



浙江工商大学

硕士学位论文

论文题目：基于扩散模型的三维药物分子设计框架

作者姓名：徐璨

学科专业：统计学

研究方向：数理统计

指导教师：王伟刚

提交日期：2024 年 1 月

**Dissertation Submitted to Zhejiang Gongshang University
for Master's Degree of Science**

A Diffusion-based 3D Molecule Generative Framework

Author: Can Xu

Major: Statistics

Supervisor: Prof. Weigang Wang



Jan. 2024

School of Statistics and Mathematics

Zhejiang Gongshang University

Hangzhou, 310018, P. R. China

基于扩散模型的三维药物分子设计框架

摘 要

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。

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。

关键词: 扩散模型; 分子生成; 几何神经网络

A DIFFUSION-BASED 3D MOLECULE GENERATIVE FRAMEWORK

ABSTRACT

AAAAAAAAAAAAAA

KEYWORDS: Diffusion model; Molecule generation; Geometry neural network

第1章 引言

1.1 选题背景与研究意义

1.1.1 选题背景

除图像、视频，音频与自然语言处理等领域，AI 技术的快速发展也带动相关交叉学科的发展。AI4Science 近年在计算生物、计算化学、材料设计、计算天文，计算育种等都有广泛应用，相关 AI 技术的应用能够大幅加速相关科学研究进展。在计算制药领域，近年来相关 AI 技术在药物性质预测，生成，开发，实验等领域的运用不仅加速相关研究的进展，也能够降低相关研究的研发成本。

基于 AI 的生成模型近十年来也被广泛研究，他们包括变分自编码器 (Variational autoencoders / VAEs)^[1]，生成对抗模型 (Generative adversarial networks / GANs)^[2]，流形模型 (Normalizing flow)^[3-4]，自回归 (Auto-regressive)^[5]，强化学习与扩散模型 (Diffusion)^[6-7] 等。相关方法在图像，文字等方面也有了许多成功应用。

深度学习在分子化学领域近年来也有着成功的应用。以分子学习为例，化学分子常以简化分子线性输入规范字符串 (Simplified molecular-input line-entry system / SMILES)^[8] 存储，每一个分子式对应一个 SMILES 字符串。随着早期深度学习方法，如卷积神经网络 (Convolutional neural networks / CNNs)^[9] 和循环神经网络 (Recurrent neural networks / RNNs)^[10-12] 的发展，一些研究试图运用这些深度学习算法，对以字符串形式存在的分子式进行学习，以获得预测特定原子或是分子整体的性质的能力。随着图神经网络 (Graph neural networks / GNNs)^[13-15] 的出现，其对非结构化数据的建模能力和对节点间拓扑关系学习的能力被证明十分优异。分子作为自然界中天然存在的图结构，原子和键对应着图中的节点和边，这为分子学习提供了新的思路与方法。从最早的图卷积神经网络开始，相关研究者致力于提出新的图学习算法，以提升对分子图学习的性能。随着相关化学模拟技术的发展，让三维分子建模成为可能。这也在拓扑结构信息以外，提供了更丰富的几何构型信息，这也驱动着相关研究拓展至几何图神经网络上。分子三维构象允许研究者对分子进行更准确的研究，同时也推动更多复杂任务的出现，包括分子生成，Ligand 生成，Protacs 生成，药物亲和力预测，蛋白质预测等等。

1.1.2 研究意义

人工智能在智能计算相关研究中开始扮演越来越重要的角色，相关模型在药物发现、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。由于深度学习技术应用具备为药物研发的多阶段降本增效的潜力，在 2022 年，AI 制药赛道相关企业融资总金额达百亿美元。在这一赛道竞逐的有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。

1.2 文献综述

1.2.1 基于深度学习的分子学习

分子最早被表示为简化分子线性输入规范字符串 (SMILES)^[8]，随着早期深度学习模型卷积神经网络 (CNN) 和循环神经网络 (RNN) 的发展，相关模型利用 CNNs 和 RNNs 对分子性质做出学习。Hirohara 等^[9]的研究提出使用 CNN 对分子级别的特征和分子基团性质进行有效学习。由于 RNNs 在早期自然语言处理任务上有良好表现，Bjerrum^[10]提出 LSTM-QSAR 模型用于学习分子性质。伴随图神经网络 (GNNs)

Transformer^[16-18]

1.2.2 生成式人工智能

1.2.3 分子生成

现有的基于扩散的分子生成模型有且只有 EDM^[19]和 MDM^[20]。

1.3 创新点

本文聚焦于扩散模型在分子生成领域的前沿研究，意在提出一个能够设计出有效、稳定分子的生成模型框架。本文利用扩散模型作为生成框架的骨架，在扩散模型的去噪内核设计上，本文提出了一个全新的图学习算法。该算法能够对几何信息，拓扑信息和原子化学性质进行分别建模，在保证模型等变形的条件下，实现对多种信息的有效学习与利用。

1.4 基本框架

第2章 基于扩散模型的分子生成

参考文献

- [1] KINGMA D P, WELLING M. Auto-encoding variational bayes[A]. 2013. arXiv: 1312.6114.
- [2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems: Vol. 27. Curran Associates, Inc., 2014.
- [3] DINH L, KRUEGER D, BENGIO Y. Nice: Non-linear independent components estimation. arxiv e-prints, 2014[C]//Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego. 2015.
- [4] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real NVP[C]//International Conference on Learning Representations. 2017.
- [5] VAN DEN OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel recurrent neural networks[C]//BALCAN M F, WEINBERGER K Q. Proceedings of Machine Learning Research: Vol. 48 Proceedings of The 33rd International Conference on Machine Learning. New York, New York, USA: PMLR, 2016: 1747-1756.
- [6] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//BACH F, BLEI D. Proceedings of Machine Learning Research: Vol. 37 Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015: 2256-2265.
- [7] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[C]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems: Vol. 32. Curran Associates, Inc., 2019.
- [8] WEININGER D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules[J/OL]. Journal of Chemical

-
- Information and Computer Sciences, 1988, 28(1): 31-36. DOI: 10.1021/ci00057a005.
- [9] HIROHARA M, SAITO Y, KODA Y, et al. Convolutional neural network based on smiles representation of compounds for detecting chemical motif [J/OL]. BMC bioinformatics, 2018, 19: 83-94. DOI: 10.1186/s12859-018-2523-5.
- [10] BJERRUM E J. Smiles enumeration as data augmentation for neural network modeling of molecules[A]. 2017. arXiv: 1703.07076.
- [11] LIU S, ALNAMMI M, ERICKSEN S S, et al. Practical model selection for prospective virtual screening[J/OL]. Journal of Chemical Information and Modeling, 2019, 59(1): 282-293. DOI: 10.1021/acs.jcim.8b00363.
- [12] HUANG K, FU T, GLASS L M, et al. Deeppurpose: a deep learning library for drug-target interaction prediction[J/OL]. Bioinformatics, 2020, 36(22-23): 5545-5547. DOI: 10.1093/bioinformatics/btaa1005.
- [13] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations. 2017.
- [14] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.
- [15] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks? [C]//International Conference on Learning Representations. 2019.
- [16] HONDA S, SHI S, UEDA H R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery[A]. 2019. arXiv: 1911.04738.
- [17] WANG S, GUO Y, WANG Y, et al. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction[C/OL]//BCB '19: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational

Biology and Health Informatics. New York, NY, USA: Association for Computing Machinery, 2019: 429–436. DOI: 10.1145/3307339.3342186.

- [18] CHITHRANANDA S, GRAND G, RAMSUNDAR B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction[A]. 2020. arXiv: 2010.09885.
- [19] HOOGEBOOM E, SATORRAS V G, VIGNAC C, et al. Equivariant diffusion for molecule generation in 3D[C]//Proceedings of Machine Learning Research: Vol. 162 Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022: 8867-8887.
- [20] HUANG L, ZHANG H, XU T, et al. Mdm: Molecular diffusion model for 3d molecule generation[A]. 2022.

硕士研究生期间的科研成果

论文:

XU C, ZHANG Y, WANG W, DONG L. Pursuit and evasion strategy of a differential game based on deep reinforcement learning[J]. Frontiers in Bioengineering and Biotechnology, 2022, 10: 827408.

ZHANG Y, **XU C**, WU X, ZHANG Y, DONG L, WANG W. LFGCF: Light folksonomy graph collaborative filtering for tag-aware recommendation[J]. Expert Systems with Applications, 2022, Under Review.

XU C, ZHANG Y, CHEN H, DONG L, WANG W. A fairness-aware graph contrastive learning recommender framework for social tagging systems[J]. Information Sciences, 2023, Accepted for Publication.

课题:

面向分布式异构计算系统内存池化关键技术, 国家重点研发计划之先进计算与新兴软件。

竞赛:

“华为杯”第十八届中国研究生数学建模竞赛, 三等奖, 排名: 1/3。

第五届全国应用统计专业学位研究生案例大赛, 三等奖, 排名: 1/3。

OGB-LSC @NeurIPS 2022 (PCQM4Mv2 Track), NO.11, 排名: 1/5。