# MOLLIFT: LIFTING SMILES PRIORS INTO LOCAL FRAMES FOR 3D MOLECULE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Pocket-conditioned 3D molecule generation requires capturing both protein–ligand geometry and the chemical regularities underlying molecular structure. While 3D representations naturally describe conformations and spatial interactions, SMILES provides complementary information about molecular connectivity and chemical organization. We propose **MolLIFT**, a bidirectional 1D–3D Bayesian flow framework that incorporates evolving SMILES priors into pocket-conditioned 3D generation while preserving SE(3)-equivariance. MolLIFT constructs geometry-aware atom representations from local protein–ligand environments and uses atom-centered local frames to lift invariant SMILES-derived messages into explicit equivariant coordinate updates. The evolving 3D geometry is then projected back into the same local frames to provide orientation-invariant feedback for refining the SMILES belief, forming a closed bidirectional interaction throughout generation. On the CrossDocked benchmark, MolLIFT improves binding affinity, molecular quality, diversity, and success rate over the underlying property-agnostic generator. Ablation studies show that these gains stem from geometry-aware cross-modal interaction and explicit coordinate lifting rather than increased model capacity. The joint 1D–3D state further enables gradient-guided multi-objective optimization, yielding additional improvements in affinity and molecular quality. Our results establish local-frame lifting as an effective mechanism for transferring sequence-level chemical priors into equivariant 3D molecular generation.

## 1 INTRODUCTION

Structure-based drug design (SBDD) aims to generate chemically plausible ligands that adopt physically meaningful conformations and form favorable interactions with a target protein pocket. Recent pocket-conditioned diffusion models and Bayesian Flow Networks (BFNs) have substantially advanced this setting by directly modeling ligand coordinates and atom types in an equivariant manner (Guan et al., 2023a;b; Qu et al., 2024). Property-guided methods further improve generated molecules through learned predictors or molecular rewards, while preserving pocket compatibility and chemical validity (Gu et al., 2024; Dorna et al., 2024; Qiu et al., 2025).

However, spatial geometry captures only one aspect of molecular structure; molecular identity is also governed by topological and chemical regularities (Luo & Deng, 2025). These two aspects are emphasized differently by 3D and sequence representations. While 3D representations explicitly characterize molecular conformations and, together with the protein pocket context, are naturally suited to modeling spatial complementarity and binding interactions, SMILES provides a complementary view of molecular connectivity, functional-group composition, branching patterns, ring systems, and other sequence-level chemical regularities (Martínez Galindo et al., 2026). We empirically probe this distinction in Figure 1: 3D representations are more informative for pocket-dependent binding-affinity prediction, whereas SMILES representations yield better predictions of ligand-intrinsic properties such as synthetic accessibility (SA) and drug-likeness (QED). This complementary behavior suggests that SMILES can provide chemical priors not readily captured by geometric representations alone. This observation naturally motivates our central question: *can sequence-level chemical priors be incorporated into pocket-conditioned 3D molecule generation without weakening geometric reasoning or SE(3)-equivariance?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

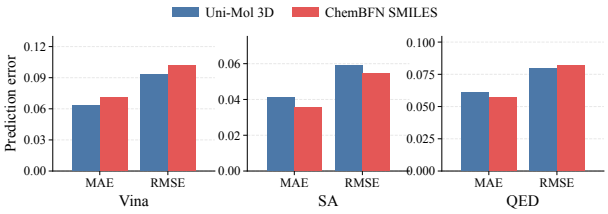



Figure 1: Complementarity between 3D and SMILES representations. Details are provided in Appendix G.1.

Bridging these two representations, however, requires resolving three fundamental problems. First, SMILES tokens are organized by molecular serialization, whereas the relevance of a 3D atom is governed by its local protein–ligand environment, making geometry-aware context essential for meaningful cross-modal association. Second, SMILES representations are inherently invariant and scalar, while coordinate denoising requires equivariant vector updates; sequence information must therefore be translated into spatial directions without introducing a global reference frame. Third, the two modalities evolve jointly throughout generation, so cross-modal interaction should be bidirectional: sequence priors should guide the 3D trajectory, while the evolving geometry should provide orientation-invariant feedback to refine the sequence representation.

To this end, we propose MOLLIFT, a bidirectional 1D–3D Bayesian flow framework for pocket-conditioned 3D molecule generation. MolLIFT maintains SMILES belief alongside ligand-coordinate and atom-type beliefs throughout the Bayesian flow. Its cross-modal interface comprises three components. First, Geometric Context Encoding (GCE) constructs a rotation-and translation-invariant representation of each ligand atom's local protein–ligand environment, providing geometry-aware queries for associating atoms with relevant SMILES information. Second, Local-Frame SMILES-to-3D Lifting (LFL) builds atom-centered frames from equivariant relative vectors and expresses SMILES-derived messages as invariant coefficients in these frames, thereby lifting sequence information into explicit SE(3)-equivariant coordinate updates. Third, Invariant 3D-to-SMILES Feedback (IFB) projects geometric changes onto the local frames and converts them into invariant atom memory for refining the evolving SMILES belief. Together, these modules establish a closed bidirectional interaction between sequence-level chemical priors and 3D denoising while preserving the required transformation properties.

We evaluate MolLIFT on pocket-conditioned molecular generation using the CrossDocked benchmark. MolLIFT improves binding-affinity and molecular-property metrics over the underlying MolCRAFT generator without using explicit property guidance. A matched-capacity baseline shows that these gains cannot be explained by parameter count alone. Ablations of the local-frame interaction and explicit vector message further demonstrate the respective contributions of geometry-aware scalar fusion and direct coordinate lifting. Our main contributions are as follows:

- We formulate a joint 1D–3D Bayesian flow in which SMILES belief provides dynamic chemical priors for pocket-conditioned 3D ligand generation.
- We introduce a bidirectional local-frame interface that associates SMILES information with local protein–ligand environments, lifts invariant sequence messages into SE(3)-equivariant coordinate updates, and returns invariant geometric feedback to the sequence branch.
- We provide generation, matched-capacity, ablation, and mechanistic experiments showing that geometry-aware sequence integration improves both molecular generation performance and the reliability of SMILES-induced coordinate corrections.

## 2 PRELIMINARIES

**Pocket-conditioned 3D molecule generation.** Given a protein pocket $\mathcal{P} = (\mathbf{X}^P, \mathbf{V}^P)$, where $\mathbf{X}^P \in \mathbb{R}^{N_P \times 3}$ denotes pocket atom coordinates and $\mathbf{V}^P \in \mathbb{R}^{N_P \times D_P}$ denotes pocket atom features, the goal of structure-based molecule generation is to generate a 3D ligand $\mathbf{m}^{3D} = (\mathbf{X}, \mathbf{V})$, where $\mathbf{X} \in \mathbb{R}^{N \times 3}$ and $\mathbf{V} \in \{1, \dots, K_v\}^N$ denote ligand atom coordinates and atom types, respectively. The desired model learns the pocket-conditioned distribution $p_\theta(\mathbf{X}, \mathbf{V} \mid \mathcal{P})$. In this work, we further introduce a tokenized SMILES sequence $\mathbf{s} \in \{1, \dots, K_s\}^L$ as a jointly modeled generative modality that provides complementary chemical priors beyond the 3D representation. By coupling the evolving SMILES and 3D states throughout generation, MolLIFT enables sequence-level chemical semantics to interact with pocket-conditioned geometric information. The SMILES modality

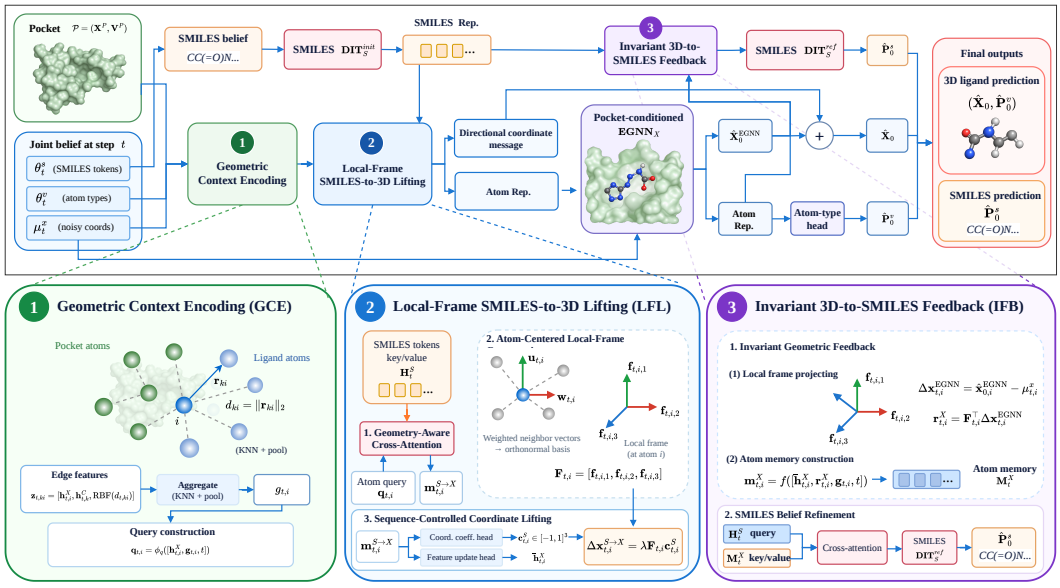



Figure 2: Overview of MolLIFT. The joint 1D–3D belief and pocket context are processed by GCE, LFL, and a pocket-conditioned EGNN to produce the 3D ligand prediction. IFB feeds invariant geometric information back to refine SMILES prediction.

thereby serves as a chemically informative representation that cooperatively guides 3D denoising, while the generated 3D ligand remains the primary output of the model.

**Bayesian Flow Networks.** Bayesian Flow Networks (BFNs) model generation as iterative communication between a sender in data space and a receiver in parameter space. For a clean variable $\mathbf{z}$, the sender transmits a noisy observation $\mathbf{y}_i$ with accuracy $\alpha_i$: $\mathbf{y}_i \sim p_S(\mathbf{y}_i \mid \mathbf{z}; \alpha_i)$. The receiver maintains a belief parameter $\boldsymbol{\theta}_{i-1}$ and updates it through a Bayesian update function $h$:

$$p_U(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{i-1}, \mathbf{z}; \alpha_i) = \mathbb{E}_{p_S(\mathbf{y}_i|\mathbf{z};\alpha_i)} \delta(\boldsymbol{\theta}_i - h(\boldsymbol{\theta}_{i-1}, \mathbf{y}_i, \alpha_i)) \,. \tag{1}$$

A neural network $\Phi_\theta$ predicts the clean data distribution from the current belief, which induces a receiver distribution

$$p_R(\mathbf{y}_i \mid \boldsymbol{\theta}_{i-1}; t_i, \alpha_i) = \mathbb{E}_{p_O(\hat{\mathbf{z}}|\Phi_\theta(\boldsymbol{\theta}_{i-1}, t_i, \mathcal{P}))} p_S(\mathbf{y}_i \mid \hat{\mathbf{z}}; \alpha_i). \tag{2}$$

The training objective minimizes the sender–receiver discrepancy:

$$\mathcal{L}^n(\mathbf{z}) = \mathbb{E} \sum_{i=1}^{n} D_{\mathrm{KL}}(p_S(\mathbf{y}_i \mid \mathbf{z}; \alpha_i) \,\|\, p_R(\mathbf{y}_i \mid \boldsymbol{\theta}_{i-1}; t_i, \alpha_i)) \,. \tag{3}$$

**Hybrid 1D–3D Bayesian flow states.** MolLIFT maintains Bayesian flow states for ligand coordinates, ligand atom types, and SMILES tokens: $\boldsymbol{\Theta}_t = (\boldsymbol{\mu}_t^x, \boldsymbol{\theta}_t^v, \boldsymbol{\theta}_t^s)$. For coordinates, the continuous belief is parameterized by a noisy mean

$$\boldsymbol{\mu}_t^x = \gamma_x(t)\mathbf{X} + \sqrt{\gamma_x(t)(1-\gamma_x(t))}\boldsymbol{\epsilon}^x, \qquad \boldsymbol{\epsilon}^x \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4}$$

where $\gamma_x(t) = 1 - \sigma_1^{2t}$. For a categorical variable $\mathbf{a} \in \{\mathbf{V}, \mathbf{s}\}$ with vocabulary size $K_a$, the belief is

$$\boldsymbol{\theta}_t^a = \mathrm{softmax}\left(\beta_a(t)(K_a \mathbf{e}_a - \mathbf{1}) + \sqrt{\beta_a(t)K_a}\boldsymbol{\epsilon}^a\right), \qquad \beta_a(t) = \beta_1^a t^2. \tag{5}$$

These states provide the common denoising interface for the 3D ligand branch and SMILES branch.

## 3 Method

### 3.1 Overview

MolLIFT is a Bayesian Flow Network (BFN)-based framework that jointly models ligand coordinates, atom types, and a SMILES sequence within a unified generative process for pocket-conditioned 3D molecule generation. As shown in Figure 2, its central design is a bidirectional

cross-modal interface between the sequence and geometric branches, allowing the two modalities to exchange information throughout Bayesian flow denoising. MolLIFT operates on the joint Bayesian flow state $\mathbf{\Theta}_t = (\boldsymbol{\mu}_t^x, \boldsymbol{\theta}_t^v, \boldsymbol{\theta}_t^s)$ and a protein pocket $\mathcal{P}$. Given these states, MolLIFT predicts the clean ligand coordinates, atom types, and SMILES token distribution:

$$(\hat{\mathbf{X}}_0, \hat{\mathbf{P}}_0^v, \hat{\mathbf{P}}_0^s) = \mathrm{MolLIFT}(\mathbf{\Theta}_t, \mathcal{P}, t). \tag{6}$$

This joint prediction is mainly realized through three complementary components that mediate bidirectional 1D–3D interaction. In the SMILES-to-3D direction, (1) *Geometric Context Encoding (GCE)* summarizes the local protein–ligand environment of each ligand atom and constructs geometry-aware atom representations for cross-modal association. (2) *Local-Frame SMILES-to-3D Lifting (LFL)* then uses these representations to extract relevant SMILES information and lifts the resulting invariant chemical message into SE(3)-equivariant spatial guidance through atom-centered local frames. The updated atom features and coordinate messages are subsequently used by a pocket-conditioned equivariant denoiser to predict the clean 3D ligand. In the reverse direction, (3) *Invariant 3D-to-SMILES Feedback (IFB)* projects the geometric changes predicted by the 3D branch into the same local frames, yielding orientation-invariant atom memories that refine the evolving SMILES representation. The complete training and generation procedures are summarized in Appendix C.

### 3.2 Geometric Context Encoding (GCE)

Geometric Context Encoding (GCE) encodes each ligand atom's current protein–ligand environment with a rotation-and translation-invariant local descriptor, yielding a geometry-aware query for subsequent SMILES-to-3D cross-modal interaction.

The ligand atom-type belief is embedded into invariant scalar features $\mathbf{h}_{t,i}^X$. We denote the feature and coordinate of a context atom by $\mathbf{h}_{t,k}^C$ and $\mathbf{x}_{t,k}^C$, respectively, where the context may contain atoms from both the ligand and the protein pocket. For each ligand atom $i$, we construct a $k$-nearest-neighbor (KNN) graph over the current protein–ligand structure and denote its neighborhood by $\mathcal{N}(i)$. For each neighboring context atom $k$, we compute

$$\begin{aligned} \mathbf{r}_{t,ki} &= \mathbf{x}_{t,k}^C - \boldsymbol{\mu}_{t,i}^x, \qquad d_{t,ki} = \|\mathbf{r}_{t,ki}\|_2, \\ \mathbf{z}_{t,ki} &= [\mathbf{h}_{t,i}^X, \mathbf{h}_{t,k}^C, \mathrm{RBF}(d_{t,ki})]. \end{aligned} \tag{7}$$

where $\mathbf{r}_{t,ki}$ describes the relative geometry around atom $i$, while $\mathbf{z}_{t,ki}$ combines atom features with an RBF encoding of the interatomic distance, and $[\cdot]$ denotes concatenation.

The local environment is summarized by aggregating these invariant edge representations,

$$\mathbf{g}_{t,i} = \frac{1}{|\mathcal{N}(i)|} \sum_{k \in \mathcal{N}(i)} \phi_g(\mathbf{z}_{t,ki}), \tag{8}$$

where $\phi_g$ is a learnable MLP. We then combine the current atom representation, its local geometric context, and the flow time to construct the atom query

$$\mathbf{q}_{t,i} = \phi_q([\mathbf{h}_{t,i}^X, \mathbf{g}_{t,i}, t]). \tag{9}$$

where $\phi_q$ is a learnable MLP. Because $\mathbf{g}_{t,i}$ is constructed from invariant scalar features and interatomic distances, both $\mathbf{g}_{t,i}$ and $\mathbf{q}_{t,i}$ remain unchanged under global rotations and translations. The resulting geometry-aware query is passed to LFL, where it determines how the current local 3D environment interacts with the evolving SMILES representation.

### 3.3 Local-Frame SMILES-to-3D Lifting (LFL)

Local-Frame SMILES-to-3D Lifting (LFL) converts invariant SMILES information into geometry-aware atom features and explicit SE(3)-equivariant coordinate guidance. Specifically, LFL consists of three components: **Geometry-Aware Cross-Attention** extracts atom-specific chemical messages conditioned on the local geometric context, **Atom-Centered Local-Frame Construction** establishes equivariant spatial bases from the current 3D environment, and **Sequence-Controlled Coordinate Lifting** predicts invariant coefficients in these bases and lifts them into directional coordinate updates.

**Geometry-Aware Cross-Attention.** Using the geometry-aware query produced by GCE, this component aggregates token-level SMILES features into an atom-specific chemical message that can be fused with the 3D atom representation. Let $\mathrm{sg}(\cdot)$ denote stop-gradient and define $\widetilde{\mathbf{h}}^S_{t,j} = \mathrm{sg}(\mathbf{h}^S_{t,j})$. For attention head $r$ with dimension $d_h$, atom $i$ attends to valid, non-padding SMILES tokens $\mathcal{V}_S$:

$$a^{(r)}_{t,ij} = \mathrm{softmax}_{j\in\mathcal{V}_S}\left(\frac{(W^{(r)}_Q \mathbf{q}_{t,i})^\top (W^{(r)}_K \widetilde{\mathbf{h}}^S_{t,j})}{\sqrt{d_h}}\right). \tag{10}$$

The outputs of the $H$ attention heads are concatenated and projected to form the SMILES message:

$$\mathbf{m}^{S\to X}_{t,i} = W_O\,\mathrm{Concat}^H_{r=1}\left(\sum_{j\in\mathcal{V}_S} a^{(r)}_{t,ij} W^{(r)}_V \widetilde{\mathbf{h}}^S_{t,j}\right). \tag{11}$$

Using $\boldsymbol{\xi}_{t,i} = [\mathbf{h}^X_{t,i}, \mathbf{g}_{t,i}, \mathbf{m}^{S\to X}_{t,i}, t]$, we update the scalar atom representation:

$$\bar{\mathbf{h}}^X_{t,i} = \mathrm{LN}\left(\mathbf{h}^X_{t,i} + \phi_h(\boldsymbol{\xi}_{t,i})\right). \tag{12}$$

This retains the indirect route by which SMILES influences the geometric backbone, while allowing the scalar fusion to depend on both the local environment and denoising time.

**Atom-Centered Local-Frame Construction.** To provide a geometry-dependent spatial basis for the subsequent coordinate update, we construct a local reference frame for each ligand atom from its surrounding relative geometry. The frame is built entirely from equivariant relative vectors and invariant scalar weights, so it follows the current molecular geometry under rigid transformations.

Specifically, two independent scalar networks assign normalized weights to the neighboring relative vectors:

$$\begin{aligned} &\alpha_{t,ki} = \mathrm{softmax}_{k\in\mathcal{N}(i)}\,\phi_\alpha(\mathbf{z}_{t,ki}), \quad \beta_{t,ki} = \mathrm{softmax}_{k\in\mathcal{N}(i)}\,\phi_\beta(\mathbf{z}_{t,ki}), \\ &\mathbf{u}_{t,i} = \sum_{k\in\mathcal{N}(i)} \alpha_{t,ki}\mathbf{r}_{t,ki}, \qquad \mathbf{w}_{t,i} = \sum_{k\in\mathcal{N}(i)} \beta_{t,ki}\mathbf{r}_{t,ki}. \end{aligned} \tag{13}$$

We use a numerically stabilized Gram–Schmidt construction:

$$\begin{aligned} \mathbf{f}_{t,i,1} &= \frac{\mathbf{u}_{t,i}}{\|\mathbf{u}_{t,i}\|_2 + \epsilon}, \\ \widetilde{\mathbf{w}}_{t,i} &= \mathbf{w}_{t,i} - (\mathbf{f}^\top_{t,i,1}\mathbf{w}_{t,i})\mathbf{f}_{t,i,1}, \\ \mathbf{f}_{t,i,2} &= \frac{\widetilde{\mathbf{w}}_{t,i}}{\|\widetilde{\mathbf{w}}_{t,i}\|_2 + \epsilon}, \qquad \mathbf{f}_{t,i,3} = \mathbf{f}_{t,i,1} \times \mathbf{f}_{t,i,2}. \end{aligned} \tag{14}$$

The local frame is $\mathbf{F}_{t,i} = [\mathbf{f}_{t,i,1}, \mathbf{f}_{t,i,2}, \mathbf{f}_{t,i,3}]$. If the anchor vectors are degenerate, as detected by a norm-based test for a vanishing first anchor or orthogonal residual, we set $\mathbf{F}_{t,i} = \mathbf{0}$. Both the frame construction and this fallback depend only on relative vectors and invariant scalar quantities.

**Sequence-Controlled Coordinate Lifting.** With the atom-specific SMILES message and local spatial frame available, we convert the sequence-derived chemical information into a coordinate update. The fused representation predicts three coefficients along the local frame directions, which are then mapped into the 3D space through the atom-centered frame:

$$\mathbf{c}^S_{t,i} = \tanh(\phi_c(\boldsymbol{\xi}_{t,i})) \in [-1,1]^3, \qquad \Delta\mathbf{x}^{S\to X}_{t,i} = \kappa \mathbf{F}_{t,i}\mathbf{c}^S_{t,i}. \tag{15}$$

The learnable scale $\kappa$ is initialized to $0.01$. The coefficients specify how much to move along each local direction, while the frame supplies those directions in the current complex. This message is SE(3)-equivariant and the proof, including the degenerate-frame case, is given in Appendix E.

After local-frame lifting, an equivariant geometric backbone processes the noisy ligand coordinates, updated scalar features, and fixed pocket context:

$$(\hat{\mathbf{X}}^{\mathrm{EGNN}}_0, \mathbf{H}^X_0) = \mathrm{EGNN}_X(\boldsymbol{\mu}^x_t, \bar{\mathbf{H}}^X_t, \mathcal{P}, t). \tag{16}$$

Its outputs yield the final coordinate and atom-type predictions:

$$\hat{\mathbf{x}}_{0,i} = \hat{\mathbf{x}}^{\mathrm{EGNN}}_{0,i} + \Delta\mathbf{x}^{S\to X}_{t,i}, \qquad \hat{\mathbf{p}}^v_{0,i} = \mathrm{softmax}\big(\phi_v(\mathbf{h}^X_{0,i})\big). \tag{17}$$

The directional message is added *after* the backbone prediction, not to the input coordinate belief. Consequently, local-frame lifting changes the neural prediction but leaves the BFN belief states and Bayesian update rules unchanged.

### 3.4 INVARIANT 3D-TO-SMILES FEEDBACK (IFB)

Invariant 3D-to-SMILES Feedback (IFB) converts predicted geometric changes into invariant local-frame representations that refine the evolving SMILES belief. Specifically, IFB consists of two components: **Invariant Geometric Feedback** projects the predicted atomic displacements onto atom-centered local frames and combines them with atom representations to construct an invariant atom memory, while **SMILES Belief Refinement** incorporates this geometric memory into the sequence representation to predict the updated SMILES distribution.

**Invariant Geometric Feedback.** To represent the geometric adjustment of each ligand atom in an orientation-independent form, we project the displacement predicted by the 3D backbone onto the atom-centered local frame constructed by LFL. We first compute the backbone displacement before adding the SMILES-induced directional message:

$$\begin{aligned} \Delta \mathbf{x}_{t,i}^{\mathrm{EGNN}} &= \hat{\mathbf{x}}_{0,i}^{\mathrm{EGNN}} - \boldsymbol{\mu}_{t,i}^{x}, \\ \mathbf{r}_{t,i}^{X} &= \mathrm{sg}(\mathbf{F}_{t,i})^{\top}\, \mathrm{sg}(\Delta \mathbf{x}_{t,i}^{\mathrm{EGNN}}). \end{aligned} \tag{18}$$

The projected displacement $\mathbf{r}_{t,i}^{X}$ describes the geometric change along the three local frame directions and is invariant to global rotations and translations, as shown in Appendix E. We then combine this geometric signal with the final atom representation and local context to construct the atom memory:

$$\mathbf{m}_{t,i}^{X} = \mathrm{LN}\big(\mathrm{sg}(\mathbf{h}_{0,i}^{X}) + \phi_r\big([\mathrm{sg}(\mathbf{h}_{0,i}^{X}), \mathrm{sg}(\mathbf{g}_{t,i}), \mathbf{r}_{t,i}^{X}]\big)\big)\,. \tag{19}$$

The source features are detached before the learnable projection $\phi_r$, allowing the feedback mapping to be trained without propagating the SMILES loss into the 3D branch.

**SMILES Belief Refinement.** To use the geometric feedback to update the sequence branch, the evolving SMILES representation queries the invariant atom memory through multihead cross-attention. Specifically,

$$\begin{aligned} \mathbf{Z}_t^{S} &= \mathrm{CrossAttn}(Q = \mathbf{H}_t^{S}, K = \mathbf{M}_t^{X}, V = \mathbf{M}_t^{X}), \\ \hat{\mathbf{P}}_0^{s} &= \mathrm{DiT}_S^{\mathrm{ref}}\left(\mathrm{LN}(\mathbf{H}_t^{S} + \mathbf{Z}_t^{S}), t\right). \end{aligned} \tag{20}$$

Here, padded ligand atoms are masked during cross-attention. The resulting $\hat{\mathbf{P}}_0^{s}$ provides the clean SMILES prediction used to update the sequence belief at the subsequent Bayesian flow step. Together with the SMILES-to-3D pathway in LFL, IFB completes the bidirectional interaction between sequence-level chemical information and evolving 3D geometry.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

**Dataset.** We train and evaluate MolLIFT on CrossDocked dataset (Francoeur et al., 2020), which contains 22.5 million protein–ligand pairs before preprocessing. We adopt the protocol of (Luo et al., 2021), applying RMSD-based filtering and a data split based on a 30% sequence-identity threshold. The resulting benchmark comprises 99,990 training pairs and 100 test proteins.

**Baselines.** We divide the baselines into two categories according to whether they use explicit molecular-property information for guidance or optimization. **Methods without property guidance (w/o Prop.)** include GraphBP (Liu et al., 2022), AR (Luo et al., 2021), Pocket2Mol (Peng et al., 2022), TargetDiff (Guan et al., 2023a), FLAG (Zhang et al., 2023), DecompDiff (Guan et al., 2023b), DiffSBDD (Schneuing et al., 2024), IPDiff (Huang et al., 2024), and MolCRAFT (Qu et al., 2024). **Methods with property guidance or optimization (w Prop.)** include DecompOpt (Zhou et al., 2024), TAGMol (Dorna et al., 2024), MolJO (Qiu et al., 2025), and AliDiff (Gu et al., 2024). Detailed descriptions of these baselines are provided in Appendix A.

**Metrics.** We assess generated molecules along three dimensions: binding affinity, molecular properties, and sample diversity. Binding affinity is evaluated with AutoDock Vina (Eberhardt et al.,

Table 1: Summary of binding affinity and molecular properties of reference molecules and molecules generated by MolLIFT and baselines. (↑) / (↓) denotes whether a larger / smaller number is preferred. Top 2 results are **bolded** and underlined, respectively.

| Category | Methods | Vina Score ↓ | | Vina Min ↓ | | Vina Dock ↓ | | QED ↑ | SA ↑ | Div ↑ | Success |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Avg. | Med. | Avg. | Med. | Avg. | Med. | Avg. | Avg. | Avg. | Rate ↑ |
| Reference | | -6.36 | -6.46 | -6.71 | -6.49 | -7.45 | -7.26 | 0.48 | 0.73 | - | 25.0% |
| w/o Prop. | GraphBP | - | - | - | - | -4.80 | -4.70 | 0.43 | 0.49 | 0.79 | 0.1% |
| | AR | -5.75 | -5.64 | -6.18 | -5.88 | -6.75 | -6.62 | 0.51 | 0.63 | 0.70 | 6.9% |
| | Pocket2Mol | -5.14 | -4.70 | -6.42 | -5.82 | -7.15 | -6.79 | 0.57 | **0.76** | 0.69 | 24.4% |
| | TargetDiff | -5.47 | -6.30 | -6.64 | -6.83 | -7.80 | -7.91 | 0.48 | 0.58 | 0.72 | 10.5% |
| | FLAG | 45.85 | 36.52 | 9.71 | -2.43 | -4.84 | -5.56 | **0.61** | 0.63 | 0.70 | 1.8% |
| | DecompDiff | -5.19 | -5.27 | -6.03 | -6.00 | -7.03 | -7.16 | 0.51 | 0.66 | 0.73 | 14.9% |
| | DiffSBDD | -1.44 | -4.91 | -4.52 | -5.84 | -7.14 | -7.30 | 0.47 | 0.58 | 0.73 | 7.9% |
| | MolCRAFT | -6.55 | -6.95 | -7.21 | -7.14 | -7.67 | -7.82 | 0.50 | 0.67 | 0.70 | 26.8% |
| | MolLIFT | **-7.04** | **-7.10** | **-7.41** | **-7.35** | **-7.95** | **-8.10** | 0.54 | 0.69 | **0.85** | **28.2%** |
| w Prop. | DecompOpt | -5.75 | -5.97 | -6.58 | -6.70 | -7.63 | -8.02 | 0.56 | 0.73 | 0.63 | 39.4% |
| | IPDiff | -6.41 | -7.01 | -7.45 | -7.48 | -8.57 | -8.51 | 0.52 | 0.59 | 0.74 | 16.5% |
| | TAGMol | -7.02 | -7.77 | -7.95 | -8.07 | -8.59 | -8.69 | 0.55 | 0.56 | 0.69 | 11.1% |
| | AliDiff | -7.07 | -7.95 | -8.09 | -8.17 | -8.90 | -8.81 | 0.50 | 0.57 | 0.73 | 42.1% |
| | MolJO | **-7.52** | -8.02 | -8.33 | -8.34 | -9.05 | -9.13 | 0.56 | **0.78** | 0.66 | 51.3% |
| | MolLIFT$_{\text{grad}}$ | -7.43 | **-8.11** | **-8.39** | **-8.41** | **-9.10** | **-9.16** | **0.57** | 0.75 | **0.79** | **53.0%** |

2021) under three protocols: **Vina Score** evaluates the generated pose directly in the protein pocket; **Vina Min** scores the pose after local energy minimization; and **Vina Dock** reports the lowest energy found through a more extensive docking search. Following (Guan et al., 2023a), we also report the **Success Rate**, defined as the proportion of generated molecules that simultaneously satisfy $\text{Vina Dock} < -8.18$, $\text{QED} > 0.25$, and $\text{SA} > 0.59$. Molecular properties are measured using **QED** for drug-likeness and **SA** for synthetic accessibility, while sample diversity is quantified by **Div** (Luo et al., 2021).

### 4.2 Experiment Results

**How Well Does MolLIFT Perform in Pocket-Conditioned 3D Molecule Generation?** For each test protein, we generate 100 molecules for evaluation. Table 1 summarizes the performance of MolLIFT and prior pocket-conditioned generation methods.

**(i) Binding affinity and molecular properties of generated molecules.** Without property optimization, MolLIFT achieves the best binding-affinity results among the compared generative models, with average Vina Score, Vina Min, and Vina Dock of $-7.04$, $-7.41$, and $-7.95$, respectively. Compared with the strongest non-property-guided baseline MolCRAFT, these affinity metrics improve by $7.5\%$, $2.8\%$, and $3.7\%$, respectively. MolLIFT also improves QED by $8.0\%$, SA by $3.0\%$, diversity by $21.4\%$, and success rate by $1.4$ percentage points. When jointly optimizing the Vina and SA objectives, MolLIFT$_{\text{grad}}$ further achieves the best median Vina Score, the best average and median Vina Min and Vina Dock, and the best QED, diversity, and success rate, while attaining the second-best SA. Overall, these results show that integrating SMILES-derived chemical priors with local geometric information improves pocket-conditioned 3D generation without explicit property guidance, while gradient-based optimization can further strengthen affinity and molecular quality.

**(ii) Qualitative comparison.** Figure 3 compares generated ligands from TargetDiff, MolCRAFT, and MolLIFT within two representative protein pockets. In both examples, MolLIFT generates chemically plausible ligands with favorable docking scores while maintaining competitive QED and SA values. These examples qualitatively illustrate the balance between pocket compatibility and molecular quality observed in the quantitative results. Additional examples are provided in Appendix I.1.

**Does the SMILES Branch Produce Chemically Meaningful Outputs?** Although MolLIFT ultimately targets 3D ligand generation, its jointly evolving SMILES branch produces chemically meaningful sequence outputs. Figure 4 compares generated SMILES with the corresponding generated

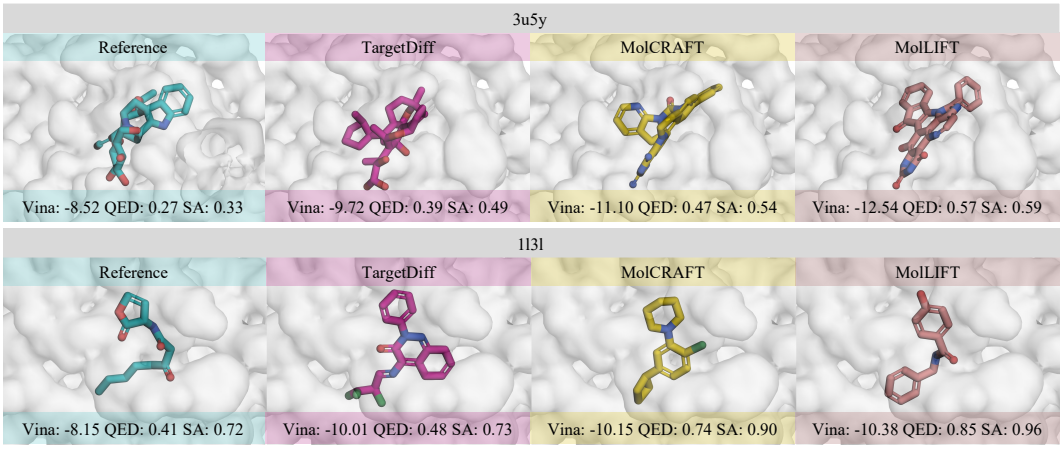



Figure 3: Qualitative comparison on two representative protein pockets. Each row shows the reference ligand and molecules generated by TargetDiff, MolCRAFT, and MolLIFT.

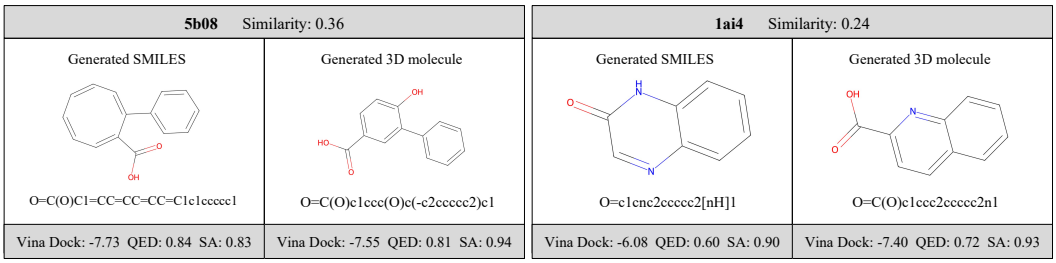



Figure 4: Generated SMILES and the corresponding generated 3D molecules for two representative protein pockets. Each pair reports Vina Dock, QED, SA, and the molecular similarity between the decoded SMILES and generated 3D molecule, computed as Morgan–Tanimoto similarity with radius 2 and 2048 bits.

3D molecules for two representative pockets. The decoded SMILES sequences exhibit favorable QED and SA values while retaining measurable molecular similarity to the generated 3D molecules, despite the absence of an explicit 1D–3D consistency constraint. This observation supports the role of the SMILES branch as a dynamic chemical prior that evolves jointly with the 3D branch during generation. Additional examples are provided in Appendix I.2.

**What Contributes to the Performance Gains of MolLIFT?** To understand where the improvements of MolLIFT come from, we examine model capacity, local-frame interaction, explicit coordinate lifting, and the effect of SMILES guidance on coordinate denoising. Each configuration is evaluated by generating 10 molecules for each of the 100 test proteins.

**(i) Is the gain simply due to increased model capacity?** To separate the effect of architecture from model size, we deepen MolCRAFT's EGNN to obtain a baseline with a similar, slightly larger parameter budget, denoted MolCRAFT$_{26L}$. As shown in Table 2, MolLIFT consistently outperforms this enlarged baseline across the reported affinity and molecular-property metrics. This comparison indicates that the observed gains are not explained by increased parameter count alone, but are associated with the proposed cross-modal interaction design.

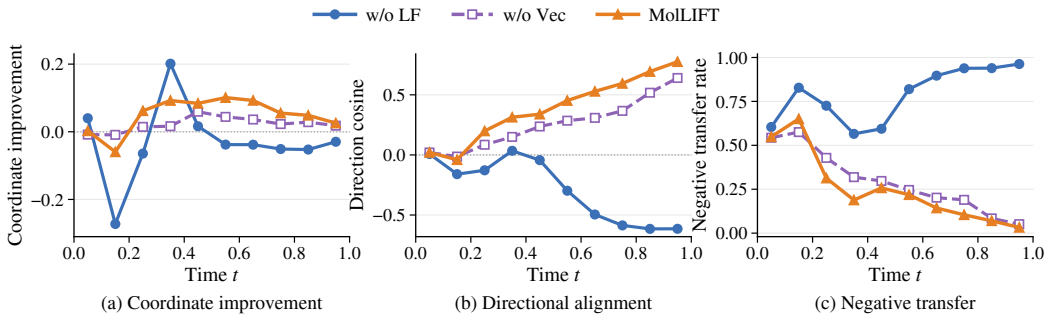
**(ii) How does local-frame interaction contribute?** MolLIFT w/o Local Frame (w/o LF) replaces the proposed local-frame interface with vanilla cross-attention, such that SMILES information is incorporated through scalar atom features without the proposed local spatial construction. Removing the local-frame interaction degrades all reported Vina metrics and reduces QED and SA. These results show that incorporating local geometric context into 1D–3D interaction is important for effectively using SMILES information during 3D generation.

**(iii) What is the effect of explicit coordinate lifting?** MolLIFT w/o Vector Message (w/o Vec) retains local-frame-aware scalar interaction but removes the explicit coordinate message $\Delta \mathbf{x}_{t,i}^{S \to X}$. Compared with w/o LF, this variant improves the Vina metrics and QED, indicating that geometry-aware scalar fusion already provides useful information to the geometric backbone. Restoring the explicit vector message leads to further improvements, showing that directly lifting sequence information into coordinate updates provides an additional benefit beyond scalar feature fusion.

**(iv) Does SMILES guidance improve coordinate denoising?** We further investigate how SMILES information affects the denoising trajectory by comparing predictions obtained with correct and null SMILES under identical pockets, noisy 3D states, and denoising times. Figure 5 reports coordinate improvement, directional alignment, and negative transfer rate across ten equal-width time bins over $[0, 1]$. Across most intermediate and late denoising stages, MolLIFT achieves larger coordinate improvements and stronger alignment with the residual correction direction than either ablation. The w/o LF variant shows less stable coordinate improvements and more frequent negative transfer, while w/o Vec alleviates these effects through geometry-aware scalar interaction. The full model further improves alignment and generally reduces harmful SMILES-induced updates. These results indicate that local geometric conditioning and explicit coordinate lifting make sequence guidance more effective and reliable throughout denoising. Complementary attention visualizations further reveal element-level correspondences between SMILES tokens and 3D ligand atoms, providing qualitative evidence that the cross-modal interaction captures chemically meaningful associations (Appendix I.3). The perturbation protocol and metric definitions are provided in Appendix F.

Table 2: Parameter-budget comparison and MolLIFT ablations.

| Method | Vina Score ↓ | | Vina Min ↓ | | QED ↑ | SA ↑ |
|---|---|---|---|---|---|---|
| | Avg. | Med. | Avg. | Med. | Avg. | Avg. |
| MolCRAFT$_{26L}$ | -6.51 | -6.68 | -7.01 | -6.89 | 0.49 | 0.67 |
| w/o LF | -6.68 | -6.82 | -7.13 | -7.04 | 0.51 | 0.67 |
| w/o Vec | -6.85 | -6.98 | -7.25 | -7.17 | 0.52 | 0.68 |
| **MolLIFT** | -7.01 | -7.12 | -7.37 | -7.29 | 0.54 | 0.68 |




(a) Coordinate improvement (b) Directional alignment (c) Negative transfer

Figure 5: Time-resolved analysis of SMILES guidance using correct-versus-null predictions under matched 3D states. Results are grouped into ten equal-width time bins. Higher coordinate improvement and directional alignment are better; lower negative transfer rate is better.

## 5 Conclusion

We introduced **MolLIFT**, a bidirectional 1D–3D Bayesian flow framework that integrates SMILES-derived chemical priors into pocket-conditioned 3D generation through geometry-aware cross-modal interaction and local-frame coordinate lifting. Experiments demonstrate improvements in binding affinity, molecular quality, diversity, and success rate, while matched-capacity ablations confirm that these gains arise from the proposed interaction design rather than increased model size. The joint 1D–3D state also supports gradient-guided property optimization. MolLIFT remains limited by imperfect consistency between decoded SMILES and generated 3D graphs, as well as the additional computational cost of joint sequence modeling, motivating more consistent and efficient cross-modal generation in future work.

## 6 AI USE STATEMENT

We used generative AI tools to assist with language editing, rephrasing, proofreading, and presentation polishing of the manuscript. We also used these tools to assist with plotting code and the visual refinement of experimental figures, based exclusively on experimental results produced by the authors. Generative AI was not used to fabricate or alter experimental data. All AI-assisted text, code, and figures were manually reviewed, edited, and verified by the authors against the underlying experiments and source data. The authors made and approved all scientific decisions, interpretations, and final claims, and take full responsibility for the content and artifacts presented in this work.

## REFERENCES

Vineeth Dorna, D Subhalingam, Keshav Kolluru, Shreshth Tuli, Mrityunjay Singh, Saurabh Singal, NM Krishnan, and Sayan Ranu. Tagmol: target-aware gradient-guided molecule generation. In *International Conference on Machine Learning*, 2024.

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, 2021.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. 3d convolutional neural networks and a crossdocked dataset for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200, 2020.

Tianfan Fu, Wenhao Gao, Connor Coley, and Jimeng Sun. Reinforced genetic algorithm for structure-based drug design. *Advances in Neural Information Processing Systems*, 2022.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Bowei He, Haichuan Tan, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. CIDD: Collaborative intelligence for structure-based drug design empowered by LLMs. In *Advances in Neural Information Processing Systems*, 2025.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. In *Advances in Neural Information Processing Systems*, 2024.

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 *International Conference on Learning Representations*, 2023a.

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 *International Conference on Machine Learning*, 2023b.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin Cui, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *International Conference on Learning Representations*, 2024.

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*, 2021.

Yue Luo and Lei Deng. Molcl-sp: a multimodal contrastive learning framework with non-overlapping substructure perturbations for molecular property prediction. *Bioinformatics*, 41(10): btaf507, 2025.

Marcos Martínez Galindo, Marco Luca Sbodio, Mykhaylo Zayats, Rodrigo Ordonez-Hurtado, Raúl Fernández-Díaz, Vanessa López García, and Hoang Thanh Lam. Impact of molecular multimodality on neural network models for prediction tasks related to drug discovery. *Nature Communications*, 2026.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, 2022.

Keyue Qiu, Yuxuan Song, Jie Yu, Hongbo Ma, Ziyao Cao, Zhilong Zhang, Yushuai Wu, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Empower structure-based molecule optimization with gradient guided bayesian flow networks. In *International Conference on Machine Learning*, 2025.

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. In *International Conference on Machine Learning*, 2024.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Jacob O Spiegel and Jacob D Durrant. Autogrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of cheminformatics*, 12(1):25, 2020.

Nianze Tao and Minori Abe. Bayesian flow network framework for chemistry tasks. *Journal of Chemical Information and Modeling*, 65(3):1178–1187, 2025.

Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *International Conference on Learning Representations*, 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 *International Conference on Learning Representations*, 2023.

Xiangxin Zhou, Xiwei Cheng, Yuwei Yang, Yu Bao, Liang Wang, and Quanquan Gu. Decompopt: Controllable and decomposed diffusion models for structure-based molecular optimization. In *International Conference on Learning Representations*, 2024.

# A Related Work

Structure-based drug design seeks molecules that combine pocket compatibility with favorable chemical properties. We distinguish property-agnostic 3D generation from property-aware generation and optimization according to whether explicit property information is used during training, sampling, or candidate refinement. The former learns from molecular structures and pocket context without property labels, property predictors, or property-based optimization objectives; the latter incorporates property supervision, guidance, or feedback to improve molecular quality.

**Property-agnostic 3D molecule generation.** These methods construct ligands from molecular structures and the geometric context of a protein pocket. Autoregressive approaches build molecules incrementally: AR (Luo et al., 2021) samples atoms from predicted voxel-wise densities using Markov chain Monte Carlo; GraphBP (Liu et al., 2022) combines atom-wise generation with normalizing flows and equivariant geometric context; and Pocket2Mol (Peng et al., 2022) predicts frontier atoms and generates new atoms and bonds with an E(3)-equivariant network. FLAG (Zhang et al., 2023) instead assembles fragments using predicted coordinates and torsion angles. Diffusion models generate molecular structures through iterative denoising. DiffSBDD (Schneuing et al., 2024) applies continuous diffusion to coordinates and continuous atom-type representations, whereas TargetDiff (Guan et al., 2023a) uses separate continuous and discrete diffusion processes for coordinates and atom types. DecompDiff (Guan et al., 2023b) incorporates structural priors obtained by decomposing ligands into pocket-contacting arms and linking scaffolds. Beyond diffusion, MolCRAFT (Qu et al., 2024) uses Bayesian Flow Networks (BFNs) to model coordinates and atom types in continuous parameter space. MolLIFT builds on this direction by jointly modeling a SMILES generation pathway to capture complementary chemical priors without explicit property supervision.

**Property-aware molecular generation and optimization.** Property information can enter molecular design through model training, sampling, or candidate refinement. At the model level, IPDiff (Huang et al., 2024) pretrains an interaction prior network with binding-affinity supervision and incorporates the learned priors into forward diffusion and sampling. AliDiff (Gu et al., 2024) uses preference optimization to align pretrained diffusion models with molecular rewards. During sampling, TAGMol (Dorna et al., 2024) applies property-predictor gradients to the coordinate diffusion process of a pretrained TargetDiff model, while MolJO (Qiu et al., 2025) jointly guides coordinates and atom types in BFN parameter space. Candidate-based optimization instead alternates molecular proposals with evaluation and selection. DecompOpt (Zhou et al., 2024) iteratively resamples molecules conditioned on top-ranked molecular arms selected by property oracles. AutoGrow4 (Spiegel & Durrant, 2020) combines genetic operations with docking-based selection, and RGA (Fu et al., 2022) learns target-aware policies to guide evolutionary search toward improved binding affinity. CIDD (Gao et al., 2025) couples 3D candidate generation with LLM-based refinement to improve chemical viability while preserving favorable protein–ligand interactions.

# B Training and Sampling

## B.1 Training Objective

MolLIFT is trained with joint 3D generation and SMILES denoising objectives. The coordinate loss is

$$\mathcal{L}_x = \mathbb{E}_t \left[ \sum_{i=1}^{N} \|\hat{\mathbf{x}}_{0,i} - \mathbf{x}_i\|_2^2 \right], \tag{21}$$

and the atom-type loss is

$$\mathcal{L}_v = \mathbb{E}_t \left[ \sum_{i=1}^{N} \mathrm{CE}\left(\mathbf{v}_i, \hat{\mathbf{p}}_{0,i}^{v}\right) \right]. \tag{22}$$

For SMILES denoising, we use a token-level denoising loss:

$$\mathcal{L}_s = \mathbb{E}_t \left[ \sum_{j=1}^{L} \mathrm{CE}\left(s_j, \hat{\mathbf{p}}_{0,j}^{s}\right) \right]. \tag{23}$$

The full objective is

$$\mathcal{L} = \mathcal{L}_x + \lambda_v \mathcal{L}_v + \lambda_s \mathcal{L}_s. \tag{24}$$

The SMILES denoising objective strengthens the sequence-level chemical representation used during denoising, while the final evaluation remains focused on the generated 3D ligand $(\hat{\mathbf{X}}_0, \hat{\mathbf{V}}_0)$.

**Cross-modal gradient isolation.** The SMILES-to-3D attention uses detached token features as keys and values. The 3D losses therefore train the geometric encoder, frame construction, coefficient predictor, attention module, learnable scale, and 3D backbone without updating the initial SMILES DiT through this path. In the reverse direction, final atom features, local descriptors, backbone displacements, and frames are detached before forming the learnable feedback memory. The SMILES loss can train the feedback projection, reverse attention, and both SMILES DiTs, but does not backpropagate into the 3D branch through that interface. Detaching the source features does not freeze the cross-modal mapping layers.

## B.2 Sampling

At inference time, MolLIFT initializes $\boldsymbol{\mu}_0^x$, $\boldsymbol{\theta}_0^v$, and $\boldsymbol{\theta}_0^s$ from uninformative Bayesian flow states. At each step, the initial SMILES DiT encodes the current SMILES belief, the local-frame module lifts SMILES priors into the 3D branch, the EGNN predicts clean ligand coordinates and atom types, and the reverse cross-attention refines the evolving SMILES belief. The Bayesian flow update is then applied to obtain the next states

$$(\boldsymbol{\mu}_{t+\Delta t}^x, \boldsymbol{\theta}_{t+\Delta t}^v, \boldsymbol{\theta}_{t+\Delta t}^s). \tag{25}$$

After the final step, MolLIFT outputs only the 3D ligand

$$\hat{\mathbf{m}}^{3D} = (\hat{\mathbf{X}}_0, \hat{\mathbf{V}}_0). \tag{26}$$

The generated SMILES sequence can be decoded for analysis, but it is not required to be exactly graph-consistent with the final generated 3D molecule.

# C ALGORITHM

## C.1 JOINT TRAINING

For each complex, we sample one time $t$ and independently form the continuous coordinate belief and the categorical atom and SMILES beliefs using their BFN sender distributions. The network predicts their clean endpoints and is trained using the joint BFN objective in Equation 24. Algorithm 1 makes the forward order and gradient routing explicit.

**Algorithm 1** Joint training of MolLIFT

**Require:** Complex $(\mathcal{P}, \mathbf{X}_0, \mathbf{V}_0, \mathbf{S}_0)$; parameters $\psi$; weights $\lambda_v, \lambda_s$

Sample $t \sim \mathcal{U}(t_{\min}, 1)$
Corrupt $(\mathbf{X}_0, \mathbf{V}_0, \mathbf{S}_0)$ to obtain $(\boldsymbol{\mu}_t^x, \boldsymbol{\theta}_t^v, \boldsymbol{\theta}_t^s)$
$\mathbf{H}_t^S \leftarrow D_S^{\text{init}}(\boldsymbol{\theta}_t^s, t)$
$(\mathbf{g}_t, \mathbf{q}_t, \mathbf{F}_t) \leftarrow \mathrm{GCE}(\boldsymbol{\mu}_t^x, \boldsymbol{\theta}_t^v, \mathcal{P}, t)$
$(\bar{\mathbf{H}}_t^X, \Delta\mathbf{X}_t^{S\to X}) \leftarrow \mathrm{LFL}(\mathbf{q}_t, \mathrm{sg}(\mathbf{H}_t^S), \mathbf{F}_t)$
$(\hat{\mathbf{X}}_0^{\mathrm{3D}}, \hat{\mathbf{P}}_0^v, \mathbf{H}_0^X) \leftarrow \mathrm{EGNN}(\bar{\mathbf{H}}_t^X, \boldsymbol{\mu}_t^x, \mathcal{P}, t)$
$\hat{\mathbf{X}}_0 \leftarrow \hat{\mathbf{X}}_0^{\mathrm{3D}} + \Delta\mathbf{X}_t^{S\to X}$
$\mathbf{M}_t^X \leftarrow \mathrm{IFB}\left(\mathrm{sg}(\mathbf{H}_0^X), \mathrm{sg}(\mathbf{F}_t), \mathrm{sg}(\hat{\mathbf{X}}_0^{\mathrm{3D}} - \boldsymbol{\mu}_t^x)\right)$
$\hat{\mathbf{P}}_0^s \leftarrow D_S^{\text{ref}}\left(\mathrm{MHA}(\mathbf{H}_t^S, \mathbf{M}_t^X, \mathbf{M}_t^X), t\right)$
Update $\psi$ using Eq. equation 24

## C.2 GENERATION

At inference time, the three beliefs are initialized from their uninformative priors. At every discretized time step, MolLIFT predicts clean coordinate, atom-type and token distributions, after which the corresponding BFN receiver updates all three beliefs. Importantly, LFL and IFB are recomputed from the current state rather than applied once: sequence information is repeatedly grounded in the evolving 3D geometry, and the resulting geometric progress is re-encoded for the next token prediction. The procedure is given in Algorithm 2.

**Algorithm 2** Joint Bayesian-flow generation with MolLIFT

**Require:** Pocket $\mathcal{P}$, ligand atom count $N_L$, number of steps $T$

Initialize $\boldsymbol{\mu}_0^x \leftarrow \mathbf{0}$ and uniform categorical beliefs $\boldsymbol{\theta}_0^v, \boldsymbol{\theta}_0^s$
**for** $n = 0, \ldots, T-1$ **do**
    $t_n \leftarrow (n+1)/T$
    $(\hat{\mathbf{X}}_0, \hat{\mathbf{P}}_0^v, \hat{\mathbf{P}}_0^s) \leftarrow \mathrm{MolLIFT}(\boldsymbol{\Theta}_{t_n}, \mathcal{P}, t_n)$
    Sample sender observations from the predicted clean distributions
    Update $(\boldsymbol{\mu}^x, \boldsymbol{\theta}^v, \boldsymbol{\theta}^s)$ with the continuous and categorical BFN receiver rules
**end for**
Decode atom types and return the generated 3D ligand

# D IMPLEMENTATION DETAILS

**Architecture.** MolLIFT is implemented as a joint 3D–1D Bayesian flow model that simultaneously generates ligand conformations, atom types, and SMILES sequences. The 3D branch adopts an equivariant UniTransformer backbone, consisting of one block with nine message-passing layers. The hidden dimension is set to 128, and each layer uses 16 attention heads. A $k$-nearest-neighbor graph with $k = 32$ is constructed over protein and ligand atoms. Geometric edge features are encoded using 20 Gaussian radial basis functions with a cutoff radius of 10 Å. The input dimensions of the protein and ligand atom features are 27 and 13, respectively.

The SMILES branch contains an eight-layer ChemBFN Transformer followed by an eight-layer refinement Transformer. Both Transformers use a hidden dimension of 128, eight attention heads, a feed-forward dimension of 512, a dropout rate of 0.1, and rotary positional embeddings. The maximum SMILES sequence length is 128, and the vocabulary contains 246 tokens.

To exchange information between the two modalities, we introduce bidirectional cross-modal interactions. First, the SMILES token representations are projected to ligand atoms through a local-frame cross-attention module. The module constructs equivariant local coordinate frames from relative atomic positions and predicts both invariant atom features and equivariant coordinate messages. The coordinate messages are transformed from the local frames to the global coordinate system and added to the coordinate prediction of the 3D denoiser. After 3D message passing, the updated ligand representations are propagated back to the SMILES branch using a multi-head attention layer with eight heads, followed by residual normalization and the refinement Transformer. We stop the gradients from the SMILES loss at the 3D-to-SMILES interface to prevent the discrete sequence objective from destabilizing the coordinate denoising backbone. The complete model contains approximately 8.0 million parameters.

**Training Details.** We train MolLIFT from scratch on the processed CrossDocked dataset. The training split contains 99,990 protein–ligand complexes. All model components, including the 3D denoiser, the SMILES Transformers, and the local-frame interaction module, are jointly optimized.

We use the Adam optimizer with an initial learning rate of $5 \times 10^{-4}$, $\beta_1 = 0.95$, $\beta_2 = 0.999$, and zero weight decay. The model is trained with a batch size of 16 for 15 epochs, corresponding to 93,750 optimization steps. The learning rate is adjusted using a `ReduceLROnPlateau` scheduler with a decay factor of 0.6, a patience of 10 validation intervals, and a minimum learning rate of $10^{-6}$. An exponential moving average of the model parameters is maintained with a decay rate of 0.999.

For the Bayesian flow process, we set the terminal noise parameter of the continuous coordinate channel to $\sigma_1 = 0.03$ and the accuracy parameter of the categorical atom-type channel to $\beta_1 = 1.5$. The minimum continuous time is set to $10^{-4}$, and the flow process is discretized into 1,000 time steps. The coordinate, atom-type, and SMILES objectives are assigned equal weights:

$$\mathcal{L} = \mathcal{L}_x + \mathcal{L}_v + \mathcal{L}_s. \tag{27}$$

We employ adaptive gradient clipping. Specifically, the clipping threshold is determined from a queue containing the most recent 3,000 gradient norms:

$$\tau_{\text{grad}} = 1.5\mu_{\text{grad}} + 2\sigma_{\text{grad}}, \tag{28}$$

where $\mu_{\text{grad}}$ and $\sigma_{\text{grad}}$ denote the running mean and standard deviation, respectively. No additional coordinate noise or random rotation augmentation is used. All experiments use a random seed of 1234. The model is trained on a single NVIDIA A800-SXM4-80GB GPU and completes training in approximately 17 hours.

**Sampling Details.** During inference, we discretize the Bayesian flow sampling trajectory into 100 steps. The number of generated ligand atoms is set to the atom count of the corresponding reference ligand. For evaluation, we generate 100 ligand candidates for each protein pocket. Binding affinity is evaluated using AutoDock Vina in the Vina Score mode with an exhaustiveness value of 16. Unless otherwise specified, neither backward correction nor classifier guidance is applied during unconditional sampling.

## E Equivariance Analysis

We establish equivariance of the cross-modal denoising interface in the direct clean-coordinate prediction parameterization. Consider a fixed denoiser input transformed by $\boldsymbol{\mu}^x_{t,i} \mapsto R\boldsymbol{\mu}^x_{t,i} + \mathbf{b}$ and $\mathbf{x}^C_{t,k} \mapsto R\mathbf{x}^C_{t,k} + \mathbf{b}$, where $R \in SO(3)$ and $\mathbf{b} \in \mathbb{R}^3$. We assume invariant scalar input features, a distance-based neighborhood rule with coordinate-independent tie-breaking, and an SE(3)-equivariant 3D backbone. The argument applies to nonempty neighborhoods; degenerate local frames are handled by the zero-frame fallback described in the main text.

**Invariant context and attention.** The relative vectors and distances satisfy

$$\mathbf{r}'_{t,ki} = R\mathbf{r}_{t,ki}, \qquad d'_{t,ki} = d_{t,ki}. \tag{29}$$

Thus, the edge inputs, local descriptors, geometry-augmented queries, attention weights, and aggregated SMILES messages are invariant. The scalar feature update and local coefficients are consequently invariant as well:

$$\bar{\mathbf{h}}^{X\prime}_{t,i} = \bar{\mathbf{h}}^{X}_{t,i}, \qquad \mathbf{c}^{S\prime}_{t,i} = \mathbf{c}^{S}_{t,i}. \tag{30}$$

Stop-gradient changes derivatives but not forward values, and therefore does not affect these transformation properties. Vanilla scalar cross-attention is invariant for the same reason; local-frame lifting adds an explicit equivariant vector channel rather than correcting an equivariance violation in vanilla attention.

**Equivariant local frames and coordinate predictions.** The anchor weights are invariant, so $\mathbf{u}'_{t,i} = R\mathbf{u}_{t,i}$ and $\mathbf{w}'_{t,i} = R\mathbf{w}_{t,i}$. Norms and inner products are preserved by $R$, including in the numerically stabilized normalization. Hence the first two frame vectors transform by $R$. For $R \in SO(3)$, the cross product satisfies $(R\mathbf{a}) \times (R\mathbf{b}) = R(\mathbf{a} \times \mathbf{b})$, giving

$$\mathbf{F}'_{t,i} = R\mathbf{F}_{t,i}. \tag{31}$$

The norm-based degeneracy test is invariant. If it sets the frame to zero, the transformation rule still holds because $R\mathbf{0} = \mathbf{0}$. This argument does not require the stabilized frame to be exactly orthonormal. Combining the frame transformation with invariant coefficients gives

$$\Delta\mathbf{x}^{S\to X\prime}_{t,i} = \kappa(R\mathbf{F}_{t,i})\mathbf{c}^{S}_{t,i} = R\Delta\mathbf{x}^{S\to X}_{t,i}. \tag{32}$$

Since the backbone satisfies $\hat{\mathbf{x}}^{\mathrm{EGNN}\prime}_{0,i} = R\hat{\mathbf{x}}^{\mathrm{EGNN}}_{0,i} + \mathbf{b}$, the complete coordinate prediction obeys

$$\begin{aligned} \hat{\mathbf{x}}'_{0,i} &= R\hat{\mathbf{x}}^{\mathrm{EGNN}}_{0,i} + \mathbf{b} + R\Delta\mathbf{x}^{S\to X}_{t,i} \\ &= R\hat{\mathbf{x}}_{0,i} + \mathbf{b}. \end{aligned} \tag{33}$$

Atom-type probabilities remain invariant because they are predicted from invariant scalar features. The result guarantees SE(3)-equivariance, not equivariance to reflections: the cross product acquires an additional determinant factor under improper rotations.

**Invariant feedback to SMILES.** The backbone displacement transforms as $\Delta\mathbf{x}^{\mathrm{EGNN}\prime}_{t,i} = R\Delta\mathbf{x}^{\mathrm{EGNN}}_{t,i}$. Consequently, ignoring stop-gradient notation, which leaves forward values unchanged,

$$\mathbf{r}^{X\prime}_{t,i} = (R\mathbf{F}_{t,i})^{\top} R\Delta\mathbf{x}^{\mathrm{EGNN}}_{t,i} = \mathbf{F}^{\top}_{t,i}\Delta\mathbf{x}^{\mathrm{EGNN}}_{t,i} = \mathbf{r}^{X}_{t,i}. \tag{34}$$

Both the atom memory and the subsequent token-query attention therefore remain invariant. The bidirectional interface preserves equivariant coordinate predictions and invariant atom-type and SMILES predictions.

## F Mechanistic Analysis of SMILES Guidance

**Controlled perturbations.** We evaluate trained models without further optimization. For each protein pocket and noisy 3D state, the protocol considers four SMILES inputs: the correct sequence, a sequence randomly reassigned within the batch, a null input (fully masked or empty SMILES), and the correct sequence with randomly permuted token order. These controls probe sequence identity, the presence of SMILES information, and token order. We hold the pocket, denoising noise, time $t$, initial coordinates, initial atom-type belief, and random seed fixed across paired evaluations. The corresponding coordinate predictions are denoted by $\hat{\mathbf{X}}^{\mathrm{true}}$, $\hat{\mathbf{X}}^{\mathrm{shuffle}}$, $\hat{\mathbf{X}}^{\mathrm{null}}$, and $\hat{\mathbf{X}}^{\mathrm{perm}}$. Figure 5 reports only the correct-versus-null comparison; the other perturbations are not shown.

**Coordinate improvement.** For evaluation instance $n$ containing $N_n$ ligand atoms, let $\mathbf{X}^{\mathrm{clean}}_n$ be the clean reference coordinates. We define

$$\begin{aligned} E^a_n &= \mathrm{MSE}(\hat{\mathbf{X}}^a_n, \mathbf{X}^{\mathrm{clean}}_n) = \frac{1}{3N_n}\sum_{i=1}^{N_n} \|\hat{\mathbf{x}}^a_{n,i} - \mathbf{x}^{\mathrm{clean}}_{n,i}\|_2^2, \\ \Delta_{\mathrm{coord},n} &= E^{\mathrm{null}}_n - E^{\mathrm{true}}_n. \end{aligned} \tag{35}$$

Positive coordinate improvement means that the correct SMILES reduces coordinate error relative to the null input; negative values indicate deterioration.

**Negative transfer rate.** For a set $\mathcal{B}$ of matched evaluation instances, we define

$$\mathrm{NTR}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \mathbb{I}[E_n^{\mathrm{true}} > E_n^{\mathrm{null}}]. \tag{36}$$

NTR measures the fraction of instances for which correct SMILES worsens coordinate prediction. Lower values indicate more reliable guidance. This complements mean coordinate improvement by measuring the frequency, rather than the magnitude, of harmful changes.

**Directional alignment.** Using the null prediction as the reference, we define the desired residual correction and the SMILES-induced displacement for each atom:

$$\mathbf{r}_{n,i} = \mathbf{x}_{n,i}^{\mathrm{clean}} - \hat{\mathbf{x}}_{n,i}^{\mathrm{null}}, \qquad \mathbf{d}_{n,i}^{S} = \hat{\mathbf{x}}_{n,i}^{\mathrm{true}} - \hat{\mathbf{x}}_{n,i}^{\mathrm{null}}. \tag{37}$$

Their stabilized cosine similarity is

$$C_{n,i} = \frac{(\mathbf{d}_{n,i}^{S})^{\top} \mathbf{r}_{n,i}}{\|\mathbf{d}_{n,i}^{S}\|_2 \|\mathbf{r}_{n,i}\|_2 + \epsilon}, \tag{38}$$

where $\epsilon > 0$ stabilizes the denominator. Positive values indicate a displacement component toward the clean coordinates, negative values indicate a component in the opposite direction, and near-zero values indicate weak alignment or negligible displacement. Positive alignment alone does not guarantee reduced coordinate error because the displacement may overshoot. Importantly, $\mathbf{d}_{n,i}^{S}$ measures the total effect of changing the SMILES input through both scalar and vector pathways; it is not the explicit vector message in isolation.

Let $A_{\mathcal{B}} = \sum_{n \in \mathcal{B}} N_n$ denote the number of evaluated atoms. The mean alignment and the fraction of positively aligned atoms are

$$\overline{C}(\mathcal{B}) = \frac{1}{A_{\mathcal{B}}} \sum_{n \in \mathcal{B}} \sum_{i=1}^{N_n} C_{n,i}, \qquad P_{\mathrm{positive}}(\mathcal{B}) = \frac{1}{A_{\mathcal{B}}} \sum_{n \in \mathcal{B}} \sum_{i=1}^{N_n} \mathbb{I}[C_{n,i} > 0]. \tag{39}$$

The directional panel reports $\overline{C}$; $P_{\mathrm{positive}}$ is a complementary statistic not displayed in the figure.

**Time-resolved aggregation.** We divide the time interval $[0, 1]$ into ten bins of equal width. Each bin includes its left endpoint but excludes its right endpoint, except for the final bin, which includes $t = 1$. Within each bin, we compute mean coordinate improvement and NTR over matched instances and mean directional alignment over their atoms. Results are summarized separately for w/o LF (ordinary cross-attention), w/o Vec, and MolLIFT, revealing how the usefulness of SMILES guidance varies along the denoising trajectory.

# G More Experimental Results

## G.1 Property Prediction with 3D and SMILES Representations

To investigate the complementary information captured by the two modalities, we extract molecular representations using pretrained Uni-Mol(Zhou et al., 2023) 3D and ChemBFN(Tao & Abe, 2025) SMILES encoders and train MLP predictors with the same parameter budget. For Vina prediction, which depends on both the ligand and its binding environment, we additionally encode the protein pocket using the pretrained Uni-Mol pocket encoder and combine the resulting pocket embedding with the molecular representation from each modality. For the ligand-intrinsic properties SA and QED, only the corresponding molecular representations are used. Prediction performance is evaluated using mean absolute error (MAE) and root mean squared error (RMSE).

As shown in Figure 1, 3D representations are more predictive of Vina-related binding affinity, whereas SMILES representations better capture SA and QED. This distinction is consistent with the information emphasized by the two modalities: 3D representations explicitly characterize molecular geometry and, when combined with pocket representations, provide information relevant to spatial complementarity and protein–ligand interactions. In contrast, SMILES representations encode sequential and topological chemical patterns, including functional-group composition, ring systems, synthetic regularities, and drug-like substructures. These results indicate that SMILES captures chemical priors that are complementary to 3D geometric information, motivating its integration as a complementary modality in 3D generation.

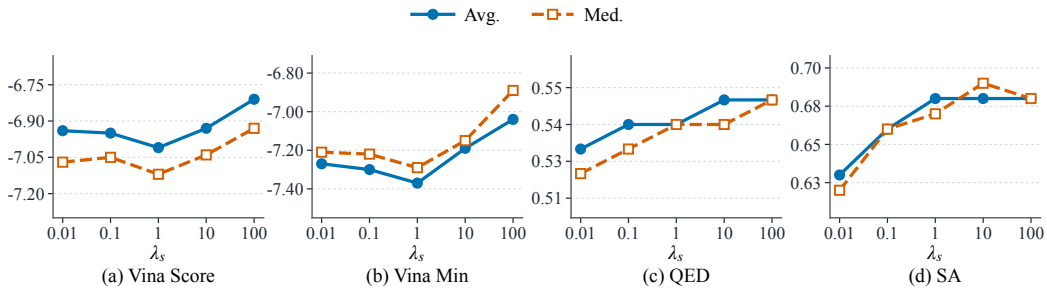



Figure 6: Parameter sensitivity analysis of the SMILES loss weight $\lambda_s$ in MolLIFT.

### G.2 Parameter Analysis

We set $\lambda_v = 1$, following the default configuration of MolCRAFT, and therefore focus our analysis on the SMILES loss weight $\lambda_s$. As shown in Figure 6, QED and SA generally improve as $\lambda_s$ increases, demonstrating that incorporating SMILES supervision provides useful chemical priors for drug-likeness and synthetic accessibility. However, these gains gradually saturate at larger values of $\lambda_s$, suggesting that the relevant sequence-level information has been sufficiently exploited and that further emphasizing the SMILES objective provides limited additional benefit.

The affinity-related metrics exhibit a non-monotonic trend: Vina Score and Vina Min initially improve as $\lambda_s$ increases, but deteriorate when the SMILES loss is assigned an excessively large weight. This result indicates that moderate SMILES supervision can also benefit affinity-oriented 3D structure generation by encouraging chemically plausible ligand patterns. In contrast, excessive sequence supervision can shift optimization away from the geometric objective required to model protein–ligand interactions. Overall, $\lambda_s = 1$ provides the best balance between molecular properties and binding affinity, and is therefore used as the default setting in our experiments.

## H Gradient-Guided Multi-Objective Optimization

MolLIFT can be extended from property-agnostic generation to plug-and-play molecular optimization without retraining the generative model. Inspired by gradient-guided Bayesian flow optimization (Qiu et al., 2025), we apply property gradients directly to the continuous BFN belief parameters of ligand coordinates, atom types, and the jointly modeled SMILES tokens. This guided variant is denoted by $\text{MolLIFT}_{\text{grad}}$. Unlike post-generation filtering, the property signal participates in every sampling step and jointly steers the three modalities while retaining the pretrained MolLIFT prior.

**Differentiable property surrogates.** Following MolJO (Qiu et al., 2025), we restrict the main guided experiments to binding affinity and synthetic accessibility. MolJO reports that jointly optimizing affinity, QED, and SA yields a more balanced property profile but reduces peak affinity performance; it therefore uses the Affinity+SA objectives in its main experiments. We adopt the same two-objective setting to retain strong affinity optimization while maintaining synthesizability and to enable a direct comparison with MolJO. For each target property $q \in \{\text{Vina}, \text{SA}\}$, we train a differentiable surrogate $f_q$ on randomly corrupted joint BFN states. The predictor uses a copy of the MolLIFT backbone to encode the current pocket-conditioned 3D state and the relaxed SMILES state. Mean-pooled ligand-atom features, masked-mean-pooled token features, and the current time are concatenated and passed through a two-layer prediction head:

$$r_q(\Theta_t, \mathcal{P}, t) = \sigma\Big(\text{MLP}_q\big[\,\text{MeanPool}(H_t^X),\ \text{MaskedMean}(H_t^S), t\big]\Big), \tag{40}$$

where $r_q \in (0, 1)$ and a larger value is always preferred. Raw Vina energies are clipped to $[-16, 0]$ and transformed as $r_{\text{Vina}} = -\text{clip}(E_{\text{Vina}}, -16, 0)/16$, whereas SA is already represented on a

higher-is-better $[0, 1]$ scale. During guided sampling, all surrogate parameters are frozen and gradients are propagated only to the current BFN state.

For multi-objective optimization, we combine the normalized predictions through a weighted log-utility:

$$\begin{aligned} \mathcal{J}_t &= \frac{1}{\sum_{q\in\mathcal{Q}} \omega_q} \sum_{q\in\mathcal{Q}} \omega_q \log(r_q(\Theta_t, \mathcal{P}, t) + \varepsilon)\,, \\ \mathcal{Q} &= \{\mathrm{Vina}, \mathrm{SA}\}. \end{aligned} \tag{41}$$

The logarithm prevents a property with an already large normalized score from dominating the optimization and penalizes candidates that perform poorly on either objective. We use equal objective weights, $\omega_{\mathrm{Vina}} = \omega_{\mathrm{SA}} = 1$.

**Joint guidance in the hybrid BFN state.** Let $\ell_t^v = \log(\theta_t^v + \varepsilon)$ denote the atom-type logits associated with the categorical belief. We compute gradients with respect to the continuous coordinate mean, atom-type logits, and relaxed SMILES belief:

$$g_t^x = \nabla_{\mu_t^x} \mathcal{J}_t, \qquad g_t^v = \nabla_{\ell_t^v} \mathcal{J}_t, \qquad g_t^s = \nabla_{\theta_t^s} \mathcal{J}_t. \tag{42}$$

Suppose the MolLIFT denoiser predicts clean coordinates $\widehat{X}_0$, atom-type probabilities $\widehat{P}_0^v$, and SMILES-token probabilities $\widehat{P}_0^s$. We modify these predictions before the Bayesian update:

$$\widetilde{X}_0 = \widehat{X}_0 + \eta_x\big(1 - \gamma_x(t)\big) g_t^x, \tag{43}$$

$$\widetilde{P}_0^v = \mathrm{softmax}\Big(\log(\widehat{P}_0^v + \varepsilon) + \eta_v g_t^v\Big)\,, \tag{44}$$

$$\widetilde{P}_0^s = \mathrm{softmax}\Big(\log(\widehat{P}_0^s + \varepsilon) + \eta_s M_s \odot g_t^s\Big)\,, \tag{45}$$

where $\gamma_x(t) = 1 - \sigma_1^{2t}$, $M_s$ masks padded or otherwise fixed token positions, and $(\eta_x, \eta_v, \eta_s)$ are modality-specific guidance scales. The factor $1-\gamma_x(t)$ adapts coordinate guidance to the uncertainty of the continuous belief. For categorical variables, guidance is applied in log-probability space followed by a softmax, which keeps atom-type and token predictions on their probability simplices. After SMILES guidance, the vocabulary mask, sampled-length constraint, and fixed `BOS`/`EOS`/`PAD` positions are applied again.

**Guided Bayesian update.** At sampling step $i$, let $t_i = i/n$. The `end_back_pmf` sampler used in our experiments reconstructs the next coordinate and atom-type beliefs from the guided clean-state predictions:

$$\mu_{t_i}^x = \gamma_x(t_i)\widetilde{X}_0 + \sqrt{\gamma_x(t_i)\big(1 - \gamma_x(t_i)\big)}\,\epsilon_i^x, \tag{46}$$

$$\theta_{t_i}^v = \mathrm{softmax}\Big(\beta_v(t_i)\big(K_v \widetilde{P}_0^v - \mathbf{1}\big) + \sqrt{\beta_v(t_i) K_v}\,\epsilon_i^v\Big)\,, \tag{47}$$

where $\epsilon_i^x, \epsilon_i^v \sim \mathcal{N}(0, I)$ and $\beta_v(t) = \beta_1^v t^2$. For SMILES, a token is sampled from $\widetilde{P}_0^s$ and communicated through the categorical BFN sender. With $\alpha_i^s = \beta_1^s(t_i^2 - t_{i-1}^2)$, the update is

$$\begin{aligned} &\widetilde{s}_i \sim \mathrm{Cat}(\widetilde{P}_0^s), \\ &y_i^s \sim \mathcal{N}(\alpha_i^s(K_s e_{\widetilde{s}_i} - \mathbf{1}), \alpha_i^s K_s I)\,, \end{aligned} \tag{48}$$

$$\theta_{t_i}^s = \frac{\theta_{t_{i-1}}^s \odot \exp(y_i^s)}{\mathbf{1}^\top\Big(\theta_{t_{i-1}}^s \odot \exp(y_i^s)\Big)}. \tag{49}$$

The updated beliefs are then passed to the next MolLIFT denoising step. After the final step, only the 3D ligand $(\widehat{X}_0, \widehat{V}_0)$ is retained as the model output; SMILES remains a sequence-level optimization channel.

**Sampling procedure and implementation.** Starting from uninformative coordinate, atom-type, and SMILES beliefs, each iteration performs four operations: (1) MolLIFT predicts the clean 3D ligand and joint SMILES distribution; (2) the frozen Vina and SA surrogates evaluate the current relaxed joint state; (3) gradients of Equation 41 modify the three clean-state predictions according

to Equation 43–45; and (4) the guided predictions are incorporated into the next beliefs using Equation 46– 49. The first update is unconditional because the belief at $t = 0$ is uninformative; guidance is applied at all remaining steps.

For the reported MOLLIFT$_{\text{grad}}$ experiment, we use 200 sampling steps, sample the ligand atom count from the reference ligand prior, and set $\eta_x = \eta_v = \eta_s = 50$. We generate 100 molecules for each pocket. The Vina and SA predictors use separately fine-tuned full MolLIFT backbones and the same joint 1D–3D mean-pooling head in Equation 40. No gradient clipping is applied in this run.

**Ablation of guidance channels.** Table 3 evaluates the contribution of coordinate, atom-type, and SMILES guidance by disabling the property gradient for one modality at a time. The complete MOLLIFT$_{\text{grad}}$ model achieves the best result on every reported metric, showing that the three guidance channels provide complementary signals.

Table 3: Ablation of gradient-guidance channels in MOLLIFT$_{\text{grad}}$. Each w/o variant disables the property gradient for the indicated modality while retaining guidance for the other two. Each configuration is evaluated by generating 10 molecules for each of the 100 test proteins. Top 2 results are **bolded** and underlined, respectively.

| Method | Vina Score ↓ | | Vina Min ↓ | | QED ↑ | SA ↑ |
|---|---|---|---|---|---|---|
| | Avg. | Med. | Avg. | Med. | Avg. | Avg. |
| w/o type | <u>-7.23</u> | <u>-7.79</u> | <u>-7.82</u> | <u>-7.89</u> | 0.54 | 0.70 |
| w/o coord | -6.91 | -7.47 | -7.16 | -7.24 | <u>0.55</u> | <u>0.73</u> |
| w/o SMILES | -7.19 | -7.62 | -7.35 | -7.38 | 0.52 | 0.68 |
| MolLIFT$_{\text{grad}}$ | **-7.41** | **-8.07** | **-8.35** | **-8.39** | **0.56** | **0.75** |

Removing coordinate guidance causes the largest degradation in Vina Score and Vina Min while retaining relatively strong QED and SA, confirming that coordinate gradients are the primary channel for improving pocket-aware binding geometry. Removing SMILES guidance produces the lowest QED and SA and also weakens affinity, indicating that sequence-level gradients supply chemical-quality priors that complement geometric optimization. Disabling atom-type guidance leads to smaller but consistent losses across both affinity and molecular-property metrics. Together, these results show that coordinate guidance mainly controls spatial interactions, SMILES guidance strengthens molecular quality, and atom-type guidance provides an additional compositional signal; jointly guiding all three modalities yields the most balanced performance.

**Equivariance.** The property surrogates combine an equivariant geometric backbone with invariant pooling and scalar property heads. Therefore, for a rigid transformation $(R, b)$ of the protein–ligand complex,

$$\mathcal{J}_t(R\mu_t^x + b, RX^P + b, \theta_t^v, \theta_t^s) = \mathcal{J}_t(\mu_t^x, X^P, \theta_t^v, \theta_t^s), \tag{50}$$

which implies $g_t^{x\prime} = Rg_t^x$. Hence, Equation 43 preserves SE(3)-equivariance, while atom-type and SMILES guidance remain invariant.

**Relation to backward correction.** MolJO additionally proposes a sliding-window backward correction that rebuilds the current belief from an earlier state $\Theta_{t_{i-k}}$ and a guided aggregate observation over $(t_{i-k}, t_i]$. This mechanism can reduce the influence of suboptimal past observations and provides an exploration–exploitation trade-off controlled by the window size $k$. Our implementation supports this optional update for coordinates and atom types. However, the experiment reported here disables backward correction ($k = 0$) and uses `end_back_pmf`; consequently, all results attributed to MOLLIFT$_{\text{grad}}$ arise from joint coordinate–type–SMILES gradient guidance rather than sliding-window history correction.

# I MORE VISUALIZATIONS

## I.1 ADDITIONAL 3D MOLECULE VISUALIZATIONS

Figure 7 provides additional qualitative comparisons across six protein pockets. The examples complement Figure 3 and illustrate the pocket-conditioned structures generated by MolLIFT relative to TargetDiff and MolCRAFT.

## I.2 ADDITIONAL SMILES–3D VISUALIZATIONS

Figure 8 provides additional examples of the jointly generated SMILES and 3D ligand outputs. Across diverse pockets, the decoded SMILES sequences maintain favorable molecular-property values and show nonzero structural similarity to their 3D counterparts, further illustrating their role as informative co-evolving chemical states.

## I.3 CROSS-MODAL ATTENTION VISUALIZATION

We visualize the attention weights between generated SMILES tokens and 3D ligand atoms, averaging the weights across all eight cross-attention heads. Figure 9 presents nine examples, with SMILES tokens along the rows and 3D atoms along the columns. The heatmaps reveal a noticeable element-level correspondence: tokens representing carbon, nitrogen, oxygen, and sulfur often assign higher attention weights to 3D atoms of the corresponding element. This pattern suggests that the cross-modal interface learns chemically meaningful associations between sequence and 3D representations.

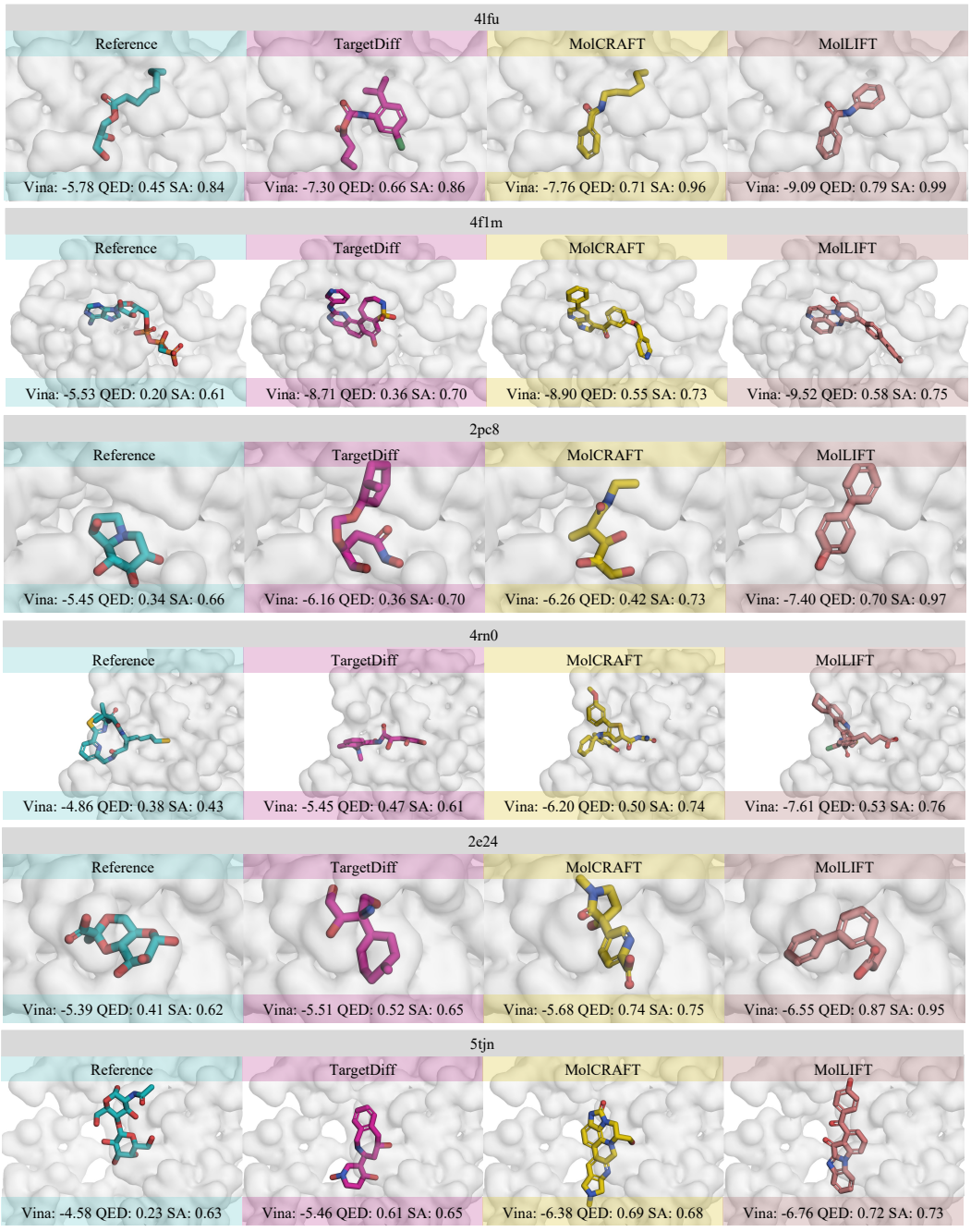



Figure 7: Additional qualitative comparisons across six protein pockets. Each row shows the reference ligand and outputs from TargetDiff, MolCRAFT, and MolLIFT, with the corresponding Vina, QED, and SA values.

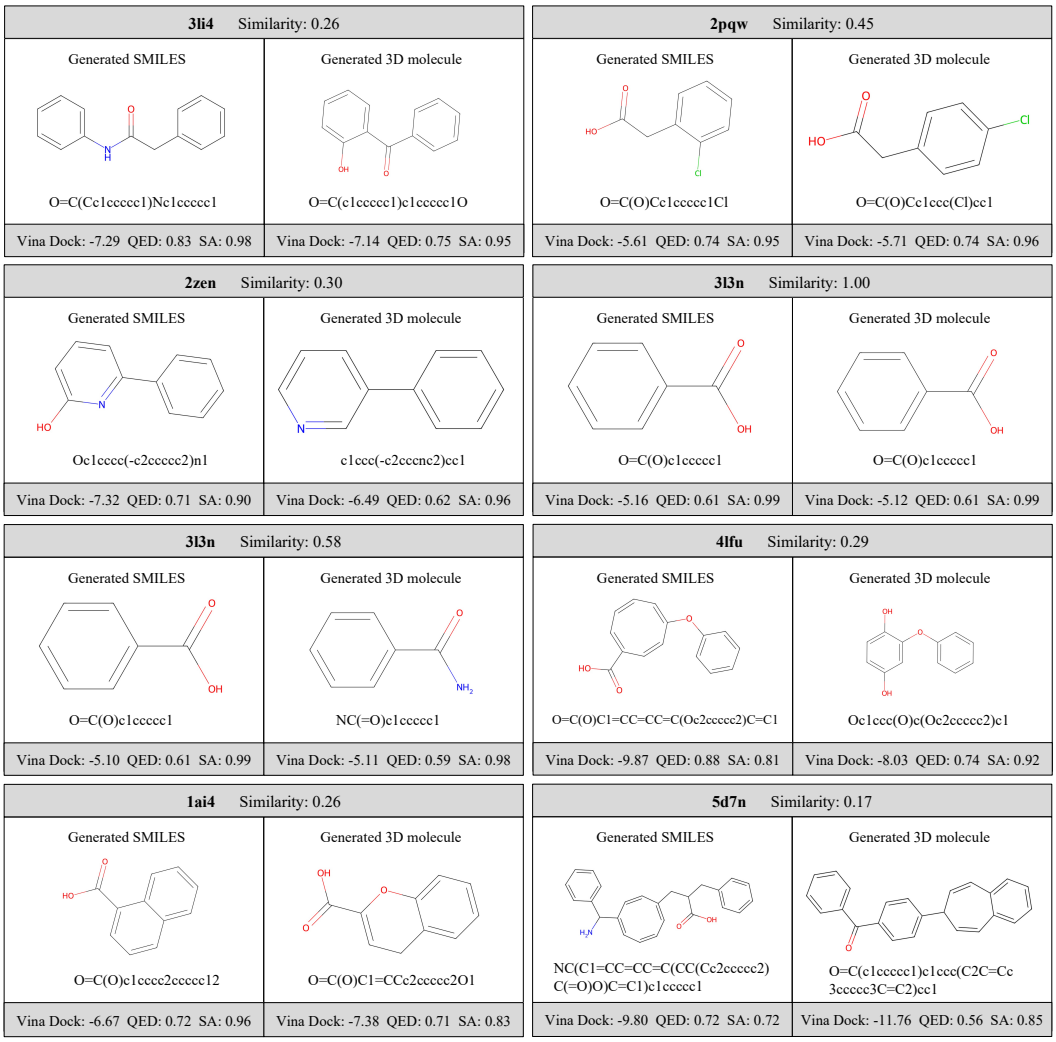



Figure 8: Additional generated SMILES–3D pairs. Each pair shows the decoded SMILES and generated 3D molecule, with their Vina Dock, QED, SA, and Morgan–Tanimoto similarity (radius 2, 2048 bits).

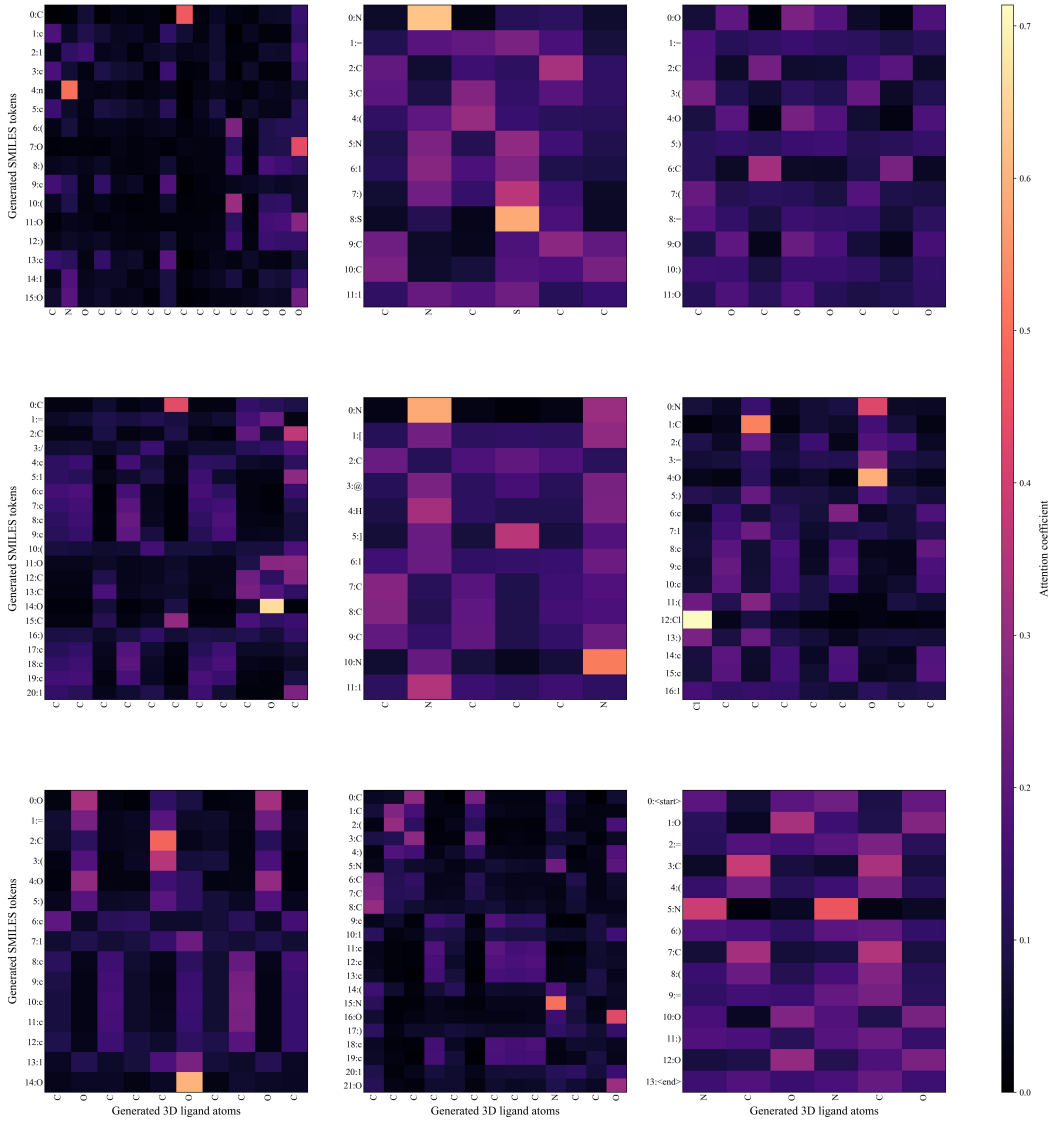



Figure 9: Cross-modal attention between generated SMILES tokens (rows) and 3D ligand atoms (columns) in nine examples. We average the attention weights over all eight heads. Brighter colors indicate higher attention weights, highlighting element-level correspondences across the two modalities.