ChemFlow: Traversing Chemical Space with Latent Potential Flows

Guanghao Wei [1*] Yining Huang [2*] Chenru Duan [3] Yue Song [4‡] Yuanqi Du [1‡]

[1]Cornell University [2]Harvard University [3]Massachusetts Institute of Technology [4]University of Trento [*]Equal Contribution [‡]Corresponding

Overview

We propose a new framework, ***ChemFlow***, based on potential flows to efficiently explore the latent structure of molecule generative models.

- We unify previous studies on molecule latent space traversal under the realm of flow that transforms data density along time via a vector field.
- We validate the efficacy of ChemFlow on molecule manipulation and single- and multi-objective molecule optimization tasks under both supervised and unsupervised molecular discovery settings.

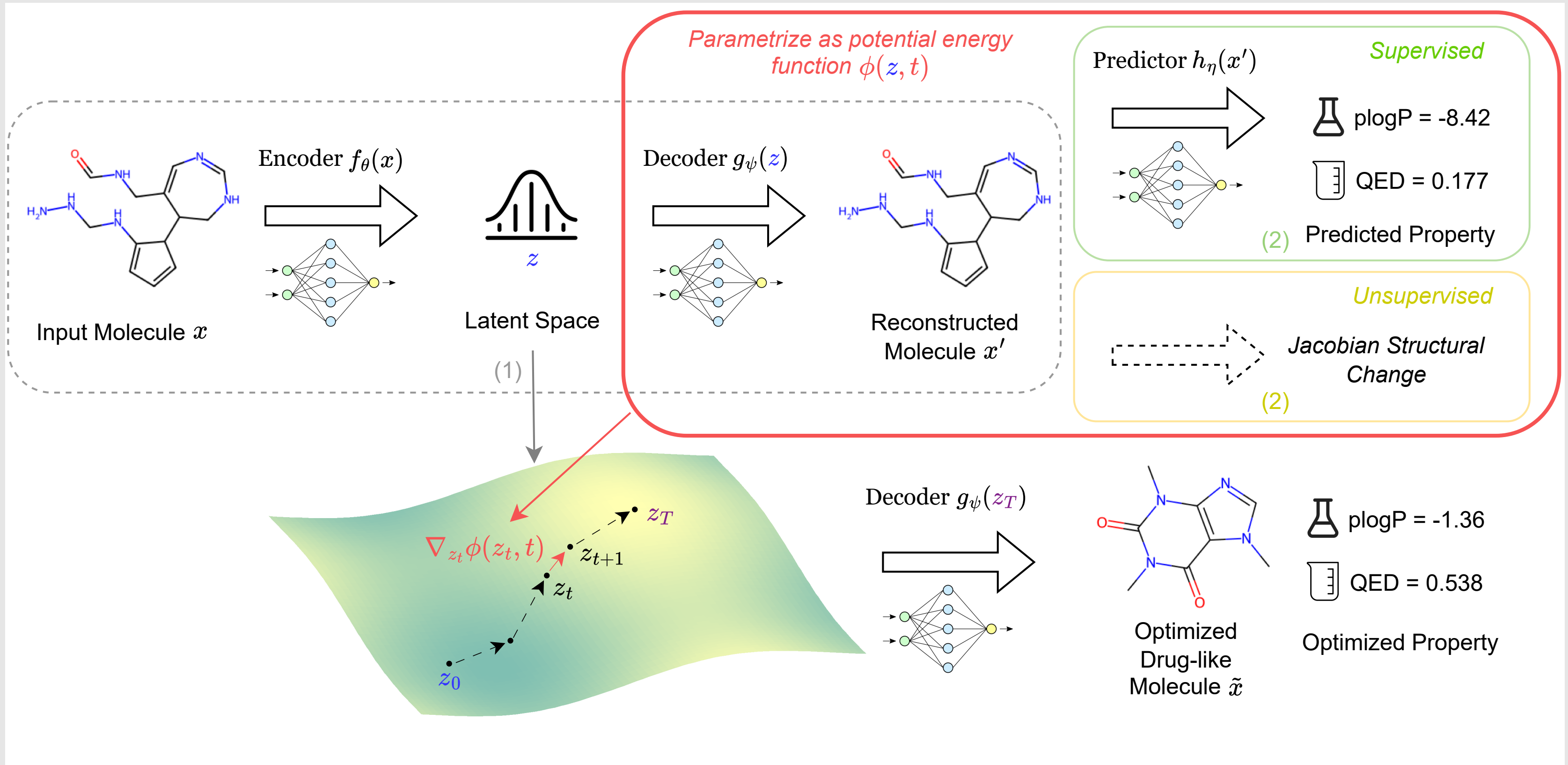

Figure 1. *ChemFlow* framework: (1) a pre-trained encoder $f_\theta(\cdot)$ and decoder $g_\psi(\cdot)$ that maps between molecules $\boldsymbol{x}$ and latent vectors $\boldsymbol{z}$, (2) we use a property predictor $h_\eta(\cdot)$ (green box) or a "Jacobian control" (yellow box) as the guidance to learn a vector field $\nabla\phi(\boldsymbol{z}_t,t)$ that maximizes the change in certain molecular properties (e.g. plogP, QED) or molecular structures, (3) during the training process, we add additional dynamical regularization on the flow. The learned flows move the latent samples to change the structures and properties of the molecules smoothly.

Motivation

Deep generative models have stimulated significant interest in more structured data generation such as molecules. However, beyond generating new random molecules, efficient exploration and a comprehensive understanding of the vast chemical space are of great importance to molecular science and applications in drug and materials discovery.

Background

Gradient-based optimization follows the direction of steepest descent of the potential energy function $h(\cdot)$ and discretized as the following ODE:

$$\boldsymbol{z}'(t) = -\nabla h(\boldsymbol{z}(t)) \tag{1}$$

Wasserstein Potential Flows. The commonly used L_2 Wasserstein distance has the following dynamic formulation:

$$W_2(\rho_0,\rho_1)^2 = \min_{\rho,v}\left\{\int\int \frac{1}{2}\rho(\boldsymbol{x},t)|v(\boldsymbol{x},t)|^2\,dx\,dt : \partial_t\rho(\boldsymbol{x},t) = -\nabla\cdot(v(\boldsymbol{x},t)\rho(\boldsymbol{x},t))\right\} \tag{2}$$

Solving Eq. (2) by by Karush–Kuhn–Tucker (KKT) conditions will give the optimal solution — the Hamilton-Jacobi Equation (HJE):

$$\frac{\partial}{\partial t}\phi(\boldsymbol{z},t) + \frac{1}{2}||\nabla_{\boldsymbol{z}}\phi(\boldsymbol{z},t)||^2 = 0 \tag{3}$$

Alternatively, we can sacrifice the optimal transport property and restrict the advection term to enforce other types of dynamics for smooth spatiotemporal dynamics, e.g., second-order Wave equation with wave coefficient c:

$$r(\boldsymbol{z}_t,t) = \frac{\partial^2}{\partial t^2}\phi(\boldsymbol{z}_t,t) - c^2\nabla^2_{\boldsymbol{z}}\phi(\boldsymbol{z}_t,t) \tag{4}$$

ChemFlow Framework

Given learned molecule generative model with encoder $f_\theta(\cdot)$, decoder $g_\psi(\cdot)$, and latent space $\mathcal{Z}$, we aim to parameterize a set of K scalar potential energies $\phi^k = \texttt{MLP}_{\theta^k}(\boldsymbol{z},t) \in \mathbb{R}$ to be a Physics-informed Neural Networks (PINN) and use the potential flow $\nabla_{\boldsymbol{z}}\phi$ to traverse the latent samples as the following procedure:

$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} + \nabla_{\boldsymbol{z}}\phi^k(\boldsymbol{z}_{t-1}, t-1) \tag{5}$$

Our PINN objective involves minimizing the PDE residual and initial condition terms:

$$\mathcal{L}_r = \frac{1}{T}\sum_{t=0}^{T-1}||r^k(\boldsymbol{z}_t,t)||_2^2,\ \mathcal{L}_\phi = ||\nabla_{\boldsymbol{z}}\phi^k(\boldsymbol{z}_0,0)||_2^2 \tag{6}$$

Supervised Semantic Potential Guidance. When labeled data of the semantic of interest is available, we use a pre-trained surrogate model $h_\eta : \mathcal{X} \to \mathbb{R}$ to predict the corresponding molecular property to guide the potential flow as the following:

$$d = \langle -\nabla_{\boldsymbol{z}} h_\eta(g_\psi(\boldsymbol{z}_t)), \nabla_{\boldsymbol{z}}\phi^k(\boldsymbol{z}_t,t)\rangle,\ \mathcal{L}_{\mathcal{P}} = -\operatorname{sign}(d)\|d\|_2^2 \tag{7}$$

Unsupervised Structure Diversity Guidance. [4] When no explicit potential energy function is provided, we devise a potential energy that maximizes the continuous Jacobian structure change of the generated molecules:

$$\mathcal{L}_{\mathcal{J}} = -\left\|\frac{\partial g(\boldsymbol{z}_t)}{\partial \boldsymbol{z}_t}\nabla_{\boldsymbol{z}}\phi^k(\boldsymbol{z}_t,t)\right\|_2^2 \tag{8}$$

To prevent K potential flows from collapsing into identical directions that correspond to the maximum Jacobian change, we adopt an auxiliary classifier l_γ to predict the potential index and use the cross-entropy loss to optimize it:

$$\hat{k}{=}l_\gamma(g_\psi(\boldsymbol{z}_t); g_\psi(\boldsymbol{z}_{t+1})),\ \mathcal{L}_k = \mathcal{L}_{CE}(\hat{k},k) \tag{9}$$

In conclusion, the overall objective of PINN under supervised and unsupervised settings are

$$\mathcal{L}_{\text{spv}} = \mathcal{L}_r + \mathcal{L}_\phi + \mathcal{L}_{\mathcal{P}},\ \mathcal{L}_{\text{unsup}} = \mathcal{L}_r + \mathcal{L}_\phi + \mathcal{L}_{\mathcal{J}} + \mathcal{L}_k \tag{10}$$

Connection with Langevin Dynamics. In case we adhere to the Fokker-Planck equation, our approach can be interpreted as employing a learned potential energy function h_η to simulate Langevin Dynamics for global optimization [3], where

$$\begin{aligned} \mathrm{d}\boldsymbol{z}_t &= -\nabla_z h_\eta(\boldsymbol{z}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{w}_t \\ \boldsymbol{z}_t &= \boldsymbol{z}_{t-1} - \nabla_z h_\eta(\boldsymbol{z}_{t-1})\mathrm{d}t + \sqrt{2\mathrm{d}t}\mathcal{N}(0,I) \end{aligned} \tag{11}$$

Unconstrained Single-objective Optimization

We randomly sample 100K molecules (10K molecules for docking tasks) from the latent space and report the top 3 scores after 10 steps of manipulation of each method.

Table 1. **Unconstrained plogP, QED maximization, and docking score minimization.** (SPV denotes supervised scenarios, UNSUP denotes unsupervised scenarios). Boldface highlights the highest-performing generation for each property within each rank.

Method	plogP ↑			QED ↑			ESR1 Docking ↓			ACAA1 Docking ↓		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Random	3.52	3.43	3.37	0.940	0.933	0.932	-10.32	-10.18	-10.03	-9.86	-9.50	-9.34
ChemSpace	3.74	3.69	3.64	0.941	0.936	0.933	**-11.66**	-10.52	-10.43	-9.81	-9.72	-9.63
Gradient Flow	4.06	3.69	3.54	0.944	0.941	0.941	-11.00	-10.67	-10.46	-9.90	-9.64	-9.61
Wave (spv)	4.76	3.78	3.71	**0.947**	0.934	0.932	-11.05	**-10.71**	**-10.68**	**-10.48**	**-10.04**	**-9.88**
Wave (unsup)	**5.30**	**5.22**	**5.14**	0.905	0.902	0.978	-10.22	-10.06	-9.97	-9.69	-9.64	-9.57
HJ (spv)	4.39	3.70	3.48	0.946	0.941	0.940	-10.68	-10.56	-10.52	-9.89	-9.61	-9.60
HJ (unsup)	4.26	4.10	4.07	0.930	0.928	0.927	-10.24	-9.96	-9.92	-9.73	-9.31	-9.24
LD	4.74	3.61	3.55	**0.947**	**0.947**	**0.942**	-10.68	-10.29	-10.28	-10.34	-9.74	-9.64

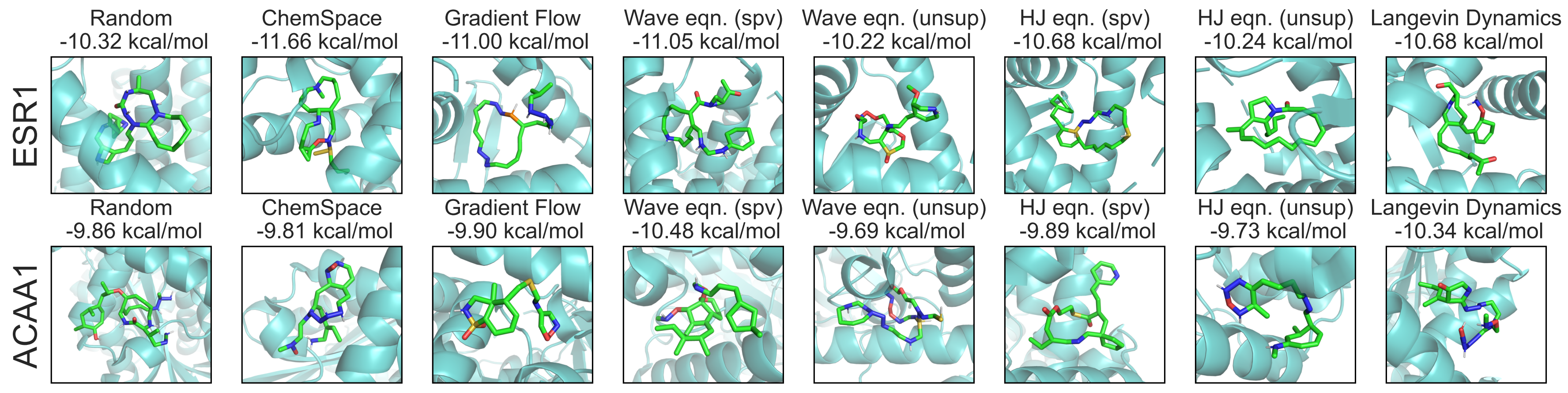
ESR1
Random -10.32 kcal/mol
ChemSpace -11.66 kcal/mol
Gradient Flow -11.00 kcal/mol
Wave eqn. (spv) -11.05 kcal/mol
Wave eqn. (unsup) -10.22 kcal/mol
HJ eqn. (spv) -10.68 kcal/mol
HJ eqn. (unsup) -10.24 kcal/mol
Langevin Dynamics -10.68 kcal/mol
ACAA1
Random -9.86 kcal/mol
ChemSpace -9.81 kcal/mol
Gradient Flow -9.90 kcal/mol
Wave eqn. (spv) -10.48 kcal/mol
Wave eqn. (unsup) -9.69 kcal/mol
HJ eqn. (spv) -9.89 kcal/mol
HJ eqn. (unsup) -9.73 kcal/mol
Langevin Dynamics -10.34 kcal/mol

Figure 2. Visualization of generated ligands docked against target ESR1 and ACAA1.

Similarity-constrained Molecule Optimization

We select the 800 molecules with the lowest QED scores in the ZINC250k dataset and perform 1,000 steps of optimization until all methods are converged. Molecules are cut off by levels of constraints δ.

Table 2. **Similarity-constrained plogP maximization.** For each method with minimum similarity constraint δ, the results in reported in format mean ± standard derivation (success rate %) of absolute improvement, where the mean and standard derivation are calculated among molecules that satisfy the similarity constraint.

Method	$\delta = 0$	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.6$
Random	11.76 ± 6.18 (99.0)	7.64 ± 6.38 (80.0)	5.03 ± 5.70 (52.1)	2.37 ± 3.71 (21.1)
ChemSpace	12.13 ± 6.41 (99.8)	9.07 ± 6.80 (90.2)	**7.52 ± 6.29 (59.4)**	**5.70 ± 5.84 (20.2)**
Gradient Flow	7.88 ± 7.28 (60.4)	7.20 ± 6.98 (56.5)	5.45 ± 6.45 (41.9)	3.60 ± 5.50 (18.4)
Wave (spv)	6.83 ± 7.15 (59.6)	5.62 ± 6.42 (54.9)	4.31 ± 5.55 (41.9)	2.47 ± 4.21 (20.6)
Wave (unsup)	19.76 ± 13.62 (99.6)	7.47 ± 9.62 (50.2)	2.06 ± 4.37 (27.3)	0.77 ± 2.21 (16.8)
HJ (spv)	8.58 ± 8.08 (68.0)	6.62 ± 7.44 (60.0)	4.27 ± 5.40 (40.6)	2.39 ± 4.10 (18.5)
HJ (unsup)	**20.64 ± 12.93 (98.0)**	8.57 ± 9.69 (50.1)	2.12 ± 3.55 (19.5)	0.67 ± 0.86 (8.6)
Langevin Dynamics	12.98 ± 6.23 (99.6)	**9.70 ± 6.21 (94.4)**	6.14 ± 5.99 (70.9)	2.94 ± 4.34 (35.4)

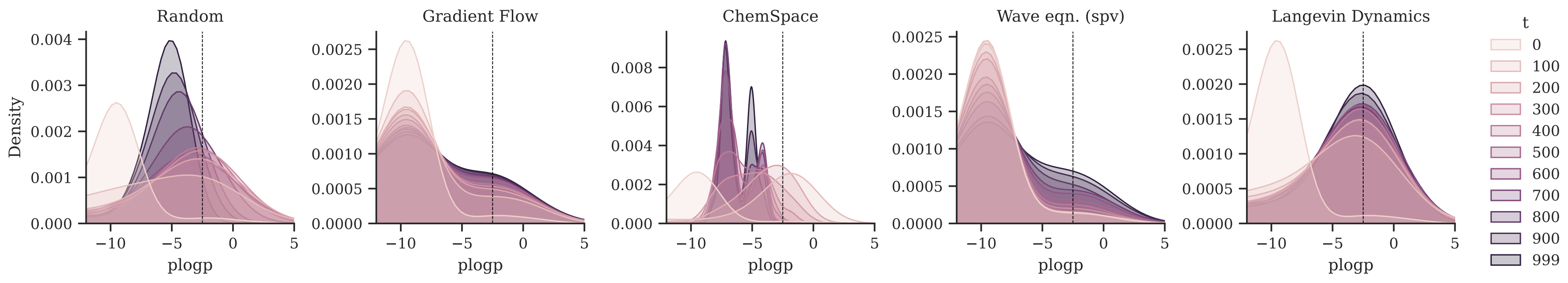

Figure 3. **Molecular property distribution shifts following the latent traversal path.** The results correspond to a similarity-constrained molecule optimization task with $\delta = 0$. LD effectively effectively shifts the distributions, while ChemSpace results in Out-of-distribution (OOD) issues.

Manipulation & Optimization Trajectories

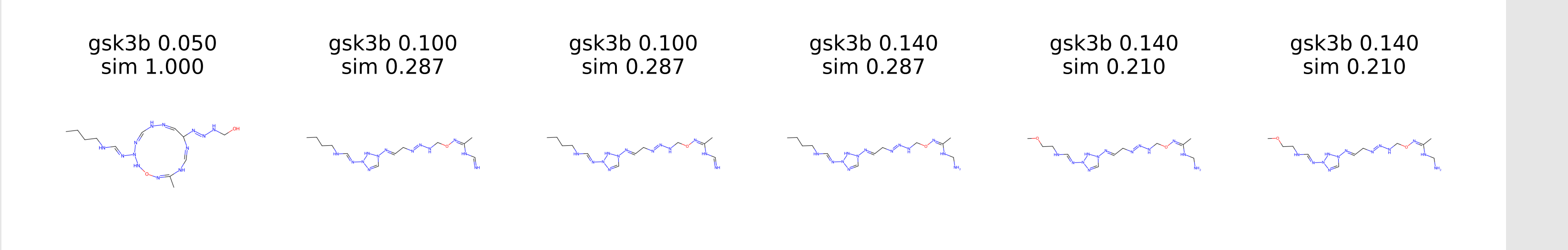

Figure 4. **Molecule Manipulation Trajectory**. The figure shows a full 10 step manipulation by gradient flow on plogP. Each molecule in the figure represents a step in the path.

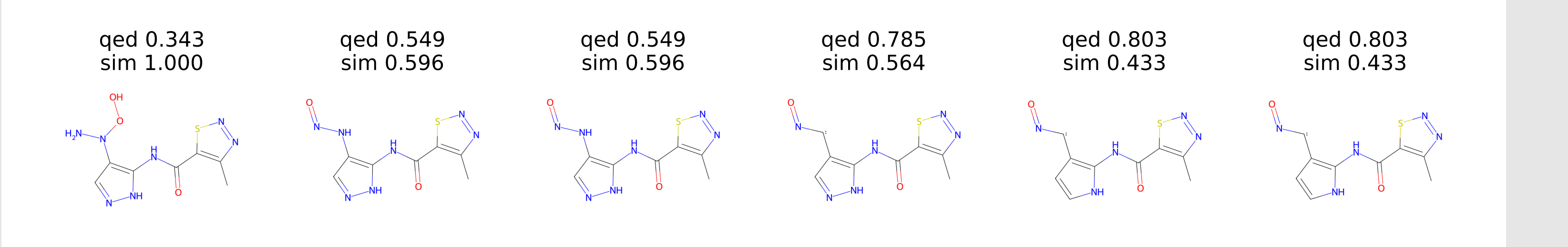

Figure 5. **Molecule Optimization Trajectory**. From left to right, each molecule is a step selected from a full 1000 step optimization trajectory by Fokker Planck flow on plogP. Only 10 intermediate steps, during which the molecules underwent changes, are shown in this gifure.

Conclusion & Limitation

- Formulate the traversal process as a flow that learns a vector flow to transport the molecular distribution through time
- Propose a variety of regularizations on the dynamics that exhibit different properties

References

[1] Yuanqi Du, Xian Liu, Nilay Mahesh Shah, Shengchao Liu, Jieyu Zhang, and Bolei Zhou. Chemspace: Interpretable and interactive chemical space exploration. *Trans. Mach. Learn. Res.*, 2023, 2023.

[2] Peter Eckmann, Kunyang Sun, Bo Zhao, Mudong Feng, Michael Gilson, and Rose Yu. Limo: Latent inceptionism for targeted molecule generation. In *International Conference on Machine Learning*, pages 5777–5792. PMLR, 2022.

[3] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

[4] Yue Song, Andy Keller, Nicu Sebe, and Max Welling. Flow factorzied representation learning. In *NeurIPS*, 2023.

Email: gw338@cornell.edu