Target-aware Molecular Graph Generation

Cheng Tan*, Zhangyang Gao*, Stan Z. Li ✉

Zhejiang University
AI Lab, Research Center for Industries of the Future, Westlake University
{tancheng, gaozhangyang, stan.zq.li}@westlake.edu.cn

Abstract. Generating molecules with desired biological activities has attracted growing attention in drug discovery. Previous molecular generation models are designed as chemocentric methods that hardly consider the drug-target interaction, limiting their practical applications. In this paper, we aim to generate molecular drugs in a target-aware manner that bridges biological activity and molecular design. To solve this problem, we compile a benchmark dataset from several publicly available datasets and build baselines in a unified framework. Building on the recent advantages of flow-based molecular generation models, we propose *SiamFlow*, which forces the flow to fit the distribution of target sequence embeddings in latent space. Specifically, we employ an alignment loss and a uniform loss to bring target sequence embeddings and drug graph embeddings into agreements while avoiding collapse. Furthermore, we formulate the alignment into a one-to-many problem by learning spaces of target sequence embeddings. Experiments quantitatively show that our proposed method learns meaningful representations in the latent space toward the target-aware molecular graph generation and provides an alternative approach to bridge biology and chemistry in drug discovery.

Keywords: AI for Science · Bioinformatics · Molecular Generation · Graph Neural Networks.

1 Introduction

Drug discovery, which focuses on finding candidate molecules with desirable properties for therapeutic applications, is a long-period and expensive process with a high failure rate. The challenge primarily stems from the actuality that only a tiny fraction of the theoretical possible drug-like molecules may have practical effects. Specifically, the entire search space is as large as $10^{23} \sim 10^{60}$, while only 10^8 of them are therapeutically relevant [45]. In the face of such difficulty, traditional methods like high-throughput screening [19] fail in terms of efficiency because of the large number of resources required in producing minor hit compounds. One alternative is using computational methods [44] such as virtual screening [51] to identify hit compounds from virtual libraries through similarity-based searches or molecular docking. Another alternative is automated molecule design, such as inverse QSAR [53], structure-based de novo design [52], or genetic algorithms [3].

* Equal Contribution

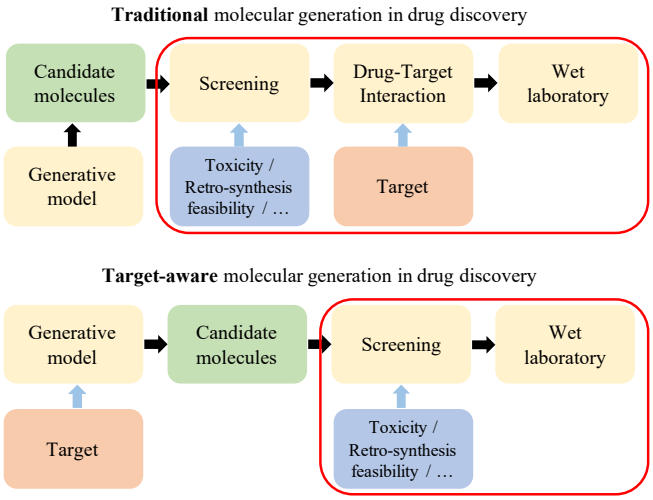

Fig. 1. The computational drug discovery pipelines of traditional chemocentric and target-aware molecular generation. The black arrows denote the main steps, the blue arrows denote external considerations, and the red boxes denote the post-processing process of generated molecules.

Recent deep generative models have demonstrated the potential to promote drug discovery by exploring huge chemical space in a data-driven manner. Various forms of variational autoencoder (VAE) [55], generative adversarial networks (GAN) [47], autoregressive (AR) [58, 46, 62], and normalizing flow (NF) [11, 12, 40, 54, 39] have been proposed to generate molecular SMILES or graphs. Though these approaches can generate valid and novel molecules to some extent, they remain inefficient because the generated candidate molecules need further screened against given targets. As the primary goal of these chemocentric methods is to generate drug-like molecules that satisfy specific properties, directly applying them in drug discovery requires extra effort in predicting the binding affinities between candidate molecules and target proteins.

While previous molecular generation methods scarcely take biological drug-target interactions into account, we aim to generate candidate molecules based on a biological perspective. This paper proposes target-aware molecular generation to bridge biological activity and chemical molecular design that generate valid molecules conditioned on specific targets and thus facilitate the development of drug discovery. As shown in Fig. 1, the pipeline of computational drug discovery is supposed to be simplified to a great extent with the help of target-aware molecular generation. Our main contributions are summarized as follows:

– We propose a target-aware molecular generation manner from a biological perspective, while prior works on chemocentric molecular generation are inefficient in practical drug discovery.

– We establish a new benchmark for the target-aware molecular generation containing abundant drug-target pairs for evaluating generative models.
– We propose SiamFlow, a siamese network architecture for the conditional generation of flow-based models. While the sequence encoder and the generative flow align in the latent space, a uniformity regularization is imposed to avoid collapse.

2 Related work

2.1 De Novo Molecular Generation

VAE-based VAE has been attractive in molecular generation in the virtue of its latent space is potentially operatable. CharVAE [14] first proposes to learn from molecular data in a data-driven manner and generate with a VAE model. GVAE [31] represents each data as a parse tree from a context-free grammar, and directly encodes to and decodes from these parse trees to ensure the validity of generated molecules. Inspired by syntax-directed translation in complier theory, SD-VAE [7] proposes to convert the offline syntax-directed translation check into on-the-fly generated guidance for ensuring both syntactical and semantical correctness. JT-VAE [23] first realize the direct generation of molecular graphs instead of linear SMILES (Simplified Molecular-Input Line-Entry System) strings.

GAN-based An alternative is to implement GAN in molecular generation. ORGAN [16] adds expert-based rewards under the framework of WGAN [2]. ORGANIC [50] improves the above work for inverse design chemistry and implements the molecular generation towards specific properties. MolGAN [10] proposes GAN-based models to generate molecular graphs rather than SMILES. Motivated by cycle-consistent GAN [64], Mol-CycleGAN [41] generates optimized compounds with high structural similarity to the original ones.

Flow-based Molecular generation with the normalizing flow is promising as its invertible mapping can reconstruct the data exactly. GraphNVP [40] and GRF [20] are the early works on flow-based molecular generation. GraphAF [54] combines the advantages of both autoregressive and flow-based approaches to iteratively generate molecules. MolFlow [63] proposes a variant of Glow [26] to generate atoms and bonds in a one-shot manner. MolGrow [32] constrains optimization of properties by using latent variables of the model, and recursively splits nodes.

Though these approaches have achieved significant performance, we recognize them as chemocentric molecular generation methods that lack biological connections. We aim to bridge biological and chemical perspectives in molecular generation for practical drug discovery.

2.2 Drug-target Interaction

Recent progress in artificial intelligence has inspired researchers to utilize deep learning techniques in drug-target interaction prediction. DeepDTA [43] and

DeepAffinity [24] are representatives of deep-learning methods that take SMILES of drugs and primary sequences of proteins as input, from which neural networks are employed to predict affinities. InterpretableDTIP [13] predicts DTI directly from low-level representations and provides biological interpretation using a two-way attention mechanism. DeepRelations [25] embeds protein sequences by hierarchical recurrent neural network and drug graphs by graph neural networks with joint attention between protein residues and compound atoms. MONN [33] predicts binding affinities with extra supervision from the labels extracted from available high-quality three-dimensional structures. Our proposed target-aware molecular generation builds on the recent advances in data-driven drug-target interaction prediction. We connect chemical molecular generation with biological drug-target interaction to promote the efficiency of drug discovery.

2.3 Conditional Molecular Generation

Generating molecules with the consideration of some external conditions is a promising field. CVAE [14] jointly trains VAE with a predictor that predicts properties from the latent representations of VAE. [34] proposes applying conditional VAE to generate drug-like molecules satisfying properties at the same time. [15] employs constrained Bayesian optimization to control the latent space of VAE in order to find molecules that score highly under a specified objective function. CogMol [5] and CLaSS [8] pretrain the latent space with SMILES and train property classifiers from the latent representations. They sample from the latent space that satisfies high scores from property classifiers to generate molecules. Though recent molecular generation methods [23, 40, 63, 39] also present property optimization experiments, they still barely take account of drug-target interaction. [42] proposes stacks of conditional GAN to generate hit-like molecules from gene expression signature. While this work focuses on drug-gene relationships, we instead focus on the drug-protein case.

3 Background and Preliminaries

3.1 Problem Statement

Let $\mathcal{T} = \{T_i\}_{i=1}^{t}$ be a set of targets, and there exists a set of drugs $\mathcal{M}_{T_i} = \{M_j^{(T_i)}\}_{j=1}^{d_i}$ that bind to each target T_i. $S(T, M)$ is defined as a function measuring the interaction between target T and drug M. The target-aware molecular generation aims to learning a generation model $p_\theta(\cdot|T_i)$ from each drug-target pair $(M_j^{(T_i)}, T_i)$ so as to maximize $\mathbb{E}_{M|T_i \sim p_\theta}[S(M, T_i)]$.

3.2 The Flow Framework

A flow model is a sequence of parametric invertible mapping $f_\Theta = f_Q \circ ... \circ f_1$ from the data point $x \in \mathbb{R}^D$ to the latent variable $z \in \mathbb{R}^D$, where $x \sim P_X(x), z \sim P_Z(z)$. The latent distribution P_Z is usually predefined as a simple distribution,

e.g., a normal distribution. The complex data in the original space is modelled by using the change-of-variable formula:

$$P_X(x) = P_Z(z)\left|\det\frac{\partial Z}{\partial X}\right|, \tag{1}$$

and its log-likelihood:

$$\begin{aligned}\log P_X(x) &= \log P_Z(z) + \log\left|\det\frac{\partial Z}{\partial X}\right| \\ &= \log P_Z(z) + \sum_{q=1}^{Q}\log\left|\det\frac{\partial f_q(z^{(q-1)})}{\partial z^{(q-1)}}\right|,\end{aligned} \tag{2}$$

where $z^{(q)} = f_q(z^{(q-1)})$, and we represent the input $z^{(0)}$ by using z for notation simplicity.

As the calculation of the Jacobian determinant for f_Θ is expensive for arbitrary functions, NICE [11] and RealNVP [12] develop an affine coupling transformation $z = f_\Theta(x)$ with expressive structures and efficient computation of the Jacobian determinant.

For given D-dimensional input x and $d < D$, the output y of an affine coupling transformation is defined as:

$$\begin{aligned}y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} \odot \exp(S_\Theta(x_{1:d})) + T_\Theta(x_{1:d}),\end{aligned} \tag{3}$$

where $S_\Theta : \mathbb{R}^d \to \mathbb{R}^{D-d}$ and $T_\Theta : \mathbb{R}^d \to \mathbb{R}^{D-d}$ stand for scale function and transformation function. For the sake of the numerical stability of cascading multiple flow layers, we follow Moflow [63] to replace the exponential function for the S_Θ with the Sigmoid function:

$$\begin{aligned}y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= x_{d+1:D} \odot \mathrm{Sigmoid}(S_\Theta(x_{1:d})) + T_\Theta(x_{1:d}),\end{aligned} \tag{4}$$

and the invertibility is guaranteed by:

$$\begin{aligned}x_{1:d} &= y_{1:d} \\ x_{d+1:D} &= (y_{d+1:D} - T_\Theta(y_{1:d}))/\mathrm{Sigmoid}(S_\Theta(y_{1:d})).\end{aligned} \tag{5}$$

The logarithmic Jacobian determinant is:

$$\begin{aligned}\log\left|\det\frac{\partial y}{\partial x}\right| &= \log\left|\det\left(\begin{bmatrix}\mathbb{I} & 0 \\ \frac{\partial y_{d+1:D}}{\partial x_{1:d}} & \mathrm{Sigmoid}(S_\Theta(x_{1:d}))\end{bmatrix}\right)\right| \\ &= \log\mathrm{Sigmoid}(S_\Theta(x_{1:d})).\end{aligned} \tag{6}$$

To further improve the invertible mapping with more expressive structures and high numerical stability, Glow [26] proposes using invertible 1×1 convolution to learn an optimal partition and actnorm layer to normalize dimensions in each

channel over a batch by an affine transformation. Invertible 1×1 convolution is initialized as a random rotation matrix with zero log-determinant and works as a generalization of a permutation of channels. Act norm initializes the scale and the bias such that the post-actnorm activations per-channel have zero mean and unit variance and learns these parameters in training instead of using batch statistics as batch normalization does.

3.3 Flow on the Molecular Graph

Prior works on flow-based molecular graph generation are well developed. Inspired by the graph normalizing flows of GRevNets [35], GraphNVP [40] proposes to generate atom features conditioned on the pre-generated adjacency tensors, which is then followed by other one-shot flow-based molecular graph generation approaches, e.g., GRF [20] and Moflow [63]. Our proposed SiamFlow follows this manner, that is, firstly transforms the bonds B of molecules to the latent variables Z_B with Glow [26], and then transforms the atom features A given B into the conditional latent variable $Z_{A|B}$ with a graph conditional flow.

Let N, K, C be the number of nodes, node types, and edge types, respectively. A molecular graph $G = (A, B)$ is defined by an atom matrix $A \in \{0,1\}^{N \times K}$ and a bond tensor $B \in \{0,1\}^{C \times N \times N}$, which correspond to nodes and edges in the vanilla graph. $A[i,k] = 1$ represents the i-th atom i has atom type k, and $B[c,i,j] = 1$ represents there is a bond with type c between the i-th atom and j-th atom.

Flow-based molecular graph generation methods decompose the generative model into two parts:

$$P(G) = P((A,B)) \approx P(A|B;\theta_{A|B})P(B;\theta_B), \tag{7}$$

where θ_B is learned by the bond flow model h_B, and $\theta_{A|B}$ is learned by the atom flow model $h_{A|B}$ conditioned on the bond tensor B.

With the strengths of the flow, the optimal parameters $\theta^*_{A|B}$ and θ^*_B maximize the exact likelihood estimation:

$$\arg\max_{\theta_{A|B},\theta_B} \mathbb{E}_{(A,B)\sim P_G}[\log P(A|B;\theta_{A|B}) + \log P(B;\theta_B)] \tag{8}$$

Our work follows the one-shot molecular graph generation manner [40, 20, 63] that employs Glow [26] as the bond flow model h_B and graph conditional flow as the atom flow model $h_{A|B}$.

4 SiamFlow

4.1 Overview

While current flow-based molecular graph generation methods [40, 20, 54, 63, 32, 39] learn from drug-like datasets and generate without the invention of targets, our proposed SiamFlow aims to serve as a conditional flow toward molecular

graph generation. Though the conditional flow has been well developed in computer vision [36, 28, 1, 30, 48], there are limited works that can fit graph generation, especially when it comes to the molecular graph.

In this section, we introduce SiamFlow, a novel molecular graph generative model conditioned on specific targets. As shown in Fig. 2, SiamFlow learns the distribution of sequence embedding instead of the isotropic Gaussian distribution like other flow-based methods.

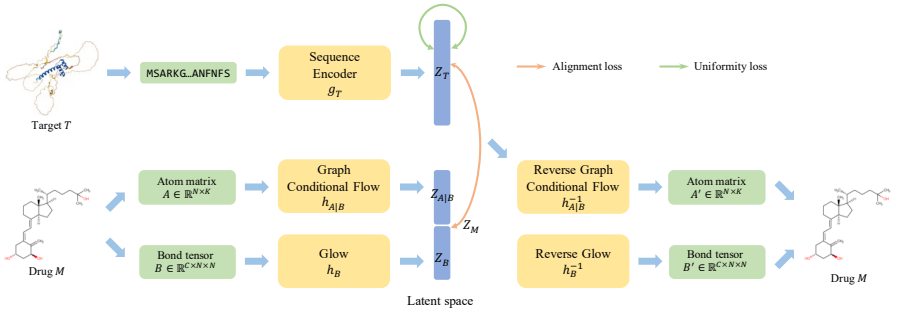

Fig. 2. The framework of our proposed SiamFlow. In the training phase, the target sequence embedding Z_T aligns with the drug graph embedding Z_M, while a uniformity regularization term forces its distribution as a spherical uniform distribution. In the generation phase, the target sequence embedding Z_T is fed into reverse flows to generate the desired drug.

4.2 Alignment Loss

Given a pair of target T and drug M, we decompose the drug M into an atom matrix $A \in \mathbb{R}^{N\times K}$ and a bond tensor $B \in \mathbb{R}^{C\times N\times N}$. The sequence encoder g_T can be arbitrary mapping that maps the target sequence T into the sequence embedding $Z_T \in \mathbb{R}^D$. The flow model contains a glow $h_B : \mathbb{R}^{C\times N\times N} \to \mathbb{R}^{\frac{D}{2}}$ and a graph conditional flow $h_{A|B} : \mathbb{R}^{N\times K} \to \mathbb{R}^{\frac{D}{2}}$. The drug graph embedding Z_M is the concatenation of $Z_{A|B}$ and Z_B.

Instead of directly learning the isotropic Gaussian distribution, we impose alignment loss between the target sequence embedding Z_T and the drug graph embedding Z_M so that Z_T can be used as the input of the generation process. Thus, the generated atom matrix and the bond tensor are:

$$A' = h^{-1}_{A|B}(Z_T[1 : \frac{D}{2}]),\ B' = h^{-1}_B(Z_T[\frac{D}{2} : D]). \tag{9}$$

While traditional flow-based models assume the latent variables follow the Gaussian distribution, SiamFlow forces the flow model to learn the distribution of the

condition information instead of a predefined distribution. We define the alignment loss $\mathcal{L}_{align}$ as:

$$\begin{aligned}\mathcal{L}_{align} &:= \mathbb{E}_{(T,M)\sim P_{\text{data}}}||Z_T - Z_M||_2 \\ &= \mathbb{E}_{(T,M)\sim P_{\text{data}}}||Z_T - [Z_{A|B}, Z_B]||_2\end{aligned} \tag{10}$$

where $[Z_{A|B}, Z_B]$ denotes the concatenation of the atom embedding $Z_{A|B}$ and the bond embedding Z_B, and the pair of protein target T and molecular drug M is sampled from the data P_{data}.

The alignment loss bridges the connections between the target sequence embedding Z_T and the drug graph embedding Z_M in the latent space, but there are still challenges that will be revealed in Sec. 4.3 and Sec. 4.4.

4.3 Uniformity Loss

Simply aligning the target sequence embedding Z_T and the drug graph embedding Z_M is not enough. There still remains three challenges: (1) the distribution of Z_T is uncertain, so that the alignment learning may be difficult to converge; (2) sampling from an unknown distribution is indefinite in the generation process; (3) the alignment loss alone admits collapsed solutions, e.g., outputting the same representation for all targets.

To overcome the above issues, we design an objective to force the target sequence embedding Z_T to follow a specific distribution, in our case the uniform distribution on the unit hypersphere [49, 29, 18, 60]. We recognize angles of embeddings are the critical element that preserves the most abundant and discriminative information. By fitting the hyperspherical uniform distribution, the projections of target sequence embeddings on the hypersphere are kept as far away from each other as possible; thus, discriminations are imposed. Specifically, we project the target sequence embedding Z_T into a unit hypersphere $\mathbb{S}^{D-1}$ by L2 normalization and require the embeddings uniformly distributed on this hypersphere, as shown in Fig. 3.

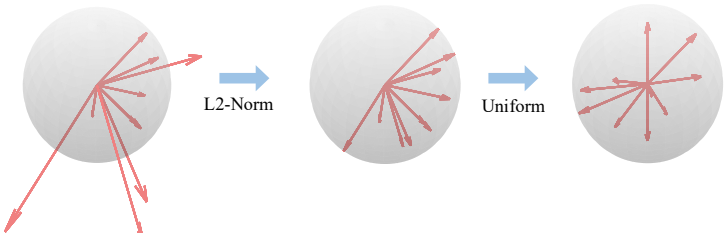

Fig. 3. The schematic diagram of the uniformity loss.

The uniform hypersphere distribution can be formulated as a minimizing pairwise potential energy problem [38, 4, 60] while higher energy implies less dis-

criminations. Let $\widehat{Z}_T = \frac{Z_T}{\|Z_T\|} \in \mathcal{C}$, and $\mathcal{C}$ is a finite subset of the unit hypersphere $\mathbb{S}^{D-1} \in \mathbb{R}^D$. We define the f-potential energy [6] of $\mathcal{C}$ to be:

$$\sum_{\widehat{Z}_T^{(x)}, \widehat{Z}_T^{(y)} \in \mathcal{C}, x \neq y} f(|\widehat{Z}_T^{(x)} - \widehat{Z}_T^{(y)}|^2). \tag{11}$$

where $\widehat{Z}_T^{(x)}$ and $\widehat{Z}_T^{(y)}$ denote normalized sequence embeddings with index x, y.

Definition. (Universally optimal [6]). A finite subset $\mathcal{C} \subset \mathbb{S}^{D-1}$ is *universally optimal* if it (weakly) minimizes potential energy among all configurations of $|\mathcal{C}|$ points on $\mathbb{S}^{D-1}$ for each completely monotonic potential function.

In SiamFlow, we consider the Gaussian function kernel $G_t(x, y) : \mathbb{S}^{D-1} \times \mathbb{S}^{D-1} \to \mathbb{R}$ as the potential function f, which is defined as:

$$G_t(x, y) = e^{-t|x-y|^2}. \tag{12}$$

This kernel function is closely related to the universally optimal configuration, and distributions of points convergence weak* to the uniform distribution by minimizing the expected pairwise potential.

Theorem. (Strictly positive definite kernels on $\mathbb{S}^D$ [4]). Consider kernel $K_f : \mathbb{S}^D \times \mathbb{S}^D \to (-\infty, +\infty]$ of the form $K_f(x, y) := f(|x - y|^2)$, if K_f is *strictly positive definite* on $\mathbb{S}^D \times \mathbb{S}^D$ and the energy $I_{K_f}[\sigma_D]$ is finite, then σ_D is the unique measure on Borel subsets of $\mathbb{S}^D$ in the solution of $\min_{\mu \in \mathcal{M}(\mathbb{S}^D)} I_{K_f}(\mu)$, and the normalized counting measure associated with any K_f-energy minimizing sequence of point configurations on $\mathbb{S}^D$ converges weak* to σ_D.

This theorem reveals the connections between strictly positive definite kernels and the energy minimizing problem. The Gaussian function is strictly positive definite on $\mathbb{S}^D \times \mathbb{S}^D$, thus well tied with the uniform distribution on the unit hypersphere.

Proposition 1. (Strictly positive definite of the Gaussian function) For any $t > 0$, the Gaussian function kernel $G_t(x, y)$ is strictly positive definte on $\mathbb{S}^D \times \mathbb{S}^D$.

Though Riesz s-kernels $R_s(x, y) := |x - y|^{-s}$ are commonly used as potential functions, we argue that the Gaussian function is expressive because it maps distances to infinite dimensions like radial basis functions, benefiting from the Taylor expansion of exponential functions. Moreover, the Gaussian function is a general case of Riesz s-kernels and can represent Riesz s-kernels by:

$$R_s(x, y) = \frac{1}{\Gamma(s/2)} \int_0^\infty G_t(x, y) t^{s/2-1} dt. \tag{13}$$

where $\Gamma(s/2) = \int_0^\infty e^{-t} t^{s/2-1}$ for $s > 0$.

As the Gaussian function kernel is an ideal choice of potential functions, we define the uniformity loss as the logarithm of the pairwise Gaussian potential's expectation:

$$\mathcal{L}_{unif} := \log \mathbb{E}_{(T^{(x)}, T^{(y)}) \sim P_{\mathcal{T}}} [G_t(\widehat{Z}_T^{(x)}, \widehat{Z}_T^{(y)})], \tag{14}$$

where $T^{(x)}$ and $T^{(y)}$ are two different targets sampled from the target data $P_{\mathcal{T}}$.

4.4 One Target to Many Drugs

Implementing the alignment loss and the uniformity loss above, the flow model can already generate validated molecular drugs conditioned on specific targets. However, there are multiple affinable drugs for a single target in most cases. To deal with this *one-to-many* problem, we reformulate learning target embeddings into learning spaces of target embeddings in the latent space, as shown in Fig. 4.

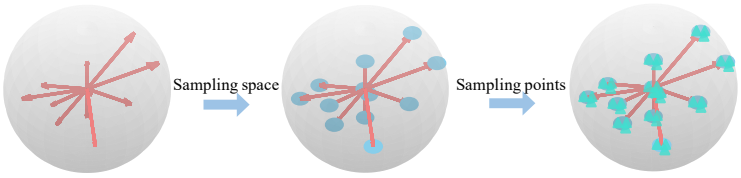

Fig. 4. The schematic diagram of the one-to-many strategy. The blue circles denote the possible spaces around the target sequence embeddings, and the green triangles denote the instances sampled from the possible spaces.

As the target embeddings have been pushed by the uniformity loss to stay as far away as possible on the hypersphere, they preserve abundant and discriminative information to a large extent. We design an adaptive space learning strategy that holds the discriminative angle information with a limited scope. For a set of target sequence embeddings $\mathcal{Z}_\mathcal{T} = \{Z_T^{(0)}, ..., Z_T^{(L)}\}$, we first calculate their standard deviation by:

$$\sigma(\mathcal{Z}_\mathcal{T}) = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(Z_T^{(i)} - \mu(\mathcal{Z}_\mathcal{T}))}, \tag{15}$$

where $\mu(\mathcal{Z}_\mathcal{T}) = \frac{1}{L}\sum_{i=1}^{L} Z_T^{(i)}$ is the mean of the set $\mathcal{Z}_\mathcal{T}$. Then, we define a space for each target sequence embedding:

$$\Omega(Z_T) = \{Z_T + Z'_T | Z'_T \in \mathcal{N}(0, \lambda\sigma^2(\mathcal{Z}_\mathcal{T}))\}, \tag{16}$$

where λ is the hyperparameter that controls the scale of the space and is empirically set as 0.1.

Note that we define the space on Z_T instead of the normalized $\widehat{Z}_T$, as normalized embeddings lose the length information to the extent that the available space is limited. Thus, we modify the alignment loss as:

$$\mathcal{L}_{align} = \mathbb{E}_{(T,M)\sim P_{\text{data}}} |\Omega(Z_T) - Z_M|. \tag{17}$$

In the generation process, sampling from the same space is permissible to generate desired drugs.

In summary, the objective is a linear combination of the modified alignment loss and uniform loss:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \mathcal{L}_{unif} \tag{18}$$

5 Experiments

Baselines Since we present a novel generative approach conditioned on targets, we primarily compare our approach to other conditional generative models, i.e., conditional VAE (CVAE) [56], CSVAE [27], PCVAE [17]. Furthermore, an attention-based Seq2seq [57, 59] neural translation model between the target protein sequence and drug SMILES is considered a straightforward solution in our setting. An explainable substructure partition fingerprint [22] is employed for sequential drug SMILES and protein sequences. We also involve GraphAF [54], GraphDF [39], and MolGrow [32] in the generative comparison.

Datasets To evaluate the ability of our proposed SiamFlow, we collect a dataset based on four drug-target interaction datasets, including BIOSNAP [65], BindingDB [37], DAVIS [9], and DrugBank [61]. We remove all the negative samples in the original datasets, and only keep the positive samples. Our dataset contains 24,669 unique drug-target pairs with 10,539 molecular drugs and 2,766 proteins. The maximum number of atoms in a molecular drug is 100 while 11 types of common atoms are considered. We split drug-target pairs by target protein sequence identity at 30%, and define the dataloader to ensure zero overlap protein in the training, validation, and test set.

Metrics To comprehensively evaluate the conditional generative models in terms of target-aware molecular generation, we design metrics from two perspectives: (1) Generative metrics. Following the common molecular generation settings, we apply metrics including: **Validity** which is the percentage of chemically valid molecules in all the generated molecules, **Uniqueness** which is the percentage of unique valid molecules in all the generated molecules, **Novelty** which is the percentage of generated valid molecules which are not in the training dataset. (2) Biochemical metrics. We evaluate the similarities between the generated drugs and the nearest drugs in the training set including: **Tanimoto similarity** which is calculated based on hashed binary features, **Fraggle similarity** which focus on the fragment-level similarity, **MACCS similarity** which employs 166-bit 2D structure fingerprints, and **Binding Score** predicted by DeepPurpose [21].

Empirical Running Time We implement our proposed method SiamFlow and the other two baselines Seq2seq, CVAE by Pytorch-1.8.1 framework. We train them with Adam optimizer with a learning rate of 0.001, batch size 16, and 100 epochs on a single NVIDIA Tesla V100 GPU. To evaluate the validity and chemical similarities, we employ the cheminformatics toolkit RDKit in the assessment phase. Our SiamFlow completes the training process of 100 epochs in an average of 1.06 hours (38 seconds/ epoch), while CVAE and Seq2seq take an average of 1.14 hours (41 seconds/ epoch) and 8.33 hours (5 minutes/ epoch) respectively.

5.1 Target-aware Molecular Graph Generation

We conduct experiments on molecular drug generation with specific targets for comparisons. For each experiment, we repeat three trials with different random seeds and report the mean and standard deviation.

Table 1 shows the results on generative metrics of our SiamFlow model in comparison to the baselines. Our proposed SiamFlow inherits the strengths of the flow and far surpasses other baselines in generative metrics. It can be seen that Seq2seq suffers from low validity, uniqueness, and novelty, which indicates Seq2seq's generation relies on its memorization. CVAE has higher uniqueness and novelty than Seq2seq though its validity is even lower. Besides, the standard deviations of metrics on CVAE are relatively high, suggesting it is volatile to train. Moreover, compared to other baselines, SiamFlow obtains superior performance with relatively low volatility.

Table 1. Evaluation results on generative metrics of SiamFlow v.s. baselines; high is better for all three metrics.

Method	% Validity	% Uniqueness	% Novelty
Seq2seq	16.08±4.14	13.87±1.74	14.89±11.41
CVAE	12.54±7.56	72.30±20.33	99.72±0.39
CSVAE	76.53±2.4	60.31±6.56	99.37±0.59
PCVAE	78.81±2.4	89.32±2.74	99.59±0.32
GraphAF	100.00±0.00	98.68±0.40	100.00±0.00
GraphDF	100.00±0.00	96.97±0.23	100.00±0.00
MolGrow	100.00±0.00	99.57±0.01	100.00±0.00
SiamFlow	**100.00**±0.00	**99.61**±0.16	**100.00**±0.00

In addition to generative metrics, we also report chemical metrics in Table 2. The generated molecular drugs are expected to have a chemical structure similar to the ground-truth drugs in order to have a high binding affinity to the target. SiamFlow is consistently better than other baselines in both the Tanimoto and Fraggle similarity while obtaining relatively lower MACCS similarity than Seq2seq. Considering that MACCS measures the similarity of encodings of molecules, the sequence partition rules of Seq2seq may help it. Thus, we pay more attention to the Tanimoto and Fraggle similarity because they are structure-centric metrics.

We visualize the distribution of the Tanimoto similarity and the Fraggle similarity evaluated on these methods in Fig. 6. SiamFlow consistently outperforms other methods and generates desirable molecular drugs. The examples of generated drugs are shown in Fig. 5.

Table 2. Evaluation results on biochemical metrics of SiamFlow v.s. baselines.

Method	% Tanimoto (↑)	% Fraggle (↑)	% MACCS (↑)	Binding Score (↓)
Seq2seq	26.27±9.91	25.84±7.27	**37.98**±7.70	8.83±4.70
CVAE	7.76±6.61	12.31±5.81	16.42±7.17	10.92±5.28
CSVAE	18.49±3.92	16.67±2.71	17.91±3.21	6.91±3.10
PCVAE	39.59±2.17	24.56±3.17	25.74±1.14	4.87±2.34
SiamFlow	**48.55**±0.97	**34.41**±0.35	29.30±1.07	**2.07**±0.15

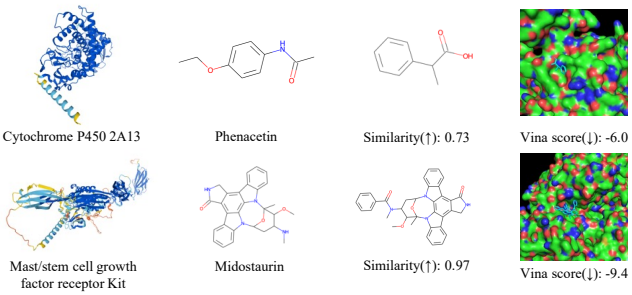

Fig. 5. Examples of the generated drugs.

5.2 Ablation Study

We conduct the ablation study and report the results in Table 3 and Table 4. It can be seen from Table 3 that simply aligning the target sequence embedding and drug graph embedding will result in extremely low uniqueness. Our one-to-many strategy enriches the latent space so that one target can map to different drugs. The absence of $\mathcal{L}_{unif}$ does not harm the generative metrics because it only constrains the distribution of target sequence embeddings but has a limited impact on the generation process.

Table 3. Ablation results on generative metrics.

Method	% Validity	% Uniqueness	% Novelty
SiamFlow	100.00	99.39	100.00
w/o one-to-many	100.00	12.55	100.00
w/o $\mathcal{L}_{unif}$	100.00	100.00	100.00

Table 4 demonstrates the chemical metrics are well without the one-to-many strategy. If we generate only one drug for a particular target, the nearest drug

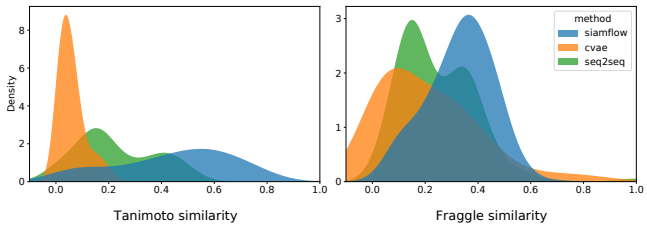

Fig. 6. The distribution of generative metrics evaluated on SiamFlow and baselines.

similarity degrades to a special case, i.e., comparing the generated drug with its corresponding one in the training set. Moreover, removing $\mathcal{L}_{unif}$ severely impairs the chemical performance, suggesting uniformity loss promotes the expressive abilities of target sequence embeddings.

Table 4. Ablation results on chemical metrics.

Method	% Tanimoto	% Fraggle	% MACCS
SiamFlow	49.43	34.62	29.55
w/o one-to-many	48.83	34.93	31.23
w/o $\mathcal{L}_{unif}$	18.49	15.70	17.91

6 Conclusion and Discussion

In this paper, we delve into the topic of target-aware molecular graph generation, which involves creating drugs that are specifically conditioned on particular targets. While existing methods focus on developing drugs similar to those found in drug-like datasets, target-aware molecular generation combines drug-like molecular generation with target-specific screening to simplify the drug-target interaction step. To thoroughly explore this problem, we compile a benchmark dataset using several public datasets. Furthermore, we leverage recent progress in flow-based molecular graph generation methods and propose SiamFlow as a solution for target-aware molecular generation. Through the use of alignment and uniform loss, our proposed method can effectively generate molecular drugs conditioned on protein targets. Additionally, we address the challenge of generating multiple drugs for a single target by aligning the embedding space, rather than relying on a single embedding. Extensive experiments and analyses demonstrate that SiamFlow is a highly promising solution for target-aware molecular generation.

Acknowledgements

This work was supported by the National Key R&D Program of China (Project 2022ZD0115100), the National Natural Science Foundation of China (Project U21A20427), the Research Center for Industries of the Future (Project WU2022C043).

Ethical Statement

Our submission does not involve any ethical issues, including but not limited to privacy, security, etc.

References

1. Abdelhamed, A., Brubaker, M.A., Brown, M.S.: Noise flow: Noise modeling with conditional normalizing flows. In: ICCV (2019)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML (2017)
3. Bandholtz, S., Wichard, J., Kühne, R., Grötzinger, C.: Molecular evolution of a peptide gpcr ligand driven by artificial neural networks. PloS one (2012)
4. Borodachov, S.V., Hardin, D.P., Saff, E.B.: Discrete energy on rectifiable sets (2019)
5. Chenthamarakshan, V., Das, P., et al.: Cogmol: target-specific and selective drug design for covid-19 using deep generative models. NeurIPS (2020)
6. Cohn, H., Kumar, A.: Universally optimal distribution of points on spheres. J. Amer. Math. Soc (2007)
7. Dai, H., Tian, Y., Dai, B., Skiena, S., Song, L.: Syntax-directed variational autoencoder for molecule generation. In: ICLR (2018)
8. Das, P., Sercu, T., et al.: Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nature Biomedical Engineering (2021)
9. Davis, I.M., Hunt, P.J., et al.: Comprehensive analysis of kinase inhibitor selectivity. Nature BioTechnology (2011)
10. De Cao, N., Kipf, T.: MolGAN: An implicit generative model for small molecular graphs. ICML Workshop (2018)
11. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. In: ICLR Workshop (2015)
12. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. In: ICLR (2017)
13. Gao, K.Y., Fokoue, A., et al.: Interpretable drug target prediction using deep neural representation. In: IJCAI (2018)
14. Gómez-Bombarelli, R., Wei, J.N., et al.: Automatic chemical design using a data-driven continuous representation of molecules. ACS central science (2018)
15. Griffiths, R.R., Hernández-Lobato, J.M.: Constrained bayesian optimization for automatic chemical design using variational autoencoders. Chemical science (2020)
16. Guimaraes, G.L., Sanchez-Lengeling, B., et al.: Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint arXiv:1705.10843 (2017)
17. Guo, X., Du, Y., Zhao, L.: Property controllable variational autoencoder via invertible mutual dependence. In: ICLR (2021)
18. Hardin, D.P., Saff, E.B., et al.: Discretizing manifolds via minimum energy points. Notices of the AMS (2004)
19. Hert, J., Irwin, J.J., et al.: Quantifying biogenic bias in screening libraries. Nature chemical biology (2009)
20. Honda, S., Akita, H., et al.: Graph residual flow for molecular graph generation. arXiv preprint arXiv:1909.13521 (2019)

21. Huang, K., Fu, T., et al.: Deeppurpose: a deep learning library for drug–target interaction prediction. Bioinformatics (2020)
22. Huang, K., Xiao, C., Glass, L., Sun, J.: Explainable substructure partition fingerprint for protein, drug, and more. In: NeurIPS (2019)
23. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. In: ICML (2018)
24. Karimi, M., Wu, D., et al.: Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics (2019)
25. Karimi, M., Wu, D., et al.: Explainable deep relational networks for predicting compound–protein affinities and contacts. J. Chem. Inf. Model. (2020)
26. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. NeurIPS (2018)
27. Klys, J., Snell, J., et al.: Learning latent subspaces in variational autoencoders. NIPS (2018)
28. Kondo, R., Kawano, K., Koide, S., Kutsuna, T.: Flow-based image-to-image translation with feature disentanglement. NeurIPS (2019)
29. Kuijlaars, A., Saff, E.: Asymptotics for minimal discrete energy on the sphere. Trans. Amer. Math. Soc. (1998)
30. Kumar, M., Babaeizadeh, M., et al.: Videoflow: A conditional flow-based model for stochastic video generation. In: ICLR (2019)
31. Kusner, M.J., Paige, B., Hernández-Lobato, J.M.: Grammar variational autoencoder. In: ICML (2017)
32. Kuznetsov, M., Polykovskiy, D.: Molgrow: A graph normalizing flow for hierarchical molecular generation. In: AAAI (2021)
33. Li, S., Wan, F., et al.: Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. Cell Systems (2020)
34. Lim, J., Ryu, S., et al.: Molecular generative model based on conditional variational autoencoder for de novo molecular design. J. Cheminform (2018)
35. Liu, J., Kumar, A., et al.: Graph normalizing flows. NeurIPS (2019)
36. Liu, R., Liu, Y., Gong, X., Wang, X., Li, H.: Conditional adversarial generative flow for controllable image synthesis. In: CVPR (2019)
37. Liu, T., Lin, Y., et al.: Bindingdb: A web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Research (2007)
38. Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., Song, L.: Learning towards minimum hyperspherical energy. NeurIPS (2018)
39. Luo, Y., Yan, K., Ji, S.: Graphdf: A discrete flow model for molecular graph generation. arXiv preprint arXiv:2102.01189 (2021)
40. Madhawa, K., Ishiguro, K., et al.: Graphnvp: An invertible flow model for generating molecular graphs. arXiv preprint arXiv:1905.11600 (2019)
41. Maziarka, Ł., Pocha, A., et al.: Mol-cyclegan: a generative model for molecular optimization. J. Cheminform (2020)
42. Méndez-Lucio, O., Baillif, B., et al.: De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature communications (2020)
43. Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics (2018)
44. Phatak, S.S., Stephan, C.C., et al.: High-throughput and in silico screenings in drug discovery. Expert opinion on drug discovery (2009)

45. Polishchuk, P.G., Madzhidov, T.I., Varnek, A.: Estimation of the size of drug-like chemical space based on gdb-17 data. Journal of computer-aided molecular design (2013)
46. Popova, M., Shvets, M., Oliva, J., Isayev, O.: Molecularrnn: Generating realistic molecular graphs with optimized properties. arXiv preprint arXiv:1905.13372 (2019)
47. Prykhodko, O., Johansson, S.V., et al.: A de novo molecular generation method using latent vector based generative adversarial network. J. Cheminform (2019)
48. Pumarola, A., Popov, S., et al.: C-flow: Conditional generative flow models for images and 3d point clouds. In: CVPR (2020)
49. Saff, E.B., Kuijlaars, A.B.: Distributing many points on a sphere. The mathematical intelligencer (1997)
50. Sanchez-Lengeling, B., Outeiral, C., et al.: Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry. ChemRxiv (2017)
51. Schneider, G.: Virtual screening: an endless staircase? Nature Reviews Drug Discovery (2010)
52. Schneider, G.: De novo molecular design (2013)
53. Schneider, P., Schneider, G.: De novo design at the edge of chaos: Miniperspective. Journal of medicinal chemistry (2016)
54. Shi, C., Xu, M., et al.: Graphaf: a flow-based autoregressive model for molecular graph generation. In: ICLR (2019)
55. Simonovsky, M., Komodakis, N.: Graphvae: Towards generation of small graphs using variational autoencoders. In: ICANN (2018)
56. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. NeurIPS (2015)
57. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NeurIPS (2014)
58. Van Oord, A., Kalchbrenner, N., et al.: Pixel recurrent neural networks. In: ICML (2016)
59. Vaswani, A., Shazeer, N., et al.: Attention is all you need. In: NeurIPS (2017)
60. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML (2020)
61. Wishart, S.D., Feunang, D.Y., et al.: Drugbank 5.0: A major update to the drugbank database for 2018. Nucleic Acids Research (2018)
62. You, J., Liu, B., et al.: Graph convolutional policy network for goal-directed molecular graph generation. In: NeurIPS (2018)
63. Zang, C., Wang, F.: Moflow: an invertible flow model for generating molecular graphs. In: SIGKDD (2020)
64. Zhu, J.Y., Park, T., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In: ICCV (2017)
65. Zitnik, M., Sosic, R., et al.: Biosnap datasets: Stanford biomedical network dataset collection (2018)