Interaction-based Retrieval-augmented Diffusion Models for Protein-specific 3D Molecule Generation

Zhilin Huang *[1 2] Ling Yang *[3] Xiangxin Zhou [4] Chujun Qin [5] Yijie Yu [1 2]
Xiawu Zheng [2 6] Zikun Zhou [2] Wentao Zhang [3] Yu Wang [2] Wenming Yang [1 2]

Abstract

Generating ligand molecules that bind to specific protein targets via generative models holds substantial promise for advancing structure-based drug design. Existing methods generate molecules from scratch without reference or template ligands, which poses challenges in model optimization and may yield suboptimal outcomes. To address this problem, we propose an innovative interaction-based retrieval-augmented 3D molecular diffusion model named IRDIFF to facilitate target-aware molecule generation. IRDIFF leverages a curated set of ligand references, *i.e.*, those with desired properties such as high binding affinity, to steer the diffusion model towards synthesizing ligands that satisfy design criteria. Specifically, we design a geometric protein-molecule interaction network (*PMINet*), and pretrain it with binding affinity signals to: (i) retrieve target-aware ligand molecules with high binding affinity to serve as references, and (ii) incorporate essential protein-ligand binding structures for steering molecular diffusion generation with two effective augmentation mechanisms, *i.e.*, *retrieval augmentation* and *self augmentation*. Empirical studies on CrossDocked2020 dataset show IRDIFF can generate molecules with more realistic 3D structures and achieve state-of-the-art binding affinities towards the protein targets, while maintaining proper molecular properties. The codes and models are available at https://github.com/YangLing0818/IRDiff.

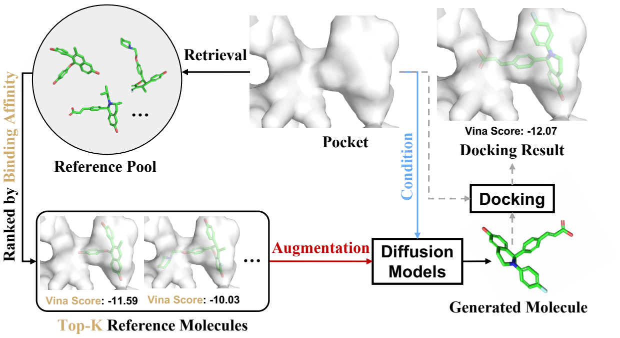

Figure 1. Interaction-based Retrieval-augmentation SBDD with the ligands with high binding affinities towards target protein.

1. Introduction

Designing ligand molecules that can bind to specific protein targets and modulate their function, also known as *structure-based drug design* (SBDD) (Anderson, 2003; Batool et al., 2019), is a fundamental problem in drug discovery and can lead to significant therapeutic benefits. SBDD requires models to synthesize drug-like molecules with stable 3D structures and high binding affinities to the target. Nevertheless, it is challenging and involves massive computational efforts because of the enormous space of synthetically feasible chemicals (Ragoza et al., 2022a) and freedom degree of both compound and protein structures (Hawkins, 2017).

Recently, several new generative methods have been proposed for the SBDD task (Li et al., 2021; Luo et al., 2021; Peng et al., 2022; Powers et al., 2022; Ragoza et al., 2022b; Zhang et al., 2023), which learn to generate ligand molecules by modeling the complex spatial and chemical interaction features of the binding site. For instance, some methods adopt autoregressive models (ARMs) (Luo & Ji, 2021; Liu et al., 2022; Peng et al., 2022) and show promising results in SBDD tasks, which generate 3D molecules by iteratively adding atoms or bonds based on the target binding site. However, ARMs tend to suffer from error accumulation, and it is difficult to find an optimal generation order, which are both nontrivial for 3D molecular graphs. Aiming to address these limitations of ARMs, recent works (Guan et al., 2023a; Schneuing et al., 2022; Lin et al., 2022) adopt

*Equal contribution [1]Shenzhen International Graduate School, Tsinghua University [2]Peng Cheng Laboratory [3]Peking University [4]School of Artificial Intelligence, University of Chinese Academy of Sciences [5]China Southern Power Grid [6]Xiamen University. Correspondence to: Yu Wang <wangy20@pcl.ac.cn>, Wenming Yang <yangelwm@163.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

diffusion models (Ho et al., 2020) to model the distribution of atom types and positions from a standard Gaussian prior with a post-processing to assign bonds. These diffusion-based SBDD methods develop SE(3)-equivariant diffusion models (Hoogeboom et al., 2022) to capture both local and global spatial interactions between atoms and have achieved more promising performance compared with previous autoregressive models. Despite achieving state-of-the-art performance, it is still difficult for existing diffusion-based methods to generate molecules that satisfies biological metrics such as binding affinity. This difficulty mainly arises from the extensive search space of poses within and between molecules. Moreover, generating molecules from scratch makes the generation process more challenging to optimize and may lead to suboptimal performance.

To overcome these challenges, we propose a novel **I**nteraction-based **R**etrieval-augmented **Diff**usion model (**IRDiff**) for SBDD task as in Figure 1. IRDiff is inspired by the recent significant advancement in machine learning particularly in generative modeling, called retrieval-augmented generation or in-context learning (Liu et al., 2023; Gu et al., 2022; Rubin et al., 2022), which enables (large) models (Brown et al., 2020; OpenAI, 2023) to generalize well to previously-unseen tasks with proper task-specific references. Herein, unlike previous methods that solely depend on the generalization capacity of generative models for new target proteins, IRDiff explicitly utilizes a small set of target-aware molecular ligand references with high binding affinity to the specific protein. By leveraging the protein-molecule interaction information between references and the given protein to steer the diffusion model toward generating ligands, our IRDiff is capable of generating molecules that bind tightly to the target pocket. Specifically, we introduce a protein-molecule interaction network named ***PMINet*** and pre-train it with binding affinity signals, which is parameterized with SE(3)-equivariant and attention layers to capture interaction information between protein-molecule pairs. Then we utilize the pre-trained PMINet to (1) retrieve protein-aware ligand molecules with high binding affinity to serve as molecular references, and (2) incorporate essential protein-molecule binding structures for steering molecular diffusion generation with two effective augmentation mechanisms, *i.e.*, ***retrieval augmentation*** and ***self augmentation***, conditioned on both the molecular ligand reference set and target protein. Significantly, our IRDiff effectively leverages the protein-molecule interactions modeled by PMINet for 3D protein-specific molecule generation, even when PMINet solely focuses on modeling sequence-level interactions between proteins and molecular ligands. This capability enhances the potential for our method to be widely applicable and scalable. Extensive experiments on CrossDocked2020 dataset demonstrate the effectiveness of our IRDiff, achieving new state-of-the-art performance regarding binding-related metrics.

We highlight our main contributions as follows: **(i):** We propose an interaction-based retrieval-augmented 3D molecular diffusion model named IRDiff for SBDD tasks. This model guides 3D molecular generation using informative external target-aware references, bridging the binding affinity prediction task and its inverse problem. **(ii):** We design two novel augmentation mechanisms, *i.e., retrieval augmentation* and *self augmentation*, to incorporate essential protein-molecule binding structures for target-aware molecular generation. **(iii):** Our IRDiff can generate ligands that not only bind tightly to target pockets but also maintain proper molecular properties. Empirical results on the CrossDocked2020 dataset show that our model achieves **-6.03** Avg. Vina Score and **0.53** Avg. QED score, indicating a prominent trade-off between binding- and property-related metrics.

2. Related Work

Structure-Based Drug Design As the increasing availability of 3D-structure protein-ligand data (Kinnings et al., 2011), structure-based drug design (SBDD) becomes a hot research area and it aims to generate diverse molecules with high binding affinity to specific protein targets (Luo et al., 2021; Yang et al., 2022; Schneuing et al., 2022; Tan et al., 2022). Early attempts learn to generate SMILES strings or 2D molecular graphs given protein contexts (Skalic et al., 2019; Xu et al., 2021a). However, it is uncertain whether the resulting compounds with generated strings or graphs could really fit the geometric landscape of the 3D structural pockets. More works start to involve 3D structures of both proteins and molecules (Li et al., 2021; Ragoza et al., 2022b; Zhang et al., 2023; Zhang & Liu, 2023). Luo et al. (2021), Liu et al. (2022), and Peng et al. (2022) adopt autoregressive models to generate 3D molecules in an atom-wise manner. Recently, powerful diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) begin to play a role in SBDD, and have achieved promising generation results with non-autoregressive sampling (Lin et al., 2022; Schneuing et al., 2022; Guan et al., 2023a). TargetDiff (Guan et al., 2023a), DiffBP (Lin et al., 2022), and DiffSBDD (Schneuing et al., 2022) utilize E(n)-equivariant GNNs (Satorras et al., 2021) to parameterize conditional diffusion models for protein-aware 3D molecular generation. Despite progress, existing methods mainly generate molecules from scratch without informative template or reference ligands for unseen target proteins, which may lead to hard optimization and poor binding affinity. In this paper, IRDiff for the first time utilize the external ligands with high binding affinity to steer molecular diffusion generation.

Retrieval-Augmented Generation The concept of retrieval augmentation has been well studied for exploiting the generalization ability of generative models, including natural language processing (Liu et al., 2023; Siriwardhana et al., 2023; Rubin et al., 2022; Yang et al., 2023), and computer vision (Alayrac et al., 2022; Gan et al., 2022; Jia et al., 2022; Sheynin et al., 2022; Blattmann et al., 2022). Numerous works have explored techniques to adapt models to novel tasks using a few examples as references. Some methods (Chen et al., 2022; Rubin et al., 2022) take input as reference and retrieve similar examples for further augmentation. For example, EPR (Rubin et al., 2022) utilizes a dense retriever to retrieve training examples as references for sequence-to-sequence generation. Recently, in the field of drug discovery, RetMol (Wang et al., 2022) employ a retrieval-based framework to control 2D molecule generation for ligand-based drug design (LBDD) (Bacilieri & Moro, 2006). In contrast, our pioneering contribution is to discover proper protein-aware 3D molecule references for the purpose of addressing SBDD tasks, which is the first to incorporate complex cross-modal (protein-molecule) interactions in reference design, thereby introducing a new dimension to the field.

3. Preliminary

Notations The SBDD task from the perspective of generative models can be defined as generating ligand molecules which can bind to a given protein binding site. The target (protein) and ligand molecule can be represented as $\mathcal{P} = \{(\boldsymbol{x}_P^{(i)}, \boldsymbol{v}_P^{(i)})\}_{i=1}^{N_P}$ and $\mathcal{M} = \{(\boldsymbol{x}_M^{(i)}, \boldsymbol{v}_M^{(i)})\}_{i=1}^{N_M}$, respectively. Here N_P (resp. N_M) refers to the number of atoms of the protein $\mathcal{P}$ (resp. the ligand molecule $\mathcal{M}$). $\boldsymbol{x} \in \mathbb{R}^3$ and $\boldsymbol{v} \in \mathbb{R}^K$ denote the position and type of the atom respectively. In the sequel, matrices are denoted by uppercase boldface. For a matrix $\mathbf{X}$, $\mathbf{x}_i$ denotes the vector on its i-th row, and $\mathbf{X}_{1:N}$ denotes the submatrix comprising its 1-st to N-th rows. For brevity, the ligand molecule is denoted as $\mathbf{M} = [\mathbf{X}_M, \mathbf{V}_M]$ where $\mathbf{X}_M \in \mathbb{R}^{N_M \times 3}$ and $\mathbf{V}_M \in \mathbb{R}^{N_M \times K}$, and the protein is denoted as $\mathbf{P} = [\mathbf{X}_P, \mathbf{V}_P]$ where $\mathbf{X}_P \in \mathbb{R}^{N_P \times 3}$ and $\mathbf{V}_P \in \mathbb{R}^{N_P \times K}$. The task can be formulated as modeling the conditional distribution $p(\mathbf{M}|\mathbf{P})$.

DDPMs in SBDD Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) equipped with SE(3)-invariant prior and SE(3)-equivariant transition kernel have been applied on the SBDD task (Guan et al., 2023a; Schneuing et al., 2022; Lin et al., 2022). More specifically, types and positions of the ligand molecule are modeled by DDPM, while the number of atoms N_M is usually sampled from an empirical distribution (Hoogeboom et al., 2022; Guan et al., 2023a) or predicted by a neural network (Lin et al., 2022), and bonds are determined as post-processing.

Target-aware Molecular Diffusion In the forward diffusion process, a small Gaussian noise is gradually injected into data as a Markov chain. Because noises are only added on ligand molecules but not proteins in the diffusion process, we denote the atom positions and types of the ligand molecule at time step t as $\mathbf{X}_t$ and $\mathbf{V}_t$ and omit the subscript M without ambiguity. The diffusion transition kernel can be defined as follows:

$$q(\mathbf{M}_t|\mathbf{M}_{t-1}, \mathbf{P}) = \prod_{i=1}^{N_M} \mathcal{N}(\mathbf{x}_{i,t}; \sqrt{1-\beta_t}\mathbf{x}_{i,t-1}, \beta_t \boldsymbol{I}) \cdot \\ \mathcal{C}(\mathbf{v}_{i,t}|(1-\beta_t)\mathbf{v}_{i,t-1} + \beta_t/K), \tag{1}$$

where $\mathcal{N}$ and $\mathcal{C}$ stand for the Gaussian and categorical distribution respectively, β_t is defined by fixed variance schedules. The corresponding posterior can be analytically derived as follows:

$$q(\mathbf{M}_{t-1}|\mathbf{M}_t, \mathbf{M}_0, \mathbf{P}) = \prod_{i=1}^{N_M} \mathcal{N}(\mathbf{x}_{i,t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_{i,t}, \mathbf{x}_{i,0}), \tilde{\beta}_t \boldsymbol{I}) \cdot \\ \mathcal{C}(\mathbf{v}_{i,t-1}|\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t}, \mathbf{v}_{i,0})), \tag{2}$$

where $\tilde{\boldsymbol{\mu}}(\mathbf{x}_{i,t}, \mathbf{x}_{i,0}) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_{i,0} + \frac{\sqrt{\alpha}_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_{i,t}$, $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, $\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t}, \mathbf{v}_{i,0}) = \frac{\boldsymbol{c}^*}{\sum_{k=1}^K c_k^*}$, and $\boldsymbol{c}^*(\mathbf{v}_{i,t}, \mathbf{v}_{i,0}) = [\alpha_t \mathbf{v}_{i,t} + (1 - \alpha_t)/K] \odot [\bar{\alpha}_{t-1}\mathbf{v}_{i,0} + (1-\bar{\alpha}_{t-1})/K]$.

In the approximated reverse process, also known as the generative process, a neural network parameterized by θ learns to recover data by iteratively denoising. The reverse transition kernel can be approximated with predicted atom types $\hat{\mathbf{v}}_{i,0}$ and atom positions $\hat{\mathbf{x}}_{i,0}$ as follows:

$$p_\theta(\mathbf{M}_{t-1}|\mathbf{M}_t, \mathbf{P}) = \prod_{i=1}^{N_M} \mathcal{N}(\mathbf{x}_{i,t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_{i,t}, \hat{\mathbf{x}}_{i,0}), \tilde{\beta}_t \boldsymbol{I}) \cdot \\ \mathcal{C}(\mathbf{v}_{i,t-1}|\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t}, \hat{\mathbf{v}}_{i,0})). \tag{3}$$

4. Methods

We propose IRDiff, a novel interaction-based retrieval-augmented diffusion framework (as demonstrated in Figure 2) for target-aware 3D molecule generation. We first introduce the pre-trained binding-affinity models as PMINet which is fully modelling complex interaction information between proteins and ligands and can be used as a retriever to discover protein-aware ligand references in the molecular database (Section 4.2). Finally, we propose exemplar augmentation and self augmentation to utilize the ligand references with the pre-trained PMINet for facilitating protein-aware 3D molecular generation (Section 4.3).

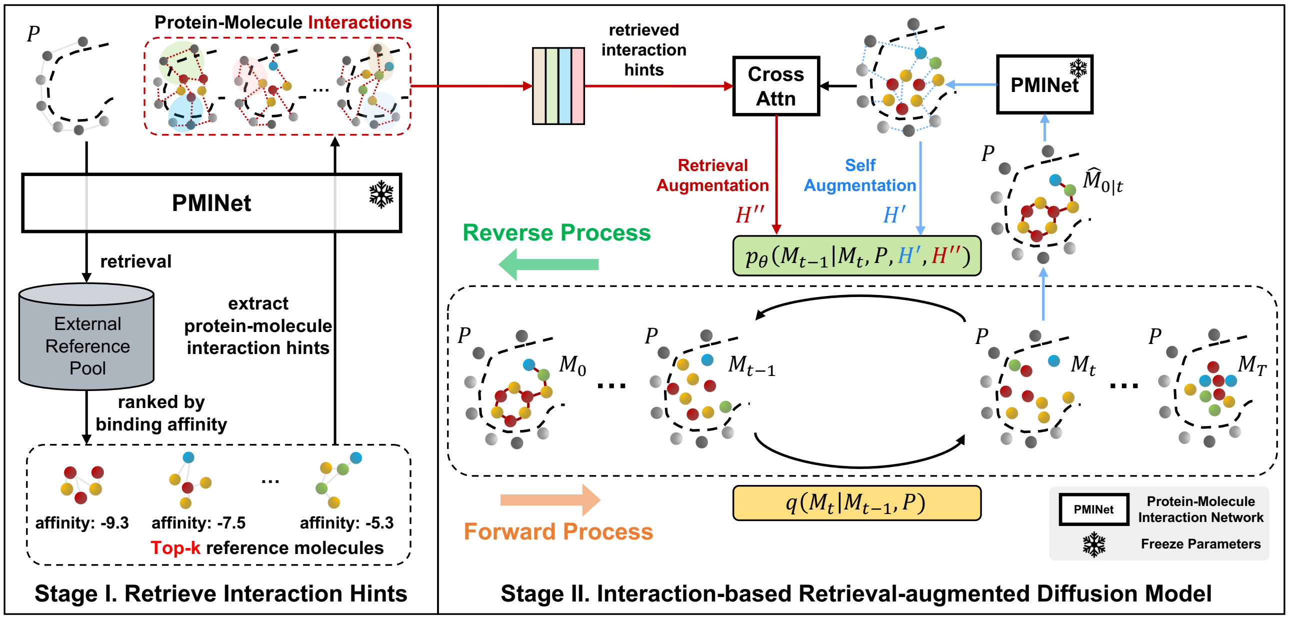
P
Protein-Molecule Interactions
PMINet
retrieval
External Reference Pool
extract protein-molecule interaction hints
ranked by binding affinity
affinity: -9.3
affinity: -7.5
affinity: -5.3
Top-k reference molecules
Stage I. Retrieve Interaction Hints
retrieved interaction hints
Cross Attn
PMINet
Retrieval Augmentation
Self Augmentation
H''
H'
$\widehat{M}_{0|t}$
Reverse Process
$p_\theta(M_{t-1}|M_t, P, H', H'')$
M_0
M_{t-1}
M_t
M_T
Forward Process
$q(M_t|M_{t-1}, P)$
PMINet Protein-Molecule Interaction Network
Freeze Parameters
Stage II. Interaction-based Retrieval-augmented Diffusion Model

Figure 2. The overall schematic diagram of IRDIFF. It first utilizes the pre-trained **PMINet** (in Section 4.1) to retrieve binding-aware molecule references (in Section 4.2), and then steers the target-aware 3D molecular diffusion generation with **retrieval and self augmentation** (in Section 4.3).

4.1. Modeling Protein-Ligand Interactions with PMINet

To capture the interactions between proteins and ligand molecules, we introduce a protein-ligand interaction network (namely PMINet) to model the binding affinity of protein-ligand pairs, consisting of SE(3)-equivariant neural networks (Satorras et al., 2021) and cross-attention layers (Borgeaud et al., 2022; Hou et al., 2019). Two shallow SE(3)-equivariant neural networks are applied on the fully-connected graphs of the protein $\mathcal{G}_P$ and ligand molecule $\mathcal{G}_M$ to model the inner-molecule interactions. Given a ligand graph $\mathcal{G}_M$, the l-th graph attention layer works as follows:

$$\mathbf{h}_{M,i}^{l+1} = \mathbf{h}_{M,i}^{l} + \sum_{j\in\mathcal{N}_M(i)} f_{M,h}^{l}\left(\left\|\Delta\mathbf{x}_{M,ij}^{l}\right\|, \mathbf{h}_{M,i}^{l}, \mathbf{h}_{M,j}^{l}\right), \quad (4)$$

$$\mathbf{x}_{M,i}^{l+1} = \mathbf{x}_{M,i}^{l} + \sum_{j\in\mathcal{N}_M(i)} \Delta\mathbf{x}_{M,ij}^{l} f_{M,x}^{l}\left(\left\|\Delta\mathbf{x}_{M,ij}^{l}\right\|, \mathbf{h}_{M,i}^{l+1}, \mathbf{h}_{M,j}^{l+1}\right) \quad (5)$$

where $\Delta\mathbf{x}_{M,ij}^{l} := \mathbf{x}_{M,i}^{l} - \mathbf{x}_{M,j}^{l}$, $\mathbf{h}_{M,i}^{l+1} \in \mathbb{R}^d$ and $\mathbf{x}_{M,i}^{l+1} \in \mathbb{R}^3$ are the SE(3)-invariant and SE(3)-equivariant hidden states of the atom i of the ligand after the l-th SE(3)-equivariant layer, respectively. $\mathcal{N}_M(i)$ stands for the set of neighbors of atom i on $\mathcal{G}_M$, and the initial hidden state $\mathbf{h}_{M,i}^{0}$ is obtained by an embedding layer that encodes atoms. Given a protein graph $\mathcal{G}_P$, $\mathbf{h}_{P,i}^{l}$ and $\mathbf{x}_{P,i}^{l}$ can be derived in the same way.

And then, an atom-wise cross-attention based interaction layer is proposed to learn the inter-molecule interactions between protein-ligand pairs, which essentially accounts for the binding affinity. Finally, the SE(3)-invariant features $\mathbf{H}_M^L \in \mathbb{R}^{N_M\times d}$ and $\mathbf{H}_P^L \in \mathbb{R}^{N_P\times d}$ (whose i-th rows are $\mathbf{h}_{M,i}^L$ and $\mathbf{h}_{P,i}^L$ respectively) are used as inputs to the cross-attention (Vaswani et al., 2017) layer for extracting binding-aware interactive representations:

$$\mathbf{Int}_M = \mathrm{softmax}((\mathbf{W}_Q\mathbf{H}_M^L)(\mathbf{W}_K\mathbf{H}_P^L)^T)\mathbf{W}_V\mathbf{H}_P^L, \quad (6)$$

$$\mathbf{Int}_P = \mathrm{softmax}((\mathbf{W}_Q\mathbf{H}_P^L)(\mathbf{W}_K\mathbf{H}_M^L)^T)\mathbf{W}_V\mathbf{H}_M^L, \quad (7)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices. The enhanced features of the protein and molecule (*i.e.*, $\mathbf{Int}_P$ and $\mathbf{Int}_M$) are further aggregated into a global features to predict the binding affinity: $S_{\mathrm{Aff}}(\mathcal{M}, \mathcal{P}) :=$ $\mathrm{PMINet}(\mathcal{M}, \mathcal{P})$. Please refer to Appendix B.1 for details.

4.2. Constructing Target-Aware Ligand References

Inspired by the recent success of retrieval-augmented generation (Yang et al., 2023), we hope to design an effective retrieval augmentation strategy for structure-based drug design to guide the generation of target-aware ligands with desired properties. Hence, we utilize the structure-based protein-molecule interaction prior learned by PMINet to identify the top candidates (*i.e.*, molecular ligands with high-binding affinity), which are most suitable for enhancing the subsequent target-aware molecule design.

More concretely, given a target $\mathcal{P}$ and an external database of molecular ligands $\mathcal{D} := \{\mathcal{M}_i\}_{i=1}^N$, we use the pre-trained $\mathrm{PMINet}(\cdot, \mathcal{P})$ introduced in Section 4.1 to scan the database and retrieve the molecular ligands with **top-k**

high predicted binding affinity to this target as references:

$$\mathcal{D}(\mathcal{P}, k) \coloneqq \mathrm{top}_k(\{\mathcal{M}_i\}_{i=1}^{N}, \mathrm{PMINet}(\cdot,\ \mathcal{P})). \tag{8}$$

Here we denote the reference pool as $\mathcal{D}(\mathcal{P}, k)$, the size of reference pool are denoted as N. The ligand molecules in the external database are real, so they are expected to provide some valid substructures as references and thus promote the validity of the generated ligand molecules. For example, the ligands in the reference pool can be viewed as probes to explore how the target interacts with ligands, which is supposed to offer useful clues for molecule generation. Due to their high binding affinity, they can potentially reveal the critical locations (*e.g.*, promising hydrogen donors or acceptors) to support strong inter-molecular forces. At time step t, we extend the reference pool to $\{\mathcal{M}_{t+1}^{\mathrm{pred}}\} \cup \mathcal{D}(\mathcal{P}, k)$, where $\mathcal{M}_{t+1}^{\mathrm{pred}}$ denotes predicted atom positions and types (*i.e.*, estimated $[\hat{\mathbf{X}}_0, \hat{\mathbf{V}}_0]$) at time step $t + 1$. This can be regarded as self augmentation which will be described next.

4.3. Interaction-based Retrieval-augmented 3D Equivariant Molecular Diffusion

In this subsection, we describe how to introduce the bingding-aware ligand references into the design of the neural network ϕ_θ which predicts (*i.e.*, reconstructs) $[\mathbf{X}_0, \mathbf{V}_0]$ in the reverse generation process (we highlight the critical parts of our augmentation mechanisms in violet):

$$[\hat{\mathbf{X}}_0, \hat{\mathbf{V}}_0] = \phi_\theta([\mathbf{X}_t, \mathbf{V}_t], t, \mathbf{P}, \{\mathcal{M}_{t+1}^{\mathrm{pred}}\} \cup \mathcal{D}(\mathcal{P}, k)). \tag{9}$$

Self Augmentation We first extract atom-wise embeddings $\mathbf{H}_M \in \mathbb{R}^{N_M \times d}$ and $\mathbf{H}_P \in \mathbb{R}^{N_P \times d}$ of the ligand molecule being generated and target protein, respectively. The molecule being generated, $\mathcal{M}_{t+1}^{\mathrm{pred}}$, itself is supposed to be a candidate ligand with high binding affinity to the target, especially when t is large (*i.e.*, the generative process nearly ends). To maximize the exploitation of protein-ligand interaction prior emerged in the reverse diffusion trajectories, we leverage the enhanced molecular atom embedding $\mathbf{Int}_{M_{t+1}^{\mathrm{pred}}}$ and protein atom embedding $\mathbf{Int}_P$ produced by the interaction layer of $\mathrm{PMINet}(\mathcal{M}_{t+1}^{\mathrm{pred}}, \mathcal{P})$ to self augment the generative process as follows:

$$\mathbf{H}'_M = \mathrm{MLP}\left([\mathbf{H}_M, \mathbf{Int}_{M_{t+1}^{\mathrm{pred}}}]\right), \tag{10}$$

$$\mathbf{H}'_P = \mathrm{MLP}\left([\mathbf{H}_P, \mathbf{Int}_P]\right). \tag{11}$$

In training, due to the inaccessibility of $\mathcal{M}_{t+1}^{\mathrm{pred}}$, we directly use ground truth molecule $\mathcal{M}$ to substitute it in a teacher-forcing fashion. We illustrate more insights about self augmentation in Appendix A.

Retrieval Augmentation We further propose retrieval augmentation to leverage the reference ligands for steering the reverse generation process. The pre-trained PMINet is reused here to extract interactive structural context information between the target protein $\mathcal{P}$ and ligands in reference pool $\mathcal{D}(\mathcal{P}, k)$ to enhance the protein representations:

$$\mathbf{H}''_P = \mathrm{Pool}(\{\mathrm{MLP}([\mathbf{H}'_P, \mathbf{Int}_P^i])\}_{i=1}^{k}) \tag{12}$$

where k is the number of candidate ligands with top-k highest binding affinities, $\mathbf{Int}_P^i$ is the binding-aware protein feature produced by the interaction layer of $\mathrm{PMINet}(\mathcal{M}_i, \mathcal{P})$ (as Equation (7)) and $\mathcal{M}_i$ is the i-th exemplar ligand in the reference pool $\mathcal{D}(\mathcal{M}, k)$. Besides, in order to augment the molecular diffusion generation with possible binding structures in exemplar ligands, we merge the enhanced embeddings of exemplar ligands and generated molecules with a trainable cross attention mechanism:

$$\mathbf{H}''_M = \mathrm{Pool}(\{\mathrm{softmax}((\mathbf{W}_Q \mathbf{H}'_M)(\mathbf{W}_K \mathbf{Int}_{M_i})^T)\mathbf{W}_V \mathbf{Int}_{M_i}\}_{i=1}^{k}) \tag{13}$$

where $\mathbf{Int}_{M_i}$ is the binding-aware exemplar ligand feature produced by the interaction layer of $\mathrm{PMINet}(\mathcal{M}_i, \mathcal{P})$ (as Equation (6)). Our IRDiff uses retrieval and self augmentation to sufficiently leverage both the informative binding prior in external ligands and the protein-aware interaction context for 3D equivariant molecular diffusion generation.

3D Equivariant Molecular Diffusion We then apply an SE(3)-equivariant neural network on the k-nn graph of the protein-ligand complex (denoted as $\mathbf{C} = [\![\mathbf{M}, \mathbf{P}]\!]$, where $[\![\cdot]\!]$ denotes concatenation along the first dimension) to learn the atom-wise protein-molecule interactions in generative process. The SE(3)-invariant hidden state $\mathbf{H}_C$ and SE(3)-equivariant positions $\mathbf{X}_C$ are updated as follows:

$$\mathbf{h}_{C,i}^{l+1} = \sum_{j \in \mathcal{N}_C(i)} f_{C,h}^{l}\left(\left\|\Delta \mathbf{x}_{C,ij}^{l}\right\|, \mathbf{h}_{C,i}^{l}, \mathbf{h}_{C,j}^{l}, \mathbf{e}_{C,ij}\right) + \mathbf{h}_{C,i}^{l}, \tag{14}$$

$$\begin{aligned}\mathbf{x}_{C,i}^{l+1} = &\sum_{j \in \mathcal{N}_C(i)} \Delta \mathbf{x}_{C,ij}^{l} \cdot f_{C,x}^{l}\left(\left\|\Delta \mathbf{x}_{C,ij}^{l}\right\|, \mathbf{h}_{C,i}^{l+1}, \mathbf{h}_{C,j}^{l+1}, \mathbf{e}_{C,ij}\right) \cdot \mathbb{1}_{\mathrm{mol}} \\ &+ \mathbf{x}_{C,i}^{l}\end{aligned} \tag{15}$$

where $\Delta \mathbf{x}_{C,ij}^{l} \coloneqq \mathbf{x}_{C,i}^{l} - \mathbf{x}_{C,j}^{l}$, $\mathcal{N}_C(i)$ stands for the set of k-nearest neighbors of atom i on the protein-ligand complex graph, $\mathbf{e}_{C,ij}$ indicates the atom i and atom j are both protein atoms or both ligand atoms or one protein atom and one ligand atom, and $\mathbb{1}_{\mathrm{mol}}$ is the ligand atom mask since the protein atom coordinates are known and thus supposed to remain unchanged during this update. We let $\mathbf{H}_C^0 \coloneqq [\![\mathbf{H}''_M, \mathbf{H}''_P]\!]$ to incorporate the information contained in the reference pool $\{\mathcal{M}_{t+1}^{*}\} \cup \mathcal{D}(\mathcal{P}, k)$. Finally, we use $\hat{\mathbf{V}}_0 = \mathrm{softmax}(\mathrm{MLP}(\mathbf{H}_{C,1:N_M}^{L}))$ and $\hat{\mathbf{X}}_0 = \mathbf{X}_{C,1:N_M}^{L}$ as the final prediction. Shifting the Center of Mass (CoM) of protein atoms to zero (Xu et al., 2021b; Hoogeboom et al., 2022; Guan et al., 2023a) and the design of EGNN (Satorras et al., 2021) ensure SE(3)-equivariance of the reverse transition kernel $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{X}_P)$. Augmenting the generative process only augments the SE(3)-invariant hidden states without breaking SE(3)-equivariance.

Training and Sampling To train IRDIFF (*i.e.*, optimize the evidence lower bound induced by IRDIFF), we use the same objective function as Guan et al. (2023a). The atom position loss and atom type loss at time step $t-1$ are defined as follows respectively:

$$\begin{aligned} L^{(x)}_{t-1} &= \frac{1}{2\tilde{\beta}_t^2}\sum_{i=1}^{N_M}\|\tilde{\boldsymbol{\mu}}(\mathbf{x}_{i,t},\mathbf{x}_{i,0}) - \tilde{\boldsymbol{\mu}}(\mathbf{x}_{i,t},\hat{\mathbf{x}}_{i,0})\|^2 \\ &= \gamma_t \sum_{i=1}^{N_M}\|\mathbf{x}_{i,0}-\hat{\mathbf{x}}_{i,0}\|; \end{aligned} \tag{16}$$

$$L^{(v)}_{t-1} = \sum_{i=1}^{N_M}\sum_{k=1}^{K}\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t},\mathbf{v}_{i,0})_k \log\frac{\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t},\mathbf{v}_{i,0})_k}{\tilde{\boldsymbol{c}}(\mathbf{v}_{i,t},\hat{\mathbf{v}}_{i,0})_k}; \tag{17}$$

where $\hat{\mathbf{X}}_0$ and $\hat{\mathbf{V}}_0$ are predicted from $\mathbf{X}_t$ and $\mathbf{V}_t$, and $\gamma_t = \frac{\bar{\alpha}_{t-1}\beta_t^2}{2\tilde{\beta}_t^2(1-\bar{\alpha}_t)^2}$. Kindly recall that $\mathbf{x}_{i,t}$, $\mathbf{v}_{i,t}$, $\hat{\mathbf{x}}_{i,0}$, and $\hat{\mathbf{v}}_{i,0}$ correspond to the i-th row of $\mathbf{X}_t$, $\mathbf{V}_t$, $\hat{\mathbf{X}}_0$, and $\hat{\mathbf{V}}_0$, respectively. The final loss combines the above two losses with a hyperparameter λ as: $L = L^{(x)}_{t-1} + \lambda L^{(v)}_{t-1}$. We summarize the training procedure of IRDIFF in Algorithm 1 and highlight the differences from its counterpart, TargetDiff (Guan et al., 2023a), in violet.

Algorithm 1 Training Procedure of IRDIFF

```
Input: Protein-ligand binding dataset $\{\mathcal{P},\mathcal{M}\}_{i=1}^N$, neural network $\phi_\theta$, external database $\mathcal{D}$, pre-trained PMINet, number of exemplar ligands in each retrieval pool $k$
Screen $\{\mathcal{P},\mathcal{M}\}_{i=1}^N$ and retrieve ligands with top-$k$ high binding affinity from $\mathcal{D}$ using PMINet to obtain $\{\mathcal{P},\mathcal{M},\mathcal{D}(\mathcal{P},k)\}_{i=1}^N$ as described in Section 4.2
while $\phi_\theta$ not converge do
    Sample diffusion time $t \in \mathcal{U}(0,\dots,T)$
    Move the complex to make CoM of protein atoms zero
    Perturb $[\mathbf{X}_0,\mathbf{V}_0]$ to obtain $[\mathbf{X}_t,\mathbf{V}_t]$
    Embed $\mathbf{V}_t$ into $\mathbf{H}_M^0$, and embed $\mathbf{V}_P$ into $\mathbf{H}_P^0$
    Obtain features $\mathbf{H}'_M$ and $\mathbf{H}'_P$ with self augmentation based on $[\mathbf{X}_0,\mathbf{V}_0]$    (Equation (10))
    Obtain enhanced protein atom feature $\mathbf{H}''_P$ augmented by $\mathcal{D}(\mathcal{P},k)$    (Equation (12))
    Obtain enhanced ligand atom feature $\mathbf{H}''_M$ augmented by $\mathcal{D}(\mathcal{P},k)$    (Equation (13))
    Predict $[\hat{\mathbf{X}}_0,\hat{\mathbf{V}}_0]$ from $[\mathbf{X}_t,\mathbf{H}''_M]$ and $[\mathbf{X}_P,\mathbf{H}''_P]$ (Equations (14) and (15))
    Compute loss $L$ with $[\hat{\mathbf{X}}_0,\hat{\mathbf{V}}_0]$ and $[\mathbf{X}_M,\mathbf{V}_M]$ (Equations (16) and (17))
    Update $\theta$ by minimizing $L$
end while
```

Given a protein $\mathcal{P}$, the molecules can be sampled as Algorithm 2. The differences from previous molecular diffusion models, are highlighted in violet.

Algorithm 2 Sampling Procedure of IRDIFF

```
Input: The protein binding site $\mathcal{P}$, the learned model $\phi_\theta$, external databse $\mathcal{D}$, pre-trained PMINet, the number of exemplar ligands in each reference pool $k$
Output: Generated ligand molecule $\mathcal{M}$ that binds to the protein pocket $\mathcal{P}$
Sample the number of atoms $N_M$ of the ligand molecule $\mathcal{M}$ as described in Section 3
Move CoM of protein atoms to zero
Sample initial ligand atom coordinates $\mathbf{x}_T$ and atom types $\mathbf{v}_T$
Let $\mathbf{M}^* := [\mathbf{0},\mathbf{0}]$
Embed $\mathbf{V}_P$ into $\mathbf{H}_P$
for $t$ in $T, T-1, \dots, 1$ do
    Embed $\mathbf{V}_t$ into $\mathbf{H}_M$
    Obtain $\mathbf{H}'_M$, $\mathbf{H}'_P$ with self augmentation    (Eq. 10)
    Obtain $\mathbf{H}''_M$, $\mathbf{H}''_P$ augmented by $\mathcal{D}(\mathcal{P},k)$    (Eq. 12)
    Predict $[\hat{\mathbf{X}}_0,\hat{\mathbf{V}}_0]$ from $[\mathbf{X}_t,\mathbf{H}''_M]$ and $[\mathbf{X}_P,\mathbf{H}''_P]$    (Eqs. 14 and 15)
    Sample $\mathbf{X}_{t-1}$, $\mathbf{V}_{t-1}$ from the posterior $p_\theta$    (Eq. 3)
    Let $\mathbf{M}^* := [\hat{\mathbf{X}}_0,\hat{\mathbf{V}}_0]$
end for
```

5. Experiments

Datasets and Baseline Methods To pretrain PMINet with binding affinity signals, we use the PDBbind v2016 dataset (Liu et al., 2015), which is most frequently used in binding-affinity prediction tasks. Specifically, 3767 complexes are selected as training set, and the other 290 complexes are selected as testing set. As for molecular generation, following the previous work (Luo et al., 2021; Peng et al., 2022; Guan et al., 2023a), we train and evaluate IRDIFF on the CrossDocked2020 dataset (Francoeur et al., 2020). We follow the same data preparation and splitting as (Luo et al., 2021), where the 22.5 million docked binding complexes are refined to high-quality docking poses (RMSD between the docked pose and the ground truth < 1Å) and diverse proteins (sequence identity $< 30\%$). This produces $100,000$ protein-ligand pairs for training and 100 proteins for testing.

We randomly choose 128 ligands from training set for retrieval, and select the ligand of top-1 predicted binding affinity as reference for each protein. We compare our model with recent representative methods for SBDD. **LiGAN** (Ragoza et al., 2022a) is a conditional VAE model trained on an atomic density grid representation of protein-ligand structures. **AR** (Luo et al., 2021) and **Pocket2Mol** (Peng et al., 2022) are autoregressive schemes that generate 3D molecules atoms conditioned on the protein pocket and previous generated atoms. **TargetDiff** (Guan et al., 2023a) and **DecomposeDiff** (Guan et al., 2023b) are recent state-of-the-art non-autoregressive diffusion-based SBDD models.

Evaluation We comprehensively evaluate the generated molecules from three perspectives: **molecular structures**, **target binding affinity** and **molecular properties**. In terms of **molecular structures**, we calculate the Jensen-Shannon divergences (JSD) in empirical distributions of atom/bond distances between generated molecules and ground-truth ones provided in the test set. To estimate the **target binding affinity**, following previous work (Luo et al., 2021; Ragoza et al., 2022b; Guan et al., 2023a), we adopt AutoDock Vina (Eberhardt et al., 2021) to compute and report the mean and median of binding-related metrics, including *Vina Score*, *Vina Min*, *Vina Dock* and *High Affinity*. Vina Score directly estimates the binding affinity based on generated 3D molecules; Vina Min performs a local structure minimization before estimation; Vina Dock involves an additional re-docking process and reflects the best possible binding affinity; High affinity measures the ratio of how many generated molecules binds better than the ground-truth molecule per test protein. To evaluate **molecular properties**, we utilize the *QED*, *SA*, *Diversity* as metrics following (Luo et al., 2021; Ragoza et al., 2022a). QED is a simple quantitative estimation of drug-likeness combining several desirable molecular properties; SA is a measurement of the difficulty of synthesizing ligands; Diversity is computed as average pairwise dissimilarity between all generated ligands.

5.1. Main Results

Generated Molecular Structures We compare our IRDiff and the representative methods in terms of molecular structures. We plot the all-atom pairwise distance distribution of the generated molecules in Figure 3. IRDiff achieves JSD 0.08 to ground-truth molecules in the all-atom pairwise distance distribution of the generated molecules, which is better than two strong baseline methods Pocket2Mol and TargetDiff, indicating it effectively captures real atomic distances. We compute different bond distributions of the generated molecules and compare them against the corresponding ground-truth empirical distributions in Table 1. Our model has a comparable performance with DecompDiff and substantially outperforms other baselines across most major bond types, indicating the great potential of IRDiff for generating stable molecular structures. We attribute this to our augmentation mechanisms which directly provides realistic 3D ligand templates for steering molecule generation.

Target Binding Affinity and Molecule Properties We evaluate the effectiveness of IRDiff in terms of binding affinity. We can see in Table 2 that our IRDiff outperforms baselines in binding-related metrics. Specifically, IRDiff surpasses strong autoregressive method Pocket2Mol by a large margin of **17.3%** and **46.6%** in Avg. and Med. Vina Score, and surpasses strong diffusion-based method DecompDiff by **6.3%** and **14.1%** in Avg. and Med. Vina Score. In terms of high-affinity binder, we find that on average **67.4%** of the IRDiff molecules show better binding affinity than the ground-truth molecule in the test set, which is significantly better than other baselines. These gains demonstrate that the IRDiff effectively utilizes the external protein-ligand interactions to enable generating molecules with higher target binding affinity.

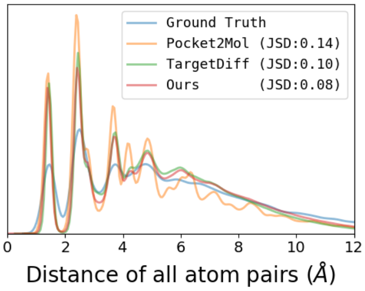

Figure 3. Comparing the distribution for distances of all-atom for ground-truth molecules in the test set (blue) and model generated molecules (color). Jensen-Shannon divergence (JSD) between two distributions is reported.

Table 1. Jensen-Shannon divergence between bond distance distributions of the ground-truth molecules and the generated molecules, and lower values indicate better performances. “-”, “=”, and “:” represent single, double, and aromatic bonds, respectively.

Bond	liGAN	AR	Pocket2 Mol	Target Diff	Decomp Diff	ours
C−C	0.601	0.609	0.496	0.369	**0.359**	0.439
C=C	0.665	0.620	0.561	0.505	0.537	**0.272**
C−N	0.634	0.474	0.416	0.363	0.344	**0.302**
C=N	0.749	0.635	0.629	0.550	0.584	**0.255**
C−O	0.656	0.492	0.454	0.421	0.376	**0.371**
C=O	0.661	0.558	0.516	0.461	0.374	**0.361**
C:C	0.497	0.451	0.416	0.263	0.251	**0.214**
C:N	0.638	0.552	0.487	0.235	0.269	**0.209**

Ideally, using whole training set for retrieval can significantly improve the both binding- and property-related metrics as demonstrated in Table 4, but it would increase computational burden. Thus we randomly choose 128 molecules for retrieval in all experiments. Moreover, we can see a trade-off between property-related metrics (QED and SA) and binding-related metrics in previous methods. TargetDiff and DecompDiff perform better than AR and Pocket2Mol in binding-related metrics, but fall behind them in QED and SA scores. In contrast, our IRDiff not only achieves the state-of-the-art binding-related scores but also maintains proper QED score, achieving a better trade-off than TargetDiff and DecompDiff. Nevertheless, we put less emphasis on QED and SA because they are often applied as rough

Table 2. Summary of different properties of ground-truth molecules in the test set and molecules generated by our model and other non-diffusion and diffusion-based baselines. n and k denote the size of reference pool and the number of utilized references, respectively. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively.

Methods		Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)	
		Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Ground Truth		-6.36	-6.46	-6.71	-6.49	-7.45	-7.26	-	-	0.48	0.47	0.73	0.74	-	-
Compare with Non-Diffusion	LiGAN	-	-	-	-	-6.33	-6.20	21.1%	11.1%	0.39	0.39	0.59	0.57	0.66	0.67
	GraphBP	-	-	-	-	-4.80	-4.70	14.2%	6.7%	0.43	0.45	0.49	0.48	**0.79**	**0.78**
	AR	-5.75	-5.64	-6.18	-5.88	-6.75	-6.62	37.9%	31.0%	0.51	0.50	0.63	0.63	0.70	0.70
	Pocket2Mol	-5.14	-4.70	-6.42	-5.82	-7.15	-6.79	48.4%	51.0%	**0.56**	**0.57**	**0.74**	**0.75**	0.69	0.71
	IRDIFF (n=128, k=1)	-5.86	-6.51	-7.14	-7.27	-8.33	**-8.49**	66.8%	**73.9%**	0.53	0.54	0.58	0.58	0.72	0.72
	IRDIFF (n=128, k=3)	**-6.03**	**-6.89**	**-7.27**	**-7.37**	**-8.42**	-8.42	**67.4%**	72.7%	0.53	0.54	0.59	0.58	0.72	0.72
Compare with Diffusion	TargetDiff	-5.47	-6.30	-6.64	-6.83	-7.80	-7.91	58.1%	59.1%	0.48	0.48	0.58	0.58	0.72	0.71
	DecompDiff	-5.67	-6.04	-7.04	-7.09	-8.39	-8.43	64.4%	71.0%	0.45	0.43	**0.61**	**0.60**	0.68	0.68
	IRDIFF (n=128, k=1)	-5.86	-6.51	-7.14	-7.27	-8.33	**-8.49**	66.8%	**73.9%**	**0.53**	**0.54**	0.58	0.58	**0.74**	**0.72**
	IRDIFF (n=128, k=3)	**-6.03**	**-6.89**	**-7.27**	**-7.37**	**-8.42**	-8.42	**67.4%**	72.7%	**0.53**	**0.54**	0.59	0.58	0.72	**0.72**

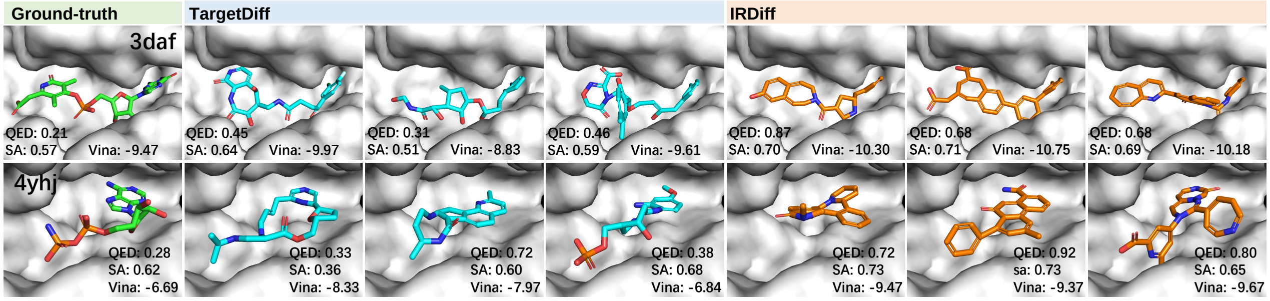

Figure 4. Ground-truth ligands and generated ligand molecules of TargetDiff (Guan et al., 2023a) and IRDIFF for 3daf (top row) and 4yhj (bottom row). We report QED, SA, and Vina Dock score for each molecule.

screening metrics in real drug discovery scenarios, and it would be fine as long as they are within a reasonable range. Figure 4 shows some examples of generated molecules and their properties. The molecules generated by our model have valid structures and reasonable binding poses to the target, which are supposed to be promising candidate ligands. More visualization are provided in Appendix D.

5.2. Model Analysis

Influence of Self and Retrieval Augmentation We investigate the impact of self augmentation and retrieval augmentation of IRDIFF. We showcase the efficacy of self augmentation and retrieval augmentation in our IRDIFF, and put results in Table 3. In particular, we remove our augmentation mechanisms from IRDIFF and use it as baseline.

We observe that simply applying augmentation mechanism without our pre-trained PMINet even hurt the generation performance, because it does not include informative protein-ligand interactions for self refinement. In contrast, our self augmentation significantly improve both binding- and property-related metrics due to the informative interaction knowledge brought by our pre-trained PMINet. Besides, our retrieval augmentation also has a notable improvement over baseline, revealing that the external target-aware references indeed provide a suitable reference and facilitate the optimization for molecular generation. Retrieval augmentation does not help much in property-related metrics, because we only focus on useful binding structures in exemplars.

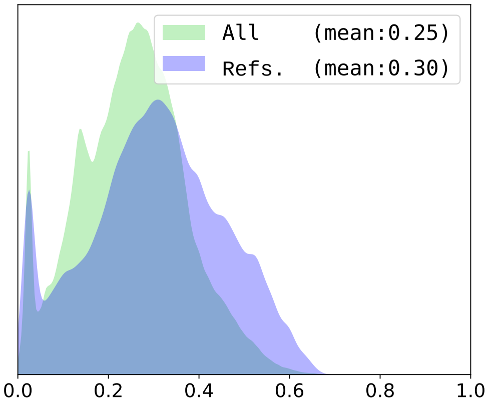

Figure 5. The distributions of Tanimoto similarity between generated ligands and (a) all ligands in database, and (b) corresponding references.

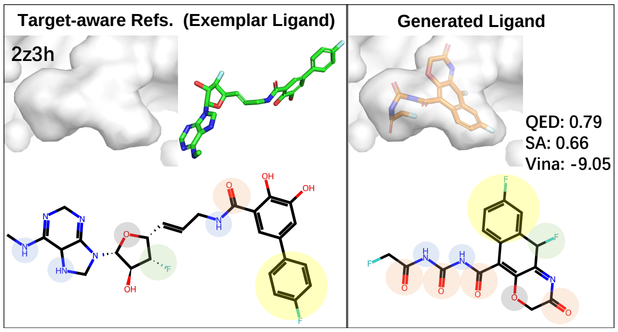

Figure 6. Example of a generated ligand molecule and its corresponding reference. Important substructures shared by these two molecules are highlighted in the same colors.

Table 3. Ablation studies of self augmentation and retrieval augmentation in IRDIFF (n=128, $k = 1$). Please refer to the Table 9 for the complete results.

Methods	Vina Score (↓)		Vina Min (↓)		High Affinity (↑)		QED (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-5.04	-5.75	-6.38	-6.52	54.2%	54.1%	0.46	0.46
augmentation without PMINet	-4.14	-5.64	-6.11	-6.36	54.9%	57.5%	0.47	0.48
+ self augmentation	-4.91	-6.03	-6.53	-6.82	61.5%	64.4%	**0.52**	**0.53**
+ retrieval augmentation	-5.39	-6.28	-6.40	-6.67	64.0%	71.5%	0.51	0.52
+ both augmentations	**-5.86**	**-6.51**	**-7.14**	**-7.27**	**66.8%**	**73.9%**	**0.53**	**0.54**

Furthermore, using both retrieval and self augmentation achieves the best binding-related metrics, demonstrating the effectiveness of the two complementary augmentation mechanisms in IRDIFF. Figures 5 and 6 provide both quantitative and qualitative analyses of the similarity between the generated ligands and their corresponding references, indicating that the references indeed serve as exemplars and guide the generation. More specifically, Figure 6 shows the model may automatically select critical substructures for high binding affinity to the target from reference ligands and reassemble them in proper positions to generate molecules.

Effect of Molecule Retrieval Database We investigate the influence of the 3D molecule database for retrieval in IRDIFF through ablation study on two variables n and k, where n denotes the size of the molecule database and k denotes the number of reference exemplars. From the results in Table 4, we observe that larger n can benefit IRDIFF in terms of binding-related metrics, because higher diversity allows for more binding-related cues (substructures) that can augment the generation process. Simply increasing k does not have an obvious improvement because leveraging more molecule references would also introduce more noises into generation process. Kindly note that the retrieval database is fixed during both training and testing, and we further evaluate the robustness of our model to the choices of retrieval database in Appendix C.

Table 4. The effect of hyper-parameter n and k. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively. Please refer to the 11 for the complete results.

Methods	Vina Score (↓)		Vina Min (↓)		High Affinity (↑)		QED (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
$n = 32$	-5.67	-6.21	-7.01	-7.13	65.2%	72.0%	0.51	0.52
$n = 64$	-5.74	-6.39	-7.07	-7.19	65.9%	72.8%	0.52	0.53
$n = 128$	**-5.86**	**-6.51**	**-7.14**	**-7.27**	**66.8%**	**73.9%**	**0.53**	**0.54**
$k = 1$	-5.86	-6.51	-7.14	-7.27	66.8%	73.9%	**0.53**	**0.54**
$k = 2$	-5.88	-6.62	-7.25	-7.29	66.6%	**74.4%**	**0.53**	**0.54**
$k = 3$	**-6.03**	**-6.89**	**-7.27**	**-7.37**	**67.4%**	72.7%	**0.53**	**0.54**

Effectiveness of Interaction-based Retrieval We investigate the impact of the reference ligands' binding affinity on the generation performance of IRDIFF. In our experiments, we choose ligands with the highest (i.e., k=1) target binding affinity as references, and we set the size n of reference pool to 128. Here, the reference ligand is replaced by the one with the lowest binding affinity based on the ranking list provided by PMINet. The experiments are conducted on the IRDIFF **without** utilizing self augmentation mechanism. As indicated in Table 5, utilizing low-affinity references (with an average predicted binding affinity of 3.95 by PMINet between the reference molecule and the corresponding protein pocket) in the molecule database results in poorer performance on binding-related metrics compared to using high-affinity references (with an average binding affinity of 7.24). This demonstrates the effectiveness of our designs for interaction-based retrieval and reference utilization in IRDIFF.

Table 5. The impact of molecule database with different properties. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively.

Methods	Avg. Affinity of Ref. (↑)	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		QED (↑)	
		Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-	-5.04	-5.75	-6.38	-6.52	-7.55	-7.72	0.46	0.46
low-affinity ref.	3.95	-4.12	-5.58	-6.09	-6.52	-7.84	-7.91	0.48	0.47
high-affinity ref.	**7.24**	**-5.39**	-6.28	**-6.40**	**-6.67**	**-8.14**	**-8.37**	**0.51**	**0.52**

6. Conclusion

In this work, we for the first time propose a interaction-based retrieval-augmented 3D molecular diffusion model IRDIFF for SBDD. We leverage the target-aware reference ligands to enhance the 3D molecular diffusion generation with effective self augmentation and retrieval augmentation mechanisms, significantly improving the binding affinity measured by Vina while maintaining proper molecular properties. For future work, we will incorporate other binding-related information into the generation process.

Acknowledgments

The work was partly supported by the National Natural Science Foundation of China (No.62171251), the Major Key Project of GZL under Grant SRPG22-001, the Major Key Research Project of PCL under Grant PCL2023A08 and PCL2023A09.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anderson, A. C. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.

Bacilieri, M. and Moro, S. Ligand-based drug design methodologies in drug discovery process: an overview. *Current drug discovery technologies*, 3(3):155–165, 2006.

Batool, M., Ahmad, B., and Choi, S. A structure-based drug discovery paradigm. *International journal of molecular sciences*, 20(11):2783, 2019.

Blattmann, A., Rombach, R., Oktay, K., Müller, J., and Ommer, B. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35: 15309–15324, 2022.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, X., Li, L., Zhang, N., Liang, X., Deng, S., Tan, C., Huang, F., Si, L., and Chen, H. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. In *Advances in Neural Information Processing Systems*, 2022.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. 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.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. 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.

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J., et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

Gu, Y., Han, X., Liu, Z., and Huang, M. Ppt: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8410–8423, 2022.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023a.

Guan, J., Zhou, X., Yang, Y., Bao, Y., Peng, J., Ma, J., Liu, Q., Wang, L., and Gu, Q. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (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 2023b. URL https://proceedings.mlr.press/v202/guan23a.html.

Hawkins, P. C. Conformation generation: the state of the art. *Journal of Chemical Information and Modeling*, 57 (8):1747–1756, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Hou, R., Chang, H., Ma, B., Shan, S., and Chen, X. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual prompt tuning. In

Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII, pp. 709–727. Springer, 2022.

Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., and Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling*, 51(2):408–419, 2011.

Li, Y., Pei, J., and Lai, L. Structure-based de novo drug design using 3d deep generative models. *Chemical science*, 12(41):13664–13675, 2021.

Lin, H., Huang, Y., Liu, M., Li, X., Ji, S., and Li, S. Z. Diffbp: Generative diffusion of 3d molecules for target protein binding. *arXiv preprint arXiv:2211.11214*, 2022.

Liu, M., Luo, Y., Uchino, K., Maruhashi, K., and Ji, S. Generating 3d molecules for target protein binding. In *International Conference on Machine Learning*, 2022.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Luo, Y. and Ji, S. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations*, 2021.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

OpenAI. Gpt-4 technical report. 2023. URL https://openai.com/research/gpt-4.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. *arXiv preprint arXiv:2205.07249*, 2022.

Powers, A. S., Yu, H. H., Suriana, P., and Dror, R. O. Fragment-based ligand generation guided by geometric deep learning on protein-ligand structure. *bioRxiv*, pp. 2022–03, 2022.

Ragoza, M., Masuda, T., and Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem Sci*, 13:2701–2713, Feb 2022a. doi: 10.1039/D1SC05976A.

Ragoza, M., Masuda, T., and Koes, D. R. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022b.

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schneuing, A., Du, Y., Harris, C., Jamasb, A., Igashov, I., Du, W., Blundell, T., Lió, P., Gomes, C., Welling, M., et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., and Nanayakkara, S. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.

Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. Shape-based generative modeling for de novo drug design. *Journal of chemical information and modeling*, 59(3):1205–1214, 2019.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Tan, C., Gao, Z., and Li, S. Z. Target-aware molecular graph generation. *arXiv preprint arXiv:2202.04829*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. Graph attention networks. *stat*, 1050 (20):10–48550, 2017.

Wang, Z., Nie, W., Qiao, Z., Xiao, C., Baraniuk, R., and Anandkumar, A. Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126*, 2022.

Xu, M., Ran, T., and Chen, H. De novo molecule design through the molecular generative model conditioned by 3d information of protein binding sites. *Journal of Chemical Information and Modeling*, 61(7):3240–3254, 2021a.

Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021b.

Yang, Y., Ouyang, S., Dang, M., Zheng, M., Li, L., and Zhou, H. Knowledge guided geometric editing for unsupervised drug design, 2022. URL https://openreview.net/forum?id=91muTwt1_t5.

Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D.-A., Fan, L., Yu, Z., Lan, S., Li, B., et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023.

Zhang, Z. and Liu, Q. Learning subpocket prototypes for generalizable structure-based drug design. *arXiv preprint arXiv:2305.13997*, 2023.

Zhang, Z., Min, Y., Zheng, S., and Liu, Q. Molecule generation for target protein binding with structural motifs. In *International Conference on Learning Representations*, 2023.

A. Self Augmentation

Here we offer more insights about self augmentation. During sampling, at time step t, we utilize the protein-ligand interaction information embedded in $\mathrm{PMINet}(\mathcal{M}_{t+1}^{\mathrm{pred}}, \mathcal{P})$ to guide the generative process itself. For efficient training, given a protein-ligand pair $(\mathcal{P}, \mathcal{M})$, due to inaccessibility of $\mathcal{M}_{t+1}^{\mathrm{pred}}$, we directly replace it with the ground truth ligand molecule $\mathcal{M}$.

Because the training objective is to generate $\mathcal{M}$ given $\mathcal{P}$, a straightforward question is whether using $\mathcal{M}$ as input would provide a shortcut signal for the model and lead to its training collapse. Thanks to the design of PMINet, the model cannot naively rely on the input $\mathcal{M}$ to generate $\mathcal{M}$. More specifically, in PMINet, $\mathcal{M}$ and $\mathcal{P}$ are first input into two separate EGNNs, and only the produced SE(3)-invariant features, which are agnostic to the coordinate systems, are further input in the cross-attention layer to capture the protein-ligand interaction information. Thus, in the output produced by the cross-attention layer of PMINet, the relative positions and poses between the protein and ligand molecule in the physical world are eliminated, and only the protein-ligand interaction information in the feature space is kept. This means no shortcut signal is left for the model during training and the model still needs to normally learn from the protein context to generate the ligand molecule.

B. Implementation Details

B.1. Details of PMINet

Input Initialization To represent each protein atom, we use a one-hot element indicator {H, C, N, O, S, Se} and one-hot amino acid type indicator (20 types). Similarly, we represent each ligand atom with a one-hot element indicator {C, N, O, F, P, S, Cl}. Additionally, we introduce a one-dimensional flag to indicate whether the atoms belong to the protein or ligand. Two 1-layer MLPs are used to map the input protein and ligand into 128-dim latent spaces respectively.

Model Architectures We aim to use PMINet to model the complex 3D interactions between the atoms of proteins and ligands. To achieve this, we use two shallow SE(3)-equivariant neural networks for geometric message passing on the fully-connected graphs of the protein and ligand, respectively. We then apply a cross attention layer to the paired protein-ligand graph for learning the inter-molecule interactions. Finally, we use a sum-pooling layer to extract a global representation of the protein-ligand pair by pooling all atom nodes. And a two-layer MLP is introduced to predict the binding affinity S_{Aff}. More details about the model architecture are provided in Table 7.

Training Details During the training, we use the Mean Squared Error (MSE) loss with respect to the difference between the predicted and ground truth binding affinity scores as the optimization objective. The binding affinity values of protein-ligand pairs range from 2.0 to 11.92. For avoiding information leakage, we filter the training set by calculating the Tanimoto similarity with the molecules in the testing set of CrossDocked2020, and the similarity threshold was set to 0.1. As a result, there are 23 complexes filtered out from the training set. We train PMINet on a single NVIDIA V100 GPU, and we use the Adam as our optimizer with learning rate 0.001, $betas = (0.95, 0.999)$, batch size 16. The experiments are conducted on PDBBind v2016 dataset as mentioned in the main text.

Evaluation of PMINet We evaluate PMINet's effectiveness in predicting binding affinity, and compare it with a baseline model which is specifically designed for binding affinity prediction, *i.e.*, GraphDTA (Nguyen et al., 2021). Following Li et al. (2021), we select Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Pearson's correlation coefficient (R) and the standard deviation (SD) in regression to measure the prediction error. The testing results are present in Table 6, indicating the rationality of our model design. In IRDIFF, we use PMINet to serve as a binding-aware ligand retriever and a protein-ligand interaction information extractor.

B.2. Details of Interaction-based Retrieval-augmented Diffusion Model

Input Initialization For balancing the computational burden and the generation performance, we construct the retrieval 3D molecule database by randomly sampling 128 ligand molecules from the training set of CrossDocked2020, and the database is fixed in both training and testing. Then for each target protein, we use PMINet to scan the retrieval database and only select one ligand with the top-1 binding affinity predicted as the reference molecule. To represent each protein atom, we use a one-hot element indicator {H, C, N, O, S, Se} and one-hot amino acid type indicator (20 types). Similarly, we represent

Table 6. Performance of PMINet in binding affinity prediction. (↑) / (↓) denotes a larger / smaller number is better. Top 1 results are highlighted with **bold text**.

Methods	RMSE (↓)	MAE (↓)	SD (↓)	R (↑)
GraphDTA	1.562	**1.191**	1.558	0.697
PMINet	**1.554**	1.193	**1.520**	**0.716**

each ligand atom using a one-hot element indicator {C, N, O, F, P, S, Cl}. Additionally, we introduce a one-dimensional flag to indicate whether the atoms belong to the protein or ligand. Two 1-layer MLPs are introduced to map the inputs of protein and ligand into 128-dim spaces respectively. For representing the connection between atoms, we introduce a 4-dim one-hot vector to indicate four bond types: bond between protein atoms, ligand atoms, protein-ligand atoms or ligand-protein atoms. And we introduce distance embeddings by using the distance with radial basis functions located at 20 centers between 0 Å and 10 Å. Finally we calculate the outer products of distance embedding and bond types to obtain the edge features.

Model Architectures At the l-th layer, we dynamically construct the protein-ligand complex with a k-nearest neighbors (knn) graph based on coordinates of the given protein and the ligand from previous layer. In practice, we set the number of neighbors $k_n = 32$. As mentioned in Section 4.3, we apply an SE(3)-equivariant neural network for message passing. The 9-layer equivariant neural network consists of Transformer layers with 128-dim hidden layer and 16 attention heads. Following Guan et al. (2023a), in the diffusion process, we select the fixed sigmoid β schedule with $\beta_1 = 1\text{e}{-7}$ and $\beta_T = 2\text{e}{-3}$ as variance schedule for atom coordinates, and the cosine β schedule with $s = 0.01$ for atom types. The number of diffusion steps are set to 1000.

Training Details We use the Adam as our optimizer with learning rate 0.001, $betas = (0.95, 0.999)$, batch size 4 and clipped gradient norm 8. We balance the atom type loss and atom position loss by multiplying a scaling factor $\lambda = 100$ on the atom type loss. We train the parameterized diffusion denoising model of our IRDIFF on a single NVIDIA V100 GPU, and it could converge within 200k steps.

Table 7. Details of both PMINet and Interaction-based Retrieval-augmented Diffusion Model in our IRDIFF

Network	Module	Backbone	Input Dimensions	Output Dimensions	Blocks
PMINet	Protein Encoder	EGNN	$N_P \times 128$	$N_P \times 128$	2
	Ligand Encoder	EGNN	$N_M \times 128$	$N_M \times 128$	2
	Interaction Layer	Graph Attention Layer	$(N_P + N_M) \times 128$	$(N_P + N_M) \times 128$	1
	Pooling	Sum-pooling	$(N_P + N_M) \times 128$	1×128	1
Sequence-based PMINet	Protein Encoder	Graph Attention Layer	$N_P \times 128$	$N_P \times 128$	2
	Ligand Encoder	Graph Attention Layer	$N_M \times 128$	$N_M \times 128$	2
	Interaction Layer	Graph Attention Layer	$(N_P + N_M) \times 128$	$(N_P + N_M) \times 128$	1
	Pooling	Sum-pooling	$(N_P + N_M) \times 128$	1×128	1
Interaction-based Retrieval-augmented Diffusion Model	Position Dynamics	Transformer	$(N_P + N_M) \times 3$	$(N_P + N_M) \times 3$	9
	Atom Type Dynamics	Transformer	$(N_P + N_M) \times 128$	$(N_P + N_M) \times 128$	9
	Protein Fusion Layer	MLP	$N_P \times (128 + 128)$	$N_P \times 128$	1
	Ligand Fusion Layer	CrossAttention	$\{N_M \times 128, N_{\text{ref}} \times 128\}$	$N_M \times 128$	1

C. Ablation Study

C.1. Effectiveness of Interaction-based Retrieval

We investigate the impact of the reference ligands' binding affinity on the generation performance of IRDIFF. In our experiments, we choose ligands with the highest (i.e., k=1) target binding affinity as references, and we set the size n of reference pool to 128. Here, the reference ligand is replaced by the one with the lowest binding affinity based on the ranking list provided by PMINet. The experiments are conducted on the IRDiff **without** utilizing self augmentation mechanism. As indicated in Table 8, utilizing low-affinity references (with an average predicted binding affinity of 4.63 by PMINet between the reference molecule and the corresponding protein pocket) in the molecule database results in poorer performance on binding-related metrics compared to using high-affinity references (with an average binding affinity of 7.24). This demonstrates the effectiveness of our designs for interaction-based retrieval and reference utilization in IRDiff.

Table 8. The impact of molecule database with different properties. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively.

Methods	Avg. Affinity of Ref. (↑)	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)	
		Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-	-5.04	-5.75	-6.38	-6.52	-7.55	-7.72	54.2%	54.1%	0.46	0.46	0.57	0.57	0.71	0.69
low-affinity references	3.95	-4.12	-5.58	-6.09	-6.52	-7.84	-7.91	59.3%	60.2%	0.48	0.47	**0.58**	**0.59**	**0.73**	0.71
high-affinity references	**7.24**	**-5.39**	-6.28	**-6.40**	**-6.67**	**-8.14**	**-8.37**	**64.0%**	**71.5%**	**0.51**	**0.52**	0.57	0.57	0.72	**0.72**

C.2. Reusing a Set of Reference Molecules During the Generation Process

Table 9 demonstrates that the diversity of our IRDiff did not significantly decrease after the introduction of retrieval augmentation. The reason for this phenomenon may be attributed to the fact that our retrieval augmentation primarily relies on atom-wise cross-attention. Consequently, for the same protein pocket, we can repeatedly use the same reference molecule and apply retrieval augmentation to improve the performance of the molecule generation in IRDiff. This outcome indicates that our method does not require the retrieval operation in the first stage in every generation process, thereby avoiding a corresponding increase in the complexity of our algorithm with the addition of the retrieval process.

Table 9. The complete ablation study results of self augmentation and retrieval augmentation in IRDiff (n=128, k = 1).

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-5.04	-5.75	-6.38	-6.52	-7.55	-7.72	54.2%	54.1%	0.46	0.46	0.57	0.57	0.71	0.69
augmentation without PMINet	-4.14	-5.64	-6.11	-6.36	-7.60	-7.67	54.9%	57.5%	0.47	0.48	0.57	0.57	0.70	0.70
+ self augmentation	-4.91	-6.03	-6.53	-6.82	-7.95	-8.14	61.5%	64.4%	**0.52**	**0.53**	**0.59**	**0.58**	0.72	0.71
+ retrieval augmentation	-5.39	-6.28	-6.40	-6.67	-8.14	-8.37	64.0%	71.5%	0.51	0.52	0.57	0.57	**0.74**	**0.73**
+ both augmentations	**-5.86**	**-6.51**	**-7.14**	**-7.27**	**-8.33**	**-8.49**	**66.8%**	**73.9%**	**0.53**	**0.54**	**0.58**	**0.58**	**0.74**	0.72

C.3. Augmentation Position

We investigate what the best position of our augmentation mechanisms is in IRDiff. *early*, *middle*, *late* and *layer-wise* means we conduct augmentation in the first, middle, last and each layer of our diffusion denoising networks in Equations (14) and (15), respectively. To simplify the experiments, we only use self augmentation. As presented in Table 10, the augmentation position does not have much impact on property-related metrics, consistently improving the baseline. Regarding binding-related metrics, the augmentation position plays a role in final performance. Thus, we select **early augmentation** in practice for better trade-off between binding- and property-related metrics.

Table 10. The effect of augmentation position in IRDiff.

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-5.04	-5.75	-6.38	-6.52	-7.55	-7.72	54.2%	54.1%	0.46	0.46	0.57	0.57	0.71	0.69
early augmentation	-4.91	-6.03	-6.53	-6.82	**-7.95**	**-8.14**	61.5%	64.4%	**0.55**	**0.57**	**0.62**	**0.61**	0.72	0.71
middle augmentation	**-5.07**	-6.05	-6.49	-6.64	-7.87	-7.97	**62.2%**	**65.7%**	0.51	0.52	0.60	0.59	**0.74**	0.72
late augmentation	-4.90	**-6.17**	**-6.57**	**-6.85**	-7.79	-8.09	61.6%	64.0%	0.53	0.54	0.60	0.58	0.72	0.70
layer-wise augmentation	-4.16	-5.78	-6.20	-6.56	-7.74	-7.97	60.0%	60.3%	0.53	0.55	0.61	0.59	0.73	**0.73**

C.4. Complete Results of the Ablation Study for Hyper-parameters n and k

We present the comprehensive results of the ablation study for hyper-parameters n and k in 11, where n and k denote the size of reference pool and the number of utilized references, respectively.

Table 11. The effect of the hyper-parameter n and k in IRDIFF, where n and k denote the size of reference pool and the number of utilized references, respectively. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively.

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
$n = 32$	-5.67	-6.21	-7.01	-7.13	-8.13	-8.30	65.2%	72.0%	0.51	0.52	0.56	0.57
$n = 64$	-5.74	-6.39	-7.07	-7.19	-8.21	-8.35	65.9%	72.8%	0.52	0.53	0.57	**0.58**
$n = 128$	**-5.86**	**-6.51**	**-7.14**	**-7.27**	**-8.33**	**-8.49**	66.8%	**73.9%**	**0.53**	**0.54**	**0.58**	**0.58**
$k = 1$	-5.86	-6.51	-7.14	-7.27	-8.33	**-8.49**	**66.8%**	73.9%	**0.53**	**0.54**	0.58	**0.58**
$k = 2$	-5.88	-6.62	-7.25	-7.29	-8.34	-8.42	66.6%	**74.4%**	**0.53**	**0.54**	0.57	0.56
$k = 3$	**-6.03**	**-6.89**	**-7.27**	**-7.37**	**-8.42**	-8.42	**67.4%**	72.7%	**0.53**	**0.54**	**0.59**	**0.58**

C.5. Sequence-based Protein-Molecule Interaction Networks

Significantly, our IRDIFF effectively leverages the protein-molecule interactions modeled by PMINet for 3D protein-specific molecule generation, even when PMINet solely focuses on modeling sequence-level interactions between proteins and ligands. This capability enhances the potential for our method to be widely applicable and scalable. For example, it enables PMINet to utilize self-supervised learning techniques to capture interactions from a vast number of protein-molecule sequence pairs without relying on labeled binding affinities in the future work. To verify it, we introduce the sequence-based PMINet, denoted as PMINet-Seq in Table 7. PMINet-Seq consisting of graph attention layers (Velickovic et al., 2017) and cross-attention layers (Borgeaud et al., 2022; Hou et al., 2019). Two graph attention layers are applied on the fully-connected graphs of the protein $\mathcal{G}_P$ and ligand molecule $\mathcal{G}_M$ to model the inner-molecule interactions. Given a molecule graph $\mathcal{G}_M$, the l-th graph attention layer works as follows:

$$\mathbf{h}_{M,i}^{l+1} = \mathbf{h}_{M,i}^{l} + \sum_{j \in \mathcal{N}_M(i)} f_{M,h}^{l}\left(\mathbf{h}_{M,i}^{l}, \mathbf{h}_{M,j}^{l}\right) \tag{18}$$

where $\mathbf{h}_{M,i}^{l+1} \in \mathbb{R}^d$ is the SE(3)-invariant hidden states of the atom i of the ligand after the l-th layer. $\mathcal{N}_M(i)$ stands for the set of neighbors of atom i on $\mathcal{G}_M$, and the initial hidden state $\mathbf{h}_{M,i}^{0}$ is obtained by an embedding layer that encodes atom information. Given a protein graph $\mathcal{G}_P$ and $\mathbf{h}_{P,i}^{l}$ can be derived in the same way. Following the training procedure of PMINet, we pre-train PMINet-Seq on PDBBind v2016 with the supervision of binding affinity to capture protein-molecule interactions. Table 12 presents the result of replacing the PMINet with the PMINet-Seq in IRDIFF for 3D molecule generation. It demonstrates that our IRDIFF is able to capture interaction hints from the pre-trained protein-molecule interaction networks for 3D molecule generation, even when the protein-molecule interaction network solely focuses on the sequence-level interactions between protein-molecule pairs.

Table 12. The performance of utilizing PMINet-Seq in IRDIFF. (↑) / (↓) denotes a larger / smaller number is better. Top 2 results are highlighted with **bold text** and underlined text, respectively.

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		High Affinity (↑)		QED (↑)		SA (↑)		Diversity (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
baseline	-5.04	-5.75	-6.38	-6.52	-7.55	-7.72	54.2%	54.1%	0.46	0.46	0.57	0.57	0.71	0.69
TargetDiff	-5.47	-6.30	-6.64	-6.83	-7.80	-7.91	58.1%	59.1%	0.48	0.48	0.58	0.58	0.72	0.71
IRDIFF with PMINet	**-5.86**	-6.51	**-7.14**	**-7.27**	**-8.33**	**-8.49**	**66.8%**	**73.9%**	**0.53**	**0.54**	**0.58**	**0.58**	**0.74**	**0.72**
IRDIFF with PMINet-Seq	-5.80	**-6.61**	-6.94	-7.09	-8.06	-8.14	62.1%	67.4%	0.51	0.52	0.58	0.57	0.71	0.70

D. More Visualization Results

We provide the visualization of more ligand molecules generated by IRDIFF, comparing to both reference and TargetDiff (Guan et al., 2023a), as shown in Figure 7.

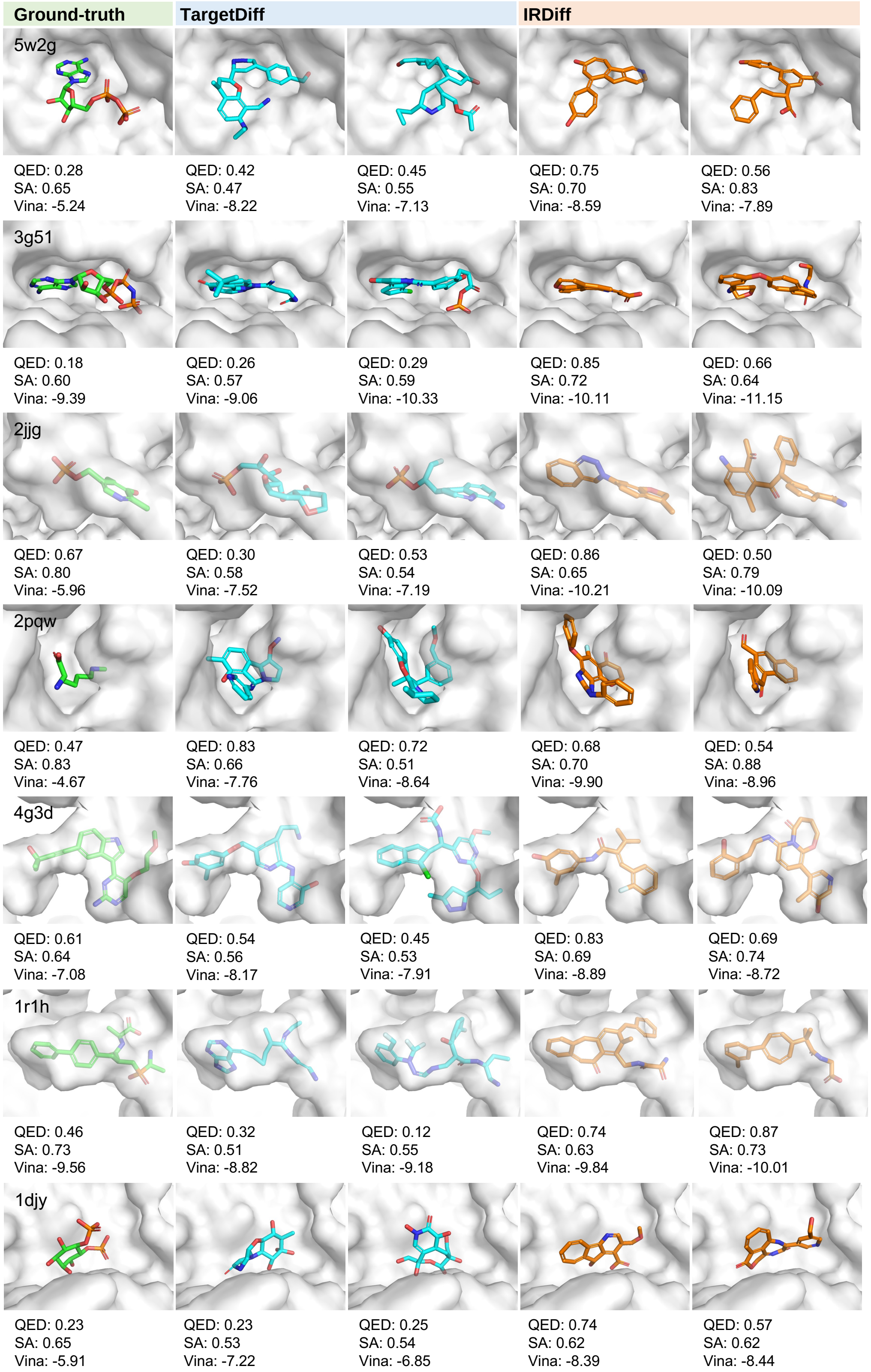

Figure 7. Examples of generated ligands. Carbon atoms in grouth-truth ligands of test set, ligands generated by TargetDiff (Guan et al., 2023a) and our model are visualized in green, cyan, and orange respectively. We report QED, SA, and Vina Dock score for each molecule.