基于扩散模型的三维分子设计

毕业设计开题报告

徐璨

浙江工商大学统计与数学学院

2023 年 3 月 15 日

导师：王伟刚教授

研究背景

- 近年来基于深度学习的生成模型也在多领域有成功应用，例如 AI 在图画、语音、视频、对话等应用上的优秀表现引发了社会对人工智能新一轮广泛关注与热烈讨论。
- 在智能计算的计算医药相关研究中，深度学习模型在药物发现、药物属性预测等应用中已经展现出良好的性能和极大的潜力。人工智能技术应用能够为药物研发的多个阶段降本增效。过去的 2022 年，AI 制药赛道相关融资总金额达百亿美元。国内互联网巨头如百度百图生科、华为 EIHealth、腾讯云深智药，及初创企业晶泰科技，剂泰医药，星药科技等，相关成果已经展现出深度学习在该领域的强大性能和广阔前景。

研究意义

- 近来大火的生成模型被广泛应用于智能计算领域，人工智能算法有望根据人类的要求生成理想的结果，帮助提升药物发现与设计的效率与质量。
- 在计算化学的相关研究中，深度学习模型的成功落地能够推动制药企业减少湿实验成本，助力靶点确认、药物发现、分子生成、化学反应设计、化合物筛选、临床试验、风险评估等多阶段。
- 本文聚焦于深度学习算法在三维药物分子发现这一主题，意在利用时下最优的生成模型算法，创新性设计出更符合三位药物理化性质的算法，提升全新药物分子设计的性能与效率。

生成模型

- 生成模型近年来在作画、视频创作、空间建模、文本生成等多方面都有了成功的应用。早期的生成模型研究主要集中在变分自编码器（VAE）、生成对抗网络（GAN）、流型模型（Flow-based models）、自回归模型（Auto-regressive models）及深度强化学习领域。近两年来，基于扩散模型的生成模型（Diffusion-based models）以其优异的性能、计算资源消耗相对较少在学术圈与工业界爆火。
- 相关 Diffusion 模型的研究包括对干扰、去噪过程的改进（DDPMs，SGMs，Score SDEs），采样策略的改进，损失函数，各类型数据应用（计算机视觉、自然语言处理、时序建模、信号传递、多模态学习、图建模、医学影像）等。

计算医药

- 生成模型的发展也带动了其在智能计算领域的应用。5 年前，药物分子、蛋白质配体分子等分子学习任务主要聚焦于 2D 分子结构，但实际上，分子结构信息远不止原子间的拓扑结构信息这么简单，因此某一分子在自然界中可能存在种类繁多的同分异构体，故近来对分子的研究拓展到了结合 3D 信息对不同分子构型的相关研究。因此在 3D 分子生成任务上，目标不再是生成简单的可能有效的分子，更要生成具备理想属性和性质稳定的分子 3D 构型。
- 3D 分子生成任务又可分为 3D 构型生成与全新分子生成。3D 构型生成的目标是根据给定的分子式，生成理想的三维构型。全新分子生成的目标则是凭空生成全新的、有效的分子。
- 3D 构型生成的研究有基于 VAE 的 CVGAE，GraphDG，ConfVAE，基于扩散模型的 ConfGF，EVFN，GeoDiff 等。

计算医药

- 2D 分子生成的研究有基于 VAE 的 TransVAE，JT-VAE，基于 GAN 的 MolGAN，基于 Flow 和 AR 的 GraphDF，GraphAF，MoFlow，基于能量函数的 GraphEBM 等。
- 3D 分子生成的研究有基于 AR 的 G-SchNet，GEN3D，基于 Flow 的 E-NFs，G-SphereNet，基于强化学习的 MolGym，基于 VAE 的 3DMolNet。基于这些模型的算法存在着计算需求大，效果有限的问题。
- 目前基于 Diffusion 的 3D 分子生成模型仅有两个，EDM（ICML-22）[?] 和 MDM（AAAI-22）[?] 在扩散模型的内核设计上仍沿用了之前研究的设计如 SchNet[?] 等，本人认为传统的图网络在空间几何建模上存在劣势，且不符合扩散模型的内在机理。

前期准备

- 本文意在利用生成模型，提出创新的算法，生成全新且有效的分子。
- 本文首先选取 GEOM-QM9 和 GEOM-Drugs 作为研究数据集。GEOM-QM9 包含 130831 个分子，平均每个分子由 19 个分子构成。
- GEOM-Drugs 包含超过 600W 个分子构型，将每个分子的不同构型经过筛选后，处理得到 29W 个稳定的分子构型，平均每个分子由 44 个原子构成。

生成模型构建

- 本文基于 DDPMs 构建分子生成模型框架，包括扩散过程与去噪过程。
- 扩散过程本质是通过给初始数据有规律地增加噪声，使其经过 T 时间步后收敛至高斯噪声。
- 去噪过程的本质是通过设计的去噪神经网络内核，将高斯噪声还原至初始状态，使其与初始数据相似。

等变图网络构建

- 之前的研究基于的图网络基本比较简单，如 SchNet[?]。
- 先前的图网络对于 3D 几何信息的建模能力有限，且不符合扩散模型的内在本质。本文提出运用全局注意力机制代替图卷积，对 3D 几何信息进行建模。
- 本文等变图网络对三维空间信息的建模分为两个并行模块：inter-atomic 级别的和 pair-wise 级别的。在每个 Interblock 中，两个并行模块分别对原子和原子对的特征通过全局注意力机制进行学习，并通过交换模块对信息进行统一。

生成过程示例

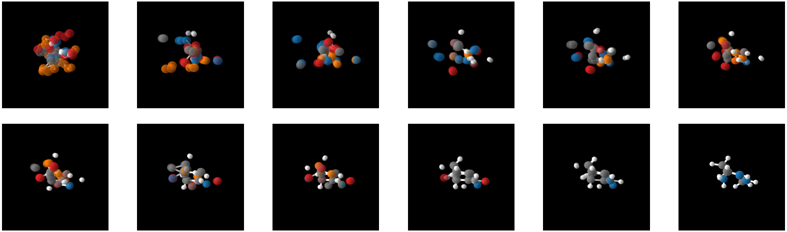

图 1: 生成流程示意图 “CNC1=C(N)CCN1”

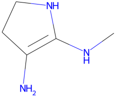

图 2: 分子 “CNC1=C(N)CCN1” 理论结构

生成结果示例

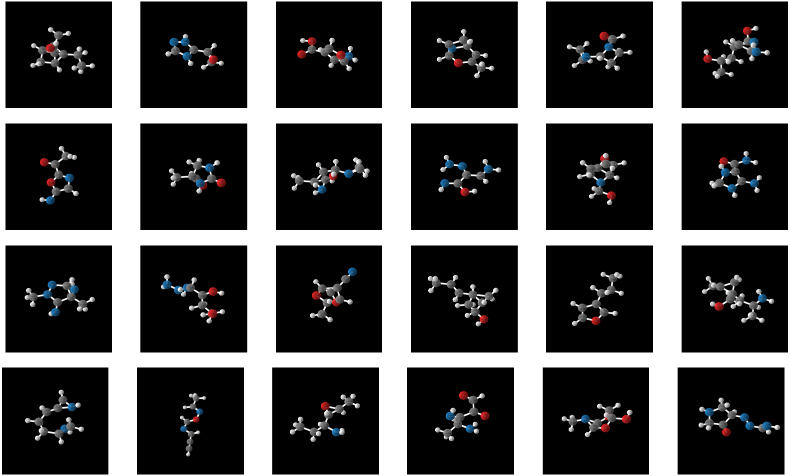

图 3: 生成分子示例

- 2023.01-2023.03：前期调研与相关文献阅读
- 2023.03-2023.04：模型搭建
- 2023.04-2023.05：模型训练，性能调优与对比试验
- 2023.05-2023.07：文章撰写
- 2023.07-2023.09：文章修改

Thanks!