SURFSPEC: ENHANCING OFF-TARGET-AGNOSTIC SPECIFICITY BY BOUNDING POCKET-LIGAND GEOMETRIC MISMATCH

Anonymous authors
Paper under double-blind review

ABSTRACT

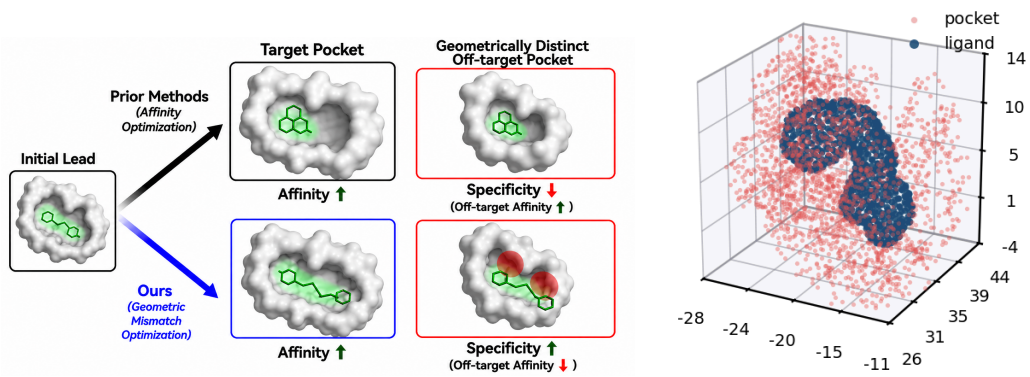
Lead optimization with generative AI for drug discovery aims to improve target binding while avoiding unintended interactions with off-target pockets. However, existing affinity-driven methods do not explicitly control specificity, whereas current specificity-aware approaches commonly require prior knowledge of off-target structures. We address off-target-agnostic specificity-aware lead optimization by analyzing the geometric mismatch between a ligand and the target pocket. Under separation and calibration assumptions, we derive a conservative specificity lower bound for an off-target-agnostic optimization setting. By metricizing pocket–ligand mismatch, the triangle inequality shows that reducing target–ligand mismatch weakly improves a conservative lower bound on mismatch to a separated off-target class, which can be translated into a specificity lower bound through an empirical geometry–affinity calibration. Motivated by this analysis, we introduce **SurfSpec**, an off-target-agnostic lead optimization framework that iteratively grows ligands toward under-occupied regions of the target pocket surface. SurfSpec alternates between linker generation toward selected target-surface patches, which provides geometric pseudo-labels, and refinement under a pocket-conditioned ligand prior to recover chemically valid ligands. On the CrossDocked2020 test set, SurfSpec reduces geometric mismatch and outperforms evaluated off-target-agnostic lead optimization baselines in empirical specificity, while maintaining competitive target-affinity improvement.

1 INTRODUCTION

Lead optimization in structure-based drug design (SBDD) seeks to improve an initial ligand for a target pocket. Beyond increasing target-pocket affinity, an ideal optimized ligand should also avoid strong interactions with undesirable off-target pockets, thereby achieving ligand specificity. Many existing lead optimization methods Huang et al. (2024); Qiao et al. (2025); Chen et al. (2025); Zhou et al. (2024), however, primarily optimize affinity to the target pocket. As highlighted by recent work Gao et al. (2024), high target affinity alone does not guarantee specificity, because an optimized ligand may also bind strongly to unintended pockets. Such off-target interactions can reduce selectivity and cause adverse effects Keiser et al. (2009), motivating specificity-aware generative methods for more reliable early-stage drug discovery.

Recent specificity-aware methods address this limitation using explicit negative supervision, either through off-target pocket information or activity labels for undesired targets. SBE-Diff Gao et al. (2024), for example, learns specific binding energy from positive and in-batch negative pocket–ligand pairs, whereas ActivityDiff Zhou et al. (2025) employs separately trained drug–target activity classifiers for desired target activity and undesired off-target activity. Although these methods demonstrate the value of incorporating specificity into molecular optimization, their applicability depends on the availability and representativeness of negative target information. In practice, potential off-target interactions form an open-ended set that may include unknown, structurally unavailable, or out-of-distribution targets. Larger surrogate negative sets and stronger activity predictors may improve coverage of observed interactions, but their performance on unseen or structurally unavailable off-targets remains uncertain. This motivates an off-target-agnostic formulation of specificity-aware lead optimization, in which only the target pocket is used during optimization.

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.



(a) Specificity principle

(b) Surface representation

Figure 1: Target-only specificity optimization and its surface representation. (a) Surface-directed growth seeks target compatibility while reducing compatibility with geometrically distinct off-targets. (b) Ligand (blue) and ligand-oriented pocket (red) surface samples. Their induced probability measures define geometric mismatch; Figure 3 shows the corresponding densities.

Instead of relying on off-target supervision, we seek a target-only geometric principle that can be analyzed without accessing off-targets during optimization. We define geometric mismatch between a ligand and a pocket using the Jensen–Shannon distance between surface-induced probability measures. Because this mismatch is metricized, the triangle inequality provides a conservative lower bound on ligand mismatch conditional on the geometric separation of an off-target class. Assuming that an empirical geometry–affinity calibration transfers to the evaluation pairs, this yields a specificity lower-bound analysis for separated off-target classes.

Motivated by this analysis, we propose **SurfSpec**, an off-target-agnostic lead optimization framework that progressively expands an initial ligand toward under-occupied regions of the target pocket surface. SurfSpec alternates between surface-directed pseudo-label generation and molecular-prior-based recovery. At each iteration, it identifies a nearby target-surface patch that is insufficiently covered by the current ligand and constructs a pseudo-label by extending the ligand toward the selected region. A pretrained pocket-conditioned diffusion prior then refines the pseudo-label while balancing validity and geometric faithfulness. Through this iterative procedure, SurfSpec improves coverage of accessible target-pocket regions and encourages target-pocket surface complementarity while maintaining consistency with the learned molecular distribution.

To keep pseudo-label recovery simple and stable within the ligand-growth pipeline, SurfSpec uses a task-specific low-noise recovery procedure. General inverse-problem solvers can be adapted to this task through clean-state posterior approximations Chung et al. (2023); Kim et al. (2025) or repeated consistency and prior updates Zhang et al. (2025); Li et al. (2024); Mardani et al. (2024). These procedures introduce approximation error or sequential correction steps. SurfSpec is motivated by the localized nature of the anchored recovery task: the desired ligand is expected to remain close to the geometric pseudo-label while lying on the highly structured pocket-conditioned ligand distribution, making the corresponding anchored distribution locally concentrated. Based on this observation, SurfSpec performs gradient-flow mode recovery for a localized pseudo-label-conditioned distribution at a single intermediate noise level, re-noises around the recovered mode, and completes sampling with a low-noise unguided reverse SDE. This design avoids high-noise anchor-conditioned estimation and reduces dependence on repeated correction stages, serving as a practical recovery component for surface-directed ligand growth rather than a new general-purpose refinement principle.

On the CrossDocked2020 test set Francoeur et al. (2020), SurfSpec achieves higher empirical specificity than evaluated off-target-agnostic lead optimization baselines while maintaining competitive target-pocket affinity, as shown in Table 1. It also attains higher empirical specificity than ActivityDiff under the evaluated surrogate off-target configuration. Further analyses show that SurfSpec improves pocket coverage and geometric fit, with low clash rates and a tradeoff between pseudo-label validity and faithfulness.

Our main contributions are as follows:

- We formulate off-target-agnostic specificity-aware lead optimization through a target-only geometric mismatch metric, defined as the Jensen–Shannon distance between probability measures induced by the ligand surface and the ligand-oriented target-pocket surface. Under a geometric separation condition on the off-target class, the triangle inequality shows that reducing target–ligand mismatch weakly improves a conservative lower bound on ligand mismatch to separated off-targets, which yields a specificity lower bound under an assumed transferable geometry–affinity calibration.
- We introduce **SurfSpec**, an off-target-agnostic lead optimization framework motivated by this target-only geometric principle. SurfSpec iteratively grows ligands toward under-occupied regions of the target pocket surface by alternating linker generation toward selected surface patches and task-specific low-noise anchored recovery under a pocket-conditioned ligand prior.
- We evaluate SurfSpec against off-target-agnostic lead optimization baselines using held-out off-target pockets that are never accessed during optimization. On CrossDocked2020, SurfSpec achieves higher empirical specificity while maintaining competitive target-pocket docking performance, with improved pocket coverage and geometric fit.

2 Related Work

Lead Optimization and Molecular Specificity. Recent SBDD models support pocket-conditioned molecular generation and lead optimization through autoregressive generation, equivariant diffusion models, and controllable molecular optimization Peng et al. (2022); Guan et al. (2023); Gu et al. (2024); Schneuing et al. (2024); Huang et al. (2024); Zhou et al. (2024); Chen et al. (2025). Methods such as PMDM, DecompOpt, and Delete Huang et al. (2024); Zhou et al. (2024); Chen et al. (2025) optimize an initial ligand or ligand fragment for target-pocket affinity, drug-likeness, and validity, but do not explicitly optimize molecular specificity. More recent approaches, including SBE-Diff and ActivityDiff Gao et al. (2024); Zhou et al. (2025), incorporate negative pockets or activity signals to improve specificity through off-target supervision. SurfSpec studies specificity-aware optimization using target geometry without off-target supervision.

Geometric Complementarity. Classical docking methods Kuntz et al. (1982); Shoichet & Kuntz (1993) exploit pocket–ligand surface complementarity to identify favorable poses. Protein-pocket-conditioned SBDD models Peng et al. (2022); Guan et al. (2023); Schneuing et al. (2024); Zhou et al. (2024); Huang et al. (2024); Lin et al. (2023) use target pocket geometry for molecular generation, elaboration, or optimization. Other geometry-conditioned methods use reference molecular shapes, multimodal ligand profiles, or learned geometric representations, as in Diff-Shape Li et al. (2026a), ShEPhERD Adams et al. (2025), and GeoRCG Li et al. (2025), while UniGuide Ayadi et al. (2024) applies user-specified geometric conditions through training-free inference-time guidance. However, these approaches do not explicitly formulate pocket–ligand surface complementarity as a metricized mismatch objective linked to molecular specificity. SurfSpec uses this mismatch to connect geometric complementarity to molecular specificity optimization.

Training-Free Sample Refinement. Training-free refinement methods use several overlapping strategies. *Reverse-path* methods, including SDEdit, DDNM, MCG, DPS, ΠGDM, FlowChef, and FlowDPS Meng et al. (2022); Wang et al. (2023); Chung et al. (2022, 2023); Song et al. (2023); Patel et al. (2025); Kim et al. (2025), modify samples along a reverse SDE/ODE trajectory. SDEdit Meng et al. (2022) constructs this state by directly perturbing an input anchor, whereas conditional reverse-path methods Chung et al. (2022, 2023); Song et al. (2023); Kim et al. (2025) incorporate observations or constraints through guidance terms computed from denoised estimates. *Optimization-based* methods, such as DiffPIR, DDS, DAPS, DCDP, PnP-Flow, FLOWER, FlowLPS, and FAST-DIPS Zhu et al. (2023); Chung et al. (2024); Zhang et al. (2025); Li et al. (2024); Martin et al. (2025); Pourya et al. (2026); Park & Ye (2025); Kim et al. (2026), interleave prior-based denoising with explicit consistency-oriented optimization, projection, or proximal correction. *Variational posterior* methods, including RED-Diff, RSD, and FLAIR Mardani et al. (2024); Zilberstein et al. (2025); Erbach et al. (2025), instead optimize a parameterized variational distribution toward the target

posterior. These approaches provide general and flexible mechanisms for training-free refinement across inverse problems and continuous data domains. ACE Lee et al. (2026) studies the validity of composed diffusion-steering paths. SurfSpec adopts a task-specific refinement strategy tailored to geometrically constructed molecular pseudo-labels: it performs anchored recovery at a low noise level under a pocket-conditioned molecular prior, reducing reliance on high-noise clean-state posterior approximation or multi-stage posterior optimization.

3 PRELIMINARIES

3.1 PROBLEM SETUP: SPECIFIC LEAD OPTIMIZATION WITHOUT OFF-TARGET ACCESS

We consider lead optimization for a target pocket P_{tgt} when off-target information is unavailable. A ligand is represented as $L = (X_L, H_L)$, where $X_L \in \mathbb{R}^{3\times N_L}$ and $H_L \in \{0,1\}^{n_{\mathrm{types}}\times N_L}$ denote its coordinates and atom-type features, respectively. A pocket $P = (X_P, H_P)$ is defined analogously. For the methodology derivation in Section 4, we use a rigid-body abstraction and identify ligands and pockets up to separate rigid transformations: $L \sim \rho_g(L)$ and $P \sim \rho_h(P)$ for $g, h \in \mathrm{SE}(3)$.

Given the target pocket P_{tgt} and a set of possible off-target pockets $\mathcal{O}$, we define specificity as the affinity gap between the target and the strongest off-target:

$$\mathrm{Spec}(L; P_{\mathrm{tgt}}, \mathcal{O}) = A_{P_{\mathrm{tgt}}}(L) - \max_{O\in\mathcal{O}} A_O(L), \tag{1}$$

where $A_P(L)$ denotes the affinity between ligand L and pocket P, with larger values indicating stronger binding. In the methodology derivation, we instantiate $A_P(L)$ under the rigid-body abstraction as $\max_{g\in\mathrm{SE}(3)} a(\rho_g(P), L)$, where a is the pose-level affinity. In experiments, we evaluate affinity using the negative AutoDock Vina score Trott & Olson (2010), allowing ligand torsional changes during docking. A ligand is therefore considered specific when it achieves strong target-pocket affinity without exhibiting similarly strong affinity for competing off-target pockets.

Specific Lead Optimization without Off-Target Access. Directly optimizing Eq. (1) is generally impractical because the true off-target set $\mathcal{O}$ is unavailable during lead optimization. Existing methods therefore either optimize target-pocket affinity alone or approximate the off-target set using a predefined collection of pockets. Methods that optimize target-pocket affinity alone often yield limited specificity improvement, because increasing target-pocket affinity can also increase off-target affinity Gao et al. (2024). Methods that rely on a predefined off-target set Gao et al. (2024); Zhou et al. (2025) explicitly account for specificity, but may fail to capture challenging off-targets outside the constructed set. We therefore seek a target-only optimization principle that improves specificity without requiring access to off-target pockets, where the guarantees apply to off-target pockets that are geometrically distinguishable from the target pocket.

3.2 JENSEN–SHANNON DISTANCE AND METRIC STRUCTURE

The Jensen–Shannon distance measures the discrepancy between two probability distributions through a symmetrized and smoothed KL divergence. For probability measures μ and ν defined on a common domain Ω, let

$$\bar{\mu} = \frac{1}{2}(\mu + \nu).$$

The Jensen–Shannon distance is defined as

$$d_{\mathrm{JS}}(\mu,\nu) = \left[\frac{1}{2}\,\mathrm{KL}(\mu\|\bar{\mu}) + \frac{1}{2}\,\mathrm{KL}(\nu\|\bar{\mu})\right]^{1/2}. \tag{2}$$

The square-root Jensen–Shannon divergence is a metric on probability distributions Endres & Schindelin (2003) and satisfies the triangle inequality:

$$d_{\mathrm{JS}}(\mu,\rho) \le d_{\mathrm{JS}}(\mu,\nu) + d_{\mathrm{JS}}(\nu,\rho). \tag{3}$$

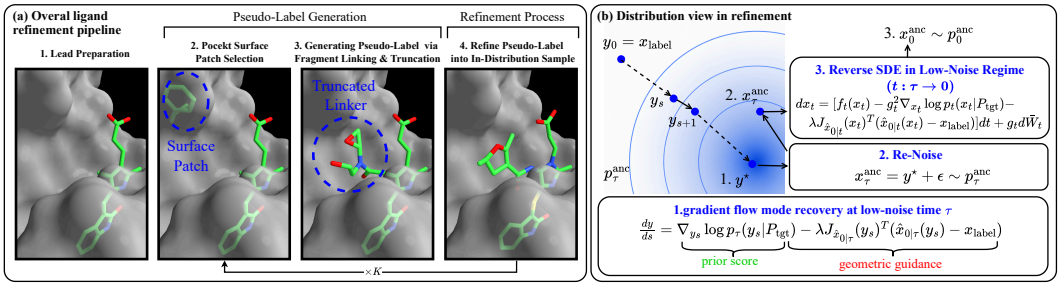


Figure 2: Overview of the ligand refinement pipeline. The left panel illustrates iterative pocket-surface-targeted ligand growth, while the right panel presents the prior-guided refinement process that maps an out-of-distribution pseudo-label toward an in-distribution ligand sample.

4 METHODOLOGY

4.1 GEOMETRIC MISMATCH BOUNDS SPECIFICITY AGAINST GEOMETRICALLY DISTINCT OFF-TARGETS

We first analyze metricized geometric mismatch between the target pocket and the ligand as a target-only principle for specificity-aware lead optimization. The key intuition is that a ligand with small mismatch to the target pocket is encouraged to fit the native pocket, while its compatibility with geometrically distinct off-target pockets can be conservatively bounded through the metric structure. This objective is not intended to replace affinity optimization, but rather to optimize a geometric component that contributes to molecular recognition and favorable binding. Motivated by classical geometric docking and shape-complementarity studies Kuntz et al. (1982); Shoichet & Kuntz (1993), we formalize this principle by defining a computable mismatch metric, examining its empirical relationship with docking affinity, and deriving a conditional specificity bound.

Definition of geometric mismatch. We first express each pocket in the ligand-centered affinity-optimal frame. For ligand L and pocket P, define $P^L = \rho_{g^\star(L,P)}(P)$, where $g^\star(L,P) \in \arg\max_{g\in \mathrm{SE}(3)} a(\rho_g(P), L)$. Let $A_L = \{a_i\}_{i=1}^{N_L}$ denote the heavy-atom coordinates of L, and define the ligand-local domain $\Omega_L = \{x \in \mathbb{R}^3 : \min_i \|x - a_i\|_2 \le 8\,\text{Å}\}$. We define the ligand surface $\mathcal{S}_L$ as the van der Waals surface of the ligand heavy atoms. For the aligned pocket P^L, we define the ligand-oriented pocket surface $\mathcal{S}_P^L$ as the van der Waals surface of P^L restricted to surface points within 8 Å of any ligand heavy atom.

To compare these surfaces, we convert $\mathcal{S}_L$ and $\mathcal{S}_P^L$ into probability measures on the common domain Ω_L. Let μ_L and ν_P^L denote the probability measures induced by $\mathcal{S}_L$ and $\mathcal{S}_P^L$, respectively, through the signed-distance-based Boltzmann densities defined in Appendix B.1. We define geometric mismatch as the Jensen–Shannon (JS) distance between the two measures:

$$\begin{aligned} d_{\mathrm{gm}}(L,P) &= d_{\mathrm{JS}}(\mu_L, \nu_P^L) \\ &= \left[\frac{1}{2}\,\mathrm{KL}(\mu_L\|\bar{\mu}) + \frac{1}{2}\,\mathrm{KL}(\nu_P^L\|\bar{\mu})\right]^{1/2}, \end{aligned} \tag{4}$$

where $\bar{\mu} = (\mu_L + \nu_P^L)/2$. A smaller $d_{\mathrm{gm}}(L,P)$ indicates stronger agreement between the ligand surface field and the ligand-oriented pocket surface field. Figure 1b illustrates the surfaces; the induced densities are shown in Appendix B.1.

Correlation with affinity. We formalize the empirical relationship between geometric mismatch and binding affinity through a quantile-calibrated geometry–affinity envelope. The empirical envelope allows residual variation beyond geometric mismatch; its probabilistic interpretation requires the coverage assumption below.

Assumption 1 (Quantile-calibrated geometry–affinity envelope). *Let $\mathcal{D}_{\mathrm{CD}}$ denote the empirical ligand–pocket distribution induced by CrossDocked2020. For a tail probability $\eta \in (0,1)$, there exist monotone decreasing functions f_-^η and f_+^η such that, for $(L,P) \sim \mathcal{D}_{\mathrm{CD}}$,*

$$\Pr\Big[f_-^\eta(d_{\mathrm{gm}}(L,P)) \le A_P(L) \le f_+^\eta(d_{\mathrm{gm}}(L,P))\Big] \ge 1-\eta. \tag{5}$$

When applying the certificate to optimized ligands, we additionally assume that the same $1-\eta$ envelope coverage holds marginally for the target pair and each considered off-target pair under the joint distribution induced by the optimization and evaluation protocol, with a fixed finite number of off-targets. This coverage transfer is assumed, not established by the CrossDocked2020 calibration experiment.

The fitted 90%, 95%, and 99% envelopes cover 91.0%, 92.0%, and 97.0% of the 100 CrossDocked2020 test samples, respectively; the calibration procedure and full results are provided in Appendix D.1.

Specificity certificate for geometrically separated off-targets. We analyze off-target pockets that remain geometrically distinguishable from the target pocket under ligand-oriented surface representations. Let $\mathcal{L}_{\mathrm{opt}}$ denote the set of ligands reachable by optimization. For $L \in \mathcal{L}_{\mathrm{opt}}$ and $O \in \mathcal{O}$, let $m_L(O) = d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O)$ denote the ligand-oriented geometric separation between the target and off-target pocket surfaces, where $\nu^L_{P_{\mathrm{tgt}}}$ and ν^L_O are the ligand-oriented pocket measures induced by the corresponding cropped pocket surfaces. We call $\mathcal{O}$ a δ-separated off-target class if

$$\inf_{L\in\mathcal{L}_{\mathrm{opt}}} \inf_{O\in\mathcal{O}} m_L(O) \ge \delta.$$

This condition requires every considered off-target pocket to remain at least δ away from the target pocket in the ligand-oriented metric throughout optimization.

Theorem 1 (Specificity certificate for separated off-targets). *Let $L \in \mathcal{L}_{\mathrm{opt}}$ be the output of the optimization protocol, and let $\epsilon = d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})$. Suppose that $\mathcal{O}$ is a δ-separated off-target class. Under Assumption 1, including its coverage-transfer condition, with probability at least $1-(|\mathcal{O}|+1)\eta$ over the joint optimization and evaluation protocol,*

$$\mathrm{Spec}(L; P_{\mathrm{tgt}}, \mathcal{O}) \ge f_-^\eta(\epsilon) - f_+^\eta([\delta-\epsilon]_+)\,. \tag{6}$$

Consequently, this high-probability specificity lower bound weakly increases as the target mismatch ϵ decreases.

The theorem formalizes the target-only role of geometric mismatch. Although off-target pockets are not accessed during optimization, their geometric separation from the target pocket provides a specificity certificate: as the ligand better matches the target surface, the certified lower bound on its separation from geometrically distinct off-targets weakly increases. The proof applies the triangle inequality of d_{JS} to the ligand-induced measure, the target-pocket measure, and each off-target pocket measure, yielding $d_{\mathrm{gm}}(L,O) \ge [\delta-\epsilon]_+$. The quantile-calibrated geometry–affinity envelope then converts this geometric certificate into an affinity-gap certificate, where the target affinity is lower bounded by $f_-^\eta(\epsilon)$ and each off-target affinity is upper bounded by $f_+^\eta([\delta-\epsilon]_+)$. A union bound over the target and off-target pockets gives the stated high-probability lower bound. The complete proof is provided in Appendix B.2.

4.2 Ligand Growth via Geometric Pseudo-Labeling and Refinement

Motivated by the target-only specificity analysis, we design SurfSpec as an iterative ligand-growth pipeline that improves target-pocket surface fit without accessing off-target pockets during optimization. SurfSpec uses geometric mismatch as a practical design principle and operationalizes it through surface-directed ligand growth toward under-occupied regions of the target pocket.

As illustrated in Figure 2, SurfSpec alternates between two stages. First, it generates a geometric pseudo-label by extending the current ligand toward a nearby under-occupied target-pocket surface region. Second, it refines this pseudo-label into a valid ligand using a pretrained diffusion prior over

Table 1: Lead optimization on CrossDocked2020 with 10 off-targets per target and an 1800 s docking timeout. Blue columns highlight specificity, the primary metric. Specificity and Vina scores are in kcal/mol. Bold and underline indicate best and second-best values, including ties; atom count is descriptive. Full statistics are in Appendix D.4.

Method	Off-target agnostic	**Primary metric:** $\widehat{\text{Spec}}\uparrow$ Mean ± SD	>0.2	>0.4	>0.6	Geom. mismatch ↓	Occ. ↑	Target Vina ↓	Clash rate ↓	# Atoms	Dock success ↑
Initial Lead	O	<u>-0.81</u> ± 1.43	0.15	0.14	0.11	0.540	0.26	-7.17	**0.00**	22.75	**1.00**
Delete	O	<u>-0.81</u> ± 1.54	<u>0.19</u>	0.14	0.13	0.539	0.26	-7.23	**0.00**	22.85	**1.00**
PMDM	O	-1.10 ± 2.03	0.15	0.12	0.12	0.501	<u>0.33</u>	-6.47	0.16	31.00	0.92
DecompOpt	O	-0.93 ± 1.43	0.13	0.12	0.11	0.542	0.25	-6.08	<u>0.02</u>	22.75	0.91
DiffShape	O	-0.88 ± 1.58	0.13	0.13	0.12	0.585	0.14	-6.90	0.11	21.91	<u>0.99</u>
ActivityDiff	X	-0.96 ± 2.15	0.13	0.11	0.10	0.538	0.27	-7.07	0.11	22.89	**1.00**
PMDM–MS2	O	-1.38 ± 2.19	0.13	0.13	0.12	0.499	0.31	**-8.95**	0.25	33.42	0.91
PMDM–MS3	O	-2.05 ± 2.93	0.07	0.07	0.06	0.496	0.31	-8.38	0.31	35.83	0.82
PMDM–MS+ES	O	-1.02 ± 2.08	<u>0.19</u>	<u>0.17</u>	<u>0.14</u>	0.500	<u>0.33</u>	-8.61	0.16	31.37	0.87
DecompOpt–MS2	O	-1.03 ± 1.33	0.11	0.11	0.08	0.542	0.25	-7.08	0.03	22.75	0.96
DecompOpt–MS3	O	-0.99 ± 1.39	0.12	0.09	0.08	0.541	0.25	-7.00	0.04	22.76	0.97
DecompOpt–MS+ES	O	-0.83 ± 1.59	0.17	0.14	<u>0.14</u>	0.541	0.25	-7.40	0.03	22.75	0.95
DiffSBDD–SizeExt(+20)	O	-1.51 ± 2.24	0.14	0.11	0.10	<u>0.480</u>	0.28	-7.57	0.45	35.88	0.93
DiffSBDD–SizeExt(+30)	O	-2.20 ± 3.76	0.08	0.07	0.07	**0.468**	0.32	-7.14	0.62	41.58	0.81
Ours	O	**-0.37** ± 2.54	**0.31**	**0.29**	**0.25**	<u>0.480</u>	**0.36**	<u>-8.63</u>	**0.00**	34.99	**1.00**

the pocket-conditioned ligand distribution. Repeating these stages encourages coverage of accessible target-pocket regions while maintaining consistency with the learned ligand distribution. The full procedure is described in Algorithm 1.

Geometric pseudo-label generation. We generate a pseudo-label by exploiting a pretrained linker generation model p_{linker}. At iteration k, we select the closest unoccupied surface patch $r^{(k)}$ whose distance from the current ligand $x^{(k)}$ falls within 4–10 Å. The detailed filtering rule for unoccupied surface patches is provided in Appendix C.2. Given $r^{(k)}$, we sample a linker that connects the current ligand and the selected surface patch. The sampled linker is truncated to exclude pocket-clashing atoms falling within the 2 Å clashing threshold, forming a geometric pseudo-label $x_{\text{label}}^{(k)}$. The resulting $x_{\text{label}}^{(k)}$ serves as a directed anchor for reducing the target pocket mismatch.

Pseudo-label refinement. Given the pseudo-label x_{label}, the refinement task is to recover a valid ligand with the same number of atoms that remains close to x_{label} while staying likely under the pocket-conditioned ligand prior. Formally, we target the anchored clean density

$$p_0^{\text{anc}}(x \mid P_{\text{tgt}}, x_{\text{label}}) \propto p_0(x \mid P_{\text{tgt}}) \exp\left(-\frac{\lambda}{2}\|x - x_{\text{label}}\|^2\right), \tag{7}$$

where $p_0(x \mid P_{\text{tgt}})$ is the pretrained ligand prior and $\lambda > 0$ controls anchor fidelity. Let p_t^{anc} denote the distribution obtained by diffusing p_0^{anc} to time t.

General-purpose inverse-problem solvers discussed in Section 2 can be applied to this refinement task, but their design choices are not tailored to geometrically constructed molecular pseudo-labels. Optimization-based and variational posterior methods often introduce multiple coupled correction or optimization stages, which require careful coordination of interaction hyperparameters in this molecular setting. Reverse-path methods avoid such multi-stage optimization, but typically depend on either denoised clean-state approximations, such as the DPS plug-in approximation $p(x_0 \mid y, P_{\text{tgt}}, \tau) \approx \delta_{\hat{x}_{0|\tau}(y)}(x_0)$, or direct noising of the pseudo-label as in SDEdit. These approximations can become unreliable when applied at high noise levels or when the geometrically constructed pseudo-label lies outside the typical support of the pretrained molecular prior.

We therefore use a simpler task-specific low-noise recovery procedure that avoids multi-stage posterior optimization and avoids applying clean-state posterior approximation in the high-noise regime. The procedure is motivated by the local concentration of the anchored distribution under the combined

constraints of pocket-conditioned validity and similarity to the pseudo-label. In Appendix B.3, we formalize this intuition by assuming that the anchored clean distribution is concentrated in a small ball around its mean. Under this condition, at a low-noise time τ where the local DPS score approximation in Assumption 3 holds, Proposition 2 shows that the anchored score flow converges to a neighborhood of the mode of p_τ^{anc}, with residual error controlled by the implemented score approximation error. We therefore estimate this low-noise anchored point using the anchored score flow

$$\begin{aligned}\frac{dy_s}{ds} = &\nabla_{y_s} \log p_\tau(y_s \mid P_{\mathrm{tgt}}) \\ &- \lambda J_{\hat{x}_{0|\tau}}(y_s)^\top \left(\hat{x}_{0|\tau}(y_s) - x_{\mathrm{label}}\right).\end{aligned} \tag{8}$$

Let $y^\star$ denote the recovered point. Proposition 4 further shows that re-noising around this recovered point,

$$x_\tau^{\mathrm{anc}} = y^\star + \sigma_\tau \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{9}$$

approximates sampling from the low-noise anchored marginal p_τ^{anc}. Starting from x_τ^{anc}, the idealized anchored formulation uses a low-noise guided reverse SDE:

$$\begin{aligned}dx_t = \Big[f_t(x_t) - g_t^2\{&\nabla_{x_t} \log p_t(x_t \mid P_{\mathrm{tgt}}) \\ &- \lambda J_{\hat{x}_{0|t}}(x_t)^\top \left(\hat{x}_{0|t}(x_t) - x_{\mathrm{label}}\right)\}\Big] dt + g_t \, d\bar{W}_t,\end{aligned} \tag{10}$$

where f_t and g_t denote the drift and diffusion coefficients of the forward SDE defining the pretrained pocket-conditioned ligand prior $p_t(\cdot \mid P_{\mathrm{tgt}})$, and $\bar{W}_t$ is the reverse-time Brownian motion. In the experiments, we use the unguided reverse variant described in Appendix C.1. The recovery and initialization analysis in Appendix B.3 does not establish anchored-posterior accuracy for this practical variant.

The refined ligand is passed to the next growth iteration, and the pipeline alternates between geometric expansion and refinement until the maximum number of iterations is reached or the target-pocket Vina score after local optimization increases relative to the previous iteration.

5 Experiments

5.1 Experimental Setup

Benchmark and Implementation. We evaluate SurfSpec on the 100 test complexes from CrossDocked2020 Francoeur et al. (2020). For each complex, the reference ligand paired with the target pocket is used as the initial lead, and each method generates one optimized molecule. SurfSpec runs at most $K = 3$ growth iterations, using DiffLinker Igashov et al. (2024) for surface-directed pseudo-label generation and the pretrained DiffSBDD Schneuing et al. (2024) prior for refinement, with low-noise recovery time $\tau = 0.15$. Off-target pockets are never accessed by off-target-agnostic methods during optimization; for evaluation, each target is paired with an off-target set $\mathcal{O}_{\mathrm{eval}}$ of 10 randomly sampled pockets from the remaining CrossDocked2020 test pockets. The primary evaluation uses an 1800 s timeout per AutoDock Vina run.

Baselines and Metrics. Baselines comprise Delete, PMDM, DecompOpt, DiffShape, ActivityDiff, multi-step PMDM/DecompOpt variants, and DiffSBDD-based size-extension baselines Chen et al. (2025); Huang et al. (2024); Zhou et al. (2024); Li et al. (2026a); Zhou et al. (2025). We evaluate empirical specificity using target and off-target AutoDock Vina scores, and additionally report target-pocket Vina score, target-pocket geometric mismatch, pocket occupancy, clash rate, ligand size, and docking success rate. Full protocols and metric definitions are in Appendices C.1 and C.3.

5.2 Performance on Specificity Enhancement

Table 1 reports the primary results. SurfSpec achieves the highest mean empirical specificity (-0.37 versus -0.81 for the initial lead), although the mean remains negative. Its success rates at the three thresholds are also highest: 0.31, 0.29, and 0.25, versus 0.15, 0.14, and 0.11 for the initial lead. SurfSpec combines the highest mean occupancy (0.36) with low geometric mismatch (0.480),

Table 2: Pseudo-label refinement on CrossDocked2020 ($K = 1$). Affinity is the negative Vina score (kcal/mol; higher is better). Repulsion and affinity are means; RMSD (Å) and topological similarity also include Q1/Q3. Bold/underline mark best/second-best results (ties included; input excluded).

Method	Prior Consistency									Pseudo-Label Faithfulness					
	Pocket Clash ↓			Affinity ↑		Validity ↑	Bond MMD ↓			RMSD ↓			Topo. Sim. ↑		
	Rate	Local rep.	Global rep.	Local	Global		C-C	C-N	C-O	Mean	Q1	Q3	Mean	Q1	Q3
Input	0.00	1.46	7.61	0.23	6.75	0.84	0.06	0.01	0.09	0.00	0.00	0.00	1.00	1.00	1.00
DCDP	<u>0.02</u>	0.83	5.38	<u>0.43</u>	<u>7.69</u>	**1.00**	**0.08**	**0.03**	<u>0.10</u>	<u>0.62</u>	<u>0.55</u>	<u>0.67</u>	0.48	0.35	0.62
DPS	0.03	**0.20**	<u>4.88</u>	0.39	6.49	**1.00**	0.11	<u>0.04</u>	0.11	6.17	5.44	6.79	0.35	0.25	0.46
SDEdit	0.03	1.29	6.31	0.28	7.13	**1.00**	<u>0.09</u>	**0.03**	0.13	**0.31**	**0.28**	**0.35**	0.53	0.36	<u>0.68</u>
RED-Diff	0.51	5.04	42.39	0.14	1.67	**1.00**	0.12	0.05	0.16	1.47	0.97	2.00	**0.95**	**0.99**	**1.00**
RSD	0.77	9.54	106.75	0.12	0.69	**1.00**	0.11	<u>0.04</u>	0.15	2.77	1.98	3.62	<u>0.93</u>	<u>0.91</u>	**1.00**
Ours	**0.01**	<u>0.33</u>	**4.40**	**0.46**	**8.00**	**1.00**	**0.08**	**0.03**	**0.08**	0.78	0.66	0.90	0.48	0.34	0.64

competitive target Vina score (-8.63 kcal/mol), zero clash rate, and full docking success. With a second random seed, SurfSpec also has the highest mean specificity and success rates (Appendix D.4).

Multi-step and early-stopping variants of PMDM and DecompOpt do not match SurfSpec's specificity, suggesting that repeated baseline application alone does not explain the gain. The DiffSBDD size-extension baselines achieve low geometric mismatch but high clash rates and weak specificity, showing that naive size growth can produce invalid pocket occupation. Together, these comparisons support surface-directed growth as a way to improve target-pocket coverage and empirical target-over-off-target preference without accessing off-targets during optimization.

Activation of the geometric lower bound. In Appendix D.3, we evaluate the certificate margin $M_L(O) = d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O) - d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})$, where $d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O)$ is the ligand-oriented target–off-target pocket separation. When $M_L(O) > 0$, the triangle-inequality lower bound on ligand–off-target mismatch is nonzero. Empirically, 65.96% of the filtered usable sampled pairs have positive margin. This establishes a nonzero geometric lower-bound term for these pairs, without establishing a positive affinity-specificity bound.

5.3 Validation of Pseudo-Label Refinement

We isolate one-step recovery ($K = 1$) on 100 CrossDocked2020 test complexes, fixing the DiffLinker pseudo-labels and DiffSBDD prior. Each method produces one sample per complex. Baselines are DCDP, DPS, SDEdit, RED-Diff, and RSD Li et al. (2024); Chung et al. (2023); Meng et al. (2022); Mardani et al. (2024); Zilberstein et al. (2025). Table 2 reports clash rate (any protein–ligand atom pair below 2 Å), unweighted Vina repulsion, affinity (negative Vina), valence validity, and bond-length MMD. Faithfulness uses RMSD with corresponding atom indices and Tanimoto similarity of RDKit path fingerprints. Protocols and metrics are in Appendices C.1 and C.3.

SurfSpec achieves the lowest clash rate (0.01) and global repulsion (4.40), highest local and global affinity (0.46 and 8.00 kcal/mol), and best or tied-best bond-length MMD among refinement methods. SDEdit and DCDP have lower RMSD; RED-Diff and RSD retain higher topological similarity but produce more clashes. DPS has lower local repulsion but substantially higher RMSD (Appendix D.2).

6 Conclusion

We studied off-target-agnostic specificity-aware lead optimization using only the target pocket. Under separation and calibration assumptions, reducing target pocket–ligand surface mismatch weakly improves a conservative specificity lower bound for geometrically separated off-targets. Motivated by this analysis, **SurfSpec** grows ligands toward under-occupied target-pocket regions. On CrossDocked2020, it achieves higher empirical specificity, improves target-pocket occupancy, and maintains competitive docking performance with low geometric mismatch and zero clash rate.

AI USE STATEMENT

We used generative AI tools to polish the manuscript for clarity and readability and to assist with implementing proof-of-concept experiments. The authors take responsibility for the final manuscript, experimental implementation, and reported findings.

ETHICS STATEMENT

SurfSpec aims to support early-stage drug discovery through specificity-aware molecular optimization. The reported evidence is computational: improvements in docking scores and geometric compatibility do not establish biological selectivity, safety, or therapeutic efficacy. Generated molecules require appropriate experimental validation before downstream use. As with other molecular design methods, potential misuse warrants responsible application and assessment of candidate compounds.

REPRODUCIBILITY STATEMENT

The manuscript and appendices describe the assumptions and proofs underlying our theoretical analysis, the ligand-growth and refinement procedures, and the experimental evaluation protocol. Additional details include the pretrained models, hyperparameters, baseline configurations, and metric definitions used in our experiments.

REFERENCES

Keir Adams, Kento Abeywardane, Jenna Fromer, and Connor W. Coley. ShEPhERD: Diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KSLkFYHlYg.

Sirine Ayadi, Leon Hetzel, Johanna Sommer, Fabian Theis, and Stephan Günnemann. Unified guidance for geometry-conditioned molecular generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 138891–138924. Curran Associates, Inc., 2024. doi: 10.52202/079017-4407. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/faa6276ea12d7afeb3e42b210c86f688-Paper-Conference.pdf.

Shicheng Chen, Odin Zhang, Chenran Jiang, Huifeng Zhao, Xujun Zhang, Mengting Chen, Yun Liu, Qun Su, Zhenxing Wu, Xinyue Wang, Wanglin Qu, Yuanyi Ye, Xin Chai, Ning Wang, Tianyue Wang, Yuan An, Guanlin Wu, Qianqian Yang, Jiean Chen, Wei Xie, Haitao Lin, Dan Li, Chang-Yu Hsieh, Yong Huang, Yu Kang, Tingjun Hou, and Peichen Pan. Deep lead optimization enveloped in protein pocket and its application in designing potent and selective ligands targeting LTK protein. *Nature Machine Intelligence*, 7(3):448–458, 2025. doi: 10.1038/s42256-025-00997-w. URL https://doi.org/10.1038/s42256-025-00997-w.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25683–25696. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a48e5877c7bf86a513950ab23b360498-Paper-Conference.pdf.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DsEhqQtfAG.

Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003. doi: 10.1109/TIT.2003.813506.

Julius Erbach, Dominik Narnhofer, Andreas Dombos, Bernt Schiele, Jan Eric Lenssen, and Konrad Schindler. Solving inverse problems with FLAIR. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 136667–136704. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c7ae6e9659f0c99582c2e8214ba0b413-Paper-Conference.pdf.

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. *J. Chem. Inf. Model.*, 60(9):4200–4215, 2020. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411.

Bowen Gao, Minsi Ren, Yuyan Ni, Yanwen Huang, Bo Qiang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. Rethinking specificity in SBDD: Leveraging delta score and energy-guided diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14811–14825. PMLR, 2024. URL https://proceedings.mlr.press/v235/gao24k.html.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 44040–44063. Curran Associates, Inc., 2024. doi: 10.52202/079017-1398. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4ddfe69f164eae70abc86f0f9cbed7e8-Paper-Conference.pdf.

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*, 2023.

Qiaoyu Hu, Changzhi Sun, Huan He, Jiazheng Xu, Danlin Liu, Wenqing Zhang, Sumeng Shi, Kai Zhang, and Honglin Li. Target-aware 3D molecular generation based on guided equivariant diffusion. *Nature Communications*, 16:7928, 2025. doi: 10.1038/s41467-025-63245-0. URL https://www.nature.com/articles/s41467-025-63245-0.

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(1):2657, 2024. doi: 10.1038/s41467-024-46569-1. URL https://doi.org/10.1038/s41467-024-46569-1.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024. doi: 10.1038/s42256-024-00815-9. URL https://doi.org/10.1038/s42256-024-00815-9.

Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009. doi: 10.1038/nature08506. URL https://doi.org/10.1038/nature08506.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. FlowDPS : Flow-driven posterior sampling for inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12328–12337, 2025.

Minwoo Kim, Seunghyeok Shin, and Hongki Lim. FAST-DIPS: Adjoint-free analytic steps and hard-constrained likelihood correction for diffusion-prior inverse problems. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=voMeZVAkKL.

Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161 (2):269–288, 1982. doi: https://doi.org/10.1016/0022-2836(82)90153-X. URL https://www.sciencedirect.com/science/article/pii/002228368290153X.

Ziseok Lee, Minyeong Hwang, Wooyeol Lee, Sanghyun Jo, Jihyung Ko, Young Bin Park, Jae-Mun Choi, Eunho Yang, and Kyungsu Kim. On the collapse of generative paths: A criterion and correction for diffusion steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=emv2qsi3TG.

Bohao Li, Xinyu Wu, Yu Cao, Jie Lin, Jingpeng Zhong, Hua Chen, Yongzhi Lu, Miru Tang, Jinping Lei, Ting Ran, Yuedong Yang, Mingyuan Xu, Jinsai Shang, and Hongming Chen. De novo molecular design via shape-constrained diffusion models. *J. Chem. Inf. Model.*, 66(8):4592–4606, 2026a. doi: 10.1021/acs.jcim.6c00044. URL https://doi.org/10.1021/acs.jcim.6c00044.

Xiang Li, Soo Min Kwon, Shijun Liang, Ismail R. Alkhouri, Saiprasad Ravishankar, and Qing Qu. Decoupled data consistency with diffusion purification for image restoration, 2024. URL https://arxiv.org/abs/2403.06054.

Yanheng Li, Haojia Dong, Xiaohan Lin, Yize Hao, Yue Xue, Jun Zhang, Yundong Wu, Jinming Zhou, and Yi Qin Gao. MolSculptor: An adaptive diffusion–evolution framework enabling generative drug design for multitarget affinity and selectivity. *Journal of Chemical Theory and Computation*, 2026b. doi: 10.1021/acs.jctc.6c00395. URL https://doi.org/10.1021/acs.jctc.6c00395.

Zian Li, Cai Zhou, Xiyuan Wang, Xingang Peng, and Muhan Zhang. Geometric representation condition improves equivariant molecule generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36921–36953. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25dz.html.

Haitao Lin, Yufei Huang, Odin Zhang, Yunfan Liu, Lirong Wu, Siyuan Li, Zhiyuan Chen, and Stan Z. Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34603–34626. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6cdd4ce9330025967dd1ed0bed3010f5-Paper-Conference.pdf.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 28027–28053, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/7711026951f3dc3afb2be427bf98bc73-Paper-Conference.pdf.

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/708e58b0b99e3e62d42022b4564bad7a-Abstract-Conference.html.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Jonghyun Park and Jong Chul Ye. FlowLPS: Langevin-proximal sampling for flow-based inverse problem solvers, 2025. URL https://arxiv.org/abs/2512.07150.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. FlowChef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15308–15318, 2025.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *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.

Xingang Peng, Ruihan Guo, Fenglin Guo, Ziyi Wang, Jiayu Sun, Jiaqi Guan, Yinjun Jia, Yan Xu, Yanwen Huang, Muhan Zhang, Jian Peng, Xinquan Wang, Chuanhui Han, Zihua Wang, and Jianzhu Ma. Unified modeling of 3D molecular generation via atomic interactions with PocketXMol. *Cell*, 2026. doi: 10.1016/j.cell.2026.01.003. URL https://doi.org/10.1016/j.cell.2026.01.003.

Mehrsa Pourya, Bassam El Rawas, and Michael Unser. FLOWER: A flow-matching solver for inverse problems. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QGd34p02mI.

Anjie Qiao, Yuting Chen, Junjie Xie, Weifeng Huang, Hao Zhang, Qirui Deng, Jiahua Rao, Ji Deng, Fanbo Meng, Zhen Wang, Mingyuan Xu, Hongming Chen, Jiancong Xie, Shuangjia Zheng, Yuedong Yang, Guo-Bo Li, and Jinping Lei. A 3D pocket-aware lead optimization model with knowledge guidance and its application for discovery of new glutaminyl cyclase inhibitors. *Briefings in Bioinformatics*, 26(4):bbaf345, 2025. doi: 10.1093/bib/bbaf345. URL https://doi.org/10.1093/bib/bbaf345.

Keyue Qiu, Yuxuan Song, Jie Yu, Hongbo Ma, Ziyao Cao, Zhilong Zhang, Yushuai Wu, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Empower structure-based molecule optimization with gradient guided Bayesian flow networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50645–50671, 2025. URL https://proceedings.mlr.press/v267/qiu25h.html.

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.

Brian K. Shoichet and Irwin D. Kuntz. Matching chemistry and shape in molecular docking. *Protein Engineering*, 6(7):723–732, 1993. doi: 10.1093/protein/6.7.723. URL https://doi.org/10.1093/protein/6.7.723.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

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.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=mRieQgMtNTQ.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20895–20905, 2025.

Renyi Zhou, Huimin Zhu, Jing Tang, and Min Li. ActivityDiff: A diffusion model with positive and negative activity guidance for de novo drug design, 2025. URL https://arxiv.org/abs/2508.06364.

Xiangxin Zhou, Xiwei Cheng, Yuwei Yang, Yu Bao, Liang Wang, and Quanquan Gu. DecompOpt: Controllable and decomposed diffusion models for structure-based molecular optimization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Y3BbxvAQS9.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1219–1229, 2023.

Nicolas Zilberstein, Morteza Mardani, and Santiago Segarra. Repulsive latent score distillation for solving inverse problems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 60181–60224, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/96d328a1f6d8396d8c8a62f2beee252a-Paper-Conference.pdf.

Table 3: Comparison of the cited molecular design and sample-refinement formulations. A checkmark denotes an explicit component, a triangle partial or variant-dependent support, and a cross a component not explicitly formulated. Definitions and scope are given in Appendix A.

	Method family	Representative methods	No objective-specific retraining	Inference-time sample control	Explicit geometric control	Pocket–ligand surface complementarity	Target-only specificity	OOD pseudo-label refinement
Lead Optimization	**Off-target-agnostic specificity-aware optimization**	**SurfSpec (Ours)**	✓	✓	✓	✓	✓	✓
	Off-target-guided specificity	SBE-Diff (2024), ActivityDiff (2025)	×	✓	×	×	×	×
		MolSculptor (2026b)	△	✓	×	×	×	×
	Target-only optimization	DecompOpt (2024)	✓	✓	×	×	×	×
		Delete (2025)	×	×	×	×	×	×
		PMDM (2024)	△	△	×	×	×	×
		Diffleop (2025), MolJO (2025)	×	✓	△	×	×	×
Sample refinement	**Reverse-path refinement**	SDEdit (2022), DDNM (2023), MCG (2022), DPS (2023), ΠGDM (2023), FlowChef (2025), FlowDPS (2025)	✓	✓	×	×	×	×
	Optimization-based refinement	DiffPIR (2023), DDS (2024), DAPS (2025), DCDP (2024), PnP-Flow (2025), FLOWER (2026), FlowLPS (2025)	✓	✓	×	×	×	×
	Variational refinement	RED-Diff (2024), RSD (2025), FLAIR (2025)	✓	✓	×	×	×	×
	Constrained sample refinement	FAST-DIPS (2026)	✓	✓	×	×	×	×
Geometric control	**Shape-conditioned generation**	Diff-Shape (2026a)	×	×	✓	×	×	×
	Multimodal geometric generation	ShEPhERD (2025)	×	×	✓	×	×	×
	Representation-conditioned generation	GeoRCG (2025)	×	×	✓	×	×	×
	Training-free geometric guidance	UniGuide (2024)	✓	✓	✓	×	×	×
Molecular design	**Pocket-conditioned generation**	Pocket2Mol (2022), TargetDiff (2023), AliDiff (2024)	×	×	×	×	×	×
	Multi-task SBDD generation	DiffSBDD (2024)	✓	✓	△	×	×	×
		DiffGui (2025), PocketXMol (2026)	△	✓	△	×	×	×
	Molecular elaboration	D3FG (2023)	×	×	△	×	×	×

A Literature Context of the Work

Table 3 compares the cited formulations, not their capacity for future adaptation. "No objective-specific retraining" allows pretraining but excludes fitting a new objective model; partial support includes pipelines or grouped methods with different training requirements. "Inference-time sample control" denotes explicit guidance, correction, or optimization beyond ordinary conditioning on a pocket or reference input. "Explicit geometric control" refers to a prescribed shape, substructure, or spatial constraint. "Surface complementarity" denotes an explicit pocket–ligand surface mismatch objective, and "target-only specificity" a specificity analysis or objective that does not access off-targets during optimization. The final column concerns the construction and recovery of molecular geometric pseudo-labels, rather than generic image editing. SurfSpec combines these components in one ligand-growth pipeline; its specificity analysis remains conditional on the separation and calibration assumptions in Section 4.1.

B Proofs and Additional Theoretical Details

B.1 Surface-Oriented Probability Fields

We construct the surface-induced probability measures used in Eq. (4) from signed-distance functions. Figure 3 illustrates the surfaces and induced densities.

Let $A_L = \{a_i\}_{i=1}^{N_L}$ denote the heavy-atom coordinates of the ligand L. As in the main text, we define the continuous ligand-local domain

$$\Omega_L = \left\{ x \in \mathbb{R}^3 : \min_i \|x - a_i\|_2 \le 8\,\text{Å} \right\}. \tag{11}$$

Let $\mathcal{S}_L$ denote the van der Waals surface of the ligand heavy atoms, and let $\mathcal{S}_P^L$ denote the ligand-oriented pocket surface, obtained by aligning the pocket to the ligand frame and restricting the pocket van der Waals surface to points within $8\,$Å of any ligand heavy atom.

We define signed-distance functions on Ω_L. For the ligand, let B_L be the ligand van der Waals solid, whose boundary is $\mathcal{S}_L$. The signed distance to the ligand surface is

$$\phi_L(x) = \begin{cases} \inf_{z\in\mathcal{S}_L} \|x - z\|_2, & x \notin B_L, \\ -\inf_{z\in\mathcal{S}_L} \|x - z\|_2, & x \in B_L. \end{cases} \tag{12}$$

For the pocket, let B_P^L be the aligned protein van der Waals solid. We define

$$\phi_P(x) = \begin{cases} \inf_{z\in\mathcal{S}_P^L} \|x - z\|_2, & x \notin B_P^L, \\ -\inf_{z\in\mathcal{S}_P^L} \|x - z\|_2, & x \in B_P^L. \end{cases} \tag{13}$$

Thus, $\phi_P(x) < 0$ indicates that x lies inside the protein van der Waals volume, corresponding to a clash region.

We define the ligand and pocket surface energies as

$$\begin{aligned} E_L(x) &= \frac{\phi_L(x)^2}{2\sigma_L^2}, \\ E_P(x) &= \frac{\phi_P(x)^2}{2\sigma_P^2} + \lambda_{\text{clash}}[-\phi_P(x)]_+, \end{aligned} \tag{14}$$

where $[u]_+ = \max(u, 0)$. The clash term discourages probability mass inside the protein side of the pocket surface.

We convert these energies into normalized Boltzmann densities on the common continuous domain Ω_L:

$$q_L(x) = \frac{\exp(-\beta E_L(x))}{\int_{\Omega_L} \exp(-\beta E_L(x'))\,dx'}, \tag{15}$$

$$q_P(x) = \frac{\exp(-\beta E_P(x))}{\int_{\Omega_L} \exp(-\beta E_P(x'))\,dx'}. \tag{16}$$

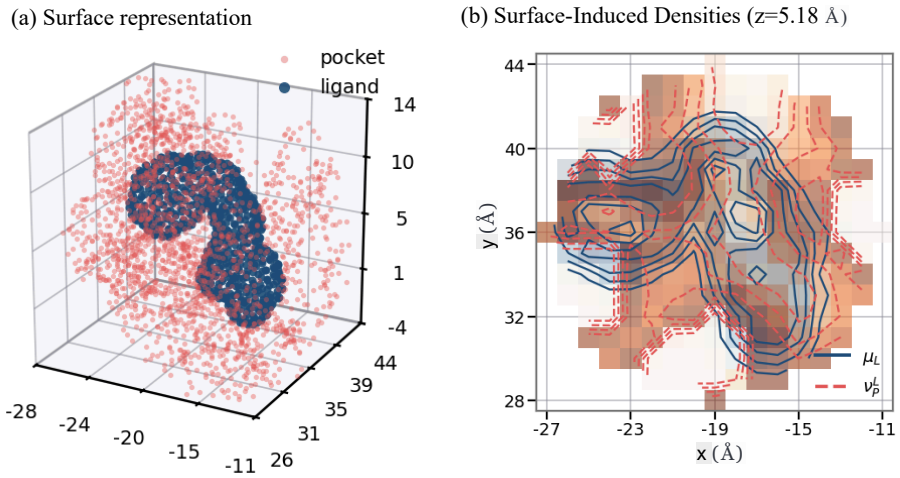


Figure 3: Surface and measure representations used for geometric mismatch. (a) The ligand surface $\mathcal{S}_L$ and the ligand-oriented pocket surface $\mathcal{S}_P^L$, obtained by cropping the aligned pocket surface around the ligand. (b) A two-dimensional slice of the signed-distance-based Boltzmann densities induced by $\mathcal{S}_L$ and $\mathcal{S}_P^L$, which define the probability measures μ_L and ν_P^L, respectively. The geometric mismatch $d_{\mathrm{gm}}(L, P)$ is computed as the Jensen–Shannon distance between these induced probability measures.

We set $\beta = 2$ in all experiments. The corresponding probability measures are

$$d\mu_L(x) = q_L(x)\,dx, \qquad d\nu_P^L(x) = q_P(x)\,dx. \tag{17}$$

The geometric mismatch in Eq. (4) is then computed as the Jensen–Shannon distance between μ_L and ν_P^L, which are defined on the same ligand-local domain.

In implementation, the integrals in Eq. (16) are approximated by evaluating the unnormalized Boltzmann densities on a ligand-centered voxel grid over Ω_L and normalizing the resulting discrete weights.

B.2 Proof of the Specificity Certificate

We first restate the quantile-calibrated geometry–affinity assumption used in the main text. This assumption provides an empirical bridge from geometric mismatch to affinity, while allowing residual variation due to non-geometric factors such as electrostatics, hydrogen bonding, and desolvation.

Assumption 1 (Quantile-calibrated geometry–affinity envelope). *Let $\mathcal{D}_{\mathrm{CD}}$ denote the empirical ligand–pocket distribution induced by CrossDocked2020. For a tail probability $\eta \in (0, 1)$, there exist monotone decreasing functions f_-^η and f_+^η such that, for $(L, P) \sim \mathcal{D}_{\mathrm{CD}}$,*

$$\Pr\left[f_-^\eta(d_{\mathrm{gm}}(L, P)) \le A_P(L) \le f_+^\eta(d_{\mathrm{gm}}(L, P))\right] \ge 1 - \eta. \tag{18}$$

Here, $A_P(L)$ is an affinity-like score for ligand L against pocket P, where larger values indicate stronger binding. When applying the certificate to optimized ligands, we additionally assume that the same $1 - \eta$ envelope coverage holds marginally for the target pair and each considered off-target pair under the joint distribution induced by the optimization and evaluation protocol, with a fixed finite number of off-targets. This coverage transfer is assumed, not established by the CrossDocked2020 calibration experiment.

This assumption does not require geometry to fully determine affinity. Rather, it states that geometric mismatch constrains the attainable affinity range with high probability. The percentile envelope absorbs affinity variation caused by non-geometric interactions.

We next formalize the geometric part of the certificate. Let $\mathcal{L}_{\mathrm{opt}}$ denote the set of ligands reachable by optimization. For $L \in \mathcal{L}_{\mathrm{opt}}$ and $O \in \mathcal{O}$, let

$$m_L(O) = d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O)$$

denote the ligand-oriented geometric separation between the target and off-target pocket surfaces, where $\nu^L_{P_{\mathrm{tgt}}}$ and ν^L_O are the corresponding ligand-oriented pocket surface measures. We say that $\mathcal{O}$ is a δ-separated off-target class if

$$\inf_{L \in \mathcal{L}_{\mathrm{opt}}} \inf_{O \in \mathcal{O}} m_L(O) \geq \delta.$$

This means that every considered off-target pocket remains at least δ away from the target pocket under the ligand-oriented surface metric throughout optimization.

Lemma 1 (Triangle lower bound for off-target mismatch). *Let $L \in \mathcal{L}_{\mathrm{opt}}$, and let*

$$\epsilon = d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})$$

be the target mismatch. Then, for each off-target pocket $O \in \mathcal{O}$,

$$d_{\mathrm{gm}}(L, O) \geq [m_L(O) - \epsilon]_+. \tag{19}$$

In particular, if $\mathcal{O}$ is a δ-separated off-target class, then

$$d_{\mathrm{gm}}(L, O) \geq [\delta - \epsilon]_+ \qquad \textit{for all } O \in \mathcal{O}. \tag{20}$$

Proof. By definition of the ligand-oriented mismatch,

$$d_{\mathrm{gm}}(L, P_{\mathrm{tgt}}) = d_{\mathrm{JS}}(\mu_L, \nu^L_{P_{\mathrm{tgt}}}) = \epsilon,$$

and

$$d_{\mathrm{gm}}(L, O) = d_{\mathrm{JS}}(\mu_L, \nu^L_O).$$

Since d_{JS} is a metric on probability measures, it satisfies the triangle inequality:

$$d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O) \leq d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \mu_L) + d_{\mathrm{JS}}(\mu_L, \nu^L_O).$$

Therefore,

$$m_L(O) \leq \epsilon + d_{\mathrm{gm}}(L, O).$$

Rearranging gives

$$d_{\mathrm{gm}}(L, O) \geq m_L(O) - \epsilon.$$

Since $d_{\mathrm{gm}}(L, O) \geq 0$, we obtain

$$d_{\mathrm{gm}}(L, O) \geq [m_L(O) - \epsilon]_+.$$

If $\mathcal{O}$ is δ-separated, then $m_L(O) \geq \delta$ for every $L \in \mathcal{L}_{\mathrm{opt}}$ and $O \in \mathcal{O}$. Hence

$$[m_L(O) - \epsilon]_+ \geq [\delta - \epsilon]_+,$$

which yields

$$d_{\mathrm{gm}}(L, O) \geq [\delta - \epsilon]_+.$$

□

Theorem 1 (Specificity certificate for separated off-targets). *Let $L \in \mathcal{L}_{\mathrm{opt}}$ be the output of the optimization protocol, and let $\epsilon = d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})$. Suppose that $\mathcal{O}$ is a δ-separated off-target class. Under Assumption 1, including its coverage-transfer condition, with probability at least $1 - (|\mathcal{O}| + 1)\eta$ over the joint optimization and evaluation protocol,*

$$\mathrm{Spec}(L; P_{\mathrm{tgt}}, \mathcal{O}) \geq f^{\eta}_{-}(\epsilon) - f^{\eta}_{+}([\delta - \epsilon]_+)\,. \tag{21}$$

Consequently, reducing the target mismatch ϵ weakly improves this high-probability specificity lower bound.

Proof. The proof has two steps. First, we obtain a geometric lower bound on the mismatch between the ligand and each off-target pocket. Second, we convert this geometric lower bound into an affinity-level specificity certificate using the percentile-calibrated geometry–affinity envelope.

By Lemma 1, for every off-target pocket $O \in \mathcal{O}$,

$$d_{\mathrm{gm}}(L, O) \geq [\delta - \epsilon]_+. \tag{22}$$

Now consider the calibrated affinity events under the joint optimization and evaluation protocol. By the coverage-transfer condition in Assumption 1, the target lower-envelope event holds with probability at least $1 - \eta$,

$$A_{P_{\mathrm{tgt}}}(L) \geq f_-^\eta(d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})) = f_-^\eta(\epsilon).$$

For each off-target pocket $O \in \mathcal{O}$, the upper envelope gives, with probability at least $1 - \eta$,

$$A_O(L) \leq f_+^\eta(d_{\mathrm{gm}}(L, O)).$$

Because f_+^η is monotone decreasing and $d_{\mathrm{gm}}(L, O) \geq [\delta - \epsilon]_+$, we have

$$f_+^\eta(d_{\mathrm{gm}}(L, O)) \leq f_+^\eta([\delta - \epsilon]_+)\,.$$

Therefore, for every $O \in \mathcal{O}$,

$$A_O(L) \leq f_+^\eta([\delta - \epsilon]_+)\,.$$

Taking the maximum over off-target pockets yields

$$\max_{O \in \mathcal{O}} A_O(L) \leq f_+^\eta([\delta - \epsilon]_+)\,.$$

Combining the target lower bound and the off-target upper bound gives

$$\begin{aligned} \mathrm{Spec}(L; P_{\mathrm{tgt}}, \mathcal{O}) &= A_{P_{\mathrm{tgt}}}(L) - \max_{O \in \mathcal{O}} A_O(L) \\ &\geq f_-^\eta(\epsilon) - f_+^\eta([\delta - \epsilon]_+)\,. \end{aligned}$$

It remains to account for the probability of the calibrated events. The target lower-envelope event fails with probability at most η. For each of the $|\mathcal{O}|$ off-target pockets, the upper-envelope event fails with probability at most η. By the union bound, all calibrated events hold simultaneously with probability at least

$$1 - (|\mathcal{O}| + 1)\eta.$$

Thus Eq. (21) holds with the stated probability.

Finally, define the certified lower bound as

$$B(\epsilon) = f_-^\eta(\epsilon) - f_+^\eta([\delta - \epsilon]_+)\,.$$

Since f_-^η is monotone decreasing, reducing ϵ weakly increases $f_-^\eta(\epsilon)$. Also, reducing ϵ weakly increases $[\delta - \epsilon]_+$, and since f_+^η is monotone decreasing, this weakly decreases $f_+^\eta([\delta - \epsilon]_+)$. Therefore, reducing the target mismatch ϵ weakly improves the high-probability lower bound $B(\epsilon)$. □

B.3 Low-Noise Analysis of Anchored Refinement

We analyze the low-noise anchored recovery step for a pseudo-label $a = x_{\mathrm{label}}$. Let

$$V_a(x) = \frac{\lambda}{2}\|x - a\|_2^2$$

denote the anchor potential. The anchored clean distribution is defined as

$$p_0^{\mathrm{anc}}(x \mid P_{\mathrm{tgt}}, a) = \frac{1}{Z_a} p_0(x \mid P_{\mathrm{tgt}}) \exp(-V_a(x)), \tag{23}$$

where Z_a is the normalizing constant. Let μ_0^{anc} and Σ_0^{anc} denote the mean and covariance of $p_0^{\mathrm{anc}}(\cdot \mid P_{\mathrm{tgt}}, a)$, respectively.

For diffusion time τ, consider the Gaussian forward perturbation

$$Y_\tau = \alpha_\tau X_0 + \sigma_\tau Z, \qquad Z \sim \mathcal{N}(0, I_d), \tag{24}$$

where $X_0 \sim p_0^{\mathrm{anc}}(\cdot \mid P_{\mathrm{tgt}}, a)$ and Z is independent of X_0. We denote the distribution of Y_τ by p_τ^{anc}.

Assumption 2 (Anchored concentration). *The anchored clean distribution is concentrated in a small ball around its mean:*

$$\operatorname{supp}\left(p_0^{\mathrm{anc}}(\cdot \mid P_{\mathrm{tgt}}, a)\right) \subset B(\mu_0^{\mathrm{anc}}, r_0). \tag{25}$$

This assumption formalizes the intuition that imposing both pocket-conditioned validity and similarity to the pseudo-label leaves a locally concentrated set of feasible clean ligands. It also implies the covariance bound

$$\operatorname{tr}(\Sigma_0^{\mathrm{anc}}) = \mathbb{E}\left[\|X_0 - \mu_0^{\mathrm{anc}}\|_2^2\right] \le r_0^2. \tag{26}$$

This covariance bound is a consequence of Assumption 2, not an additional assumption.

Proposition 1 (Mode recovery under anchored concentration). *Suppose Assumption 2 holds. Let $y_\tau^\star$ be an interior mode of p_τ^{anc}. Then*

$$\|y_\tau^\star - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2 \le |\alpha_\tau| r_0. \tag{27}$$

Consequently, if the low-noise anchored score flow recovers the mode $y_\tau^\star$, then its convergence point lies within $|\alpha_\tau| r_0$ of the forward mean $\alpha_\tau \mu_0^{\mathrm{anc}}$.

Proof. For the Gaussian forward perturbation in Eq. (24), the score identity is

$$\nabla_y \log p_\tau^{\mathrm{anc}}(y) = -\frac{1}{\sigma_\tau^2}\left(y - \alpha_\tau \mathbb{E}[X_0 \mid Y_\tau = y, P_{\mathrm{tgt}}, a]\right). \tag{28}$$

Since $y_\tau^\star$ is an interior mode, $\nabla_y \log p_\tau^{\mathrm{anc}}(y_\tau^\star) = 0$. Therefore,

$$y_\tau^\star = \alpha_\tau \mathbb{E}[X_0 \mid Y_\tau = y_\tau^\star, P_{\mathrm{tgt}}, a].$$

Conditioning on $Y_\tau = y_\tau^\star$ only reweights the clean distribution and does not enlarge its support. Hence the posterior distribution of $X_0 \mid Y_\tau = y_\tau^\star$ remains supported in $B(\mu_0^{\mathrm{anc}}, r_0)$. Since this ball is convex, the posterior mean also lies in the same ball:

$$\|\mathbb{E}[X_0 \mid Y_\tau = y_\tau^\star, P_{\mathrm{tgt}}, a] - \mu_0^{\mathrm{anc}}\|_2 \le r_0.$$

Multiplying by $|\alpha_\tau|$ proves Eq. (30). □

Assumption 3 (Local DPS score approximation). *Let*

$$\widetilde{s}_\tau^{\mathrm{anc}}(y) = \nabla_y \log p_\tau(y \mid P_{\mathrm{tgt}}) - \lambda J_{\hat{x}_{0|\tau}}(y)^\top \left(\hat{x}_{0|\tau}(y) - x_{\mathrm{label}}\right)$$

denote the DPS-style anchored score estimator used in Eq. (8)*. On the region B visited by the low-noise anchored score flow, we assume that this estimator is locally close to the exact anchored score:*

$$\sup_{y \in B} \|\widetilde{s}_\tau^{\mathrm{anc}}(y) - \nabla_y \log p_\tau^{\mathrm{anc}}(y)\|_2 \le \xi_\tau. \tag{29}$$

Proposition 2 (Mode recovery under anchored concentration). *Suppose Assumption 2 holds and define*

$$\kappa_\tau = \frac{\alpha_\tau^2 r_0^2}{\sigma_\tau^2}.$$

If $\kappa_\tau < 1$, then $\log p_\tau^{\mathrm{anc}}$ is strongly concave with constant

$$m_\tau = \frac{1 - \kappa_\tau}{\sigma_\tau^2}.$$

Consequently, p_τ^{anc} has a unique mode $y_\tau^\star$, and

$$\|y_\tau^\star - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2 \le |\alpha_\tau| r_0. \tag{30}$$

Assume further that the approximate flow

$$\frac{dy_s}{ds} = \widetilde{s}_\tau^{\mathrm{anc}}(y_s)$$

remains in a region B, and that the implemented score satisfies

$$\sup_{y\in B}\|\widetilde{s}_\tau^{\mathrm{anc}}(y)-\nabla_y\log p_\tau^{\mathrm{anc}}(y)\|_2\le\xi_\tau.$$

Then

$$\|y_s-y_\tau^\star\|_2\le e^{-m_\tau s}\|y_0-y_\tau^\star\|_2+\frac{\xi_\tau}{m_\tau}\left(1-e^{-m_\tau s}\right). \tag{31}$$

Proof. For a Gaussian forward perturbation, the Hessian identity gives

$$\nabla_y^2\log p_\tau^{\mathrm{anc}}(y)=-\frac{1}{\sigma_\tau^2}I_d+\frac{\alpha_\tau^2}{\sigma_\tau^4}\operatorname{Cov}\left(X_0\mid Y_\tau=y,P_{\mathrm{tgt}},a\right).$$

Under Assumption 2, the conditional support of $X_0\mid Y_\tau=y$ remains inside $B(\mu_0^{\mathrm{anc}},r_0)$. Therefore,

$$\|\operatorname{Cov}\left(X_0\mid Y_\tau=y,P_{\mathrm{tgt}},a\right)\|_{\mathrm{op}}\le r_0^2.$$

Hence the largest eigenvalue of $\nabla_y^2\log p_\tau^{\mathrm{anc}}(y)$ is at most

$$-\frac{1}{\sigma_\tau^2}+\frac{\alpha_\tau^2 r_0^2}{\sigma_\tau^4}=-\frac{1-\kappa_\tau}{\sigma_\tau^2}=-m_\tau,$$

so $\log p_\tau^{\mathrm{anc}}$ is m_τ-strongly concave.

The score identity for Gaussian perturbations is

$$\nabla_y\log p_\tau^{\mathrm{anc}}(y)=-\frac{1}{\sigma_\tau^2}\left(y-\alpha_\tau\mathbb{E}[X_0\mid Y_\tau=y,P_{\mathrm{tgt}},a]\right).$$

At the mode $y_\tau^\star$, the score vanishes, so

$$y_\tau^\star=\alpha_\tau\mathbb{E}[X_0\mid Y_\tau=y_\tau^\star,P_{\mathrm{tgt}},a].$$

The posterior mean lies in the same convex ball $B(\mu_0^{\mathrm{anc}},r_0)$, which gives Eq. (30).

Let

$$s_\tau^{\mathrm{anc}}(y)=\nabla_y\log p_\tau^{\mathrm{anc}}(y),\qquad \widetilde{s}_\tau^{\mathrm{anc}}(y)=s_\tau^{\mathrm{anc}}(y)+\Delta_\tau(y),$$

with $\|\Delta_\tau(y)\|_2\le\xi_\tau$ on B. Since $s_\tau^{\mathrm{anc}}(y_\tau^\star)=0$, strong concavity implies

$$\langle y_s-y_\tau^\star,s_\tau^{\mathrm{anc}}(y_s)-s_\tau^{\mathrm{anc}}(y_\tau^\star)\rangle\le-m_\tau\|y_s-y_\tau^\star\|_2^2.$$

Therefore,

$$\frac{d}{ds}\|y_s-y_\tau^\star\|_2\le-m_\tau\|y_s-y_\tau^\star\|_2+\xi_\tau.$$

Gronwall's inequality gives Eq. (31). □

Proposition 3 (Accuracy of the low-noise Gaussian surrogate). *Define the Gaussian surrogate*

$$\widetilde{p}_\tau^{\mathrm{anc}}=\mathcal{N}\left(\alpha_\tau\mu_0^{\mathrm{anc}},\sigma_\tau^2I_d\right), \tag{32}$$

and the residual-dispersion ratio

$$\rho_\tau=\frac{\alpha_\tau^2\operatorname{tr}(\Sigma_0^{\mathrm{anc}})}{\sigma_\tau^2}. \tag{33}$$

Then

$$D_{\mathrm{KL}}\left(p_\tau^{\mathrm{anc}}\|\widetilde{p}_\tau^{\mathrm{anc}}\right)\le\frac{\rho_\tau}{2}, \tag{34}$$

$$W_2^2\left(p_\tau^{\mathrm{anc}},\widetilde{p}_\tau^{\mathrm{anc}}\right)\le\alpha_\tau^2\operatorname{tr}(\Sigma_0^{\mathrm{anc}}), \tag{35}$$

and

$$\mathrm{TV}\left(p_\tau^{\mathrm{anc}},\widetilde{p}_\tau^{\mathrm{anc}}\right)\le\frac{1}{2}\sqrt{\rho_\tau}. \tag{36}$$

Under Assumption 2,

$$\rho_\tau\le\frac{\alpha_\tau^2r_0^2}{\sigma_\tau^2}. \tag{37}$$

Thus, when $|\alpha_\tau|r_0\ll\sigma_\tau$, p_τ^{anc} is well approximated by the Gaussian surrogate centered at $\alpha_\tau\mu_0^{\mathrm{anc}}$.

Proof. Conditioned on $X_0 = x$, the forward distribution is

$$Y_\tau \mid X_0 = x \sim \mathcal{N}(\alpha_\tau x, \sigma_\tau^2 I_d).$$

By convexity of KL divergence in its first argument,

$$\begin{aligned} &D_{\mathrm{KL}}\left(p_\tau^{\mathrm{anc}} \| \widetilde{p}_\tau^{\mathrm{anc}}\right) \\ &\le \mathbb{E}_{X_0}\left[D_{\mathrm{KL}}\left(\mathcal{N}(\alpha_\tau X_0, \sigma_\tau^2 I_d) \| \mathcal{N}(\alpha_\tau \mu_0^{\mathrm{anc}}, \sigma_\tau^2 I_d)\right)\right] \\ &= \frac{\alpha_\tau^2}{2\sigma_\tau^2}\mathbb{E}\left[\|X_0 - \mu_0^{\mathrm{anc}}\|_2^2\right] = \frac{\rho_\tau}{2}. \end{aligned}$$

For the Wasserstein bound, couple

$$Y_\tau = \alpha_\tau X_0 + \sigma_\tau Z, \qquad \widetilde{Y}_\tau = \alpha_\tau \mu_0^{\mathrm{anc}} + \sigma_\tau Z$$

using the same Gaussian noise Z. Then

$$\mathbb{E}\left[\|Y_\tau - \widetilde{Y}_\tau\|_2^2\right] = \alpha_\tau^2 \operatorname{tr}(\Sigma_0^{\mathrm{anc}}),$$

which proves Eq. (35). The total-variation bound follows from Pinsker's inequality. Finally, Eq. (37) follows from Eq. (26). □

Proposition 4 (Re-noising around the recovered point). *Suppose Assumption 2 holds, and suppose the conditions of Proposition 2 hold. Let y_s be the point obtained by the approximate anchored score flow at flow time s, and define*

$$\widehat{X}_\tau = y_s + \sigma_\tau \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d).$$

Equivalently,

$$\widehat{X}_\tau \sim \mathcal{N}(y_s, \sigma_\tau^2 I_d).$$

Then

$$\begin{aligned} &W_2\left(p_\tau^{\mathrm{anc}}, \mathcal{N}(y_s, \sigma_\tau^2 I_d)\right) \\ &\le 2|\alpha_\tau| r_0 + e^{-m_\tau s}\|y_0 - y_\tau^\star\|_2 + \frac{\xi_\tau}{m_\tau}\left(1 - e^{-m_\tau s}\right). \end{aligned} \tag{38}$$

In particular, if the recovered point is the exact mode $y_s = y_\tau^\star$, then

$$W_2\left(p_\tau^{\mathrm{anc}}, \mathcal{N}(y_\tau^\star, \sigma_\tau^2 I_d)\right) \le 2|\alpha_\tau| r_0. \tag{39}$$

Proof. By the triangle inequality,

$$\begin{aligned} &W_2\left(p_\tau^{\mathrm{anc}}, \mathcal{N}(y_s, \sigma_\tau^2 I_d)\right) \\ &\le W_2\left(p_\tau^{\mathrm{anc}}, \widetilde{p}_\tau^{\mathrm{anc}}\right) + W_2\left(\widetilde{p}_\tau^{\mathrm{anc}}, \mathcal{N}(y_s, \sigma_\tau^2 I_d)\right). \end{aligned}$$

The first term is bounded by

$$W_2\left(p_\tau^{\mathrm{anc}}, \widetilde{p}_\tau^{\mathrm{anc}}\right) \le |\alpha_\tau| \sqrt{\operatorname{tr}(\Sigma_0^{\mathrm{anc}})} \le |\alpha_\tau| r_0,$$

where the first inequality follows from Proposition 3, and the second follows from Assumption 2.

For the second term, recall that

$$\widetilde{p}_\tau^{\mathrm{anc}} = \mathcal{N}\left(\alpha_\tau \mu_0^{\mathrm{anc}}, \sigma_\tau^2 I_d\right).$$

Since the two Gaussian distributions have the same covariance,

$$W_2\left(\widetilde{p}_\tau^{\mathrm{anc}}, \mathcal{N}(y_s, \sigma_\tau^2 I_d)\right) = \|y_s - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2.$$

By the triangle inequality,

$$\|y_s - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2 \le \|y_s - y_\tau^\star\|_2 + \|y_\tau^\star - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2.$$

Proposition 2 gives

$$\|y_s - y_\tau^\star\|_2 \le e^{-m_\tau s}\|y_0 - y_\tau^\star\|_2 + \frac{\xi_\tau}{m_\tau}\left(1 - e^{-m_\tau s}\right),$$

and also

$$\|y_\tau^\star - \alpha_\tau \mu_0^{\mathrm{anc}}\|_2 \le |\alpha_\tau| r_0.$$

Combining these bounds proves Eq. (38). If $y_s = y_\tau^\star$, the flow-error term vanishes, yielding Eq. (39). □

C Experimental Details

C.1 Implementation Details for Ligand Growth and Refinement

Benchmark protocol. All ligand-growth experiments are conducted on the 100 test complexes from CrossDocked2020 Francoeur et al. (2020). For each complex, the reference ligand paired with the target pocket is used as the initial lead, and each method generates one optimized molecule. For each target pocket, we construct the evaluated off-target set $\mathcal{O}_{\text{eval}}$ by randomly sampling 10 pockets from the remaining 99 CrossDocked2020 test pockets. For off-target-agnostic methods, these off-target pockets are never accessed during optimization and are used only for evaluation. For ActivityDiff, which requires negative targets, we randomly assign 10 pockets as surrogate off-targets, reflecting the assumed inaccessibility of the true off-target set.

Docking timeout. The primary evaluation in Tables 1 and 6 uses an 1800 s timeout per AutoDock Vina run. Timeouts and missing scores count as failures in the docking-success statistic; Vina-score summaries use finite scores. A valid target-pocket docking score is required to compute empirical specificity, so samples whose target-pocket docking fails are excluded from specificity aggregation. Because docking success varies across methods, we report it alongside specificity in the primary evaluation.

Shared outer ligand-growth loop. SurfSpec uses at most $K = 3$ outer ligand-growth iterations. At each iteration, we select the closest feasible unoccupied surface patch within 4–10 Å, generate a linker of size 10 using DiffLinker, and refine the resulting pseudo-label using low-noise anchored recovery. The separate refinement comparison uses $K = 1$ and fixed pseudo-labels. The procedure terminates early if no feasible surface patch exists or if the target-pocket Vina score after local optimization worsens relative to the previous iteration.

Baseline variants. We evaluate PMDM and DecompOpt both in their standard forms and in augmented variants. $\mathrm{MS2}$ and $\mathrm{MS3}$ denote two-step and three-step applications of the corresponding baseline. $\mathrm{MS} + \mathrm{ES}$ denotes the multi-step variant with early stopping, where the stopping rule uses the same target-pocket Vina-score criterion as SurfSpec. These variants test whether SurfSpec's performance can be explained by repeated application of existing lead-optimization baselines or by early stopping alone.

Naive size extension. We include $\mathrm{DiffSBDD\text{-}SizeExt}$ as a naive size-extension baseline because DiffSBDD is the pocket-conditioned diffusion prior used inside SurfSpec. This baseline increases the generated molecule size using DiffSBDD without selecting under-occupied target-surface regions. It therefore tests whether improved geometric mismatch can be obtained simply by adding more atoms, rather than through controlled surface-directed growth.

Surrogate off-target-aware baseline. ActivityDiff is evaluated as a surrogate off-target-aware specificity baseline. Because the true off-target set is unavailable during optimization, we construct its negative guidance using randomly selected negative targets. This setting compares specificity under the selected surrogate negative targets.

Ours. SurfSpec uses the same DiffLinker proposal mechanism as the refinement baselines, but applies a low-noise anchored recovery procedure. After DiffLinker proposes a candidate pseudo-label, we first remove atoms whose distance to any pocket atom is less than 2 Å. This clash-truncation step produces the anchor used for refinement. We then run the low-noise mode-recovery update at $\tau = 0.15$ for 100 steps with step size $\gamma_{\text{proj}} = 0.3$. The anchor-guidance coefficient is set to 0.001. The anchor target is the truncated pseudo-label, and the guidance score is evaluated at the current denoising state. After mode recovery, we perform 50 reverse-SDE steps from $t = \tau$ to $t = 0$ to obtain the refined ligand. In our experiments, this simpler SDEdit-style reverse process was sufficient for refinement, so we omit additional guidance along the reverse trajectory. Finally, the ligand is locally optimized using AutoDock Vina before proceeding to the next outer iteration. The complete procedure is described in Algorithm 1, and Figure 7 illustrates the ligand-growth steps.

Algorithm 1: Full Ligand Growth Pipeline with Geometric Pseudo-Label Generation and Low-Noise Refinement

Require: native target pocket P_{tgt}, initial ligand $x^{(0)}$, pretrained linker generation model p_{linker}, pretrained pocket-conditioned ligand prior $p(x \mid P_{\mathrm{tgt}})$, maximum number of growth iterations K, low noise level τ, anchor strength λ

Ensure: optimized ligand x^{final}

for $k = 0, \ldots, K-1$ **do**
  Construct the current ligand-oriented pocket surface for $x^{(k)}$
  Identify unoccupied pocket surface patches around the current ligand
  Select the closest unoccupied surface patch $r^{(k)}$ whose distance from $x^{(k)}$ lies within 4–10 Å
  if no feasible surface patch $r^{(k)}$ exists **then**
    break
  end if
  Sample a linker from the pretrained linker generation model:

$$z^{(k)} \sim p_{\mathrm{linker}}\Big(z \mid x^{(k)}, r^{(k)}, P_{\mathrm{tgt}}\Big)$$

  Attach the sampled linker to the current ligand
  Remove linker atoms that clash with the pocket under the 2 Å clash threshold
  Obtain the geometric pseudo-label $x^{(k)}_{\mathrm{label}}$
  Define the anchored density around the geometric pseudo-label:

$$p^{\mathrm{anc}}(x \mid P_{\mathrm{tgt}}, x^{(k)}_{\mathrm{label}}) \propto p(x \mid P_{\mathrm{tgt}}) \exp\left(-\frac{\lambda}{2}\|x - x^{(k)}_{\mathrm{label}}\|^2\right)$$

  Recover the low-noise mode $y^{\star(k)}$ of p^{anc}_{τ} by solving the gradient flow:

$$\begin{aligned}\frac{dy_s}{ds} =& \nabla_{y_s} \log p_\tau(y_s \mid P_{\mathrm{tgt}}) \\ & - \lambda J_{\hat{x}_{0|\tau}}(y_s)^\top \left(\hat{x}_{0|\tau}(y_s) - x^{(k)}_{\mathrm{label}}\right)\end{aligned}$$

  Re-noise the recovered low-noise mode:

$$x^{\mathrm{anc},(k)}_{\tau} = y^{\star(k)} + \sigma_\tau \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I)$$

  Starting from $x^{\mathrm{anc},(k)}_{\tau}$, run the reverse stochastic differential equation from $t = \tau$ to $t = 0$:

$$dx_t = \left[f_t(x_t) - g_t^2 \nabla_{x_t} \log p_t(x_t \mid P_{\mathrm{tgt}})\right] dt + g_t \, d\bar{W}_t$$

  Set the refined ligand as the next ligand $x^{(k+1)}$
  Locally optimize the refined ligand $x^{(k+1)}$ with AutoDock Vina (Vina Min)
  if $k > 0$ and affinity drop is observed **then**
    break
  end if
  $x^{\mathrm{final}} = x^{(k+1)}$
end for
return the last refined ligand as x^{final}

Refinement baselines. For Table 2, all refinement baselines are evaluated under a shared one-step pseudo-label recovery setting. They consume the same DiffLinker pseudo-labels and use the same DiffSBDD pocket-conditioned ligand prior, and each method generates one refined sample for each CrossDocked2020 test complex. Thus, the comparison isolates the effect of the recovery procedure rather than the pseudo-label generator, diffusion prior, or outer growth policy. The shared pseudo-labels use the linker size and closest-patch selection rule described above; the recovery comparison holds these inputs fixed. The baseline-specific refinement hyperparameters are summarized in Table 4.

Table 4: Hyperparameters for refinement baselines. Settings for the shared one-step recovery comparison in Table 2. All methods use the same DiffLinker pseudo-labels and DiffSBDD prior.

Method	Key steps	Other hyperparameters
SDEdit	50 reverse steps	$\tau = 0.15$
DPS	1000 reverse steps from $t = 1$	Anchor weight $= 0.001$
DCDP	10 purification iterations; 100 inner consistency steps; 50 reverse steps per purification	Consistency LR $= 0.01$; consistency weight $= 1.0$; purification $\tau : 0.15 \to 0.05$; raw anchor gradients
RED-Diff	50 optimization steps	LR $= 0.001$; anchor weight $= 0.001$; prior weight $= 0.5$; raw anchor gradients
RSD	50 optimization steps; 4 particles	LR $= 0.001$; anchor weight $= 0.001$; prior weight $= 0.5$; repulsion weight $= 0.001$; kernel bandwidth $= 1.0$; raw anchor gradients

C.2 Surface Patch Selection Policy

At each ligand-growth iteration, we select a protein surface patch as the target pseudo-fragment for linker generation. As the selection policy, we first construct visibility-filtered surface candidates and then select the candidate closest to the current ligand. The selected candidate is used as the surface-side fragment for DiffLinker with fixed linker size 10.

Candidate construction. Let $\mathcal{L}$ be the set of ligand heavy-atom coordinates and let $\mathcal{R}$ be the receptor heavy atoms. We first define the local pocket atom set

$$\mathcal{P} = \left\{ x \in \mathcal{R} : \min_{\ell \in \mathcal{L}} \|x - \ell\|_2 \le 10\,\text{Å} \right\}.$$

From $\mathcal{P}$, we extract an outer surface shell $\mathcal{S}$ by retaining the outer 20% of atoms with respect to distance from the pocket centroid. Specifically, for $c_{\mathcal{P}} = |\mathcal{P}|^{-1} \sum_{x \in \mathcal{P}} x$ and $\rho(x) = \|x - c_{\mathcal{P}}\|_2$, we set

$$\mathcal{S} = \{ x \in \mathcal{P} : \rho(x) \ge Q_{0.80}\left(\{\rho(y) : y \in \mathcal{P}\}\right) \}.$$

We first construct a residue pool by selecting protein residues whose closest atom lies 4–10 Å from the current ligand. Each candidate fragment r is represented by the connected carbon atoms forming the backbone of the corresponding residue.

We compute its distance to the extracted surface shell as

$$d_{\text{surf}}(r) = \min_{x \in r,\ s \in \mathcal{S}} \|x - s\|_2.$$

If at least one candidate satisfies $d_{\text{surf}}(r) \le 2.5\,\text{Å}$, we keep only those surface-proximal candidates.

Visibility filtering. We further require each candidate to have line-of-sight to at least one ligand atom. For a candidate fragment r, a ligand atom ℓ_i, and a representative candidate atom $x \in r$, we sample the segment

$$p(t) = (1 - t)\ell_i + tx, \qquad t \in \{0.1, 0.2, \ldots, 0.9\}.$$

Blockers are receptor heavy atoms excluding atoms from the same as r. With clash cutoff $\delta = 2.5\,\text{Å}$, we define

$$f_{\text{block}}(r, \ell_i) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbf{1}\left[\min_b \|p(t) - b\|_2 < \delta \right],$$
$$\mathcal{T} = \{0.1, 0.2, \ldots, 0.9\}.$$

A candidate is retained if there exists a ligand atom ℓ_i such that

$$f_{\text{block}}(r, \ell_i) \le 0.05.$$

Let $\mathcal{C}$ denote the resulting visibility-filtered candidate set.

Closest surface policy. Among the final candidates, we select the fragment with minimum ligand–fragment distance:

$$r^\star = \arg\min_{r\in\mathcal{C}} d_{\mathrm{lig}}(r), \qquad d_{\mathrm{lig}}(r) = \min_{\ell\in\mathcal{L},\ x\in r} \|x-\ell\|_2.$$

If $\mathcal{C}$ is empty, the growth iteration stops. Otherwise, $r^\star$ is used as the surface patch, and DiffLinker grows a linker of size 10 toward $r^\star$.

C.3 Evaluation Metrics

Off-target agnostic flag. For ligand-growth experiments, we indicate whether each method is *off-target agnostic*. A method is marked as off-target agnostic if it does not use off-target structures, scores, or activity labels during generation. This is a method-level property rather than a per-sample metric. Off-target pockets are used only for evaluation unless explicitly stated otherwise.

Clash rate. We measure protein–ligand clashes using a distance-based criterion. A generated ligand is considered clashing if any ligand atom is within 2 Å of any protein atom. The reported clash rate is the fraction of generated ligands that satisfy this condition:

$$\mathrm{ClashRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\min_{a\in L_i,\ b\in P_i} \|a-b\|_2 < 2\,\text{Å}\right].$$

Lower clash rate indicates fewer physically implausible protein–ligand overlaps.

Ligand size. We report the mean number of heavy atoms in the generated ligand. Atom counts are computed using RDKit from the generated SDF files, with hydrogens excluded.

Empirical specificity. We evaluate cross-pocket specificity using AutoDock Vina docking scores. For each generated ligand, we dock it to the target pocket and to a finite evaluated off-target set $\mathcal{O}_{\mathrm{eval}}$. Let $S_{\mathrm{tgt}}(L)$ denote the Vina docking score on the target pocket, and let $S_O(L)$ denote the score on off-target pocket O. Because lower Vina scores indicate stronger predicted binding, we define the empirical specificity estimate as

$$\widehat{\mathrm{Spec}}(L) = -\left(S_{\mathrm{tgt}}(L) - \min_{O\in\mathcal{O}_{\mathrm{eval}}} S_O(L)\right).$$

This metric compares the target-pocket score against the strongest predicted off-target binder among the evaluated off-targets. In addition to the sample average, we report thresholded success rates for $\widehat{\mathrm{Spec}} > 0.2$, $\widehat{\mathrm{Spec}} > 0.4$, and $\widehat{\mathrm{Spec}} > 0.6$.

Geometric mismatch. Geometric mismatch measures ligand–pocket surface complementarity using the Jensen–Shannon distance defined in Eq. (4). For each ligand–pocket pair, we construct the ligand surface and the ligand-oriented target-pocket surface restricted to points within 8 Å of any ligand heavy atom. These surfaces induce probability measures on the common ligand-local domain through the signed-distance-based Boltzmann construction in Appendix B.1. Lower geometric mismatch indicates better agreement between the ligand surface-induced measure and the ligand-oriented target-pocket measure.

Pocket occupancy. Pocket occupancy measures how much of the target pocket is covered by the generated ligand. We define occupancy as the fraction of target-pocket grid points that are occupied by the ligand, where a grid point is counted as occupied if it lies within the predefined ligand fill radius of any ligand atom. We report the mean in Table 1, with first and third quartiles in Table 6. Higher occupancy indicates broader coverage of the target-pocket region.

Target-pocket Vina docking. We report the AutoDock Vina docking score of each generated ligand on the target pocket. This metric is distinguished from the Vina-derived score used during optimization, as it is computed by re-docking the final generated ligand with AutoDock Vina under the evaluation protocol. Because lower Vina scores indicate stronger predicted binding, lower values correspond to better target-pocket docking performance. We report the mean in Table 1, with first and third quartiles in Table 6.

Docking success. Docking success is the fraction of attempted AutoDock Vina dockings that return a finite score under the evaluation timeout. Timeouts and missing scores count as failures. This statistic describes completion of the docking evaluation and is distinct from the specificity threshold success rates.

Refinement prior consistency. For refinement experiments, we evaluate whether a refined ligand remains consistent with the pocket-conditioned ligand prior. We report protein clash rate using the same 2 Å criterion as above, RDKit valence validity, Vina score-only repulsion diagnostics, and bond-length MMD. The repulsion diagnostics use the unweighted repulsion term from Vina `-score_only`. We compute this diagnostic globally on the full receptor and locally on receptor atoms within 4 Å of the fragment neighborhood. Larger repulsion indicates stronger steric conflict and is not interpreted as docking affinity. Local and global affinity values in Table 2 are the negatives of the corresponding Vina scores, $A = -S_{\mathrm{Vina}}$, in kcal/mol; higher values indicate more favorable predicted binding. Table 1 instead reports Vina scores directly, for which lower values are better.

Valence validity. Valence validity is the fraction of expected samples that pass RDKit valence and sanitization checks. Missing ligands and invalid molecules are counted as failures. Higher validity indicates better chemical consistency.

Bond-length MMD. Bond-length MMD measures the discrepancy between the bond-length distribution of generated ligands and the reference distribution from the dataset. We compute maximum mean discrepancy separately for C–C, C–N, and C–O bonds. Lower values indicate that the generated bond geometry is closer to the reference molecular geometry.

Pseudo-label faithfulness. We evaluate whether refinement preserves the intended pseudo-label using RMSD and topological similarity. RMSD is computed against the pseudo-label SDF using the same atom indexing, without MCS remapping:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \|x_j - \tilde{x}_j\|_2^2},$$

where x_j and $\tilde{x}_j$ denote corresponding atoms in the refined ligand and pseudo-label, respectively. Lower RMSD indicates better preservation of pseudo-label geometry. Topological similarity is computed as the Tanimoto similarity between RDKit path fingerprints of the refined ligand and pseudo-label using `RDKFingerprint` with `maxPath=7`. Higher topological similarity indicates better preservation of pseudo-label topology.

D Additional Results

D.1 Geometry–Affinity Calibration

Figure 4 shows the CrossDocked2020 calibration results. We fit a linear trend $\hat{a}(d) = \beta_0 + \beta_1 d$ using the random pairs from the CrossDocked2020 training split ($n = 531$), and define the central $(1-\eta)$-quantile envelope by $f_-^\eta(d) = \hat{a}(d) + q_{\eta/2}$ and $f_+^\eta(d) = \hat{a}(d) + q_{1-\eta/2}$, where q_α is the empirical α-quantile of the training residuals. The fitted 90%, 95%, and 99% envelopes cover 89.8%, 94.7%, and 98.9% of the training samples, respectively, and cover 91.0%, 92.0%, and 97.0% of the 100 CrossDocked2020 test samples. These empirical rates do not establish the nominal coverage levels or coverage transfer to optimized ligands and unseen off-target pairs.

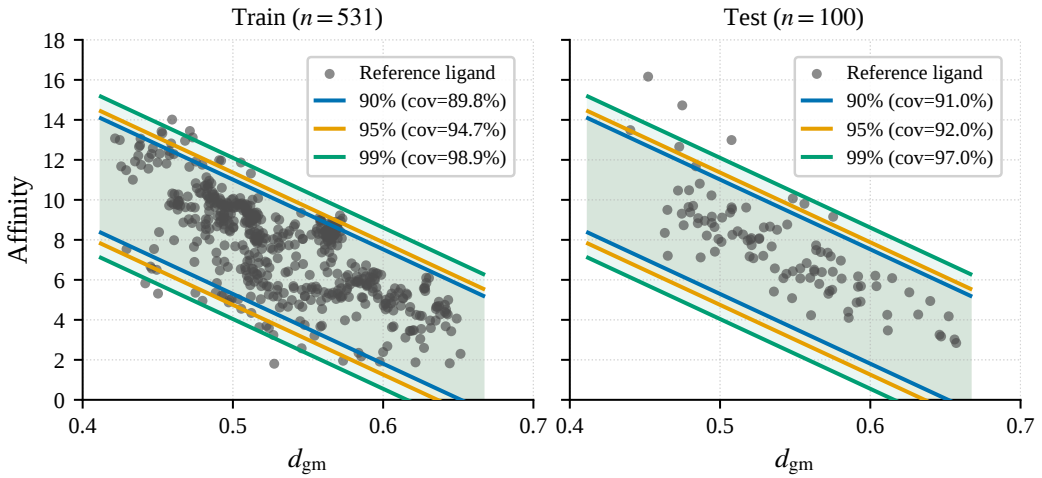


Figure 4: Geometry–affinity calibration on CrossDocked2020. Gray points denote reference ligand–pocket pairs, where affinity is measured as the negative AutoDock Vina score and d_{gm} is the proposed geometric mismatch. Lines show the fitted linear geometry–affinity trend with central 90%, 95%, and 99% empirical quantile envelopes constructed from training residuals.

D.2 Additional Analysis of Pseudo-Label Refinement

The one-step comparison in Table 2 fixes the DiffLinker pseudo-labels and DiffSBDD prior, with $K = 1$ and one refined sample per CrossDocked2020 test complex. Local and global repulsion reveal different aspects of compatibility: DPS has the lowest local repulsion (0.20), whereas SurfSpec has lower global repulsion (4.40 versus 4.88) and a lower clash rate (0.01 versus 0.03).

SurfSpec attains the lowest clash rate and global repulsion among refinement methods, the highest local and global affinity, and best or tied-best bond-length MMD values. All refinement methods have valence validity 1.00, compared with 0.84 for the pseudo-label inputs. SurfSpec's mean RMSD (0.78 Å) is lower than DPS, RED-Diff, and RSD, but higher than SDEdit and DCDP. Its mean topological similarity (0.48) matches DCDP and is below SDEdit, RED-Diff, and RSD, illustrating the tradeoff between structural compatibility and anchor faithfulness.

D.3 Empirical Activation of the Geometric Specificity Certificate

The specificity certificate in Theorem 1 is based on a geometric condition: the target–off-target pocket separation should exceed the target–ligand mismatch. We empirically evaluate how often this condition is activated on randomly sampled target–off-target pocket pairs from the CrossDocked2020 test complexes. For each pair, we use the reference ligand L associated with the target pocket to construct the ligand-oriented pocket measures $\nu^L_{P_{\mathrm{tgt}}}$ and ν^L_O, and compute the target–ligand mismatch $d_{\mathrm{gm}}(L, P_{\mathrm{tgt}})$ and the ligand-oriented target–off-target separation

$$m_L(O) = d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O).$$

We define the geometric certificate margin as

$$M_L(O) = m_L(O) - d_{\mathrm{gm}}(L, P_{\mathrm{tgt}}) \tag{40}$$

$$= d_{\mathrm{JS}}(\nu^L_{P_{\mathrm{tgt}}}, \nu^L_O) - d_{\mathrm{gm}}(L, P_{\mathrm{tgt}}). \tag{41}$$

When $M_L(O) > 0$, the triangle-inequality lower bound

$$d_{\mathrm{gm}}(L, O) \geq [M_L(O)]_+$$

is nonzero. Thus, $M_L(O) > 0$ measures activation of the geometric lower-bound term.

Figure 5 reports the empirical CDF of $M_L(O)$ over the filtered usable sampled pairs. We find that 65.96% of pairs have positive margin, showing that the geometric certificate is activated for a

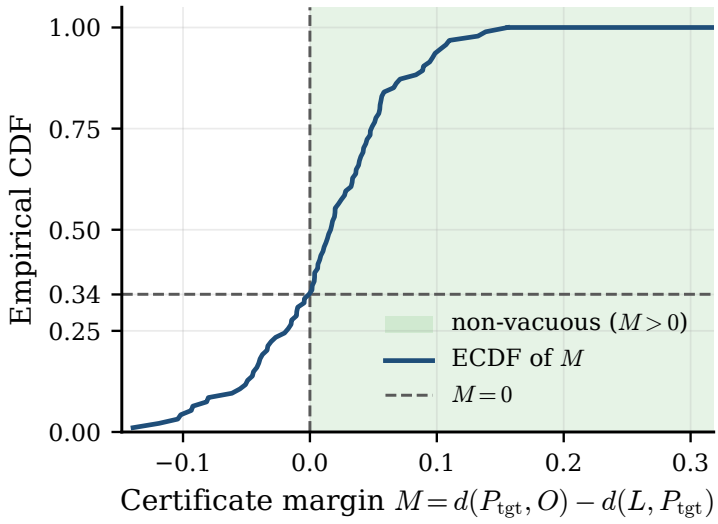


Figure 5: Empirical activation of the geometric specificity certificate. The plot shows the empirical CDF of the certificate margin $M_L(O) = m_L(O) - d_{\text{gm}}(L, P_{\text{tgt}})$, where $m_L(O) = d_{\text{JS}}(\nu^L_{P_{\text{tgt}}}, \nu^L_O)$ is the ligand-oriented target–off-target pocket separation. The shaded region indicates $M_L(O) > 0$, where the triangle-inequality lower bound on ligand–off-target mismatch is nonzero. Overall, 65.96% of filtered usable pairs fall in this activated region.

Table 5: Geometric activation statistics of the specificity certificate on the filtered usable target–off-target pairs. Here $M = M_L(O)$ and $\text{LB} = [M]_+$ denotes the triangle-inequality lower bound on ligand–off-target geometric mismatch.

Metric	Avg.	Std.	Q1	Med.	Q3
$\Pr[M > 0]$	65.96%	–	–	–	–
Margin M	0.014	0.057	-0.017	0.017	0.049
Lower bound LB	0.031	0.037	0.000	0.017	0.049

substantial fraction of evaluated target–off-target pairs. This indicates that the triangle-inequality certificate yields a nonzero ligand–off-target mismatch lower bound in many evaluated cases.

Table 5 summarizes the resulting geometric lower-bound statistics. The positive activation rate shows that the conservative geometric lower-bound term is not purely vacuous on this benchmark.

D.4 Additional Specificity Results

Complete evaluation statistics. Table 6 provides the full statistics for the primary 1800 s evaluation summarized in Table 1. Each target is evaluated against 10 randomly sampled off-target pockets from the remaining CrossDocked2020 test pockets. SurfSpec achieves the highest mean empirical specificity and the highest success rates at all three thresholds, together with the highest mean occupancy, zero clash rate, and full docking success. The size-extension baselines attain low geometric mismatch but have higher clash rates and weaker empirical specificity.

Random-seed sensitivity. Table 7 evaluates a subset of methods under a second random seed for ligand generation and off-target sampling, using the primary 1800 s timeout. SurfSpec has the highest mean empirical specificity and success rates at all three thresholds among the compared methods. This comparison provides additional evidence that the observed ranking persists under the tested change of seed.

Table 6: Full statistics for the primary 1800 s evaluation summarized in Table 1, using the same sampled off-target set. The second line in each specificity cell gives the relative change from the initial lead, $100 \times (x - x_0)/|x_0|$, except for the standard deviation. Blue and red percentages indicate positive and negative changes, respectively; bold percentages mark the largest change in each specificity column. For metric values, bold and underline indicate the best and second-best reported values, including ties; ligand size is descriptive. Docking success is the fraction of attempted Vina dockings returning a finite score under the 1800 s timeout; timeouts and missing scores count as failures.

Method	Off-Target	**Primary Metric:** $\widehat{\text{Spec}}\uparrow$					Geometric Mismatch ↓			Occupancy ↑			Vina Dock ↓			Clash Rate ↓	# Atoms	Dock Success ↑
	Agnostic	Avg.	Std.	> 0.2	> 0.4	> 0.6	Avg.	Std.	Med.	Avg.	Q1	Q3	Avg.	Q1	Q3			
Initial Lead	O	-0.81	1.43	0.15	0.14	0.11	0.540	0.051	0.536	0.26	0.22	0.30	-7.17	-8.67	-5.73	**0.00**	22.75	**1.00**
Delete	O	-0.81 +0%	1.54	0.19 +27%	0.14 +0%	0.13 +18%	0.539	0.052	0.535	0.26	0.22	0.30	-7.23	-8.65	-5.73	**0.00**	22.85	**1.00**
PMDM	O	-1.10 −36%	2.03	0.15 +0%	0.12 −14%	0.12 +9%	0.501	0.056	0.489	0.33	0.27	0.38	-6.47	-8.87	-5.10	0.16	31.00	0.92
DecompOpt	O	-0.93 −14%	1.43	0.13 −13%	0.12 −14%	0.11 +0%	0.542	0.053	0.536	0.25	0.22	0.29	-6.08	-8.26	-4.49	0.02	22.75	0.91
DiffShape	O	-0.88 −9%	1.58	0.13 −13%	0.13 −7%	0.12 +9%	0.585	0.115	0.555	0.14	0.00	0.25	-6.90	-8.33	-5.20	0.11	21.91	0.99
ActivityDiff	X	-0.96 −18%	2.15	0.13 −13%	0.11 −21%	0.10 −9%	0.538	0.054	0.529	0.27	0.23	0.31	-7.07	-8.35	-5.39	0.11	22.89	**1.00**
PMDM–MS2	O	-1.38 −70%	2.19	0.13 −13%	0.13 −7%	0.12 +9%	0.499	0.063	0.482	0.31	0.26	0.37	**-8.95**	-10.04	**-7.01**	0.25	33.42	0.91
PMDM–MS3	O	-2.05 −152%	2.93	0.07 −53%	0.07 −50%	0.06 −45%	0.496	0.069	0.480	0.31	0.26	0.37	-8.38	**-10.51**	-6.43	0.31	35.83	0.82
PMDM–MS+ES	O	-1.02 −26%	2.08	0.19 +27%	0.17 +21%	0.14 +27%	0.500	0.057	0.489	0.33	0.28	0.38	-8.61	-10.14	-6.56	0.16	31.37	0.87
DecompOpt–MS2	O	-1.03 −27%	1.33	0.11 −27%	0.11 −21%	0.08 −27%	0.542	0.053	0.538	0.25	0.22	0.28	-7.08	-8.79	-5.62	0.03	22.75	0.96
DecompOpt–MS3	O	-0.99 −22%	1.39	0.12 −20%	0.09 −36%	0.08 −27%	0.541	0.053	0.537	0.25	0.21	0.28	-7.00	-8.57	-5.72	0.04	22.76	0.97
DecompOpt–MS+ES	O	-0.83 −2%	1.59	0.17 +13%	0.14 +0%	0.14 +27%	0.541	0.052	0.534	0.25	0.22	0.28	-7.40	-8.91	-5.95	0.03	22.75	0.95
DiffSBDD–SizeExt(+20)	O	-1.51 −86%	2.24	0.14 −7%	0.11 −21%	0.10 −9%	0.480	0.037	0.476	0.28	0.19	0.38	-7.57	-8.25	-6.37	0.45	35.88	0.93
DiffSBDD–SizeExt(+30)	O	-2.20 −172%	3.76	0.08 −47%	0.07 −50%	0.07 −36%	**0.468**	0.046	**0.461**	0.32	0.23	**0.42**	-7.14	-8.40	-5.44	0.62	41.58	0.81
Ours	O	**-0.37** +54%	2.54	**0.31** +107%	**0.29** +107%	**0.25** +127%	0.480	0.050	0.467	**0.36**	**0.30**	0.41	-8.63	-10.43	-6.31	**0.00**	34.99	**1.00**

Table 7: Specificity under a second random seed for ligand generation and off-target sampling, using the same 1800 s timeout as the primary evaluation in Table 1. SurfSpec has the highest mean specificity and all three thresholded success rates among the compared methods. Bold indicates the best value; underline indicates the second-best value, including ties.

Method	**Primary Metric:** $\widehat{\text{Spec}}\uparrow$				
	Avg.	Std.	> 0.2	> 0.4	> 0.6
Delete	-0.82	1.59	0.19	0.13	0.12
PMDM–MS+ES	-1.02	1.86	0.19	0.14	0.14
DecompOpt–MS+ES	-0.93	1.36	0.17	0.15	0.14
Ours	**-0.52**	2.21	**0.29**	**0.27**	**0.25**

D.5 Qualitative Results

Figure 6 qualitatively compares optimized ligands on three representative target pockets. SurfSpec expands ligands toward under-filled pocket regions while maintaining plausible pocket-bound conformations. The inset values report the target-pocket geometric mismatch d_{gm}, where lower values indicate better surface complementarity. These examples support that SurfSpec achieves controlled surface-directed growth with low geometric mismatch.

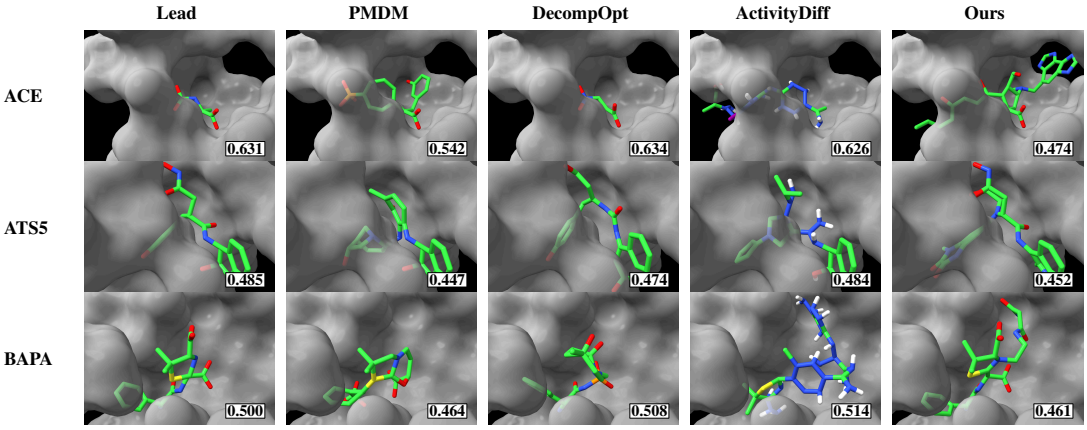


Figure 6: Qualitative comparison of lead optimization results on three representative target pockets. Rows correspond to target pockets, and columns correspond to the initial lead and optimized ligands generated by different methods. The gray surface denotes the target pocket, and the inset value reports the target-pocket geometric mismatch d_{gm}, where lower values indicate better surface complementarity. SurfSpec achieves low geometric mismatch across the shown examples while maintaining plausible pocket-bound conformations.

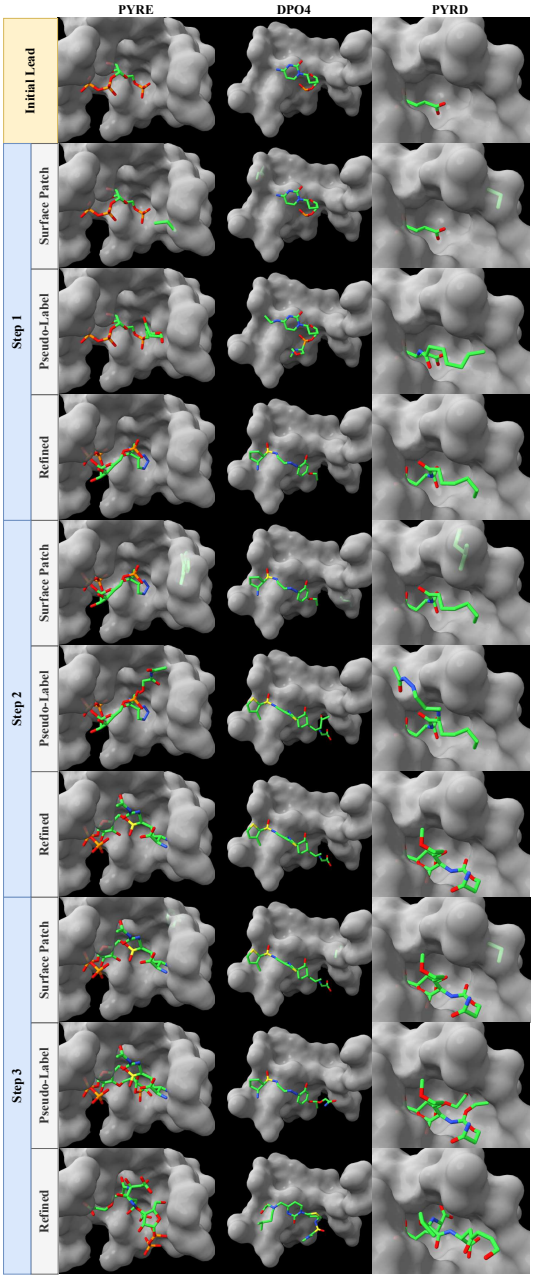


Figure 7: Qualitative visualization of SurfSpec ligand growth on representative CrossDocked2020 complexes. Columns show different target pockets, and rows show the initial lead and the intermediate states across three growth iterations. At each step, SurfSpec first selects an under-occupied target-surface patch, constructs a surface-directed pseudo-label toward that patch, and then refines the pseudo-label under the pocket-conditioned ligand prior. The gray surface denotes the target pocket, the green sticks denote the current ligand, and the pale green sticks indicate the selected surface-directed growth region. SurfSpec progressively expands the ligand toward uncovered pocket regions while maintaining plausible pocket-bound conformations after refinement.