中国科学: 数学 2025 年 第 55 卷 第 7 期: 1505 ~ 1526
SCIENTIA SINICA Mathematica
《中国科学》杂志社
SCIENCE CHINA PRESS
综 述

扩散模型及其在生物信息学中的应用

献给钱敏平教授 86 寿辰

吴泽宇[1†], 付艺伟[1†], 陈佳晓[2†], 马健文[2], 吴文睿[1], 邓明华[1,2*]

1. 北京大学数学科学学院, 北京 100871;
2. 北京大学定量生物学中心, 北京 100871
† 同等贡献
E-mail: 2301110086@pku.edu.cn, fuyw@stu.pku.edu.cn, chenjx@stu.pku.edu.cn, jianwen_ma@stu.pku.edu.cn, 2401110085@stu.pku.edu.cn, dengmh@math.pku.edu.cn

收稿日期: 2024-10-15; 接受日期: 2025-03-05; 网络出版日期: 2025-05-21; * 通信作者
国家自然科学基金 (批准号: 32270689) 资助项目

摘要 近年来, 扩散模型备受瞩目, 在计算机视觉、自然语言处理和生物信息学等领域取得了显著成果, 展现出广阔的应用前景. 本文旨在阐述扩散模型的概率学意义、扩散模型的发展过程并详细列举其在生物信息学中的应用. 首先从随机微分方程的角度定义扩散模型的前向过程和反向过程, 结合得分匹配的推导, 详细阐述去噪扩散概率模型等三种扩散模型的概率学原理. 由于其强大的生成能力, 扩散模型在生物信息学中得到了广泛的应用, 特别是在转录组和蛋白质研究领域. 扩散模型不仅在单细胞转录组和空间转录组测序数据填补与去噪方面表现突出, 同时在蛋白质设计、结构预测以及小分子和多肽药物设计中效果显著.

关键词 扩散模型 随机微分方程 转录组插补 蛋白质设计 药物设计

MSC (2020) 主题分类 60J60, 60H30, 62P10, 92B05

1 引言

生成模型, 尤其是大语言模型和扩散模型, 彻底地改变了人工智能 (artificial intelligence, AI) 领域. 扩散模型是一种生成模型, 它通过先加噪后去噪的方式来实现数据的生成. 作为扩散模型 (diffusion probabilistic model, DPM) 领域的佼佼者, OpenAI 公司开发的 Sora 结合自然语言处理模块, 能够从文本等提示中生成图片、视频等图像数据, 创造了图像生成领域的里程碑.

在扩散模型出现之前, 图像生成任务已经引起人们的广泛关注, 衍生出了众多模型. 变分自编码器 (variational autoencoder, VAE) 由 Kingma 和 Welling [32] 在 2013 年提出, 通过拟合隐空间的分布来生成样本, 拥有优越的理论性质. 2014 年, Goodfellow 等 [18] 提出了对抗生成式网络 (generative

英文引用格式: Wu Z Y, Fu Y W, Chen J X, et al. The diffusion model and its applications in bioinformatics (in Chinese). Sci Sin Math, 2025, 55: 1505–1526, doi: 10.1360/SSM-2024-0316

adversarial networks, GAN), 通过判别器和生成器对抗的方式使得生成器产出的样本和真实样本几乎无差别. GAN 的效果很大程度上超越了 VAE, 迅速成为图像生成任务的主流模型.

最早的扩散模型由 Sohl-Dickstein 等 [59] 在 2015 年提出. 他们受到非平衡热力学的启发, 提出了用三个步骤来生成新的数据. 第一步是将数据逐步加入噪声, 从而破坏原有的数据结构, 这一步骤称为扩散过程 (diffusion process, forward process); 第二步是通过原数据和扩散过程生成的数据, 学习一个包含原数据信息的反向扩散网络; 第三步是将加噪后的数据通过反向扩散网络逐步复原, 使得复原后的数据和原数据分布接近, 这一步骤称为反向过程 (reverse process). 模型在 CIFAR-10 等实际数据集上的生成效果并不卓越, 因此并没有引起足够的关注. 2020 年, Ho 等 [21] 提出了去噪扩散概率模型 (denoising diffusion probabilistic model, DDPM), 在文献 [59] 的基础上引入了新的损失函数, 并使用了 U-net 的网络架构拟合反向传播过程. 该模型在 CIFAR-10 和 LSUN 等数据集上生成的样本质量接近于 GAN. 由于生成的样本种类更加丰富, 视觉体验更佳, 因此吸引了许多学者的关注. 随着扩散模型的应用日益广泛, DDPM 训练成本和生成成本过高等问题也随之引起了人们的重视. 随后, 多种扩散模型相继被提出 (参见文献 [41,55,60]), 旨在解决这些问题, 进一步推动了扩散模型在计算机视觉、自然语言处理和生物信息学等领域的广泛应用.

扩散模型凭借其强大的生成能力和上下文理解能力, 在生物信息领域也展现出广泛的应用前景, 尤其是在转录组分析和蛋白质相关任务中 (参见文献 [20]). 在转录组数据分析中, 测得的基因表达数据通常伴有噪声 (参见文献 [27]), 而扩散模型能够有效填补缺失值并生成高质量单细胞转录组数据 (参见文献 [43,56,68,81]). 此外, 扩散模型还可用于对空间转录组数据进行去噪和提升图像分辨率, 同时利用离散切片进行组织的三维重构 (参见文献 [31,35,74]). 扩散模型在转录组数据分析中的广泛应用, 不仅提高了转录组数据的精度和可靠性, 还为探索基因调控网络和疾病机制等方面提供了新的工具, 推动了生物信息学和生物医学研究的发展. 在蛋白质分析领域, 扩散模型在无条件蛋白质设计、结合剂设计、对称结构构建、酶活性位点优化等任务中展现出卓越的性能 (参见文献 [7,34]). 扩散模型促使了高效的蛋白质结构和序列设计算法的开发 (参见文献 [2,70,76,77]), 从而生成符合生物物理规律的创新型蛋白质, 并且在实验验证中表现出高度的准确性和可操作性. 此外, 扩散模型还显著提高了蛋白质复合物结构预测的准确性, 并推动了小分子和多肽药物的设计 (参见文献 [9,24,25,53,78]). 这一系列进展显著加速了蛋白质工程以及新型药物分子的开发进程.

本文从算法层面和应用层面两方面对扩散模型进行总结. 在算法层面, 本文主要从随机微分方程、得分匹配方法和随机过程视角对扩散模型进行概率学阐述. 此外, 本文还从采样加速、架构改进、适用范围扩展等多个维数概述扩散模型的发展和改进. 在应用层面, 本文主要从转录组和蛋白质两个维数来概括扩散模型在生物信息学中的应用, 尤其是扩散模型在转录组数据插补、空间转录组图像分辨率提升、蛋白质设计和药物设计等方面的应用.

2 随机微分方程和得分匹配简介

形如 $d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, t)dt + \boldsymbol{G}(\boldsymbol{X}_t, t)d\boldsymbol{B}_t$ 的微分方程称为随机微分方程 (stochastic differential equations, SDE), 其中 $t \in [0, T]$ 表示时间, $\boldsymbol{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ 是 d 维漂移向量, $\boldsymbol{G}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ 为扩散张量, $\boldsymbol{B}_t$ 为 d 维标准 Brown 运动.

2.1 逆转时间方程

随机微分方程的边际分布 $p_t(\boldsymbol{X}_t)$ 满足以下的微分方程:

$$\partial_t p_t(\boldsymbol{x}) = -\sum_{i=1}^{d} \partial_{x_i}[\boldsymbol{f}^i(\boldsymbol{x},t)p_t(\boldsymbol{x})] + \frac{1}{2}\sum_{i,j=1}^{d} \partial_{x_i}\partial_{x_j}\left[\sum_{k=1}^{d} \boldsymbol{G}^{ik}(\boldsymbol{x},t)\boldsymbol{G}^{jk}(\boldsymbol{x},t)p_t(\boldsymbol{x})\right]. \tag{2.1}$$

1982 年, Anderson [3] 根据正向随机过程的边际分布 $p_t(\boldsymbol{X}_t)$ 给出逆转时间过程 $\bar{\boldsymbol{X}}_t$ 的构造. 如果 $\bar{\boldsymbol{X}}_t$ 满足如下逆转时间随机微分方程 (reverse-time SDE):

$$d\bar{\boldsymbol{X}}_t = \{\boldsymbol{f}(\bar{\boldsymbol{X}}_t,t) - \nabla\cdot[\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)^{\mathsf{T}}] - \boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)^{\mathsf{T}}\nabla\log p_t(\bar{\boldsymbol{X}}_t)\}dt + \boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)d\overline{\boldsymbol{B}}_t, \tag{2.2}$$

则对任意 $t\in[0,T]$ 有 $\bar{\boldsymbol{X}}_t \sim \boldsymbol{X}_t$, 其中 $\overline{\boldsymbol{B}}_t$ 是反向 Brown 运动. 值得注意的是, 随机过程 $\bar{\boldsymbol{X}}$ 与 $\boldsymbol{X}$ 并不同分布, 仅是在任意时刻 t 它们的边际分布相同, 而扩散模型恰好只关注 t 时刻的边际分布.

除了利用逆转时间 SDE 构造逆转时间过程, 也可以利用 ODE (ordinary differential equation) 构造逆转时间过程, 形式如下:

$$d\bar{\boldsymbol{X}}_t = \left\{\boldsymbol{f}(\bar{\boldsymbol{X}}_t,t) - \frac{1}{2}\nabla\cdot[\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)^{\mathsf{T}}] - \frac{1}{2}\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)\boldsymbol{G}(\bar{\boldsymbol{X}}_t,t)^{\mathsf{T}}\nabla\log p_t(\bar{\boldsymbol{X}}_t)\right\}dt. \tag{2.3}$$

逆转时间方程的出现给人们带来了反向生成给定随机过程的可能. 人们尝试通过正向过程从分布 p_0 中抽样 (p_0 是未知的, 但是可以从初始数据经验分布中抽样) 得到初始数据, 然后将样本通过正向随机过程转变成某种与数据无关的分布 p_T, 如标准正态分布和二项分布等. 接着再从分布 p_T 中抽取样本, 将样本通过逆转时间方程回到 $t=0$ 时刻, 此时这些样本服从 p_0 分布, 可以视为从初始数据取样. 假设正向随机过程、逆转时间过程的方程参数已知, 则通过这种方式生成样本的优势有以下两点.

(1) 可实现的: 上面的每一个步骤都由确定性方程给出, 不涉及对分布进行归一化等不确定操作;

(2) 灵活的: 生成的样本并不局限在初始数据, 或者由初始数据简单变换得到, 采用这种方式可以生成各式各样的样本.

2.2 得分匹配方法

在实际问题中, $\boldsymbol{f}(\boldsymbol{X}_t,t)$ 和 $\boldsymbol{G}(\boldsymbol{X}_t,t)$ 往往是预先设定好的, 所以正向随机过程中不含有任何未知参数. 反向过程包含的未知参数是 $\nabla\log p_t$, 这一项是对数似然的导数, 在统计学中被称为得分 (score). 得分不能直接由 $\boldsymbol{f}$ 和 $\boldsymbol{G}$ 的信息计算得到, 因为其包含了初始分布 p_0 的信息. 计算某分布得分的过程称作得分匹配 (score matching). 我们使用函数 $\boldsymbol{s}_\theta(\boldsymbol{X},t)$ 对 $\nabla\log p_t(\boldsymbol{X})$ 进行估计, 使用的损失函数为

$$\mathcal{L}(\theta) = \int_0^T \lambda(t)\mathbb{E}_{\boldsymbol{X}_t\sim p_t}[\|\boldsymbol{s}_\theta(\boldsymbol{X}_t,t) - \nabla_{\boldsymbol{X}_t}\log p_t(\boldsymbol{X}_t)\|^2]dt. \tag{2.4}$$

这里使用二范误差的平均作为误差估计, 其中 $\lambda(t)$ 是已知的参数, 代表时间的权重. 由于得分本身是未知的, 需要改变 $\mathcal{L}(\theta)$ 右边项的形式. 目前有两种主流的方法可以解决这个问题.

第一种方法是去噪得分匹配 (denoising score matching), 其由 Vincent [72] 于 2011 年提出, 运用加噪的方法将不好计算的得分转变为特定分布的得分. 设 $p_{t|0}(\boldsymbol{X}_t\,|\,\boldsymbol{X}_0)$ 为给定 $\boldsymbol{X}_0$ 下 $\boldsymbol{X}_t$ 的分布密度, 则可以最小化下面的损失函数对得分进行估计:

$$\mathcal{L}(\theta) = \sum_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{X}_0}[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}[\boldsymbol{s}_\theta(\boldsymbol{X}_t,t) - \nabla_{\boldsymbol{X}_t}\log p_{t|0}(\boldsymbol{X}_t\,|\,\boldsymbol{X}_0)]^2\,|\,\boldsymbol{X}_0]\,dt + C. \tag{2.5}$$

2020 年, Song 等 [63] 提出了一种新颖的得分匹配方法, 称为切片得分匹配. 相比于去噪得分匹配, 该方法适用性更广. 假设 $\boldsymbol{v}\in\mathbb{R}^n$, $\mathbb{E}_\nu[\nu_i\nu_j] = \delta_{ij}$, 则有

$$\mathcal{L}(\theta) = \int_0^T \lambda(t)\mathbb{E}_{\boldsymbol{X}_t}\left\{\|s_\theta(\boldsymbol{X}_t,t)\|^2 + 2\mathbb{E}_\nu\left[\frac{d}{dh}\nu^{\mathsf{T}}s_\theta(\boldsymbol{X}_t+h\nu,t)\Big|_{h=0}\right]\right\}dt + C. \tag{2.6}$$

通常情况下, 切片得分匹配比去噪得分匹配更加通用, 因为后者需要 $\boldsymbol{X}_t \mid \boldsymbol{X}_0$ 具有较好的形式. 实际情况中, 如果去噪得分匹配可以使用, 则它的表现会优于前者. 有两类特殊的随机微分方程能让 $\boldsymbol{X}_t \mid \boldsymbol{X}_0$ 简洁, 分别是方差爆炸随机微分方程 (variance exploding SDE, VE SDE) 和方差保留随机微分方程 (variance preserving SDE, VP SDE).

2.3 VE SDE 和 VP SDE

若 $d\boldsymbol{X}_t = \Sigma(t)d\boldsymbol{B}_t$, 其中 $\Sigma(t)$ 是 d 阶正定矩阵, 则 $\boldsymbol{X}_t$ 满足

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \Sigma(s)d\boldsymbol{B}_s. \tag{2.7}$$

根据 Itô 等距公式, 可以计算得 $\mathrm{Var}(\boldsymbol{X}_t) = \mathrm{Var}(\boldsymbol{X}_0) + \int_0^t \Sigma(s)\Sigma(s)^\top ds$. 通常情况下, $\int_0^t \Sigma(s)\Sigma(s)^\top ds$ 不收敛, 因此这类 SDE 被称为 VE SDE. 这类 SDE 中 $\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N(\boldsymbol{X}_0, \int_0^t \Sigma(s)\Sigma(s)^\top ds)$, 服从保均值的正态分布.

特别地, 若 $\Sigma(t) = \sqrt{\frac{d\sigma^2(t)}{dt}}\boldsymbol{I}$, 则

$$\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N(\boldsymbol{X}_0, \sigma^2(t)). \tag{2.8}$$

与 VE SDE 相比, VP SDE 的微分形式略有不同. 若 $d\boldsymbol{X}_t = -\frac{1}{2}\boldsymbol{A}(t)\boldsymbol{X}_t dt + \Sigma(t)d\boldsymbol{B}_t$, 其中 $\boldsymbol{A}(t)$ 为 d 阶对称矩阵, $\Sigma(t)$ 是 d 阶正定矩阵, 则 $\boldsymbol{X}_t$ 满足

$$\boldsymbol{X}_t = \exp\left\{-\frac{1}{2}\int_0^t \boldsymbol{A}(w)dw\right\}\boldsymbol{X}_0 + \exp\left\{-\frac{1}{2}\int_0^t \boldsymbol{A}(w)dw\right\}\int_0^t \exp\left\{\frac{1}{2}\int_0^s \boldsymbol{A}(w)dw\right\}\Sigma(s)d\boldsymbol{B}_s. \tag{2.9}$$

类似地, 可以计算 $\boldsymbol{X}_t$ 的方差

$$\begin{aligned}\mathrm{Var}(\boldsymbol{X}_t) = \exp\left\{-\frac{1}{2}\int_0^t \boldsymbol{A}(w)dw\right\}&\left(\mathrm{Var}(\boldsymbol{X}_0) + \int_0^t \exp\left\{\int_0^s \boldsymbol{A}(w)dw\right\}\Sigma^2(s)ds\right)\\ &\times \exp\left\{-\frac{1}{2}\int_0^t \boldsymbol{A}(w)dw\right\}.\end{aligned} \tag{2.10}$$

特别地, 若 $\boldsymbol{A}(t) = \beta(t)\boldsymbol{I}$, $\Sigma(t) = \sqrt{\beta(t)}$, 则

$$\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N\left(\exp\left\{-\frac{1}{2}\int_0^t \beta(w)dw\right\}\boldsymbol{X}_0, \left(1 - \exp\left\{-\int_0^t \beta(w)dw\right\}\right)\boldsymbol{I}\right). \tag{2.11}$$

3 随机过程视角下的扩散模型

以噪声条件得分网络 (noise conditional score networks, NCSN)、DDPM 和扩散隐式去噪模型 (denoising diffusion implicit models, DDIM) 为代表的经典扩散模型在许多研究领域发挥着不可替代的作用. 为了更好地理解这些扩散模型, Song 等 [64] 对这些扩散模型作了分类与阐释. 本节依照 Song 等的思路, 以 SDE 和反向 SDE 得分匹配的视角对这三个模型作出概率学解释.

3.1 NCSN

NCSN 模型由 Song 和 Ermon[62] 在 2019 年提出. 模型主要由去噪得分匹配部分和退火 Langevin 方法 (annealed Langevin dynamics) 两部分组成. Langevin 方法, 又称 SGLD (stochastic gradient Langevin dynamics) 方法, 通过过阻尼 Langevin Itô 过程 $d\boldsymbol{X}_t = \frac{1}{2}\nabla \log p_0(\boldsymbol{X}_t)dt + d\boldsymbol{B}_t$ 采样出近似 p_0 分布的样本. 这其中需要用到初始数据分布的得分 $\nabla \log p_0(\boldsymbol{X})$. 他们通过给初始数据逐步加噪, 再利用退火 Langevin 算法实现了 NCSN.

具体而言, 假设 $0 < \sigma_1 < \sigma_2 < \cdots < \sigma_T$, 第 t 步数据加噪将 $N(0, \sigma_t^2)$ 的噪声加入初始数据, 即 $\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N(\boldsymbol{0}, \sigma_t^2)$, 生成新分布 p_t. 接着通过退火 Langevin 算法, 第 i 步通过 L 次 Langevin Monte Carlo 方法采样出近似由 p_{T-i} 生成的样本. 第 i 步中只用到对 $\nabla \log p_{T-i}(\boldsymbol{X})$ 的估计. 这可以通过去噪得分匹配最小化损失函数得到. 其中损失函数为

$$\mathcal{L}(\theta) = \sum_{t=0}^{T} \lambda(t) \mathbb{E}_{\boldsymbol{X}_0}\left[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}\left[\boldsymbol{s}_\theta(\boldsymbol{X}_t, t) + \frac{\boldsymbol{X}_t - \boldsymbol{X}_0}{\sigma_t^2}\right]^2 \middle| \boldsymbol{X}_0\right] dt + C. \tag{3.1}$$

$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X}, t)$ 为 $\nabla \log p_t(\boldsymbol{X})$ 的估计, 这样规避了对初始分布得分 $\nabla \log p_0(\boldsymbol{X})$ 的估计过程. 具体采样方法见算法 1.

算法 1 NCSN 采样

1: **输入**: $\{\sigma_i\}_{i=1}^L, \epsilon, T$
2: **初始化** $\bar{\boldsymbol{X}}_0$
3: **循环开始** t 从 T 到 1:
4: $\alpha_t \leftarrow \epsilon \cdot \sigma_t^2/\sigma_T^2$ (α_t 表示步长)
5: **循环开始** i 从 1 到 L:
6: $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, I)$
7: $\bar{\boldsymbol{X}}_i \leftarrow \bar{\boldsymbol{X}}_{i-1} + \frac{\alpha_t}{2}\boldsymbol{s}_{\boldsymbol{\theta}}(\bar{\boldsymbol{X}}_{i-1}, t) + \sqrt{\alpha_t}\boldsymbol{z}_t$
8: **循环结束**
9: $\bar{\boldsymbol{X}}_0 \leftarrow \bar{\boldsymbol{X}}_L$
10: **循环结束**
11: **返回** $\bar{\boldsymbol{X}}_0$

从另一角度, NCSN 是以 VE SDE 为扩散过程的离散化扩散模型. 考虑扩散过程

$$d\boldsymbol{X}_t = \sqrt{\frac{d\sigma^2(t)}{dt}}\boldsymbol{I}d\boldsymbol{B}_t,$$

由 (2.8) 可知 $\boldsymbol{X}_t|\boldsymbol{X}_0 \sim N(\boldsymbol{X}_0, \sigma^2(t))$. 将去噪得分匹配的损失函数离散化, 得到

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{X}_0}[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}[\boldsymbol{s}_\theta(\boldsymbol{X}_t, t) - \nabla_{\boldsymbol{X}_t}\log p_{t|0}(\boldsymbol{X}_t \mid \boldsymbol{X}_0)]^2 \mid \boldsymbol{X}_0]\, dt + C \\ &= \sum_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{X}_0}\left[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}\left[\boldsymbol{s}_\theta(\boldsymbol{X}_t, t) + \frac{\boldsymbol{X}_t - \boldsymbol{X}_0}{\sigma_t^2}\right]^2 \middle| \boldsymbol{X}_0\right] dt + C\end{aligned} \tag{3.2}$$

与 NCSN 的损失函数吻合. 不过在反向过程中, NCSN 用的是退火 Langevin 采样, 而不是直接利用逆转时间方程. 如果采用后者, 则利用 $\boldsymbol{s}_\theta(\boldsymbol{X}_t, t)$ 对 $\nabla \log p_t(\boldsymbol{X}_t)$ 的估计, 由逆转时间方程, 即可得到反向过程为

$$d\boldsymbol{y}_t = \{\boldsymbol{f}(\boldsymbol{y}_t, t) - \nabla \cdot [\boldsymbol{G}(\boldsymbol{y}_t, t)\boldsymbol{G}(\boldsymbol{y}_t, t)^{\mathsf{T}}] - \boldsymbol{G}(\boldsymbol{y}_t, t)\boldsymbol{G}(\boldsymbol{y}_t, t)^{\mathsf{T}}\nabla \log p_t(\boldsymbol{y}_t)\}dt + \boldsymbol{G}(\boldsymbol{y}_t, t)d\overline{\boldsymbol{B}}_t$$

$$= -\frac{d\sigma^2(t)}{dt}\nabla \log p_t(\boldsymbol{y}_t)dt + \sqrt{\frac{d\sigma^2(t)}{dt}}d\overline{\boldsymbol{B}}_t. \tag{3.3}$$

离散化后反向过程的系数与 NCSN 有所差别.

3.2 DDPM

DDPM 是 Ho 等 [21] 2020 年提出的扩散模型, 是第一个拥有高质量图片生成能力的扩散模型, 也是最泛用的扩散模型之一. DDPM 的扩散过程和反向过程分别定义为

$$q(\boldsymbol{X}_{1:T}\boldsymbol{X}_0) := \prod_{t=1}^{T} q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}), \tag{3.4}$$

$$q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}) := \mathcal{N}(\boldsymbol{X}_t; \sqrt{1-\beta_t}\boldsymbol{X}_{t-1}, \beta_t \boldsymbol{I}), \tag{3.5}$$

$$p_\theta(\boldsymbol{X}_{0:T}) := p(\boldsymbol{X}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t), \tag{3.6}$$

$$p_\theta(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t) := \mathcal{N}(\boldsymbol{X}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{X}_t, t), \Sigma_\theta(\boldsymbol{X}_t, t)). \tag{3.7}$$

通过优化变分下界 (evidence lower bound) 的方法, DDPM 的损失函数为

$$\begin{aligned} L(\theta) &:= \mathbb{E}_{x\sim q}\bigg[-\log p(\boldsymbol{X}_T) - \sum_{t\geqslant 1}\log\frac{p_\theta(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t)}{q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1})}\bigg] = \mathbb{E}_{x\sim q}\bigg[-\log\frac{p_\theta(\boldsymbol{X}_{0:T})}{q(\boldsymbol{X}_{1:T} \mid \boldsymbol{X}_0)}\bigg] \\ &\geqslant \mathbb{E}_{x_0\sim q(x_0)}[-\log p_\theta(\boldsymbol{X}_0)]. \end{aligned} \tag{3.8}$$

将均值估计项 $\boldsymbol{\mu}_\theta(\boldsymbol{X}_t, t)$ 重参数化为误差估计项 $\boldsymbol{\epsilon}_\theta(\boldsymbol{X}_t, t)$, 方差估计项 $\Sigma_\theta(\boldsymbol{X}_t, t)$ 设计为常数 $\Sigma_\theta(\boldsymbol{X}_t, t) = \gamma_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, $\alpha_t := 1-\beta_t$, $\overline{\alpha}_t := \prod_{s=1}^{t}\alpha_s$, 损失函数化简为

$$L(\theta) = \mathbb{E}_{t,\boldsymbol{X}_0,\boldsymbol{\epsilon}}\bigg[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2\bigg]. \tag{3.9}$$

DDPM 的算法分为训练 (见算法 2) 和采样 (见算法 3) 两部分.

算法 2 DDPM 训练

1: **重复以下过程**:
2: 从数据分布采样 $\boldsymbol{X}_0 \sim q(\boldsymbol{X}_0)$
3: 均匀采样时间步 $t \sim \text{Uniform}(\{1, 2, \ldots, T\})$
4: 采样噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
5: **计算梯度并更新参数**:
6: $\nabla_{\boldsymbol{\theta}}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2$
7: **直到收敛**

注意到 DDPM 是以 VP SDE 为扩散过程的离散化扩散模型. 考虑扩散过程 $d\boldsymbol{X}_t = -\frac{1}{2}\beta(t)\boldsymbol{X}_t dt + \sqrt{\beta(t)}d\boldsymbol{B}_t$, 由 (2.11) 可得

$$\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N\bigg(\exp\bigg\{-\frac{1}{2}\int_0^t \beta(w)dw\bigg\}\boldsymbol{X}_0, \boldsymbol{I} - \exp\bigg\{-\int_0^t \beta(s)ds\bigg\}\boldsymbol{I}\bigg). \tag{3.10}$$

当 $\beta(t)$ 光滑且取值较小时, 扩散过程步长取为 1, 对扩散过程离散化, 可以近似为

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \frac{1}{2}\beta(t)\boldsymbol{X}_t + \sqrt{\beta(t)}\epsilon(t) \approx \sqrt{1-\beta(t)}\boldsymbol{X}_t + \sqrt{\beta(t)}\epsilon(t),$$

算法 3 DDPM 取样

$\bar{\boldsymbol{X}}_T \sim N(\boldsymbol{0}, \boldsymbol{I})$
循环开始 t 从 T 到 1:
如果 $t = 1$, 则 $\boldsymbol{z} = 0$, 否则 $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$
$\bar{\boldsymbol{X}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\bar{\boldsymbol{X}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t)) + \gamma_t \boldsymbol{z}$
循环结束
返回 $\bar{\boldsymbol{X}}_0$

其中 $\epsilon(t) \sim N(0, \boldsymbol{I})$, 这与 DDPM 的扩散过程吻合. 进一步地, $\exp\{-\int_0^t \beta(w)dw\} \approx \exp\{-\sum_{w=0}^t \beta(w)\} \approx \prod_{w=0}^t (1-\beta_w) = \bar{\alpha}_t$, 进而 $\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$, 与 DDPM 扩散过程吻合.

定义 $\sigma_t = \sqrt{1-\bar{\alpha}_t}$, $\epsilon_\theta(\boldsymbol{x}, t) = \frac{1}{\sigma_t}\boldsymbol{s}_\theta(\boldsymbol{x}, t)$. 将去噪得分匹配的损失函数化简, 可得

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{X}_0}[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}[\boldsymbol{s}_\theta(\boldsymbol{X}_t, t) - \nabla_{\boldsymbol{X}_t}\log p_{t|0}(\boldsymbol{X}t \mid \boldsymbol{X}_0)]^2 \mid \boldsymbol{X}_0]\, dt + C \\
&= \sum_{t=0}^{T} \lambda(t)\mathbb{E}_{\boldsymbol{X}_0}\left[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}\left[\frac{1}{\sigma_t}\epsilon_\theta(\boldsymbol{X}_t, t) - \frac{\boldsymbol{X}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{X}_0}{\sigma_t^2}\right]^2 \,\middle|\, \boldsymbol{X}_0\right] dt + C \\
&= \sum_{t=0}^{T} \frac{\lambda(t)}{\sigma_t^2}\mathbb{E}_{\boldsymbol{X}_0}\left[\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}\left[\epsilon_\theta(\boldsymbol{X}_t, t) - \frac{\boldsymbol{X}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{X}_0}{\sigma_t}\right]^2 \,\middle|\, \boldsymbol{X}_0\right] dt + C. \qquad (3.11)
\end{aligned}$$

利用重期望公式以及 $\frac{\boldsymbol{X}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{X}_0}{\sigma_t} \sim N(\boldsymbol{0}, \boldsymbol{I})$ 且与 $\boldsymbol{X}_0$ 独立, 进而上式可以化简为

$$\mathcal{L}(\theta) = \sum_{t=0}^{T} \frac{\lambda(t)}{\sigma_t^2}\mathbb{E}_{\boldsymbol{X}_0,\epsilon}[\epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) - \epsilon]^2\, dt + C, \qquad (3.12)$$

其中 $\epsilon \sim N(\boldsymbol{0}, \boldsymbol{I})$. 可见 (3.12) 与 DDPM 的训练损失函数吻合, 同时 DDPM 的取样步骤可以由逆转时间方程推导得到. 根据逆转时间方程, 可得

$$d\bar{\boldsymbol{X}}_t = \left(-\frac{\beta(t)}{\sigma_t}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) - \frac{\beta(t)}{2}\bar{\boldsymbol{X}}_t\right)dt + \sqrt{\beta(t)}d\bar{\boldsymbol{B}}_t. \qquad (3.13)$$

以步长为 -1 离散化, 结合已有的近似项, 近似得到

$$\begin{aligned}
\bar{\boldsymbol{X}}_{t-1} &= \bar{\boldsymbol{X}}_t + \left(\frac{\beta(t)}{\sqrt{1-\exp(-\int_0^t \beta(s)ds)}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) + \frac{\beta(t)}{2}\bar{\boldsymbol{X}}_t\right) + \sqrt{\beta(t)}\boldsymbol{Z}_t \\
&= \left(1 + \frac{\beta(t)}{2}\right)\bar{\boldsymbol{X}}_t + \frac{\beta(t)}{\sqrt{1-\exp(-\int_0^t \beta(s)ds)}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) + \sqrt{\beta(t)}\boldsymbol{Z}_t \\
&\approx \frac{1}{\sqrt{1-\beta_t}}\left(\bar{\boldsymbol{X}}_t + \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t)\right) + \gamma_t \boldsymbol{Z}_t, \qquad (3.14)
\end{aligned}$$

其中 $\boldsymbol{Z}_t \sim N(\boldsymbol{0}, \boldsymbol{I})$. 这与 DDPM 的逆向过程取样方法吻合.

3.3 DDIM

DDIM 由 Song 等 [60] 提出, 他们将 DDPM 扩散过程中的 Markov 假设摒弃, 换来更高效的训练和取样效率. DDIM 的扩散过程定义如下:

$$q(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T \mid \boldsymbol{X}_0) = q(\boldsymbol{X}_T \mid \boldsymbol{X}_0)\prod_{t=1}^{T-1} q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t+1}, \boldsymbol{X}_0), \qquad (3.15)$$

并且,

$$q(\boldsymbol{X}_T \mid \boldsymbol{X}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_T}\boldsymbol{X}_0, (1-\bar{\alpha}_T)I), \tag{3.16}$$

$$q(\boldsymbol{X}_t \mid \boldsymbol{X}_{t+1}, \boldsymbol{X}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \frac{\sqrt{1-\bar{\alpha}_t-\rho_{t+1}^2}}{\sqrt{1-\bar{\alpha}_{t+1}}}(\boldsymbol{X}_{t+1} - \sqrt{\bar{\alpha}_{t+1}}\boldsymbol{X}_0), \rho_{t+1}^2 I\right), \tag{3.17}$$

其中 $\rho_t^2 \geqslant 0$ 为参数. 尽管形式上与 DDPM 不一样, 但两者扩散过程的边际分布都满足 $\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim N(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$. 这样的前向过程设计能够更好地匹配正反过程, DDIM 和 DDPM 拥有同样的变分下界, 由变分下界推导得到的损失函数都是

$$L(\theta) = \mathbb{E}_{t,\boldsymbol{X}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2\right]. \tag{3.18}$$

因此, DDIM 的训练过程与 DDPM 相同. 但是由于 DDIM 扩散过程与 DDPM 的不同, 因此两者的反向过程有本质的区别. DDIM 的反向过程见算法 4.

算法 4 DDIM 取样

$\bar{\boldsymbol{X}}_T \sim N(\boldsymbol{0}, \boldsymbol{I})$
循环开始 t 从 T 到 1:
$\hat{\boldsymbol{X}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\overline{\boldsymbol{X}}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_\theta(\overline{\boldsymbol{X}}_t, t)$
如果 $t=1$, 则 $\boldsymbol{z}=0$, 否则 $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$
$\overline{\boldsymbol{X}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{X}}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\rho_t^2}\epsilon_\theta(\overline{\boldsymbol{X}}_t, t) + \rho_t\boldsymbol{z}$
循环结束
返回 $\bar{\boldsymbol{X}}_0$

可以验证, 当 $\rho_t = \gamma_t$ 时, DDIM 的取样过程与 DDPM 一致. 当 $\rho_t = 0$ 时, DDIM 的取样过程不包含随机项. 此时模型被称为确定性 DDIM (deterministic DDIM), 是最常用的 DDIM 模型. 与 DDPM 相比, 它的确定性采样能节约很多时间.

确定性 DDIM 可以视为以 VP SDE 为正向过程, 而且利用逆转时间 ODE 为反向过程的扩散模型. 由逆转时间 ODE 方程, 可得

$$\bar{\boldsymbol{X}}_t = \left(-\frac{\beta(t)}{2\sigma_t}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) - \frac{\beta(t)}{2}\bar{\boldsymbol{X}}_t\right)dt. \tag{3.19}$$

以步长为 -1 离散化, 近似得到

$$\begin{aligned}
\bar{\boldsymbol{X}}_{t-1} &= \bar{\boldsymbol{X}}_t + \left(\frac{\beta(t)}{2\sqrt{1-\exp(-\int_0^t\beta(s)ds)}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) + \frac{\beta(t)}{2}\bar{\boldsymbol{X}}_t\right) \\
&= \left(1+\frac{\beta(t)}{2}\right)\bar{\boldsymbol{X}}_t + \frac{\beta(t)}{2\sqrt{1-\exp(-\int_0^t\beta(s)ds)}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) \\
&\approx \frac{1}{\sqrt{1-\beta_t}}\bar{\boldsymbol{X}}_{t-1} + \frac{\beta_t}{2\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\bar{\boldsymbol{X}}_t, t) \\
&\approx \frac{1}{\sqrt{1-\beta_t}}\overline{X}_t - \left(\frac{\sqrt{1-\overline{\alpha}_t}}{\sqrt{1-\beta_t}} - \sqrt{1-\frac{\overline{\alpha}_t}{1-\beta_t}}\right)\epsilon_\theta(\overline{X}_t, t).
\end{aligned} \tag{3.20}$$

这与确定性 DDIM 的取样过程吻合.

综上, 我们从 SDE 视角对 NCSN, DDPM 和 DDIM 三个经典的扩散模型作了概率学的解释.

4 扩散模型的发展

自从 DDPM 问世以来, 扩散模型的潜力开始被广泛挖掘. 经典的扩散模型能生成高质量的、多种类的样本. 我们以经典的 DDPM 为例来理解扩散模型生成过程. 在生成任务中, 我们希望能够生成接近初始训练数据集的样本, 但又与这些样本有所区分. 例如, 在人脸生成任务中, 希望通过训练数据中各种人脸的图像, 形成对人脸特性的统一认知, 进而生成出新的人脸图像. 扩散模型给出的方法是将训练样本视作来源于某一分布 q_0 的抽样, 现在需要从训练样本估计 q_0 并且从该分布采样.

从标准正态分布中采样是容易实现的, DDPM 将分布 q_0 抽样的每个独立样本 $\boldsymbol{X}_0$ 分 T 步加入正态噪声, 得到噪声逐步增大样本轨迹 $\boldsymbol{X}_1,\dots,\boldsymbol{X}_T$, 直至 q_0 相关的信息被抹除, 即 $\boldsymbol{X}_T$ 视作从纯噪声 $N(\boldsymbol{0},\boldsymbol{I})$ 中的采样.

接着可以统计在第 t 步由初始样本 $\boldsymbol{X}_0$ 加噪得到的样本 $\boldsymbol{X}_t$, 通过最小化损失函数 (3.9), 使得噪声估计函数 $\boldsymbol{\epsilon}_\theta$ 准确涵盖了 $\boldsymbol{X}_t$ 分布的信息. 最后通过从纯噪声 $N(\boldsymbol{0},\boldsymbol{I})$ 中采样 $\bar{\boldsymbol{X}}_T$, 利用 $\boldsymbol{\epsilon}_\theta$ 逐步还原为 $\bar{\boldsymbol{X}}_0$.

OpenAI 的 Dhariwal 和 Nichol[12] 在不同数据集的图像生成任务中比较了扩散模型和经典的对抗生成网络的效果, 表 1 展示了部分实验的结果. 可见, 扩散模型的生成效果十分突出. 然而, 扩散模型的加噪和去噪过程需要的步数很高, 因此有训练时间长、生成时间长和内存占据大等诸多问题, 人们在解决这些问题的过程中推动了扩散模型的发展.

目前的改进模型大致可以分为三类: 第一类是针对扩散模型推理时间较长问题的加速方案, 第二类是通过修改模型架构以提升其性能, 第三类改进则是旨在扩展模型的应用范围.

4.1 采样加速

现有的采样加速方法主要分为两大类: 基于训练的加速方法和免于训练的加速方法. 基于训练的加速方法包括知识蒸馏和训练策略的优化等, 虽然这些方法通常能够显著提升性能, 但需要重新训练模型, 代价较高. 然而, 这类方法的效果通常也更加出色. 相比之下, 免于训练的加速方法无需重新训练模型, 能够作为插件应用于任何已经训练完成的模型, 且同样能够达到较好的效果. 因此, 免于训练的加速方法应用得更加广泛. 本文的重点也将放在免于训练的加速方法上.

免于训练的采样加速方法主要分为两类: SDE-solvers 和 ODE-solver. 相较于 SDE-solver, ODE-solver 在加速效果上更为显著, 并且能够借助数值分析领域中已有的各种解算方法 (尽管由于 SDE-solver 包含随机性, 生成的样本通常具有更好的多样性). 因此, 目前大部分的采样加速工作都基于 ODE-solver. 对于一个 SDE, 可以找到一个 ODE, 使其与 SDE 具有相同的边际概率分布, 这类 ODE 被称为概率流 ODE, 它定义了采样过程中从噪声到最终生成样本的轨迹.

DPM-solver[41] 的研究者发现, 扩散模型的概率流 ODE 具有半线性结构, 并基于此给出了扩散常微分方程 (diffusion ODE) 的精确解. 通过对该精确解进行不同阶数的近似, 可以得到不同阶数的

表 1 文献 [12] 中对比扩散模型和对抗生成网络的部分结果. 扩散模型选取的是 ADM 模型, 在 LSUN 数据集中对抗神经网络选取的是 StyleGAN, ImageNet 中则是 BigGAN. FID 评分是衡量生成样本质量的指标, 越低代表样本多样性越高

数据集	LSUN Bedrooms 256×256	LSUN Horses 256×256	LSUN Cats 256×256	ImageNet 64×64
GANs FID	1.90	2.57	5.57	2.07
DPMs FID	2.35	3.84	7.25	4.06

ODE-solver. 研究还表明, DDIM 与 ODE-solver 一文提出的一阶 ODE-solver 等价. 高阶 ODE-solver 可以在 10 至 20 步内生成高质量的样本. 然而, 包含 DPM-solver 在内的采样方法依赖于特定的参数设置, 这些参数未必总是最优的. 为了解决这一问题, DPM-solver 的作者进一步提出了 DPM-solver-v3 [82], 该方法基于预训练模型计算出了最优参数值, 进一步将推理步数缩减到 5 至 10 步.

2024 年, 一种名为 AMED [83] 的更快速采样方法被提出, 它能够在约 5 步内生成高质量的样本. 通过大量实验, Zhou 等 [83] 发现一个令人惊讶的事实: 采样轨迹几乎位于一个二维子空间内. 基于这一发现, 并结合积分中值定理, 他们提出了甚至可以在一步内生成样本的 ODE-solver. 尽管如此, 实验表明, 分为约 5 步的采样过程才能生成高质量的样本.

值得一提的是, 除了 ODE-solver 和 SDE-solver 之外, 近年来也有一些简洁有效的加速方法被提出. 例如, Ma 等 [44] 提出的 Skip-Tuning 方法, 通过简单调整 Unet 网络中 Skip-Connection 的权重即可实现进一步加速. 与此同时, 该方法还可以与 ODE-solver 或 SDE-solver 结合使用, 从而进一步提升采样速度.

4.2 架构改进

为了应对扩散模型在训练速度慢、采样效率低和可扩展性差等方面的挑战, 一些新颖的扩散模型架构被提出, 以改进这些问题. 针对训练时间较长的问题, Rombach 等 [55] 提出了隐扩散模型 (latent diffusion model, LDM). LDM 的核心思路是, 首先通过编码器将输入数据映射到一个低维的潜在空间中, 然后在潜在空间中执行扩散过程, 生成潜在特征图. 最后, 这些特征图通过解码器转换回原始数据空间. 由于潜在空间的维数远小于原始数据空间, 模型在训练和推理时的计算成本大幅降低. 实验表明, LDM 不仅能够显著加快训练过程, 还能生成与经典扩散模型相媲美的高质量图像. 这一方法增强了扩散模型在图像生成、图像修复和图像翻译等任务中的表现.

2023 年, Song 等 [61] 提出了一致性模型, 进一步提升了扩散模型的采样效率. 在常规扩散模型的基础上, 他们设计了一种一致性函数, 该函数依赖于样本和时间, 并且满足轨迹上所有点都具有相同值的约束条件. 在此约束下, 模型可以在采样过程中实现从任意点跳跃到另一个点, 从而支持一步生成.

除了 LDM 和一致性模型, Diffusion Transformer [52] 在视频生成模型 Sora 的推动下也得到了广泛关注. 该模型将传统的 Unet 网络替换为 Transformer 架构, 结合了扩散模型的生成优势与 Transformer 模型的可扩展性优势, 使得扩散模型在一定范围内也符合规模化法则 (scaling law). 这种架构改进为扩散模型在大规模生成任务中的应用开辟了新的可能性.

4.3 使用范围的扩大

在实际应用中, 扩散模型往往需要应对一些特殊情况, 常见的有两类: 一是根据某种条件生成符合特定要求的样本, 即条件扩散模型; 二是处理非 Euclid 空间的数据, 如流形或图结构数据, 这需要专门的模型进行适配.

条件扩散模型可以分为两类: 基于分类器引导的和基于免分类器引导的. 基于分类器引导的条件扩散模型是在无条件扩散模型基础上, 通过额外训练一个分类器引入条件信息, 从而生成符合特定条件的样本. 这种方法成本较低, 迁移性强. 免分类器引导的模型则在训练时直接将条件融入模型中, 从零开始训练, 虽然成本较高, 但通常效果更好. 以 GLIDE [50]、Imagen [57] 和 ControlNet [79] 为代表的条件扩散模型, 通过引入额外的输入条件 (如文本、边缘图、深度图或语义标签), 使得生成的样本更

加符合用户的需求. 这种方法极大增强了扩散模型的灵活性与定制性, 在图像生成、修复和转换等任务中展现了出色的性能.

针对非 Euclid 空间的数据, 研究者提出了一系列扩散模型, 专门用于处理离散数据、流形数据和图结构数据等复杂场景. 对于离散数据, 如分类标签或文本序列, 传统的扩散模型由于微小扰动难以迁移到离散空间, 表现受限. 为此, 出现了专门用于离散数据的扩散模型, 如 D3PM (discrete denoising diffusion probabilistic model) [5] 系列. 这些模型通过在离散状态空间中引入噪声扰动和扩散过程, 成功地将无条件扩散转化为适合离散数据的生成任务, 如文本生成和分类标签建模. 类似地, VQ-Diffusion [19] 及其改进版 VQ-Diffusion+ [69] 通过引入向量量化机制 (vector quantization), 在处理离散数据的复杂结构时表现优异. VQ-Diffusion 通过将输入数据编码为离散的向量索引, 并在此索引空间中进行扩散, 从而在文本生成和图像编码等任务中取得了显著成果.

在处理流形数据 (如 3D 点云和姿态数据) 方面, 研究者们开发了特定的扩散模型, 如 RGSM (Riemannian generative stochastic model) [11] 和 PNDM (projected nonlinear diffusion model) [39]. 这些模型特别设计了适用于嵌入流形空间的数据扩散过程, 充分考虑了流形数据的几何性质, 避免了传统扩散模型在非 Euclid 空间中的不一致性问题. RGSM 能够生成具有流形结构的高维数据, 如 3D 点云和动作姿态, 而 PNDM 通过投影技术将扩散过程限制在特定的流形几何结构上, 生成符合几何约束的数据. Boomerang 模型则进一步结合了流形几何和时间反向扩散的特性, 通过精准的时间步长控制, 实现了高效的流形数据生成.

在图结构数据的生成任务中, 扩散模型也取得了显著进展. 图结构数据相比 Euclid 空间数据, 有更复杂的节点和边之间的关系, 因此更难建模. EDP-GNN (edge denoising process graph neural network) [51] 结合了图神经网络与扩散模型的优点, 能够有效去噪并生成图结构数据. GDSS (graph diffusion-based stochastic sampling) [28] 模型则扩展了扩散模型的框架, 能够联合建模图中的节点和边, 使其在分子结构生成和社交网络建模等任务中表现优异. 与传统图生成方法不同, 这类扩散模型通过逐步去噪的过程生成更加稳定的图结构, 且能够有效控制图的特性, 如节点度分布和连通性等.

通过引入这些处理非 Euclid 数据的扩散模型, 研究者们大大扩展了扩散模型在不同领域的应用场景. 这些模型不仅能够处理传统的图像生成任务, 还能够在文本生成、图结构建模、三维点云和姿态生成等任务中展现强大的能力. 随着研究的深入, 这些模型有望在药物发现、3D 建模、自然语言处理等实际应用中发挥更大的作用.

5 扩散模型在转录组中的应用

扩散模型以其强大的生成能力和上下文理解能力, 在生物信息领域表现出广泛的应用前景. 其逐步加噪与去噪的机制, 使模型能够有效学习数据中的噪声形式, 实现高效去噪, 在单细胞转录组测序数据的缺失值填补、空间转录组表达数据的修复和去噪以及图像分辨率提升等生物信息领域的相关任务中取得了显著效果. 此外, 扩散模型还支持自监督学习, 减少了对标签数据的依赖, 进一步增强了其在复杂生物学数据处理中的应用潜力. 表 2 列举了几个具有代表性的扩散模型在转录组的应用. 在下文中, 我们将从单细胞转录组数据去噪、空间转录组表达数据填补以及图像分辨率提升等多个角度, 详细阐述这些扩散模型在转录组中的实际应用以及给该领域带来的突破.

5.1 单细胞 RNA 测序数据

随着单细胞转录组测序技术的快速发展, 我们能够全面描绘单个细胞的基因表达图谱, 从而对生物

表 2　第 5 节提及的模型及其应用场景

模型	使用数据	功能	年份
scIDPMs	单细胞 RNA 测序数据	预测缺失值	2024
scDiffusion	单细胞 RNA 测序数据	提升数据集的密度	2024
scVAEDer	单细胞 RNA 测序数据	学习数据的低维表征	2023
RegDiffusion	单细胞 RNA 测序数据	推断基因调控网络	2023
SpaDiT	空间转录组数据	预测缺失值, 保持细胞布局结构	2024
SpatialScope	空间转录组数据	提升数据分辨率	2023
Diff-ST	空间转录组数据	提升转录组图像分辨率	2024

体的发育和功能有了更加系统和深入的理解 (参见文献 [17,29]). 然而, 由于生物学因素和测序技术的局限性, 所测得的基因表达数据往往伴有噪声, 尤其是经常出现缺失值 (dropout) 现象 (参见文献 [27]). 近年来, 各种算法被用于数据去噪、数据插补、数据生成和批次效应校正等问题, 主要分为基于统计模型的方法和基于深度学习的方法. 基于统计模型的方法利用先验知识, 如细胞 - 细胞相互作用、基因 - 基因相互作用或两者的结合, 构建统计模型以估计缺失值, 并恢复在单细胞转录组数据中的基因表达模式. 基于深度学习的方法通过各种生成模型, 如自编码器 (autoencoder, AE) [48]、VAE [32]、GAN [18] 和 DDPM [21], 学习隐空间中的基因表达数据的分布. 其中, 去噪扩散概率模型由于其强大的生成能力在单细胞数据分析中得到了广泛的应用.

scIDPMs [81] 基于条件概率扩散模型对单细胞 RNA 测序数据进行填补, 它首先利用整体基因表达数据识别缺失值位点, 然后利用上下文的表达数据对缺失值进行推断. 具体而言, scIDPMs 先使用基于生物学的方法来识别缺失值位点, 认为同一标签的细胞有共同的非零表达基因, 再使用自监督学习的方式训练一个 DDPM 模型, 目标是学习缺失值在未缺失值作为条件下的分布. DDPM 的训练过程中需要已知初始值, 我们无法直接利用 DDPM 模型来还原缺失值. scIDPMs 采用自监督的方法克服这一困难, 将未缺失的基因表达值随机分为两部分, 一部分作为已知表达值, 另一部分作为需要被重构的目标. 基于此训练一个扩散模型, 给重构目标添加随机 Gauss 噪声, 将已知表达值及它们在基因表达矩阵中的坐标位置、细胞类型等作为已知信息加入到反向扩散学习所加噪声的网络中, 以还原出重构目标为目的进行训练. scIDPMs 反向学习噪声的网络选用带注意力机制的多层残差网络, 最终可以将完整的表达数据输入给训练好的 scIDPMs 模型, 使其预测所有缺失位点处的基因表达值.

除了应用于单细胞 RNA 测序数据的填补外, DDPM 还可用于学习单细胞数据的低维嵌入、生成高质量的单细胞数据以提高数据集密度来辅助其他模型训练等任务. scVAEDer [56] 通过结合 VAE 和 DDPM 来学习单细胞数据的低维表征, 并且能够生成新的单细胞 RNA 测序数据, 预测不同细胞类型在干扰下的反应, 识别去分化过程中的基因表达变化, 并检测生物过程中的主调控因子. scDiffusion [43] 结合扩散模型和基础模型的生成模型, 在受控条件下生成高质量的 scRNA-seq 数据. 通过将下游计算任务统一为后验估计问题, scDiff [68] 提出了一个基于 DDPM 的通用单细胞分析框架, 可以实现细胞标记、数据插补和知识迁移等多个下游任务.

5.2　空间转录组

空间转录组测序技术是一种新兴的生物技术, 旨在在组织切片的特定空间位置上测量和分析基因的表达情况. 与传统的单细胞或常规转录组技术不同, 空间转录组学不仅可以获取基因表达数据, 还能保留细胞在组织中的空间位置信息. 目前空间转录组学的数据分析主要有三个挑战: 一是空间转录

组表达数据的去噪、缺失值填补等整合工作; 二是提升空间转录组图像的分辨率; 三是空间异质性的解析, 如识别不同空间域中的细胞类型等. 在处理第一个挑战和第二个挑战时扩散模型有着卓越的表现, 并在 DDPM 模型的基础上已经诞生不少优秀的方法.

5.2.1 空间转录组数据的填补

SpaDiT [36] 是一种基于条件扩散的生成模型, 它的任务是利用单细胞 RNA 测序数据作为先验信息来增强空间转录组学数据, 目标是准确预测缺失或未知基因的表达值. SpaDiT 的输入数据有两种: 来自空间转录组学的基因表达矩阵和来自单细胞 RNA 测序技术的另一个基因表达矩阵. 利用条件扩散模型, SpaDiT 将单细胞 RNA 测序数据作为条件因子指导模型完成扩散过程, 最终为空间转录组生成目标基因表达谱. SpaDiT 架构包括三个模块: 用于预处理空间转录组数据的低维嵌入模块、用于处理单细胞 RNA 测序数据的条件嵌入模块及用于去噪的核心网络. 核心网络以 DDPM 为主心骨, 在正向扩散中往转录组的低维嵌入上添加标准 Gauss 噪声, 反向扩散中采用带自注意力机制的神经网络学习正向扩散中加入的噪声. 由于单细胞 RNA 测序数据的辅助, SpaDiT 可以为单细胞 RNA 数据中存在表达但空间转录组中却未测得表达的基因填补缺失的表达值. 经过数十个空间转录组学和单细胞 RNA 测序数据集上的测试, SpaDiT 不仅保持了细胞布局固有的复杂拓扑结构, 还实现预测的基因表达与真实数据准确对齐, 证明了其在重现转录组空间表达模式方面的稳健性.

SpatialScope [73] 利用已有单细胞 RNA 测序参考的数据集, 增强基于序列技术的空间转录组数据以达到单细胞分辨率, 并推断基于图像技术的空间转录组数据表达水平. SpatialScope 包括三个步骤. 第一步是细胞核分割, 用已有的 StarDist [58] 和 Cellpose [66] 方法定位并计算一个测量格点中细胞的个数. 第二步是细胞类型识别, 结合单细胞 RNA 参考数据集中不同细胞类型的基因表达, 构建概率模型给第一步中定位的细胞贴上类型标签. 作为核心的第三步是基因表达分解, 使用基于得分的生成模型学习单细胞 RNA 参考数据集中不同细胞类型的基因表达分布, 再对目标空间转录组数据中每个测量格点里相应类型标签的细胞预测其表达. SpatialScope 核心的第三步中基于得分的生成模型与 NCSN 本质上是同一思想方法, 而 NCSN 也是扩散模型的一种延伸, 因此 SpatialScope 方法也彰显了扩散模型为空间转录组精确填补表达值的强大能力.

5.2.2 空间转录组图像的分辨率提升

传统为空间转录组图像作超分辨率提升的方法主要有两个局限. (1) 传统方法将多模态数据视为多种扩散条件, 通过简单的拼接进行整合. 这种方法平等对待每种模态, 忽略跨模态之间的联系, 这可能无法有效利用组织学图像和基因表达中的互补信息. (2) 传统模型主要用于生成单个基因的空间表达图像, 即分别处理每个基因的空间转录组图像, 然而跨多个基因的内在表达关联可能产生不一致的整体转录组景观, 对相关特征的有效学习构成挑战.

Diff-ST [74] 是一种跨模态的条件扩散生成模型, 它的任务是利用低分辨率的空间转录组学图像和组织学图像提高空间转录组学图像的分辨率, 目标是将输入的高分辨率空间转录组图像增强为超分辨率图像. Diff-ST 方法主要包括三个部分: 第一部分是低分辨的空间转录组图像和组织学图像的跨模态整合, 通过基于课程学习 (curriculum learning) 的交叉注意力模型实现, 该策略能够从组织图像中提取分层的细胞到组织级信息, 以克服传统方法的第一个困难; 第二部分是提出一种基于共表达强度的基因相关图网络来模拟多个基因的共表达关系, 以克服传统方法的第二个困难; 完成前两部分后, 第三部分先将高分辨率的空间转录组图像进行正向扩散, 再将前两部分得到的信息加入到注意力网络中, 学

习正向扩散中加入的噪声值, 完成反向扩散. 在足够的训练后, 反向扩散最终得到的便是超高分辨率的空间转录组图像. 基于 Diff-ST 方法得到的超分辨率空间转录组学图像在三个公共数据集上进行了大量实验, 结果表明 Diff-ST 方法得到的超分辨率图像在多项实验中表现优异, 能够良好提取组织级特征并进行细胞级的图像修补.

DDPM 还可以作为空间转录组学 3D 视图构建、多模态数据整合等应用的模型基础. stDiff [35] 采用条件扩散模型, 通过两个 Markov 过程捕获单细胞 RNA 测序数据中的基因表达丰度关系, 基于此实现对空间转录组数据的增强. SpatialDiffusion [31] 基于 DDPM 模型生成多个组织的空间转录组切片, 通过在已有的有限近邻切片数据之间插值生成新的转录组切片, 可以在提高数据集密度的同时为生物体组织进行 3D 重构奠定基础. ACGDC [45] 在整合单细胞多组学数据后使用条件扩散模型生成新样本, 在真实数据集中验证发现其可以显著改良星形胶质细胞亚型的分类与识别.

5.3 基因调控网络

基因调控网络对于阐明细胞机制和推进治疗干预至关重要. 从大量表达数据推断基因调控网络的原始方法常常会面临两大问题: 高维数据中的维数灾祸和数据本身的固有噪声. RegDiffusion [84] 是一个基于去噪扩散概率模型推断基因调控网络的方法, 其重点关注基因表达数据转化而成的特征之间的调节效应. RegDiffusion 先使用编码和多层感知器将表达数据投影到低维空间中, 解决维数灾祸的问题, 再在低维空间上修复数据的固有噪声. 在正向扩散过程中, RegDiffusion 反复将 Gauss 噪声加入低维编码, 在反向扩散过程中再使用具有参数化邻接矩阵的神经网络预测先前添加的噪声. 在基准实验中, RegDiffusion 于多个数据集中表现出比其他方法更优异的性能, RegDiffusion 可以在不到 5 分钟的时间内从包含超过 15,000 个基因的真实单细胞数据集中推断出具有生物学意义的基因调控网络, 反映出 DDPM 在应用于推断基因调控网络方面也有不俗表现.

6 扩散模型在蛋白质中的应用

扩散模型在图像生成任务中有着出色的表现 [10], 凭借其在处理高维离散数据和生成任务中的卓越表现, 迅速被应用于生物信息学领域 [20]. 该模型在无条件蛋白质设计、结合剂设计、对称结构构建、酶活性位点优化等任务中展现出出色性能, 能够生成符合生物物理规律的创新型蛋白质. 通过扩散模型, 还可以根据简单的分子输入精准设计出复杂的蛋白质结构, 并且在实验验证中表现出高度的准确性和可操作性, 这显著加速了蛋白质工程以及新型生物分子的开发进程 [1,33]. 在下文中, 我们将从蛋白质设计、复合物结构预测、多肽药物设计等多个角度, 详细综述扩散模型在蛋白质领域中的实际应用及其为该领域带来的突破性进展. 审视扩散模型如何在蛋白质设计中发挥作用, 包括其在生成具有特定功能或性质的蛋白质方面的能力. 我们将探讨扩散模型如何通过高效的结构生成和优化算法, 推动新型蛋白质的设计, 满足各种生物技术需求. 分析扩散模型在蛋白质 - 蛋白质和蛋白质 - 小分子复合物结构预测中的应用, 探讨扩散模型如何提高复合物结构预测的准确性, 优化蛋白质与其他分子相互作用的模拟, 从而促进药物开发和生物医学研究. 讨论扩散模型在多肽药物设计中的应用, 包括如何利用其生成潜在的药物候选分子, 并预测其与靶标的结合能力. 表 3 给出本节提及的模型.

6.1 蛋白质设计

蛋白质通过与其他生物分子相互作用, 参与和调控多种生物过程. 随着长期对蛋白质生物化学和

表 3 第 6 节提及的模型及其应用场景

模型	使用数据	功能	年份
Rfdiffusion	蛋白质序列和结构	蛋白质设计	2024
EvoDiff	蛋白质序列	蛋白质设计	2023
FoldingDiff	蛋白质结构	蛋白质设计	2024
Chroma	蛋白质序列和结构	蛋白质复合物结构预测	2023
RFAA	蛋白质序列和结构	蛋白质复合物结构预测	2024
Alphafold3	蛋白质序列和结构	蛋白质复合物结构预测	2024
AMP-diffusion	抗菌肽序列	多肽设计	2024
DiffLinker	小分子片段	小分子设计	2024

生物物理性质的知识积累, 蛋白质设计在技术上是可以实现的 (参见文献 [23]). 蛋白质设计的目标是通过修饰天然蛋白质或者从头设计新的蛋白质, 创造出具有特定性质和功能的生物分子 [80]. 由于蛋白质具有功能多样性、纳米级大小和可生物降解等特性, 大量人类设计的蛋白质已经被广泛应用于生物学、医学、农业和制造业等领域 (参见文献 [16]).

早期的蛋白质设计如定向进化以及后续的理性工程等, 主要关注的是模仿或加速自然进化过程. 通过多轮突变文库构建和高通量筛选, 这些方法能够偶然获得性能改善甚至具有新功能的蛋白质. 然而, 这些方法始终面临实验精度与筛选通量之间的权衡, 更重要的是, 它们的探索仍然局限于最初的天然蛋白质周围. 随着计算设备和算法的发展, 前述的不足逐渐被计算机辅助蛋白质工程克服, 旨在生成自然界中不存在的新蛋白质的从头设计受到了广泛的关注. 凭借众多有价值的成果, 蛋白质从头设计在 2016 年被《科学》杂志评为年度十大突破之一. 蛋白质从头设计是指设计一个物理上合理的蛋白质骨架及其对应序列, 而不依赖现有的天然蛋白质作为基础 (参见文献 [23]). 近年来, 随着深度学习算法, 尤其是扩散模型的发展, 蛋白质领域发生了革命性的变化. 深度学习算法不仅显著提升了蛋白质结构的精准预测能力 [1,6], 还推动了具有特定性质和功能的蛋白质设计, 例如设计折叠成特定拓扑结构的蛋白质 [34]、设计与特定靶标进行结合的蛋白质 [7] 等. 值得注意的是, 扩散模型具有输出高度多样性, 可以基于特定目标进行逐步迭代生成, 并且能够在全局框架下对 3D 结构进行建模, 因此在蛋白设计领域得到了广泛的应用 (参见文献 [76]).

蛋白质设计算法包括基于结构的蛋白设计和基于序列的结构设计 (参见文献 [13]). 基于结构的蛋白设计首先设计具有特定拓扑结构的骨架, 然后需要对一些可能的序列进行优化, 使其能更好地适应这个结构并发挥功能. 基于序列的蛋白设计在隐空间学习序列表征分布, 并根据从该分布推导出的推测性表示, 在实际空间中生成新的蛋白质序列. RFdiffusion [76] 通过微调 RoseTTAFold [6] 结构预测网络用于蛋白质结构去噪任务, 能够得到一个蛋白质骨架生成模型. RFdiffusion 可以实现无条件和拓扑约束的蛋白单体设计、蛋白质结合物设计、对称寡聚体设计、酶活性位点支架设计等多个蛋白设计任务. 基于扩散模型的算法 SMCDiff [70] 和 FADiff [38] 用于解决基序 - 支架问题, 即设计一个给定蛋白质基序的支架结构. SMCDiff 首先利用 ProtDiff 来预测蛋白质骨架, 然后以基序为指导生成带有基序的蛋白质. FADiff 首次解决了相对位置未知的多基序问题, 它确保基序存在并且通过促进基序的刚性运动自主设计基序位置. 受蛋白质自然折叠过程启发, FoldingDiff [77] 利用扩散模型生成稳定折叠的蛋白质骨架结构. 它将蛋白质骨架结构描述为一个角度序列, 捕捉组成骨架原子的相对方向, 通过从随机非折叠状态去噪, 逐步生成稳定的折叠结构.

蛋白质设计的主要任务是找到能够稳定地显示出期望性质并且执行预期功能的序列. 原则上, 直

接映射蛋白质序列和功能的空间似乎比需要预先确定拓扑结构的设计具有一定的优势. 更重要的是, 测序技术的进步带来了比结构数据丰富得多的大量的蛋白质序列数据. 这些蛋白质序列的数据结合深度学习算法强大的特征提取、模式识别和目标生成能力, 可以直接探索序列空间和改进蛋白质设计范式. EvoDiff[2] 利用进化数据集和扩散模型, 仅使用序列信息进行可控的蛋白质设计. 在 EvoDiff 中, 离散扩散框架通过前向过程逐步扰乱蛋白质序列的氨基酸组成, 而神经网络参数化的逆向过程预测每一步的变化, 从而可以从随机噪声生成新的蛋白质序列. 与基于结构的蛋白设计方法相比, EvoDiff 还能生成基于结构模型无法生成的蛋白质, 如具有无序区域的蛋白质, 同时保持设计功能性结构基序的能力. TaxDiff[37] 是一种用于可控蛋白质序列生成的分类学引导扩散模型, 它结合生物物种信息和扩散模型的生成能力在序列空间生成结构稳定的蛋白质.

扩散模型通过对随机噪声的去噪, 能够生成多样化的蛋白质结构和序列, 创造出与自然蛋白质不同但满足设计目标的新的蛋白质. 扩散模型还允许在设计中加入控制变量, 从而在特定的需求下生成特定的性质和功能的蛋白质. 扩散模型还能够处理蛋白质序列与结构之间的复杂关系, 通过考虑蛋白质骨架的生物化学和生物物理特性, 从而高效地生成结构上折叠稳定、功能上满足需求的蛋白质. 总之, 扩散模型在蛋白设计领域有诸多优势, 具有广泛的应用前景.

6.2 复合物结构预测

蛋白质复合物结构预测是计算生物学和结构生物学中的一项核心任务, 旨在确定多种蛋白质如何在细胞内相互作用并形成功能性复合物 (参见文献 [14]). 蛋白质复合物在细胞的各种生理过程中发挥关键作用, 包括信号转导、代谢调控、基因表达调节和细胞结构维持等. 因此, 准确预测蛋白质复合物的三维结构对于深入理解生物过程和疾病机制具有重要意义 (参见文献 [34]).

尽管 AlphaFold2 (AF2) 和 RoseTTAFold (RF) 通过高精度的蛋白质结构建模革新了结构生物学 (参见文献 [6,30]), 但它们无法模拟共价修饰或蛋白质与小分子及其他非蛋白质分子的相互作用, 而这些共价修饰和相互作用在生物功能中起着至关重要的作用 (参见文献 [4]). 此时确定一个复合体的三维结构仍然是一个挑战, 亟需研究者开发能高精度预测复合物结构的方法, 从而在三维结构预测上取得新突破.

扩散模型作为一种通过对噪声数据的去噪学习的方法被引入了蛋白质研究领域, 被用于预测蛋白质复合物的结构. 作为一种创新的蛋白质研究方法, 结构生物学研究者也将其应用于蛋白质复合物预测领域. 在后 AlphaFold 时代, 扩散模型的加入极大程度地推动了复合物结构预测, 并取得了预测精度的突破.

6.2.1 蛋白质 - 蛋白质复合物结构预测

蛋白质 - 蛋白质复合物在细胞内执行多种重要功能, 如信号转导、基因表达调控和代谢过程 (参见文献 [46]). 同时许多药物的作用机制是通过与特定的蛋白质 - 蛋白质复合物相互作用来实现的 (参见文献 [65]). 准确预测复合物结构有助于识别潜在的药物靶点, 优化药物分子与靶点的结合能力, 从而加速药物开发过程, 提高药物研发的效率和成功率. 预测这些复合物的结构, 可以揭示细胞内的生物学机制和疾病的分子基础. 扩散模型在预测蛋白质 - 蛋白质复合物结构中有着显著优势: 擅长处理高维数据, 能够有效地捕捉蛋白质 - 蛋白质相互作用中的复杂空间和结构特征.

Ingraham 等[26] 2023 年发表在 *Nature* 上的 Chroma 是一种蛋白质 - 蛋白质复合物的生成模型. 它引入了一种可以维持聚合物整体构象统计的扩散过程、一种有效的分子系统神经结构以及一种用

于扩散模型的通用低温采样算法. 该方法对 310 种蛋白质进行了实验验证, 结果表明 Chroma 采样得到的蛋白质高度表达、折叠并具有良好的生物物理特性. 除了对蛋白以蛋白序列为输入, 还有方法如 DIFFMASIF[67] 提出了使用基于蛋白质分子表面的编码器 - 解码器架构来有效地学习物理互补性. 因此, 它在结构新颖的界面和低序列守恒上有着良好的性能. 在对主链构象取得较好预测突破后, 研究者希望进一步探索扩散模型在蛋白侧链中的应用. SidechainDiff[40] 利用了未标记的实验蛋白质结构结合 Riemann 扩散模型来学习侧链构象的生成过程, 同时, SidechainDiff 是第一个基于扩散的侧链生成模型.

2024 年, 为了更好地模拟蛋白质与其他生物分子的相互作用, 开发了 RoseTTAFold (RF) 的 David Baker 组在 RF 的基础上作了进一步的扩展, 通过对扩散去噪任务进行微调, 开发了 RoseTTAFold All-Atom (RFAA) 全原子模型[33]. 它可以模拟包含蛋白质、核酸、小分子、金属以及共价修饰的完整生物组件, 并基于聚合物的序列和小分子的原子键合几何形状及其共价修饰进行预测. 该模型利用 RFAA 设计并通过实验验证了与治疗心脏病的地高辛、酶辅因子血红素以及光学活性胆磷脂分子结合的蛋白质, 将该方法应用于复杂生物分子系统的建模和设计. AF2 团队也在同时期推导出了 Alphafold3[1], 能预测包含更广泛的生物分子, 包括配体、离子、核酸和修饰残基的复合物的结构. 它基于 AF2 的网络架构作出了重要变革, 用扩散模型模块取代了作用于氨基酸特定框架和侧链扭转角的结构模块, 从而实现直接预测原子坐标.

6.2.2 蛋白质小分子复合物结构预测

预测小分子配体与蛋白质的结合结构对于药物发现与优化、理解生物机制有着至关重要的作用 (参见文献 [15, 49]). 传统的对接方法可以很好地完成复杂状态下已知蛋白质结合袋构象的再对接任务 (参见文献 [71]). 然而, 在不知道新配体的蛋白质结合构象的现实对接场景中, 由于灵活对接的计算成本高且不准确, 准确建模结合复杂结构仍然具有挑战性. 将对接作为回归问题的深度学习方法减少了运行时间, 但却在准确度上仍然没有明显提高 (参见文献 [47]). 随着扩散模型的兴起, 蛋白质小分子结构预测的准确度得到了新的推动力. 扩散模型通过逐步生成和优化结构, 能够在复杂的结合场景中提供更高的准确性.

DiffDock[9] 作为一个基于配体姿态的非 Euclid 流形的扩散生成模型, 在 PDB-Bind 上获得 38% 的 top-1 成功率 (均方根偏差 RMSD < 2A), 显著优于之前的传统对接方法. GeoDiff[78] 将每个原子视为一个粒子, 并学习直接逆转扩散过程 (即从噪声分布转变为稳定构象) 作为 Markov 链. NeuralPLexer[53] 则是使用了原子级分辨率对结合复合物的三维结构及其构象变化进行采样, 仅使用蛋白质序列和配体分子图输入直接预测蛋白质 - 配体复合物结构. 该模型基于扩散过程, 该过程结合了基本的生物物理约束和多尺度几何深度学习系统, 以分层方式迭代采样残余级接触图和所有重原子坐标. 与所有现有方法相比, NeuralPLexer 在蛋白质 - 配体盲对接和灵活结合位点结构恢复方面都达到了优于 Alphafold2 的最先进的性能. NeuralPLexer 的预测与酶工程和药物发现中重要靶点的结构确定实验一致, 表明其在蛋白质组级加速功能蛋白和小分子设计的潜力. DynamicBind[42] 使用等变几何扩散网络来构建平滑的能量景观, 促进不同平衡状态之间的有效转换. 同时, DynamicBind 可以在看不见的蛋白质靶标中识别隐藏的口袋. DiffBindFR[85] 是一种基于全原子扩散的柔性对接模型, 可以在配体整体运动和柔性以及口袋侧链扭转变化的产物空间上运行. 在 Apo 和 AlphaFold2 模型结构中, DiffBindFR 在准确的配体结合姿态和蛋白质结合构象预测方面表现出优越的优势, 适用于基于 Apo 和 AlphaFold2 结构的药物设计.

综上所述, 通过对蛋白质结构的精确模拟, 扩散模型能够处理复杂的多蛋白质复合物, 并对其进行精细建模. 同时, 扩散模型有潜力可以识别和优化潜在的结合位点. 通过预测小分子与蛋白质的结合模式, 扩散模型可以加速药物筛选过程. 这使得模型在预测和优化多蛋白质复合物的结构和功能方面表现出色, 对于研究复杂的生物过程和开发新的生物技术工具具有重要意义.

6.3 小分子和多肽药物设计

小分子药物设计在现代医学和药物研发中扮演着关键角色 (参见文献 [15]). 小分子药物设计是药物研究和开发领域的关键组成部分. 小分子的设计和开发通常以其低分子量和结构简单为特征, 对广泛疾病的治疗干预具有深远的意义. 扩散模型在该领域中除了直接预测蛋白质复合物结构, 还可以进行有蛋白质约束的分子生成. 该类方法进一步提升了基于口袋的分子生成的有效性.

IPDiff [24] 的主要目标是用于蛋白质特异性的 3D 分子生成. 该方法结合亲和信号, 预训练一个蛋白质 - 配体相互作用先验网络 (IPNet). 随后利用预训练的先验网络将目标蛋白与分子配体之间的相互作用整合到适应分子扩散轨迹的正向过程中, 以及增强结合感知的分子采样过程. 2024 年发表在 *Nature* 上的 DiffLinker [25] 是一种基于片段黏合 (binder) 设计的方法, 利用了 E(3)- 等变三维条件扩散模型, 用于黏合设计. 该模型不像以前的方法只能连接成对的分子片段, 该方法可以连接任意数量的片段. PMDM 用于 3D 分子生成拟合指定目标. PMDM 由具有局部和全局分子动力学的条件等变扩散模型组成, 使 PMDM 能够考虑条件蛋白质信息以有效地生成分子. 模型应用于实际药物设计场景, 对 SARS-CoV-2 主蛋白酶 (Mpro) 和细胞周期蛋白依赖性激酶 2 (CDK2) 进行了先导化合物优化, 更进一步地说明了扩散模型对药物设计的实用性 (参见文献 [22]).

治疗性肽设计是生物技术和人工智能的融合药物开发的新领域. 肽是一种短链氨基酸, 具有良好的生物相容性和高选择性, 可用于开发新型的药物和治疗性分子. 与小分子药物相比, 具有靶向治疗和最小副作用等优点. 多肽药物在治疗癌症、糖尿病、心血管疾病等方面展现了广泛的应用潜力.

AMP-diffusion [8] 是一种为抗菌肽 (antimicrobial peptide, AMP) 设计量身定制的潜在空间扩散模型, 利用最先进的 pLM ESM-2 的能力, 重新生成功能性 AMP, 用于下游实验应用. 该方法与众多方法不同, 它并不基于结构、图或者离散序列, 而是基于蛋白质语言模型 (pLMs). 同样类似的方法还有 ProT-Diff [75], 该方法将预训练的蛋白质语言模型与扩散模型结合起来, 能够在数小时内快速生成具有 32 种不同长度的数千个 AMP. HYDRA [54] 结合了扩散模型的分布建模能力与结合亲和力最大化算法, 使其能够从头设计针对各种靶受体的肽结合物. 在实际应用中, 该方法设计了针对恶性疟原虫红细胞膜蛋白 1 (PfEMP1) 的治疗性肽, 这些肽针对 PfEMP1 基因表达的蛋白质进行设计.

综上所述, 扩散模型能够在蛋白结构或蛋白口袋的约束下生成新的小分子或多肽结构, 这些结构可能具有潜在的生物活性和药物价值. 扩散模型能够高效生成和筛选大量药物候选分子, 显著加快药物设计和开发的进程. 它的生成能力和优化算法减少了实验筛选的工作量, 缩短了药物研发的周期.

7 结论

扩散模型于 2015 年首次被提出, 并在 2020 年因 DDPM 模型生成的高质量样本而受到广泛关注. 扩散模型以随机过程理论为基础, 具有强大的生成能力, 在计算机视觉、自然语言处理和生物信息学等领域展现出广阔的应用前景. 本文重点关注扩散模型的概率学阐释, 以及三种经典的扩散模型 NCSN、DDPM 和 DDIM 的概率学原理和算法实现过程. 值得注意的是, 传统的扩散模型存在训练时

间长、计算资源占用多等问题. 为了解决这些问题, 研究人员提出了许多改进模型, 这在一定程度上也促进了扩散模型的蓬勃发展和广泛应用. 目前, 改进的模型大致可以分为三类: 针对扩散模型推理时间长的加速方法、修改模型架构以提高其性能的方法以及扩展模型使用范围的方法.

扩散模型在生物信息学应用广泛, 本文重点关注扩散模型在转录组和蛋白质相关任务上的应用. 一方面, 扩散模型凭借其生成能力, 能有效学习转录组数据中噪声形式, 实现高效去噪, 在单细胞转录组测序数据的缺失值填补、空间转录组数据去噪以及空间转录组图像分辨率提升等问题中取得了显著的成绩. 另一方面, 扩散模型可以生成符合生物物理规律的蛋白质, 广泛应用于蛋白质设计方法. 与此同时, 扩散模型可以提高蛋白质复合物结构预测的准确性, 优化蛋白质与其他分子的相互作用, 从而促进结构预测和药物设计等方面的发展, 在蛋白质相关任务中有着广阔的应用前景.

参考文献

1 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493–500

2 Alamdari S, Thakkar N, van den Berg R, et al. Protein generation with evolutionary diffusion: Sequence is all you need. bioRxiv:2023.09.11.556673, 2023

3 Anderson B D O. Reverse-time diffusion equation models. Stochastic Process Appl, 1982, 12: 313–326

4 Arkin M R, Wells J A. Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov, 2004, 3: 301–317

5 Austin J, Johnson D D, Ho J, et al. Structured denoising diffusion models in discrete state-spaces. arXiv:2107.03006, 2023

6 Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373: 871–876

7 Cao L, Coventry B, Goreshnik I, et al. Design of protein-binding proteins from the target structure alone. Nature, 2022, 605: 551–560

8 Chen T, Vure P, Pulugurta R, et al. AMP-diffusion: Integrating latent diffusion with protein language models for antimicrobial peptide generation. bioRxiv:2024.03.03.583201, 2024

9 Corso G, Stärk H, Jing B, et al. Diffdock: Diffusion steps, twists, and turns for molecular docking. arXiv:2210.01776, 2022

10 Croitoru F A, Hondru V, Ionescu R T, et al. Diffusion models in vision: A survey. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 10850–10869

11 De Bortoli V, Mathieu E, Hutchinson M, et al. Riemannian score-based generative modelling. arXiv:2202.02763, 2022

12 Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst, 2021, 34: 8780–8794

13 Ding W, Nakai K, Gong H. Protein design via deep learning. Brief Bioinform, 2022, 23: bbac102

14 Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-multimer. bioRxiv:2021.10.04.463034, 2021

15 Ferreira L G, Dos Santos R N, Oliva G, et al. Molecular docking and structure-based drug design strategies. Molecules, 2015, 20: 13384–13421

16 Ferruz N, Heinzinger M, Akdel M, et al. From sequence to function through structure: Deep learning for protein design. Comput Struct Biotechnol J, 2023, 21: 238–250

17 Gohil S H, Iorgulescu J B, Braun D A, et al. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. Nat Rev Clin Oncol, 2021, 18: 244–256

18 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. arXiv:1406.2661, 2014

19 Gu S, Chen D, Bao J, et al. Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022, 10696–10606

20 Guo Z, Liu J, Wang Y, et al. Diffusion models in bioinformatics and computational biology. Nat Rev Bioeng, 2024, 2: 136–154

21 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 2020, 33: 6840–6851

22 Huang L, Xu T, Yu Y, et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. Nat Commun, 2024, 15: 2657
23 Huang P S, Boyken S E, Baker D. The coming of age of de novo protein design. Nature, 2016, 537: 320–327
24 Huang Z, Yang L, Zhou X, et al. Protein-ligand interaction prior for binding-aware 3D molecule diffusion models. In: The Twelfth International Conference on Learning Representations. Https://openreview.net/forum?id=qH9nrMNTIW, 2024
25 Igashov I, Stärk H, Vignac C, et al. Equivariant 3D-conditional diffusion model for molecular linker design. Nat Mach Intell, 2024, 6: 417–427
26 Ingraham J B, Baranov M, Costello Z, et al. Illuminating protein space with a programmable generative model. Nature, 2023, 623: 1070–1078
27 Jia C, Hu Y, Kelly D, et al. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res, 2017, 45: 10978–10988
28 Jo J, Lee S, Hwang S J. Score-based generative modeling of graphs via the system of stochastic differential equations. arXiv:2202.02514, 2022
29 Jovic D, Liang X, Zeng H, et al. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med, 2022, 12: e694
30 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
31 Khan S A, Lagani V, Lehmann R, et al. SpatialDiffusion: Predicting spatial transcriptomics with denoising diffusion probabilistic models. bioRxiv:2024.05.21.595094, 2024
32 Kingma D P, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114, 2013
33 Krishna R, Wang J, Ahern W, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science, 2024, 384: eadl2528
34 Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol, 2019, 20: 681–697
35 Li K, Li J, Tao Y, et al. stDiff: A diffusion model for imputing spatial transcriptomics through single-cell transcriptomics. Brief Bioinform, 2024, 25: bbae171
36 Li X, Zhu F, Min W. SpaDiT: Diffusion transformer for spatial gene expression prediction using scRNA-seq. arXiv:2407.13182, 2024
37 Lin Z Y, Li H, Lv L Z H, et al. TaxDiff: Taxonomic-guided diffusion model for protein sequence generation. arXiv: 2402.17156, 2024
38 Liu K, Mao W, Shen S, et al. Floating anchor diffusion model for multi-motif scaffolding. arXiv:2406.03141, 2024
39 Liu L, Ren Y, Lin Z, et al. Pseudo numerical methods for diffusion models on manifolds. arXiv:2202.09778, 2022
40 Liu S, Zhu T, Ren M, et al. Predicting mutational effects on protein-protein binding via a side-chain diffusion probabilistic model. Adv Neural Inf Process Syst, 2024, 36: 72495
41 Lu C, Zhou Y, Bao F, et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Adv Neural Inf Process Syst, 2022, 35: 5775–5787
42 Lu W, Zhang J, Huang W, et al. DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. Nat Commun, 2024, 15: 1071
43 Luo E, Hao M, Wei L, et al. scDiffusion: Conditional generation of high-quality single-cell data using diffusion model. Bioinformatics, 2024, 40: btae518
44 Ma J, Xue S, Hu T, et al. The surprising effectiveness of skip-tuning in diffusion sampling. arXiv:2402.15170, 2024
45 Mao J, Wang J, Zeb A, et al. Multimodal generation of astrocyte by integrating single-cell multi-omics data via deep learning. bioRxiv:2023.11.30.569500, 2023
46 Marsh J A, Teichmann S A. Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem, 2015, 84: 551–575
47 McNutt A T, Francoeur P, Aggarwal R, et al. GNINA 1.0: Molecular docking with deep learning. J Cheminform, 2021, 13: 43
48 Michelucci U. An introduction to autoencoders. arXiv:2201.03898, 2022
49 Morris G M, Lim-Wilby M. Molecular docking. Mol Model Proteins, 2008, 443: 365–382
50 Nichol A, Dhariwal P, Ramesh A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided

diffusion models. arXiv:2112.10741, 2022
51 Niu C, Song Y, Song J, et al. Permutation invariant graph generation via score-based generative modeling. arXiv: 2003.00638, 2020
52 Peebles W, Xie S. Scalable diffusion models with transformers. arXiv:2212.09748, 2022
53 Qiao Z, Nie W, Vahdat A, et al. State-specific protein-ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell, 2024, 6: 195–208
54 Ramasubramanian V S, Choudhuri S, Ghosh B. A hybrid diffusion model for stable, affinity-driven, receptor-aware peptide generation. bioRxiv:2024.03.14.584934, 2024
55 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Soc, 2022, 10684–10695
56 Sadria M, Layton A. The power of two: Integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis. bioRxiv:2023.04.13.536789, 2023
57 Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487, 2022
58 Schmidt U, Weigert M, Broaddus C, et al. Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention. Lecture Notes in Computer Science, vol. 11071. Cham: Springer, 2018, 265–273
59 Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: PMLR, 2015, 2256–2265
60 Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502, 2020
61 Song Y, Dhariwal P, Chen M, et al. Consistency models. arXiv:2303.01469, 2023
62 Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019, 11918–11930
63 Song Y, Garg S, Shi J, et al. Sliced score matching: A scalable approach to density and score estimation. arXiv: 1905.07088, 2019
64 Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations. arXiv:2011.13456, 2020
65 Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA, 2003, 100: 12123–12128
66 Stringer C, Wang T, Michaelos M, et al. Cellpose: A generalist algorithm for cellular segmentation. Nat Methods, 2021, 18: 100–106
67 Sverrisson F, Akdel M, Abramson D, et al. DiffMaSIF: Surface-based protein-protein docking with diffusion models. In: Machine Learning in Structural Biology Workshop at NeurIPS 2023. Https://hal.science/hal-04360638v1, 2023
68 Tang W, Liu R, Wen H, et al. A general single-cell analysis framework via conditional diffusion generative models. bioRxiv:2023.10.13.562243, 2023
69 Tang Z, Gu S, Bao J, et al. Improved vector quantized diffusion models. arXiv:2205.16007, 2022
70 Trippe B L, Yim J, Tischer D, et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. arXiv:2206.04119, 2022
71 Trott O, Olson A J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem, 2010, 31: 455–461
72 Vincent P. A connection between score matching and denoising autoencoders. Neural Comput, 2011, 23: 1661–1674
73 Wan X, Xiao J, Tam S S T, et al. Integrating spatial and single-cell transcriptomics data using deep generative models with SpatialScope. Nat Commun, 2023, 14: 7848
74 Wang X F, Huang X X, Price S J, et al. Cross-modal diffusion modelling for super-resolved spatial transcriptomics. arXiv:2404.12973, 2024
75 Wang X F, Tang J Y, Liang H, et al. ProT-Diff: A modularized and efficient approach to de novo generation of antimicrobial peptide sequences through integration of protein language model and diffusion model. bioRxiv: 2024.02.22.581480, 2024
76 Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion.

Nature, 2023, 620: 1089–1100
77 Wu K E, Yang K K, van den Berg R, et al. Protein structure generation via folding diffusion. Nat Commun, 2024, 15: 1059
78 Xu M, Yu L, Song Y, et al. Geodiff: A geometric diffusion model for molecular conformation generation. arXiv: 2203.02923, 2022
79 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the International Conference on Computer Vision (ICCV). Paris: ICCV, 2023, 3836–3847
80 Zhang S, Jiang Z, Huang R, et al. PRO-LDM: Protein sequence generation with a conditional latent diffusion model. bioRxiv:2023.08.22.554145, 2023
81 Zhang Z, Liu L. scIDPMs: Single-cell RNA-seq imputation using diffusion probabilistic models. bioRxiv:2024.02.29.582870, 2024
82 Zheng K, Lu C, Chen J, et al. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. Adv Neural Inf Process Syst, 2023, 36: 55502–55542
83 Zhou Z, Chen D, Wang C, et al. Fast ode-based sampling for diffusion models in around 5 steps. In: Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition. Seattle: CVPR, 2024, 7777–7786
84 Zhu H, Slonim D K. From noise to knowledge: Diffusion probabilistic model-based neural inference of gene regulatory networks. bioRxiv:2023.11.05.565675, 2023
85 Zhu J, Gu Z, Pei J, et al. DiffBindFR: An SE(3) equivariant network for flexible protein-ligand docking. Chem Sci, 2024, 15: 7926–7942

The diffusion model and its applications in bioinformatics

Zeyu Wu, Yiwei Fu, Jiaxiao Chen, Jianwen Ma, Wenrui Wu & Minghua Deng

Abstract In recent years, diffusion models have garnered significant attention, achieving remarkable success in many fields such as computer vision, natural language processing, and bioinformatics, while also demonstrating broad application prospects. In this survey paper, we elucidate the probabilistic significance of diffusion models, outline their development processes, and provide a detailed account of their applications in bioinformatics. In the beginning, we define the forward and backward processes of diffusion models from the perspective of stochastic differential equations, and we elaborate on the probabilistic principles of three types of diffusion models: NCSN, DDPM, and DDIM, in conjunction with the derivation of score matching. Due to their powerful generative capabilities, diffusion models have found extensive applications in bioinformatics, particularly in the areas of transcriptomics and protein research. Furthermore, these models not only excel in imputing and denoising single-cell and spatial transcriptomic sequencing data but also demonstrate significant effectiveness in protein design, structure prediction, and the design of small molecules and peptide drugs.

Keywords **diffusion model, stochastic differential equation, transcriptome imputation, protein design, drug design**

MSC(2020) **60J60, 60H30, 62P10, 92B05**

doi: 10.1360/SSM-2024-0316