Graph and Geometry Generative Modeling for Drug Discovery

Minkai Xu*
minkai@cs.stanford.edu
Stanford University
CA, USA

Meng Liu*
mengliu@tamu.edu
Texas A&M University
TX, USA

Wengong Jin*
wengong@csail.mit.edu
Broad Institute of MIT and Harvard
MA, USA

Shuiwang Ji*
sji@tamu.edu
Texas A&M University
TX, USA

Jure Leskovec*
jure@cs.stanford.edu
Stanford University
CA, USA

Stefano Ermon*
ermon@cs.stanford.edu
Stanford University
CA, USA

ABSTRACT

With the recent progress in geometric deep learning, generative modeling, and the availability of large-scale biological datasets, molecular graph and geometry generative modeling have emerged as a highly promising direction for scientific discovery such as drug design. These generative methods enable efficient chemical space exploration and potential drug candidate generation. However, by representing molecules as 2D graphs or 3D geometries, there exist many both fundamental and challenging problems for modeling the distribution of these irregular and complex relational data. In this tutorial, we will introduce participants to the latest key developments in this field, covering important topics including 2D molecular graph generation, 3D molecular geometry generation, 2D graph to 3D geometry generation, and conditional 3D molecular geometry generation. We further include antibody generation, where we particularly consider large-size antibody molecules. For each topic, we will outline the underlying problem characteristics, summarize key challenges, present unified views of the representative approaches, and highlight future research direction and potential impacts. We anticipate this lecture-style tutorial would attract a broad audience of researchers and practitioners.

ACM Reference Format:
Minkai Xu, Meng Liu, Wengong Jin, Shuiwang Ji, Jure Leskovec, and Stefano Ermon. 2023. Graph and Geometry Generative Modeling for Drug Discovery. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599559

1 TUTORIAL DESCRIPTION

1.1 Target Audience and Goals

In this tutorial, we will introduce the audience to the state-of-the-art approaches in molecular graph and geometry generative modeling for drug discovery. The tutorial will cover a wide range of topics, including 2D molecular graph generation, 3D molecular geometry generation, 2D graph to 3D geometry generation, conditional 3D molecular geometry generation, and antibody generation. Our goal is to provide the audience with an in-depth understanding of the underlying principles, challenges, and recent advancements in the field. This target audience includes (1) researchers who work on or are interested in advanced graph and geometry generative modeling, AI for science, or drug discovery, and (2) industry professionals who are interested in the latest techniques and tools in graph and geometry generative modeling applied to drug discovery.

1.2 Outline

We will first introduce the required preliminary knowledge and then introduce the latest developments in the covered topics.

Preliminaries. To help participants better understand the delivered tutorial, we will first introduce the necessary preliminaries. We will introduce the key ML methods including graph and geometry representation learning [1] and the generative modeling methods such as diffusion models [10]. We will also provide basic concepts of molecular representations in 2D and 3D formats.

2D molecular graph generation. Machine learning for drug discovery aims to automate the design of molecules with desirable properties, and an intrinsic and informative way to represent molecules is the molecular graph, where nodes and edges are labeled with atom and bond type respectively. In this tutorial, we will cover the most recent progress in this area, such as junction-tree VAE (JT-VAE) [2] and graph flows [8, 9].

3D molecular geometry generation. Generating molecules in 3D space is another critical problem since 3D molecular geometries are directly related with many fundamental molecular properties. This part will introduce two categories of 3D molecule generation methods. One type directly generates the atomic coordinates in 3D space, such as Geometric Latent Diffusions [11]. The other type of method considers generating invariant variables for obtaining coordinates (distances, angles, and torsion angles) such as G-SphereNet [7].

2D graph to 3D geometry generation. Predicting 3D molecular geometries from 2D molecular graphs is another fundamental problem in cheminformatics and drug discovery such as molecular docking. In this tutorial, we will cover the recent progress in this field, including ConfVAE [12] and the most recent GeoDiff [13], which generates 3D molecular geometries by directly running diffusion processes on the coordinates.

Conditional 3D molecular geometry generation. Structure-based drug design (SBDD) seeks to design 3D molecules that bind to given target proteins. In this tutorial, we will introduce the most recent advancements in this direction. We will cover GraphBP [6]

*All authors contributed equally.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599559

which adopts autoregressive models to generate atoms in the target protein context one by one and recent methods which generate 3D molecules in a one-shot fashion with diffusion models [5].

Antibody generation. In the last part, we will cover antibody generation, which are large size molecular structures with high realistic values for drug discovery. Antibodies are versatile proteins that bind to antigens like viruses and stimulate the adaptive immune system. The goal is to generate amino acid sequences with the right 3D structures to bind to a target antigen. We will introduce some recent works in the tutorial, including RefineGNN [4], HERN [3].

2 BRIEF BIOGRAPHIES OF TUTORS

Minkai Xu is currently a Ph.D. student in the Computer Science Department at Stanford University. His research lies in generative models and machine learning for scientific discovery. He has published papers on the above topics in top machine learning conferences such as ICML, NeurIPS, ICLR, AAAI, and AAMAS. He served as a session chair of ML4Bio division on ICML 2021, and regularly served as the program committee member for major machine learning conferences such as ICML, NeurIPS, ICLR, and AAAI.

Meng Liu is currently a Ph.D. student in the Department of Computer Science and Engineering, Texas A&M University. His research focuses on graph representation learning, generative modeling, and AI for scientific problems. He also delivered a tutorial "Frontiers of Graph Neural Networks with DIG" on KDD 2022. He served as a session chair of ICML 2022, and regularly served as a program committee member for major machine learning conferences and journals, such as ICML, NeurIPS, ICLR, KDD, JMLR, and TPAMI.

Wengong Jin is a Postdoctoral Fellow at Eric and Wendy Schmidt Center of Broad Institute. He received his Ph.D. from MIT CSAIL in 2021. His research focuses on developing generative models to design novel therapeutic molecules such as small-molecule drugs and antibodies. He is a recipient of MIT EECS Outstanding Ph.D. Thesis Award, Dimitris N. Chorafas Prize, and Koch Institute Frontier Award. His research is published in top AI conferences (ICML, ICLR, and NeurIPS) and biology journals (Cell and PNAS).

Shuiwang Ji is a Professor and Presidential Impact Fellow in the Department of Computer Science and Engineering, Texas A&M University. His research interests include machine learning for graphs, molecules, materials, and quantum systems. He received the NSF CAREER Award in 2014. He is currently an Associate Editor for IEEE TPAMI, ACM TKDD, and ACM Computing Surveys. He regularly serves as an Area Chair or equivalent roles for data mining and machine learning conferences, including AAAI, ICLR, ICML, IJCAI, KDD, and NeurIPS. He is a Fellow of IEEE and AIMBE, and a Distinguished Member of ACM.

Jure Leskovec is an Associate Professor of Computer Science at Stanford University where he is a member of the InfoLab and the AI lab. His general research area is applied machine learning for large interconnected systems. His research has won several awards including a Lagrange Prize, Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship, and numerous best paper and test of time awards. It has also been featured in popular press outlets such as the New York Times and the Wall Street Journal.

Stefano Ermon is an Associate Professor in the Department of Computer Science at Stanford University. His research is mainly around developing innovative solutions to problems of broad societal relevance through advances in probabilistic modeling, learning, and inference. Among others, he co-organized the 2021 KDD Tutorial on "Challenges in KDD and ML for Sustainable Development", the NeurIPS 2019 workshop on Information Theory and Machine Learning, the ICML 2022 Workshop "Adaptive Experimental Design and Active Learning in the Real World".

ACKNOWLEDGMENTS

MX and JL gratefully acknowledge the support of DARPA under Nos. HR00112190039 (TAMI), N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), NIH under No. 3U54HG010426-04S1 (HuBMAP), Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Amazon, Docomo, GSK, Hitachi, Intel, JPMorgan Chase, Juniper Networks, KDDI, NEC, and Toshiba. MX and SE gratefully acknowledge the support of NSF (#1651565), ARO (W911NF-21-1-0125), ONR (N00014-23-1-2159), CZ Biohub, Stanford HAI. MX thanks the generous support of Sequoia Capital Stanford Graduate Fellowship. ML and SJ gratefully acknowledge the support of NSF under Nos. IIS-1955189, IIS-1908220, and IIS-2006861; NIH under No. U01AG070112; Cisco, and Texas A&M Presidential Impact Fellowship. WJ gratefully acknowledges the support of Eric and Wendy Schmidt Center at Broad Institute.

REFERENCES

[1] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* (2017).
[2] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*. PMLR, 2323–2332.
[3] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2022. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*. PMLR, 10217–10227.
[4] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi S Jaakkola. 2022. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. In *International Conference on Learning Representations*.
[5] Haitao Lin, Yufei Huang, Meng Liu, Xuanjing Li, Shuiwang Ji, and Stan Z Li. 2022. DiffBP: generative diffusion of 3D molecules for target protein binding. *arXiv preprint arXiv:2211.11214* (2022).
[6] Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. 2022. Generating 3D Molecules for Target Protein Binding. In *International Conference on Machine Learning*. PMLR, 13912–13924.
[7] Youzhi Luo and Shuiwang Ji. 2022. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations (ICLR)*.
[8] Youzhi Luo, Keqiang Yan, and Shuiwang Ji. 2021. GraphDF: A discrete flow model for molecular graph generation. In *International Conference on Machine Learning*. PMLR, 7192–7203.
[9] Chence Shi*, Minkai Xu*, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. 2020. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. In *International Conference on Learning Representations*.
[10] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
[11] Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. 2023. Geometric Latent Diffusion Models for 3D Molecule Generation. In *International Conference on Machine Learning*. PMLR.
[12] Minkai Xu, Wujie Wang, Shitong Luo, Chence Shi, Yoshua Bengio, Rafael Gomez-Bombarelli, and Jian Tang. 2021. An end-to-end framework for molecular conformation generation via bilevel programming. In *International Conference on Machine Learning*. PMLR, 11537–11547.
[13] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*.