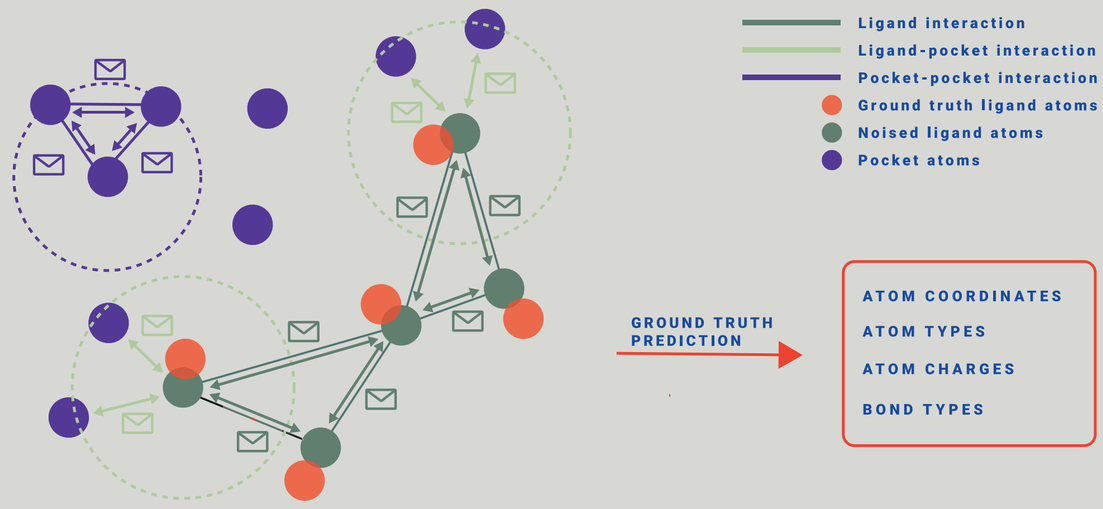
PoLiGenX: Latent-Conditioned Equivariant Diffusion for Controlled Target-Aware De Novo Ligand Design

TUAN LE*[1,2], JULIAN CREMER*[1,3], DJORK-ARNÉ CLEVERT[1] AND KRISTOF T. SCHÜTT[1]

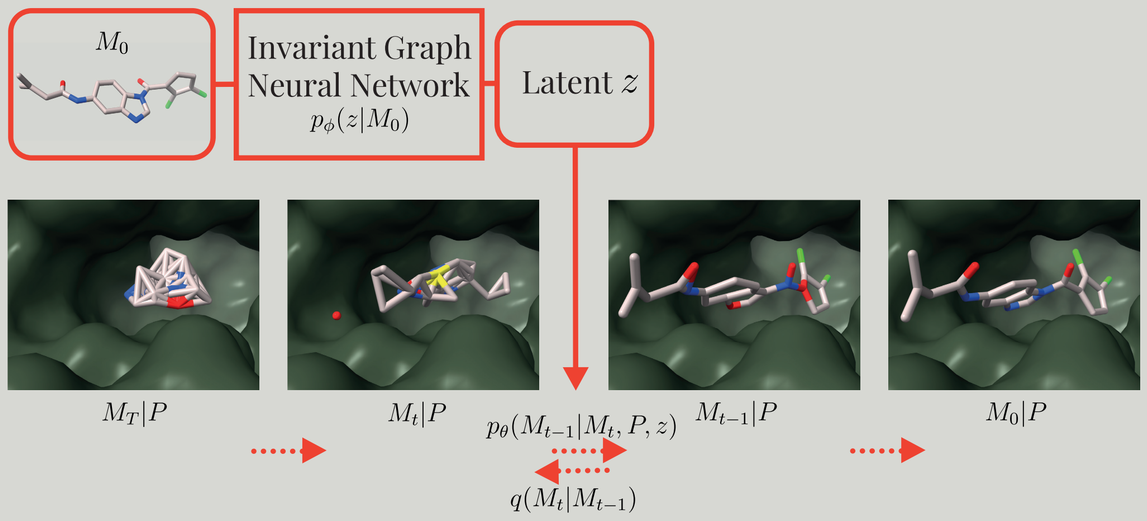
[1] Pfizer Research & Development
[2] Freie Universität Berlin
[3] Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB)
* Equal contribution

Overview

EQUIVARIANT TARGET-AWARE DIFFUSION



TRAINING AND INFERENCE



Motivation and Background

1. Failing to consider the essential chemical properties for target binding can lead to a significant lack of specificity and result in ineffective drug candidates. Moreover, drug candidates must exhibit favorable absorption, distribution, metabolism, excretion (ADME), and toxicity profiles.
2. But, the respective data is often too sparse and too noisy for developing effective machine learning models. Thus, designing ligands from scratch without addressing these critical properties may produce molecules with poor bioavailability or potential toxicity, thereby limiting their therapeutic potential.
3. However, can we use machine learning during the hit expansion phase of drug discovery? This crucial stage involves enhancing and exploring the chemical space around promising hits that are already identified through high-throughput screening or other methods and might provide significantly better starting points for generative models.
4. Can we perform hit expansion by just adding one layer of control to a generative diffusion model? We propose a latent-conditional training and sampling. Here, a seed molecule is given to the model in form of a jointly learned latent embedding to steer the diffusion process.

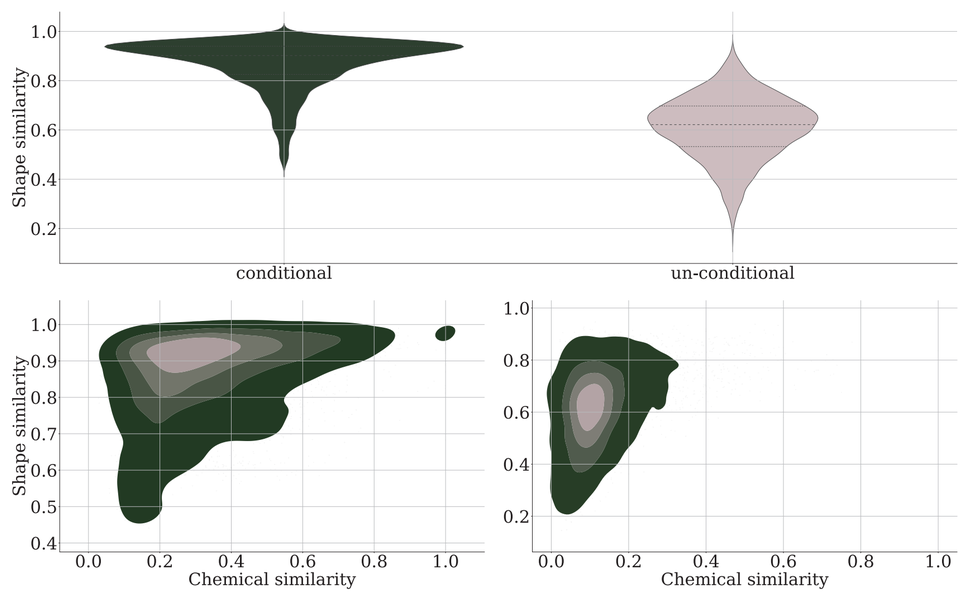
We introduce PoLiGenX (**Po**cket-based **Li**gand **Gen**erator for hit e**X**pansion), a novel latent-controlled *de novo* generative model, denoted as $p_\theta(M|P, z)$, designed for generating 3D ligands represented by $x = (\mathbf{H}, \mathbf{X}, \mathbf{E})$, with $\mathbf{H} \in \{0,1\}^{N \times K_a}$, $\mathbf{E} \in \{0,1\}^{N \times N \times K_b}$, $\mathbf{X} \in \mathbb{R}^{N \times 3}$. This model processes both continuous and discrete variables and is conditioned on a specific protein pocket P and a seed molecule embedding z.

Experiments

SHAPE PRESERVATION

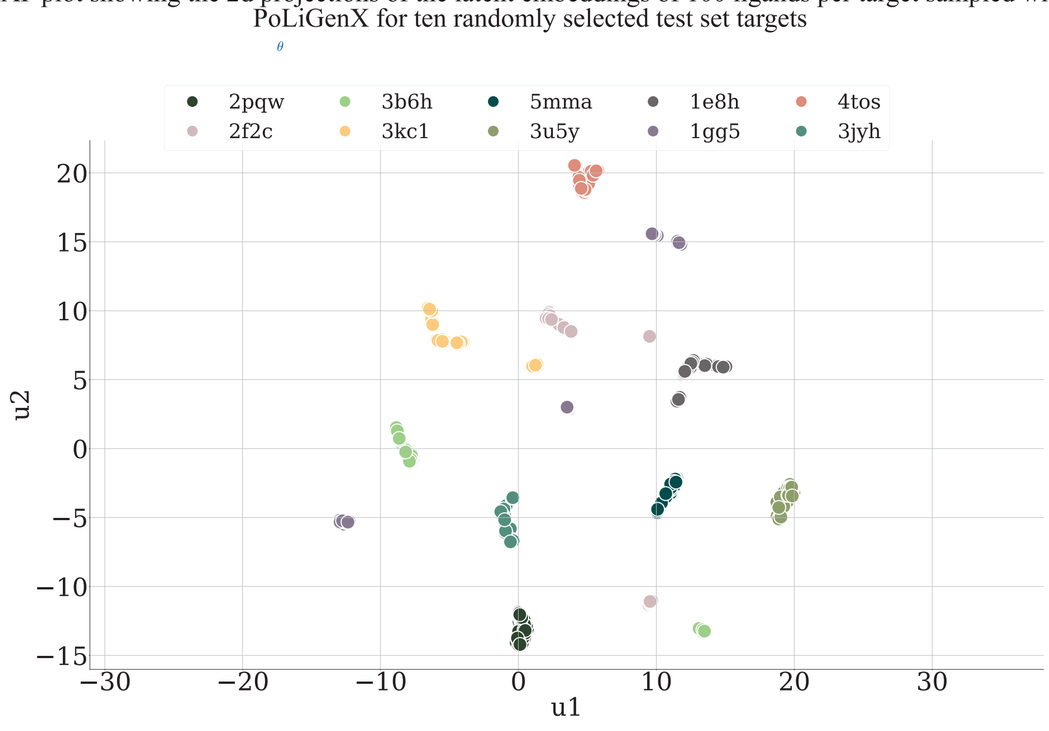
top: Violin plot of the Tanimoto shape similarity evaluated across 100 ligands per CrossDocked2020 test set target. PoLiGenX (left) is compared to EQGAT-diff (right)

bottom: Heatmap histogram comparing PoLiGenX (left) with EQGAT-diff (right) with respect to Tanimoto shape and chemical similarity. The brighter the more molecules.



LATENT EMBEDDINGS

UMAP plot showing the 2d projections of the latent embeddings of 100 ligands per target sampled with PoLiGenX for ten randomly selected test set targets

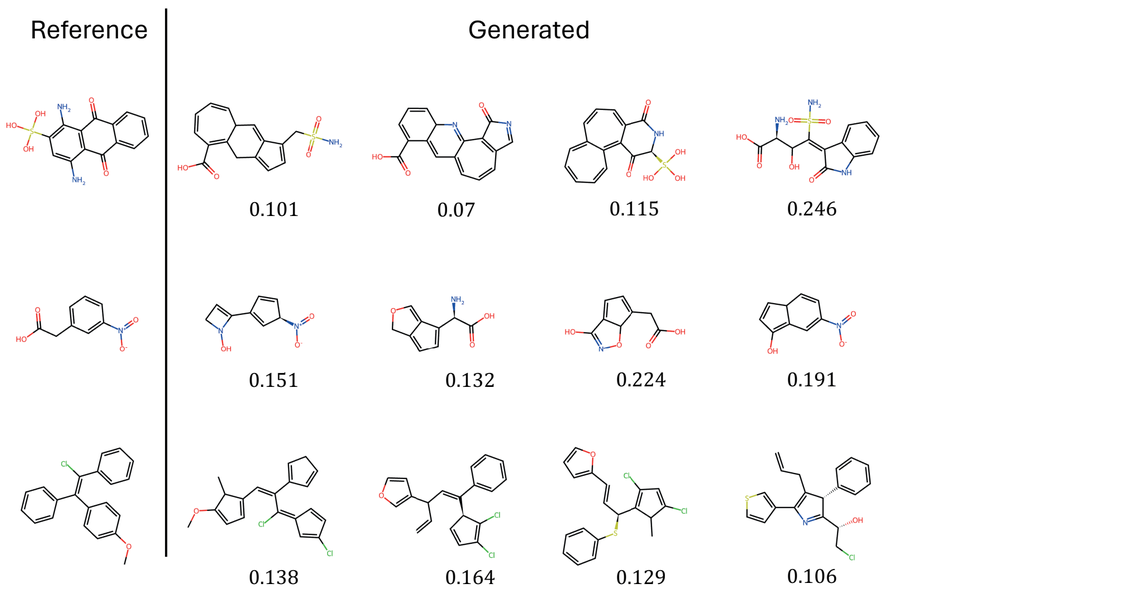


HIT EXPANSION

Docking performance on the CrossDocked2020 test set with ligands generated using PoLiGenX. We measure the docking score as proxy for binding affinity with QVina2. Other chemical properties like drug-likeness or the octanol-water partition coefficient are measured with RDKit.

Data	QVina2 (All) ↓	QVina2 (Top-10%) ↓	QED ↑	logP ↑	MolWt ↑	H-acceptors ↑	H-donors ↑	Lipinski ↑
CrossDocked test set	$-6.85_{\pm 2.33}$	-	$0.47_{\pm 0.20}$	0.79	0.85	0.84	0.8	$3.35_{\pm 1.14}$
PoLiGenX	$\mathbf{-7.21}_{\pm 2.22}$	$\mathbf{-8.04}_{\pm 2.44}$	$\mathbf{0.59}_{\pm 0.20}$	**0.91**	**0.87**	**0.85**	**0.91**	$\mathbf{3.57}_{\pm 0.93}$

Reference molecules extracted from the CrossDocked2020 test set (left) and four molecules sampled randomly with PoLiGenX. The chemical similarity between reference and sampled molecule are stated



CONTROLLING THE LATENT

Density plot for chemical similarity of ligands for varying control parameter each generated by PoLiGenX. The control parameter regulates the importance of the latent embedding