PILOTING STRUCTURE-BASED DRUG DESIGN VIA MODALITY-SPECIFIC OPTIMAL SCHEDULE

Keyue Qiu[1,2*] **Yuxuan Song**[1,2*] **Zhehuan Fan**[3] **Peidong Liu**[4] **Zhe Zhang**[2]
Mingyue Zheng[3] **Hao Zhou**[1†] & **Wei-Ying Ma**[1]
[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] Department of Computer Science and Technology, Tsinghua University
[3] Shanghai Institute of Materia Medica, Chinese Academy of Sciences
[4] Sichuan University
qky22@mails.tsinghua.edu.cn, {zhouhao,maweiying}@air.tsinghua.edu.cn

ABSTRACT

Structure-Based Drug Design (SBDD) is crucial for identifying bioactive molecules. Recent deep generative models are faced with challenges in geometric structure modeling. A major bottleneck lies in the twisted probability path of multi-modalities—continuous 3D positions and discrete 2D topologies—which jointly determine molecular geometries. By establishing the fact that noise schedules decide the Variational Lower Bound (VLB) for the twisted probability path, we propose VLB-Optimal Scheduling (VOS) strategy in this under-explored area, which optimizes VLB as a path integral for SBDD. Our model effectively enhances molecular geometries and interaction modeling, achieving state-of-the-art PoseBusters passing rate of 95.9% on CrossDock, more than 10% improvement upon strong baselines, while unlocking the potential of repurposing SBDD model as docking method without retraining, with 44.0% RMSD $<$ 2Å on PoseBusters V2.

1 INTRODUCTION

Structure-Based Drug Design (SBDD) plays a pivotal role in the discovery of bioactive molecules, leveraging the knowledge of protein-ligand interactions to identify potential therapeutic compounds (Isert et al., 2023). Despite its importance, achieving high-fidelity interaction modeling remains a significant challenge. Recent advances in geometric deep generative models have centered on non-autoregressive methods such as diffusion (Guan et al., 2023) and Bayesian Flow Network (BFN) (Qu et al., 2024), showing promise by capturing the structures at the global level. However, Fig. 1 shows that current generation strategies may inadequately capture the fine-grained molecular geometries.

One key factor in geometric structure modeling is the intertwined probability path of different modalities: continuous atom positions and discrete molecular topologies. While these modalities jointly determine molecular geometries and protein-ligand interaction types, there lacks a systematic understanding in designing the twisted probability path for SBDD. Prior works (Peng et al., 2023; Vignac et al., 2023; Guan et al., 2023) adopt sophisticated noise schedules adapted for different modalities, and propose to generate 3D positions first. In our preliminary studies (Sec. 2), we identify the potential problem of prioritizing the 3D modality, where the model cannot effectively utilize 2D topology information in generation, suggesting current noise schedule is suboptimal. Despite its importance, the design of optimal noise schedules remains largely under-explored.

To address the gap left by previous methods, we aim at a systematic solution for *the optimal schedule for the twisted probability path of different modalities*, and emphasize the need for principled metrics relevant to generative modeling in order to evaluate schedule quality. EquiFM (Song et al., 2024b) introduces improved hybrid probability paths aligned with information-theoretic heuristics but lacks a theoretical foundation for optimality. To rigorously define optimality, we analyze the Variational Lower Bound (VLB) under varying noise schedules. Crucially, we prove that in the multi-modality

*Equal contribution. Correspondence to Hao Zhou (zhouhao@air.tsinghua.edu.cn).

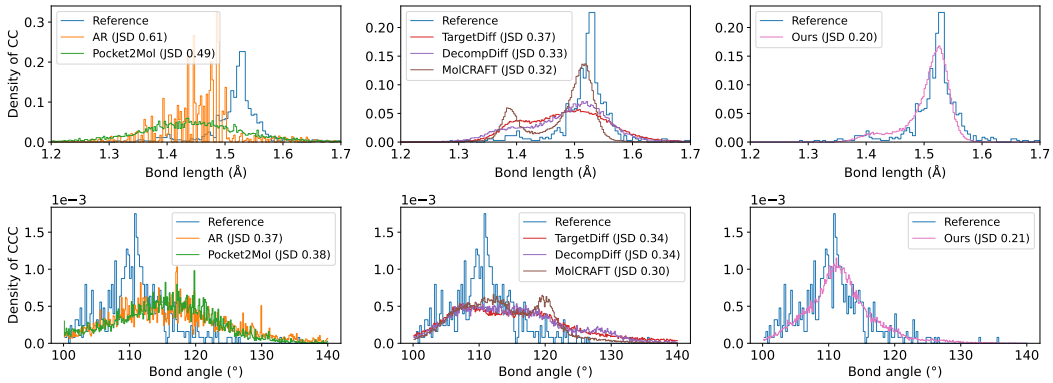


Figure 1: Visualization of the bond length and bond angle distribution for the most frequent bond types in CrossDock test. More types are shown in Appendix D.2, where our MolPilot consistently captures the molecular geometries the most accurately.

generation, the VLB becomes a path-dependent integral—its value depends on the entire noise schedule, not just endpoints (Kingma et al., 2021).

This motivates our **VLB-Optimal Scheduling** (VOS), eliminating heuristic noise schedule design and directly linking it to the theoretical guarantees of the VLB landscape. Specifically, we combat the *combinatorial complexity in schedule design space* when navigating and optimizing for VLB. To make it tractable to exhaust the possibilities, we demonstrate that the design space of multi-modality noise schedules can be reduced to a two-dimensional plane. Building on this insight, we develop a generalized objective, showing that decoupling timesteps during training allows a single model to implicitly cover this space, generalizing beyond fixed schedules. This framework enables efficient interpolation and extrapolation of schedules at inference time, bypassing costly retraining for new design constraints. By advancing the under-explored area of modality-specific optimal scheduling, we address key shortcomings in SBDD models such as strained conformations and suboptimal interactions, substantially improving the geometric structure modeling. Comprehensive evaluations demonstrate that our method achieves state-of-the-art conformation quality and binding affinities, with competitive docking accuracy.

Our contributions can be summarized as follows:

- We introduce VLB-Optimal Scheduling (VOS), a novel method for systematic noise schedule design in SBDD, achieving fine-grained control over multi-modality interdependence and improved interaction modeling.
- We establish the theoretical link between noise schedules and VLB in multi-modality probabilistic modeling. Unlike prior heuristic approaches, our principled method reveals path-dependent VLB dynamics and provides insights for modality-specific scheduling.
- Integrated with advanced frameworks, our proposed MolPilot achieves SOTA in de novo design with a remarkable PoseBusters passing rate of 95.9% on CrossDock, and competitive in local docking with 44.0% RMSD $<$ 2Å on PoseBusters.

2 Issues with Current Probability Path

In this section, we introduce the formulation of Structure-Based Drug Design (SBDD) and demonstrate that the default noise schedule in current generative models leads to a 3D-dominant probability path. Through preliminary experiments, we demonstrate that the model trained along this path cannot adequately capture the interdependence between 2D and 3D modalities, which motivates the need for an optimal probability path facilitated by corresponding noise schedule.

2.1 Problem Formulation

Structure-based Drug Design (SBDD) involves modeling the conditional probability $P(\mathbf{x}_M|\mathbf{x}_P)$, where $\mathbf{x}_M = (\boldsymbol{r}_M, \boldsymbol{h}_M, \boldsymbol{A}_M)$ represents the N-atom molecular geometry, and $\mathbf{x}_P = (\boldsymbol{r}_P, \boldsymbol{h}_P)$ represents the protein target. Here, $\boldsymbol{r} \in \mathbb{R}^{N\times 3}$ denotes continuous atom positions, $\boldsymbol{h} \in \mathbb{R}^{N\times K_h}$ encodes discrete atom types, and $\boldsymbol{A} \in \mathbb{R}^{N\times(N-1)\times K_A}$ encodes discrete bond types, with K_h and K_A denoting the number of atom types and bond types, respectively. In molecular generation, $\boldsymbol{r}$ from the continuous modality is usually described as the 3D geometry, and $(\boldsymbol{h}, \boldsymbol{A})$ from the discrete modality as the 2D topology of the molecular graph.

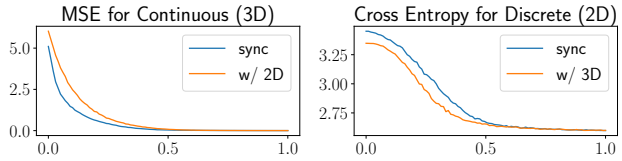


Figure 2: Validation loss curves on default schedule w.r.t. time in generation. *Sync:* Modalities at the same timestep. *w/ 2D:* Discrete modality at $t = 1$. *w/ 3D:* Continuous modality at $t = 1$.

2.2 3D-driven Probability Path

We visualize the modality-specific validation losses for a vanilla BFN (Qu et al., 2024) trained with default noise schedule (details of the background in Sec. 3). As shown in Fig. 2, the continuous modality loss decreases before the discrete modality, verifying that the twisted probability path is driven by 3D modality. However, the problem lies in that the model leverages cleaner 3D input to denoise the 2D topology effectively (*Right*), but fails the other way around (*Left*), as it performs worse when exposed to less noisy 2D input, suggesting that it is unable to benefit from cleaner 2D information. This highlights that the current noise schedule favors a twisted probability path driven by 3D modality, inducing a generative process with lower-noise 3D and higher-noise 2D data.

Intuitively, a well-balanced probability path should allow the models to leverage cleaner input from either modality to inform the generative process effectively, thereby accurately capturing the molecular geometries required for effective drug design. The problem of current twisted probability path is strengthened by the observation of inaccurate geometric structure modeling, where SBDD baselines exhibit abnormal bond lengths and angles (Fig. 7-9), suggesting inconsistency between modalities. We hypothesize that the distorted molecular geometry originates from the gap between modalities in the twisted probability path, which inspires us for an optimal scheduling.

We desire the optimal probability path defined by the optimal noise schedule to benefit from mutually informed modalities, with the potential to improve the generation of chemically valid and spatially accurate molecules, ultimately enhancing protein-ligand interaction modeling.

3 Background

Our work is built within the framework of Bayesian Flow Network (BFN) (Graves et al., 2023), the SOTA model for 3D molecular generation (Song et al., 2024a; Qu et al., 2024) that shows superiority over diffusion counterparts. In this section, we introduce the noise schedule and define the corresponding noising process, paving the way for Variational Lower Bound (VLB) analysis w.r.t. noise schedules.

Similar to diffusion, BFN involves a noising process q for data $\mathbf{x}$ to obtain a temporal sequence of observed latents $\mathbf{y}_{1:n} := \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, and optimizes the VLB:

$$\begin{aligned}\log p(\mathbf{x}) \geq \text{VLB} = \mathbb{E}_{q(\mathbf{y}_{1:n}|\mathbf{x})}[\log p(\mathbf{x} \mid \mathbf{y}_{1:n}) \\ - D_{\text{KL}}(q(\mathbf{y}_{1:n}|\mathbf{x}) \| p(\mathbf{y}_{1:n}))]\end{aligned} \tag{1}$$

where the noising process q is defined by the variational distribution $q(\mathbf{y}_{1:n} \mid \mathbf{x}) = \prod_{i=1}^{n} q(\mathbf{y}_i \mid \mathbf{x})$ that depends on the noise schedule over timesteps:

$$q(\mathbf{y} \mid \mathbf{x}) = \begin{cases} \mathcal{N}(\mathbf{x}, \alpha^{-1}\boldsymbol{I}), & \text{continuous data} \\ \mathcal{N}(\alpha(K\mathbf{e}_{\mathbf{x}} - \mathbf{1}), \alpha K\boldsymbol{I}), & \text{discrete one-hot} \end{cases} \tag{2}$$

The modality-specific noise schedules are defined as monotonically increasing differentiable functions over $t \in [0, 1]$:

$$\beta_c(t) = \sigma_1^{-2t} - 1, \quad \beta_d(t) = \beta_1 t^2, \tag{3}$$

where $\sigma_1 \in \mathbb{R}^+$ controls the noise for continuous atomic positions, and $\beta_1 \in \mathbb{R}^+$ is the hyperparameter for discrete topology. The noise level $\alpha(t) = \beta'(t) := d\beta(t)/dt$.

BFN differs from diffusions in the generative process, where instead of noisy latents $\mathbf{y}$, the network is informed by a lower-variance posterior $\boldsymbol{\theta}$ given the observed latents, from the Bayesian flow distribution of Gaussian or Dirac δ:

$$p_F(\boldsymbol{\theta} \mid \mathbf{x}; t) = \begin{cases} \mathcal{N}(\gamma\mathbf{x}, \gamma(1-\gamma)\boldsymbol{I}), & \text{continuous} \\ \delta(\boldsymbol{\theta} - \text{softmax}(\mathbf{y})), & \text{discrete} \end{cases} \tag{4}$$

where $\gamma := \frac{\beta_c(t)}{1+\beta_c(t)}$, $\mathbf{y}$ follows the discrete case in Eq. 2.

The network $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$[1] is trained to denoise $\mathbf{x}$ given $\boldsymbol{\theta}$ instead of $\mathbf{y}$, and thus constructs a receiver distribution $p_R(\mathbf{y} \mid \boldsymbol{\theta}; t) = q(\mathbf{y} \mid \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t))$. We show in Appendix B.1 that the single-modality VLB objective simplifies to:

$$\begin{aligned} \mathcal{L}_{\text{VLB}}(\mathbf{x}) &= -\mathbb{E}_{q(\mathbf{y}_{1:n}|\mathbf{x})} \log p(\mathbf{x} \mid \mathbf{y}_{1:n}) \\ &+ \sum_{i=1}^{n} D_{\text{KL}}(q(\mathbf{y}_i \mid \mathbf{x}) \parallel p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1})) \end{aligned} \tag{5}$$

and can be optimized by the continuous-time loss:

$$\begin{aligned} \mathcal{L}^{\infty}(\mathbf{x}) &= \mathop{\mathbb{E}}_{t\sim U(0,1), p_F(\boldsymbol{\theta}|\mathbf{x};t)} D_{\text{KL}}(q(\mathbf{y} \mid \mathbf{x}; t) \parallel p_R(\mathbf{y} \mid \boldsymbol{\theta}; t)) \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{t\sim U(0,1), p_F(\boldsymbol{\theta}|\mathbf{x};t)} \beta'(t) \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 \end{aligned} \tag{6}$$

where we redefine $\mathbf{x} := \sqrt{K}\mathbf{e}_{\mathbf{x}}$ for the discrete data for simplicity. Based on the monotonic and thus invertible function $u \equiv \beta(t)$, we perform a change of variables similarly to Kingma et al. (2021) as $t = \beta^{-1}(u)$, and rewrite the network as $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)$. Therefore, the loss in Eq. 6 is equivalent to the continuous-time loss expressed by noise schedule $\beta(t) : [0, 1] \to \mathbb{R}^+$:

$$\mathcal{L}^{\infty}(\mathbf{x}) = \frac{1}{2} \int_{\beta(0)}^{\beta(1)} \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};u)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)\|^2 du. \tag{7}$$

In the single-modality case, this states that the VLB is invariant to the shape of noise schedule functions except for the endpoints (Kingma et al., 2021). However, we will see in the following section that this invariance no longer holds for multi-modality cases, where the VLB becomes:

$$\begin{aligned} \mathcal{L}_{\text{VLB}}(\mathbf{x}) = \frac{1}{2} \mathbb{E}_{p_F(\boldsymbol{\theta}|\mathbf{x};t)} \int_0^1 [\beta_c'(t) \|\mathbf{x}_c - \tilde{\mathbf{x}}_{\phi,c}(\boldsymbol{\theta}, \beta(t))\|^2 \\ + \beta_d'(t) \|\mathbf{x}_d - \tilde{\mathbf{x}}_{\phi,d}(\boldsymbol{\theta}, \beta(t))\|^2] dt, \end{aligned} \tag{8}$$

which is simply summed along each modality due to the factorized nature of q, yet this results in path-dependent VLB which is closely related to the design of joint noise schedule $\beta(t) : [0, 1] \to (\mathbb{R}^+)^2$.

4 METHODOLOGY

We propose VLB-Optimal Scheduling (VOS), a principled methodology for optimal probability path in SBDD, by analyzing and identifying the optimal noise schedule for both modalities.

[1]The actual noise prediction model is specified as $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t, \mathbf{x}_P)$ with the protein structure input. We omit $\mathbf{x}_P$ for simplicity.

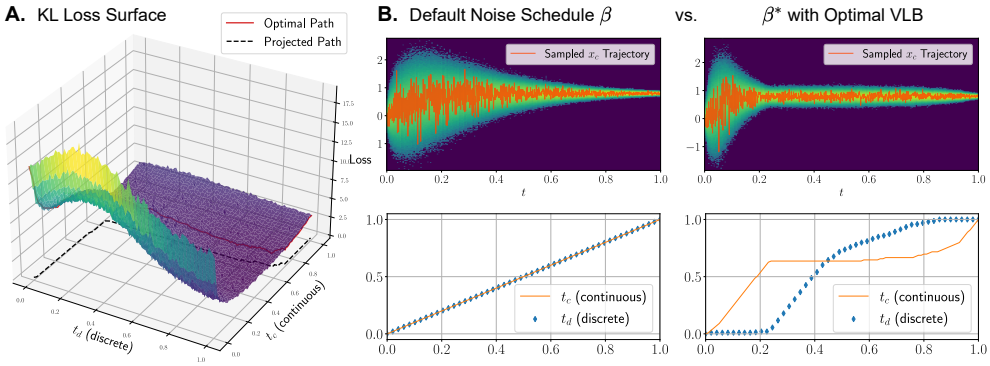


Figure 3: Our proposed VLB-Optimal Scheduling (VOS) that works by estimating the loss landscape and deriving an optimal noise schedule. **A.** Visualization of the loss surface over the function space. **B.** *Upper:* Visualization of the probability path of continuous data $\mathbf{x}_c$. *Lower:* The equivalent time-rescaling functions $t_c \equiv f(t), t_d \equiv g(t)$.

Overview (1) In Sec. 4.1, we theoretically analyze the path-dependent VLB for a given twisted probability path induced by specific noise schedule in that function space. (2) In Sec. 4.2, we formalize the entire function space of possible noise schedules, and propose to navigate the space by time rescaling. (3) In Sec. 4.3, we propose a path-invariant generalized loss objective that enables efficient estimation of the VLB landscape over the entire function space of noise schedules. Then, we describe the method to search for the optimal schedule on the landscape that maximizes the VLB.

4.1 Path-dependent VLB for Joint Noise Schedule

In this section, we establish a key result for the twisted probability path, where the joint noise schedule $\beta(t) : [0, 1] \to (\mathbb{R}^+)^2$ induces path-dependent VLB.

A key foundation in single modality is that the continuous-time loss, i.e., VLB, remains invariant to the shape of noise schedule function $\beta(t)$, except for the endpoints $\beta(0), \beta(1)$ (Kingma et al., 2021), allowing the design of different schedules for efficient training. As a natural extension, we generalize this invariance to multi-modalities given $\beta_c \neq \beta_d$:

$$\dot{\mathcal{L}}^{\infty}(\mathbf{x}) = \frac{1}{2} \int_{\beta_c(0)}^{\beta_c(1)} \int_{\beta_d(0)}^{\beta_d(1)} \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};\beta)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)\|^2 d\beta. \tag{9}$$

This equation shows that the generalized loss is invariant to the decoupled schedules as a surface integral over the plane $[\beta_c(0), \beta_c(1)] \times [\beta_d(0), \beta_d(1)]$.

However, this generalized loss $\dot{L}$ no longer corresponds to the objective of generative modeling, where the VLB should be a path integral along a line on such plane:

$$\mathcal{L}^{\infty}(\mathbf{x}) = \frac{1}{2} \int_{\beta_c, \beta_d} \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};\beta)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)\|^2 d\beta. \tag{10}$$

This curve is equivalent to a specific noise schedule β in the possible design space $\mathcal{Z}$ (formalized in Sec. 4.2). Therefore, we draw the conclusion that the VLB varies for different coupling even with the same endpoints $\beta(0) = (\beta_c(0), \beta_d(0))$ and $\beta(1) = (\beta_c(1), \beta_d(1))$, and it is no longer agnostic to the intermediate trajectory of $\beta(t)$, which is essentially different from the generative modeling within single modality.

The fundamental challenge, then, lies in identifying an optimal joint schedule β^* in the design space $\mathcal{Z}$ with the best possible VLB integrated along that path:

$$\beta^* = \arg\min_{\beta \in \mathcal{Z}} \int_{\beta} \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};\beta)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)\|^2 d\beta. \tag{11}$$

4.2 DESIGN SPACE OF JOINT NOISE SCHEDULES

To formalize the design space of noise schedules, we define the space of monotonically increasing functions $\mathcal{X} : [0, 1] \to \mathbb{R}^+$, thus the coupled function space $\mathcal{Z}$ is:

$$\{\beta(t) = (\beta_c(t), \beta_d(t)) \mid \forall \beta_c, \beta_d \in \mathcal{X}, \beta_c' > 0, \beta_d' > 0\}, \tag{12}$$

where $\beta_c(t)$ and $\beta_d(t)$ are monotonic schedules for the continuous and discrete modalities, respectively.

To explore arbitrary schedule configurations, we introduce time-rescaling functions $t_c \equiv f(t)$, $t_d \equiv g(t)$, allowing for arbitrary β expressed in predefined $\tilde{\beta}_c, \tilde{\beta}_d$ as in Eq. 3 and varying forms of implicit functions f and g by:

$$\beta(t) = (\tilde{\beta}_c(f(t)), \tilde{\beta}_d(g(t))) \tag{13}$$

for which we have the following theorem that guarantees this general form of β expressed in terms of t_c, t_d is sufficient to cover the entire function space $\mathcal{Z}$.

Theorem 4.1. *Suppose we have a monotonic function $\tilde{\beta}_m(t) : [0, 1] \to \mathbb{R}^+$, and let $\beta_m(t)$ be any such monotonic function. Then there exists a time-rescaling function $t_m \equiv f(t)$ such that $\beta_m(t) = \tilde{\beta}_m(t_m)$. In fact, $f(t) = \tilde{\beta}_m^{-1}(\beta_m(t))$ has the same monotonicity as $\beta_m(t), \tilde{\beta}_m(t)$.*

Remark 4.2. It follows from the monotonicity of the schedule functions that we can obtain arbitrary combination of noise levels simply by setting $t_c, t_d \in [0, 1]$ separately.

4.3 NAVIGATING FOR THE VLB-OPTIMAL SCHEDULE

The first obstacle in identifying the optimal β^* is to obtain a model $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)$ that can be evaluated over all possible $\beta \in \mathcal{Z}$. We facilitate the VLB analysis of different joint noise schedules by training $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$ to minimize the generalized loss $\dot{\mathcal{L}}$ described in Eq. 9 through the change of variables $t_c \equiv \beta_c^{-1}(t)$, $t_d \equiv \beta_d^{-1}(t)$, thereby the objective becomes:

$$\dot{\mathcal{L}}^\infty(\mathbf{x}) = \frac{1}{2} \int_0^1 \int_0^1 \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};t)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 dt_c dt_d, \tag{14}$$

for which the proposition holds (proof in Appendix B.4):

Proposition 4.3. *Suppose we have a model $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$ trained by Eq. 14, and let $\beta_c(t), \beta_d(t)$ be any monotonically increasing functions in $\mathcal{X}$. Then the line integral in Eq. 10 corresponds to the negative VLB for $\beta(t) = (\beta_c(t), \beta_d(t))$.*

We reframe finding the VLB-optimal β^* as a search problem for the solution with the minimal cumulative cost along the discretized N-step trajectory $\{t_c, t_d\}$ from $[0, 0]$ to $[1, 1]$, which can be solved by dynamic programming (Appendix A.1).

We visualize the derived optimal schedule in Fig. 3B. Intuitively, this time rescaling corresponds to a two-stage probability path: (1) *Shape-driven sketching:* First, the generation predominantly focuses on generating the continuous atom positions and largely ignores the discrete topology. Then, when $t \in [0.3, 0.8]$, the model starts to fit a possible 2D molecular graph into the rough shape of fixed 3D conformation. (2) *Topology-driven docking:* at the last stage $t > 0.8$, the generation enters into the docking stage, altering the conformation according to the discrete molecular topologies. By facilitating such a principled probability path, the optimal noise schedule effectively harnesses different modalities, achieving the best VLB and sample quality.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We conduct two main experiments in the broader scope of SBDD: (1) de novo design on CrossDock (Francoeur et al., 2020), and (2) molecular docking on PoseBusters V2 (Buttenschoen et al., 2024). The same model checkpoint trained with generalized loss is evaluated throughout both main experiments. Detailed setups are in Appendix C.1, with more evaluation results in Appendix D.

Table 1: Performance on CrossDock, where MolPilot shows superior results. ♡: results cited from Qu et al. (2024). †: results calculated by us using the official samples. Top-2 results are highlighted in **bold** and underlined, respectively.

Methods	PB-Valid†	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		scRMSD	Energy	Connected†	QED	SA	Div	Size
	Avg. (↑)	Avg.	Med.	Avg.	Med.	Avg.	Med.	<2 Å (↑)	Passed† (↑)	Avg. (↑)	Avg.	Avg.	Avg.	Avg.
CrossDock♡	95.0%	-6.36	-6.46	-6.71	-6.49	-7.45	-7.26	34.0%	98.0%	-	0.48	0.73	-	22.8
AR♡	59.0%	-5.75	-5.64	-6.18	-5.88	-6.75	-6.62	36.5%	84.9%	93.5%	0.51	0.63	0.70	17.7
Pocket2Mol♡	72.3%	-5.14	-4.70	-6.42	-5.82	-7.15	-6.79	32.0%	97.3%	96.3%	0.57	0.76	0.69	17.7
FLAG♡	16.0%	45.85	36.52	9.71	-2.43	-4.84	-5.56	0.3%	83.4%	97.1%	0.61	0.63	0.70	16.7
TargetDiff♡	50.5%	-5.47	-6.30	-6.64	-6.83	-7.80	-7.91	37.1%	69.8%	90.4%	0.48	0.58	0.72	24.2
DiffSBDD†	37.6%	-1.44	-4.91	-4.52	-5.84	-7.14	-7.3	18.7%	74.0%	93.2%	0.47	0.58	0.73	24.4
DecompDiff♡	71.7%	-5.19	-5.27	-6.03	-6.00	-7.03	-7.16	24.2%	84.1%	82.9%	0.51	0.66	0.73	21.2
MolCRAFT♡	84.6%	-6.55	-6.95	-7.21	-7.14	-7.67	-7.82	**46.8%**	91.1%	96.7%	0.50	0.67	0.72	22.7
MolPilot (Ours)	**95.9%**	**-6.88**	**-7.03**	**-7.23**	**-7.27**	**-7.92**	**-7.92**	41.1%	**98.5%**	**97.9%**	0.56	0.74	0.69	22.6

Dataset. Following SBDD conventions (Luo et al., 2021), we adopt the same split of CrossDock (Francoeur et al., 2020) to train and validate our model, which consists of 100,000 training poses and 100 test proteins for de novo design. For molecular docking, after removing 10 proteins that cannot be processed for non-standard residues from the PoseBusters V2, we test on the remaining 298 protein-ligand complexes as ground-truth-based evaluation.

Baselines. We consider the following SBDD baselines: (1) Autoregressive models including AR (Luo et al., 2021), Pocket2Mol (Peng et al., 2022), FLAG (Zhang et al., 2023), (2) Diffusion-based TargetDiff (Guan et al., 2022), DiffSBDD (Schneuing et al., 2022), DecompDiff (Guan et al., 2023), and (3) BFN-based MolCRAFT (Qu et al., 2024).

Metrics. We employ the following metrics: (1) For conformation quality, we report PoseBusters passing rate (**PB-Valid**), and **Strain Energy Passed** denotes the ratio of molecules passing PoseBusters internal energy check at the global level, and Jensen-Shannon divergence (**JSD**) between reference and generated distributions of bond length, bond angle and torsion angle at the substructure level. (2) For interaction modeling, we emphasize **RMSD**, the root-mean-squared distance between generated and ground truth pose in docking. We also report the self-consistency RMSD (**scRMSD**) in de novo design calculated between generated and Vina Dock pose as an indicator of binding mode consistency upon redocking. Affinity metrics are calculated by AutoDock Vina (Eberhardt et al., 2021), including binding affinities where **Vina Score** directly scores the generated molecular pose in the pocket, **Vina Min** quickly optimizes the pose in-place and scores the minimized pose, and **Vina Dock** exhaustively searches for optimal pose to obtain the lowest energy. (3) For molecular properties, we report drug-likeness (**QED**) and synthetic accessibility (**SA**), which are desired to fall within reasonable ranges. We additionally report the percentage of fully connected molecules (**Connected**) to show the performance of generation successes, together with the average number of atoms of those successfully generated molecules (**Size**).

5.2 De novo Design

We report the main results in Table 1 by sampling 100 molecules for each of the test proteins.

MolPilot achieves the most accurate molecular geometries. For intramolecular validity, we achieve the highest passing rate of internal energy, demonstrating our conformation stability. We additionally report the bond length, angle, and torsion angle distributions in Fig. 7, 8, 9, further underscoring its superiority.

MolPilot generates the best binding poses. For protein-ligand intermolecular validity, we excel at the highest PB-Valid of 95.9%, ensuring reasonable binding poses. For binding affinities, our method outperforms all other models with an average Vina Score of -6.88 kcal/mol.

5.3 Molecular Docking

We note that the model trained with $\dot{\mathcal{L}}$ in Eq. 9 is also available for multimarginal generative modeling (Campbell et al., 2024), i.e., it can be used for molecular docking by an induced $P(\boldsymbol{r} \mid \{\boldsymbol{h}, \boldsymbol{A}\}, \mathbf{x}_P)$.

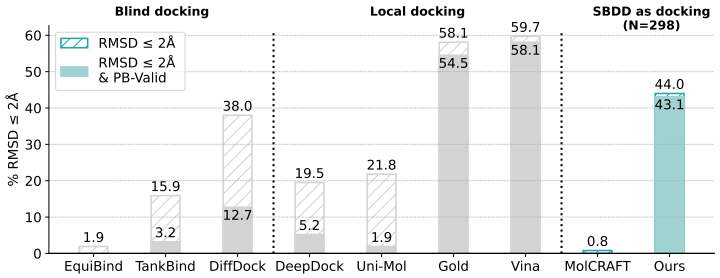


Figure 4: PoseBusters V2 structural accuracy (RMSD) and validity (PB-Valid) for docking methods (in gray, cited from Abramson et al. (2024)) and SBDD methods (in blue, calculated by us).

For a deeper understanding of our model's ability in capturing genuine spatial interaction, we additionally evaluate its docking accuracy on PoseBusters V2 dataset with holo pocket residue structures specified, and report the RMSD results. While we did not specifically optimize our design w.r.t. molecular docking, Fig. 4 suggests that our method is a competitive SBDD model even compared with the local docking methods. This capability in recovering ground truth binding poses serves as an indicator of capturing true interactions, which is essential for designing bioactive molecules with potential biological efficacy. We believe that incorporating molecular docking as a subtask of SBDD also sheds light on the model's ability of interaction modeling.

Table 2: Ablation study of the optimal noise schedule. The same model trained with generalized loss (Eq. 9) is evaluated given different noise schedules β for test time only.

Schedule	Vina Score	Vina Min	QED	SA	PB-Valid	Connect
Default	-6.16	-6.59	0.49	0.72	0.95	0.98
Optimal	**-7.02**	**-7.29**	**0.53**	**0.75**	**0.99**	0.98

5.4 Ablation

We report the ablation study results in Table 2. When training under the generalized objective $\dot{\mathcal{L}}$ and sampling under different noise schedules β, it can be seen that optimal schedule β^* consistently achieves better performance than default noise schedule.

6 Related Works

Structure-based Drug Design (SBDD) SBDD generative models focus on the joint generation of discrete molecular topology and continuous conformation conditioned on protein-binding pockets. Recent approaches have centered on capturing the critical protein-ligand interactions. Autoregressive models (Luo et al., 2021; Peng et al., 2022; Liu et al., 2022) generate molecules atom-by-atom while preserving geometric equivariance but are computationally expensive. Fragment-based methods (Powers et al., 2022; Zhang et al., 2023; Lin et al.) improve efficiency by generating motifs instead of atoms, though they often require post-processing to mitigate error accumulation. Non-autoregressive models such as diffusion-based approaches (Schneuing et al., 2022; Guan et al., 2022; 2023) and Bayesian Flow Networks (BFNs) (Qu et al., 2024) focus on full-atom generation, enabling scalability and improved controllability. While significant progress has been made, challenges in conformation quality and interaction modeling remain central to advancing structure-based drug design (Harris et al., 2023).

Molecular Docking In the context of SBDD, molecular docking concerns predicting the 3D conformation given 2D molecular topology and the protein pocket structures. Traditional search-based local-docking approaches include AutoDock Vina (Eberhardt et al., 2021), Gold (Jones et al., 1997). Deep learning-based methods are divided into blind docking methods with holo protein structures specified, such as DiffDock (Corso et al., 2023), and local docking methods such as DeepDock (Méndez-Lucio et al., 2021) and Uni-Mol (Zhou et al., 2023). These methods additionally rely on the RDKit-initialized molecular conformation as input.

7 CONCLUSION

This work addresses a critical challenge in Structure-Based Drug Design (SBDD) by introducing a better-balanced dynamics in the generative process of multi-modalities. Equipped with an optimal noise schedule in terms of Variational Lower Bound (VLB), MolPilot achieves a SOTA PoseBusters passing rate of 95.9% on CrossDock and unlocks the potential of repurposing SBDD model as docking method, offering a promising step forward in the field of drug discovery.

ACKNOWLEDGMENTS

This work is supported by the Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5453–5512. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/campbell24a.html`.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)*, 2023.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11827–11846. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/guan23a.html`.

Charles Harris, Kieran Didi, Arian R Jamasb, Chaitanya K Joshi, Simon V Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models? *arXiv preprint arXiv:2308.07413*, 2023.

Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, April 2023. ISSN 0959440X. doi: 10.1016/j.sbi.2023.102548. URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X23000222.

Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology*, 267(3): 727–748, 1997.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Haitao Lin, Yufei Huang, Odin Zhang, Yunfan Liu, Lirong Wu, Siyuan Li, Zhiyuan Chen, and Stan Z. Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, pp. 34603–34626. Curran Associates, Inc.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. In *International Conference on Machine Learning*, 2022.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3D Generative Model for Structure-Based Drug Design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021. URL http://arxiv.org/abs/2203.10446.

Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/peng22b.html.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. In *International Conference on Machine Learning*, pp. 27611–27629. PMLR, 2023.

Alexander S Powers, Helen H. Yu, Patricia Adriana Suriana, and Ron O. Dror. Fragment-based ligand generation guided by geometric deep learning on protein-ligand structures. In *ICLR2022 Machine Learning for Drug Discovery*, 2022. URL https://openreview.net/forum?id=192L9cr-8HU.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Molcraft: Structure-based drug design in continuous parameter space. *arXiv preprint arXiv:2404.12141*, 2024.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based Drug Design with Equivariant Diffusion Models, October 2022. URL http://arxiv.org/abs/2210.13695. arXiv:2210.13695 [cs, q-bio].

Yuxuan Song, Jingjing Gong, Yanru Qu, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3d molecules via bayesian flow networks. *arXiv preprint arXiv:2403.15441*, 2024a.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation, 2023. URL https://arxiv.org/abs/2302.09048.

Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule Generation For Target Protein Binding with Structural Motifs. 2023.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Algorithm 1: Deriving Optimal Schedule

Input: Multi-modality data $\mathbf{x}$, default noise schedule $\beta = (\tilde{\beta}_c, \tilde{\beta}_d)$, grid resolution M, step N, step scale K.
Output: Optimal schedule β^*, generative model $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$.
Train $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$ to minimize the generalized loss $\dot{\mathcal{L}}$ (Eq. 14);
$\dot{\mathcal{L}}^\infty(\mathbf{x}) = \frac{1}{2}\int_0^1\int_0^1 \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};t)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 dt_c dt_d$
Discretized grid $g = (t_c, t_d)_{M\times M} \leftarrow [\frac{1}{M}, \dots, 1]^2$;
foreach (t_c, t_d) *in* g **do**
  Compute cost $C(t_c, t_d) \leftarrow \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2$
Optimal path $\{t_c^* \equiv f(t), t_d^* \equiv g(t)\}, J^* \leftarrow \mathrm{DP}(C, N, K)$;
Optimal schedule $\beta^*(t) \leftarrow (\tilde{\beta}_c(t_c^*), \tilde{\beta}_d(t_d^*))$;

A Details about the Proposed Method

A.1 Dynamic Programming for Optimal Schedules

We present the method to navigate the function space of arbitrary β. Note that from Remark 4.2, we can discretize the function space through discretized combinations of t_c, t_d. Therefore, we reframe finding the VLB-optimal β^* as a search problem for the solution with the minimal cumulative cost along the discretized N-step trajectory $\{t_c, t_d\}$ from $[0, 0]$ to $[1, 1]$.

Depicted in Fig. 3, we estimate the cost matrix across a grid of possible t_c and t_d values by evaluating the KL divergence over a batch of samples given the $\dot{\mathcal{L}}$-trained model $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)$, and fit a smooth loss surface using B-spline interpolation.

Definition A.1 (Cost)**.** The cost at a given noise level t_c, t_d as implicit functions over t is defined as:

$$C(t_c, t_d) = \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, \beta)\|^2 \tag{15}$$

We employ dynamic programming to solve the search problem for a minimal cumulative cost J:

$$J(t_c, t_d) = \min_{(\epsilon_c, \epsilon_d)} \big(J(t_c - \epsilon_c, t_d - \epsilon_d) + \alpha C(t_c, t_d)\big). \tag{16}$$

where $\alpha := \frac{\beta(t) - \beta(t-\epsilon)}{\epsilon}$ is the discretized noise level, and the valid ranges of (ϵ_c, ϵ_d) are approximated by the gradient of the smooth surface. It is guaranteed that this yields the VLB-optimal schedule:

Remark A.2. $J(1, 1)$ corresponds to an unbiased Monte-Carlo estimate of the optimal VLB among all generative models trained by arbitrary $\beta \in \mathcal{Z}$, and backtracking the path $\{t_c = f(t), t_d = g(t)\}$ from $[1, 1]$ to $[0, 0]$ yields the optimal schedule $\beta^*(t) = (\tilde{\beta}_c(t_c), \tilde{\beta}_d(t_d))$.

Algorithm 2: Dynamic Programming for Optimal Path

Input: Cost matrix $C \in \mathbb{R}^{M \times M \times 2}$, step budget L, step scale K, default noise schedule $\beta(t)$
Output: Optimal $path$, minimal cumulative cost J^*
Initialization:
Fit a smooth loss surface $\tilde{C}$ using B-spline interpolation.;
$J[: M, : M, : N+1] \leftarrow \infty$, $J[0,0,0] \leftarrow 0$
prev$[: M, : M, : N+1, : 2] \leftarrow -1$
active_states $\leftarrow \{(0,0,0)\}$
size $\leftarrow \frac{K}{1+\|\nabla \tilde{C}\|}$, where $\|\nabla \tilde{C}\| = \sqrt{(\frac{\partial \tilde{C}}{\partial x})^2 + (\frac{\partial \tilde{C}}{\partial y})^2}$
Dynamic Programming:
for $l \leftarrow 0$ ***to*** $L-1$ **do**
  states $\leftarrow \emptyset$
  foreach $(x, y, l) \in$ *active_states* **do**
    foreach $(\epsilon_x, \epsilon_y) \in$ *valid_steps*(*size*$[x, y]$) **do**
      $x_t, y_t \leftarrow x + \epsilon_x, y + \epsilon_y$
      $\alpha \leftarrow [\frac{\beta_c(x_t) - \beta_c(x)}{\epsilon_x}, \frac{\beta_d(y_t) - \beta_d(y)}{\epsilon_y}]$
      $cost \leftarrow J[x, y, l] + \alpha \cdot C[x_t, y_t]$
      if $J[x_t, y_t, l+1] > cost$ **then**
        $J[x_t, y_t, l+1] \leftarrow cost$
        prev$[x_t, y_t, l+1] \leftarrow (x, y)$
        states $\leftarrow$ states $\cup \{(x_t, y_t, l+1)\}$
  active_states $\leftarrow$ states
$x, y, path \leftarrow M-1, M-1, \emptyset$
$l \leftarrow \arg\min J[x, y, :]$
while $l \geq 0$ **do**
  $path$.append$((x, y))$
  $(x, y) \leftarrow$ prev$[x, y, l]$, $l \leftarrow l - 1$
return $path, \min J[M-1, M-1, :]$

Convergence guarantee As the grid resolution $N \to \infty$, the discrete solution converges to the continuous VLB-optimal path $\beta^*(t)$. Since the VLB for a joint schedule $\beta(t)$ is a line integral over the model's loss field in the 2D noise space (Appendix B.4), and training with the generalized loss $\dot{\mathcal{L}}^\infty$ ensures the model's predictions are accurate everywhere in this space, enabling the evaluation of VLBs for arbitrary paths $\beta(t)$ (Appendix B.5), the dynamic programming solution recovers the optimal path asymptotically, maximizing the VLB over the function space $\mathcal{Z}$.

Time complexity The computational complexity is $O(NM^2K)$, which does not add too much computational overhead and can be solved within a few minutes on a single CPU. Once solved, the optimal schedule can be adopted for test time, enabling the VLB-optimal generative process, where the multi-modalities effectively inform each other at the optimal accuracy level. However, in order to estimate the cost matrix over $M \times M$ discretized grid requires the same number of evaluations of the generative model $\tilde{\mathbf{x}}_\phi$, which scales quadratically to the discretization M and involves considerable GPU computation. Empirically, we found that $M = 20$ would yield an accurate enough estimate, taking less than 20 minutes.

A.2 Why Learning for Optimal Schedule Might Fail

Given the obtained optimal schedule $\beta^*(t)$, we can train the generative model with invariant optimal VLB (Eq. 10) more effectively, allocating the model capacity towards learning to denoise only along the selected path. We found this results in earlier convergence, taking around half the previous training time for generalized objective (Eq. 9) with similar performance.

We would like to take one step further to see if we can directly optimize the joint noise schedule $\beta(t)$ in generative training with the negative VLB objective in Eq. 10, i.e., optimizing the generative model and the noise schedule simultaneously through the same loss objective.

We parametrize the learnable time-rescaling function by

$$f(t) = t + (1-t)t \cdot \text{sigmoid}(h_\psi(t)) \tag{17}$$

where $h_\psi(x)$ is a monotonic neural network which we choose to be a three-layered MLP with Softplus activation before the final output. This rescaling function is strictly monotonically increasing with endpoint constraints, i.e., $f(0) = 0, f(1) = 1$.

However, we found that the learned noise schedule does not change much from its initialized shape, and the performance proves to be suboptimal. We hypothesize that there are several challenges in learning for the optimal schedule during training: (1) Large combinatorial space with limited exposure to possible noise schedules in the process of active learning. Unlike single-modality models, where the range of noise levels is easily explored within the modality (Dieleman et al., 2022), multi-modality generative models can suffer from exposure to only a subset of the possible noise combinations. This incomplete exploration makes it challenging to optimize directly for the best schedule, as large regions of the design space remain unvisited. It impacts the VLB estimation since the model $\tilde{\mathbf{x}}_\phi$ is still learning to adapt to the scaled noisy inputs during training, and is not invariant w.r.t. noisy input at different scales. (2) Optimization difficulties. Directing taking the gradient w.r.t. t_c, t_d introduces numerical instability, particularly when $\tilde{\beta}_c$ involves exponentiation, which can grow or shrink very rapidly for certain values of t_c, leading to vanishing or exploding gradients. Moreover, the VLB optimization is sensitive to the trajectory defined by the noise schedule, and the optimization landscape of may have many local minima, especially due to the coupling of noise schedules, which can make it difficult to find the global optimum.

Addressing these might require more sophisticated methods such as bilevel optimization, which we leave for future work.

A.3 Model Architecture

We employ an equivariant Graph Transformer architecture (Vignac et al., 2023), with heterogeneous message passing between protein-ligand atoms similar to (Guan et al., 2023).

Denoting node positions as $\mathbf{r} \in \mathbb{R}^{N\times 3}$, one-hot atom types as $\mathbf{h} \in \mathbb{R}^{N\times K_h}$, and one-hot bond types as $\mathbf{A} \in \mathbb{R}^{N\times(N-1)\times K_A}$, where K_h and K_A are the numbers of atom and bond types, respectively.

Protein-Ligand Message Passing: For interactions between a protein node i and a ligand node j, we define:

$$\mathbf{K} = \boldsymbol{X}\boldsymbol{W}_k, \mathbf{V} = \boldsymbol{X}\boldsymbol{W}_v, \mathbf{Q} = \boldsymbol{X}_q\boldsymbol{W}_q.$$

where the input features $\boldsymbol{X} = \text{concat}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_{\text{edge}})$, $\boldsymbol{X}_q = \text{concat}(\mathbf{h}_i, \mathbf{h}_{\text{edge}})$. The attention scores are computed using:

$$\alpha_{ij} = \frac{(\mathbf{q}_i \odot \mathbf{k}_j \odot \mathbf{e}_{ij})}{\sqrt{d_h}}, \quad \mathbf{e}_{ij} = \tanh(\mathbf{W}_{\text{edge}}\mathbf{d}_{ij}),$$

where d_h is the head dimension, and $\mathbf{e}_{ij}$ is the edge-specific weights modulating the contribution of distant neighbors, $\mathbf{d}_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$ is the pairwise distance. $\mathbf{h}_{\text{edge}}$ describes the edge types between protein-ligand atoms.

Intra-Ligand Message Passing: For interactions within the same ligand (between atoms i and j), denoting the flattened bond matrix as $\mathbf{h}_{\text{bond}}$, the input features for message passing is defined as $\boldsymbol{X} = \text{concat}(\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_{\text{bond}}, \mathbf{r}_{\text{feat}}, \mathbf{a}_{\text{feat}})$, $\boldsymbol{X}_q = \text{concat}(\mathbf{h}_i, \mathbf{h}_{\text{bond}})$. Here, attention is computed in a similar fashion:

$$\alpha_{ij}^{\text{ligand}} = \frac{(\mathbf{q}_i \odot \mathbf{k}_j \odot \mathbf{e}_{ij}^{\text{ligand}})}{\sqrt{d_h}}, \quad \mathbf{e}_{ij}^{\text{ligand}} = \tanh(\mathbf{W}_{\text{edge}}\mathbf{r}_{\text{feat}}).$$

where $\mathbf{r}_{\text{feat}}$ encodes the pairwise distance through Gaussian smearing, and $\mathbf{a}_{\text{feat}}$ is the angular encoding for the bond triplets.

Message Update The aggregated message for each node i is computed as:

$$\mathbf{m}_{ij} = \sum_{k\in\mathcal{N}(j)} \alpha_{kji}\mathbf{v}_j,$$

where $\mathcal{N}(j)$ denotes the set of neighbors of node j.

Feature and Position Updates Node features at layer l are updated as:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}_{\text{out}}^{(l)}\left(\mathbf{h}_i^{(l)} \| \sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}\right)\right),$$

where σ is a non-linear activation function, and $\|$ denotes concatenation.

The bond feature $\mathbf{h}_{\text{bond}}^{(l+1)}$ is updated as:

$$\mathbf{h}_{\text{bond},ij}^{(l+1)} = \sigma\left(\mathbf{W}_{\text{bond}}^{(l)}\left(\mathbf{h}_{\text{bond},ij}^{(l)} \| \mathbf{m}_{ij}^{\text{ligand}}\right)\right).$$

The spatial positions are updated to ensure equivariance:

$$\mathbf{r}_i^{(l+1)} = \mathbf{r}_i^{(l)} + \Delta\mathbf{r}_i^{(l)},$$

where the update $\Delta\mathbf{r}_i^{(l)}$ is computed as a function of the aggregated message and neighboring bond features:

$$\Delta\mathbf{r}_i^{(l)} = \sum_{j\in\mathcal{N}_{\text{ligand}}(i)} (\mathbf{r}_i^{(l)} - \mathbf{r}_j^{(l)})\mathbf{m}_{ij}^{\text{ligand}} + \sum_{k\in\mathcal{N}_{\text{complex}}(i)} (\mathbf{r}_i^{(l)} - \mathbf{r}_k^{(l)})\mathbf{m}_{ik}^{\text{complex}}$$

where $\mathbf{m}_{ij}^{\text{ligand}}$, $\mathbf{m}_{ik}^{\text{complex}}$ are the aggregated message under the similar message passing mechanism to protein-ligand and intra-ligand, respectively, given the updated node features $\mathbf{h}_{ij}^{(l+1)}$ and bond features $\mathbf{h}_{\text{bond},ij}^{(l+1)}$.

A.4 Implementation Details

Training and Inference We use Adam optimizer with learning rate 5e-4, batch size of 16, and fit the model with 3.1 million parameters on one NVIDIA 80GB A100 GPU. We set $\beta_1 = 1.5$ for discrete atom types and bond types, $\sigma_1 = 0.05$ for atom coordinates. The training converges in 200K steps (around 24 hours). For inference, we use the exponential moving average of the weights from training that is updated at every optimization step with a decay factor of 0.999. We run inference with 100 sampling steps with the same variance reduction sampling strategy as Qu et al. (2024).

Hyperparameters for Network We set the network to be kNN graphs with $k = 32$, $N = 9$ layers with $d = 128$ hidden dimension, 16-headed attention, and dropout rate 0.1.

Featurization We describe our atom-level and edge-level featurization for the protein-ligand complexes. At the atom-level, each protein atom is represented with a one-hot element indicator (H, C, N, O, S, Se), a 20-dimensional one-hot amino acid type indicator, and a 1-dimensional backbone flag. Ligand atoms are featurized with a one-hot element indicator (C, N, O, F, P, S, Cl) coupled with an aromaticity flag. For edge-level featurzation in the heterogeneous protein-ligand graph, edge types are encoded as a 4-dimensional one-hot vector indicating whether the edge is between ligand atoms, protein atoms, ligand-protein atoms, or protein-ligand atoms. For the ligand graph, bonds are represented with a 4-dimensional one-hot bond type vector (non-bond, single, double, triple), where aromatic bonds are kekulized using RDKit.

B Proof

B.1 Derivation of Eq. 6 as single-modality VLB (Eq. 5)

We begin with the negative Variational Lower Bound (VLB) in Eq. 5:

$$\mathcal{L}_{\text{VLB}}(\mathbf{x}) \;=\; -\mathbb{E}_{q(\mathbf{y}_{1:n}|\mathbf{x})}\big[\log p(\mathbf{x} \mid \mathbf{y}_{1:n}) \;+\; \sum_{i=1}^{n} D_{\text{KL}}\Big(q(\mathbf{y}_i \mid \mathbf{x}) \;\big\|\; p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1})\Big)\big].$$

In further analysis, we focus on the treatment of the second term on the right-hand side. For convenience in the subsequent derivation, we define:

$$\begin{aligned}\mathcal{L}^n(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{y}_{1:n}|\mathbf{x})}\Big[\sum_{i=1}^{n} D_{\mathrm{KL}}\big(q(\mathbf{y}_i \mid \mathbf{x}) \parallel p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{x})\big)\Big] \\ &= n\, \mathbb{E}_{i\sim U(1,n), q(\mathbf{y}_{1:n}|\mathbf{x})}\Big[D_{\mathrm{KL}}\big(q(\mathbf{y}_i \mid \mathbf{x}) \parallel p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{x})\big)\Big] \end{aligned} \tag{18}$$

Next, to specify the conditional distribution $p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{x})$, we utilize $p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{x}) = \frac{p(\mathbf{y}_{1:i}|\mathbf{x})}{p(\mathbf{y}_{1:i-1}|\mathbf{x})}$, and incorporate the following decomposition:

$$p(\mathbf{y}_{1:i} \mid \mathbf{x}) = \prod_{j=1}^{i} \mathbb{E}_{p_F(\boldsymbol{\theta}_{j-1}|\mathbf{x}, t_{j-1})}\big[p_R(\mathbf{y}_j \mid \boldsymbol{\theta}_{j-1}, t_{j-1})\big] \tag{19}$$

It follows that:

$$p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{x}) = \mathbb{E}_{p_F(\boldsymbol{\theta}_{i-1}|\mathbf{x}, t_{i-1})}\Big[p_R(\mathbf{y}_i \mid \boldsymbol{\theta}_{i-1}, t_{i-1})\Big] \tag{20}$$

Substituting this result into Eq. 18, we obtain:

$$\mathcal{L}^n(\mathbf{x}) = n\, \mathbb{E}_{i\sim U(1,n), \boldsymbol{\theta}\sim p_F(\boldsymbol{\theta}|\mathbf{x}, t_{i-1})}\Big[D_{\mathrm{KL}}\big(q(\mathbf{y}_i \mid \mathbf{x}) \parallel p_R(\mathbf{y}_i \mid \boldsymbol{\theta}, t_{i-1})\big)\Big] \tag{21}$$

Given our parametrization with neural network, we can further replace $p_R(\mathbf{y}_i \mid \boldsymbol{\theta}, t_{i-1})$ with $q\big(\mathbf{y}_i \mid \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\big)$:

$$\mathcal{L}^n(\mathbf{x}) = n\, \mathbb{E}_{i\sim U(1,n), \boldsymbol{\theta}\sim p_F(\boldsymbol{\theta}|\mathbf{x}, t_{i-1})}\Big[D_{\mathrm{KL}}\big(q(\mathbf{y}_i \mid \mathbf{x}) \parallel q(\mathbf{y}_i \mid \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t))\big)\Big] \tag{22}$$

Following Graves et al. (2023), different noising processes are specified for q for different modalities. For continuous data, $q(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{x}, \alpha^{-1}\boldsymbol{I})$. For discrete data, $q(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\alpha(K\mathbf{e}_{\mathbf{x}} - \mathbf{1}), \alpha K\boldsymbol{I})$, where we define $\mathbf{x} := \sqrt{K}\mathbf{e}_{\mathbf{x}}$. Denoting $\mathcal{L}^\infty(\mathbf{x}) = \lim_{n\to\infty} \mathcal{L}^n(\mathbf{x}), \alpha(t, \epsilon) = \beta(t) - \beta\big(t - \epsilon\big)$, we have

$$\begin{aligned}\mathcal{L}^\infty(\mathbf{x}) &= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\, \mathbb{E}_{i\sim U(1,n), p_F(\boldsymbol{\theta}|x, t_{i-1})} D_{\mathrm{KL}}(q(\mathbf{y}_i \mid \mathbf{x}) \parallel q(\mathbf{y}_i \mid \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t))) \\ &= \lim_{\epsilon\to 0} \mathbb{E}_{i\sim U(1,n), p_F(\boldsymbol{\theta}|x, t_{i-1})} \frac{\alpha(t, \epsilon)}{2\epsilon} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 \\ &= \frac{1}{2}\, \mathbb{E}_{i\sim U(1,n), p_F(\boldsymbol{\theta}|x, t_{i-1})} \beta'(t) \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 \end{aligned} \tag{23}$$

which establishes the fact that Eq. 6 corresponds to the single-modality VLB.

B.2 Equivalence between Eq. 6 and Eq. 7

Rewrite the expectation in the above equation into an integral form and make the substitution $u = \beta(t)$. This substitution can be transformed as follows: $u = \beta(t) \Rightarrow du = \beta'(t)dt \Rightarrow dt = \frac{du}{\beta'(t)}$, and we directly rewrite $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)$ as $\tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)$.

$$\begin{aligned}\mathcal{L}^\infty(\mathbf{x}) &= \frac{1}{2}\int_0^1 \int_{p_F(\boldsymbol{\theta}|x,t)} \beta'(t) \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, t)\|^2 d\boldsymbol{\theta}\, dt \\ &= \frac{1}{2}\int_{\beta(0)}^{\beta(1)} \int_{p_F(\boldsymbol{\theta}|x,u)} \beta'(\beta^{-1}(u)) \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)\|^2 \frac{du}{\beta'(\beta^{-1}(u))} d\boldsymbol{\theta} \\ &= \frac{1}{2}\int_{\beta(0)}^{\beta(1)} \int_{p_F(\boldsymbol{\theta}|x,u)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)\|^2 du\, d\boldsymbol{\theta} \\ &= \frac{1}{2}\int_{\beta(0)}^{\beta(1)} \mathbb{E}_{p_F(\boldsymbol{\theta}|x,u)} \|\mathbf{x} - \tilde{\mathbf{x}}_\phi(\boldsymbol{\theta}, u)\|^2 du \end{aligned} \tag{24}$$

which gives the form of Eq. 7.

B.3 Equivalence between Eq. 9 and 14

Recall that the generalized loss over $\beta \in \mathcal{Z}$ is:

$$\mathcal{L}^{\infty}(\mathbf{x}) = \frac{1}{2}\int_{\beta_c,\beta_d} \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};\beta)} \|\mathbf{x} - \tilde{\mathbf{x}}_{\phi}(\boldsymbol{\theta},\beta)\|^2 d\beta.$$

and the function space $\mathcal{Z}$ is reparameterized into a product of uniform distributions over t_c, t_d, yielding

$$\dot{\mathcal{L}}^{\infty}(\mathbf{x}) = \frac{1}{2}\int_0^1\int_0^1 \mathop{\mathbb{E}}_{p_F(\boldsymbol{\theta}|\mathbf{x};t)} \|\mathbf{x} - \tilde{\mathbf{x}}_{\phi}(\boldsymbol{\theta},t)\|^2 dt_c dt_d,$$

this is equivalent to implicitly sampling $f(t), g(t)$, covering all possible time-rescaling functions as stated in Remark 4.2.

The joint distribution over (β_c, β_d) is thus

$$p(\beta_c, \beta_d) = p(f, g) = U(0,1) \times U(0,1), \tag{25}$$

where $\beta_c = \tilde{\beta}_c(f(t))$ and $\beta_d = \tilde{\beta}_d(g(t))$.

Therefore, we see that the time-rescaling functions $f(t), g(t)$ act as latent variables defining the coupling of β_c, β_t. Integrating over (t_c, t_d) marginalizes over all possible β, ensuring the model is trained to denoise all possible combinations of noise levels. By demonstrating this equivalence, we reframe the invariant objective described in Eq. 9 as Eq. 14.

B.4 Derivation of Eq. 10 as multi-modality VLB

To rigorously prove that the VLB corresponds to a line integral along a 1D submanifold (path) in the 2D noise schedule space, recall that the joint noise schedule be a parameterized path in the joint function space $\mathcal{Z}$:

$$\beta(t) = \big(\beta_c(t), \beta_d(t)\big), \quad t \in [0,1],$$

where $\beta_c(t), \beta_d(t)$ are monotonically increasing functions with fixed endpoints:

$$\beta_c(0) = \beta_{c,0}, \quad \beta_c(1) = \beta_{c,1}, \quad \beta_d(0) = \beta_{d,0}, \quad \beta_d(1) = \beta_{d,1}.$$

Following Appendix B.1 and the factorized nature of q, the VLB for a joint schedule $\beta(t)$ and multi-modality $\mathbf{x}$ is:

$$\mathcal{L}_{\text{VLB}}(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{p_F(\boldsymbol{\theta}|\mathbf{x};t)} \int_0^1 \left[\beta_c'(t)\|\mathbf{x}_c - \tilde{\mathbf{x}}_{\phi,c}(\boldsymbol{\theta},\beta(t))\|^2 + \beta_d'(t)\|\mathbf{x}_d - \tilde{\mathbf{x}}_{\phi,d}(\boldsymbol{\theta},\beta(t))\|^2\right] dt,$$

where $\tilde{\mathbf{x}}_{\phi} = (\tilde{\mathbf{x}}_{\phi,c}, \tilde{\mathbf{x}}_{\phi,d})$ is the denoising model output for continuous (c) and discrete (d) modalities.

The VLB can be interpreted as a line integral over the trajectory $\beta(t) \in \mathcal{Z}$. Define the vector field $\mathbf{F}(\beta_c, \beta_d)$ as:

$$\mathbf{F}(\beta_c, \beta_d) = \begin{pmatrix} \|\mathbf{x}_c - \tilde{\mathbf{x}}_{\phi,c}(\boldsymbol{\theta},\beta)\|^2 \\ \|\mathbf{x}_d - \tilde{\mathbf{x}}_{\phi,d}(\boldsymbol{\theta},\beta)\|^2 \end{pmatrix}.$$

The VLB then becomes:

$$\mathcal{L}_{\text{VLB}}(\mathbf{x}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}} \int_{\beta(t)} \mathbf{F}(\beta_c, \beta_d) \cdot d\boldsymbol{\beta},$$

where $d\boldsymbol{\beta} = (\beta_c'(t)dt, \beta_d'(t)dt)$ is the differential vector along the path $\beta(t)$, and $\cdot$ denotes inner product. By redefining $\tilde{\mathbf{x}}_{\phi}(\boldsymbol{\theta},\beta) := \mathbf{F}(\beta_c, \beta_d)$, we obtain Eq. 10.

B.5 Derivation of Proposition 4.3

Proof. Since the model $\tilde{\mathbf{x}}_{\phi}(\boldsymbol{\theta}, t)$ is trained by the generalized loss $\dot{\mathcal{L}}^{\infty}$, which is equivalent to integrating the scalar field $\|\mathbf{x} - \tilde{\mathbf{x}}_{\phi}\|^2$ over the entire function space $\mathcal{Z}$, it is exposed to all pairs $(\beta_c, \beta_d) \in \mathcal{Z}$ within the region $[\tilde{\beta}_c(0), \tilde{\beta}_c(1)] \times [\tilde{\beta}_d(0), \tilde{\beta}_d(1)]$, with $\tilde{\beta}_c, \tilde{\beta}_d$ defined in Eq. 3. Therefore,

the model achieves minimal prediction error everywhere in the joint noise space, allowing accurate computation for any specific path $\beta(t)$.

Following Appendix B.4, the line integral $\mathcal{L}^{\infty}(\mathbf{x})$ corresponds to the VLB for a given path $\beta(t)$. Since the integral is equivalent to restricting the generalized loss $\dot{\mathcal{L}}^{\infty}(\mathbf{x})$ to the 1D manifold defined by $\beta(t)$, and the model's predictions are accurate along any submanifold within the space, evaluating $\mathcal{L}^{\infty}(\mathbf{x})$ for a specific $\beta(t)$ uses the model's pre-optimized predictions at each point on the path, thereby yielding a valid VLB estimate. This establishes that the generalized training enables estimating VLBs for arbitrary joint schedules $\beta \in \mathcal{Z}$. □

C Details on Benchmarking

C.1 Experimental Setup

Baselines We provide a brief overview of all SBDD baselines as follows:

- **Autoregressive methods:** AR (Luo et al., 2021) utilizes MCMC sampling to reconstruct molecules atom-by-atom based on voxel-wise density predictions. Pocket2Mol (Peng et al., 2022) generates molecules atom-by-atom with bonds using an E(3)-equivariant network, predicting frontier atoms to improve sampling efficiency. FLAG (Zhang et al., 2023) is a fragment-based model that assembles molecular fragments by predicting their positions and torsion angles.
- **Diffusion:** DiffSBDD (Schneuing et al., 2022) constructs an E(3)-equivariant continuous diffusion model for full-atom generation, applying noise to both atom types and coordinates, while TargetDiff (Guan et al., 2022) adopts a hybrid diffusion process to separately handle continuous coordinates and discrete atom types. DecompDiff (Guan et al., 2023) incorporates chemical priors by decomposing molecules into scaffolds and contact arms.
- **BFN:** MolCRAFT (Qu et al., 2024) uses Bayesian Flow Network (BFN) with advanced variance reduction sampling technique, demonstrating notable improvements over diffusion-based models in molecular design.

Metrics We evaluate the generated molecules using the following commonly adopted metrics:

- **PB-Valid** denotes the PoseBusters passing rate, where **Strain Energy Passed** denotes the ratio of molecules passing PoseBusters internal energy check.
- **Affinity Metrics** are calculated using AutoDock Vina (Eberhardt et al., 2021), these include *Vina Score* as the raw binding energy of a molecular pose in the pocket, *Vina Min* as the binding energy after local energy minimization of the molecular pose, and *Vina Dock* as the lowest binding energy obtained after an extended search for the optimal pose.
- **Molecular Properties** including **QED** for drug-likeness and **SA** (synthetic accessibility score) are calculated using RDKit. They are desired to fall within reasonable ranges.
- **Connected Ratio** is the percentage of fully connected molecules.
- **Diversity** assesses the variety of generated molecules for each binding site by averaging Tanimoto similarity over Morgan fingerprints across all test proteins, following Luo et al. (2021). It is worth noting that this is not necessarily the higher the better, as the desirable bioactive compounds against a protein target often cluster in the molecular space.
- **RMSD** reports the percentage of molecules where the RMSD between generated poses and ground-truth or Vina redocked poses is within 2 Å, indicating consistent binding modes. For better differentiation, we refer to the latter as self-consistency RMSD (scRMSD).

D More Evaluation Results

D.1 Dynamics of Better-balanced Modalities

We visualize the modality-specific validation loss curves for the model trained under generalized loss (Eq. 9) in Fig. 6, which shows that the denoising model effectively learns to mutually benefit from cleaner information either modality.

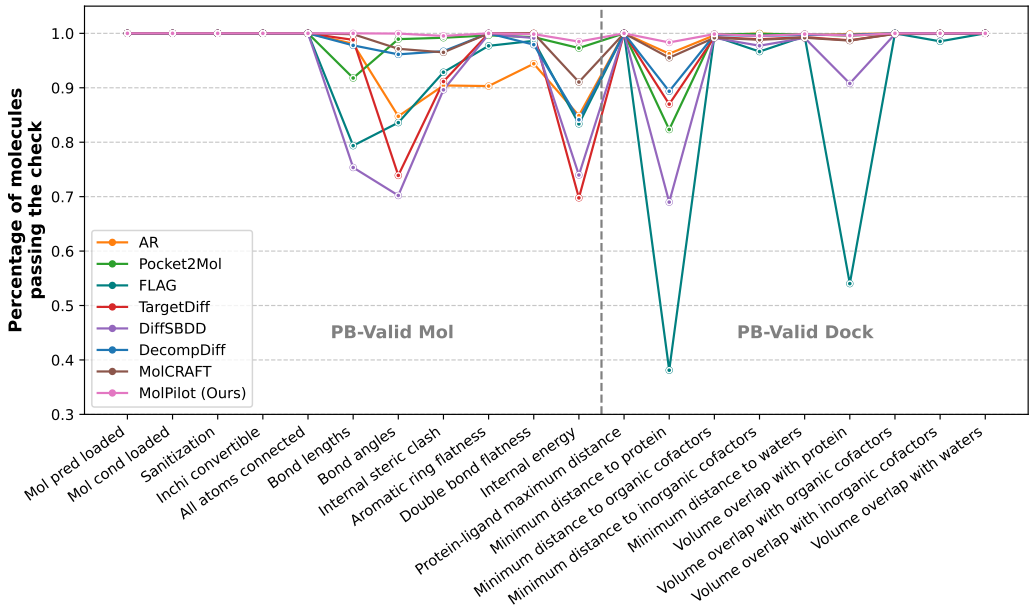


Figure 5: Percentage of generated molecules that have passed the PoseBusters validity checks on CrossDock test set. *PB-Valid Mol:* intramolecular validity. *PB-Valid Dock:* intermolecular validity. Reported *PB-Valid*: PB-Valid Mol & PB-Valid Dock.

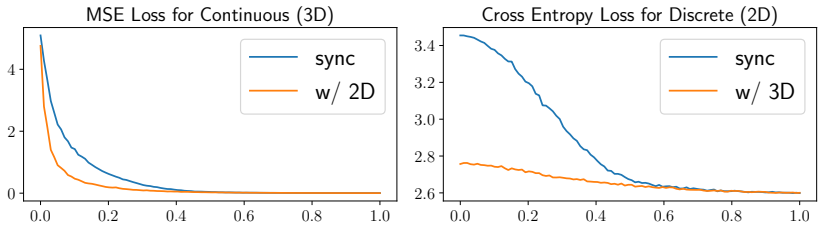


Figure 6: Validation loss curves for the model trained under generalized objective w.r.t. timestep in the generative process. *Sync:* Modalities at the same timestep. *w/ 2D:* Discrete modality always at $t = 1$. *w/ 3D:* Continuous modality always at $t = 1$.

Table 3: Results of Jenson-Shannon Divergence (JSD) of bond length, bond angle, torsion angle distributions between generated molecules and CrossDock reference molecules, averaged over all types with a frequency > 100.

Methods	Length (↓) Avg. JSD	Angle (↓) Avg. JSD	Torsion (↓) Avg. JSD
AR	0.544	0.507	0.545
Pocket2Mol	0.472	0.482	0.467
TargetDiff	0.365	0.435	0.411
DecompDiff	0.332	0.410	0.338
MolCRAFT	0.318	0.384	0.322
Ours	**0.252**	**0.351**	**0.287**

D.2 De novo Design

We report the results on CrossDock (Francoeur et al., 2020) for de novo design. We sample 100 molecules for each of the 100 test proteins, and the results in Table 1 show that our model consistently performs exceptionally across most metrics—leading in PB-Valid, overall Vina affinities, and achieving comparable drug-like properties w.r.t. reference molecules.

For conformation quality, our model achieves the best performance with the highest PoseBusters passing rate, closely matching the reference value of 95.0%. Details of each validity check are shown in Fig. 5, and it can be seen that the leading factors affecting the overall performance are bond lengths and angles (AR, TargetDiff), internal energy (MolCRAFT, DecompDiff) for intramolecular validity, and minimum distance to protein (Pocket2Mol) for intermolecular validity.

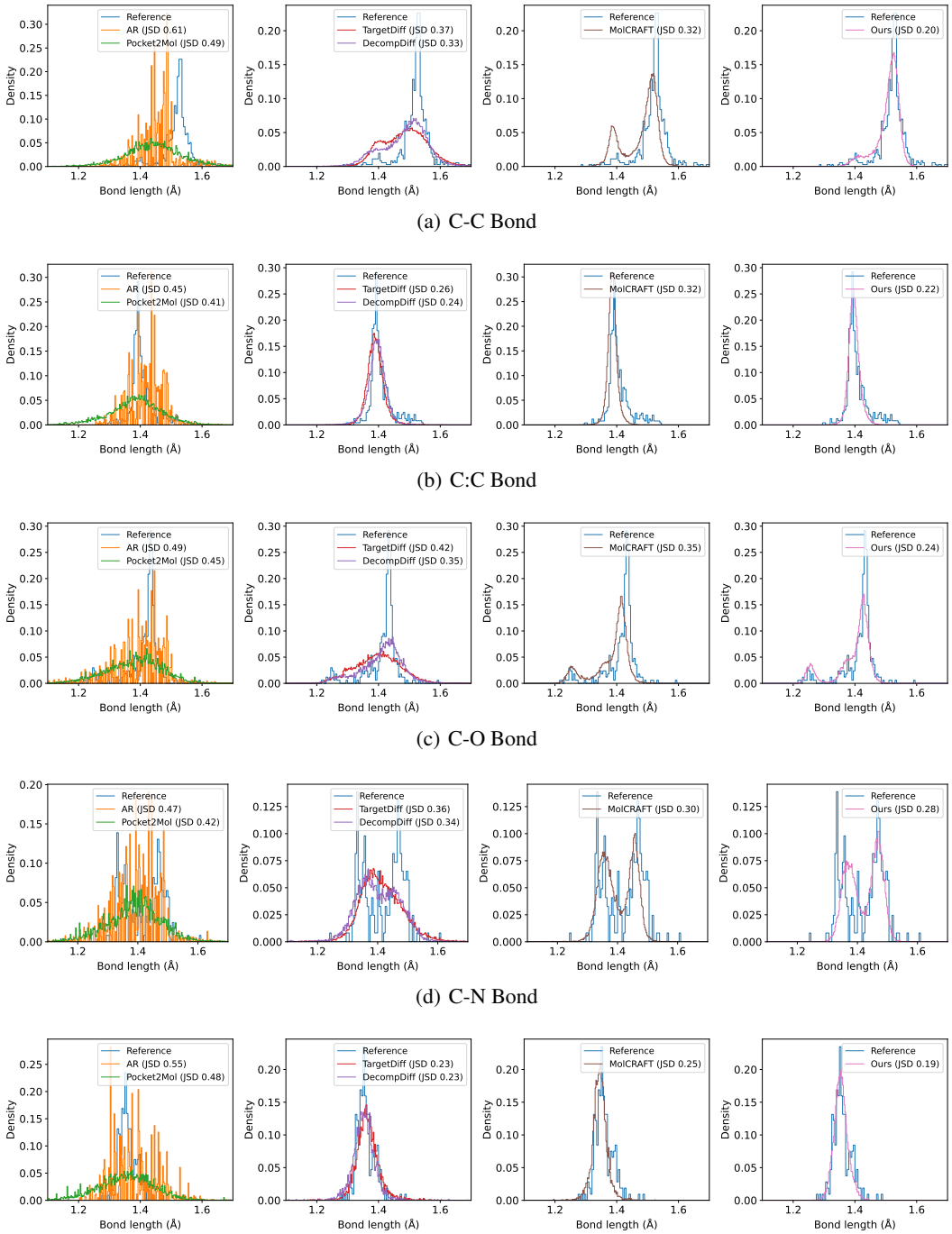
For molecular geometries, we visualize the distributions of bond length, bond angle, and torsion angle, between generated molecules and reference molecules in the test set in Fig. 7, 8, 9 for the top-5 frequent types, and summarize the overall Jensen-Shannon Divergence (JSD) in Table 3, averaged over all types with a frequency > 100. While the previous strong-performing method MolCRAFT captures the bond length distributions relatively well, it cannot fit all the bond angle or torsion angle distributions, displaying for example non-standard C-C-C bond angle, smoothed C-C-N bond angle, and diverges more severely in C-C-C-C and C-C-N-C torsion angles compared with reference. Our method excels at modeling all these distributions, underscoring its ability to capture the molecular geometries.

For binding affinities, our model demonstrates the leading performance with an average Vina score of -6.88 kcal/mol, maintaining the closest gap between Vina Score, Vina Min, and Vina Dock. This shows our model's superiority in capturing spatial interactions in the generated pose.

For molecular properties like QED and SA, our model displays competitive performance, and ranks at the top among all non-autoregressive models.

D.3 Molecular Docking

We repurpose the SBDD models from the joint codesign model $P(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{A} \mid \mathbf{x}_P)$ to the conditional marginal $P(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{A}, \mathbf{x}_P)$. This actually corresponds to another specific line on the loss surface described by Eq. 10, with $\beta_d(t) \equiv \beta_d(1)$. We employ the joint schedule $\beta(t) = (\tilde{\beta}_c(t), \tilde{\beta}_d(1))$ with $\tilde{\beta}$ being the default noise schedule for continuous modality in Eq. 3.



(a) C-C Bond

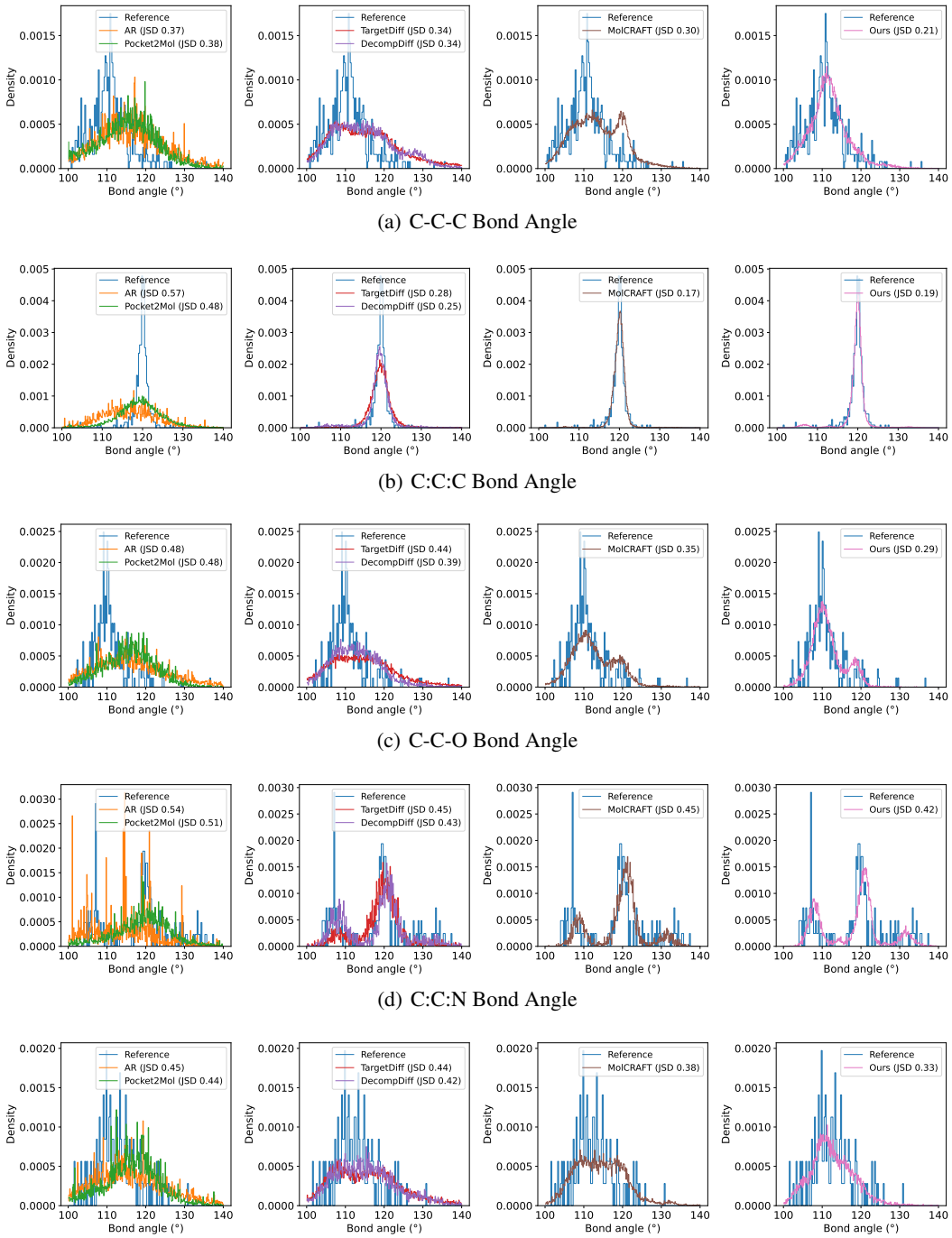
(b) C:C Bond

(c) C-O Bond

(d) C-N Bond

(e) C:N Bond

Figure 7: Top-5 frequent bond length distribution of generated molecules compared with CrossDock reference molecules.



(a) C-C-C Bond Angle

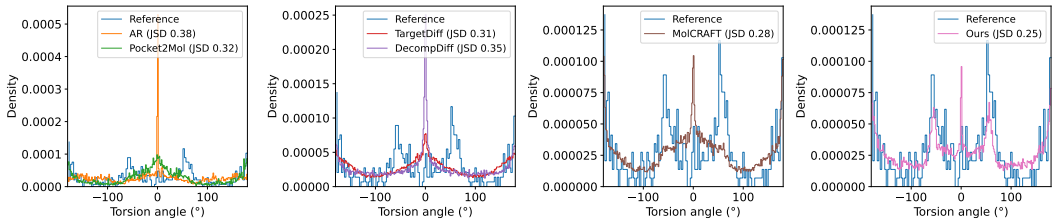
(b) C:C:C Bond Angle

(c) C-C-O Bond Angle

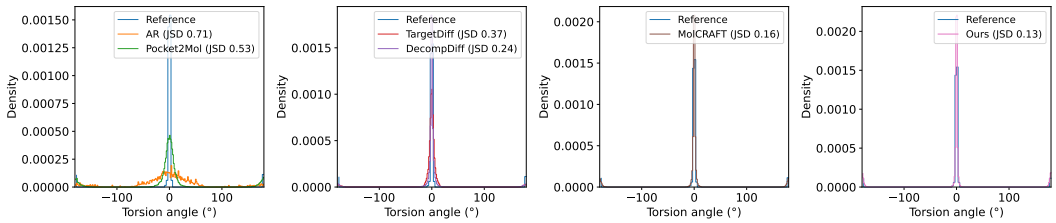
(d) C:C:N Bond Angle

(e) C-C-N Bond Angle

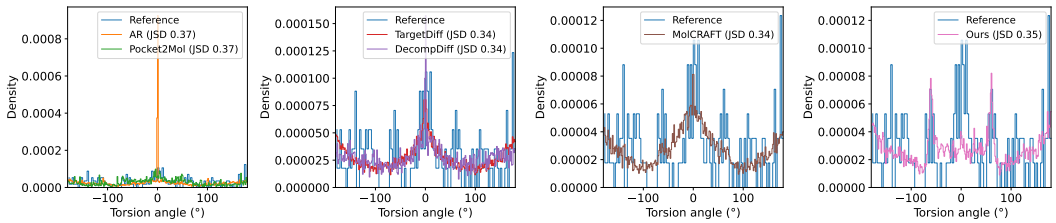
Figure 8: Top-5 frequent bond angle distribution of generated molecules compared with CrossDock reference molecules.



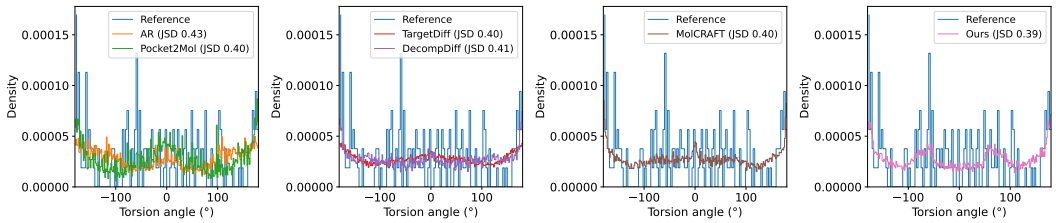
(a) C-C-C-C Torsion Angle



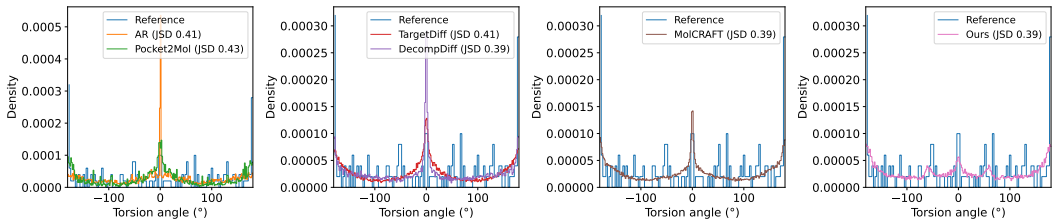
(b) C:C:C:C Torsion Angle



(c) C-C-O-C Torsion Angle



(d) C-C-C-O Torsion Angle



(e) C-C-N-C Torsion Angle

Figure 9: Top-5 frequent torsion angle distribution of generated molecules compared with CrossDock reference molecules.