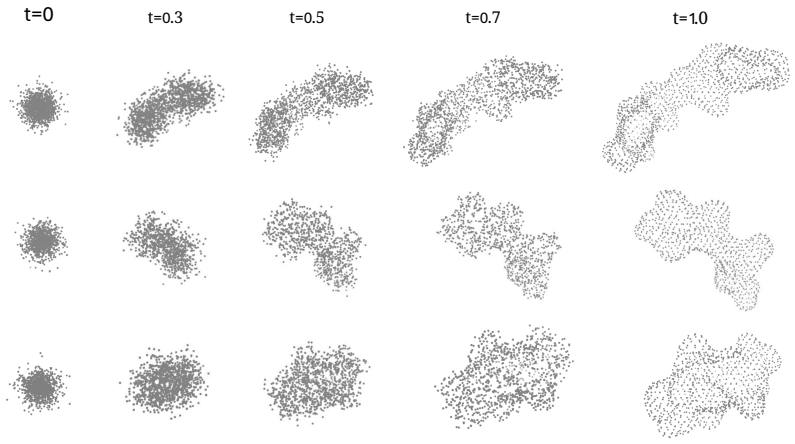


图 6：表面点位置随时间演化的可视化 $t \in [0, 1]$。

C.2 计算效率比较

由于表面特征计算可能带来额外的计算开销，我们将本方法与包括 RFDiffusion 在内的竞争性最先进方法进行了实际运行时间比较。结果汇总于表 5，显示尽管 SurfFlow 因计算表面特征而产生了一定的计算开销，但设计质量的提升足以证明这部分额外耗时是合理的。具体而言，表面特征提取步骤约占总运行时间的 12.8%，但该计算可以在多个核心上高效并行，从而将其在实际应用中的影响降至最低。此外，值得注意的是，RFdiffusion 最初采用 200 个离散时间步进行训练。然而，近期的改进使得减少推理时间步成为可能。RFdiffusion 声称，在许多情况下，使用约 20 步即可获得与使用 200 步时计算机模拟质量相当的输出（速度提升 10 倍）。但为确保公平比较，我们在此仍使用 200 步。此分析凸显了流模型的速度优势，这与 FrameFlow 的发现一致 [69]。它还表明了 SurfFlow 在计算成本与设计质量之间的权衡，强调了丰富特征表示对于推进蛋白质生成建模的重要性。

表 5：不同方法在 PepMerge测试集上的推理时比较。每种方法均在一块配备 80GB 显存的 NVIDIA H100 GPU 上以默认设置运行。

模型	推断步数	时间（分钟）
RFdiffusion	200	182.5
PepFlow	200	25.6
SurfFlow	200	28.9

D 局限性与未来工作

尽管我们的 SurfFlow 改进了原有的全原子设计机制，但仍有进一步探索的空间。例如，如果将受体表面的信息纳入联合分布建模，预计还能取得进一步改进。也就是说，我们的目标变为 $C_{\text{rec}} = \{a_j, O_j, \boldsymbol{x}_j, \boldsymbol{\chi}_j\}_{j=1}^{n_{\text{rec}}} \cup \{\boldsymbol{x}_i^s, \boldsymbol{n}_i^s, \boldsymbol{\tau}_i^s, \Upsilon_i^s\}_{i=1}^{m'}$，其中 m' 是受体表面点的数量。此外，RFDiffusion 的成功表明，在 PDB 中的常规蛋白质上进行预训练可能会有所裨益。