密级：公开 中图分类号：O212.1

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

AAAAAAAAAAAAA

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 flows / NFs）[3-4]，自回归（Autoregressive models / ARs）[5] 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 在早期自然语言护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）的发展，由于分子的形式天然的属于图结构，一些研究开始使用 GNNs 进行分子学习。CGCNN[16]将图卷积神经网络引入到分子性质学习，用以模拟并替代复杂的 DFT 计算。Xiong 等[17]提出 Attentive FP，一个结合注意力机制的图神经网络实现分子性质的有效学习与预测。GraSeq[18]提出在运用图神经网络学习拓扑结构同时，利用双向 LSTM 模型对 SMILES 分子式进行学习，通过两个通道联合预测分子性质。伴随相关技术的发展，研究者可以不再拘泥于原子拓扑结构，进而实现对分子三维结构的建模与学习。鉴于某一分子对应大量的同分异构体，而不同构象对应属性不尽相同，因此对三维构象的有效学习是十分必要的。EGNN[19]在流行的图网络基础上，保留最基本的几何信息，即原子间距离，其简单的设计也成为了一些。在分子预训练框架 GEM[20]中，研究者提出 GeoGNN 图网络，将键长视作原子节点图的边特征，又对原子键构图，并将键角作为原子键图的边特征，通过在两个网络上的信息传递实现分子局部空间几何性质学习。Schütt 等[21]在继承 GNNs 的信息传递范式的同时，将原子间距离用径向基函数（Radial basis funciton / RBF）建模后融入边特征，使算法对三维几何信息的有效学习的同时保证了等变性要求。SphereNet[22]提出了基于球坐标系的信息传递范式 SMP。通过一系列参考原子或键的规则，SMP 在保证对原子

对距离，键角和键扭转角这三个空间几何信息完整提取的同时，避免计算复杂度的爆炸式增长。ComENet[23]在 ShpereNet 的基础上，简化了空间几何信息的提取范式，在保证利用完整空间信息的前提下，以损失部分精度为代价，大幅度提升计算速度。

伴随着 Transformer[24]相关研究在图像与文本领域的兴起，相关研究[25-27]也利用 Transformer 对 SMILES 分子式进行学习。随着 GTN[28]将 Transformer 引入图学习，越来越多的研究也将多头注意力机制用于分子图学习领域。大规模分子图预训练框架 GROVER[29]中，分子学习内核运用了 GTransformer，同时学习分子中的原子与键的节点嵌入（embedding）或边嵌入。在分子预训练框架 MPG[30]中提出的图学习内核 MolGNet 放弃了对边嵌入的学习，仅利用多头注意力学习节点嵌入，结果证明了该图学习算法的有效性。

1.2.2 基于深度生成模型的分子设计

自深度学习研究兴起以来，深度生成模型一直是研究者重点研究的对象。作画、翻译、对话，渲染等应用能够直接服务于广大用户。主流的深度生成模型有变分自编码器（Variational autoencoders / VAEs)[1]，生成对抗模型（Generative adversarial networks / GANs)[2]，流形模型（Normalizing flows / NFs)[3-4]，自回归（Autoregressive models / ARs)[5]与扩散模型（Diffusion)[6-7]等。

与图学习的演进过程相似，基于深度生成模型的分子设计也经历了从二维图结构到三维几何构象的演进。主流分子设计任务具体又可以被细分为：分子生成，分子优化，构象生成，蛋白质配体生成，蛋白质生成，蛋白降解靶向嵌合体生成。

分子生成的任务就是根据给定分子数据，使模型具备凭空生成全新且有效的药物分子图或三维结构。考虑到复杂药物分子主要由官能团等子结构组成，JT-VAE[31]基于 VAE 生成树结构骨架，而后利用树结构骨架逐步生成分子图结构。GraphVAE[32]是早期的基于 VAE 的图生成研究，为避免离散化结构的线性表示的相关障碍，使其中解码器直接输出预设的最大概率的全连接图。基于 NF 模型，MoFlow[33]将隐式表征逐步映射到条件流过程中，模型首先生成连接原子的键，随后通过图条件流生成原子，并最终组成有效的分子图。

现有的基于扩散的分子生成模型有且只有 EDM[34]和 MDM[35]。

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/ci0005 7a005.

[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/s12859-018-2 523-5.

[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] XIE T, GROSSMAN J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J/OL]. Phys. Rev. Lett., 2018, 120: 145301. DOI: 10.1103/PhysRevLett.120.145301.

[17] XIONG Z, WANG D, LIU X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism[J/OL].

Journal of Medicinal Chemistry, 2020, 63(16): 8749-8760. DOI: 10.1021/acs.jmedchem.9b00959.

[18] GUO Z, YU W, ZHANG C, et al. Graseq: Graph and sequence fusion learning for molecular property prediction[C/OL]//CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York, NY, USA: Association for Computing Machinery, 2020: 435–443. DOI: 10.1145/3340531.3411981.

[19] SATORRAS V G, HOOGEBOOM E, WELLING M. E(n) equivariant graph neural networks[C]//MEILA M, ZHANG T. Proceedings of Machine Learning Research: Vol. 139 Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 9323-9332.

[20] FANG X, LIU L, LEI J, et al. Geometry-enhanced molecular representation learning for property prediction[J]. Nature Machine Intelligence, 2022, 4(2): 127-134.

[21] SCHüTT K, KINDERMANS P J, SAUCEDA FELIX H E, et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions[C]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[22] LIU Y, WANG L, LIU M, et al. Spherical message passing for 3d molecular graphs[C]//International Conference on Learning Representations. 2022.

[23] WANG L, LIU Y, LIN Y, et al. ComENet: Towards complete and efficient message passing for 3d molecular graphs[C]//OH A H, AGARWAL A, BELGRAVE D, et al. Advances in Neural Information Processing Systems. 2022.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[25] HONDA S, SHI S, UEDA H R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery[A]. 2019. arXiv: 1911.04738.

[26] 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.

[27] CHITHRANANDA S, GRAND G, RAMSUNDAR B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction[A]. 2020. arXiv: 2010.09885.

[28] YUN S, JEONG M, KIM R, et al. Graph transformer networks[C]// WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems: Vol. 32. Curran Associates, Inc., 2019.

[29] RONG Y, BIAN Y, XU T, et al. Self-supervised graph transformer on large-scale molecular data[C]//Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 12559-12571.

[30] LI P, WANG J, QIAO Y, et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery[J/OL]. Briefings in Bioinformatics, 2021, 22(6). DOI: 10.1093/bib/bbab109.

[31] JIN W, BARZILAY R, JAAKKOLA T. Junction tree variational autoencoder for molecular graph generation[C]//DY J, KRAUSE A. Proceedings of Machine Learning Research: Vol. 80 Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 2323-2332.

[32] SIMONOVSKY M, KOMODAKIS N. Graphvae: Towards generation of small graphs using variational autoencoders[C]//Artificial Neural Networks and Machine Learning – ICANN 2018. Cham: Springer International Publishing, 2018: 412-422.

[33] ZANG C, WANG F. Moflow: An invertible flow model for generating molecular graphs[C/OL]//KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York,

NY, USA: Association for Computing Machinery, 2020: 617–626. DOI: 10.1145/3394486.3403104.

[34] 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.

[35] 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: 119064.

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。