AFIMOL: ADAPTIVE FRAGMENTATION AND FRAGMENT–POCKET INTERACTION LEARNING FOR 3D MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating 3D ligands conditioned on protein pockets remains a central challenge in structure-based drug design. However, fragment-based generators often rely on fixed chemical rules for molecular decomposition and lack explicit supervision for evaluating candidate fragments according to their residue-level interaction preferences. Consequently, predefined fragment boundaries may fail to capture molecule-specific structural organization, while fragment selection may not fully exploit localized interaction evidence from the protein pocket. To address these limitations, we propose AFIMol, an autoregressive framework for 3D molecular generation that integrates adaptive fragmentation with fragment-pocket interaction learning. Rather than determining fragment boundaries solely through fixed cutting rules, AFIMol learns molecule-specific boundary scores from atom-level structural responses while restricting cuts to chemically admissible bonds. At each generation step, AFIMol combines the current generation state with each candidate fragment to predict its interaction preferences over pocket residues. The interaction predictor is supervised using contact annotations from protein-ligand complexes, and the predicted residue-level interactions are aggregated into compatibility scores that adjust candidate logits during inference.This mechanism guides the generator toward fragments that better match the local pocket environment. Experiments on CrossDocked2020 show favorable synthetic accessibility and molecular diversity while maintaining competitive docking performance. Ablation studies validate the contributions of adaptive fragmentation and fragment-pocket interaction learning, while structural case studies illustrate the localized interaction patterns of the generated ligands.

1 INTRODUCTION

Designing ligands for protein pockets requires coordinated control of molecular connectivity, three-dimensional conformation, and the functional groups exposed to surrounding residues. Structure-based drug design (SBDD) leverages target structures to guide this search Anderson (2003), but chemical validity, predicted binding, drug-likeness, and synthetic feasibility are not interchangeable objectives. A useful generator must therefore jointly model molecular structure and the spatially heterogeneous pocket environment, while capturing complex structure–interaction dependencies.

Existing pocket-conditioned generators operate mainly at the atom level or through iterative denoising paradigms. Autoregressive methods such as Pocket2Mol Peng et al. (2022) predict atoms, coordinates, and bonds sequentially, whereas TargetDiff Guan et al. (2023a) and DiffSBDD Schneuing et al. (2024) generate 3D structures through pocket-conditioned denoising. Fragment-based models instead treat connected substructures as reusable building blocks, preserving local connectivity while shortening the generation sequence and reducing search. FLAG Zhang et al. (2023b), DrugGPS Zhang & Liu (2023), and Frag2Seq Fu et al. (2025) demonstrate the promise of this representation. Its effectiveness, however, depends on two coupled choices: how fragments are defined and how the next fragment is selected for a given pocket with structural and interaction awareness.

Conventional decompositions determine fragment boundaries using predefined chemical criteria. BRICS applies chemically motivated cleavage and recombination rules Degen et al. (2008), while

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.

Frag2Seq cuts bonds satisfying fixed rotatability criteria Fu et al. (2025). Although chemically useful, these rules do not rank alternative admissible boundaries according to their molecular context. Learning molecule-dependent cutting priorities can preserve more appropriate local substructures while retaining chemical constraints on where cuts are allowed.

Fragment selection poses a complementary challenge: plausible continuations of the same partial ligand may expose very different polar, hydrophobic, or aromatic groups to nearby residues. Recent methods demonstrate the value of explicit interaction modeling through affinity supervision, residue prompts, or joint interaction generation Huang et al. (2024b); Wu et al. (2024b); Lee et al. (2025); Cremer et al. (2026). Yet a fragment-autoregressive backbone does not explicitly predict candidate–residue contacts for each possible next token. Such localized interaction evidence can provide an additional criterion for candidate selection and improve pocket-specific generation.

These observations motivate **AFIMol**, a structure-adaptive and interaction-aware framework for autoregressive 3D fragment generation. Its **Structure-aware Adaptive Fragmentation (SAF)** module predicts molecule-specific boundary scores from atom-level structural responses and selects cuts within a chemically admissible set. Its **Fragment Interaction-aware Pocket Residue Adapter (FIPRA)** combines the current generation state with each candidate fragment, predicts residue-level interaction preferences under contact supervision, and aggregates them into a compatibility score that adjusts the backbone logits. On CrossDocked2020, AFIMol attains an SA score of 0.788 and a diversity score of 0.771 while maintaining a competitive Vina Score of -7.241.

Our main contributions are summarized as follows:

- We propose **AFIMol**, a unified structure-adaptive and interaction-aware framework for pocket-conditioned 3D fragment generation. By jointly modeling molecule-specific structural organization and localized protein-pocket compatibility, AFIMol establishes a flexible and effective fragment-based generation paradigm that better aligns molecular construction with the target binding environment while retaining the chemical sequence prior of the autoregressive backbone and providing a comprehensive structure-aware perspective.
- We develop AFIMol to address these limitations through two complementary designs. For **fragment construction**, we develop a structure-aware adaptive fragmentation strategy that determines boundaries from molecule-specific structural responses, with chemical constraints only restricting admissible cuts. For **fragment selection**, we introduce a fragment interaction-aware pocket residue adapter that combines the current generation state with each candidate fragment to predict its interaction preferences over pocket residues.
- We evaluate **AFIMol** on CrossDocked2020, where it achieves the highest reported synthetic accessibility and diversity scores among the compared methods while maintaining competitive docking performance. Comprehensive ablation studies systematically assess the effects of the two components, and detailed structural case studies further illustrate the interaction patterns of generated ligands in distinct protein pockets across diverse targets.

2 Related Work

Structure-Based Drug Design.Structure-based drug design Liu et al. (2026) aims to generate molecules that can bind to the given protein pocket. Early approaches like LiGAN Ragoza et al. (2022) represent protein-ligand structures using atomic density grids. Subsequent works Liu et al. (2022); Luo et al. (2021); Peng et al. (2022) utilize autoregressive models to generate molecules sequentially. For instance, GraphBP Liu et al. (2022) employs normalizing flow for equivariant atom position generation, while Pocket2Mol Peng et al. (2022) further integrates bond generation to ensure structural validity. Recent advances leverage diffusion models Schneuing et al. (2024); Guan et al. (2023a;b) to generate all atoms in a one-shot denoising process. To improve structural realism, fragment-based methods like FLAG Zhang et al. (2023b) and DrugGPS Zhang & Liu (2023) generate molecules motif-by-motif. However, most of these methods focus on explicit 3D coordinate generation, often ignoring the rich chemical syntax captured by language models.

Language Models for Chemistry. Chemical Language Models (CLMs) have demonstrated remarkable efficacy in molecular representation learning Bran & Schwaller (2024); Janakarajan et al. (2024); Bajorath (2024); Zhang et al. (2025). Works such as DrugGPT Li et al. (2023),

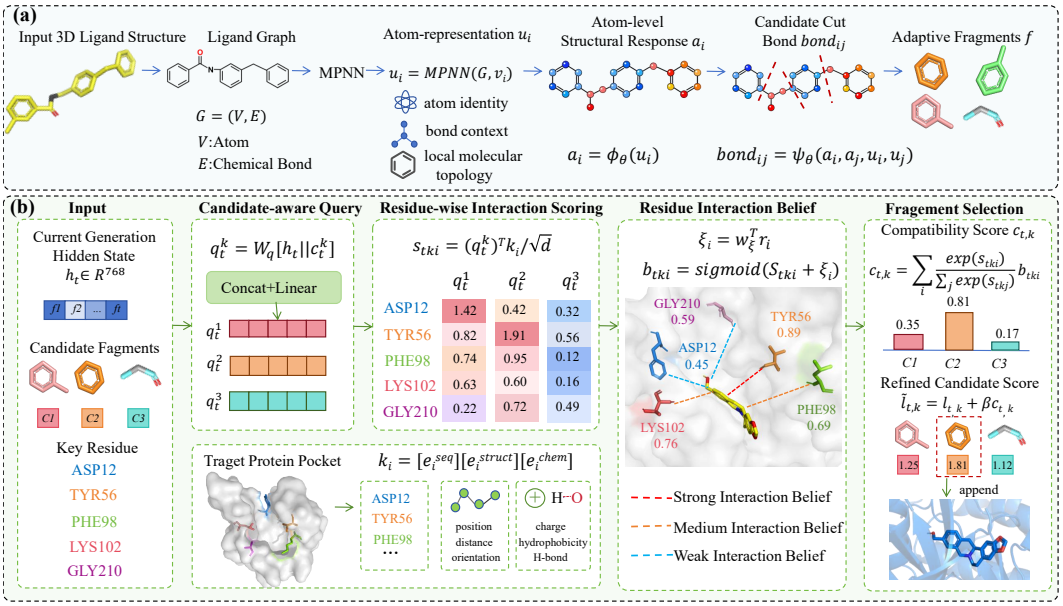


Figure 1: **Overview of AFIMol.** (a) SAF predicts molecule-specific boundaries over chemically admissible bonds to construct adaptive fragments. (b) FIPRA estimates candidate-residue interactions and aggregates them into a compatibility score $c_{t,k}$ that refines the backbone logit $\ell_{t,k}$ for next-fragment selection.

MoleculeGPT Zhang et al. (2023a), and InstructMol Cao et al. (2025) adapt LMs to process molecular SMILES strings through large-scale self-supervised pre-training. While these models excel at capturing chemical syntax, they predominantly operate in the 2D domain. To bridge the gap with structural biology, some studies incorporate geometric encoders such as GNNs to effectively inject 3D information Xia et al. (2023); Zhang et al. (2023a); Cao et al. (2025). Nonetheless, these approaches typically rely on complex multi-modal alignment strategies and decouple the generative process from explicit 3D geometry. Crucially, existing CLM-based SBDD works, such as TamGen Wu et al. (2024a), Lingo3DMol Feng et al. (2024), and TSMGen Chen et al. (2026), either limit their scope to 2D SMILES generation or predict coordinates post-hoc, failing to directly integrate 3D geometric regression into the autoregressive generation pipeline.

3 Preliminary

Let a protein pocket be a set of atoms $P = \{(\mathbf{x}_i^P, \mathbf{v}_i^P)\}_{i=1}^{N_P}$ and a ligand be $M = \{(\mathbf{x}_j^M, \mathbf{v}_j^M)\}_{j=1}^{N_M}$, where $\mathbf{x} \in \mathbb{R}^3$ is an atom coordinate and $\mathbf{v}$ is its atom type. SBDD models the conditional distribution $p(M \mid P)$.

AFIMol represents a ligand as an ordered sequence of fragments $F = (F_1, \ldots, F_T)$. Each fragment F_t contains a discrete identity token and the geometric tokens used by the underlying GPT generator. The conditional generation objective is

$$p_\theta(F \mid P) = \prod_{t=1}^{T} p_\theta(F_t \mid F_{<t}, P), \tag{1}$$

where $F_{<t}$ denotes the generated prefix. At step t, the generator produces a hidden state $\mathbf{h}_t$ and a candidate set $\mathcal{C}_t = \{f_t^1, \ldots, f_t^K\}$. AFIMol keeps this chemical language-model backbone and adds two operations: SAF defines the fragment units and FIPRA supplies candidate-specific protein evidence.

4 METHOD

4.1 OVERVIEW

AFIMol augments a fragment-autoregressive generator with adaptive fragment construction and candidate-conditioned protein interaction reasoning. Given a pocket P and prefix $F_{<t}$, the backbone outputs a hidden state $\mathbf{h}t$, candidates $\mathcal{C}t = f_t^1, \ldots, f_t^K$, and base logits $\ell t, k = \text{logit}\,\theta(f_t^k \mid F_{<t}, P)$. SAF learns molecule-specific fragment boundaries within chemically admissible cuts and preserves meaningful molecular substructures. FIPRA predicts candidate–residue interaction preferences under contact supervision and uses their aggregated evidence to refine the base logits accordingly.

4.2 STRUCTURE-AWARE ADAPTIVE FRAGMENTATION

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a molecular graph and $\mathbf{u}_i$ the representation of atom i produced by a message-passing encoder. SAF identifies fragment boundaries from structural variation within each molecule rather than prescribing cut locations through fixed chemical rules. It first computes an atom-level response $a_i = \phi_\theta(\mathbf{u}_i)$ and constructs the bond descriptor $\boldsymbol{\xi}_{ij} = [a_i, a_j, |a_i - a_j|, \mathbf{u}_i \| \mathbf{u}_j]$, which combines neighboring responses with their molecular context. A learnable pair encoder maps this descriptor to a transition score d_{ij}, which is converted into a boundary response using a learnable offset μ_b and a positive temperature τ_b. The resulting $b_{ij} \in (0, 1)$ measures the evidence for cutting bond (i, j):

$$d_{ij} = \psi_\theta(\boldsymbol{\xi}_{ij}), \qquad b_{ij} = \sigma\left(\frac{d_{ij} - \mu_b}{\tau_b}\right). \tag{2}$$

The boundary responses determine which bonds are selected, while chemical constraints restrict selection to the admissible set $\mathcal{E}_{\text{cut}} \subseteq \mathcal{E}$, such as single non-ring bonds satisfying valence constraints. This distinction allows bonds satisfying the same chemical criteria to receive different cutting priorities according to their molecular context. Given a cut budget B, the operator $\text{Top}_B(\mathcal{E}_{\text{cut}}; b)$ returns the edge set containing the $\min(B, |\mathcal{E}_{\text{cut}}|)$ largest boundary responses. Removing these edges produces a partition into connected components, each of which defines a fragment:

$$\mathcal{S} = \text{Top}_B(\mathcal{E}_{\text{cut}}; b), \qquad \mathcal{F} = \text{CC}(\mathcal{V}, \mathcal{E} \setminus \mathcal{S}). \tag{3}$$

SAF regularizes the relaxed boundary field toward a target cut ratio while discouraging ambiguous responses. Let $m = |\mathcal{E}_{\text{cut}}|$ and $\bar{b} = m^{-1} \sum_{(i,j)\in\mathcal{E}_{\text{cut}}} b_{ij}$. A target ratio ρ_B controls the mean response, while $\gamma \geq 0$ weights the penalty on values far from binary decisions. A straight-through estimator uses hard decisions in the forward pass and relaxed responses in the backward pass. For $m = 0$, no bond is removed and the loss is zero; otherwise,

$$\mathcal{L}_{\text{SAF}} = (\bar{b} - \rho_B)^2 + \frac{\gamma}{m} \sum_{(i,j)\in\mathcal{E}_{\text{cut}}} b_{ij}(1 - b_{ij}). \tag{4}$$

4.3 FRAGMENT INTERACTION-AWARE POCKET RESIDUE ADAPTER

FIPRA predicts candidate preferences over pocket residues. Let $\mathbf{r}_i \in \mathbb{R}^{d_r}$ be the feature of residue i and $\mathbf{e}_t^k$ the embedding of candidate f_t^k. The adapter combines $\mathbf{e}_t^k$ with generation state $\mathbf{h}_t$ and projects residue features into the same d-dimensional space:

$$\mathbf{q}_t^k = W_q[\mathbf{h}_t \| \mathbf{e}_t^k], \qquad \mathbf{k}_i = W_k \mathbf{r}_i. \tag{5}$$

For each candidate–residue pair, FIPRA combines scaled similarity $s_{tki} = (\mathbf{q}_t^k)^\top \mathbf{k}_i / \sqrt{d}$ with the residue-dependent bias $\delta_i = \mathbf{w}_\delta^\top \mathbf{r}_i$. The resulting belief b_{tki} depends on both the candidate and generated prefix, and the full vector is retained for contact supervision and analysis:

$$b_{tki} = \sigma(s_{tki} + \delta_i), \qquad \mathbf{b}_{t,k} = [b_{tk1}, \ldots, b_{tkL}]. \tag{6}$$

FIPRA converts residue-level beliefs into a scalar candidate score by normalizing s_{tki} over pocket residues and using the resulting weights to aggregate b_{tki}. The compatibility score $c_{t,k}$ therefore

emphasizes residues most relevant to the current candidate and can directly refine its backbone logit:

$$\alpha_{tki} = \frac{\exp(s_{tki})}{\sum_{j=1}^{L} \exp(s_{tkj})}, \qquad c_{t,k} = \sum_{i=1}^{L} \alpha_{tki} b_{tki}. \tag{7}$$

We supervise the interaction predictor using contact annotations from reference protein–ligand complexes. For the teacher-forced fragment $f_t^\star$, atom-level hydrogen bonds, hydrophobic contacts, and aromatic contacts involving that fragment are mapped to the corresponding pocket residues and combined into soft targets $y_{ti} \in [0, 1]$. The ground-truth fragment is included in the training candidate set at index $k_t^\star$, and its predicted belief is denoted by $b_{ti}^\star = b_{tk_t^\star i}$. Class weights ω_{ti} compensate for the sparsity of contacting residues, giving the following interaction loss:

$$\mathcal{L}_{\mathrm{FI}} = -\sum_{t=1}^{T} \sum_{i=1}^{L} \omega_{ti} \left[y_{ti} \log b_{ti}^\star + (1 - y_{ti}) \log(1 - b_{ti}^\star) \right]. \tag{8}$$

Training combines the autoregressive objective $\mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{T} \log p_\theta(f_t^\star \mid F_{<t}, P)$ with the contact-supervised interaction loss and the SAF boundary regularizer. Nonnegative coefficients λ_{FI} and λ_{SAF} balance the three terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{FI}} \mathcal{L}_{\mathrm{FI}} + \lambda_{\mathrm{SAF}} \mathcal{L}_{\mathrm{SAF}}. \tag{9}$$

During inference, FIPRA predicts interaction evidence from the pocket, the generated prefix, and the candidate fragment without requiring reference contact annotations. The compatibility score $c_{t,k}$ is added to the base logit $\ell_{t,k}$, with $\beta \geq 0$ controlling the strength of interaction guidance. The adjusted logits are then normalized over $\mathcal{C}_t$ to obtain the sampling distribution. This allows local residue interaction preferences to affect the next-fragment decision while preserving the backbone's contribution to candidate ranking; setting $\beta = 0$ recovers the base distribution over the same candidate set:

$$\widetilde{\ell}_{t,k} = \ell_{t,k} + \beta c_{t,k}, \qquad p(f_t^k \mid F_{<t}, P) = \mathrm{softmax}_k(\widetilde{\ell}_{t,k}). \tag{10}$$

The complete training and inference procedure of FIPRA is summarized in Algorithm 1 in Appendix A.

5 Experiments

We evaluate overall performance, component contributions, the intended representation and geometry changes, and molecule-level structural and synthesis-planning hypotheses under a comprehensive experimental protocol. Accordingly, benchmark results establish the outcome profile, controlled ablations identify component-level effects, mechanistic diagnostics examine the intended behavior, and the case study provides only qualitative illustration of representative generated molecules.

5.1 Experimental Settings

Dataset. Following previous structure-based drug design (SBDD) works, we evaluate our method on the standard processed CrossDocked2020 benchmark Luo et al. (2021). The dataset contains protein-ligand complexes collected from experimentally determined structures and further filtered according to binding pose quality. Specifically, we adopt the commonly-used curated version with a binding pose RMSD threshold of 1 Å, and split the dataset according to 30% protein sequence identity to avoid information leakage between training and testing proteins. The final dataset contains approximately 100,000 protein-ligand pairs for training and 100 protein pockets for evaluation. Each sample consists of a target protein pocket and its corresponding ligand, where the objective is to generate novel ligands conditioned on the given protein structure.

Baselines. We compare AFIMol with nine pocket-conditioned generators evaluated on CrossDocked2020. FLAG Zhang et al. (2023b) and DrugGPS Zhang & Liu (2023) perform motif-level autoregressive generation. PMDM Huang et al. (2024a), AMDiff Li et al. (2025), and Apo2Mol Zheng et al. (2026) are diffusion-based methods; AMDiff couples atom- and motif-level

Table 1: Performance comparison of AFIMol with existing SBDD methods on CrossDocked2020. ↑/↓ indicates that higher/lower values are better. The best results are highlighted in bold, while the second-best results are underlined.

Model	Vina Score (↓)	High Affinity (↑)	QED (↑)	SA (↑)	LogP	Lipinski (↑)	Diversity (↑)
FLAG Zhang et al. (2023b)	-7.247	0.580	0.495	<u>0.745</u>	0.630	4.943	0.720
DrugGPS Zhang & Liu (2023)	-7.276	0.565	0.613	0.743	0.913	4.917	–
PMDM Huang et al. (2024a)	-7.572	**0.628**	0.594	0.611	0.301	<u>4.975</u>	0.709
TamGen Wu et al. (2024a)	<u>-7.745</u>	0.559	**0.771**	–	–	–	<u>0.747</u>
AMDiff Li et al. (2025)	-7.466	–	0.479	0.681	0.814	4.758	0.729
Token-Mol Wang et al. (2025)	-6.862	–	0.568	0.648	2.319	4.666	0.711
Frag2Seq Fu et al. (2025)	-7.366	–	<u>0.645</u>	0.642	–	**4.989**	0.672
Apo2Mol Zheng et al. (2026)	**-7.860**	0.529	0.610	0.590	3.490	4.760	–
MolCRAFT Qu et al. (2024)	-6.590	–	0.500	0.690	–	–	0.720
AFIMol (Ours)	-7.241	<u>0.589</u>	0.596	**0.788**	3.003	4.959	**0.771**

views, whereas Apo2Mol additionally models pocket conformational changes. TamGen Wu et al. (2024a) and Token-Mol Wang et al. (2025) use Transformer-based chemical language models, while Frag2Seq Fu et al. (2025) combines fragment-aware geometric tokenization with autoregressive sequence generation. MolCRAFT Qu et al. (2024) instead operates in a continuous parameter space. Together, these baselines cover diverse paradigms including motif assembly, diffusion, continuous-space modeling, and token-based generation. We retain available CrossDocked2020 metrics and denote unreported values by "–". Because published implementations may differ in sampling and docking details, we interpret Table 1 as a benchmark-level comparison under consistent evaluation criteria; component effects are assessed with the controlled ablations below.

Evaluation Metrics. Following previous SBDD studies, we evaluate generated molecules from multiple perspectives, including binding affinity, drug-likeness, synthetic feasibility, and structural diversity. Specifically, we use Vina Score to estimate the binding affinity between generated ligands and target protein pockets, where a lower score indicates stronger predicted binding. We further report High Affinity, which measures the proportion of generated molecules achieving better binding affinity than the reference ligand. To evaluate molecular properties, we adopt QED for drug-likeness assessment, SA for synthetic accessibility estimation, and Lipinski score to measure compliance with Lipinski's rule of five. Moreover, Diversity is calculated based on pairwise molecular similarity to evaluate the overall structural diversity of generated samples. Vina scores are calculated using QVina, while molecular properties are obtained using RDKit implementations.

Implementation Details. Our model is implemented based on an autoregressive transformer architecture. During training, the model is optimized with the combination of generation loss and interaction-guided adaptation loss introduced in our framework. We use the AdamW optimizer with cosine learning rate scheduling and a linear warm-up strategy. All experiments are conducted on NVIDIA GPUs under the same computational setting. During inference, we generate multiple candidate molecules for each protein pocket using autoregressive sampling and select the final molecules according to the predicted generation probability from the model.

5.2 Main Results

5.2.1 Comparison with State-of-the-Art Methods

Table 1 compares AFIMol with representative fragment-, diffusion-, and sequence-based SBDD methods. AFIMol obtains a Vina Score of -7.241 and a High Affinity ratio of 0.589. Although several baselines have lower mean Vina scores, AFIMol ranks second in High Affinity among methods that report it. Its main advantage is the joint property profile: AFIMol achieves the best SA (0.788) and Diversity (0.771), exceeding the second-best reported values by 0.043 and 0.024, while retaining QED (0.596) and Lipinski (4.959) values comparable to several baselines. Thus, Table 1 supports a favorable trade-off among synthetic accessibility, sample diversity, and docking-based performance rather than docking dominance. The ablations below test whether SAF and FIPRA account for this trade-off. A metric-wise comparison is provided in Appendix B.

Table 2: Ablation study of different components in AFIMol. ↑/↓ indicates that higher/lower values are better. The best results are highlighted in bold, while the second-best results are underlined.

Model	Vina Score (↓)	High Affinity (↑)	QED (↑)	SA (↑)	LogP	Lipinski (↑)	Diversity (↑)
w/o SAF	<u>-7.226</u>	0.523	0.589	0.765	2.663	4.926	0.753
w/o FIPRA	-7.082	0.520	**0.625**	<u>0.771</u>	2.915	**4.989**	**0.782**
w/o SAF+FIPRA	-6.789	<u>0.531</u>	<u>0.618</u>	0.671	2.293	<u>4.976</u>	0.711
AFIMol (Ours)	**-7.241**	**0.589**	0.596	**0.788**	3.003	4.959	<u>0.771</u>

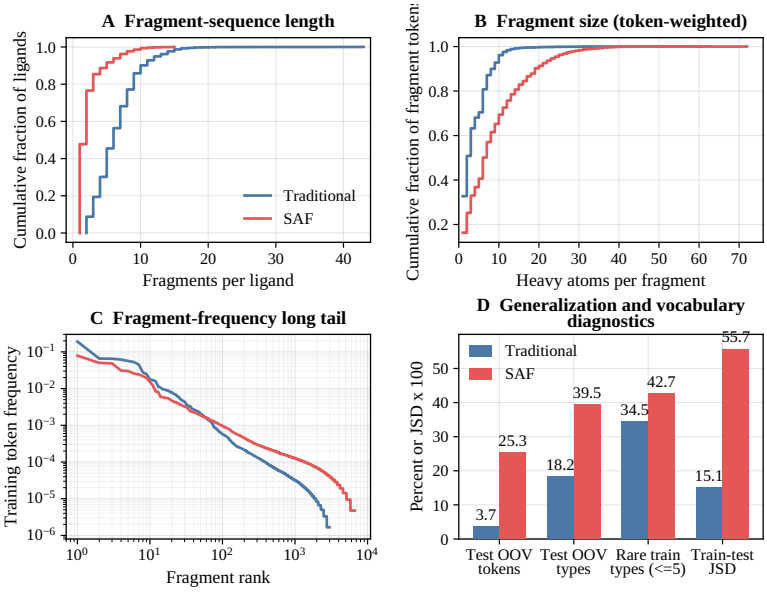


Figure 2: **Fragment-space analysis of SAF.** (a) Fragment-sequence length, (b) token-weighted fragment size, (c) fragment-frequency long tail, and (d) vocabulary diagnostics.

5.2.2 Ablation Study

Table 2 isolates the two modules. Removing FIPRA weakens Vina Score from -7.241 to -7.082 and High Affinity from 0.589 to 0.520, while QED and Diversity increase. Its point-estimate benefit is therefore specific to pocket-directed metrics rather than every property. Removing SAF lowers SA from 0.788 to 0.765, Diversity from 0.771 to 0.753, and High Affinity from 0.589 to 0.523, supporting its role in fragment quality and generation. Joint removal gives the least favorable Vina Score (-6.789), SA (0.671), and Diversity (0.711), although High Affinity is slightly above the single ablations. The effects are complementary but not strictly additive. Appendix C provides a property-wise interpretation of these trade-offs.

5.2.3 Mechanistic Analysis of SAF and FIPRA

Effect of SAF on molecular fragmentation. Figure 2 tests the intended representational change. Relative to rule-based fragmentation, SAF produces fewer, larger fragments. Its heavier tail, higher OOV, and greater train–test divergence reveal the corresponding cost: greater molecule specificity reduces token reuse. The ablation supplies outcome-level evidence for this changed fragmentation.

Effect of FIPRA on protein–ligand geometry. Figure 3 compares paired pocket-level statistics with and without FIPRA. Nearby-residue density per ligand heavy atom and C–C proximity increase by medians of $+0.060$ and $+0.047$, indicating a shift in local packing.

The ligand N/O fraction rises (median $+0.063$), whereas N/O proximity changes modestly ($+0.007$); composition alone therefore does not explain the geometric shifts. Together with the ablation, these results associate FIPRA with target-facing metrics and geometry. Proximity does

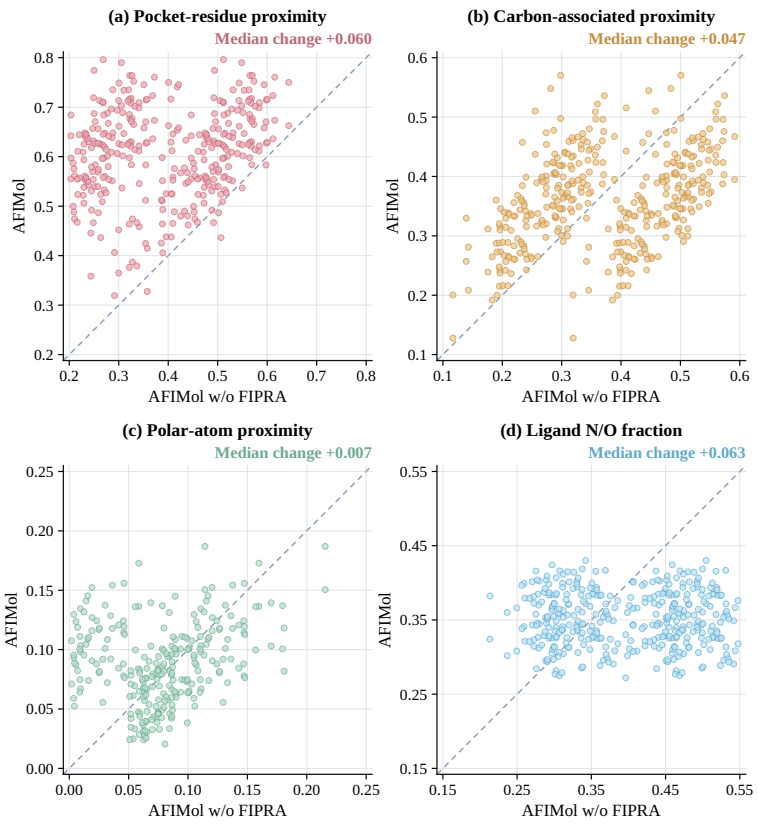


Figure 3: **Pocket-level geometric analysis of FIPRA.** AFIMol w/o FIPRA (x-axis) versus AFIMol (y-axis): (a) nearby residues per ligand heavy atom within 4 Å, (b) C–C nearby-residue density within 4 Å, (c) N/O nearby-residue density within 3.5 Å, and (d) ligand N/O fraction. The dashed diagonal denotes no change.

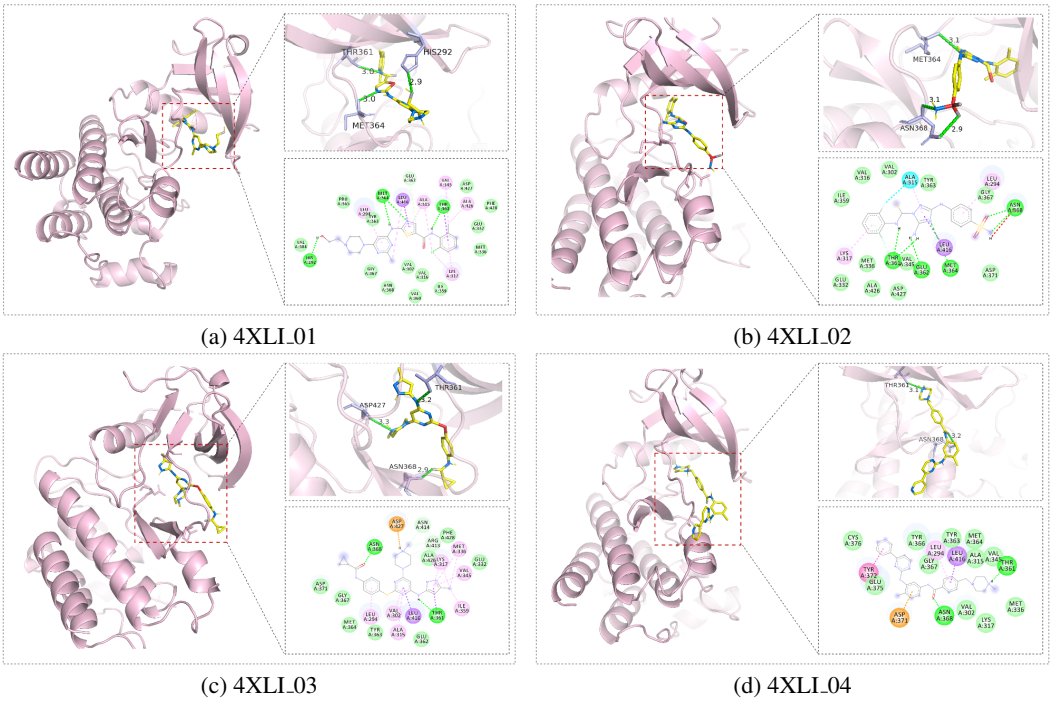
not establish non-covalent interactions, and the following case study remains qualitative. Additional SAF and FIPRA diagnostics are discussed in Appendix D.

5.3 Case Study

To examine whether FIPRA's population-level geometric effects yield residue-resolved hypotheses, we study PDB 4XLI, a 2.5 Å X-ray structure of mouse ABL2/ARG in a ligand-bound DFG-in conformation. Four generated candidates are docked into the crystallographic ATP-binding site, with best Vina scores of -9.77, -9.65, -8.93, and -9.58 kcal/mol. Because the narrow 0.85 kcal/mol range precludes reliable affinity ranking, we use the poses only to compare local pocket geometries.

Figure 4 visualizes the selected 3D poses and 2D interaction maps. The predicted geometries recur near kinase landmarks— gatekeeper Thr361, hinge-region Met364, and Asn368—while extensions toward His292 or DFG residue Asp427 distinguish candidates. This connects FIPRA-guided selection to structurally meaningful ATP-pocket regions, providing residue-level interpretability beyond a global docking score. Detailed contacts are reported in Appendix E. These static poses define testable hypotheses, not evidence of interaction persistence, ABL2 selectivity, or biological activity.

Retrosynthetic analysis. To connect these residue-level hypotheses with experimentally actionable follow-up studies, we further evaluate route-search accessibility for all four generated candidates with AiZynthFinder Genheden et al. (2020), using the `full` USPTO template-based expansion policy, 17,422,831 buyable compounds, a maximum depth of six transformations, and 100 iterations per molecule under the specified search configuration. A route is considered successfully solved only when every terminal leaf occurs in the configured stock database.



(a) 4XLI_01 (b) 4XLI_02

(c) 4XLI_03 (d) 4XLI_04

Figure 4: Predicted docking poses and interaction maps of four generated candidates for 4XLI, illustrating candidate-specific arrangements around recognizable ATP-pocket residues.

Table 3: AiZynthFinder retrosynthetic route-search results for the four generated molecules under the stated search configuration.

Candidate	Search time (s)	Route steps	Solved routes / total	Top score	Stock status
4XLI_01	4.52	2	35 / 125 (28.0%)	0.994	3/3 precursors in stock
4XLI_02	3.24	4	42 / 130 (32.3%)	0.738	4/4 precursors in stock
4XLI_03	4.26	3	58 / 153 (37.9%)	0.987	4/4 precursors in stock
4XLI_04	3.17	2	88 / 110 (80.0%)	0.994	3/3 precursors in stock

All four candidates have top-ranked two- to four-step routes ending in stock-listed precursors. 4XLI_01 and 4XLI_04 have the shortest routes, while 4XLI_04 has the largest solved-route fraction. The route search therefore adds an experimental-accessibility dimension to the structural interpretation by pairing candidate-specific pocket hypotheses with concrete synthesis plans that can support subsequent binding and functional assays. This does not validate the predicted contacts, because planner scores rank search solutions rather than estimating reaction success, yield, or binding and remain dependent on the template library, stock, and search budget. Together, the docking and retrosynthetic analyses provide complementary explanations of where the candidates may engage the pocket and how those hypotheses could be made experimentally testable. Extended pose-, property-, and route-level interpretations are provided in Appendix E.

6 Conclusion

We presented AFIMol, a pocket-conditioned 3D molecular generator that couples structure-aware adaptive fragmentation with fragment–residue interaction learning. SAF learns molecule-specific fragment boundaries within chemically admissible cuts, while FIPRA uses predicted residue-level interactions to refine candidate selection. On CrossDocked2020, AFIMol achieves the best synthetic accessibility and diversity among the compared methods while maintaining competitive docking performance. Ablations and mechanistic analyses further show that SAF changes fragment organization and that FIPRA alters local protein–ligand geometry. These results support adaptive fragment construction and localized pocket guidance as complementary directions for 3D molecular generation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX preparation, citation formatting, and presentation of the metric formulation. The authors reviewed the AI-assisted material and take responsibility for the final text, derivations, claims, citations, and artifacts.

REFERENCES

Amy C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, 2003. doi: 10.1016/j.chembiol.2003.09.002.

Jürgen Bajorath. Chemical language models for molecular design. *Molecular Informatics*, 43(1): e202300288, 2024. doi: 10.1002/minf.202300288.

Andres M. Bran and Philippe Schwaller. Transformers and large language models for chemistry and drug discovery. In *Drug Development Supported by Informatics*, pp. 143–163. Springer Nature Singapore, Singapore, 2024. doi: 10.1007/978-981-97-4828-0_8.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 354–379, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.25/.

Yaoyu Chen, Xiaoli Lin, Ziyi Gong, and Jun Pang. TSMGen: Target-specific molecule generation via higher-order structural dependencies and context-aware bidirectional fusion. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=T5eEBKqIT3.

Julian Cremer, Ross Irwin, Alessandro Tibo, Jon Paul Janet, Simon Olsson, and Djork-Arné Clevert. FLOWR: Flow matching for structure-aware de novo, interaction- and fragment-based ligand generation. *Nature Computational Science*, 6(6):565–574, 2026. doi: 10.1038/s43588-026-00998-8.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using drug-like chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: 10.1002/cmdc.200800178.

Wei Feng, Lvwei Wang, Zaiyun Lin, Yanhao Zhu, Han Wang, Jianqiang Dong, Rong Bai, Huting Wang, Jielong Zhou, Wei Peng, Bo Huang, and Wenbiao Zhou. Generation of 3d molecules in pockets via a language model. *Nature Machine Intelligence*, 6(1):62–73, 2024. doi: 10.1038/s42256-023-00775-6.

Cong Fu, Xiner Li, Blake Olson, Heng Ji, and Shuiwang Ji. Fragment and geometry aware tokenization of molecules for structure-based drug design using language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mMhZS7qt0U.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12:70, 2020. doi: 10.1186/s13321-020-00472-1.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=kJqXEPXMsE0.

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. URL https://proceedings.mlr.press/v202/guan23a.html.

Lei Huang, Tingyang Xu, Yang Yu, Peilin Zhao, Xingjian Chen, Jing Han, Zhi Xie, Hailong Li, Wenge Zhong, Ka-Chun Wong, and Hengtong Zhang. A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications*, 15: 2657, 2024a. doi: 10.1038/s41467-024-46569-1.

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 *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=qH9nrMNTIW.

Nikita Janakarajan, Tim Erdmann, Sarath Swaminathan, Teodoro Laino, and Jannis Born. Language models in molecular discovery. In *Drug Development Supported by Informatics*, pp. 121–141. Springer Nature Singapore, Singapore, 2024. doi: 10.1007/978-981-97-4828-0_7.

Joongwon Lee, Wonho Zhung, Jisu Seo, and Woo Youn Kim. BInD: Bond and interaction-generating diffusion model for multi-objective structure-based drug design. *Advanced Science*, 12(35):e02702, 2025. doi: 10.1002/advs.202502702.

Guanlue Li, Chenran Jiang, Ziqi Gao, Yu Liu, Chenyang Liu, Jiean Chen, Yong Huang, and Jia Li. Molecule generation for target protein binding with hierarchical consistency diffusion model. *Chemical Science*, 16(39):18263–18277, 2025. doi: 10.1039/D5SC02113H.

Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu, and Suxia Han. DrugGPT: A GPT-based strategy for designing potential ligands targeting specific proteins. *bioRxiv*, 2023. doi: 10.1101/2023.06.29.543848.

Haoran Liu, Xiaoli Lin, Jing Hu, Yu Zou, and Xiaolong Zhang. Geometric pocket-centric protein encoding for polypharmacology-guided multi-target drug design. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DuaA7vJGF6.

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. URL https://proceedings.mlr.press/v162/liu22m.html.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in neural information processing systems*, 34:6229–6239, 2021.

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. URL https://proceedings.mlr.press/v162/peng22b.html.

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. URL https://proceedings.mlr.press/v235/qu24a.html.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical Science*, 13(9):2701–2713, 2022. doi: 10.1039/D1SC05976A.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L. Blundell, Pietro Liò, 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.

Jike Wang, Rui Qin, Mingyang Wang, Meijing Fang, Yangyang Zhang, Yuchen Zhu, Qun Su, Qiaolin Gou, Chao Shen, Odin Zhang, Zhenxing Wu, Dejun Jiang, Xujun Zhang, Huifeng Zhao,

Jingxuan Ge, Zhourui Wu, Yu Kang, Chang-Yu Hsieh, and Tingjun Hou. Token-Mol 1.0: Tokenized drug design with large language models. *Nature Communications*, 16:4416, 2025. doi: 10.1038/s41467-025-59628-y.

Kehan Wu, Yingce Xia, Pan Deng, Renhe Liu, Yuan Zhang, Han Guo, Yumeng Cui, Qizhi Pei, Lijun Wu, Shufang Xie, Si Chen, Xi Lu, Song Hu, Jinzhi Wu, Chi-Kin Chan, Shawn Chen, Liangliang Zhou, Nenghai Yu, Enhong Chen, Haiguang Liu, Jinjiang Guo, Tao Qin, and Tie-Yan Liu. TamGen: Drug design with target-aware molecule generation through a chemical language model. *Nature Communications*, 15:9360, 2024a. doi: 10.1038/s41467-024-53632-4.

Peng Wu, Huabin Du, Yingchao Yan, Tzong-Yi Lee, Chen Bai, and Song Wu. Guided diffusion for molecular generation with interaction prompt. *Briefings in Bioinformatics*, 25(3):bbae174, 2024b. doi: 10.1093/bib/bbae174.

Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z. Li. A systematic survey of chemical pre-trained models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6787–6795, 2023. doi: 10.24963/ijcai.2023/760.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Renjun Xu, Hongyang Chen, Xiaohui Fan, Huabin Xing, and Huajun Chen. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):161:1–161:38, 2025. doi: 10.1145/3715318.

Weitong Zhang, Xiaoyun Wang, Weili Nie, Joe Eaton, Brad Rees, and Quanquan Gu. MoleculeGPT: Instruction following large language models for molecular property prediction. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*, 2023a. URL `https://ai4d3.github.io/2023/papers/34.pdf`.

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. URL `https://proceedings.mlr.press/v202/zhang23z.html`.

Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations*, 2023b. URL `https://openreview.net/forum?id=Rq13idF0F73`.

Xinzhe Zheng, Shiyu Jiang, Gustavo Seabra, Chenglong Li, and Yanjun Li. Apo2Mol: 3d molecule generation via dynamic pocket-aware diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):1614–1622, 2026. doi: 10.1609/aaai.v40i2.37138.

Algorithm 1 Fragment Interaction-aware Pocket Residue Adapter (FIPRA)

Input: Generation state $\mathbf{h}_t$
Input: Candidate fragments $\mathcal{C}_t = \{f_t^k\}_{k=1}^K$ and embeddings $\{\mathbf{e}_t^k\}_{k=1}^K$
Input: Residue features $\{\mathbf{r}_i\}_{i=1}^L$ and mask $\mathbf{m}$
Input: Base logits $\{\ell_{t,k}\}_{k=1}^K$
Input: Interaction targets $\mathbf{y}_t$ and teacher-forced index $k_t^\star$ during training
Output: Candidate distribution $\{p(f_t^k \mid F_{<t}, P)\}_{k=1}^K$

for each residue i with $m_i = 1$ **do**
  $\mathbf{k}_i \leftarrow W_k \mathbf{r}_i$
  $\delta_i \leftarrow \mathbf{w}_\delta^\top \mathbf{r}_i$
end for
for $k = 1, \ldots, K$ **do**
  $\mathbf{q}_t^k \leftarrow W_q[\mathbf{h}_t \| \mathbf{e}_t^k]$
  for each residue i with $m_i = 1$ **do**
    $s_{tki} \leftarrow (\mathbf{q}_t^k)^\top \mathbf{k}_i / \sqrt{d}$
    $b_{tki} \leftarrow \sigma(s_{tki} + \delta_i)$
  end for
  $\{\alpha_{tki}\}_{i=1}^L \leftarrow \mathrm{MaskedSoftmax}_i\big(\{s_{tki}\}_{i=1}^L, \mathbf{m}\big)$
  $c_{t,k} \leftarrow \sum_{i=1}^L \alpha_{tki} b_{tki}$
end for
if training **then**
  $\mathbf{z}_t^\star \leftarrow \{s_{tk_t^\star i} + \delta_i\}_{i=1}^L$
  $\mathcal{L}_{\mathrm{FI}} \leftarrow \mathrm{WBCEWithLogits}(\mathbf{z}_t^\star, \mathbf{y}_t, \mathbf{m})$
end if
for $k = 1, \ldots, K$ **do**
  $\widetilde{\ell}_{t,k} \leftarrow \ell_{t,k} + \beta c_{t,k}$
end for
$\{p(f_t^k \mid F_{<t}, P)\}_{k=1}^K \leftarrow \mathrm{Softmax}_k\big(\{\widetilde{\ell}_{t,k}\}_{k=1}^K\big)$
return Candidate distribution and, during training, $\mathcal{L}_{\mathrm{FI}}$

A Detailed FIPRA Procedure

Algorithm 1 summarizes candidate-specific interaction prediction, contact supervision, and logit refinement in FIPRA.

B Extended Benchmark Interpretation

Table 1 reveals a complementary performance profile rather than uniform dominance on every metric. AFIMol obtains a Vina Score of -7.241 and High Affinity of 0.589. Several baselines have lower mean Vina scores, but only PMDM reports a higher High Affinity value among the methods with available entries. This distinction matters because the mean docking score and the fraction of molecules outperforming the corresponding reference ligand summarize different aspects of the generated population.

The clearest gains occur in molecular feasibility and coverage. AFIMol achieves the highest normalized SA (0.788) and Diversity (0.771), improving over the second-best reported values by 0.043 and 0.024, respectively. Its QED (0.596) and Lipinski score (4.959) remain within the range of the compared methods, but are not the best values. The result therefore supports a balance between docking-based performance, synthetic accessibility, and structural diversity, rather than a claim of universal superiority. Because some baseline entries are unavailable and published evaluation pipelines can differ, this table provides benchmark-level context; the controlled ablations provide the more direct evidence for the roles of SAF and FIPRA.

C Extended Ablation Analysis

Table 2 separates the roles of SAF and FIPRA beyond the compact main-text summary. Removing FIPRA changes Vina Score from -7.241 to -7.082 and High Affinity from 0.589 to 0.520, with a smaller decrease in SA from 0.788 to 0.771. In contrast, QED, Lipinski, and Diversity increase from 0.596, 4.959, and 0.771 to 0.625, 4.989, and 0.782, respectively. Thus, the effect of FIPRA

is concentrated on the pocket-directed metrics rather than uniformly improving every molecular descriptor. This pattern is consistent with candidate–residue guidance favoring a more pocket-specific subset of fragments, although the aggregate descriptors alone do not establish that mechanism.

Removing SAF produces a different profile. SA decreases from 0.788 to 0.765, Diversity from 0.771 to 0.753, and High Affinity from 0.589 to 0.523, whereas the mean Vina Score changes only from -7.241 to -7.226. The larger changes in SA and Diversity support the intended role of SAF in altering the fragment units available for assembly. The unequal response of Vina Score and High Affinity also shows why the mean score and the fraction of samples outperforming their reference ligand provide complementary views.

The joint ablation gives the least favorable Vina Score (-6.789), SA (0.671), and Diversity (0.711). Its High Affinity value (0.531) is below the full model but slightly above either single ablation, showing that the two modules are complementary without implying strictly additive effects. QED and Lipinski likewise vary non-monotonically. LogP changes from 3.003 for AFIMol to 2.663, 2.915, and 2.293 after removing SAF, FIPRA, and both modules; because no single target interval is specified, these shifts are interpreted as changes in molecular composition rather than improvements by themselves. Together, the ablations associate SAF most clearly with fragment quality and coverage, and FIPRA with pocket-facing selection, while also exposing the property trade-offs produced by these objectives. These comparisons concern the reported point estimates; repeated runs or per-pocket distributions would be required to quantify uncertainty in their magnitudes.

D Additional Mechanistic Analyses

The following diagnostics complement the outcome-level ablation by examining the representations produced by SAF and the pocket-level geometries associated with FIPRA. They are intended to test whether the modules alter the quantities they were designed to affect, rather than to serve as independent evidence of binding activity.

D.1 Fragment Granularity

Panels (a) and (b) of Fig. 2 characterize the most direct representational effect of SAF. Relative to rule-based fragmentation, SAF produces fewer fragments per ligand and shifts the token-weighted size distribution toward larger units. The learned boundary scores therefore induce less aggressive decomposition and preserve larger local substructures within individual fragment tokens. This shortens the fragment sequence presented to the autoregressive generator while retaining molecule-dependent organization; the following vocabulary and retention analyses quantify the accompanying changes in reuse and atomic coverage.

D.2 Vocabulary Distribution and Fragment Generalization

Beyond fragment granularity, we further analyze how SAF changes the statistical properties of the fragment vocabulary. Figure 2(c) shows that SAF induces a broader long-tail distribution than conventional fragmentation, reflecting the greater structural specificity of its fragments. The diagnostics in Fig. 2(d) further show higher test-time out-of-vocabulary rates, a larger fraction of rare training fragments, and greater train–test divergence. SAF therefore trades some vocabulary reuse for molecule-adaptive fragment units that preserve larger local substructures.

D.3 Atom Retention Analysis

We additionally examine the ligand heavy atoms retained in SAF fragment sequences. Figure 5(a) shows substantial variation across molecules, indicating selective reorganization rather than uniform atom recovery. The element-wise results in Fig. 5(b) are also chemically non-uniform. Carbon retention is $86.92\%/83.92\%$ on the training/test sets, while nitrogen and oxygen reach $74.75\%/71.01\%$ and $56.36\%/57.20\%$. Fluorine and chlorine are largely retained ($99.17\%/100.00\%$ and $96.54\%/90.91\%$), phosphorus has the lowest rate ($34.27\%/47.37\%$), and sulfur remains stable at $86.30\%/86.96\%$.

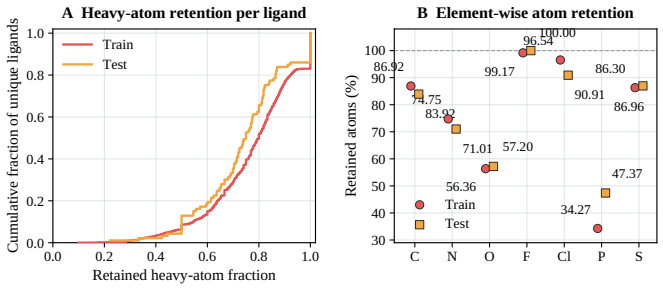


Figure 5: **Atom retention under SAF.** (a) Cumulative retained heavy-atom coverage on the training and test sets. (b) Element-wise retention for C, N, O, F, Cl, P, and S.

D.4 Extended FIPRA Geometry Analysis

Each point in Fig. 3 is a paired comparison for one test pocket: the horizontal coordinate is obtained without FIPRA and the vertical coordinate with the full model. This paired view separates the direction of the model change from the large baseline differences among pockets. The median increase of $+0.060$ nearby residues per ligand heavy atom, together with the $+0.047$ increase in C–C nearby-residue density, indicates that FIPRA is associated with denser local packing in the generated poses rather than only a change in autoregressive likelihood.

The polar descriptors provide a useful control on this interpretation. FIPRA increases the ligand N/O fraction by a median of $+0.063$, but N/O-associated pocket proximity increases by only $+0.007$. The geometric shift therefore cannot be explained solely by adding heteroatoms: whether those atoms occupy a nearby polar environment remains constrained by the three-dimensional pocket. These statistics measure proximity, not confirmed interaction types. In particular, C–C proximity does not by itself establish hydrophobic contacts, and N/O proximity does not establish hydrogen bonds. Their role is to connect the candidate-level guidance introduced by FIPRA with the observed population-level geometry of its generated ligands.

E Extended 4XLI Case Study

The 4XLI analysis complements the population-level diagnostics with candidate-resolved hypotheses. Its docking, physicochemical, and retrosynthetic results address distinct questions—pose plausibility, property profiling, and route-search accessibility—and are therefore interpreted separately below.

E.1 Residue-level Pose Interpretation

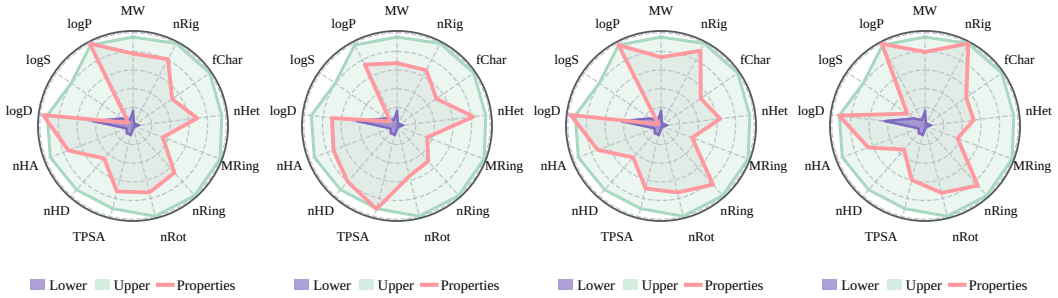
The four selected poses span only 0.85 kcal/mol, so their Vina ordering is not used as an affinity ranking. We instead use the poses to examine whether candidate-specific generation produces distinguishable residue environments within the same ATP-binding site. Table 4 summarizes the software-annotated atom-pair proximities used in this analysis.

All annotated distances lie between 2.9 and 3.3 Å. Three candidates approach gatekeeper Thr361, two approach hinge-region Met364, and three approach Asn368, supporting occupancy of recognizable ATP-site subregions. The complete patterns remain candidate-specific: 4XLI_01 approaches Thr361, His292, and Met364; 4XLI_02 approaches Met364 through one annotation and Asn368 through two distinct atom pairs; 4XLI_03 spans Thr361, Asn368, and DFG residue Asp427; and 4XLI_04 combines Thr361 and Asn368. Thus, the candidates share recurrent pocket landmarks without collapsing to a single residue-contact pattern.

This candidate-specific variation is consistent with the intended role of FIPRA: candidate fragments are evaluated against localized residue evidence rather than assigned a single pocket-level score. The number of annotated contacts does not explain the small Vina-score differences, since the docking objective depends on the full pose. Moreover, static distances do not establish contact persistence, kinase selectivity, or inhibition. The poses therefore provide interpretable structural hypotheses that

Table 4: Residue-level proximity annotations for the selected 4XLI docking poses. Distances are geometric measurements between software-annotated atom pairs and do not by themselves establish an interaction type or persistence. Repeated candidate–residue entries correspond to distinct atom pairs.

Contacts			Contacts		
Candidate	Residue	Distance (Å)	Candidate	Residue	Distance (Å)
4XLI_01	Thr361	3.0	4XLI_03	Thr361	3.2
4XLI_01	His292	2.9	4XLI_03	Asp427	3.3
4XLI_01	Met364	3.0	4XLI_03	Asn368	2.9
4XLI_02	Met364	3.1	4XLI_04	Thr361	3.1
4XLI_02	Asn368	3.1	4XLI_04	Asn368	3.2
4XLI_02	Asn368	2.9			



(a) 4XLI_01 **(b) 4XLI_02** **(c) 4XLI_03** **(d) 4XLI_04**

Figure 6: Normalized physicochemical descriptor profiles of the four generated candidates for the 4XLI target.

can guide subsequent dynamic or experimental evaluation, rather than biological validation on their own.

E.2 Physicochemical Profile Comparison

Figure 6 provides additional physicochemical characterization of the four generated molecules in the 4XLI case study. The normalized profiles compare molecular size, ring and bond statistics, hydrogen-bond capacity, polar surface area, lipophilicity, and predicted solubility. The complementary shapes show that the candidates occupy distinct property profiles despite targeting the same pocket, providing an additional view of the chemical diversity preserved during pocket-conditioned generation. Because the visualization is normalized, it does not replace the raw values, units, molecular-state definitions, or normalization rule. We therefore use it for comparative prioritization only and do not infer favorable ADME, toxicity, oral bioavailability, potency, or binding from the radar envelopes.

E.3 Retrosynthetic Route-search Analysis

Table 3 in the main paper reports short AiZynthFinder routes for all four candidates. Search time ranges from 3.17 to 4.52 seconds, and every top-ranked route terminates in stock-listed precursors within two to four transformations. Candidates 4XLI_01 and 4XLI_04 have the shortest top routes, whereas 4XLI_02 requires four transformations and has the lowest top score. Candidate 4XLI_04 yields 88 solved route instances among 110 enumerated instances (80.0%), the largest fraction in this search.

These outcomes add a practical medicinal-chemistry layer to the structural analysis: each docking hypothesis is paired with at least one concrete stock-terminated synthesis plan that can be inspected before experimental follow-up. However, enumerated solutions may share disconnections or precursor sets, so the solved fraction measures search-tree coverage rather than the number of independent synthetic strategies. The planner score ranks routes under the selected template library, stock, depth,

and iteration budget; it is not an estimate of experimental yield, reaction robustness, binding, or biological activity. Consequently, the route search establishes computational accessibility and candidate prioritization, while experimental feasibility still requires condition selection and laboratory validation.

E.4 Integrated Interpretation and Evidence Boundaries

The three parts of the case study address complementary questions rather than serving as interchangeable validation. Docking and residue-level inspection test whether the generated structures admit plausible ATP-pocket poses; the descriptor profiles compare their early chemical-property signatures; and the route search asks whether template-based planning can connect them to stocked precursors. Considered together, they turn four generated structures into concrete structural and synthesis hypotheses for follow-up while retaining the distinct evidential scope of each analysis.

None of these computational results establishes binding stability, ABL2 selectivity, pharmacokinetic behavior, laboratory yield, or biological activity. The narrow docking-score range also does not support a confident affinity ranking. Replicated docking and pose-recovery controls, molecular dynamics or free-energy calculations, reaction-condition assessment, experimental synthesis, and biochemical assays would be required to evaluate the structural, synthetic, and activity hypotheses directly.