DiffAlign: Diffusion-Based Molecular Alignment with Pocket-Aware Guidance

Iljung Kim
Hanyang University
iljung0810@hanyang.ac.kr

Keehyoung Joo[†]
Korea Institute for Advanced Study
newton@kias.re.kr

Yung-Kyun Noh[†]
Hanyang University / Korea Institute for Advanced Study
nohyung@hanyang.ac.kr

Abstract

We introduce DiffAlign, a conditional $E(3)$-equivariant diffusion framework that incorporates receptor context directly into the sampling process. We demonstrate its application to flexible molecular alignment, which requires jointly exploring a ligand's intramolecular conformation and a pocket-compatible pose relative to a reference ligand. At inference time, DiffAlign injects universal force field (UFF) gradients on the ligand–pocket system, steering trajectories toward low-energy, clash-free, pocket-aware poses without retraining. This inference-time physical guidance contrasts with approaches that incorporate receptor information only post hoc, leaving the generative trajectory unguided. On the *DISCO* cross-docking benchmark (a curated set of protein-ligand complexes), pocket-aware guidance improves Top-1 success rate over a no-steering baseline by +1.0/+1.7/+2.1 percentage points at 1/2/3 Å ($\sim$10–20% relative; up to 20% at 1 Å) and consistently outperforms ligand-only guidance. *PoseBusters*, a recent validation suite for stereochemical plausibility, further confirms chemical validity ($\geq$97% for local geometry; $\geq$95% for clash avoidance). Furthermore, DiffAlign produces diverse candidate sets, highlighting headroom for rescoring and integration into downstream tasks such as docking and virtual screening.

1 Introduction

Molecular alignment is a fundamental task in cheminformatics and structure-based drug discovery. Given a query ligand and a reference—either a known active ligand [1, 2] or a protein-binding pocket [3]—the goal is to generate a 3D superposition that preserves global shape and pharmacophoric features for quantitative structure–activity relationship modeling and ligand-based virtual screening [4, 5]. The more challenging case is flexible molecular alignment, where the intramolecular conformation must be jointly adapted with the global pose. Classical pipelines rely on pre-computed conformer libraries combined with rigid or semi-rigid superposition [6, 7, 8]. However, bioactive conformations in binding sites often deviate from low-energy gas-phase minima [9]; if pocket geometry and receptor–ligand interactions are ignored, superpositions clash with pocket atoms, break key contacts, and adopt strained torsions. These failure modes motivate a data-driven approach that learns pocket-compatible alignments rather than enumerating conformers.

Recent advances in denoising diffusion models [10], combined with $E(3)$-equivariant graph neural networks, have enabled learning expressive distributions over 3D molecular coordinates [11, 12].

[†]Co-corresponding authors: newton@kias.re.kr, nohyung@hanyang.ac.kr.
Inference code and scripts are available at https://github.com/kim-iljung/DiffAlign.

Machine Learning for Structural Biology Workshop

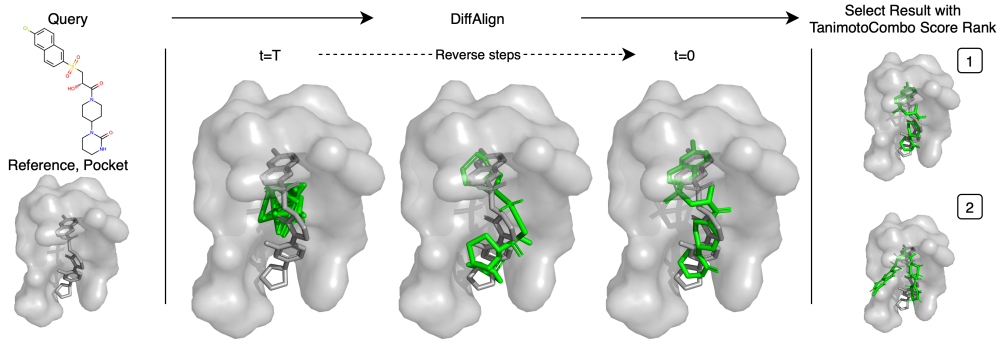

Figure 1: Overview of DiffAlign. Left: inputs are the query ligand, reference ligand, and protein pocket. Center: random initial poses are denoised via reverse diffusion. Right: sampled poses are ranked by TanimotoCombo to produce the final prediction.

Conditioning on a reference is straightforward: one can learn $p_\theta(x_{\text{query}} \mid x_{\text{ref}})$ and directly sample aligned conformations, while classifier-free guidance (CFG) [13] enforces fidelity to the conditioning signal. Yet as a purely data-driven prior, such models can still yield geometrically or physically implausible samples—non-ideal bond lengths or angles, stereochemical errors, and steric clashes—and, critically, poses that match the reference ligand while violating steric or electrostatic constraints of the binding pocket. These observations point us toward two directions: *pocket-aware* alignment, where receptor geometry and forces actively constrain generation, and *inference-time* control, i.e., imposing such constraints during sampling rather than only in training or post-processing.

This work. Previous approaches typically incorporate receptor information only after sampling, providing no pocket-aware signal to guide the denoising trajectory. In contrast, DiffAlign brings physics directly into the sampler. Our method integrates a conditional $E(3)$-equivariant diffusion model with a plug-and-play physical prior by injecting Universal Force Field (UFF) [14] gradients during denoising diffusion implicit models (DDIM) [15] updates. To the best of our knowledge, this is the first inference-time fusion of a learned diffusion prior and a pocket-aware physical prior for flexible molecular alignment, transforming receptor context into a test-time signal rather than a post-hoc correction. This in-sampler steering framework treats the diffusion model as a learned generative prior and UFF as a feasibility prior, biasing reverse-time dynamics toward low-energy, clash-free, and pocket-compatible poses without retraining, while preserving $E(3)$ equivariance and compatibility with classifier-free guidance and other differentiable scorer functions.

Contributions. Our main contributions are: (i) *Diffusion-based molecular alignment with inference-time pocket-aware steering.* We present the first diffusion-based framework for flexible molecular alignment that injects UFF energy gradients from the ligand–pocket system directly into DDIM updates, producing low-energy, clash-free, pocket-compatible poses without retraining. (ii) *Empirical validation on DISCO benchmark.* Pocket-aware guidance improves Top-1 success rates over the no-steering baseline and consistently outperforms ligand-only guidance. (iii) *Stereochemical plausibility and extensibility.* PoseBusters validation shows that physics-guided inference preserves chemical validity. Moreover, DiffAlign is a modular, plug-and-play platform extensible to alternative molecular mechanics or machine-learned score functions.

2 Methods

2.1 Problem Setup, Data, and Architecture

We study conditional 3D generation of a query ligand given a reference structure with coordinates $x_r \in \mathbb{R}^{N_r \times 3}$ and graph $G_r = (V_r, E_r)$. The model operates on (x_t, x_r, G_q, G_r), where $x_t \in \mathbb{R}^{N_q \times 3}$ are noisy query coordinates at diffusion step t and $G_q = (V_q, E_q)$ is the query graph with atom-type node embeddings and bond/distance edge features. The denoiser outputs an $E(3)$-equivariant vector field on the query atoms,

$$\hat{\varepsilon}_\theta(x_t, x_r, G_q, G_r) \in \mathbb{R}^{N_q \times 3}, \tag{1}$$

which transforms equivariantly under any joint roto-translation of (x_t, x_r). Geometry is parameterized purely in Cartesian coordinates.

For training and quantitative evaluation, we curate $\sim\mathbf{65{,}000}$ aligned pairs from GEOM-Drugs [16] by filtering with 2D Tanimoto in $[0.55, 0.85]$, initializing with conventional alignment, and retaining pairs with 3D shape score > 0.5. Each pair provides (x_0, x_r, G_q, G_r).

Recent structure-based generative models often operate in torsional space to strictly enforce rigid bond constraints [17, 18]. However, we choose to diffuse atomic Cartesian coordinates in $\mathbb{R}^{3N_q}$. From an energy–barrier perspective, operating in unconstrained Cartesian space allows the system to take short detours that transiently relax bond lengths and angles. This flexibility enables small collective displacements of many atoms, effectively lowering saddle heights between conformers and alleviating steric locks that can trap rigid-body sampling. Furthermore, Cartesian diffusion naturally couples local shape changes with global translation/rotation, which provides a unified inductive bias for alignment tasks where simultaneous pose-and-conformation adjustment is required.

In docking contexts we also consider a protein pocket defined as all protein heavy atoms within a cutoff of the reference ligand, with coordinates $x_p \in \mathbb{R}^{N_p \times 3}$ and graph $G_p = (V_p, E_p)$. The pocket (x_p, G_p) is used only by the physics prior (UFF) at inference time and is not fed to the denoiser.

2.2 Physics-Guided Sampling (Plug-and-Play UFF)

We adopt a plug-and-play physical prior to inject pocket awareness into sampling. The Universal Force Field (UFF) energy is decomposed as

$$E_{\mathrm{UFF}} \;=\; \underbrace{E_R + E_\theta + E_\phi + E_\omega}_{\text{valence (ligand)}} \;+\; \underbrace{E_{\mathrm{vdW}} + E_{\mathrm{el}}}_{\text{nonbonded (ligand–pocket)}} \quad . \tag{2}$$

Valence interactions—bond stretching (E_R), angle bending (E_θ), dihedral rotation (E_ϕ), and out-of-plane inversion (E_ω)—act within the ligand only. Nonbonded interactions—van der Waals (E_{vdW}) and, when partial charges are available, electrostatics (E_{el})—are evaluated on the joint ligand–pocket system. The pocket is treated as rigid; we update ligand coordinates only. These cross terms provide the pocket-aware forces that repel clashes and guide the ligand toward favorable regions. (We consider a single-ligand setting, so no ligand–ligand terms appear.)

Inference-Time Physical Guidance. We bias the learned prior p_θ with a physical expert evaluated at the current clean estimate:

$$\pi_\beta(x_t \mid x_r, G_q, G_r, x_p, G_p) \propto p_\theta(x_t \mid x_r, G_q, G_r)\ \exp\Big[-\beta(t)\ E_{\mathrm{UFF}}\big(\hat{x}_0(x_t);\ x_p, G_q, G_p\big)\Big], \tag{3}$$

where $\hat{x}_0(x_t)$ is the sampler's clean reconstruction at step t. To approximate sampling from this energy-weighted distribution, we apply gradient guidance directly to the predicted clean coordinates. In practice, we take a small descent step in x_0-space,

$$\tilde{x}_0 \;=\; \hat{x}_0 \;-\; \lambda(t)\,\nabla_x E_{\mathrm{UFF}}\big(\hat{x}_0;\ x_p, G_q, G_p\big), \tag{4}$$

then back-project $\tilde{x}_0$ to the sampler state and apply the usual DDIM update; see Appendix A for algorithmic details. The weight $\lambda(t)$ follows a late–strong ramp (weak early, strong late) so that physical forces act primarily once geometry is coherent; schedule details are in Appendix B.

Properties. Because UFF is $E(3)$-invariant, its gradients transform equivariantly with coordinates; injecting them at the coordinate level preserves the sampler's $E(3)$-equivariance. The mechanism is plug-and-play—no retraining—and remains compatible with classifier-free guidance and other differentiable scorers (by summing gradients). Empirically, valence terms protect stereochemistry while ligand–pocket cross terms remove clashes and improve pocket compatibility. Because chemical space is astronomically large and unevenly sampled, many valid bond/angle/torsion patterns are absent from training; a purely data-driven denoiser extrapolates poorly in these out-of-support regions, whereas UFF supplies physically grounded corrective gradients.

2.3 Training objective and augmentations.

We train with the standard ε-prediction objective

$$\mathcal{L}_\varepsilon = \mathbb{E}_{(x_q, x_r, G_q, G_r, t, \varepsilon)}\big[\|\varepsilon - \hat{\varepsilon}_\theta(x_t, x_r, G_q, G_r, t)\|_2^2\big], \tag{5}$$

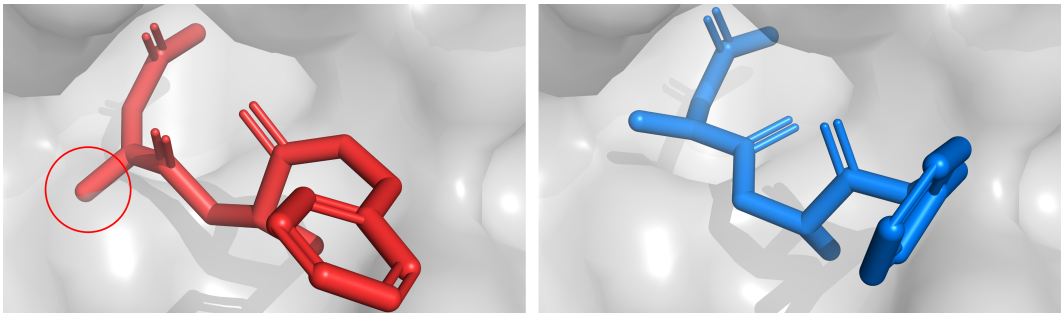

Figure 2: Pocket-aware steering resolves ligand–pocket clashes. Left: collision (red ligand vs gray pocket, circled). Right: steering yields a clash-free pose (blue).

where the expectation is taken over training pairs (x_q, x_r, G_q, G_r), the diffusion timestep t and Gaussian noise ϵ. No auxiliary losses are used: despite the absence of clash or bond-length penalties during training, the inference-time UFF steering with the late–strong schedule yields chemically plausible, low-energy samples in practice, making extra supervision unnecessary.

3 Experiments and Results

We evaluate DiffAlign on the DISCO cross-docking benchmark [19], utilizing a filtered subset of 4,035 complexes across 95 protein targets following prior work. For each target, the binding pocket is defined as all protein heavy atoms within 6 Å of the reference ligand. Given a target pocket and a query ligand, we generate 30 conditional DDIM samples using the late–strong, SNR-aware schedule described in Appendix B. We implement the ligand–pocket UFF prior in PyTorch, and generating 30 pocket-guided samples for a single complex takes about 4 s in our current setup. Note that the receptor pocket information is used solely by the inference-time physics prior (UFF) and is not explicitly provided to the denoiser network.

From the 30 generated candidates, we select a single Top-1 pose by maximizing the *TanimotoCombo* score between the generated query and the reference ligand. We employ an open-source, RDKit-based implementation of TanimotoCombo, which sums (i) *shape* similarity (RDKit shape Tanimoto) and (ii) *color* similarity (Gaussian-overlap of pharmacophore features). The resulting score ranges in $[0, 2]$. Success is reported based on the RMSD of the selected Top-1 pose relative to the crystal pose at thresholds of 1, 2, and 3 Å. For completeness, Top-k ($k \in \{3, 5\}$) and best-of-30 results are provided in Table 3 (Appendix C.2).

We compare DiffAlign against **AutoDock Vina** (docking reference) and **Flexi-LS-align** (alignment baseline), alongside four DiffAlign variants: (i) **AutoDock Vina** [20]: A widely used *ab initio* docking program, benchmarking receptor-based search performance without reference ligand guidance. (ii) **Flexi-LS-align** [21]: A classical flexible-superposition method. (iii) **DiffAlign (no steering)**: The diffusion model without physical guidance. (iv) **DiffAlign + UFF (ligand-only)**: Sampling with intramolecular UFF guidance. (v) **DiffAlign + Post-UFF (with pocket)**: Post-optimization of sampled poses against ligand–pocket UFF energy. (vi) **DiffAlign + UFF (with pocket)**: Our proposed method injecting ligand–pocket UFF gradients directly into DDIM updates. For all alignment methods (DiffAlign variants and Flexi-LS-align), the Top-1 selection follows the TanimotoCombo maximization described above, whereas Vina selects the conformation with the lowest internal energy score. All methods are initialized from random poses (or noise) to ensure an unbiased search.

Summary. Pocket-aware guidance substantially outperforms the no-steering baseline and beats the classical method (**Flexi-LS-align**) by **+3.4 points** at the strictest 1 Å threshold (Table 1). Notably, this matches the high-precision performance of the docking reference (**AutoDock Vina**), demonstrating that generative alignment can approach docking accuracy even when ranking by reference similarity. Furthermore, Table 3 highlights significant generative potential: in the Best-of-30 setting, DiffAlign achieves 10.9% success at 1 Å (more than double the baseline), while maintaining high stereochemical plausibility ($\geq$95% pass rates) as confirmed by PoseBusters.

Table 1: Success rates (%) on DISCO benchmark. We compare DiffAlign with the flexible alignment baseline (Flexi-LS-align). Bold numbers indicate the best performance among alignment-based methods.

Method	RMSD < 1 Å	RMSD < 2 Å	RMSD < 3 Å
Flexi-LS-align	2.6	14.6	27.2
DiffAlign (no steering)	5.0	16.7	25.8
DiffAlign + UFF (ligand only)	5.1	17.7	27.8
DiffAlign + Post-UFF (with pocket)	5.5	**19.2**	**29.0**
DiffAlign + UFF (with pocket)	**6.0**	18.4	27.9
*AutoDock Vina (Docking)	6.0	21.0	30.9

* AutoDock Vina is a widely used docking program and is reported here as a reference baseline.

Table 2: Stereochemistry/physical plausibility pass rates (%) on DISCO Top-1 poses, assessed with PoseBusters.

Method	Bond angles ↑	Bond lengths ↑	Double bond flatness ↑	Internal steric clash ↑
DiffAlign (no steering)	93.1	90.5	97.0	54.0
DiffAlign + UFF (ligand only)	**98.2**	**99.5**	99.1	**96.2**
DiffAlign + Post-UFF (pocket)	81.2	95.6	**99.3**	88.5
DiffAlign + UFF (pocket)	97.8	99.3	98.6	95.4

3.1 Stereochemistry and Physical Plausibility (PoseBusters)

To assess chemical validity independently of pose RMSD, we evaluate Top-1 predictions using the *PoseBusters* program [22]. We report pass rates (%) across four categories: *bond angles*, *bond lengths*, *double-bond planarity/flatness*, and *intramolecular steric clashes* (higher is better ↑).

Inference-time UFF guidance substantially improves stereochemical plausibility, raising local-geometry pass rates above 97% and steric-clash avoidance above 95% (Table 2). An important caveat is that the pocket-aware Post-UFF refinement can introduce additional internal strain: its bond-angle pass rate drops relative to ligand-only UFF, indicating an occasional trade-off between improved pocket compatibility and ideal valence geometry.

4 Conclusion & Future Work

We presented DiffAlign, a conditional $E(3)$-equivariant diffusion framework that integrates UFF energy gradients from the ligand–pocket system directly into DDIM updates, enabling pocket-aware alignment without retraining. On the DISCO benchmark, pocket-aware steering improves Top-1 accuracy over the no-steering baseline by +1.0/+1.7/+2.1 percentage points at 1/2/3 Å (Table 1), while PoseBusters validation confirms high stereochemical plausibility ($\geq$97% for local geometry; $\geq$95% for clash avoidance; see Table 2). These results demonstrate that physics-guided inference enhances pocket compatibility without compromising chemical validity. Importantly, DiffAlign is a general steering platform rather than an endorsement of UFF as the optimal energy function. The physical prior is modular and can be replaced with alternative molecular mechanics or machine-learned energies, enabling applications in template-free, pocket-conditioned de novo design and advanced sampling strategies (e.g., tempering, rare-event methods) under the same inference-time control. As future work, we plan to extend our framework to use Vina-style scoring functions as additional steering guidance, and to evaluate DiffAlign on other benchmark sets beyond DISCO to assess its robustness across diverse protein–ligand systems.

Acknowledgments and Disclosure of Funding

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) [No. RS-2021-II212068, No. RS-2023–00220628]. This work was supported by the Center for Advanced Computation at Korea Institute for Advanced Study.

References

[1] Richard D. Cramer, David E. Patterson, and Jeffrey D. Bunce. Comparative molecular field analysis (CoMFA). 1. effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18):5959–5967, 1988. PMID: 22148765.

[2] Gerhard Klebe, Ute Abraham, and Thomas Mietzner. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry*, 37(24):4130–4146, 1994. PMID: 7990113.

[3] Ajay N. Jain. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry*, 46(4):499–511, 2003. PMID: 12570372.

[4] Thomas S. Rush, J. Andrew Grant, Lidia Mosyak, and Anthony Nicholls. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry*, 48(5):1489–1495, 2005. PMID: 15743191.

[5] Hanna Geppert, Martin Vogt, and Jürgen Bajorath. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, 50(2):205–216, 2010. PMID: 20088575.

[6] J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry*, 17(14):1653–1666, 1996.

[7] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, and Matthew T. Stahl. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, 50(4):572–584, 2010. PMID: 20235588.

[8] Anthony Nicholls, Georgia B. McGaughey, Robert P. Sheridan, Andrew C. Good, Gregory Warren, Magali Mathieu, Steven W. Muchmore, Scott P. Brown, J. Andrew Grant, James A. Haigh, Neysa Nevins, Ajay N. Jain, and Brian Kelley. Molecular shape and medicinal chemistry: A perspective. *Journal of Medicinal Chemistry*, 53(10):3862–3886, 2010. PMID: 20158188.

[9] Emanuele Perola and Paul S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry*, 47(10):2499–2510, 2004. PMID: 15115393.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[11] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9558–9568. PMLR, 18–24 Jul 2021.

[12] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[14] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[16] Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

[17] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.

[18] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

[19] Shayne D. Wierbowski, Bentley M. Wingert, Jim Zheng, and Carlos J. Camacho. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Science*, 29(1):298–305, 2020.

[20] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. PMID: 34278794.

[21] Jun Hu, Zi Liu, Dong-Jun Yu, and Yang Zhang. LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics*, 34(13):2209–2218, 2018.

[22] Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130–3139, 2024.

A Algorithmic Details

A.1 Pocket-aware UFF-guided DDIM with x_0 Refinement

Algorithm A.1. Pocket-aware UFF-guided DDIM ($\eta = 1$; single sample; no UFF norm/clip). We apply UFF on the combined ligand–pocket system at the predicted clean sample; the pocket is fixed and the UFF gradient is taken w.r.t. ligand coordinates only.

`Algorithm A.1: Pocket-aware UFF-guided DDIM (`$\eta = 1$`) with` x_0 `refinement`

Input: $x_T \sim \mathcal{N}(0, I)$; ligand graph G_q; pocket (x_p, G_p);
schedule $\{\bar{\alpha}_t\}_{t=0}^{T}$ with $\bar{\alpha}_0 = 1$; CFG weight w; base step λ_0;
refinement schedule $\lambda(t) = \lambda_0\,(1 - \sigma_t)$; small constant $\varepsilon_\sigma > 0$.

for $t = T, \ldots, 1$ **do**

1. Predict $\hat{\varepsilon}_{\text{cond}}, \hat{\varepsilon}_{\text{uncond}}$ with the equivariant denoiser.
2. CFG: $\hat{\varepsilon}_{\text{cfg}} = (1 + w)\hat{\varepsilon}_{\text{cond}} - w\hat{\varepsilon}_{\text{uncond}}$.
3. Clean estimate: $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$, $\hat{x}_0^{\text{cfg}} = (x_t - \sigma_t \hat{\varepsilon}_{\text{cfg}})/\sqrt{\bar{\alpha}_t}$.
4. UFF gradient (ligand only, pocket fixed):
   $g_0 = \nabla_{x_0} E_{\text{UFF}}\big(\hat{x}_0^{\text{cfg}};\, x_p, G_q, G_p\big)$.
5. x_0-refine (update ligand only): $\tilde{x}_0 = \hat{x}_0^{\text{cfg}} - \lambda(t)\, g_0$.
6. Back-project (numerical guard): $\hat{\varepsilon}' = (x_t - \sqrt{\bar{\alpha}_t}\,\tilde{x}_0)/\max(\sigma_t, \varepsilon_\sigma)$.
7. Noise scale (stochastic DDIM, $\eta = 1$): $\tilde{\sigma}_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\left(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}\right)$.
8. Sample $z \sim \mathcal{N}(0, I)$ if $t > 1$ (else $z = 0$).
9. DDIM update: $x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\tilde{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \tilde{\sigma}_t^2}\;\hat{\varepsilon}' + \tilde{\sigma}_t\, z$.

end for; **return** $\tilde{x}_0$ (since $\bar{\alpha}_0 = 1$, $\tilde{\sigma}_1^2 = 0$ and $x_0 = \tilde{x}_0$).

B SNR-Aware Step Sizes and Stability

Noise, SNR, and the case for late–strong. We assume the standard variance-preserving parameterization $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sigma_t\, \varepsilon$ with $\bar{\alpha}_0 = 1$ and $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$. Define the step-wise signal-to-noise ratio

$$\text{SNR}_t = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t} \quad \Rightarrow \quad \sigma_t = \frac{1}{\sqrt{1 + \text{SNR}_t}}. \tag{6}$$

Early in sampling (small $\bar{\alpha}_t$, large σ_t), geometry is noise-dominated and physical forces are uninformative; near the end (large $\bar{\alpha}_t$, small σ_t), the structure is coherent and a stronger physical pull is beneficial. Hence a late–strong ramp is appropriate.

Default schedule used in our experiments. We scale the x_0-refinement step size with

$$h(t) = 1 - \sigma_t = 1 - \frac{1}{\sqrt{1 + \text{SNR}_t}}, \qquad \lambda(t) = \lambda_0\, h(t). \tag{7}$$

This choice is schedule-agnostic across common variance-preserving schedules (e.g., cosine or linear), starts near zero when the structure is noisy, and saturates smoothly as the estimate becomes clean.

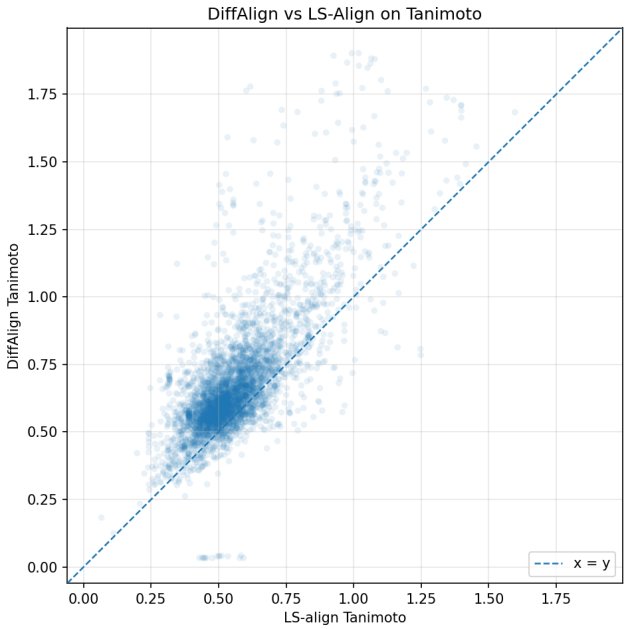

Figure 3: **DiffAlign vs. LS-align on TanimotoCombo (Top-1).** Each point corresponds to one DISCO target (4,035 total). The x-axis shows the top-1 TanimotoCombo (shape+color) score for LS-align, and the y-axis shows the top-1 score for DiffAlign. Points on or above the diagonal $y{=}x$ indicate that DiffAlign matches or outperforms LS-align on the same target, and most points lie on or above this line, indicating a stronger proposal distribution for DiffAlign. A small cluster of outliers near $y \approx 0$ marks severe failures where DiffAlign returns degenerate poses with near-zero TanimotoCombo; these typically involve query molecules that are substantially larger than those seen during training and occasional collapse toward a near single-point configuration during alignment (see Appendix C.1).

Table 3: Success rates (%) on DISCO cross-docking benchmark set (4,035 complexes).

Method	RMSD $<$ 1 Å	RMSD $<$ 2 Å	RMSD $<$ 3 Å
Flexi-LS-align (Best-of-30)	4.8	24.3	43.1
DiffAlign + UFF (pocket, Top-3)	7.9	23.0	33.8
DiffAlign + UFF (pocket, Top-5)	8.9	25.2	36.7
DiffAlign + UFF (pocket, Best-of-30)	10.9	31.2	49.4

C Detailed Results

C.1 Failure Modes and Outliers: Points Near $y \approx 0$ and Single-Point Collapse

A small fraction of targets appear as outliers near $y \approx 0$ in Figure 3, where DiffAlign's top-1 TanimotoCombo score is nearly zero, despite LS-align achieving moderate values on the same targets. These failures are primarily driven by query ligands that are substantially larger than those observed during training, representing an out-of-distribution (OOD) size regime. In such cases, the sampler can collapse into a near single-point configuration, yielding degenerate poses with negligible shape or pharmacophore overlap.

C.2 Top-k and Best-of-n Performance on *DISCO*

Table 3 extends the Top-1 results by reporting Top-k and best-of-n success rates on DISCO (4,035 complexes). We define Top-k success at threshold $\tau \in \{1, 2, 3\}$ Å as

$$S_k(\tau) \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\min_{j\le k} \mathrm{RMSD}\big(\hat{x}_i^{(j)}, x_i^{\star}\big) < \tau\right], \tag{8}$$

where $\hat{x}_i^{(j)}$ is the j-th ranked prediction for target i under the same *TanimotoCombo*-based ranker used in the main experiments. In contrast, best-of-n uses an oracle selector over $n = 30$ samples (i.e., it reports the fraction of targets whose best among 30 proposals falls below the threshold). Consequently, best-of-30 serves as an upper bound on what an ideal ranker could extract from the proposal distribution and isolates proposal quality from ranking.

Key observations. (i) **Proposal quality.** Pocket-Aware Steering (best-of-30) consistently outperforms Flexi-LS-align (best-of-30) across all thresholds: 10.9 vs. 4.8 at 1 Å ($+6.1$ abs.; $\sim$127% rel.), 31.2 vs. 24.3 at 2 Å ($+6.9$ abs.; $\sim$28.4% rel.), and 49.4 vs. 43.1 at 3 Å ($+6.3$ abs.; $\sim$14.6% rel.). This indicates that DiffAlign generates richer proposal sets containing more near-native poses than the classical baseline.

(ii) **Ranking headroom.** The gap between Top-5 and best-of-30 for Pocket-Aware Steering is modest at 1 Å ($+2.0$ points; 8.9 $\rightarrow$ 10.9) but widens at 2/3 Å ($+6.0$ and $+12.7$ points, respectively). This pattern suggests that many valid near-native poses are already present in the 30-sample pool but are not consistently promoted into the top ranks by our simple shape+color ranker, especially at looser thresholds where multiple acceptable poses exist. Stronger pocket-aware rescoring (e.g., docking or learned pose scorers) should therefore close a substantial portion of the Top-k $\rightarrow$ best-of-30 gap.

(iii) **Efficiency at small k.** Even with small budgets, Top-5 Pocket-Aware Steering already surpasses LS-align best-of-30 at 1/2 Å (8.9 vs. 4.8; 25.2 vs. 24.3), while trailing slightly at 3 Å (36.7 vs. 43.1). This reinforces that DiffAlign provides a stronger proposal distribution, with remaining deficit driven by ranking rather than generation.

Takeaway. Pocket-aware, inference-time physics improves not only Top-1 accuracy but also the diversity and quality of generated candidates. The significant best-of-30 gains highlight the strength of the proposal distribution, while the Top-k $\rightarrow$ best-of-30 gap points directly to the need for more effective rescoring strategies.

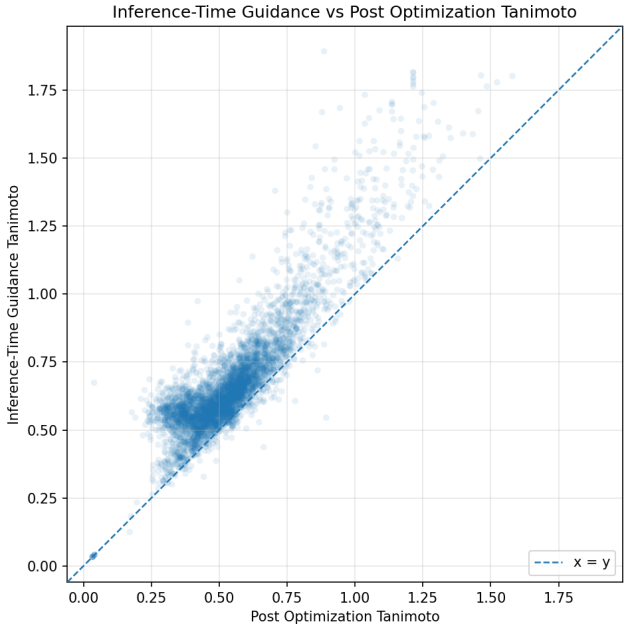

Figure 4: **Comparison of inference-time guidance vs. post optimization on Top-1 Tanimoto-Combo scores** (DISCO benchmark, 4,035 targets). Each points represents a target. The x-axis shows the Top-1 TanimotoCombo (shape+color) score for DiffAlign with post optimization, and the y-axis shows the Top-1 score for DiffAlign with inference-time guidance. Points above the diagonal $y{=}x$ indicate cases where inference-time guidance outperforms post optimization. The majority of points lie above the line, highlighting that inference-time guidance yields a consistently stronger proposal distribution.