Flow-Based Models for Active Molecular Graph Generation

Nathan C. Frey
University of Pennsylvania
n.frey@seas.upenn.edu

Bharath Ramsundar
DeepChem
bharath.ramsundar@gmail.com

Abstract

We propose a framework using normalizing-flow based models, SELF-Referencing Embedded Strings, and active learning that generates a high percentage of novel, unique, and valid molecules and efficiently identifies optimal generated samples. With an initial training set of only 200 small molecules in the QM9 dataset, 78% of the generated samples are chemically valid, not present in the training data, and unique. We define a "dissimilarity druglikeness" metric to quantify both the novelty and druglikeness of generated samples. With active learning the maximally novel, druglike generated samples are identified with an order of magnitude fewer steps than random search. Our model is significantly simpler and easier to train than other molecular generative models, and enables fast generation and identification of novel druglike molecules.

1 Introduction

The goal of generative modeling of small molecules is to discover structurally novel molecules with optimal chemical properties. Prior work using variational autoencoders (VAEs) [Gómez-Bombarelli et al. [2018]], generative adversarial networks (GANs) [De Cao and Kipf [2018]], and reinforcement learning [You et al. [2018]] has shown the promise of generative modeling in the chemical sciences. Generative models are commonly evaluated according to the percentage of valid, novel, and unique molecules they produce [Samanta et al. [2020], Brown et al. [2019]]. In the space of 10^{60} druglike compounds [Mullard [2017]], the yet-to-be-proved utility of deep generative modeling lies in finding useful molecular scaffolds and compounds that are not already well-known to medicinal chemists [Bush et al. [2020]].

Recently, normalizing flows (NFs) [Papamakarios et al. [2019]] have emerged as a promising model architecture for molecular graph generation [Zang and Wang [2020], Shi et al. [2020], Madhawa et al. [2019], Honda et al. [2019]]. Unlike the frameworks discussed above, NFs do not rely on a compressed latent space representation for generative modeling. Instead, an NF learns an invertible mapping between a simple base distribution and a target distribution. Previous work applying NFs to molecule generation [Zang and Wang [2020]] has shown that NFs with post-hoc corrections to enforce chemical validity achieve high validity, novelty, and uniqueness scores on benchmark datasets like QM9 [Ramakrishnan et al. [2014]] and ZINC250K [Sterling and Irwin [2015]].

In this paper, we present a greatly simplified NF architecture that, despite its simplicity and without post-hoc corrections, generates 78% valid, novel, and unique molecules when trained on only 200 samples from QM9. By defining a "dissimilarity druglikeness" metric and feeding the generated outputs to an active learning framework, we rapidly identify promising outliers that are both chemically dissimilar from the training data and druglike. A high percentage of valid outputs is ensured by using the 100% robust SELF-referencing Embedded Strings (SELFIES) representation [Krenn et al. [2020]]. This work presents an effective scheme for discovering novel druglike compounds, and an accessible means for researchers to do robust generative molecular modeling.

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

2 Active Flow-Based Generative Models

To circumvent a common problem with molecular generative models - invalid outputs due to chemical rules not being encoded in the model architecture - we first encode the QM9 dataset as SELFIES strings. The SELFIES grammar and bond constraints enforce chemical valency rules, guaranteeing that generated SELFIES are syntactically and semantically valid, without requiring post-hoc corrections or complex model architectures that are difficult to train.

SELFIES strings are then one-hot encoded and dequantized [Dinh et al. [2016]] by adding random noise from the interval $[0, 1)$ to each element. The original inputs can be recovered by applying a floor function, and the continuous dequantized inputs are used to train the model.

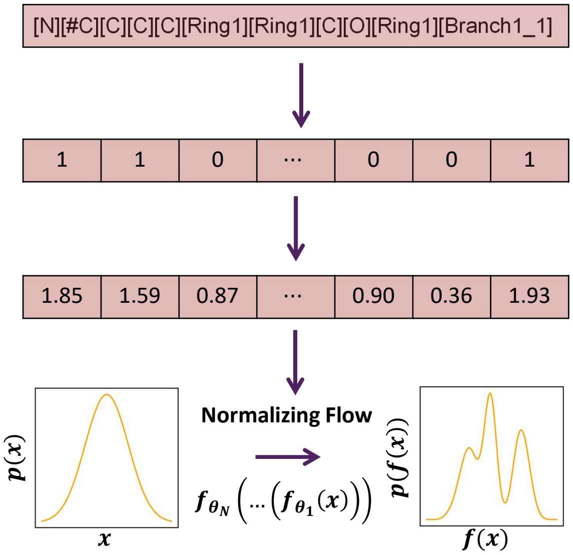
[N][#C][C][C][C][Ring1][Ring1][C][O][Ring1][Branch1_1]
1 | 1 | 0 | ··· | 0 | 0 | 1
1.85 | 1.59 | 0.87 | ··· | 0.90 | 0.36 | 1.93
Normalizing Flow
$f_{\theta_N}\left(\cdots\left(f_{\theta_1}(x)\right)\right)$
$p(x)$
x
$p(f(x))$
$f(x)$

Figure 1: Dequantized one-hot encodings of SELFIES representations are inputs to the normalizing flow. The normalizing flow maps a simple base distribution to a complex target distribution.

We use a normalizing flow (NF) to model the distribution of molecules in QM9. Because NFs are composed of bijective transformations, they provide an easily interpretable one-to-one mapping between inputs and outputs, without lossy compression to a latent space representation. NFs offer both generative sampling and exact likelihood calculation, unlike VAEs which provide only a lower-bound on log-likelihood and GANs, which do not provide likelihood estimation. Here, the NF provides a simple and straightforward means of generating new molecules. A schematic of the data pre-processing and mapping between the base and target distributions is shown in Fig. 1. The NF is comprised of eight Masked Autoregressive Flow (MAF) [Papamakarios et al. [2017]] layers, each using a Masked Autoencoder for Distribution Estimation (MADE) [Germain et al. [2015]] autoregressive network with 512, 512 hidden dimensions as the scale-and-shift operation.

The active molecular graph generation workflow is shown schematically in Fig. 2. The generative model is trained on an initial subset of training data. Samples are then generated and queried according to the active learning scheme to identify the optimal candidates by whatever metric is of interest, e.g. docking affinity to some target protein. Active learning uses a simple surrogate model like a random forest or a linear model to estimate the target property for the generated molecules. Optimal candidates are selected by the surrogate model, and these candidates are used to augment the training set. The NF is retrained and new samples are generated, with the entire procedure repeated for multiple iterations to yield generated molecules with designer properties.

3 Experiments

As a proxy target for active learning, we define the *Dissimilarity Druglikeness* (DDL) metric. Starting with the Tanimoto coefficient (a typical metric for comparing molecular fingerprints), each molecule is

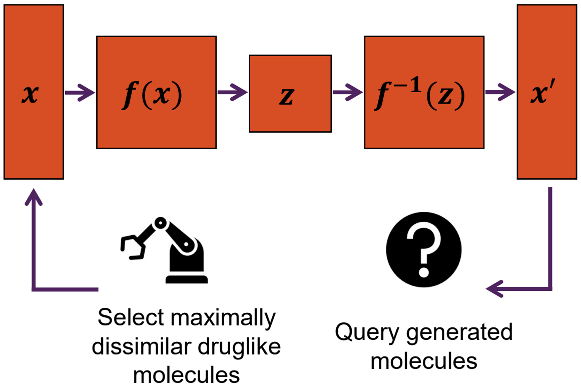
x
$f(x)$
z
$f^{-1}(z)$
x'
Select maximally dissimilar druglike molecules
Query generated molecules

Figure 2: Active molecular graph generation with normalizing flows.

assigned a "novelty score" equal to $1 - \max(\{T_C\})$, where $\{T_C\}$ is the set of all Tanimoto coefficients comparing that molecule to the training data. Druglikeness is represented by the quantitative estimate of druglikeness defined in Bickerton et al. [2012],

$$\text{QED} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln d_i\right), \tag{1}$$

where d_i are desirability functions corresponding to molecular descriptors. A scatter plot of these two metrics is shown for a subset of 5000 compounds in QM9 in Fig. 3a. We construct the joint metric

$$\text{DDL} = (1 - \max(\{T_C\})) * \text{QED}, \tag{2}$$

such that maximizing DDL corresponds to maximally druglike molecules that are also maximally dissimilar from all other molecules in the training data. A random forest is used as the surrogate model in active learning to identify generated molecules with the highest DDL score, using molecular fingerprints computed with the *RDKit* cheminformatics software. Fig. 3b is a histogram of DDL scores for the random subset of 5000 compounds from QM9. Fig. 3c shows the training and validation loss curves for training the NF on a training set of 200 compounds from QM9 (with 50 compounds in the validation set). Fig. 3d shows the results of active learning to efficiently identify the maximum DDL molecule in the first batch of 100 generated samples. For each batch of generated samples, the molecules that are valid, unique, and novel are selected. Then active learning proceeds with the surrogate model trained on only three randomly selected samples. The remaining samples are queried at random, or by using the surrogate model to choose the sample with the highest predicted DDL (expected max) or highest uncertainty in the predicted DDL (max uncertainty). The mean, median, and standard deviation of the number of queries required to identify the max DDL molecule from the generated samples with active learning (random search) are 3 (29), 5 (43), and 3.7 (16.5) over five iterations on the first set of generated molecules. This simple experiment suggests that the use of active learning could considerably speed up lead optimization efforts in real drug discovery settings, where "queries" would correspond to experimental tests.

The percentage of valid, unique, and novel generated molecules for five iterations of active molecular graph generation are given in Table 1. Our model is simple, easy to train, and performs better than or comparably to state-of-the-art flow-based generative models, with the exception of MoFlow [Zang and Wang [2020]], which has 10 coupling layers and 27 graph coupling layers and a post-hoc validity correction, while our model has only eight layers and no post-hoc processing.

4 Discussion

Our next steps include extending this work to larger, more clinically relevant datasets like ZINC15, with more expressive NF architectures. We will deploy the active molecular generation framework on target properties that are not feasible to evaluate for all generated samples, like ligand-protein binding affinities obtained through free energy perturbation calculations. This framework will then be used in collaboration with experimental groups to synthesize and validate the final generated

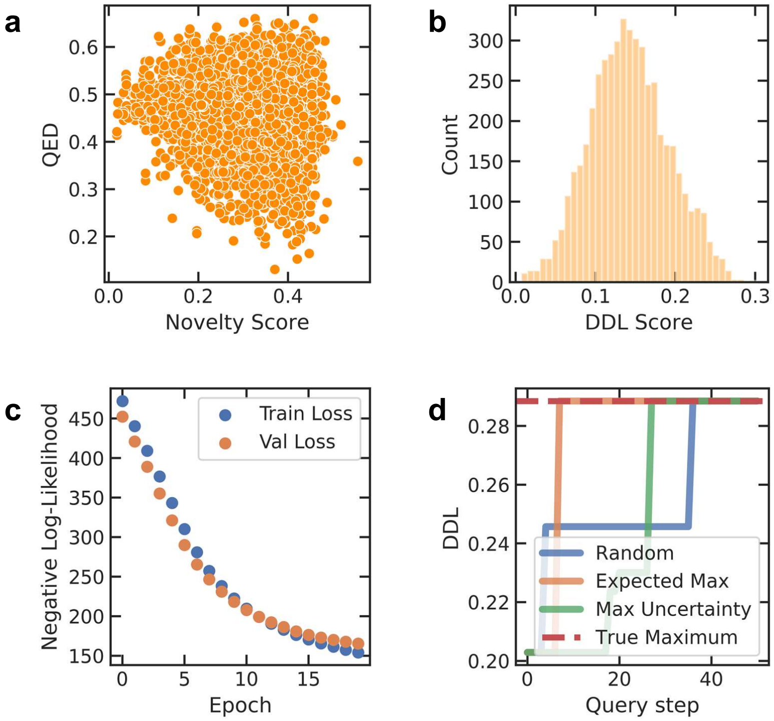

Figure 3: Active molecular graph generation for optimizing dissimilar druglikeness. A training and validation set of 200 and 50 compounds respectively were selected from the QM9 dataset. **a** Scatter plot of quantitative estimate of druglikeness versus dissimilarity. **b** Distribution of dissimilar druglikeness score in a random subset of QM9. **c** Loss curves for training and validation sets for the normalizing flow. **d** Active learning iterations to find optimal DDL molecule. The true maximum is indicated with a red dashed line. Random, expected max, and max uncertainty query strategies are shown in blue, orange, and green respectively.

Table 1: Active molecular graph generation performance on QM9.

Architecture	% Valid, Unique, and Novel Molecules	Reference
GraphNVP	47.97	Madhawa et al. [2019]
GRF	32.68	Honda et al. [2019]
GraphAF	83.95	Shi et al. [2020]
MoFlow	97.24 ± 0.21	Zang and Wang [2020]
This work	77.81 ± 0.06	

candidates. In addition to developing the active molecular generation scheme with normalizing flows, we also intend for this work to provide a simple and accessible means for researchers to do robust generative molecular modeling. Importantly, the complete code base used for deep generative modeling in this work has been made available in the *DeepChem* package [Ramsundar et al. [2019]] at https://github.com/deepchem/deepchem.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests.

References

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, feb 2012. ISSN

17554330. doi: 10.1038/nchem.1243. URL www.nature.com/naturechemistry.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, mar 2019. ISSN 15205142. doi: 10.1021/acs.jcim.8b00839.

Jacob T Bush, Peter Pogány, Stephen D. Pickett, mike Barker, andrew Baxter, Sebastien Campos, Anthony W.J. Cooper, David Jonathan Hirst, Graham Inglis, alan Nadin, Vipulkumar K. Patel, darren Poole, John Pritchard, Yoshiaki Washio, Gemma White, and Darren Green. A Turing test for molecular generators. *Journal of Medicinal Chemistry*, page acs.jmedchem.0c01148, 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.0c01148. URL https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c01148.

Nicola De Cao and Thomas Kipf. MolGAN: An implicit generative model for small molecular graphs. 2018. URL http://arxiv.org/abs/1805.11973.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, may 2016. URL http://arxiv.org/abs/1605.08803.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked Autoencoder for Distribution Estimation. *32nd International Conference on Machine Learning, ICML 2015*, 2: 881–889, feb 2015. URL http://arxiv.org/abs/1502.03509.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, feb 2018. ISSN 23747951. doi: 10.1021/acscentsci.7b00572.

Shion Honda, Hirotaka Akita, Katsuhiko Ishiguro, Toshiki Nakanishi, and Kenta Oono. Graph Residual Flow for Molecular Graph Generation. 2019. URL http://arxiv.org/abs/1909.13521.

Mario Krenn, Florian Hase, Akshatkumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 2020. doi: 10.1088/2632-2153/aba947.

Kaushalya Madhawa, Katushiko Ishiguro, Kosuke Nakago, and Motoki Abe. GraphNVP: An Invertible Flow Model for Generating Molecular Graphs. may 2019. URL http://arxiv.org/abs/1905.11600.

Asher Mullard. The drug-maker's guide to the galaxy. *Nature*, 549(7673):445–447, sep 2017. ISSN 14764687. doi: 10.1038/549445a. URL http://www.nature.com/news/the-drug-maker-s-guide-to-the-galaxy-1.22683.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. *Advances in Neural Information Processing Systems*, 2017-December:2339–2348, may 2017. URL http://arxiv.org/abs/1705.07057.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. 2019. URL http://arxiv.org/abs/1912.02762.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7, aug 2014. ISSN 20524463. doi: 10.1038/sdata.2014.22. URL www.nature.com/sdata/.

Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Drug Discovery, and More*. 2019. ISBN 9788578110796. doi: 10.1017/CBO9781107415324.004. URL https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

Bidisha Samanta, Abir De, Gourhari Jana, Vicenc Gomez, Pratim Kumar Chattaraj, Niloy Ganguly, and Manuel Gomez-Rodriguez. NEVAE: A deep generative model for molecular graphs. *Journal of Machine Learning Research*, 21(i):1–17, 2020. ISSN 15337928. URL `http://arxiv.org/abs/1802.05283`.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. 2020. URL `https://github.com/DeepGraphLearning/GraphAFhttp://arxiv.org/abs/2001.09382`.

Teague Sterling and John J. Irwin. ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, nov 2015. ISSN 15205142. doi: 10.1021/acs.jcim.5b00559. URL `https://clinicaltrials.gov`.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 6410–6421, 2018.

Chengxi Zang and Fei Wang. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 617–626, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3403104.