Molecule Meets Protein Pocket
3D-Aware Molecular Optimization for Protein Targets

Anonymous authors
Paper under double-blind review

Abstract

Lead optimization, refining drug candidates to improve binding to protein targets, is a key challenge in drug discovery. We introduce a 3D-aware generative framework that performs fragment-level molecular optimization conditioned on the geometry of the protein's binding pocket. Our model represents the molecule-protein complex as a sparse 3D graph and applies grouped vector attention to learn spatial interactions. It decomposes the molecule into a stable scaffold and generates new fragments using a Variational Autoencoder (VAE) and a SMILES-based transformer guided by local pocket structure. To handle the imbalance in fragment sizes, we incorporate a focal loss. On the CrossDock2020 benchmark, our method outperforms prior approaches in generating diverse, novel, and chemically valid candidates with improved Vina scores-while generalizing to unseen proteins.

1 Introduction

Lead optimization-tuning a candidate molecule to improve binding to a target protein-aims to modify the molecule while retaining its core structure. One frequent goal is to enhance binding to a specific protein target, but changes must also maintain favorable properties of the original compound. Recent machine learning approaches have shown promise in automating this process, but most treat it as a molecule-only task-ignoring the protein entirely and relying solely on SMILES strings or molecular graphs (Zhou et al., 2019; You et al., 2018; Barshatski & Radinsky, 2021). Some models incorporate protein sequences, such as FASTA (Kaminsky et al., 2023), but these lack the spatial detail necessary to capture molecular binding. Others leverage 3D protein information for docking or property prediction (Varadi et al., 2022; Ketata et al., 2023), and some have explored structure-aware generation (Huang et al., 2024; Schneuing et al., 2024). However, these models are often diffusion-based and primarily designed for de novo generation as opposed to optimization and require manual input, limiting automation and practical applicability.

We introduce **Mo**lecular **Do**cking-based **L**ead **O**ptimization (MODOLO), a novel method that reframes lead optimization as a 3D fragment generation task conditioned on the geometry of the protein's binding pocket. Given a target protein and a candidate molecule, we first compute their docking pose to determine how the molecule fits within the binding site. This docked complex is represented as a sparse graph, where nodes correspond to the molecule's atoms, the pocket's atoms, the alpha carbons of the protein's amino acid residues, and a special hole node representing the attachment point of the fragment. Edges connect nodes that lie within chemically meaningful distances, capturing relevant spatial interactions. A graph neural network encoder with grouped vector attention then learns a representation of this 3D chemical context. Next, we decompose the molecule into a stable scaffold using a hierarchical pruning strategy based on (Schuffenhauer et al., 2007), preserving approximately 70% of its mass and creating "holes" where the remaining fragments were attached. For each hole, the pocket-aware embedding from the graph encoder is passed to a VAE-based transformer, which generates a replacement fragment predicted to bind more effectively within the pocket. Reattaching the generated fragments results in an optimized analogue that remains structurally close to the original molecule but is fine-tuned to the 3D shape of the target binding site.

To support generalization to unseen targets, we train MODOLO using a Vina score objective on the CrossDock2020 dataset (Francoeur et al., 2020). A focal loss ensures robust learning despite the imbalance in fragment distribution, while the CVAE architecture allows for diverse fragment generation.

The experiments on synthetic CrossDocked dataset demonstrate that MODOLO can generate drug-like, synthesis-accessible, diverse molecules with high binding affinity against specific proteins and outperform the state-of-the-art (SOTA) models on multiple evaluation metrics.

The contributions of this work are summarized as follows:

- We propose MODOLO, a novel 3D-aware generative framework for lead optimization that conditions fragment generation on the geometric constraints of the protein binding pocket while preserving the core molecular scaffold.
- We design a specialized architecture that utilizes a sparse heterogeneous graph representation of the protein-ligand complex and employs grouped vector attention to capture fine-grained spatial interactions.
- We demonstrate through extensive experiments on the CrossDock2020 benchmark that our method achieves state-of-the-art performance, generating diverse, chemically valid, and high-affinity drug candidates that outperform existing diffusion and sequence-based baselines.
- We release our code and pre-trained models to support further research in structure-based drug design: `https://anonymous.4open.science/r/modolo-F13D`.

2 Related Work

In this work, we address the task of molecular optimization with the aim of generating chemically valid molecules that preserve the scaffold of an input molecule while improving its interaction with a specific protein. Unlike general molecular generation, which aims to design molecules with favorable properties from scratch, molecular optimization emphasizes structural similarity to a given molecule, creating a unique challenge. Our approach leverages 3D spatial information from the molecular docking of the molecule with the target protein to guide the optimization process.

Protein-Free Optimization Methods. Many molecular optimization models focus on modifying molecules without explicitly accounting for the protein target they are intended to bind. These approaches typically represent molecules using SMILES strings or molecular graphs (Blaschke et al., 2018; Gómez-Bombarelli et al., 2018; Harel & Radinsky, 2018; Olivecrona et al., 2017; Popova et al., 2018; You et al., 2018; Zhou et al., 2019; Fu et al., 2020; Jin et al., 2020; 2019; Liu et al., 2018). While such methods have been highly successful at generating syntactically valid and drug-like compounds, they do not directly incorporate information about the protein binding site, which can be important for optimizing target-specific bioactivity.

Our method addresses this critical gap by explicitly modeling the 3D geometry of the protein binding pocket and conditioning the fragment generation process on spatial interactions between the molecule and the target. This enables the model to generate structurally compatible analogues that are more likely to bind effectively in real biological settings.

Related techniques such as scaffold hopping aim to discover structurally novel compounds by altering the core scaffold of known actives (Böhm et al., 2004; Zheng et al., 2021). While useful for scaffold diversity, such methods are less applicable to lead optimization tasks, where preserving the original molecule's scaffold is often essential. In contrast, our approach explicitly retains the scaffold and focuses on generating new fragments tailored to the 3D shape of the protein binding site.

Protein-Aware Generative Models. Recent methods have begun incorporating protein structure into molecular generation, particularly in the context of de-novo drug design. Pocket2Mol (Peng et al., 2022) is one of the earlier works to directly condition molecule generation on 3D protein pocket structures. It employs a two-stage approach that first extracts pocket features using 3D convolutional networks and then generates molecules atom-by-atom using an autoregressive model. Different kind of approaches (Huang et al., 2024;

Schneuing et al., 2024; Lin et al., 2025; Guan et al., 2023; Chen et al., 2025) use diffusion models to generate entirely new compounds within the 3D pocket. Within this set of methods, the work of (Schneuing et al., 2024) is the only one that enables optimization through the application of partial diffusion, thus keeping the new molecule close to the origin. However, when adapted for optimization, they often discard the bonds of the original molecule, making it difficult to retain pharmacological properties and similarity to the original molecule. The method proposed by Huang et al. (2024) also faces these challenges and additionally requires manual intervention, such as choosing which atoms to modify, which limits automation. Another approach, CFOM (Kaminsky et al., 2023) improves chemical realism by editing only small fragments of the molecule to preserve similarity, and utilizes an encoding of the target protein, and thus losing a lot of the physical context of binding. MODOLO bridges this gap by combining geometric deep learning with fragment-based optimization. It models the molecule-protein complex as a 3D graph and uses graph neural networks with grouped vector attention to capture spatial interactions.

Molecular Docking. Molecular docking refers to the task of predicting how a small molecule fits into the 3D structure of a protein-essentially estimating the "pose" of the molecule within the binding site (Figure 2). There are two main approaches: classical search-based methods such as AutoDock Vina (Trott & Olson, 2010; McNutt et al., 2021), which use physics-inspired scoring functions and optimization algorithms to find likely poses, and newer machine learning-based methods that directly predict the docking pose using deep neural networks (Ketata et al., 2023; Ganea et al., 2021; Stärk et al., 2022). While docking methods are crucial for estimating how molecules interact with proteins, they are not designed to perform molecular optimization. Our work assumes a docking pose is already available and focuses instead on modifying the molecule itself to improve bioactivity-conditioned on the 3D geometry of the protein binding site. In other words, docking predicts how a molecule fits; we focus on what molecule to generate to best fit.

Protein-Drug Interaction Prediction. Numerous neural network architectures have been developed to predict interactions between small molecules and proteins (Chen et al., 2020; Huang et al., 2021; Lee et al., 2019). These models are typically multi-modal, taking as input both a molecular graph or SMILES representation and either the amino acid sequence or 3D structure of the target protein (Huang et al., 2024; Green et al., 2021). Other approaches incorporate docking poses directly to predict binding affinity based on spatial interactions between the molecule and protein pocket (Moon et al., 2024; Wang et al., 2021). In contrast to these prediction-focused methods, our work addresses a generative task: learning to modify an existing molecule to better fit a given protein binding site-thereby optimizing for biological activity rather than just predicting it.

3 MODOLO Algorithm

Given a candidate molecule and a target protein, our goal is to optimize the molecule by modifying its peripheral fragments to improve binding affinity to the protein, while preserving its core structure. We assume as input a *docking pose* (see an example in Fig 2), a 3D assignment of atomic coordinates in $\mathbb{R}^3$ for both the protein and the molecule, representing how the molecule fits into the protein's binding pocket. From this complex, we construct a graph that captures spatial relationships between atoms in both the molecule and the protein. The molecule is then decomposed into a *scaffold* (the stable core) and one or more *fragments* (the modifiable parts). We formulate the optimization task as generating new fragments that, when reattached to the scaffold, yield an improved molecule that (1) fits the protein pocket more effectively and (2) remains similar to the original molecule.

MODOLO addresses this task through a modular pipeline (see Fig 1). First, with the 3D docking pose of the molecule in the protein's binding pocket, we build a sparse, heterogeneous graph representation of the molecule-protein complex, capturing geometric and biochemical features (see Section 3.1). Next, we extract chemically meaningful scaffolds of the molecule and identify the associated fragments to be regenerated (see Section 3.2). These graphs are then processed using a geometric deep learning module with *grouped vector attention*, which learns spatial interactions across molecule and protein nodes (see Section 3.3). The original, masked fragments are embedded and used as an input to a variation autoencoder. Finally, a transformer decoder generates optimized fragments conditioned on the scaffold, the latent representation of the original

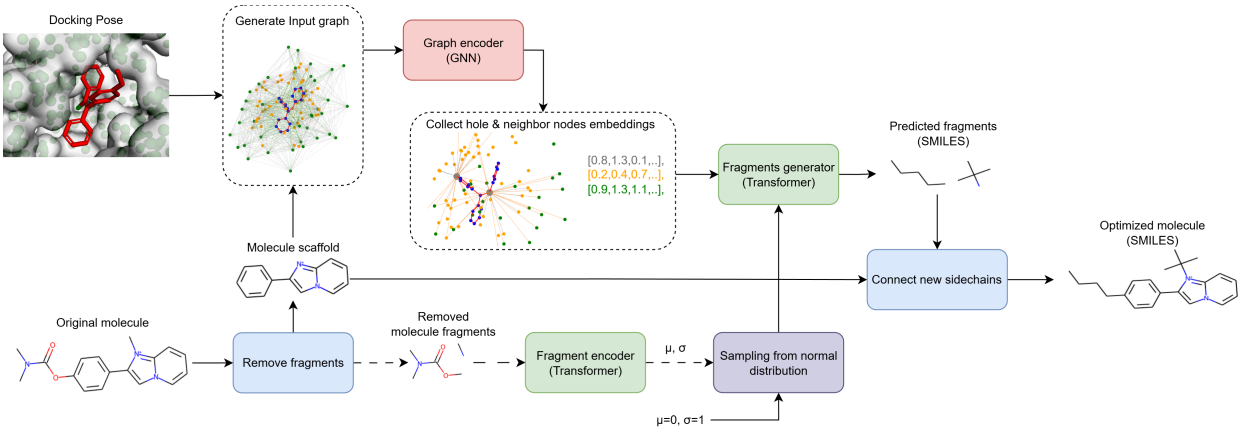

Figure 1: An overview of the Algorithm. The inputs a molecule and a docking pose of the molecule in a target protein. The molecule is split into a scaffold and fragments. The fragments are masked and the docking pose is fed into a graph based encoder. The encoded nodes of the graph are used as input to a text based fragment decoder that outputs the new, target specific fragments. Finally, the new fragments are reattached to the scaffold and the result is a new molecule that remains similar to its origin and has a stronger interaction with the target. The original, massked fragments are embedded and used as an input to a variation autoencoder. The dotted arrows represent flows of information that accure during training only.

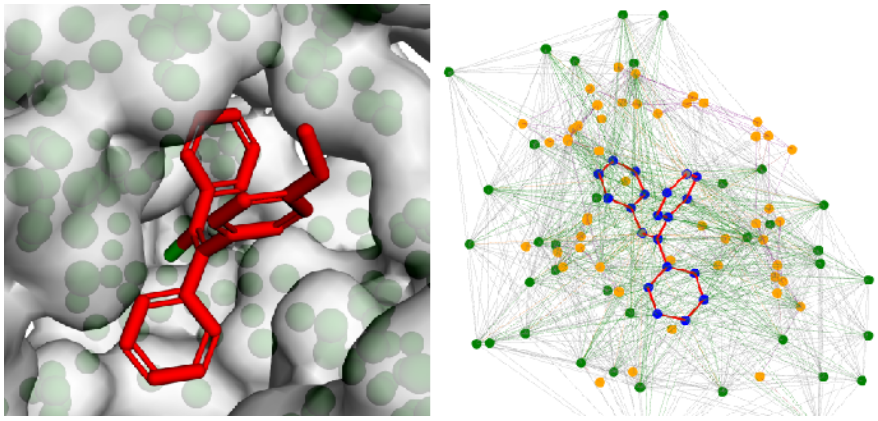

Figure 2: The image on the left depicts an input docking pose of a molecule within a pocket. The image on the right depicts the heterogeneous graph constructed from the docking pose. Ligand atoms are shown in blue, protein receptor atoms in orange, and amino acids in green. Hole nodes are shown in grey. The difference in connection distances (cutoffs) between receptor atoms and residues is illustrated by their respective edge lengths.

fragments (during training only) and the protein context (see Section 3.4). In the following subsections, we detail each component of this pipeline.

3.1 Input Graph

To represent molecular docking structures, we construct heterogeneous geometric graphs (see an example in Fig 2) that encode ligand atoms, receptor amino acids (represented spatially by their central α carbon), receptor atoms, and a special **hole** node. The hole node is positioned at the atom connecting the fragment to the scaffold. To ensure computational efficiency and incorporate useful inductive biases, the graphs are sparsely connected based on distance-dependent cutoffs. Specifically, covalent and spatial interactions between ligand and receptor atoms are connected using a 5 Å cutoff, while receptor residues use larger cutoffs of 10 Å to capture long-range contextual information. Structure-based edges are defined by these proximity rules, while ligand bonds are preserved. Node features include biochemical properties, and ESM2 embeddings (Lin et al., 2023) are used for amino acid nodes. A detailed breakdown of the cutoff values and feature definitions is provided in appendix A.

3.2 Scaffold and Fragments Extraction

To systematically deconstruct each molecule into a scaffold and its fragments, we employ a scaffold-based decomposition strategy inspired by medicinal chemistry principles. The primary motivation is to preserve the core structural and functional features of the molecule while allowing controlled modification through fragment generation.

We begin by extracting the Murcko scaffold of the molecule using RDKit (Landrum, 2016). This scaffold captures the core ring systems and linkers essential to the molecule's chemical identity. However, relying solely on the Murcko scaffold may not always provide the optimal balance between preserving key properties and allowing sufficient flexibility for optimization. Therefore, we apply the hierarchical scaffold ordering rules proposed by (Schuffenhauer et al., 2007), implemented using ScaffoldGraph (Scott & Edith Chan, 2020), to rank and prune peripheral ring systems iteratively.

This pruning process removes less-characteristic substructures while considering topological and chemical characteristics such as ring size, heteroatom content, and attachment points. From the ranked list of scaffold candidates, we select the smallest scaffold whose molecular weight constitutes at least 70% of the original compound. The 70% threshold was empirically chosen as it ensures that the retained scaffold captures the majority of the molecule's pharmacophoric features, while still allowing meaningful optimization through fragment addition.

The removed substructures are defined as fragments. Formally, for a molecule m, we denote its set of fragments as $F = \{f_1, \ldots, f_n\}$, where each f_i is a disconnected component removed during scaffold extraction. For each fragment f_i, a hole h_i is defined as an atom in the scaffold that has a bond with it in the original molecule. Each molecule has a single scaffold but may yield multiple fragments, each fragment has a single hole, due to the extraction process not splitting rings.

An illustrative example of the scaffold extraction process, including weight ratios and intermediate scaffolds, is shown in Figure 3.

3.3 Interaction Layers

After constructing the heterogeneous graph representing the molecule protein complex and decomposing the molecule into scaffold and fragments, the next step is to model the spatial and chemical interactions within this graph. To effectively capture how atoms and residues interact in 3D space, the model must go beyond simple message passing and incorporate directional, context-dependent relationships between nodes.

Our model utilizes vector attention, a mechanism where attention weights are vectors that modulate individual feature channels of node embeddings (Zhao et al., 2021). This approach allows the network to learn fine-grained interactions between nodes in the input graph, which is crucial for capturing the complex spatial and chemical relationships present in molecular docking structures.

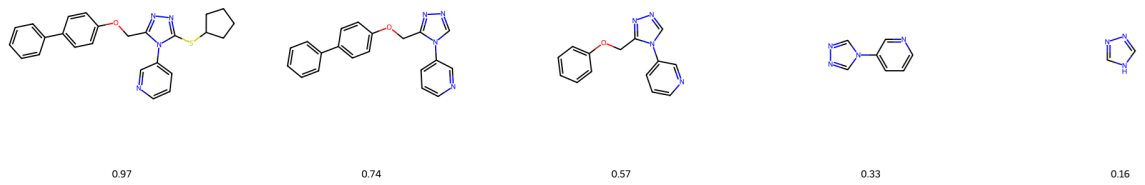

Figure 3: An example of the molecule scaffold extraction process in steps. The original molecule is shown on the left, every following scaffold is the result of a single fragment removal in the iterative process. Below are the ratios between the molecular weight of the scaffold and the original molecule.

The interaction layer operates as follows: For each node i in the graph, its updated embedding x_i is computed by aggregating information from its neighbors $x_j \in X(i)$, where $X(i)$ is the set of neighboring nodes and $E(i)$ is the set of edges connecting i to its neighbors. The relation between node i and each neighbor j is modeled using a subtraction-based relation function, which enhances the model's ability to capture directional and contextual differences.

The attention weight w^{ij} for each neighbor is computed as:

$$w^{ij} = \gamma\left(\varphi(x_i) - \psi(x_i, e_k, x_j)\right)$$

where φ and ψ are multi-layer perceptrons (MLPs) that encode the node and edge features, e_k is the edge feature between i and j, and γ is an activation function (e.g., sigmoid or softmax) that normalizes the attention weights.

The updated node embedding is then given by:

$$x_i = \sum_{\substack{x_j \in X(i) \\ e_k \in E(i)}} w^{ij} \cdot \alpha(x_i, e_k, x_j)$$

where α is a linear transformation applied to the concatenated features of x_i, e_k, and x_j.

While vector attention provides expressive power, it introduces a large number of parameters, especially in deep networks. To address this, we employ **grouped vector attention** (Wu et al., 2022), which divides the feature channels into groups. Each group shares a single attention weight, reducing the number of parameters and improving generalization.

Formally, for n feature channels divided into g groups, the updated embedding for the m-th channel is:

$$x_{i_m} = \sum_{\substack{x_j \in X(i) \\ e_k \in E(i)}} w^{ij}_{\lfloor m/g \rfloor} \cdot \alpha(x_i, e_k, x_j)_m$$

where $w^{ij} \in \mathbb{R}^{n/g}$ is the vector of attention weights for each group, and $\alpha(x_i, e_k, x_j)_m$ denotes the m-th channel of the transformed features.

This grouped attention mechanism enables the model to efficiently capture complex interactions in large graphs while maintaining computational tractability and strong generalization performance.

3.4 Fragments Generator

Once the model has encoded the spatial and chemical context of the scaffold and surrounding protein environment through the interaction layers, the next step is to generate optimized molecular fragments.

The Fragments Generator architecture consists of a variational Autoencoder (VAE) (Kingma & Welling, 2019) that conditions the generation on the structural context. The encoder of the VAE takes the fragment

SMILES (during training) and embeds them using a transformer encoder. The decoder is a transformer decoder (Vaswani et al., 2017). For input graph $\mathcal{G}$, hole h and fragment f, the input to the decoder includes:

- The embeddings of nodes that were connected to hole h in graph $\mathcal{G}$, which are added as the *memory* of the decoder.
- hole h embedding is introduced as the *first target token* for the generation sequence.

The decoder outputs the new fragment as a SMILES string, maximizing the likelihood of the valid fragment given the context. Note that each SMILES fragment is generated separately.

The goal of the model is to regenerate the fragment f SMILES from the masked ligand in a protein-ligand complex. To this end, our loss function is a combination of a reconstruction loss (Section 3.4.1) and a VAE loss (Section 3.4.2):

$$L_{total}(f, \mathcal{G}, h) = L_{focal\_recon}(f, \mathcal{G}, h) + \beta \cdot L_{KL}(f)$$

where β is a hyperparameter that controls the trade-off between the two terms.

3.4.1 Reconstruction Loss

The reconstruction loss is defined as the negative log-likelihood of the true fragment SMILES given the scaffold and hole. The reconstruction loss of a single fragment token is given by:

$$L_{recon}(f, \mathcal{G}, h, t) = -\log p(f_t | f_{<t}, h, \mathcal{G})$$

where f_t is the t-th token in the true fragment f SMILES, $f_{<t}$ are the previous tokens of the SMILES. Due to the data imbalance, we employ a **Focal Loss** to handle the data imbalance, particularly the prevalence of single-atom fragments. The focal loss is defined as:

$$L_{focal\_recon}(f, \mathcal{G}, h, t) = (1 - p(f_t | f_{<t}, h, \mathcal{G}))^{\gamma} \cdot L_{recon}(f, \mathcal{G}, h, t)$$

where γ is the focusing parameter. The focal loss puts more emphasis on hard-to-classify examples, where the model is less confident in its predictions, which are the non-single-atom fragments in this case. To further battle the imbalance, the total loss of a fragment is computed as the mean of the focal loss of each token, and weighted by the length of the fragment:

$$L_{focal\_recon}(f, \mathcal{G}, h) = \frac{\log T}{T} \sum_{t=1}^{T} L_{focal\_recon}(f, \mathcal{G}, h, t)$$

where T is the length of the fragment f.

3.4.2 KL divergence

The second component is the Kullback-Leibler (KL) divergence, which regularizes the learned latent distribution to approximate the standard normal prior $p(z) = \mathcal{N}(0, I)$:

$$L_{KL}(f) = D_{KL}\Big(q_\phi(z|f)\Big|\Big|p(z)\Big) = -\frac{1}{2}\sum_{i=1}^{D}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

This constraint ensures a continuous latent space for valid sampling. However, to prevent *posterior collapse*, where the autoregressive decoder ignores the latent code z, we employ **Cyclical KL Annealing** (Fu et al., 2019). By cyclically increasing the KL weight β from 0 to 1, we allow the decoder to learn reconstruction using unconstrained latent information before gradually enforcing the regularization.

3.5 Inference

During inference, we sample from the prior latent distribution and decode the SMILES string conditioned on the hole and context embeddings. The validity of the generated token is checked, in case of an error, the fragment is resampled.

4 Experimental Setup

4.1 Datasets

We train and evaluate our model using the standard split of the CrossDocked2020 dataset (Francoeur et al., 2020). Originating from the Protein Data Bank, this dataset initially comprises 22.5 million docked protein-ligand pairs, generated using smina via the Pocketome workflow. To ensure data quality and consistency with prior work (Huang et al., 2024), we utilize a filtered subset restricted to high-quality binding poses with a root-mean-squared deviation (RMSD) of less than 1Å. To prevent data leakage and ensure structural diversity, the dataset was clustered at 30% sequence identity using MMseqs2. The final processed dataset consists of 100,000 complexes for training and 80 complexes for evaluation.

4.2 Model Implementation

Our model is implemented using PyTorch and PyTorch Lightning. The 3D pocket encoder is a 4-layer graph neural network (GNN), where each layer includes grouped vector attention with 8 groups and applies dropout with a rate of 0.3. The decoder and encoder are both 2-layer Transformer with 4 attention heads per layer. Optimization is performed using the Adam optimizer(Kingma & Ba, 2015) with an initial learning rate of 10^-4 and a weight decay of 10^-4. The loss hyper parameters are $\beta = 0.1, \gamma = 2, M = 4$.

Experiments were run on a server equipped with 2 NVIDIA L40 GPUs (each with 48 GB of VRAM). Training on the CrossDock2020 dataset required approximately 6 hours to complete, including both training and validation phases. Inference for molecule generation takes 0.3 seconds per molecule on average using a single GPU.

4.3 Baselines

1. **DiffSBDD** (Schneuing et al., 2024). A diffusion-based model that generates molecules based on the 3D structure of the target pocket. The model can generate and optimize molecules.

2. **CFOM** (Kaminsky et al., 2023). Lead optimization model that encodes scaffold molecule Smiles and protein FASTA. Outputs interaction optimized side-chains for the molecule.

4.4 Ablations

To better understand the contribution of each component in our model, we conducted a comprehensive ablation study. For each ablation, we retrained the model under identical conditions and evaluated its performance on the CrossDocked benchmark using the same metrics as in the main experiments. Specifically, we evaluated the following variants:

- **No alpha carbons**: In this variant, we excluded the α-carbon nodes representing amino acid residues from the input graph, reducing the amount of structural and contextual information about the protein pocket. As a result, the model relies solely on the atomic-level representation of the protein, potentially limiting its ability to capture higher-level spatial relationships and long-range interactions that are important for accurate molecular optimization. Compared with the graph in Figure 2, the graph without alpha carbons is missing the green nodes.

- **No VAE**: We performed an ablation study where the VAE module was removed from the architecture. In the ablated model, the architecture is fully deterministic. This forces the decoder to map the 3D structural context of the hole and its neighbors directly to a single SMILES sequence, effectively removing the model's ability to sample diverse chemical outputs for a given geometric configuration.

- **No attention groups**: In this variant, we eliminated the use of grouped vector attention in the interaction layers, reverting to standard vector attention. This increases the number of parameters in the model and may lead to overfitting, especially in data-scarce scenarios. Grouped attention

is designed to improve generalization by sharing attention weights across groups of channels, so its removal tests the importance of this regularization mechanism.

- **Noisy docking Pose:** In this ablation, we introduce controlled perturbations to the molecular docking pose by applying random rotational transformations and random translations of different magnitudes to the molecular graph. This experiment is designed to evaluate the robustness of MODOLO to inaccuracies in the prediction of the docking pose.

4.5 Evaluation Metrics

To evaluate the structural fit within the protein pocket, we utilize the Vina Score, which estimates the binding affinity between the ligand and the target. The score is calculated by redocking the generated molecule to the protein pocket using Qvina2 (Alhossary et al., 2015). Regarding drug-likeness and developability, we report QED (Quantitative Estimation of Drug-likeness), SA (Synthetic Accessibility), and LogP (octanol-water partition coefficient), considering the range of -0.4 to 5.6 as optimal. We measure the model's generative diversity between the generated molecules and the original molecule, which is defined as

$$Diversity(org, opt) = 1 - Sim(org, opt)$$

where Sim is the Tanimoto similarity (Bajusz et al., 2015) between the two Morgan fingerprints of the molecules. We also measure success, defined by:

$$(Sim(org, opt) > 0.4) \wedge (Vina(org) > Vina(opt))$$

The definition of success is based on similar definitions used in prior works (Huang et al., 2024; Jin et al., 2019).

For each molecule in the test set, 20 new molecule where generated. The measured metric are averaged over all generated molecules.

5 Empirical Results

5.1 Baselines

The results of our evaluation on the test set are reported in Table 1. We observed that MODOLO got the best or second best results in all metrics except for SA. According to Vina score, MODOLO handles the task of keeping a high binding affinity to the target pocket while preserving a high structural similarity to the original molecule. Figure 4 shows that the distribution of Vina scores for MODOLO is more favorable than the other models as it is shifted downward in comparison to the other models. An intereseting observation is that MODOLO achieved the same mean as the reference molecules, but with a lower variance.

DiffSBDD achieved the lowest success rates among the compared methods. This is primarily due to its diffusion-based approach, which generates molecules by iteratively perturbing atomic features and 3D coordinates. During this process, explicit chemical bonds are not preserved. Instead, the final molecular graph is reconstructed by inferring bonds based on interatomic distances and chemical heuristics, such as covalent radii and valence rules. As a result, the generated molecules often lack sufficient structural similarity to the original compounds, leading to lower optimization success. Moreover, DiffSBDD's superior SA scores are attributed to its evolutionary algorithm, which explicitly monitors synthetic accessibility and retains only the best-performing candidates in each iteration, a selection process not employed by the other models.

CFOM was explicitly trained for the molecular optimization task with the objective of enhancing molecular activity against the target protein, but it has a very low diversity, which could be due the lack of a VAE component within the architecture, and a fragmenting method that retains to big of a scaffold.

The logP value of MODOLO falls within the compliance range ($-0.4 < LogP < 5.6$), which implies that the molecules generated by MODOLO hold greater promise as drug candidates, which is crucial for clinical trials.

Table 1: metrics of MODOLO and baseline models on the CrossDocked dataset.

Metric	Ref	MODOLO	CFOM	DiffSBDD
Success ↑	-	**0.36** ±0.23	0.27 ±0.20	0.01 ±0.01
Vina ↓	-7.32 ±3.48	**-7.31** ±3.53	-7.25 ±4.11	-7.11 ±4.10
Diversity ↑	-	0.40 ±0.06	0.31 ±0.09	**0.86** ±0.01
SA ↓	3.63 ±1.66	**3.62** ±1.82	3.74 ±1.75	4.08 ±1.51
QED ↑	0.48 ±0.05	0.51 ±0.05	0.50 ±0.04	**0.62** ±0.04
LogP ↑	1.02 ±8.33	1.33 ±8.20	1.59 ±8.48	2.49 ±3.21

The bold values indicate the best performance results, the underlined values indicate the second best performance results.

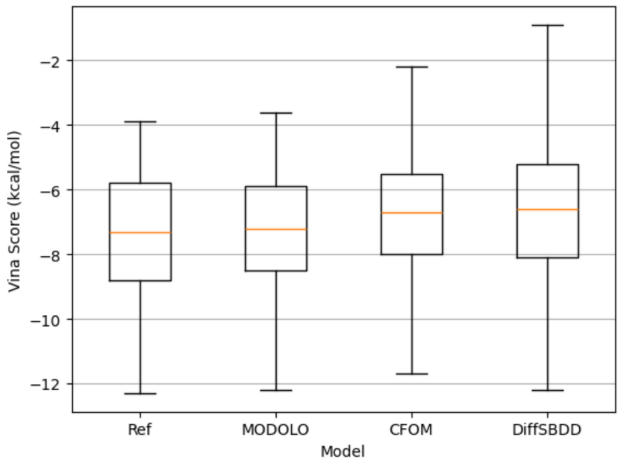

Figure 4: Distribution of Vina binding affinity scores (kcal/mol) for molecules generated by MODOLO and baseline models (CFOM, DiffSBDD) compared to the original docked poses on the CrossDock2020 test set. The box limits indicate the interquartile range (25th-75th percentiles), and the median is marked by the central line.

Fig. 5 presents a representative case study of a molecule optimized by MODOLO. As shown, the generated analogue strictly preserves the original scaffold pose while evolving the fragment to better occupy the binding cavity, resulting in improved geometric complementarity and a lower Vina score compared to the starting compound.

5.2 Ablations

The results of the ablation study are reported in Table 2. As can be seen, the full model achieved the highest and Vina score, which are the main metrics of interest. This confirms the effectiveness of combining all architectural and training components.

The removal of the grouped vector attention led to a higher model complexity, which resulted in a longer training time, but did not improve the performance. The degradation in performance observed in the VAE ablation highlights the inherent multimodality of the lead optimization task. The relationship between a specific 3D protein-ligand interface and a valid chemical fragment is one-to-many, multiple distinct chemical groups can often satisfy the same geometric and physical constraints. By removing the latent vector z, the model loses the capacity to represent this distribution of possibilities. Instead of learning a manifold of valid fragments, the deterministic model is forced to approximate a single output for given geometric inputs, likely resulting in "averaging" effects where the generated fragments fail to capture the necessary chemical diversity or validity inherent in the training data. The most apparent degradation in performance was observed during the masking of the α-carbon atoms, where the loss of the protein residue structure severely impaired the

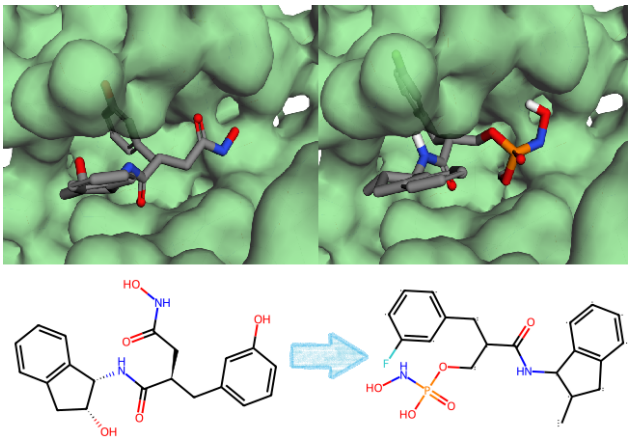

Figure 5: On the left, the original molecule and its binding pose (with a score of -9.4 kcal/mol). On the right, the generated molecule and its binding pose (with a score of -9.9 kcal/mol). The target protein is Q9UNA0 (Human).

Table 2: metrics of MODOLO and ablations on the CrossDocked dataset.

Metric	MODOLO	w/o groups	w/o VAE	w/o α-carbons
Success ↑	**0.36** ±0.23	0.29 ±0.21	<u>0.30</u> ±0.21	0.27 ±0.19
Vina ↓	**-7.31** ±3.53	<u>-7.10</u> ±3.27	-7.01 ±3.91	-6.91 ±3.66
Diversity ↑	0.40 ±0.06	<u>0.52</u> ±0.03	**0.54** ±0.03	0.49 ±0.03
SA ↓	3.62 ±1.82	3.50 ±1.91	<u>3.45</u> ±1.96	**3.38** ±1.89
QED ↑	0.50 ±0.05	<u>0.55</u> ±0.04	0.54 ±0.02	**0.56** ±0.03
Logp	1.33 ±8.20	0.71 ±7.99	1.42 ±6.63	1.35 ±8.34

The bold values indicate the best performance results, the underlined values indicate the second best performance results.

model's ability to interpret the global topology of the binding pocket. Consequently, this led to a decrease in Vina scores, indicating that the coarse-grained spatial constraints provided by the α-carbons are essential for guiding the generation of high-affinity ligands.

The introduction of docking pose noise revealed a disconnect between predicted binding affinity and structural quality in Table 3. While the Vina scores remained remarkably stable across all noise levels, suggesting the model can still locate energetic minima even when displaced, the physicochemical profile of the generated molecules degraded. We observed a notable drift from the reference active structure, as noise increased, the generated fragments became less similar to the masked, ground truth fragments, indicating that the model lost the specific geometric guidance necessary to reproduce them. This structural deviation was accompanied by a slight deterioration in both synthetic accessibility (SA) and QED. Most significantly, we observed a steady decline in lipophilicity, with LogP dropping from 1.33 to 1.05. This downward trend pushes some of the generated compounds out of the optimal lipophilicity range for drugs $1 < LogP < 3$ (Waring, 2010; Landry & Crawford, 2019), suggesting that as the generation centroid shifts from the hydrophobic pocket into the solvent, the model compensates by producing overly polar structures that fall outside the optimal profile for our target.

6 Conclusion

In this work, we introduced MODOLO, a novel architecture for molecular lead optimization that leverages both the 3D structure of protein binding pockets and the SMILES representation of molecules. By decomposing molecules into scaffold and fragments, and utilizing a graph-based encoder alongside a transformer decoder, our approach enables the generation of structurally similar molecules with enhanced protein inter-

Table 3: performance metrics of MODOLO evaluated under varying levels of docking pose noise.

Metric	Origin Pose	4Å	8Å
Success ↑	**0.36** ±0.23	0.32 ±0.22	0.32 ±0.22
Vina ↓	-7.31 ±3.53	-7.31 ±3.51	-7.31 ±3.35
Diversity ↑	0.40 ±0.06	<u>0.44</u> ±0.05	**0.46** ±0.04
SA ↓	**3.62** ±1.82	<u>3.68</u> ±1.92	3.70 ±1.98
QED ↑	**0.50** ±0.05	0.48 ±0.05	0.48 ±0.05
LogP	1.33 ±8.20	1.16 ±8.39	1.05 ±8.17

The Origin Pose column represents the original docking pose. The 4Å and 8Å columns represent docking poses perturbed by random translations (4 Å and 8 Å, respectively) and random rotations.

actions. The integration of grouped vector attention, the heterogeneous graph structure, and the variational autoencoder enabled the model to learn a more compact and disentangled representation of the molecules, which improved the model's ability to generate structurally similar molecules with enhanced protein interactions. Our experiments on the CrossDock2020 benchmark demonstrated that MODOLO consistently outperforms state-of-the-art baselines. Looking forward, future research could explore extending the architecture to multi-objective optimization, incorporating additional chemical properties such as toxicity or solubility, and adapting the model for scaffold hopping or de novo drug design. Further investigation into integrating active learning strategies to maximize performance with minimal labeled data is also promising.

References

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7, 05 2015.

Guy Barshatski and Kira Radinsky. Unpaired generative molecule-to-molecule translation for lead optimization. In *Proceedings of the 27th ACM SIGKDD Conference on knowledge discovery & data mining*, pp. 2554–2564, 2021.

Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 37(1-2):1700123, 2018.

Hans-Joachim Böhm, Alexander Flohr, and Martin Stahl. Scaffold hopping. *Drug discovery today: Technologies*, 1(3):217–224, 2004.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

Ziqi Chen, Bo Peng, Tianhua Zhai, Daniel Adu-Ampratwum, and Xia Ning. Generating 3d small binding molecules using shape-conditioned diffusion models with guidance. *Nature Machine Intelligence*, pp. 1–13, 2025.

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

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Tianfan Fu, Cao Xiao, and Jimeng Sun. Core: Automatic molecule optimization using copy & refine strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 638–645, 2020.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi Jaakkola, and Andreas Krause. Independent se (3)-equivariant models for end-to-end rigid protein docking. *arXiv preprint arXiv:2111.07786*, 2021.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Harrison Green, David R Koes, and Jacob D Durrant. Deepfrag: a deep convolutional neural network for fragment-based lead optimization. *Chemical Science*, 12(23):8036–8047, 2021.

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. *arXiv preprint arXiv:2303.03543*, 2023.

Shahar Harel and Kira Radinsky. Accelerating prototype-based drug discovery using conditional diversity networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 331–339, 2018.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

Lei Huang, Tingyang Xu, Yang Yu, Peilin Zhao, Xingjian Chen, Jing Han, Zhi Xie, Hailong Li, Wenge Zhong, Ka-Chun Wong, et al. A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications*, 15(1):2657, 2024.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecule optimization. In *International Conference on Learning Representations*, 2019.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pp. 4839–4848. PMLR, 2020.

Natan Kaminsky, Uriel Singer, and Kira Radinsky. Cfom: lead optimization for drug discovery with limited data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1056–1066, 2023.

Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

G Landrum. Rdkit: Open-source cheminformatics software, 2016.

Matthew L Landry and James J Crawford. Log d contributions of substituents commonly used in medicinal chemistry. *ACS medicinal chemistry letters*, 11(1):72–76, 2019.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

Haitao Lin, Yufei Huang, Odin Zhang, Siqi Ma, Meng Liu, Xuanjing Li, Lirong Wu, Jishui Wang, Tingjun Hou, and Stan Z Li. Diffbp: Generative diffusion of 3d molecules for target protein binding. *Chemical Science*, 16(3):1417–1431, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Seokhyun Moon, Sang-Yeon Hwang, Jaechang Lim, and Woo Youn Kim. Pignet2: a versatile deep learning-based protein–ligand interaction prediction model for binding affinity scoring and virtual screening. *Digital Discovery*, 3(2):287–299, 2024.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9:1–14, 2017.

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*, pp. 17644–17655. PMLR, 2022.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

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.

Ansgar Schuffenhauer, Peter Ertl, Silvio Roggo, Stefan Wetzel, Marcus A Koch, and Herbert Waldmann. The scaffold tree- visualization of the scaffold universe by hierarchical scaffold classification. *Journal of chemical information and modeling*, 47(1):47–58, 2007.

Oliver B Scott and AW Edith Chan. Scaffoldgraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics*, 36(12):3930–3931, 2020.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pp. 20503–20521. PMLR, 2022.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2): 455–461, 2010.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zechen Wang, Liangzhen Zheng, Yang Liu, Yuanyuan Qu, Yong-Qiang Li, Mingwen Zhao, Yuguang Mu, and Weifeng Li. Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in chemistry*, 9:753002, 2021.

Michael J Waring. Lipophilicity in drug discovery. *Expert opinion on drug discovery*, 5(3):235–248, 2010.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35: 33330–33342, 2022.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Shuangjia Zheng, Zengrong Lei, Haitao Ai, Hongming Chen, Daiguo Deng, and Yuedong Yang. Deep scaffold hopping with multimodal transformer neural networks. *Journal of cheminformatics*, 13:1–15, 2021.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.

A Construction of Input Graph

Molecular docking structures are represented as heterogeneous geometric graphs with nodes representing ligand (heavy) atoms, receptor residues (located in the position of the α-carbon atom), and receptor (heavy) atoms. To build the radius graph, we connect nodes using cutoffs that are dependent on the types of nodes they are connecting:

1. Ligand atoms-ligand atoms, receptor atoms-receptor atoms, and ligand atoms-receptor atoms interactions all use a cutoff of 5 Å, standard practice for atomic interactions. For the ligand atoms-ligand atoms interactions we also preserve the covalent bonds as separate edges with some initial embedding representing the bond type (single, double, triple and aromatic). For receptor atoms-receptor atoms interactions, we limit at 8 the maximum number of neighbors of each atom.
2. Receptor residues-receptor residues use a cutoff of 15 Å with 24 as the maximum number of neighbors for each residue.
3. Receptor residues-ligand atoms use a cutoff of 10 Å.
4. Receptor residues are connected to the receptor atoms that form the corresponding amino-acid.

Receptor residue features include the amino acid identity and a language model embedding derived from ESM2 (Lin et al., 2023). Ligand atom features include atomic number, chirality, degree, formal charge, implicit valence, number of attached hydrogens, number of radical electrons, hybridization state, aromaticity, number of rings, and six binary indicators for membership in rings of size 3 through 8. The edges are encoded based on the distance between the nodes, and the type of the bond (in ligand-ligand edges).

Moreover, during generation, for each masked fragment, a "hole" is created in the graph at the position of the scaffold's atom, to which the fragment is connected. The node is connected to all the neighbors of atoms in the masked fragment