MULTI-POSE STRUCTURAL EVIDENCE FOR LLM AGENT MOLECULAR OPTIMIZATION IN STRUCTURE-BASED DRUG DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure based molecular optimization aims to refine candidate ligands with improved binding scores and molecular quality. Existing methods typically lack explicit modeling of interaction consistency across docking poses and verification of interaction changes after optimization. This paper proposes MEDD, an LLM agent framework for structural evidence guided molecular optimization. It aggregates protein ligand interactions across multiple docking poses into a unified structural evidence representation. This representation is then used to identify stable interactions and pose sensitive regions based on cross pose consistency. The LLM then formulates testable interaction hypotheses and translates them into constrained local molecular modifications. The resulting candidates are docked again to verify whether the hypothesized interaction changes are realized while key interactions are preserved. Experiments on CrossDocked2020 show that MEDD achieves improved docking scores while maintaining QED, SA, and MRR, and achieves a higher key interaction preservation rate and a higher proportion of structurally supported optimization hypotheses. These results support the use of explicit structural evidence for reliable and verifiable LLM guided molecular optimization.

1 INTRODUCTION

Structure-based drug design (SBDD) aims to generate and optimize molecules by exploiting the three-dimensional structures of protein binding pockets, and has become an important paradigm in computational drug discovery. Recent advances in autoregressive and diffusion-based molecular generative models, such as Pocket2Mol, TargetDiff, and MolCRAFT, have significantly improved protein-conditioned molecule generation by modeling geometric compatibility between ligands and target pockets (Peng et al., 2022; Guan et al., 2023a; Qu et al., 2024). Meanwhile, large language models (LLMs) have been explored for molecular design, property reasoning, and optimization, providing a complementary approach for incorporating chemical reasoning into automated SBDD workflows (Gao et al., 2025; Averly et al., 2025).

However, current optimization strategies still rely heavily on a single predicted docking pose or a final docking score as the primary guidance signal. Protein–ligand docking often produces multiple plausible poses with distinct interaction patterns, where hydrogen bonds, hydrophobic contacts, and aromatic interactions may vary across configurations. Interactions consistently observed across multiple poses provide stronger structural support, whereas pose-dependent contacts may represent unstable signals caused by docking uncertainty. Therefore, extracting interaction information from only one pose can introduce unreliable guidance for molecular modification. Moreover, an improved docking score alone does not demonstrate that the targeted interaction has formed. For example, optimizing a molecule to recover a hydrogen bond or strengthen a hydrophobic interaction requires direct structural verification of the corresponding interaction after modification, rather than relying only on a scalar objective value.

To address these limitations, we propose MEDD, a framework for structure-based molecular optimization driven by multi-pose interaction evidence and post-editing structural verification. Given an initial ligand and a target protein pocket, MEDD first generates multiple docking poses and aggregates protein–ligand interaction patterns across these poses. Interactions repeatedly observed across poses

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

are identified as stable contacts, while pose-dependent interactions and local structural regions with optimization potential are represented as variable opportunities. This evidence representation provides the LLM agent with explicit structural information beyond a single docking configuration. Based on this evidence, the agent formulates an interaction hypothesis and translates it into a constrained local molecular modification. After editing, MEDD performs redocking and interaction analysis under the same protocol to determine whether the targeted interaction change is supported by the updated structural evidence. Thus, MEDD separates molecular optimization into evidence extraction, hypothesis-guided editing, and post-editing structural verification, where the LLM performs reasoning and planning while computational tools provide measurable structural evidence.

We evaluate MEDD on the CrossDocked2020 benchmark through comparisons with representative SBDD baselines, component ablations, and protein–ligand interaction analyses. MEDD achieves competitive predicted binding performance while consistently improving the MolCRAFT initial candidates across multiple molecular-quality metrics. Specifically, QED increases from 0.508 to 0.5506, normalized SA from 0.684 to 0.7361, MRR from 65.65% to 86.95%, and the success ratio from 11.82% to 29.97%. Further distributional, ablation, and structural analyses show that these improvements are obtained through relatively localized molecular modifications and depend on complementary contributions from multi-pose evidence, hypothesis-guided editing, and verification feedback. These results support multi-pose structural evidence and post-editing verification as useful signals for controlled and interpretable molecular refinement.

Our main contributions are:

- **Multi-pose structural evidence representation.** We introduce a representation that aggregates protein–ligand interactions across multiple docking poses and distinguishes stable contacts from variable opportunities for reliable molecular optimization.
- **Evidence-grounded molecular optimization with structural verification.** We develop an LLM agent framework that converts structural evidence into explicit interaction hypotheses and constrained molecular edits, followed by post-editing verification using updated docking and interaction evidence.
- **Evidence-guided refinement with controlled molecular changes.** We demonstrate that combining structural evidence and mechanism verification improves molecular quality while maintaining controlled molecular editing on CrossDocked2020.

2 Related Work

2.1 Three-Dimensional Molecular Generation for SBDD

Existing three-dimensional structure-based drug design methods mainly focus on generating molecules conditioned on protein binding pockets. Autoregressive approaches such as GraphBP and Pocket2Mol construct molecules sequentially while explicitly modeling local geometric relationships between ligands and protein environments (Liu et al., 2022; Peng et al., 2022). Diffusion-based methods, including TargetDiff and DiffSBDD, jointly model molecular structures and three-dimensional coordinates through equivariant diffusion processes (Guan et al., 2023a; Schneuing et al., 2024). More recently, MolCRAFT employs a Bayesian Flow Network to model heterogeneous molecular variables within a unified framework and achieves strong performance in three-dimensional molecular generation (Qu et al., 2024). These methods have substantially advanced protein-conditioned de novo generation, whereas post-generation refinement based on explicit protein–ligand structural evidence has received comparatively less attention.

2.2 LLMs for Molecular Design and Structural Reasoning

Large language models have recently been incorporated into molecular design and optimization workflows as reasoning and planning components. CIDD combines three-dimensional molecular generation with LLM-based modules for interaction analysis, molecular design, reflection, and candidate selection, improving molecular drug-likeness while maintaining competitive binding properties (Gao et al., 2025). LIDDiA further uses an LLM to coordinate molecular generation, optimization, and screening tools within an automated drug discovery workflow (Averly et al., 2025). In parallel, protein-ligand interaction representations have long been used to characterize local

molecular recognition patterns, including hydrogen bonds, hydrophobic contacts, ionic interactions, and aromatic interactions, for virtual screening, pose comparison, and affinity analysis (Brewerton, 2008; Wang et al., 2021). SPLIF further represents local structural environments around interacting atoms and compares interaction patterns between docking poses (Da & Kireev, 2014). These studies establish both LLM-based molecular reasoning and interaction-level structural representation, while MEDD further uses cross-pose interaction consistency to estimate the reliability of structural evidence before molecular editing.

2.3 Docking-Guided Optimization and Structural Verification

Molecular docking is widely used to predict protein-ligand binding poses and estimate their compatibility with target pockets. AutoDock Vina generates candidate poses and ranks them using an empirical scoring function, making docking scores common objectives in virtual screening and molecular optimization (Trott & Olson, 2010). However, docking scores and local interaction patterns characterize different aspects of a predicted protein-ligand complex, and different plausible poses may exhibit distinct hydrogen bonds, hydrophobic contacts, or aromatic interactions. Recent structural assessment studies further show that favorable docking scores do not always correspond to physically plausible binding poses, motivating evaluation beyond a single scalar score (Buttenschoen et al., 2024). MEDD builds on this perspective by combining multi-pose interaction evidence before editing with post-editing structural verification, allowing molecular refinement to be evaluated not only by endpoint properties but also by whether the targeted interaction changes are supported by the resulting docking evidence.

3 Method

3.1 Overview

Given a target protein pocket P and an initial ligand x_0, MEDD aims to iteratively refine molecular structures by exploiting reliable protein–ligand interaction evidence and verifying the structural consequences of each modification. Different from optimization strategies guided by a single docking pose or endpoint docking score, MEDD constructs structural evidence from multiple docking poses, derives interaction hypotheses from this evidence, and performs chemistry-aware molecular editing.

As illustrated in Figure 1, MEDD follows an evidence-guided molecular refinement pipeline consisting of four components: (1) multi-pose structural evidence construction, (2) evidence-grounded LLM reasoning and design, (3) chemistry-aware molecular editing, and (4) post-editing structural verification and final selection.

The overall optimization process is formulated as:

$$\begin{aligned} E_i &= \mathrm{Evidence}(x_i, P), \\ (H_i, A_i) &= \mathrm{Reason}(x_i, E_i), \\ x_{i+1} &= \mathrm{Edit}(x_i, A_i), \end{aligned} \tag{1}$$

where E_i denotes the structural evidence associated with candidate x_i, H_i denotes the interaction hypothesis generated from the evidence, and A_i denotes the corresponding molecular editing plan. After modification, the updated candidate is evaluated through structural verification:

$$V_i = \mathrm{Verify}(H_i, E_i, E_{i+1}), \tag{2}$$

where V_i summarizes whether the proposed structural change is supported by the updated protein–ligand interaction evidence.

MEDD separates language reasoning from molecular generation. The LLM performs evidence-grounded interaction reasoning and modification planning, while chemical tools execute molecular transformations and structural tools provide docking and interaction evaluation.

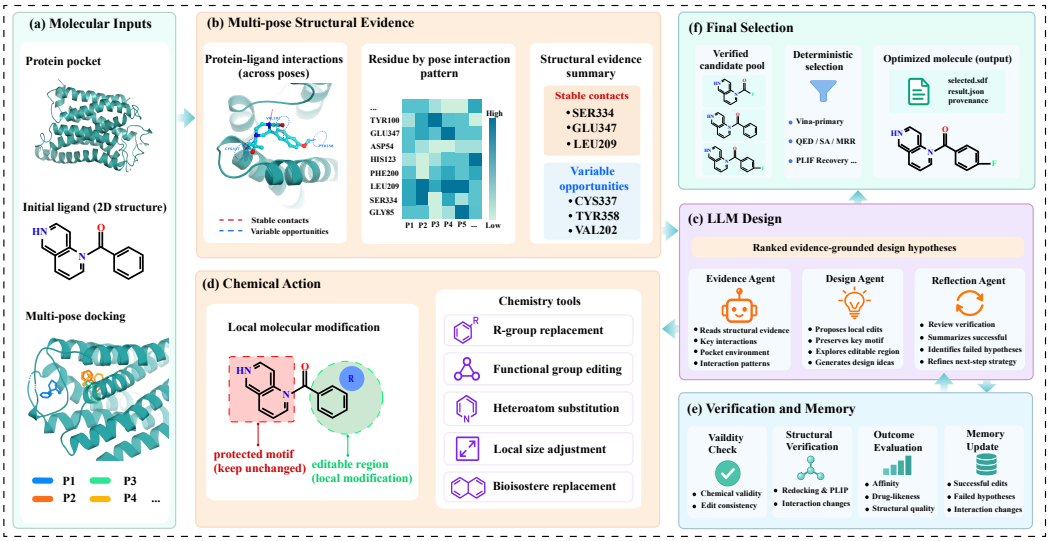


Figure 1: Overview of MEDD. Given an initial ligand and target protein pocket, MEDD first constructs multi-pose structural evidence by aggregating protein–ligand interactions across docking poses. The evidence is organized into stable contacts and variable opportunities, which guide LLM-based hypothesis generation and molecular editing. After chemical modification, MEDD performs structural verification and selects the final candidate based on docking performance, drug-like properties, and interaction evidence.

3.2 Multi-Pose Structural Evidence Construction

A single docking pose represents only one possible configuration of a protein–ligand complex. To reduce the dependence on a specific predicted pose, MEDD aggregates interaction information from multiple plausible docking poses.

For candidate molecule x_i and protein pocket P, docking produces a set of K candidate poses (Trott & Olson, 2010):

$$\mathcal{Z}_i = \mathrm{Dock}(x_i, P) = \{z_i^1, z_i^2, \ldots, z_i^K\}. \tag{3}$$

For each pose z_i^k, we extract its protein–ligand interaction profile (Salentin et al., 2015):

$$I_i^k = \mathrm{Interact}(z_i^k, P), \quad k = 1, \ldots, K. \tag{4}$$

Each interaction is represented as:

$$c = (r_l, r_p, t), \tag{5}$$

where r_l denotes the ligand atom or functional group, r_p denotes the interacting protein residue, and t denotes the interaction type, including hydrogen bonds, hydrophobic contacts, and aromatic interactions.

Instead of selecting a single docking pose, MEDD measures the consistency of each interaction across poses:

$$\rho_i(c) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[c \in I_i^k], \tag{6}$$

where $\rho_i(c)$ indicates the frequency of observing interaction c among all docking poses. Higher values indicate interactions that are less sensitive to docking uncertainty.

The structural evidence is then represented as:

$$E_i = \mathrm{Aggregate}(I_i^1, I_i^2, \ldots, I_i^K), \tag{7}$$

which contains interaction identities, cross-pose consistency statistics, and local structural information.

3.3 Stable Contacts and Variable Opportunities

Based on cross-pose statistics, MEDD organizes structural evidence into two complementary components:

$$(\mathcal{S}_i, \mathcal{V}_i) = \mathrm{Organize}(E_i), \tag{8}$$

where $\mathcal{S}_i$ denotes stable contacts and $\mathcal{V}_i$ denotes variable opportunities, defined as

$$\mathcal{S}_i = \{c \mid \rho_i(c) \geq \tau_s\}, \qquad \mathcal{V}_i = \{c \mid \tau_v < \rho_i(c) < \tau_s\}. \tag{9}$$

$\mathcal{S}_i$ captures interactions repeatedly observed across multiple docking poses, representing reliable binding features that should be preserved during molecular editing. In contrast, $\mathcal{V}_i$ captures local interactions or spatial regions with lower cross-pose consistency but potential for optimization. Rather than treating $\mathcal{V}_i$ as noise, MEDD uses it to identify potential editing regions where molecular modifications can improve pocket complementarity while maintaining the stable interactions in $\mathcal{S}_i$.

3.4 Evidence-Grounded LLM Design and Chemical Action

Given stable contacts $\mathcal{S}_i$ and variable opportunities $\mathcal{V}_i$, the LLM agent performs structured reasoning and generates an explicit interaction hypothesis together with a molecular editing plan:

$$(H_i, A_i) = \mathrm{LLM}(x_i, \mathcal{S}_i, \mathcal{V}_i, F_i), \tag{10}$$

where F_i denotes previous optimization feedback.

The hypothesis describes the expected structural change, while the editing plan specifies the corresponding chemical operation. Importantly, the LLM does not directly generate molecular structures. Instead, it interprets structural evidence and plans modifications, while chemistry tools execute the actual molecular transformation:

$$x_{i+1} = \mathrm{Edit}(x_i, A_i). \tag{11}$$

The editing operations include local fragment replacement, functional group modification, heteroatom substitution, and local structural adjustment.

3.5 Structural Verification and Final Selection

After molecular editing, MEDD reconstructs the structural evidence of the new candidate:

$$E_{i+1} = \mathrm{Evidence}(x_{i+1}, P). \tag{12}$$

The verification module compares the original evidence, updated evidence, and the proposed hypothesis:

$$V_i = \mathrm{Verify}(H_i, E_i, E_{i+1}). \tag{13}$$

The verification examines whether the targeted interaction change appears after editing, whether the updated docking poses support the proposed structural hypothesis, and whether important stable contacts are preserved.

Table 1: Test results on CrossDocked2020. We report Vina docking score, QED, normalized SA, molecular reasonability ratio (MRR), success ratio, Lipinski score, and average molecular weight (MW). For CIDD and MEDD, results are computed on the same set of 981 valid tasks. Best results are shown in bold and second-best results are underlined.

Category	Method	Vina ↓	QED ↑	SA ↑	MRR ↑	Success ↑	Lipinski ↑	MW
AR-based	AR	-6.737	0.507	0.635	56.67%	3.28%	4.76	247.50
AR-based	Pocket2Mol	-7.246	<u>0.573</u>	<u>0.758</u>	67.88%	14.60%	**4.83**	234.30
Diffusion-based	TargetDiff	-7.452	0.474	0.579	37.81%	3.04%	4.59	346.24
Diff + Inter-Guide	IPDiff	-7.745	0.511	0.627	29.83%	5.31%	4.53	328.34
Diff + Multi-Guide	TAGMol	-7.563	0.563	0.583	37.31%	3.23%	4.73	325.50
Fragment-based	DrugGPS	-7.396	0.463	0.622	54.80%	7.17%	4.45	329.88
BFN-based	MolCRAFT	-7.162	0.508	0.684	65.65%	11.82%	4.56	325.43
3DSBDD + LLM	CIDD	**-7.9701**	**0.6348**	**0.7676**	**95.21%**	**36.09%**	<u>4.80</u>	311.49
3DSBDD + LLM	MEDD	<u>-7.8956</u>	0.5506	0.7361	<u>86.95%</u>	<u>29.97%</u>	4.63	322.63

Finally, candidate molecules are ranked according to docking performance, drug-likeness properties, and structural evidence quality. The selected candidate is returned as the optimized molecular output.

4 Experiments

4.1 Experiment Settings

Dataset. We evaluate MEDD on CrossDocked2020 (Francoeur et al., 2020), following the commonly used test setting for three-dimensional structure-based drug design. The test set contains 100 protein binding pockets. MolCRAFT-generated molecules (Qu et al., 2024) are used as the initial candidates for molecular refinement. For the direct comparison among MolCRAFT, CIDD, and MEDD, we use the same set of 981 valid molecular tasks, allowing the refinement results to be compared under a matched evaluation setting.

Metrics. We evaluate molecular quality using AutoDock Vina score (Trott & Olson, 2010) for predicted binding affinity, QED (Bickerton et al., 2012) for drug-likeness, normalized SA (Ertl & Schuffenhauer, 2009) for synthetic accessibility, and molecular reasonability ratio (MRR) for structural plausibility. We additionally report the success ratio under the unified evaluation criteria, together with the Lipinski score and average molecular weight (MW). For refinement analysis, we further examine Tanimoto similarity to the initial molecule and changes in heavy-atom count to characterize the magnitude of molecular modifications.

Baseline Models. We compare MEDD with representative three-dimensional SBDD approaches spanning different generative paradigms, including the autoregressive models AR and Pocket2Mol (Luo et al., 2021; Peng et al., 2022), diffusion-based TargetDiff, IPDiff, DecompDiff, and TAGMol (Guan et al., 2023a; Huang et al., 2024; Guan et al., 2023b; Dorna et al., 2024), the fragment-based DrugGPS (Zhang & Liu, 2023), and the Bayesian flow network MolCRAFT (Qu et al., 2024). We additionally compare with CIDD (Gao et al., 2025), an LLM-assisted SBDD framework that refines molecules generated by a three-dimensional generative model.

MEDD Settings. MEDD takes a MolCRAFT-generated molecule and its target protein pocket as input. For each refinement round, multiple docking poses are used to construct cross-pose structural evidence, which is organized into stable contacts and variable opportunities. The LLM then generates an explicit interaction hypothesis and a localized modification plan. The edited candidate is redocked under the same protocol and its updated protein–ligand interactions are reanalyzed for structural verification. Detailed docking parameters and implementation settings are provided in Appendix A.

4.2 General Results

Table 1 summarizes the overall results on CrossDocked2020. Compared with MolCRAFT, MEDD improves Vina from -7.162 to -7.896, QED from 0.508 to 0.551, normalized SA from 0.684 to

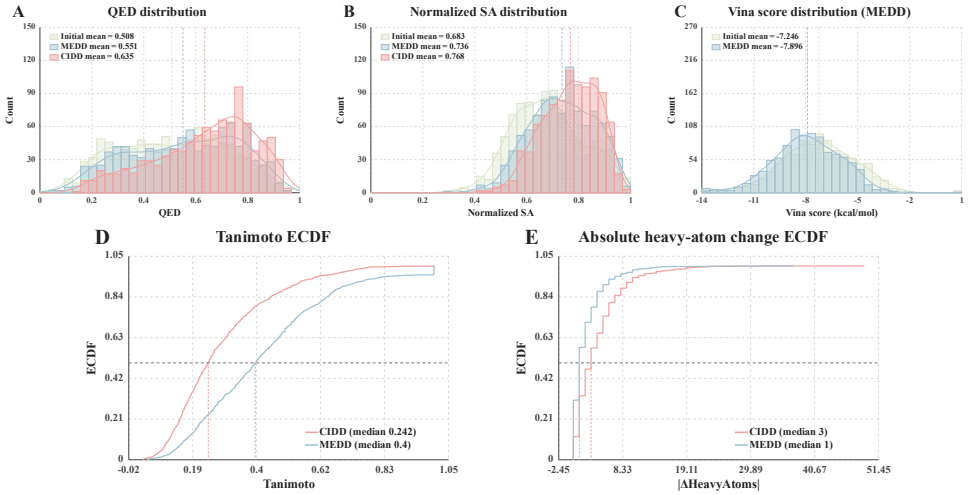


Figure 2: Distribution and editing-behavior analysis. (A–B) QED and normalized SA distributions of the initial molecules, MEDD, and CIDD. (C) Vina score distributions before and after MEDD refinement. (D–E) ECDFs of Tanimoto similarity and absolute heavy-atom changes relative to the initial molecules.

0.736, MRR from 65.65% to 86.95%, and success ratio from 11.82% to 29.97%, while maintaining a similar molecular weight. These improvements indicate that MEDD enhances both predicted binding performance and molecular quality through post-generation refinement.

Among the compared methods, CIDD achieves the strongest endpoint performance on Vina, QED, SA, MRR, and success ratio, while MEDD obtains the second-best Vina, MRR, and success ratio. In particular, MEDD reaches an MRR of 86.95% and a success ratio of 29.97%, while maintaining competitive QED and SA. These results motivate the following analyses of whether the gains are achieved through controlled molecular modifications and how the individual components of MEDD contribute to the refinement process.

4.3 Controlled Molecular Refinement Analysis

Figure 2 further analyzes the refinement behavior beyond aggregate metrics. MEDD shifts both QED and normalized SA toward more favorable regions, increasing their mean values from 0.508 to 0.551 and from 0.683 to 0.736, respectively (Figure 2A–B). The Vina distribution is also shifted toward lower scores, with the mean improving from -7.246 to -7.896 (Figure 2C). On the paired set of 981 molecules, 757 (77.2%) achieve improved Vina scores after refinement, indicating that the gain is broadly distributed across the evaluated candidates.

Figures 2D–E characterize the magnitude of structural modification. MEDD retains a median Tanimoto similarity of 0.400 to the initial molecules, compared with 0.242 for CIDD, while the median absolute heavy-atom change is **1** for MEDD and 3 for CIDD. These quantities are not optimization objectives, but indicate that MEDD tends to obtain its improvements through more localized molecular modifications.

Overall, MEDD improves molecular quality and predicted binding performance while preserving a larger fraction of the initial molecular structure. This behavior is consistent with its evidence-guided refinement design, where stable contacts constrain preserved interactions and variable opportunities identify regions for local modification.

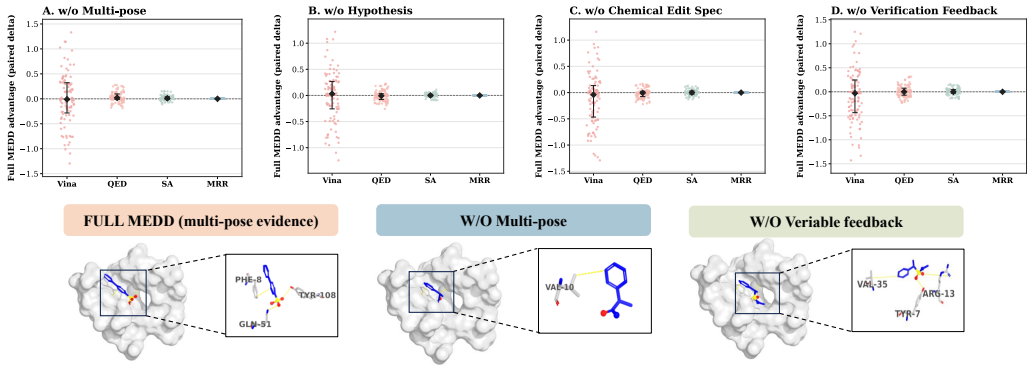


Figure 3: Component ablation of MEDD. Top: paired differences between Full MEDD and variants without multi-pose evidence, hypothesis generation, chemical edit specification, or verification feedback. Bottom: representative structural examples for Full MEDD and selected ablated variants.

4.4 Ablation and Analysis

We ablate four components of MEDD: multi-pose evidence, explicit hypothesis generation, chemical edit specification, and verification feedback. Figure 3 reports paired metric differences together with representative structural examples.

As shown in Figure 3, the ablations affect different aspects of refinement rather than uniformly degrading every endpoint metric. Removing multi-pose evidence produces the clearest loss in binding performance, with mean Vina changing from -7.931 to -7.812. Removing hypothesis generation or verification feedback causes smaller shifts in individual property means, but reduces the success ratio from 28.0% to 14.5% and 14.0%, respectively. This suggests that these components primarily improve the consistency of satisfying multiple optimization criteria rather than independently maximizing a single property.

The chemical edit ablation shows a different behavior: endpoint properties remain relatively competitive, while the mean absolute heavy-atom change increases from 1.86 to 1.99, indicating less localized modification. The structural examples in Figure 3 provide a qualitative view of the same effect: removing multi-pose evidence or verification feedback changes the local protein–ligand interaction pattern relative to Full MEDD. Together, these results support complementary roles for the four components: multi-pose evidence provides structural context, hypotheses define explicit editing objectives, edit specifications constrain modification locality, and verification feedback closes the refinement loop. Full numerical results are reported in Appendix D.

4.5 Structural Interpretability Analysis

Figure 4 shows how MEDD connects cross-pose structural evidence with molecular refinement. For the initial molecule, the four docking poses exhibit both shared and pose-dependent interaction patterns. Residues including LEU209, ASN351, SER334, and GLU347 recur across multiple interaction maps, providing more reliable structural constraints than contacts observed in only a single pose. MEDD uses these recurring patterns as evidence for preserving favorable interactions while directing molecular edits toward locally variable regions.

After refinement, the candidate is redocked using the same protocol and its interaction patterns are re-evaluated. The refined poses retain recurrent contacts while exhibiting altered local geometries, including a 2.5 Å contact with ASN210 in pose 3 and contacts around SER334 (2.2 Å), GLU347 (2.1 Å), and LEU209 (2.6 Å) in pose 4. Rather than treating any single pose or a lower docking score as sufficient evidence, MEDD evaluates these post-editing changes across the pose ensemble. The case therefore provides a direct, traceable connection between the structural evidence used for editing and the interaction patterns observed after refinement.

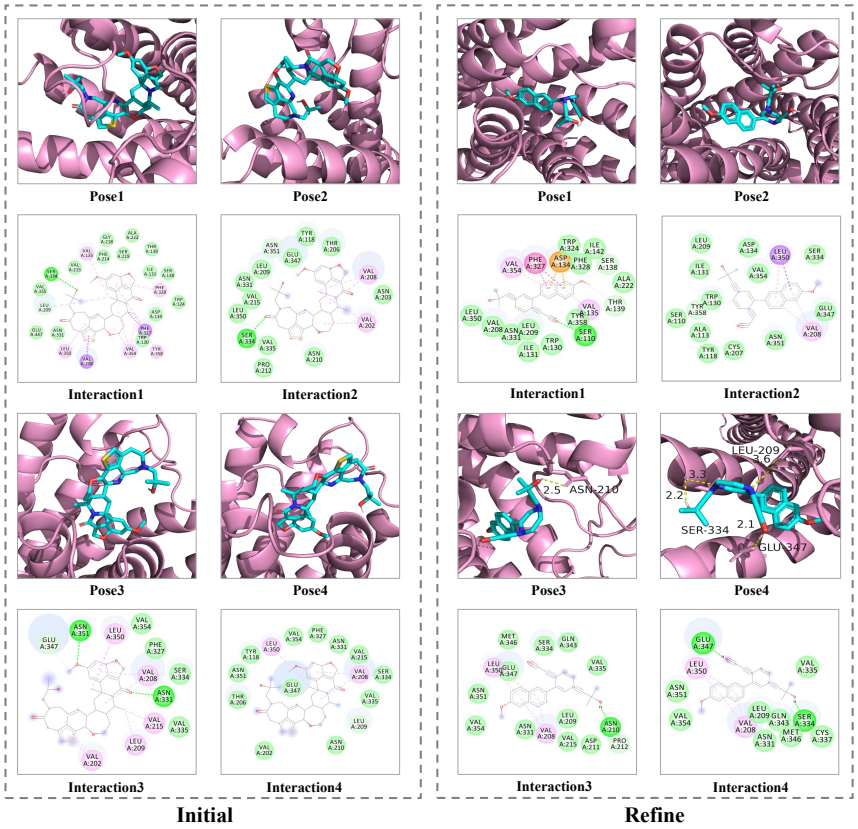


Figure 4: Representative structural analysis on the 6BQG pocket. Four docking poses and their interaction maps are shown for the initial and refined molecules. Recurrent interactions across poses provide structural evidence for refinement, while the refined poses reveal the corresponding post-editing interaction changes.

5 Limitations

MEDD remains dependent on the quality of both structural evidence and LLM-based reasoning. Although multi-pose aggregation reduces reliance on a single docking configuration, uncertainty in docking and interaction analysis can still affect subsequent hypotheses and edits. In addition, refinement quality depends on how effectively the LLM interprets structural evidence and translates it into valid local modifications. Improving uncertainty-aware evidence aggregation and verification remains a natural direction for making the refinement loop more robust.

6 Conclusion

We presented MEDD, an evidence-guided framework for structure-based molecular optimization. MEDD combines multi-pose structural evidence, LLM-based reasoning, constrained chemical editing, and post-editing verification to reduce reliance on a single docking pose or endpoint score. Experiments on CrossDocked2020 show that MEDD improves molecular reasonability, drug-likeness, and success rate over the MolCRAFT initialization while maintaining competitive binding performance and relatively localized structural changes. These results support structural evidence-guided reasoning and verification as a promising direction for more controlled and interpretable molecular optimization.

AI USE STATEMENT

Generative AI tools were used in this study for literature organization, coding assistance, analysis of experimental results, and language refinement of the manuscript. All AI assisted code, analyses, and written content were reviewed and verified by the authors against the actual implementation, execution logs, and experimental results. The structural evaluations, molecular docking, protein ligand interaction analyses, and all quantitative results reported in this paper were obtained through the corresponding computational tools and experimental pipelines rather than directly generated by AI systems. The authors take full responsibility for the methodology, experimental results, interpretation of the findings, and the final content of the manuscript.

REFERENCES

Reza Averly, Frazier N. Baker, Ian A. Watson, and Xia Ning. Liddia: Language-based intelligent drug discovery agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12004–12028. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.603.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243.

Suzanne C. Brewerton. The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery & Development*, 11(3):356–364, 2008.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130–3139, 2024. doi: 10.1039/D3SC04185A.

C. Da and Dmitri Kireev. Structural protein-ligand interaction fingerprints (splif) for structure-based virtual screening: Method and benchmark study. *Journal of Chemical Information and Modeling*, 54(9):2555–2561, 2014. doi: 10.1021/ci500319f.

Vineeth Dorna, D. Subhalingam, Keshav Kolluru, Shreshth Tuli, Mrityunjay Singh, Saurabh Singal, N. M. Anoop Krishnan, and Sayan Ranu. Tagmol: Target-aware gradient-guided molecule generation. *arXiv preprint arXiv:2406.01650*, 2024.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009. doi: 10.1186/1758-2946-1-8.

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. doi: 10.1021/acs.jcim.0c00411.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Bowei He, Haichuan Tan, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. Cidd: Collaborative intelligence for structure-based drug design empowered by llms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15293–15341, 2025. doi: 10.52202/085713-0460.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023a.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. Decompdiff: Diffusion models with decomposed priors for structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11827–11846. PMLR, 2023b.

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

Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3):3–25, 1997. doi: 10.1016/S0169-409X(96)00423-1.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13912–13924. PMLR, 2022.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6229–6239, 2021.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 2022.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Molcraft: Structure-based drug design in continuous parameter space. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41749–41768. PMLR, 2024.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. Plip: Fully automated protein-ligand interaction profiler. *Nucleic Acids Research*, 43(W1):W443–W447, 2015. doi: 10.1093/nar/gkv315.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L. Blundell, Pietro Lio, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4 (12):899–909, 2024. doi: 10.1038/s43588-024-00737-x.

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. doi: 10.1002/jcc.21334.

Debby D. Wang, Moon-Tong Chan, and Hong Yan. Structure-based protein-ligand interaction fingerprints for binding affinity prediction. *Computational and Structural Biotechnology Journal*, 19:6291–6300, 2021. doi: 10.1016/j.csbj.2021.11.018.

Zaixi Zhang and Qi Liu. Learning subpocket prototypes for generalizable structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41382–41398. PMLR, 2023.

A EXPERIMENTAL DETAILS

This section provides the experimental protocol and implementation details of MEDD, including dataset construction, docking settings, optimization procedure, evaluation metrics, failure handling, and reproducibility. Detailed definitions of multi-pose structural evidence, stable contacts, variable opportunities, and post-editing structural verification are provided in Appendix B.

A.1 DATASET AND INITIAL CANDIDATE SELECTION

We evaluate MEDD on CrossDocked2020 Francoeur et al. (2020), following the standard test setting for three-dimensional structure-based drug design with 100 protein binding pockets. For each pocket, we randomly sample 10 molecules from the corresponding MolCRAFT-generated candidate pool Qu et al. (2024), resulting in 1,000 initial refinement tasks.

The initial candidate set is sampled before optimization and remains fixed throughout the subsequent experiments, ensuring that candidate selection is independent of MEDD outcomes. MEDD does not retrain or modify MolCRAFT, but operates as a post-generation refinement framework. For paired comparisons, the same initial molecule, receptor structure, and pocket definition are used under the unified evaluation protocol.

A.2 DOCKING AND STRUCTURAL EVALUATION PROTOCOL

Docking is used both to construct structural evidence and to evaluate edited candidates. The same protocol is applied before and after molecular modification. For each molecule, three initial conformers are generated using RDKit with random seed 42. Molecular docking is performed using AutoDock Vina Trott & Olson (2010) with a $20 \times 20 \times 20$ Å search box centered on the target pocket, exhaustiveness 8, random seed 0, and up to 20 retained docking poses.

Table 2: Docking and conformer-generation settings used throughout the experiments.

Parameter	**Setting**
Docking engine	AutoDock Vina
Docking box size	$20 \times 20 \times 20$ Å
RDKit conformers	3
RDKit random seed	42
Vina exhaustiveness	8
Maximum docking poses	20
Vina random seed	0

Protein–ligand interactions are extracted independently from the retained poses using a common interaction-analysis protocol, including hydrogen bonds, hydrophobic contacts, and aromatic interactions when detected. These pose-level interaction profiles are subsequently aggregated to construct the multi-pose structural evidence used by MEDD. Detailed aggregation and evidence-selection rules are provided in Appendix B.

A.3 MEDD OPTIMIZATION PROTOCOL

Each initial molecule is refined through iterative cycles of structural evidence construction, hypothesis generation, molecular editing, and post-editing verification. Each refinement round consists of:

1. constructing multi-pose structural evidence from retained docking poses;
2. identifying stable contacts and variable opportunities;
3. generating an explicit interaction hypothesis;
4. translating the hypothesis into a constrained local molecular edit;
5. generating and structurally evaluating the edited candidate; and
6. comparing the updated structural evidence with the intended interaction change.

Each molecule undergoes up to five refinement rounds. Verification outcomes are retained as feedback for subsequent rounds, allowing unsupported modification directions to be avoided during later refinement.

A trajectory is considered complete when at least two valid modified candidates are obtained and evaluated under the unified protocol. Runs that execute normally but produce insufficient valid candidates are recorded as scientific failures, whereas failures caused by external execution or infrastructure issues are recorded separately.

A.4 Evaluation Metrics

We evaluate MEDD using molecular-quality metrics together with descriptive measures of editing behavior.

Binding affinity. AutoDock Vina score Trott & Olson (2010) is used as a computational measure of predicted protein–ligand binding compatibility, with lower values preferred.

Drug-likeness and synthetic accessibility. QED Bickerton et al. (2012) measures drug-likeness, while normalized SA Ertl & Schuffenhauer (2009) characterizes synthetic accessibility. Larger values are preferred for both metrics.

Molecular reasonability. MRR follows the evaluation protocol of CIDD Gao et al. (2025) and measures the proportion of molecules satisfying rule-based structural reasonability criteria.

Success ratio. Success ratio measures the proportion of molecules satisfying the predefined joint molecular-quality criteria under the same evaluation protocol used for the compared methods.

Lipinski score and molecular weight. We additionally report Lipinski score Lipinski et al. (1997) and molecular weight (MW) to characterize medicinal-chemistry properties and potential changes in molecular size.

Editing behavior. Tanimoto similarity based on ECFP fingerprints Rogers & Hahn (2010) is computed between the initial and final molecules. We additionally report the absolute change in heavy-atom count. These quantities are descriptive measures of structural modification rather than optimization objectives.

A.5 Failure Handling and Reproducibility

Each optimization trajectory is assigned one of three terminal outcomes:

- **Completed**: at least two valid modified candidates are obtained and evaluated under the unified protocol.
- **Scientific failure**: the optimization pipeline executes normally, but insufficient valid candidates are produced.
- **Infrastructure failure**: the trajectory cannot be completed because of an external execution problem unrelated to molecular quality.

MEDD is implemented as a modular pipeline in which LLM-based reasoning, molecular generation, docking, interaction analysis, and molecular-property evaluation are executed as separate components. Quantitative structural and molecular measurements are computed by the corresponding computational tools rather than by the language model.

Random seeds are fixed whenever supported by the underlying tools, and the initial candidate set is determined before optimization. For each trajectory, the experiment logs preserve the information required to reconstruct the refinement process, including:

- protein pocket and receptor identifiers;
- initial molecular structures;

- docking configurations and retained poses;
- pose-level interaction profiles;
- aggregated multi-pose structural evidence;
- generated interaction hypotheses and molecular edits;
- edited candidate molecules;
- post-editing docking and interaction evidence;
- verification outcomes;
- final molecular-property metrics; and
- terminal task status.

This logging scheme separates computational structural evidence, LLM reasoning decisions, and quantitative molecular evaluation, enabling the complete refinement trajectory to be inspected and reproduced.

B Additional Method Details

This section provides additional implementation details of MEDD, focusing on multi-pose structural evidence construction, the organization of stable contacts and variable opportunities, evidence-grounded hypothesis generation, and post-editing structural verification. The overall framework is described in the main paper and is not repeated here.

B.1 Multi-Pose Structural Evidence Representation

For each candidate molecule, MEDD retains multiple docking poses and extracts a protein–ligand interaction profile from each pose. Interactions across poses are mapped into a shared identity space using ligand atom mapping, protein residue identity, and interaction type, enabling direct comparison of local contacts across alternative binding configurations.

Let c denote an interaction identity and I_i^k the interaction profile extracted from the k-th docking pose of molecule x_i. The occurrence of interaction c in pose k is represented by

$$m_i^k(c) = \mathbb{I}[c \in I_i^k]. \tag{14}$$

Its cross-pose support is then computed as

$$\rho_i(c) = \frac{1}{K} \sum_{k=1}^{K} m_i^k(c), \tag{15}$$

where K denotes the number of retained docking poses.

The support value $\rho_i(c)$ measures how consistently an interaction is reproduced across plausible binding configurations. In contrast to a single-pose interaction map, this representation reduces the influence of pose-specific contacts and provides an explicit estimate of interaction consistency.

Each evidence record contains the interaction identity, interaction type, cross-pose support, and associated local structural information extracted by the interaction-analysis pipeline. These quantities are computed before LLM reasoning; the LLM receives structured evidence rather than inferring protein–ligand interactions directly from the molecular input.

B.2 Stable Contacts and Variable Opportunities

MEDD organizes the resulting structural evidence into stable contacts and variable opportunities.

Stable contacts correspond to interactions that receive consistent support across the retained docking poses. Under the implementation-specific stability criterion, the stable-contact set is written as

$$\mathcal{S}_i = \{c \mid \rho_i(c) \geq \tau_s\}, \tag{16}$$

where τ_s denotes the consistency threshold used during evidence construction. Stable contacts are treated as structural constraints during molecular refinement. They should be interpreted as interactions that are robust within the docking ensemble, rather than experimentally confirmed binding mechanisms.

Variable opportunities capture local regions whose interaction patterns vary across poses and therefore admit controlled modification. Rather than being defined as the simple complement of $\mathcal{S}_i$, they are identified from pose-dependent interaction patterns together with their local structural context. Stable contacts therefore indicate which structural features should preferentially be preserved, whereas variable opportunities indicate where local refinement can be explored.

B.3 Evidence-Grounded Hypothesis Generation

Given structural evidence $E_i = (\mathcal{S}_i, \mathcal{V}_i)$, the LLM generates an explicit interaction hypothesis before molecular editing. We represent the hypothesis as

$$H_i = (G_i, T_i, P_i), \tag{17}$$

where G_i denotes the intended interaction objective, T_i specifies the local modification region, and P_i represents structural constraints that should be preserved.

This intermediate hypothesis separates structural reasoning from molecular generation. Instead of directly optimizing an endpoint score, the agent first specifies an expected structural change and then translates it into a constrained local edit. The hypothesis subsequently provides an explicit target for post-editing verification.

B.4 Post-Editing Structural Verification

After molecular editing, the candidate is redocked using the same protocol as the parent molecule, and its protein–ligand interaction profiles are reconstructed from the resulting poses. Let E_{i+1} denote the updated structural evidence. Verification is represented as

$$R_i = \Psi(H_i, E_{i+1}), \tag{18}$$

where Ψ evaluates whether the structural change specified by H_i is supported by the post-editing docking evidence.

The verification procedure considers three complementary criteria:

1. whether the targeted interaction change is observed after editing;
2. whether the corresponding interaction receives support across multiple docking poses; and
3. whether important stable contacts are preserved.

A favorable docking score alone is therefore not treated as sufficient evidence that the intended structural modification has been achieved. MEDD instead evaluates endpoint molecular properties together with the correspondence between the proposed hypothesis and the updated interaction evidence.

B.5 Overall Optimization Procedure

Algorithm 1 summarizes the complete MEDD refinement procedure. Structural evidence extraction, LLM-based reasoning, molecular editing, and post-editing verification are executed as distinct stages within a unified iterative loop.

Overall, MEDD follows an evidence–hypothesis–edit–verification cycle: multi-pose structural evidence guides hypothesis generation, constrained chemical editing realizes the proposed modification, and redocking with interaction analysis evaluates its structural consequence.

Algorithm 1 MEDD optimization procedure with multi-pose structural evidence and verification

Require: Protein pocket P, initial molecule M, number of docking poses J, refinement budget T

for $t = 1$ to T **do**
  Dock M into P and retain J candidate poses
  Extract protein–ligand interaction profiles from each pose
  Aggregate cross-pose statistics to construct Stable contacts and Variable opportunities
  Evidence Agent analyzes structural evidence and identifies optimization targets
  Hypothesis Agent generates an evidence-grounded interaction hypothesis
  Design Agent converts the hypothesis into a constrained local molecular edit
  Generate candidate molecule M' and perform chemical validity checks
  Redock M' using the same docking protocol
  Compare pre-edit and post-edit interaction evidence
  Reflection Agent summarizes verification results and updates refinement feedback
  if M' satisfies verification criteria **then**
    Accept M' as the updated molecule
  else
    Retain current molecule and record verification feedback
  end if
end for
return Optimized molecule M

C Prompts and Structured Agent Outputs

MEDD uses the language model as an evidence-grounded reasoning component rather than as a source of structural measurements. Docking scores, protein–ligand interactions, and molecular properties are computed by the corresponding computational tools. The LLM receives these structured observations and produces molecular design decisions that can be evaluated by subsequent structural analysis.

The prompting procedure is organized around four operational stages: structural evidence analysis, interaction hypothesis generation, chemical edit specification, and verification feedback. Across these stages, the prompts enforce three general principles: molecular edits should be grounded in the available structural evidence, reliable interactions should be preserved when possible, and each proposed modification should correspond to an explicit structural objective that can be evaluated after redocking.

C.1 Structural Evidence Input

At each refinement round, the LLM receives a structured representation of the current molecular state. The input combines molecular information with multi-pose structural evidence and, when available, feedback from previous refinement rounds. A simplified representation is:

```
Molecule:
    structure
    molecular properties

Multi-pose evidence:
    interaction identity
    interaction type
    cross-pose support
    structural reference

Stable contacts:
    interaction
    supporting evidence

Variable opportunities:
    local region
    pose-dependent interaction pattern
```

```
Previous feedback:
    previous hypothesis
    verification status
    failure reason
    evidence reference
```

The structural evidence is produced before LLM reasoning. The model is therefore asked to interpret explicit interaction observations rather than infer protein–ligand interactions directly from the molecular representation.

C.2 Interaction Hypothesis Generation

Given the structured evidence, the LLM formulates a single local interaction hypothesis for the current refinement round. The hypothesis specifies what structural change should be attempted, where the modification should occur, and which existing interactions should be preserved.

The structured output follows the form:

```
hypothesis:
    intended interaction change

target_region:
    local molecular region to modify

protected_contacts:
    interactions to preserve

expected_change:
    expected post-editing structural effect

evidence_reference:
    supporting structural evidence

rationale:
    evidence-based design rationale
```

This intermediate representation separates structural reasoning from molecular generation. Rather than directly requesting a molecule with a better endpoint score, MEDD first identifies a local structural objective that can subsequently be translated into a chemical modification and evaluated after editing.

C.3 Chemical Edit Specification

The interaction hypothesis is converted into a constrained chemical edit. The LLM specifies the intended transformation, while candidate realization and subsequent quantitative evaluation are performed by the corresponding molecular-generation and computational modules.

A simplified edit specification is:

```
edit_region:
    molecular region to modify

edit_operation:
    requested chemical transformation

chemical_constraints:
    validity and drug-like constraints

protected_structure:
```

```
    scaffold or interactions to preserve

expected_effect:
    intended structural consequence
```

The prompt favors localized and incremental modifications over unrestricted molecular reconstruction. Candidate edits are additionally constrained by medicinal-chemistry and synthetic-feasibility considerations.

C.4 Chemical Design Priors

The editing stage is supplemented with a set of chemical design priors used to restrict the search toward chemically plausible modifications. These priors serve as design guidance rather than structural measurements and do not replace the docking-based evidence used by MEDD.

General design principles. The prompt encourages preservation of key hydrogen-bond and hydrophobic interactions, improved complementarity with the binding pocket, and a balance between predicted binding and molecular quality. Edits are preferably small and incremental, failed modification directions are discouraged in later rounds, and synthetically accessible chemistry is preferred.

Drug-like constraints. The prompt uses common medicinal-chemistry constraints, including molecular weight between 200 and 500 Da, no more than five hydrogen-bond donors, no more than ten hydrogen-bond acceptors, and a preferred LogP range of 2–5. Potentially unfavorable chemical motifs and unnecessarily strained or unusual ring systems are discouraged, while common drug-like ring systems are preferred.

Scaffold-level modifications. When a scaffold-level edit is required, the available strategies include ring replacement, ring opening, ring fusion, ring contraction or expansion, and bioisosteric replacement. Such modifications are instructed to preserve key pharmacophoric interactions and favor scaffolds with established drug-like and synthetic properties.

Side-chain modifications. Local side-chain optimization may involve halogen substitution, alkylation, polar-group introduction, bioisosteric replacement, aromatic substitution, or adjustment of chain length. The prompt also discourages modifications that introduce obvious metabolic or chemical liabilities.

C.5 Verification Feedback

After molecular editing, the candidate is redocked and analyzed using the same structural protocol as the parent molecule. The verification stage receives the original hypothesis together with the updated interaction evidence and records whether the expected structural change is supported.

The structured verification output is summarized as:

```
status:
    verification outcome

target_change:
    observed post-editing interaction change

supported:
    whether the updated evidence supports the hypothesis

preserved_contacts:
    retained stable contacts

reason:
    evidence-based explanation
```

Table 3: Configurations used in the MEDD component ablation.

Configuration	Description
Full MEDD	Complete framework with multi-pose structural evidence, explicit interaction hypotheses, constrained molecular editing, and verification feedback.
Without multi-pose evidence	Replaces cross-pose evidence aggregation with structural information obtained from a single docking pose.
Without hypothesis	Removes the explicit interaction hypothesis and directly proposes molecular modifications from the available structural evidence.
Without chemical edit specification	Removes explicit local-edit constraints and allows less restricted molecular modification.
Without verification feedback	Removes post-editing verification feedback from subsequent refinement rounds.

```
blocking_constraint:
    failure reason, if applicable

evidence_reference:
    supporting post-editing evidence
```

An improved docking score alone is not sufficient to establish successful verification. The post-editing evidence is additionally examined for the targeted interaction change and preservation of important stable contacts. When verification fails, the resulting reason and supporting evidence are retained as feedback for subsequent refinement rounds.

C.6 Structured Refinement Trajectory

For each refinement round, MEDD records the structural evidence available before editing, the resulting hypothesis and edit specification, the updated evidence after redocking, and the corresponding verification outcome. The complete trajectory is represented as

$$\mathcal{T} = \{(E_i, H_i, A_i, E_{i+1}, V_i)\}_{i=1}^{T}, \tag{19}$$

where E_i denotes the structural evidence before editing, H_i the interaction hypothesis, A_i the corresponding chemical edit, E_{i+1} the structural evidence obtained after redocking, and V_i the verification result.

This representation preserves the provenance of each refinement decision: the molecular edit remains associated with the structural evidence that motivated it, while the verification record captures its observed structural consequence. The resulting trajectory therefore allows each refinement decision to be inspected from its supporting evidence to its post-editing outcome.

D Ablation Analysis

This section provides the complete component-level ablation results for MEDD. We evaluate four design choices corresponding to the main stages of the framework: multi-pose structural evidence, explicit interaction hypotheses, constrained chemical editing, and verification feedback.

All variants are aligned to the same set of 200 task IDs. Since the ablation run does not contain an independent Full MEDD arm, the Full MEDD reference is taken from the main experiment using the corresponding task IDs. Endpoint property statistics are computed over valid molecular outputs for each configuration, whereas the success ratio uses all 200 tasks as the denominator.

D.1 Ablation Settings

The evaluated configurations are summarized in Table 3.

Table 4: Component ablation results on the aligned 200-task subset. Vina, QED, SA, and MRR are averaged over valid outputs in each configuration; Success is computed over all 200 tasks.

Configuration	**Valid** n	**Vina** $\downarrow$	**QED** $\uparrow$	**SA** $\uparrow$	**MRR** $\uparrow$	**Success** $\uparrow$
Full MEDD	200	-7.9309	0.5567	0.7525	86.08%	28.00%
w/o Multi-pose	200	-7.8122	0.5443	0.7474	82.14%	14.50%
w/o Hypothesis	200	-7.9130	0.5659	0.7571	82.73%	14.50%
w/o Edit Spec.	200	-7.9028	0.5671	0.7678	84.26%	18.00%
w/o Verification	200	-7.9180	0.5541	0.7525	82.73%	14.00%

Table 5: Editing behavior of MEDD and its ablated variants. Values are reported as mean / median.

Configuration	**Tanimoto**	$\lvert\Delta N_{\mathrm{heavy}}\rvert$
Full MEDD	0.4647 / 0.4305	1.8557 / 1
w/o Multi-pose	0.5467 / 0.4733	1.5000 / 1
w/o Hypothesis	0.4738 / 0.4691	1.6818 / 1
w/o Edit Spec.	0.4589 / 0.3967	1.9907 / 1
w/o Verification	0.4992 / 0.4574	1.6000 / 1

These variants isolate different functions of the refinement loop. The multi-pose ablation tests the value of cross-pose structural consistency; the hypothesis ablation tests whether an explicit intermediate objective improves decision making; the edit ablation examines control over modification locality; and the verification ablation evaluates the role of feedback across refinement rounds.

D.2 Quantitative Ablation Results

Table 4 reports the endpoint results for each configuration.

The ablations affect different aspects of refinement rather than uniformly degrading every endpoint metric. Removing multi-pose evidence produces the largest deterioration in mean Vina, from -7.9309 to -7.8122, while also reducing MRR and success ratio. This indicates that cross-pose evidence contributes primarily to reliable structural guidance rather than simply restricting the magnitude of molecular edits.

Removing explicit hypothesis generation or verification feedback has a smaller effect on individual property means but reduces the success ratio from 28.0% to 14.5% and 14.0%, respectively. Notably, the hypothesis ablation retains competitive QED and SA despite the lower joint success rate. This suggests that explicit hypotheses and verification feedback are most important for coordinating multiple refinement objectives across the complete trajectory rather than independently maximizing a single molecular property.

The edit-specification ablation exhibits a different pattern. It achieves slightly higher QED and SA than Full MEDD but a lower success ratio of 18.0%. This result is consistent with the intended role of the edit specification: it constrains how optimization is performed rather than acting as a direct property-maximization component.

D.3 Controlled Editing Behavior

We further examine whether the component ablations alter the magnitude of molecular modification. Table 5 reports Tanimoto similarity to the initial molecule and absolute heavy-atom count changes.

The editing statistics further clarify the role of the chemical edit specification. Removing this component produces the largest mean heavy-atom change (1.99) and the lowest median Tanimoto similarity (0.397), consistent with broader structural modifications. In contrast, removing multi-pose evidence leads to smaller edits but substantially worse success and Vina performance. Thus, conservative modification alone is insufficient; the structural evidence used to select where and how to edit is also important.

Together, the results indicate complementary roles for the MEDD components: multi-pose evidence determines the structural context for refinement, hypotheses convert this evidence into explicit local objectives, edit specifications constrain the modification space, and verification feedback incorporates post-editing evidence into subsequent decisions.

D.4 Evaluation Scope

The component ablations are evaluated on aligned task identities, but the number of valid molecular outputs differs across configurations. We therefore report valid sample counts explicitly and retain failed tasks in the denominator of the success ratio rather than imputing missing molecular properties.

We do not report sensitivity to the number of docking poses, refinement rounds, or evidence thresholds because controlled sweeps over these settings were not performed in the current study.

E Additional Structural Case Studies

E.1 Case Study on the 8WCC Protein Pocket

We provide an additional case study on the 8WCC protein pocket to examine how multi-pose structural evidence affects molecular refinement in a distinct protein environment. Figure 5 compares four docking poses and their corresponding protein–ligand interaction maps before and after MEDD refinement.

The initial molecule exhibits substantial pose-dependent variation in its local interaction environment. In pose 1, a short-range contact is observed around ASN283 at approximately 2.5 Å, whereas pose 2 places the ligand closer to ILE103 at approximately 2.7 Å. The corresponding interaction maps also differ markedly: pose 1 contains contacts involving ASP271, ASN283, and PRO183, pose 2 emphasizes the ILE103/SER107 region together with surrounding aromatic residues, and pose 3 introduces a distinct contact around ASP102. These differences illustrate why a single docking configuration may provide an incomplete description of the local binding environment.

After refinement, the interaction pattern becomes more consistent around specific local anchors. Most notably, ASN283 participates in the interaction network across all four refined poses, while residues such as THR268, PHE184/PHE185, and ASP271 retain pose-dependent contributions. MEDD therefore does not require all poses to converge to an identical interaction map. Instead, recurrent contacts provide the stable structural evidence used to constrain refinement, while variations in the surrounding interaction network remain available as local optimization opportunities.

This case illustrates the intended role of multi-pose evidence in MEDD: refinement strengthens cross-pose consistency around selected structural features without forcing a single binding configuration. The edited molecule is subsequently evaluated under the same docking protocol, allowing the post-editing interaction ensemble to be compared directly with the evidence that motivated the modification. This provides a concrete structural link between evidence-guided editing and post-editing verification.

F Reproducibility Details

This section summarizes the implementation details required to reproduce the reported experiments.

F.1 Computational Environment

All experiments use a consistent software environment and evaluation pipeline. Molecular docking is performed with AutoDock Vina under the unified protocol described in Appendix A, while molecular properties are computed with the RDKit-based evaluation pipeline. Protein–ligand interactions are extracted using the same interaction-analysis module throughout MEDD. LLM inference uses a fixed backend configuration together with the prompt organization described in Appendix C.

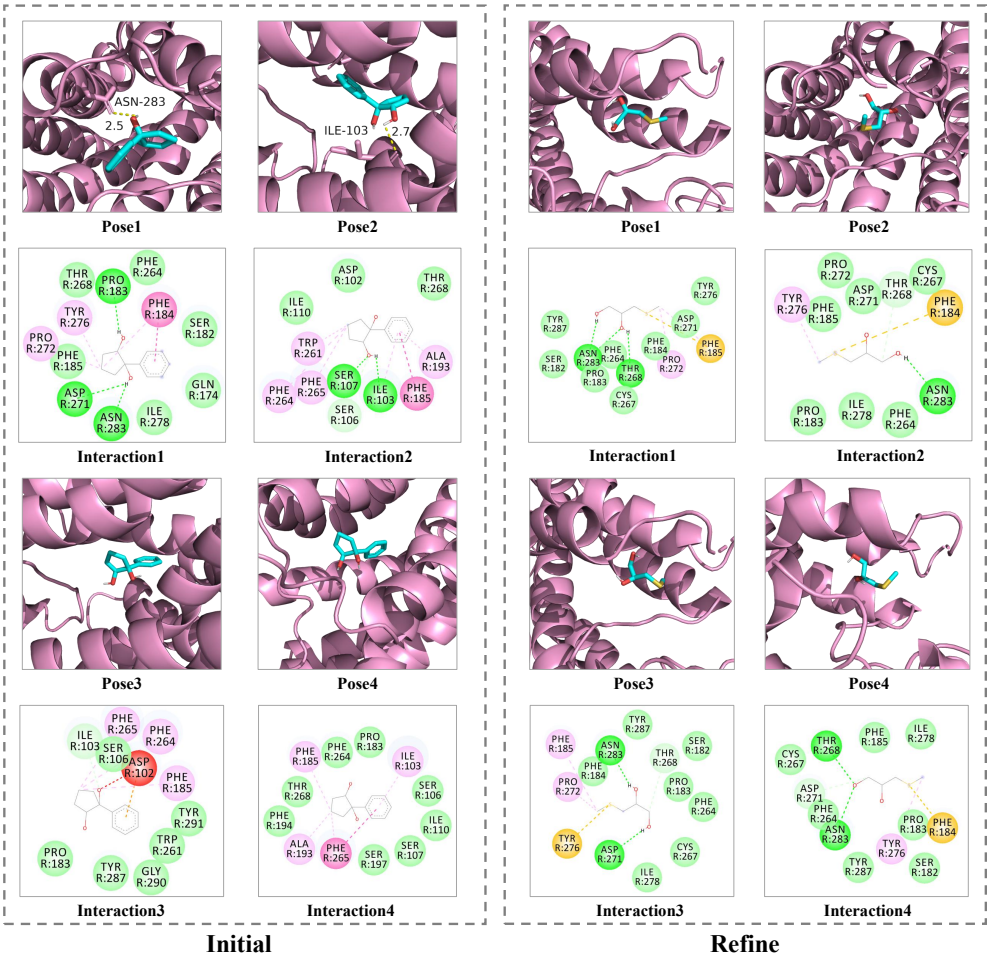


Figure 5: Structural case study on the 8WCC protein pocket. Four docking poses and corresponding interaction maps are shown for the initial and refined molecules. The initial poses exhibit heterogeneous local interaction patterns, whereas the refined poses show increased cross-pose consistency around selected local contacts while retaining pose-dependent interactions elsewhere.

F.2 Randomness Control

Random seeds are fixed whenever supported by the corresponding computational tools. The initial molecular candidates are sampled before optimization and kept fixed across subsequent evaluations. Paired comparisons therefore use the same initial molecule, protein pocket, receptor structure, and evaluation protocol whenever the corresponding outputs are available.

F.3 External Tool Configuration

All structural evaluations use the same docking boxes, sampling parameters, and interaction-analysis settings before and after molecular editing. This ensures that observed differences between initial and refined candidates are not introduced by changes in the evaluation protocol. Complete docking and conformer-generation settings are provided in Appendix A.