Garg, Vikas

Generative AI for graph-based drug design: Recent advances and the way forward

Published in:
Current Opinion in Structural Biology

DOI:
10.1016/j.sbi.2023.102769

Published: 01/02/2024

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Generative AI for graph-based drug design: Recent advances and the way forward

Vikas Garg

Abstract

Discovering new promising molecule candidates that could translate into effective drugs is a key scientific pursuit. However, factors such as the vastness and discreteness of the molecular search space pose a formidable technical challenge in this quest. AI-driven generative models can effectively learn from data, and offer hope to streamline drug design. In this article, we review state of the art in generative models that operate on molecular graphs. We also shed light on some limitations of the existing methodology and sketch directions to harness the potential of AI for drug design tasks going forward.

Addresses

Aalto University and YaiYai Ltd, Finland

Corresponding author: Garg, Vikas (vgarg@csail.mit.edu), (vikas@yaiyai.fi)

(Garg V.)

Current Opinion in Structural Biology 2024, **84**:102769

This review comes from a themed issue on **Artificial Intelligence (AI) Methodologies in Structural Biology (2024)**

Edited by **Tero A. Aittokallio** and **Evandro Fei Fang**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2023.102769

Keywords

Graph neural networks, Generative models, Molecule design, Drug design, Neural ODEs, Deep learning, Machine learning.

Introduction

Drug discovery is a cumbersome, time-consuming, and expensive process with a very low success rate. Searching for new molecules that could, e.g., bind to a protein target (available explicitly or assumed implicitly) is complicated due to myriad factors including but not limited to the overabundance of drug-like structures and the discreteness of the search space resulting in a complex and challenging landscape to optimize.

Generative models can exploit strong priors, or *inductive biases*, and replace the expensive search operation over the molecular space with a significantly easier verification step. Specifically, these models can learn to generate suggestions for molecules that are similar to a given dataset of molecules. These generated molecules can be subsequently screened based on considerations such as the feasibility of their synthesis [1], binding affinity to a target protein [2], bioactivity, and physicochemical properties.

A variety of generative models have been proposed recently in the context of drug design. Molecular data can be expressed in multiple formats, e.g., SMILES sequences, 2D/3D graphs, Morgan fingerprints, images, etc.; moreover, the protein targets may or may not be specified, so accordingly, a wide range of models and approaches have been developed to handle different formats and scenarios [3–9] including those inspired by language models and generative pre-trained transformers (GPT) [10,11]. Here, we focus specifically on generative models that are based on molecular graphs, reviewing several methods that have been proposed recently.

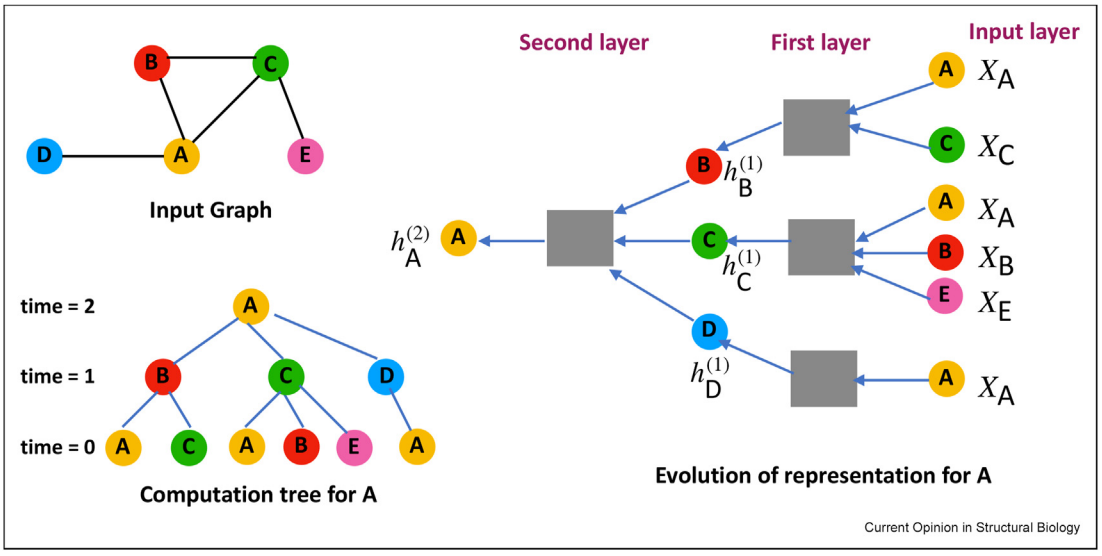
The rest of this article is organized as follows. We first discuss in Section 2 the representation of molecular data, focusing on the most prominent models for encoding graph-structured data currently, namely, the graph neural networks. We then review several recent generative models for drug design in Section 3. Finally, in Section 4, we bring attention to some shortcomings of the existing methods, charting a way forward.

Encoding molecular data

Graph neural networks (GNNs) [12–14] have continued to gain prominence as models of choice for encoding molecular graphs (Figure 1). It is well known that the standard message-passing GNNs are no more powerful than the color refinement algorithm or 1-Weisfeiler Leman (1-WL) test for isomorphism [13]. Many of the recent advances in GNNs are motivated by *expressivity* considerations, i.e., learning a class of functions with greater representational power than the existing methods (Figure 2).

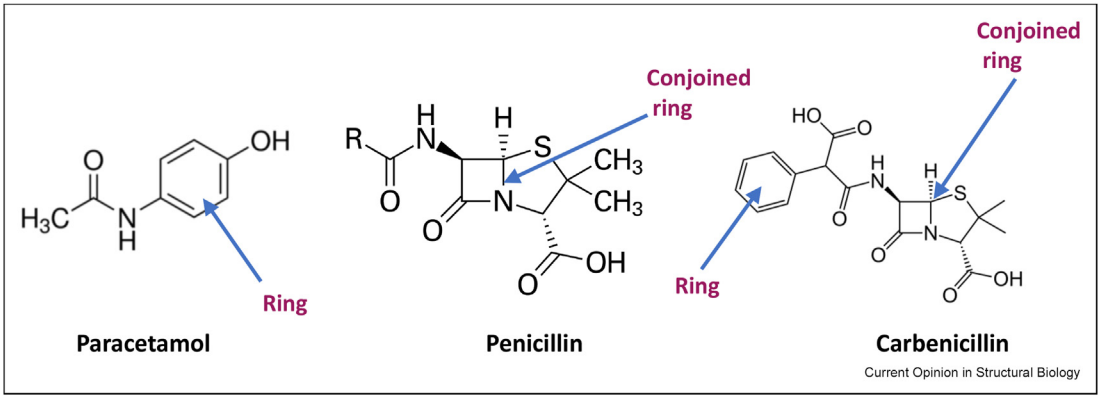
Particularly in the molecular context [14], showed inability of message-passing GNNs to compute graph properties such as counting the number of rings, or

Figure 1



Graph neural network (GNN) in action. GNNs are state-of-the-art models for embedding graphs that form the backbone of most graph-based generative models. Here, we assume a 2-layer GNN that processes the given input graph over 5 attributed nodes (e.g., atoms). Initially, at time $t = 0$, each node starts with its own feature vector as its initial embedding. During each layer, nodes exchange *messages* with their neighbors on the graph. Nodes apply transformations to aggregate the information in incoming messages and update their embeddings. Final embedding of node A after two layers, i.e., at time $t = 2$ depends on the embeddings of its neighbors *B*, *C*, and *D* at $t = 1$, which in turn depends on the embeddings of their respective neighbors at $t = 0$. Gray color indicates learnable weights used to aggregate information. Nodes share weights in each layer.

Figure 2



Obtaining effective representations for molecules is a key aspect of molecular modeling. Routinely used drugs such as Penicillin contain (conjoined) rings that standard message-passing GNNs cannot compute [14]. Therefore, more powerful models, including those that use geometric information (3D coordinates) have emerged recently.

determining when conjoined rings are present. The failure to compute substructures has also been established for so-called invariant graph networks [15]. Such negative results can be countered to an extent with 3D molecular graphs that incorporate additional spatial or geometric features [16]. In order to obtain embeddings for such 3D graphs, geometric GNN models such as H-DCPN [14], GemNet [17], and SphereNet [18] have been proposed. Such models often implement *directional* message passing taking into account distances to the neighboring nodes, and torsion angles.

Recently, several models used equivariant or invariant GNN layers to obtain representations that respect the underlying symmetries (such as rotation, translation, and reflection) [19], which has led to improved performance in various generative molecular tasks [20]. Since the standard WL test does not suffice for analysis of the

limits on the expressivity of such geometric GNNs, a geometric version of WL, namely GWL, has been introduced recently [21]. In particular, GWL can be invoked to establish the gap in expressivity between equivariant layers and invariant layers.

Graph-based generative models for molecules

We now present an overview of some graph-based generative models that have garnered interest within the AI community recently.

Flows and diffusion models

Recently, there has been a flurry of works on flow-based models [22] for molecular generation. These models can be viewed as successors of hierarchical variational autoencoders (VAEs) [23] with one key difference, namely, the encoder and the decoder are aligned with each other unlike VAEs, where low-dimensional encoding of data samples into the latent space leads to a shift between the true data distribution and the reconstructed one obtained by decoding the latent samples. Specifically, flow models perform a series of bijective, i.e., invertible transformations, typically parameterized with neural networks, to map samples from a tractable distribution (such as Gaussian) to a significantly more complex distribution. Consequently, flow-based models can provide exact likelihood estimates unlike VAEs.

To account for the discreteness of molecular graphs, some flow-based methods for molecular generation rely on noise to convert discrete data into continuous data such as *GraphAF* [24] that add small stochastic noise during encoding, and subsequently resort to a dequantization step to mitigate the effect of noise. These methods are susceptible to some other issues, e.g., they often generate invalid structures that violate chemical valency rules [27,20]. Therefore, a post-hoc step is required that checks for the validity of the generated structures (i.e., graphs) and discards those that violate the valency of any of the nodes (i.e., atoms).

Some of the issues associated with the above models can be averted with discrete flows. Specifically, a method called *GraphDF* [27] employs discrete latent variables and performs validity checks as part of the generative process. Both GraphAF and GraphDF are autoregressive procedures that incrementally yield atoms (nodes) and bonds (edges) over several steps. Another prominent method MoFlow [30] avoids autoregressive generation by combining an unconditional flow over bonds with a conditional flow over atoms given bonds, and applying post-hoc validity correction.

Diffusion-based models have also attracted considerable traction [35,32]. Unlike flows, diffusion processes obviate the need for invertible transformations. Specifically, these models incorporate two phases: in the forward pass samples from the data are treated with stochastic noise in a Markov chain setting over several steps so that eventually these samples are distributed according to a Gaussian distribution. Notably, the forward pass does not consist of any trainable parameters. In the backward pass, the objective is to map samples from the Gaussian distribution to samples resembling data using a sequence of *denoising* steps that can all be carried out using deep networks. In particular, a method based on diffusion called *EDM* [32] operates simultaneously on categorical atom types and continuous atom coordinates, and generates 3D molecules using a denoising network that is equivariant to Euclidean transformations. One major limitation of diffusion processes, however, is the enormous computational effort and time required to train the model as well as to sample from it.

Neural differential equations (neural ODEs, PDEs, and SDEs)

Differential equations have long been an important tool for modeling various physical processes and biochemical phenomena. Their adroit integration into deep neural networks [36] has fostered exciting developments in generative molecular modeling. Neural ODEs derive inspiration from residual neural networks, and typically implement continuous flows using a continuum of layers that are indexed by time unlike standard neural networks.

Graphs naturally give rise to neural PDEs: the embedding of each node evolves with time based on the instantaneous embeddings of its neighbors and its own embedding; this system of coupled ODEs pertains to a PDE. Some important models in this context include Ffjord [37], PDE-GCN [25], GRAND [28], and ModFlow [20]. For instance, ModFlow [20] associates a continuous normalizing flow with each node. These flows start independently but then repeatedly interact with each other, according to the underlying PDE, toward jointly aligned distributions that can be sampled to yield molecules. Interestingly, ModFlow shares connections with temporal graph networks [38,39].

SDE-based models have also been successfully applied, and they often form a bridge between diffusion models and the so-called score-based generative models [40,41].

Reinforcement learning (RL)-inspired methods

RL has also inspired new models recently. Prominent among these is a class of models called GFlowNets [26,42,43,29,31]. Specifically, GFlowNets train a stochastic policy or generative process for discrete objects, such as molecular graphs, as a flow network. Specifically, these models consist of two kinds of states:

terminal states that pertain to objects of interest (i.e., molecules in our case) and other incomplete, intermediate states. A reward is associated with each terminal state, and any such state is sampled proportional to its reward through a series of constructive steps (such as adding a new atom and bond to the existing structure) (Table 1, Table 2).

Structure-based drug design (SBDD)

Unlike almost all the methods discussed above, SBDD models have access to the protein target [55]. The goal of SBDD is to generate suggestions for molecules that exhibit desirable physicochemical properties besides having good binding affinity with a specified protein target. Though the unconditional generative models far outnumber SBDD models, there have been some notable developments for SBDD lately [56]. Among these are models based on autoregressive models [2,57], variational autoencoders [45], reinforcement learning [48], genetic algorithms [51], and diffusion models [53]. Interestingly, these developments in SBDD have been accompanied by exciting advances in structure-based (conditional) protein design (and binding) with models like Structured Transformers [10], TANKBind [44], Equibind [47], DiffDock [50], RFDiffusion [46], Chroma [49], ProtSeed [52], and AbODE [54]. In fact, AbODE [54] reveals connections between models for molecule design, protein design, and docking, suggesting that similar generative modeling techniques could be broadly applicable across these tasks.

Property-based molecular optimization

Generative models for drug discovery can be optimized to search for molecules with better chemical properties such as QED. Essentially, one can encode an input (generated) molecule into the learned latent space of an already-trained generative model and interpolate in this space along a direction that locally improves the property of interest, typically, via several gradient steps. Finally, a decoding step can be performed to map back the eventual latent representation into a new molecule. We refer the reader to Ref. [20] for details.

Table 1

Graph-based methods at a glance. Non-generative methods are marked by **.

Graph-based methods that do not model interactions with the target		
Flow/Diffusion	Neural Differential Equations	RL-inspired
GraphAF [24]	PDE-GCN** [25]	GFlowNets [26]
GraphDF [27]	GRAND** [28]	QM-guided [29]
MoFlow [30]	ModFlow [20]	QADD [31]
EDM [32]		Sculpting [33]
SID [34]		

Table 2

SBDD methods at a glance.

Recent structure-based drug design (SBDD) methods		
Molecule design	Protein/Antibody design	Binding
Pocket2Mol [2]	Structured Transformers [10]	TANKBind [44]
VAE-based [45]	RFDiffusion [46]	Equibind [47]
DeepLigBuilder [48]	Chroma [49]	Diffdock [50]
RGA [51]	ProtSeed [52]	
DiffSBDD [53]	AbODE [54]	

Constrained optimization

Note that property-based molecular optimization does not ensure that the output (i.e., the modified) molecule will be similar to the input molecule. Constrained optimization aims to improve a specified property while striving to keep the similarity between the input and output molecules above some threshold. Reinforcement learning approaches such as Proximal Policy Optimization (PPO) have been employed to fine-tune pretrained generative models for this purpose [27,34].

Other RL-based techniques have also been successfully employed in the literature for drug design and lead optimization. For instance, DeepLigBuilder [48] leveraged Monte Carlo tree search (MCTS) to suggest new drug-like compounds having similar binding features to those of known inhibitors for the main protease of SARS-CoV-2.

Conclusions and perspectives

We now outline below some limitations of the existing methodology, highlight emerging trends, and offer our perspective on some directions that we believe could propel molecular generative modeling over the next few years.

Generalization

In pursuit of enhanced expressivity or representational capacity, generative models for molecular design are becoming increasingly complex and likely overfit the training data. Statistical learning theory foundations state that increased complexity of the hypothesis space (i.e., the class of functions being learned) inhibits the generalizability of powerful machine learning models; i.e., their ability to do well outside the training set. The need to generalize beyond the training data cannot be overemphasized [14], and we recommend redirecting research efforts to devising generative models that strive for a tradeoff between expressivity and generalization.

Benchmarking

A prevailing trend in the field, especially being witnessed in leading machine learning conferences, is to validate new models with experiments on standard data

such as QM9 and Zinc-250k. Many recent methods already report strong performance on these data; so focusing on more challenging benchmarks such as METLIN [58] would benefit the community.

It must also be emphasized that all these datasets are extremely small, and as such, are unable to capture the immense diversity of the molecular search space. In our opinion design of scalable machine learning models that can effectively process, and learn powerful representations with, much larger datasets should be one of the priorities of our community.

Inductive bias

Equivariance and invariance are strong inductive biases. Similarly, the success of GNNs in representing molecular graphs and Transformers in encoding SMILES sequences underscores the significance of incorporating appropriate inductive bias in the model. GNNs, by design, tend to capture the local correlations whereas Transformers attend to a more global context. The sweet spot for molecular design probably lies somewhere between these two extremes and exploring alternative schemes such as leveraging random walks to learn a good representation [39,59] could be fruitful.

Interpretability

We still have little understanding about the structure of the latent space of generative models for molecules. Interpretable models can illuminate the relative importance of the learned latent subspaces towards model outcomes. Disentanglement, i.e., unraveling the complex generative factors of variation in the data, is one specific formalism for interpretability that has shown some promise in the context of molecular generation [34]. Disentanglement is known to have benefits from both generalization and sample complexity perspectives. In particular, it enables learning from fewer samples compared to models that do not learn disengaged representations.

The main idea underlying disentanglement is to learn to isolate essential information for each latent factor in few dimensions, disparate from the other factors. However, factors inherent in molecular data are often complex, and likely cannot be fully segregated from each other. Therefore, we advocate design of more flexible models such as the conditional method from Ref. [34] that could allow for partial disentanglement of molecular latent spaces.

Evaluation

Once trained, generative models can yield an extremely large number of suggestions for molecules. However, from a practical perspective, generating one promising novel candidate is way more important than suggesting several that do not hold much potential. Many recent works report results on metrics such as validity, stability, reconstruction, uniqueness, and novelty. We emphasize the need for comprehensive evaluation with more rigorous criteria including, but not limited to, molecular weight, octanol-water partition coefficient, synthetic accessibility, quantitative estimation of drug-likeness, and the MOSES metrics such as Fragment Similarity, Nearest Neighbor Similarity, and Internal Diversity.

Topological descriptors

Representational limits of GNNs have recently inspired design of novel graph representation methods that can learn multi-scale and long-range topological features, opening exciting possibilities for molecular datasets that are known to have a notable topological structure [60]. Two rather parallel lines of work - Topological Deep Learning (TDL) and Persistent Homology (PH) - have gained prominence in this context. TDL methods seek to process *part-whole* and *set-types* relations to represent complex structures or interactions between different components in data (e.g., atoms in a ring) going beyond the usual pairwise relations paradigm of message-passing GNNs [61].

Persistent Homology (PH), a key tool from Topological Data Analysis (TDA), relies on (learnable) *filtration* functions. Specifically, PH seeks to use data samples to characterize topological invariants such as connected components of an underlying manifold, yielding global topological signatures that can be integrated into, and boost the performance of, graph neural networks [62–64]. With new insights into its theoretical underpinnings [60,64], PH is set to play an important role in generative molecular modeling.

Compositionality

As generative models become increasingly complex, concerns about high costs involved in training new models from scratch have recently motivated some works on model reuse and composition [65,33]. In principle, given multiple pretrained models with each capturing some particular property, one can compose or fuse together these models to obtain a composite model with enhanced capacity. Such composite models enable complex distributions that adhere to multiple constraints, and can be sampled to generate molecules that exhibit multiple properties thereby paving way for multi-objective molecular generation [66].

Over the next few years, we anticipate significant attention toward design of compositional techniques such as *Sculpting* [33], which can coordinate the steps of powerful iterative generative processes like diffusion models and GFlowNets.

Declaration of competing interest

The author is Chief Scientist at YaiYai Ltd., a company that collaborates with pharma industry, and develops solutions for different parts of the drug discovery pipeline.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. Stanley M, Segler M: **Fake it until you make it? Generative de novo design and virtual screening of synthesizable molecules**. *Curr Opin Struct Biol* 2023, **82**:102658. ISSN 0959-440X.

2. Peng X, Luo S, Guan J, Xie Q, Peng J, Ma J: **Pocket2Mol: efficient molecular sampling based on 3D protein pockets**. In *International conference on machine learning*; 2022.

3. Baillif B, Cole J, McCabe P, Bender A: **Deep generative models for 3D molecular structure**. *Curr Opin Struct Biol* 2023, **80**: 102566, https://doi.org/10.1016/j.sbi.2023.102566. ISSN 0959-440X.

4. Grisoni F: **Chemical language models for de novo drug design: challenges and opportunities**. *Curr Opin Struct Biol* 2023, **79**:102527, https://doi.org/10.1016/j.sbi.2023.102527. ISSN 0959-440X.

5. Meyers J, Fabian B, Brown N: **De novo molecular design and generative models**. *Drug Discov Today* 2021, **26**:2707–2715. ISSN 1359-6446.

6. Hanser T: **Federated learning for molecular discovery**. *Curr Opin Struct Biol* 2023, **79**:102545. ISSN 0959-440X.

7. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, Chandak P, Liu S, Van Katwyk P, Deac A, Anandkumar A, Bergen K, Gomes C, Ho S, Kohli P, Lasenby J, Leskovec J, Liu T, Manrai A, Marks D, Ramsundar B, Song L, Sun J, Tang J, Veličković P, Welling M, Zhang L, Coley C, Bengio Y, Zitnik M: **Scientific discovery in the age of artificial intelligence**. *Nature* 2023, **620**:47–60. ISSN 0028-0836.

8. Isert C, Atz K, Schneider G: **Structure-based drug design with geometric deep learning**. *Curr Opin Struct Biol* 2023, **79**: 102548. ISSN 0959-440X.

9. Thomas M, Bender A, de Graaf C: **Integrating structure-based approaches in generative molecular design**. *Curr Opin Struct Biol* 2023, **79**:102559, https://doi.org/10.1016/j.sbi.2023.102559. ISSN 0959-440X.

10. Ingraham J, Garg V, Barzilay R, Jaakkola T: **Generative models for graph-based protein design**. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R. *Advances in neural information processing systems*, **32**. Curran Associates, Inc.; 2019.

11. Y. Wang, H. Zhao, S. Sciabola, W. Wang, cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation, Molecules 28, ISSN 1420-3049, doi: 10.3390/molecules28114430, URL https://www.mdpi.com/1420-3049/28/11/4430.

12. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G: **The graph neural network model**. *IEEE Trans Neural Network* 2009, **20**:61–80, https://doi.org/10.1109/TNN.2008.2005605.

13. Xu K, Hu W, Leskovec J, Jegelka S: **How powerful are graph neural networks?**. In *International conference on learning representations*; 2019.

14. Garg V, Jegelka S, Jaakkola T: **Generalization and representational limits of graph neural networks**. ICML 2020, 13-18 July 2020, Virtual Event. In *Proceedings of the 37th international conference on machine learning. Of proceedings of machine learning research*, **119**. PMLR; 2020:3419–3430. URL, http://proceedings.mlr.press/v119/garg20c.html.

15. Chen Z, Chen L, Villar S, Bruna J: **Can graph neural networks count substructures?**. In Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H. *Advances in neural information processing systems*, **33**. Curran Associates, Inc.; 2020:10383–10395.

16. * Bronstein MM, Bruna J, Cohen T, Velickovic P: ***Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, CoRR abs/2104***. https://arxiv.org/abs/2104.13478.
A useful resource that brings together Graph Neural Networks, Transformers, and Convolutional Neural Networks under a unified theme of geometric deep learning

17. * Klicpera J, Becker F, Günnemann S: **GemNet: universal directional graph neural networks for molecules**. In *Advances in neural information processing systems*. Edited by Beygelzimer A, Dauphin Y, Liang P, Vaughan JW; 2021.
Established that GNNs with directed message passing are universal approximators for rotationally equivariant and translation invariant predictions

18. Liu Y, Wang L, Liu M, Lin Y, Zhang X, Oztekin B, Ji S: **Spherical message passing for 3D molecular graphs**. In *International conference on learning representations*; 2022.

19. Satorras VG, Hoogeboom E, Welling M: **E(n) equivariant graph neural networks**. In *Proceedings of the 38th international conference on machine learning*. Edited by Meila M, Zhang T, *Of Proceedings of machine learning research*, **139**. PMLR; 2021: 9323–9332.

20. Verma Y, Kaski S, Heinonen M, Garg V: **Modular flows: differential molecular generation**. In *Advances in neural information processing systems*; 2022.

21. * Gemnet CK, Bodnar C, Mathis SV, Cohen T, Liò P: **On the expressive power of geometric graph neural networks**. In *International conference on machine learning*; 2023.
Introduced geometric WL and analyzed the relative power of equivariant and invariant layers such as those that are routinely used in generative models for molecules

22. Rezende D, Mohamed S: **Variational inference with normalizing flows**. In *Proceedings of the 32nd international conference on machine learning*. Edited by Bach F, Blei D, *Of Proceedings of machine learning research*, **37**. Lille, France: PMLR; 2015: 1530–1538.

23. W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, arXiv preprint arXiv: 1802.04364 .

24. Shi* C, Xu* M, Zhu Z, Zhang W, Zhang M, Tang J: **GraphAF: a flow-based autoregressive model for molecular graph generation**. In *International conference on learning representations*; 2020.

25. * M. Eliasof, E. Haber, E. Treister, Pde-gcn: novel architectures for graph neural networks motivated by partial differential equations, Adv Neural Inf Process Syst 34.
Introduced PDE-GCN to model several graph problems as discretized PDEs, and constructed deep graph convolutional networks without oversmoothing

26. * Bengio E, Jain M, Korablyov M, Precup D, Bengio Y: **Flow network based generative models for non-iterative diverse candidate generation**. In *Advances in neural information processing systems*. Edited by Beygelzimer A, Dauphin Y, Liang P, Vaughan JW; 2021. URL:https://openreview.net/forum?id=Arn2E4IRjEB.
Introduced GFlowNets, a novel class of iterative generative models for discrete objects such as molecules; can be extended to accommodate multiple objectives

27. Luo Y, Yan K, Ji S: **GraphDF: a discrete flow model for molecular graph generation**. ICML 2021, 18-24 July 2021, Virtual Event. In *Proceedings of the 38th international conference on machine learning*. Edited by Meila M, Zhang T, *Of Proceedings of machine learning research*, **139**. PMLR; 2021:7192–7203.

28. * Chamberlain B, Rowbottom J, Gorinova MI, Bronstein M, Webb S, Rossi E: **Grand: graph neural diffusion**. In *International conference on machine learning*. PMLR; 2021:1407–1418.

Developed PDE tools to analyze existing GNN architectures and propose new ones

29. Simm G, Pinsler R, Hernandez-Lobato JM: **Reinforcement learning for molecular design guided by quantum mechanics**. In *Proceedings of the 37th international conference on machine learning*. Edited by Singh A, *Of Proceedings of machine learning research*, **119**. PMLR; 2020:8959–8969.

30. Zang C, Wang F: **MoFlow: an invertible flow model for generating molecular graphs**. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*; 2020:617–626.

31. Fang Y, Pan X, Shen H-B: **De novo drug design by iterative multiobjective deep reinforcement learning with graph-based molecular quality assessment**. *Bioinformatics* 2023, **39**: 1367–4811. btad157, ISSN.

32. * Hoogeboom E, Satorras VG, Vignac C, Welling M: **Equivariant diffusion for molecule generation in 3D**. ICML 2022, 17-23 July 2022. In *International conference on machine learning*. Edited by Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S, *Of Proceedings of machine learning research*, **162**. Baltimore, Maryland, USA: PMLR; 2022:8867–8887. URL:https://proceedings.mlr.press/v162/hoogeboom22a.html.
Introduced an E(3)-equivariant diffusion model, i.e., a model equivariant to rotations, translations and reflections, for molecule generation

33. * Garipov T, Peuter SD, Yang G, Garg V, Kaski S, Jaakkola T: **Compositional sculpting of iterative generative processes**. In *Advances in neural information processing systems*; 2023.
Introduced a new technique 'Sculpting' for composing pretrained iterative generative models such as GFlowNets and Diffusion models paving way for multiobjective molecular generation

34. * Mercatali G, Freitas A, Garg V: **Symmetry-induced disentanglement on graphs**. In *Advances in neural information processing systems*. Edited by Oh AH, Agarwal A, Belgrave D, Cho K; 2022.
Introduced and formalized a new notion of conditional disentanglement for interpretable representations and showed its benefits for molecular generation

35. Ho J, Jain A, Abbeel P: **Denoising diffusion probabilistic models**. In Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H. *Advances in neural information processing systems*, **33**. Curran Associates, Inc.; 2020:6840–6851.

36. R. T. Chen, Y. Rubanova, J. Bettencourt, D. K. Duvenaud, Neural ordinary differential equations, Adv Neural Inf Process Syst 31.

37. Grathwohl W, Chen RTQ, Bettencourt J, Duvenaud D: **Scalable reversible generative models with free-form continuous dynamics**. In *International conference on learning representations*; 2019.

38. Rossi E, Chamberlain B, Frasca F, Eynard D, Monti F, Bronstein M: **Temporal graph networks for deep learning on dynamic graphs**. In *ICML 2020 workshop on graph representation learning*; 2020.

39. Souza AH, Mesquita D, Kaski S, Garg V: **Provably expressive temporal graph networks**. In *Advances in neural information processing systems*. Edited by Oh AH, Agarwal A, Belgrave D, Cho K; 2022.

40. Bao F, Zhao M, Hao Z, Li P, Li C, Zhu J: **Equivariant energy-guided SDE for inverse molecular design**. In *The eleventh international conference on learning representations*; 2023.

41. Jo J, Lee S, Hwang SJ: **Score-based generative modeling of graphs via the system of stochastic differential equations**. In *Proceedings of the 39th international conference on machine learning*. Edited by Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, *Of Proceedings of machine learning research*, **162**. PMLR; 2022:10362–10383.

42. Bengio Y, Lahlou S, Deleu T, Hu EJ, Tiwari M, Bengio E: **GFlowNet foundations**. *J Mach Learn Res* 2023, **24**:1–55.

43. Jain M, Raparthy SC, Hernandez-Garcia A, Rector-Brooks J, Bengio Y, Miret S, Bengio E: **Multi-objective GFlowNets**. In *International conference on machine learning*. PMLR; 2023: 14631–14653.

44. Lu W, Wu Q, Zhang J, Rao J, Li C, Zheng S: **TANKBind: trigonometry-aware neural NetworKs for drug-protein binding structure prediction**. In *Advances in neural information processing systems*. Edited by Oh AH, Agarwal A, Belgrave D, Cho K; 2022.

45. M. Ragoza, T. Masuda, D. Koes, Generating 3D molecules conditional on receptor binding sites with deep generative models, Chem Sci 13.

46. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D: **De novo design of protein structure and function with RFdiffusion**. *Nature* 2023:1476–4687, https://doi.org/10.1038/s41586-023-06415-8. 1–3ISSN.

47. * Stärk H, Ganea O-E, Pattanaik L, Barzilay R, Jaakkola T: **EquiBind: geometric deep learning for drug binding structure prediction**. In *International conference on machine learning*; 2022.
An important work concerning SBDD that extended graph based generative modeling to conditional settings using geometric deep learning

48. Li Y, Pei J, Lai L: **Structure-based de novo drug design using 3D deep generative models**. *Chem Sci* 2021, **12**:13664–13675.

49. J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. a. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk, G. Grigoryan, Illuminating protein space with a programmable generative model, Nature doi:10.1038/s41586-023-06728-8.

50. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola TS: **DiffDock: diffusion steps, twists, and turns for molecular docking**. In *The eleventh international conference on learning representations*; 2023.

51. Fu T, Gao W, Coley C, Sun J: **Reinforced genetic algorithm for structure-based drug design**. In Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A. *Advances in neural information processing systems*, **35**. Curran Associates, Inc.; 2022: 12325–12338.

52. Shi C, Wang C, Lu J, Zhong B, Tang J: **Protein sequence and structure Co-design with equivariant translation**. In *The eleventh international conference on learning representations*; 2023.

53. A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes, M. Welling, M. Bronstein, B. Correia, Structure-based drug design with equivariant diffusion models, [arXiv] .

54. * Verma Y, Heinonen M, Garg V: **AbODE: Ab initio antibody design using conjoined ODEs**. 23-29 July 2023. In *International conference on machine learning, ICML 2023. Of Proceedings of machine learning research*, **202**. Honolulu, Hawaii, USA: PMLR; 2023:35037–35050. URL:, https://proceedings.mlr.press/v202/verma23a.html.
Proposed a neural PDE method AbODE for antigen-conditioned antibody design, showing connections between docking, protein design, and molecule design.

55. Guan J, Qian WW, Peng X, Su Y, Peng J, Ma J: **3D equivariant diffusion for target-aware molecule generation and affinity prediction**. In *The eleventh international conference on learning representations*; 2023.

56. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF: **Generative models for molecular discovery: recent advances and challenges**. *Wiley Interdiscip Rev Comput Mol Sci* 2022, **12**, e1608.

57. Luo S, Guan J, Ma J, Peng J: **A 3D generative model for structure-based drug design**. In *Advances in neural information processing systems*; 2021.

58. Domingo-Almenara X, Guijas C, Billings E, Montenegro-Burke JR, Uritboonthai W, Aisporna AE, Chen E, Benton HP,

Siuzdak G: **The METLIN small molecule dataset for machine learning-based retention time prediction**. *Nat Commun* 2019, **10**:5811, https://doi.org/10.1038/s41467-019-13680-7. ISSN 2041-1723.

59. Wang Y, Chang Y-Y, Liu Y, Leskovec J, Li P: **Inductive representation learning in temporal networks via causal anonymous walks**. In *International conference on learning representations*; 2021.

60. B. Rieck, On the expressivity of persistent homology in graph learning, arXiv: 2302.09826 .

61. Papillon M, Sanborn S, Hajij M, Miolane N: *Architectures of topological deep learning: A Survey on Topological Neural Networks*. 2023.

62. Carriere M, Chazal F, Ike Y, Lacombe T, Royer M, Umeda Y: **PersLay: a neural network layer for persistence diagrams and new graph topological signatures**. In *Proceedings of the twenty third international conference on artificial intelligence and statistics*. Edited by Chiappa S, Calandra R, *Of Proceedings of machine learning research*, **108**. PMLR; 2020:2786–2796.

63. Horn M, De Brouwer E, Moor M, Moreau Y, Rieck B, Borgwardt K: **Topological graph neural networks**. In *International conference on learning representations*. ICLR; 2022.

64. * Immonen J, Souza A, Garg V: **Going beyond persistence homology using persistence homology**. In *Advances in neural information processing systems*; 2023.
Formalized the expressivity of persistent homology methods, and introduced more powerful models for extracting topological signatures that boosted the performance of graph neural networks on a wide range of real-world datasets

65. Du Y, Durkan C, Strudel R, Joshua SD, Tenenbaum B, Fergus R, Sohl-Dickstein J, Doucet A, Grathwohl W: **Reduce, reuse, recycle: compositional generation with energy-based diffusion models and MCMC**. In *International conference on machine learning*. ICML; 2023.

66. Luukkonen S, van den Maagdenberg HW, Emmerich MT, van Westen GJ: **Artificial intelligence in multi-objective drug design**. ISSN 0959-440X *Curr Opin Struct Biol* 2023, **79**:102537, https://doi.org/10.1016/j.sbi.2023.102537. https://www.sciencedirect.com/science/article/pii/S0959440X23000118.