MOLFRA: ONLINE ALIGNMENT OF MULTIMODAL MOLECULAR FLOWS VIA FORWARD-PROCESS REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained generative models for structure-based drug design (SBDD) learn structural regularities from protein–ligand complexes without directly optimizing task-specific molecular properties. Controllable approaches pursue these objectives through sampling-time guidance or iterative optimization, but improving properties while retaining molecular diversity remains a challenge. We introduce **MOLFRA**, an online post-training method for pocket-conditioned multimodal molecular flows. Shared pocket-relative rewards jointly guide coordinate and atom-type generation through paired positive and negative branches in endpoint and logit spaces. The logit branches preserve valid atom-type predictions, and updates retain the pretrained sampler without requiring sampling-trajectory likelihood ratios or backpropagation through sampling. On CrossDocked2020, post-training MOLFORM with a joint Vina Score and synthetic accessibility (SA) reward significantly improves mean Vina Dock (from -7.50 to -9.25 kcal/mol) and normalized SA (from 0.60 to 0.74), while fully preserving the pretrained baseline's molecular diversity (0.78 vs. 0.78).

1 INTRODUCTION

Structure-based drug design (SBDD) (Anderson, 2003) aims to generate 3D molecules that bind to a target protein pocket while balancing drug-likeness and synthetic accessibility (Segall, 2012). Pocket-conditioned generators learn structural regularities from protein–ligand complexes to produce diverse candidate structures (Peng et al., 2022; Guan et al., 2023a). Their pretraining objectives do not directly optimize task-specific property preferences. Post-training must therefore improve targeted molecular properties while retaining the pretrained generator's diversity.

Recent objective-aware approaches address this gap through sampling-time property guidance (Dorna et al., 2024; Cremer et al., 2024), gradient-guided or iterative molecular optimization (Qiu et al., 2025; Zhou et al., 2024a), and offline preference alignment (Gu et al., 2024; Cheng et al., 2025). They expose an *optimization–exploration tension*: concentrating generation on desired properties can reduce candidate diversity (Cremer et al., 2024; Choi et al., 2025). For SBDD, this balance also requires joint treatment of continuous coordinates and discrete atom identities, since molecular properties depend on both modalities (Qiu et al., 2025).

Online post-training for SBDD must accommodate both coordinate and atom-type predictions. Multimodal flows generate them through coupled coordinate dynamics and categorical jumps (Campbell et al., 2024; Huang & Zhang, 2025), but their output constraints differ: coordinates admit affine extrapolation, whereas atom-type probabilities must remain on the simplex. We seek joint molecular updates that preserve this hybrid sampler. Trajectory-based RL relies on stochastic transitions and tractable transition probabilities (Black et al., 2024; Liu et al., 2025). Forward-process reconstruction offers an alternative: DiffusionNFT (Zheng et al., 2026) learns from re-corrupted generations through implicit positive and negative branches. For SBDD, this motivates a reconstruction objective that couples both molecular outputs under shared reward feedback.

We introduce MOLFRA (Molecular Forward-process Reward Alignment) for online post-training of pocket-conditioned molecular generators. It scores fresh rollouts from an exponential moving average (EMA) anchor with black-box molecular objectives. Shared pocket-relative rewards weight

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.

paired coordinate-endpoint and atom-type-logit branches, combining self-normalized coordinate regression with categorical cross-entropy. The logit construction preserves valid atom-type predictions. Updates require neither trajectory likelihood ratios nor differentiation through the sampler. We instantiate MOLFRA on MOLFORM (Huang & Zhang, 2025), whose coupled flow/jump dynamics and 100-step log-time sampler provide a practical testbed for frequent online rollouts.

Our contributions are summarized as follows:

- **Joint reward alignment for structure-based drug design.** We jointly fine-tune coordinate and atom-type generation using pocket-relative molecular feedback, with endpoint–logit branches that retain the pretrained hybrid flow/CTMC sampler.
- **Docking and molecular-property gains on CrossDocked2020.** On CrossDocked2020, post-training MOLFORM improves mean Vina Dock from -7.50 to -9.25 kcal/mol and mean normalized SA from 0.60 to 0.74, while retaining reported molecular diversity (0.78 vs. 0.78). MOLFRA attains the lowest reported mean Vina Dock and PoseCheck clash count (2.28) among the compared methods.
- **Interpretation of the training objective.** We analyze categorical branching and its cross-entropy objective to explain valid atom-type predictions, signed feedback, and anchor-relative regularization. Reward-surrogate alignment is established under a fixed rollout–corruption law, posterior calibration, and trainable–anchor agreement.

2 RELATED WORK

Structure-based drug design. Early methods (Skalic et al., 2019) used sequence models to produce SMILES strings from protein context. Advances in equivariant networks and molecular diffusion (Satorras et al., 2021; Hoogeboom et al., 2022) support generation directly in 3D. Voxel-based models learn atomic density grids (Ragoza et al., 2022; Pinheiro et al., 2024), while autoregressive methods place atoms or fragments sequentially (Luo et al., 2021; Peng et al., 2022; Liu et al., 2022; Zhang et al., 2023; Zhang & Liu, 2023). Pocket-conditioned diffusion denoises atom types and coordinates (Guan et al., 2023a; Lin et al., 2025a; Schneuing et al., 2024; Huang et al., 2024; Guan et al., 2023b), or functional groups and linkers (Lin et al., 2023). MolCRAFT generates in continuous parameter space (Qu et al., 2024), whereas FlowSBDD uses rectified flow (Zhang et al., 2024). PocketXMol further unifies pocket-interaction tasks through atomic prompts (Peng et al., 2026).

Reinforcement learning for diffusion and flow models. Flow matching learns a time-dependent vector field that transports a simple prior toward the data distribution through simulation-free training (Lipman et al., 2023; Liu et al., 2023). Discrete flow models extend this framework to categorical state spaces through continuous-time Markov chains (CTMCs) (Campbell et al., 2024). Reward-based post-training spans language models (Ziegler et al., 2019; Ouyang et al., 2022) and diffusion models, including policy-gradient methods (Black et al., 2024; Fan et al., 2023) and control-based optimization via direct backpropagation (Uehara et al., 2024). FlowGRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025) use stochastic sampling for group-relative optimization, with updates tied to sampling transitions. DiffusionNFT (Zheng et al., 2026) instead trains implicit positive and negative prediction branches on re-corrupted generations, incorporating online rewards through forward-process reconstruction. Alternatively, DPO (Rafailov et al., 2023) and Diffusion-DPO (Wallace et al., 2024) learn directly from pairwise preferences. MOLFRA shares DiffusionNFT's forward-process negative-feedback principle. We focus on pocket-conditioned molecular generation, jointly updating coordinates and atom types with molecular rewards; our categorical analysis explains the resulting feedback and regularization.

Objective-aware molecule generation. A parallel line of work steers SBDD models toward biochemical objectives without full reinforcement learning: TAGMol (Dorna et al., 2024) guides diffusion sampling with target-aware property gradients, DecompOpt (Zhou et al., 2024a) optimizes over decomposed substructures, and MolJO (Qiu et al., 2025) uses gradient-guided Bayesian updates over both modalities. Offline preference optimization on pre-collected ranked pairs has also been applied here (Gu et al., 2024; Cheng et al., 2025; Huang & Zhang, 2025). Related energy-based preference optimization targets antibody sequence–structure co-design (Zhou et al., 2024b), while 3DMolFormer uses RL to fine-tune a dual-channel molecular transformer (Hu et al., 2025).

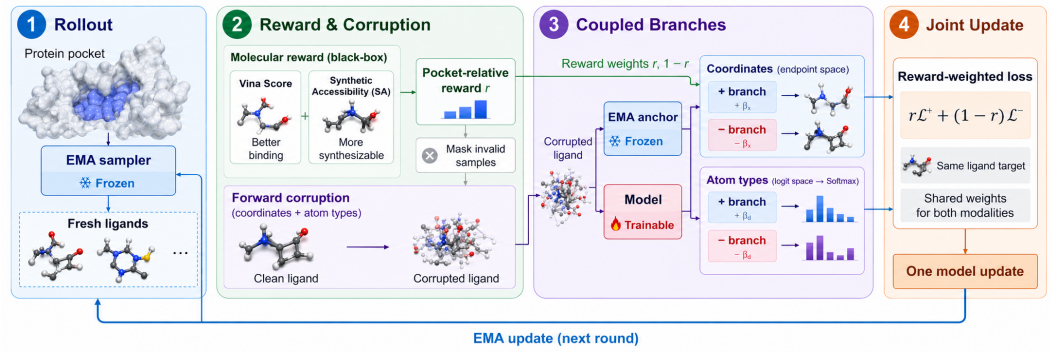

Figure 1: One round of MOLFRA. An exponential moving average (EMA) anchor generates fresh molecules and remains fixed during training. After reward evaluation and forward corruption, continuous reward coefficients r_i and $1 - r_i$ weight both modalities' branches at the same state; a separate mask w_i excludes invalid samples. One optimizer step is followed by an EMA update.

3 METHOD

MOLFRA adapts a pretrained pocket-conditioned generator to molecular design objectives through online reward feedback (Figure 1). Each round generates and evaluates candidate ligands, then uses their rewards to jointly update coordinate and atom-type generation. Below, we describe the base generator, online reward construction, and joint alignment objective.

3.1 PRETRAINED MULTIMODAL GENERATOR

We initialize the generator parameters θ from a pretrained MOLFORM-based checkpoint (Huang & Zhang, 2025). Given a protein pocket p and a specified ligand atom count N_M, it generates $M = (x_1, v_1)$ with coordinates $x_1 \in \mathbb{R}^{N_M \times 3}$ and atom types $v_1 \in \{1, \dots, C\}^{N_M}$ from C classes. We express protein and ligand coordinates relative to the centroid of the pocket atoms. We sample N_M from an empirical pocket-size-dependent prior estimated from the training set (Appendix B).

At a shared time $t \in [0, 1]$, we corrupt coordinates and independently draw each atom's type $v_t^a \sim \pi_t(\cdot \mid v_1^a)$, with $t = 0$ denoting noise and $t = 1$ data:

$$\begin{aligned} x_t &= (1-t)x_0 + t x_1, \qquad x_0 \sim \mathcal{N}(0, I), \\ \pi_t(k \mid v_1^a) &= \frac{1-t}{C} + t\,\mathbf{1}\{k = v_1^a\}, \qquad k \in \{1, \dots, C\}. \end{aligned} \tag{1}$$

Given the corrupted ligand state (x_t, v_t), the protein pocket p, and time t, the denoising network f_θ jointly predicts the clean coordinates and atom types using a shared backbone. Writing $s = (x_t, v_t, p, t)$, its outputs are a coordinate estimate $\hat{x}_{1,\theta}(s) \in \mathbb{R}^{N_M \times 3}$ of x_1 and atom-type logits $\ell_\theta(s) \in \mathbb{R}^{N_M \times C}$ for v_1. Pretraining uses coordinate MSE, atom-type cross-entropy, and Chamfer distance (Appendix A).

To generate a ligand, coordinates start from Gaussian noise and atom types from a uniform distribution at $t = 0$. Sampling couples a coordinate ODE and an atom-type CTMC (Campbell et al., 2024): between type jumps, coordinates evolve as $\dot{x}_t = \frac{\hat{x}_{1,\theta}(s) - x_t}{1-t}$ for $t < 1$. The atom-type logits determine the categorical rates. Euler steps and corrected categorical transitions numerically approximate these dynamics; Appendix B specifies the rates, discretization, and terminal prediction. MOLFRA retains this architecture, corruption process, and sampler.

3.2 ONLINE ROLLOUTS AND POCKET-RELATIVE REWARDS

Let $\pi_\theta(\cdot \mid p)$ denote the conditional ligand distribution induced by f_θ and its sampler. We maintain an exponential moving average (EMA) anchor θ_{old}, initialized from θ and updated after each optimizer step (Section 3.3). Each round holds the anchor fixed, samples training pockets, and generates

G candidates per pocket from $\pi^0(\cdot \mid p) = \pi_{\theta_{\text{old}}}(\cdot \mid p)$. Fresh rollouts make training online, although it is generally not on-policy with respect to π_θ.

For candidate M_i, let R_i denote its higher-is-better raw reward. The default reward favors lower Vina Score and higher normalized synthetic accessibility (SA), with the transformations in Eq. (36). Chemically invalid molecules and candidates without a usable reward are excluded from reward statistics and the training loss, with $w_i = 0$ for these candidates and $w_i = 1$ otherwise.

To remove pocket-specific reward offsets, we normalize within each pocket. Let $\bar{R}_p$ and σ_p be the mean and population standard deviation over its current candidates with $w_i = 1$. Let $\epsilon_r > 0$ be a standard-deviation floor and $\text{clip}(\cdot, -1, 1)$ denote clipping to $[-1, 1]$. The branch coefficient $r_i \in [0, 1]$ is

$$r_i = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\,\text{clip}\left(\dfrac{R_i - \bar{R}_p}{\max(\sigma_p, \epsilon_r)}, -1, 1\right), & w_i = 1, \\ \frac{1}{2}, & w_i = 0. \end{cases} \tag{2}$$

For i.i.d. unmasked groups, averaging these coefficients over the other group members defines the effective molecule weights used in our conditional analysis (Appendix D.1). The coefficients r_i and $1 - r_i$ weight the positive and negative losses for both coordinates and atom types (Section 3.3). If only one candidate is usable or all usable rewards are equal, $r_i = 1/2$ gives equal branch weights; these candidates remain in the loss. Groups with no usable candidates require no reward statistics and contribute no loss.

3.3 Coupled Endpoint–Logit Alignment

MolFRA jointly aligns clean-coordinate endpoints and atom-type logits under shared pocket-relative rewards. Paired positive and negative predictions convey feedback to both modalities. The continuous coordinate construction follows the negative-aware forward-process principle of Zheng et al. (2026). For atom types, negative affine combinations of probabilities may leave the simplex; we instead branch in logit space and apply softmax to obtain valid categorical predictions. These branches are used only to compute the training loss; sampling retains the original generator and hybrid sampler.

Forward-process reconstruction. For each MolFRA update, the ligands generated in Section 3.2 serve as fixed targets for reward-weighted reconstruction from forward-corrupted states. We independently sample one corruption time $t_i \sim p_{\text{train}}$ per molecule (Appendix F) and use it for all atoms and both modalities. Applying Eq. (1) yields the corrupted state s_i, at which we evaluate both the trainable model and the frozen EMA anchor. Here, forward corruption refers to corrupting a clean generated ligand, irrespective of the noise-to-data orientation of our time parameter.

Endpoint and logit branches. At s_i, let $(\hat{x}_i^0, \ell_i^0)$ and $(\hat{x}_{1,\theta,i}, \ell_{\theta,i})$ denote the frozen anchor and trainable predictions. Define the positive $(+)$ and negative $(-)$ branches as

$$\begin{aligned} \hat{x}_i^{\pm} &= (1 \mp \beta_{\text{x}})\hat{x}_i^0 \pm \beta_{\text{x}}\hat{x}_{1,\theta,i}, \\ \ell_i^{\pm} &= (1 \mp \beta_{\text{d}})\ell_i^0 \pm \beta_{\text{d}}\ell_{\theta,i}. \end{aligned} \tag{3}$$

The anchor outputs are stop-gradient quantities. The branch scales $\beta_{\text{x}}, \beta_{\text{d}} > 0$ are independently configurable; we adopt $\beta_{\text{x}} = 0.3$ and $\beta_{\text{d}} = 0.03$ as the default following the branch-scale study in Section 4.4. Only θ is optimized.

We branch in clean-endpoint space to match the pretrained generator's prediction interface; Appendix A relates this construction to velocity-space parameterizations.

For a fixed atom and context, let q^0 and q_θ denote the anchor and trainable clean-type distributions at unit temperature. Softmax of the branch logits gives normalized geometric combinations:

$$q_k^{\pm} = \frac{(q_k^0)^{1 \mp \beta_{\text{d}}}(q_{\theta,k})^{\pm\beta_{\text{d}}}}{\sum_{j=1}^{C}(q_j^0)^{1 \mp \beta_{\text{d}}}(q_{\theta,j})^{\pm\beta_{\text{d}}}}.$$

For finite logits, these distributions are symmetric exponential tilts of the anchor and remain in the interior of the categorical simplex. They provide valid predictions for both cross-entropy branches

without clipping branch probabilities, and generally differ from affine mixtures of denoising probabilities or CTMC rates. Appendix E.1 characterizes their log-odds geometry.

Both branches reconstruct the same generated ligand, weighted by r_i and $1 - r_i$. The negative branches multiply gradients passed to the trainable coordinate and logit predictions by $-\beta_{\mathrm{x}}$ and $-\beta_{\mathrm{d}}$, respectively, allowing nonnegative loss weights to convey negative feedback. When $w_i = 1$ and the trainable and anchor logits coincide, the logit gradient equals the usual cross-entropy gradient scaled by $\beta_{\mathrm{d}}(2r_i-1)$: below-average rewards ($r_i < 1/2$) reverse its direction. Away from this anchor point, the two branch gradients generally differ (Appendix E.3).

Molecule-level reconstruction objective. To control coordinate gradient scale, we adopt the self-normalized regression loss of Zheng et al. (2026). For predicted coordinates $\hat{x}$ and target coordinates x of a molecule with N_M atoms,

$$\mathcal{L}_{\mathrm{SN}}(\hat{x}, x) = \frac{N_M^{-1} \sum_{a=1}^{N_M} \|\hat{x}^a - x^a\|_2^2}{\max\Big(\epsilon_x, \mathrm{sg}\Big[(3N_M)^{-1} \sum_{a=1}^{N_M} \|\hat{x}^a - x^a\|_1\Big]\Big)}, \tag{4}$$

where $\epsilon_x > 0$ is a numerical floor and sg stops gradients through the mean absolute coordinate error. This denominator rescales the squared-error gradient using each branch's current error magnitude; it is computed separately for each branch at every forward pass. Appendix D.2 records the resulting coordinate semi-gradient, retaining the mask and both detached denominators.

Define the mean atom-type cross-entropy as

$$\mathcal{L}_{\mathrm{CE}}(\ell, v) = -\frac{1}{N_M} \sum_{a=1}^{N_M} \log[\mathrm{softmax}(\ell^a)]_{v^a}.$$

For sampled corruption times and states, the per-molecule branch losses and stochastic round-level objective are

$$\begin{aligned} \mathcal{L}_i^{\pm} &= \mathcal{L}_{\mathrm{SN}}(\hat{x}_i^{\pm}, x_{1,i}) + \lambda_v \mathcal{L}_{\mathrm{CE}}(\ell_i^{\pm}, v_{1,i}), \\ \mathcal{L}_{\mathrm{MOLFRA}}(\theta) &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \left[r_i \mathcal{L}_i^{+} + (1 - r_i) \mathcal{L}_i^{-} \right], \end{aligned} \tag{5}$$

where $\lambda_v \geq 0$ weights the atom-type loss. All losses reduce within each molecule before batch aggregation. The denominator $|\mathcal{B}|$ counts all actual collected candidates, including those masked by $w_i = 0$. The indicator w_i controls inclusion in the loss, while r_i and $1 - r_i$ control feedback. The objective averages over sampled corruption times and randomness; rollouts, rewards, inclusion indicators, and anchor predictions remain fixed during differentiation.

Update and interpretation. We accumulate gradients over $\mathcal{B}$, take one optimizer step, and update the anchor with EMA decay $\tau \in [0, 1)$:

$$\theta_{\mathrm{old}} \leftarrow \tau \theta_{\mathrm{old}} + (1 - \tau)\theta.$$

Algorithm 1 gives the full procedure. Training requires neither sampling-trajectory likelihood ratios nor backpropagation through the sampler.

Categorical feedback and regularization. For one atom with target y and reward coefficient $r \in [0, 1]$, let e_y be its one-hot vector, $\delta_\ell = \ell_\theta - \ell^0$, and $L_d = r\,\mathrm{CE}(\ell^+, y) + (1 - r)\,\mathrm{CE}(\ell^-, y)$. The branch loss satisfies

$$\begin{aligned} L_d = \mathrm{CE}(\ell^0, y) &+ \beta_{\mathrm{d}}(2r - 1)\langle q^0 - e_y, \delta_\ell \rangle \\ &+ r D_{\mathrm{KL}}(q^0 \| q^+) + (1 - r) D_{\mathrm{KL}}(q^0 \| q^-). \end{aligned} \tag{6}$$

The linear term supplies signed feedback, while both anchor-to-branch KL penalties are nonnegative. For finite logits, this identity requires neither posterior calibration nor trainable–anchor agreement (Appendix E.2). With the context, anchor, and r fixed, L_d is convex in output logits; at trainable–anchor agreement, its Hessian is independent of r. Neutral feedback ($r = 1/2$) retains both KL penalties but cancels the linear term; masking removes the entire sample contribution. These output-space properties imply neither network-parameter convexity nor molecular-diversity guarantees. Local and context-averaged reward-surrogate alignment requires a fixed rollout–corruption law, posterior calibration, and trainable–anchor agreement (Appendix E.3).

4 EXPERIMENTS

We compare MOLFRA with other SBDD methods in terms of docking scores, molecular properties, and conformation quality (Section 4.2). Reward ablations examine property trade-offs (Section 4.3). We include a published offline preference reference for alignment outcomes, while objective ablations examine the individual contributions of alignment components (Section 4.4).

4.1 EXPERIMENTAL SETUP

Dataset and splits. We conduct experiments on CrossDocked2020 (Francoeur et al., 2020) following the split of Luo et al. (2021). Our setup uses 99,900 training and 100 test complexes. Online rollouts are collected exclusively from the training pockets, strictly holding out the 100 test pockets for final evaluation.

Base generator and baselines. We select MOLFORM for its efficient 100-step log-time sampler that enables frequent online rollouts, and compare MOLFRA against its exact pretrained initialization under matched settings. External baselines comprise 11 CBGBench generators (Lin et al., 2025b) and four objective-aware methods: TAGMol (Dorna et al., 2024), DecompOpt (Zhou et al., 2024a), MolJO (Qiu et al., 2025), and AliDiff (Gu et al., 2024). We also include MOLFORM-DPO (Huang & Zhang, 2025) as a published same-backbone offline preference reference in Section 4.4.

Online post-training. The default reward combines rescaled Vina Score and SA terms (Appendix F). Main online comparisons and reward/objective ablations are trained for 1,300 rounds; each round samples 16 candidates for each of 12 training pockets from the EMA anchor and takes one Adam step (Kingma & Ba, 2015). Final evaluation uses the EMA model weights. The main run allocates 249,600 candidate generations and takes approximately 20 wall-clock hours ($\sim$55 s per round) on a single node with four GPUs (Appendix H).

Generation and evaluation. Our pretrained base and all online variants share the evaluation protocol: 100 generation attempts per test pocket with the 100-step log-time sampler, without replacement resampling for failed or invalid attempts. At test time, no reward- or property-based filtering is applied; reported metrics represent conditional summaries over the valid, successfully evaluated subsets, while diversity is computed within each pocket across its valid ligands. Specifically, docking affinity is evaluated via AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021) across generated (Score), locally minimized (Min), and redocked (Dock) poses. Molecular properties cover drug-likeness (QED), normalized synthetic accessibility (SA), and within-pocket diversity. Conformation quality is measured via PoseCheck (Harris et al., 2023), reporting median strain energy and mean clash counts on generated poses. Full evaluation protocols are detailed in Appendices B and F.

4.2 MAIN RESULTS

Addressing the optimization–diversity trade-off. MOLFRA improves docking scores and synthetic accessibility while retaining the reported diversity of its exact pretrained MOLFORM checkpoint (Table 1). Under identical evaluation protocols, post-training improves mean Vina Dock from -7.50 to -9.25 kcal/mol and mean normalized SA from 0.60 to 0.74, with diversity unchanged at 0.78. This is the highest reported diversity among the objective-aware methods in Table 1, exceeding MolJO (0.66) and TAGMol (0.63). The empirical diversity result complements the output-space regularization analysis in Section 3.3 and Appendix E.2, where symmetric categorical branching introduces nonnegative KL penalties relative to the EMA anchor.

Comprehensive docking evaluation across pose protocols. Scoring generated, locally minimized, and redocked poses probes distinct aspects of docking performance. Table 1 therefore reports all three AutoDock Vina protocols. MOLFRA achieves the lowest reported mean Vina Min (-8.36) and Vina Dock (-9.25), with the second-lowest Vina Score (-7.42) after MolJO (-7.52). Its Vina Min is close to MolJO (-8.33) and lower than AliDiff (-8.09) and TAGMol (-7.95). Beyond docking, mean SA (0.74) ranks second among the objective-aware methods, alongside mean QED 0.50 and 4.365 satisfied Lipinski criteria out of five. The reported mean ligand size is 23.0 heavy atoms, close to the reference mean of 22.8.

Table 1: **Binding affinity proxies and molecular properties on CrossDocked2020.** We benchmark MOLFRA against its exact pretrained base (MOLFORM), representative pocket-conditioned generators (top block), and objective-aware/guided methods (bottom block). (↑) and (↓) indicate higher and lower values are preferred, respectively. **Bold** denotes the best model result in each column; underline marks the second-best. † denotes larger-molecule settings (Qiu et al., 2025). MOLFRA achieves the lowest reported mean Vina Dock among the compared methods while matching the reported diversity of its pretrained base (0.78).

Model	Vina Score (↓)	Vina Min (↓)	Vina Dock (↓)	Diversity (↑)	QED (↑)		SA (↑)		Size	Lipinski
	Avg.	Avg.	Avg.	Avg.	Avg.	Med.	Avg.	Med.	Avg.	Avg.
LiGAN	-6.47	-7.14	-7.70	0.66	0.46	0.46	0.66	0.66	19.9	4.141
3DSBDD	-	-3.75	-6.45	0.70	0.48	0.48	0.63	0.63	17.7	4.751
GraphBP	-	-	-4.57	**0.79**	0.44	0.44	0.64	0.64	23.1	-
Pocket2Mol	-5.23	-6.03	-7.05	0.69	0.39	0.39	0.65	0.65	17.7	4.881
TargetDiff	-5.71	-6.43	-7.41	0.72	0.49	0.49	0.60	0.60	24.2	4.513
DiffSBDD	-	-2.15	-5.53	-	0.49	0.49	0.34	0.34	24.4	-
DiffBP	-	-	-7.34	-	0.47	0.47	0.59	0.59	-	-
FLAG	-	-	-3.65	-	0.41	0.41	0.58	0.58	21.5	-
D3FG	-	-2.59	-6.78	-	0.49	0.49	0.66	0.66	-	-
DecompDiff	-5.18	-6.04	-7.10	0.68	0.49	0.49	0.66	0.66	29.4†	4.312
MolCraft	-6.15	-6.99	-7.79	0.72	0.48	0.48	0.66	0.66	22.7	4.456
MOLFORM	-5.42	-6.42	-7.50	0.78	0.48	0.49	0.60	0.58	-	-
TAGMol	-7.02	-7.95	-8.59	0.63	0.55	0.56	0.56	0.55	24.7	4.716
DecompOpt	-5.75	-6.58	-7.63	0.67	0.48	0.45	0.65	0.65	32.9†	-
MolJO	**-7.52**	-8.33	-9.05	0.66	**0.56**	**0.57**	**0.78**	**0.77**	23.0	4.573
AliDiff	-7.07	-8.09	-8.90	0.73	0.50	0.50	0.57	0.56	24.4	-
MolFRA	-7.42	**-8.36**	**-9.25**	0.78	0.50	0.50	0.74	0.73	23.0	4.365
Reference	-6.36	-6.71	-7.45	-	0.48	0.47	0.73	0.74	22.8	4.270

Table 2: **3D conformation plausibility and structural integrity on CrossDocked2020.** PoseCheck is used to evaluate median strain energy (Energy) and average steric clashes (Clash) on generated poses. RMSD < 2Å measures pose consistency relative to redocked structures, and Connected denotes the proportion of valid single-component molecules. **Bold** marks the best model result (excluding experimental Reference). MOLFRA achieves the lowest steric clash count while maintaining high chemical validity.

Model	Energy Med. (↓)	Clash Avg. (↓)	RMSD < 2Å (↑)	Connected (↑)
Reference	114	5.46	34.0%	100%
AR	608	4.18	36.5%	93.5%
Pocket2Mol	186	6.22	31.3%	96.3%
FLAG	396	40.83	8.2%	97.1%
TargetDiff	1208	10.67	31.0%	90.4%
DecompDiff	983	14.23	25.1%	72.0%
IPDiff	5861	10.31	17.9%	90.1%
MolCraft	196	6.91	42.4%	96.7%
TAGMol	2058	7.41	37.2%	92.0%
MolJO	**163**	6.72	**43.5%**	**97.3%**
DecompOpt	861	16.6	24.3%	2.64%
MOLFRA	264	**2.28**	37.1%	94.8%

Conformation plausibility and steric clash reduction. Table 2 assesses geometric quality using PoseCheck and structural integrity metrics. MOLFRA achieves a mean protein–ligand clash count of 2.28, the lowest reported value among the compared models, including MolJO (6.72) and TargetDiff (10.67). The valid, connected molecule ratio is 94.8%, and 37.1% of generated poses have RMSD below 2Å relative to their Vina-redocked poses. Median strain energy (264) is lower than that of TAGMol (2058) and DecompOpt (861), although higher than MolJO (163) and the reference molecules (114). Figure 3 in Appendix G compares empirical distributions of all-atom and carbon–carbon bond distances with the reference (MOLFRA JSD: 0.045 and 0.196, respectively).

4.3 Ablation: Reward Design

Table 3: Reward-composition ablation on CrossDocked2020. Vina denotes Vina Score. The Vina + SA row reproduces the main configuration in Tables 1 and 2. Clash Avg. and Energy Med. denote the PoseCheck mean clash count and median strain energy.

Reward	Vina Score (↓)	Vina Min (↓)	Vina Dock (↓)	QED (↑)	SA (↑)	Clash Avg. (↓)	Energy Med. (↓)
Vina	-7.82	-8.62	-9.45	0.50	0.52	4.16	2139
Vina + SA	-7.42	-8.36	-9.25	0.50	0.74	2.28	264
Vina + QED	-6.92	-7.65	-8.48	0.62	0.57	2.78	1069
Vina + SA + QED	-5.93	-6.74	-8.23	0.62	0.68	2.93	506

We compare four reward compositions after 1,300 updates from the same pretrained checkpoint, using EMA weights and the evaluation protocol in Section 4.1 (Table 3). Reward definitions are provided in Appendix F.

SA as an empirical regularizer. Vina-only achieves the lowest docking scores but exhibits a median strain energy of 2,139. Adding SA increases normalized SA from 0.52 to 0.74, reduces median strain to 264 (about 88% lower), and decreases mean clashes from 4.16 to 2.28. These improvements accompany modest changes in Vina Score (-7.82 to -7.42) and Vina Dock (-9.45 to -9.25 kcal/mol). Neither strain nor clash count is explicitly included in the reward, supporting an empirical regularizing role for SA beyond improving its own score.

QED and remaining trade-offs. Both QED-containing rewards attain QED 0.62 but yield less favorable docking scores. Adding SA to Vina + QED also reduces median strain, from $1{,}069$ to 506, although mean clashes increase slightly (2.78 to 2.93). These results motivate Vina + SA as our default balance between docking performance, synthetic accessibility, and the measured conformation metrics. They support mitigation of the structural costs associated with Vina-only optimization, rather than establish the absence of reward hacking.

4.4 Ablation: Alignment Objective

We compare four objective variants trained for 1,300 rounds with the Vina + SA reward (Table 4).

Table 4: Alignment baselines and objective ablations on CrossDocked2020 (mean values). The pretrained MolFORM and full MolFRA rows repeat Table 1. All online variants use the Vina + SA reward and 1,300 rounds; published offline DPO results are included for context.

Method / variant	Vina Score (↓)	Vina Min (↓)	Vina Dock (↓)	QED (↑)	SA (↑)
MolFORM (pretrained)	-5.42	-6.42	-7.50	0.48	0.60
MolFORM-DPO (offline)	-6.16	-7.18	-8.13	0.50	0.65
RWR	-5.27	-6.14	-7.25	0.49	0.58
MolFRA (full)	-7.42	-8.36	-9.25	0.50	0.74
Positive-only	-0.53	-2.08	-4.77	0.22	0.44
Position-loss only	-5.91	-6.63	-7.56	0.52	0.61
Atom-type-loss only	-5.71	-6.48	-7.48	0.51	0.61

Negative-aware branching. *Positive-only* removes the negative loss from Eq. (5) while retaining the positive branch, degrading Vina Score from -7.42 to -0.53 kcal/mol and SA from 0.74 to 0.44. Reward-weighted regression (RWR) instead uses unbranched reconstruction with exponential reward weights (Appendix F) and performs worse than the pretrained base on docking and SA. These comparisons favor the full objective in this setting; RWR changes both branching and reward weighting, whereas Positive-only more directly tests the negative loss.

Joint coordinate–type alignment. *Position-loss only* removes atom-type cross-entropy, while *Atom-type-loss only* removes coordinate reconstruction. Because the backbone is shared, removing one loss does not freeze its corresponding output. Full MolFRA outperforms both variants on

all three docking metrics and SA, although their QED is slightly higher. These results support joint supervision for docking–SA optimization.

Branch-scale sensitivity. Among the five settings in Figure 2, $(\beta_{\mathrm{x}}, \beta_{\mathrm{d}}) = (0.3, 0.03)$ achieves the highest final 100-round mean training reward over the common first 1,111 rollout rounds and is therefore used as the default. The tested configuration with $\beta_{\mathrm{d}} = 3.0$ exhibits training divergence.

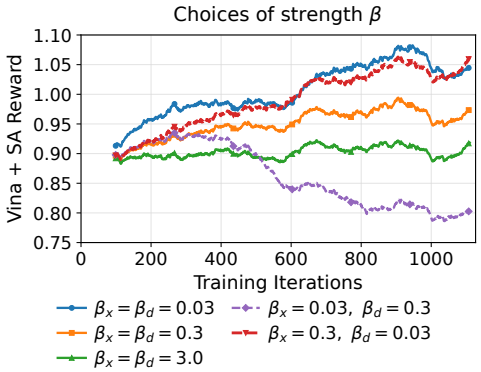

Figure 2: Choice of coordinate and atom-type branch scales. Mean raw Vina + SA training reward over the common first 1,111 rollout rounds, smoothed with a trailing 100-round moving average.

5 Conclusion

We presented MolFRA, an online post-training method for structure-based drug design. It couples coordinate-endpoint and atom-type-logit branches under shared pocket-relative rewards, without sampling-trajectory likelihood ratios or backpropagation through the sampler. An exact categorical cross-entropy decomposition provides an output-space regularization interpretation via reward-signed linear feedback and nonnegative anchor-relative KL penalties. On CrossDocked2020, post-training MolFORM with a joint Vina Score and SA reward improves mean Vina Dock from -7.50 to -9.25 kcal/mol and normalized SA from 0.60 to 0.74, while matching the pretrained base's reported diversity of 0.78. Component ablations further substantiate the role of negative-aware branching and joint geometric–chemical supervision. These findings highlight joint forward-process alignment as an effective approach to optimizing molecular property proxies while retaining generative diversity in the evaluated setting.

Limitations. Our experiments focus on a representative multimodal flow backbone and a standard SBDD benchmark, enabling controlled comparisons with the pretrained initialization and objective-aware baselines. Vina and SA, however, are computational proxies and do not provide definitive evidence of experimental binding affinity or synthetic feasibility. Reward ablations also reveal remaining property trade-offs, while median conformational strain remains above that of native complexes. Broader evaluation across model families, datasets, and target distributions, together with stronger physical constraints and prospective wet-lab validation, would further establish the generality and practical utility of MolFRA.

Reproducibility Statement

The generator interface and alignment objective are specified in Section 3, with base-model objectives in Appendix A, the sampler in Appendix B, and the full alignment procedure in Algorithm 1 (Appendix C). Reported hyperparameters — learning rates, branch-scaling coefficients, EMA decay, rollout sizes, and gradient clipping — are listed in Section 4.1 and Appendix F. Section 4.1 specifies the data split, baselines, result sources, and test generation budget; Appendix F details the evaluation protocol and metric aggregation. Source code will be released in the future.

AI Use Statement

Generative AI tools supported manuscript revision, literature verification, code review, and documentation. They were not used to originate the research ideas presented in this paper. The authors retain full responsibility for the final manuscript and all of its text, claims, code, and figures.

REFERENCES

Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/14f75513f0f1ca01de1e826b52e6b840-Paper-Conference.pdf.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

Xiwei Cheng, Xiangxin Zhou, Yu Bao, Yuwei Yang, and Quanquan Gu. Decomposed direct preference optimization for structure-based drug design. *Transactions on Machine Learning Research*, 2025.

Seungyeon Choi, Hwanhee Kim, Chihyun Park, Dahyeon Lee, Seungyong Lee, Yoonju Kim, Hyoungjoon Park, Sein Kwon, Youngwan Jo, and Sanghyun Park. Controllable 3D molecular generation for structure-based drug design through Bayesian flow networks and gradient integration. 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, Main Conference, pp. 125612–125647. Curran Associates, Inc., 2025. doi: 10.52202/085713-4187. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b6d0df730c5976ad918bbf4fb30afe7d-Paper-Conference.pdf.

Julian Cremer, Tuan Le, Frank Noé, Djork-Arné Clevert, and Kristof T. Schütt. PILOT: equivariant diffusion for pocket-conditioned de novo ligand generation with multi-objective guidance via importance sampling. *Chemical Science*, 15(36):14954–14967, 2024. ISSN 2041-6539. doi: 10.1039/d4sc03523b. URL http://dx.doi.org/10.1039/d4sc03523b.

Vineeth Dorna, D. Subhalingam, Keshav Kolluru, Shreshth Tuli, Mrityunjay Singh, Saurabh Singal, N. M. Anoop Krishnan, and Sayan Ranu. TAGMol: Target-aware gradient-guided molecule generation, 2024. URL https://arxiv.org/abs/2406.01650.

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, July 2021. ISSN 1549-960X. doi: 10.1021/acs.jcim.1c00203. URL http://dx.doi.org/10.1021/acs.jcim.1c00203.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023. doi: 10.52202/075280-3497. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fc65fab891d83433bd3c8d966edde311-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. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, August 2020. ISSN 1549-960X. doi: 10.1021/acs.jcim.0c00411. URL http://dx.doi.org/10.1021/acs.jcim.0c00411.

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*, 2023a. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11827–11846. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/guan23a.html.

Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models?, 2023. URL https://arxiv.org/abs/2308.07413.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. 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. 8867–8887. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Xiuyuan Hu, Guoqing Liu, Can Chen, Yang Zhao, Hao Zhang, and Xue Liu. 3DMolFormer: A dual-channel framework for structure-based drug discovery. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 27562–27583, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/449590dfd5789cc7043f85f8bb7afa47-Paper-Conference.pdf.

Jie Huang and Daiheng Zhang. MolFORM: Multi-modal flow matching for structure-based drug design, 2025. URL https://arxiv.org/abs/2507.05503.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin CUI, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3D molecule diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18453–18474, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50ca96a1a9ebe0b5e5688a504feb6107-Paper-Conference.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

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. doi: 10.52202/075280-1504. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6cdd4ce9330025967dd1ed0bed3010f5-Paper-Conference.pdf.

Haitao Lin, Yufei Huang, Odin Zhang, Siqi Ma, Meng Liu, Xuanjing Li, Lirong Wu, Jishui Wang, Tingjun Hou, and Stan Z. Li. DiffBP: generative diffusion of 3D molecules for target protein binding. *Chemical Science*, 16(3):1417–1431, 2025a. ISSN 2041-6539. doi: 10.1039/d4sc05894a. URL http://dx.doi.org/10.1039/d4sc05894a.

Haitao Lin, Guojiang Zhao, Odin Zhang, Yufei Huang, Lirong Wu, Cheng Tan, Zicheng Liu, Zhifeng Gao, and Stan Z Li. CBGBench: Fill in the blank of protein-molecule complex binding graph. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 19861–19887, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3275db433725892cc96ee2096e9c2d94-Paper-Conference.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di ZHANG, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. 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, Main Conference, pp. 40783–40818. Curran Associates, Inc., 2025. doi: 10.52202/085713-1362. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a10c46572628d58cb44fb705f25cbbf-Paper-Conference.pdf.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3D molecules for target protein binding. 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. 13912–13924. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22m.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3D generative model for structure-based drug design. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6229–6239. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

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, 17–23 Jul 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 PocketX-Mol. *Cell*, 189(7):1904–1922.e28, April 2026. ISSN 0092-8674. doi: 10.1016/j.cell.2026.01.003. URL http://dx.doi.org/10.1016/j.cell.2026.01.003.

Pedro O. Pinheiro, Arian Rokkum Jamasb, Omar Mahmood, Vishnu Sresht, and Saeed Saremi. Structure-based drug design by denoising voxel grids. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40795–40812. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pinheiro24a.html.

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 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. 50645–50671. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/qiu25h.html.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. MolCRAFT: Structure-based drug design in continuous parameter space. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41749–41768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/qu24a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chemical Science*, 13(9):2701–2713, 2022. ISSN 2041-6539. doi: 10.1039/d1sc05976a. URL http://dx.doi.org/10.1039/d1sc05976a.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.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, December 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00737-x. URL http://dx.doi.org/10.1038/s43588-024-00737-x.

Matthew D. Segall. Multi-parameter optimization: Identifying high quality compounds with a balance of properties. *Current Pharmaceutical Design*, 18(9):1292–1310, March 2012. ISSN 1381-6128. doi: 10.2174/138161212799436430. URL http://dx.doi.org/10.2174/138161212799436430.

Miha Skalic, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. From target to drug: Generative modeling for the multimodal structure-based ligand design. *Molecular Pharmaceutics*, 16(10):4282–4291, August 2019. ISSN 1543-8392. doi: 10.1021/acs.molpharmaceut.9b00634. URL http://dx.doi.org/10.1021/acs.molpharmaceut.9b00634.

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, January 2010. ISSN 1096-987X. doi: 10.1002/jcc.21334. URL http://dx.doi.org/10.1002/jcc.21334.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48892–48918. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/uehara24a.html.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, June 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

Daiheng Zhang, Chengyue Gong, and Qiang Liu. Rectified flow for structure based drug design, 2024. URL https://arxiv.org/abs/2412.01174.

Zaixi Zhang and Qi Liu. Learning subpocket prototypes for generalizable structure-based drug design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41382–41398. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhang23z.html.

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

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VJZ477R89F.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 5890–5912, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/180f6184a3458fa19c28c5483bc61877-Paper-Conference.pdf.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference 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. 120861–120891. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3840. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/daef77101ba5711084a57442c8cf2709-Paper-Conference.pdf.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2019. URL https://arxiv.org/abs/1909.08593.

A Base-Model Objectives and Parameterization

Prediction interface and pretraining objective. We reuse one pretrained MolFORM checkpoint (Huang & Zhang, 2025) for the base row and all online variants. Its nine-layer equivariant transformer (width 128, 16 heads, 32 neighbors) predicts clean coordinates $\hat{x}_{1,\theta}(s)$ and type logits $\ell_\theta(s)$. For minibatch $\mathcal{A}$ with $N_\mathcal{A} = \sum_{i\in\mathcal{A}} N_{M_i}$ ligand atoms, pretraining minimizes

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{pos}}^{\text{base}} + \lambda_v^{\text{base}} \mathcal{L}_{\text{type}}^{\text{base}} + \lambda_{\text{CD}} \frac{1}{|\mathcal{A}|} \sum_{i\in\mathcal{A}} \mathcal{L}_{\text{CD}}(\hat{x}_{1,i}, x_{1,i}), \tag{7}$$

with $\lambda_v^{\text{base}} = 1$ and $\lambda_{\text{CD}} = 0.5$.

Loss definitions and reductions. Coordinate MSE and type cross-entropy average over batch atoms:

$$\mathcal{L}_{\text{pos}}^{\text{base}} = \frac{\sum_{i,a} \|\hat{x}_{1,i}^a - x_{1,i}^a\|_2^2}{3N_\mathcal{A}}, \qquad \mathcal{L}_{\text{type}}^{\text{base}} = -\frac{\sum_{i,a} \log[\text{softmax}(\ell_i^a)]_{v_{1,i}^a}}{N_\mathcal{A}}. \tag{8}$$

The unsquared, two-way Chamfer term averages over molecules, with

$$\mathcal{L}_{\text{CD}}(\hat{x}, x) = \frac{1}{N}\sum_{a=1}^{N} \min_b \|\hat{x}^a - x^b\|_2 + \frac{1}{M}\sum_{b=1}^{M} \min_a \|x^b - \hat{x}^a\|_2, \tag{9}$$

where $N = |\hat{x}|$ and $M = |x|$. Post-training uses molecule-level reductions and a self-normalized coordinate loss; its λ_v is distinct from λ_v^{base} (Eqs. (4)–(5)).

Remark (Endpoint and velocity parameterization). At fixed $t < 1$, our coordinate construction is algebraically equivalent to the velocity-space branching of Zheng et al. (2026) under the endpoint–velocity reparameterization. For Eq. (1), set $u_t = x_1 - x_0$ and $\mathbf{v}_j = (\hat{x}_{1,j} - x_t)/(1-t)$, $j \in \{\theta, 0\}$. Affine branching gives, for $t < 1$,

$$\hat{x}_{1,\theta}^{\pm} = x_t + (1-t)\mathbf{v}_\theta^{\pm}, \qquad \mathbf{v}_\theta^{\pm} = (1 \mp \beta_{\text{x}})\mathbf{v}_0 \pm \beta_{\text{x}}\mathbf{v}_\theta, \tag{10}$$

$$\|\hat{x}_{1,\theta}^{\pm} - x_1\|_F^2 = (1-t)^2 \|\mathbf{v}_\theta^{\pm} - u_t\|_F^2. \tag{11}$$

Ordinary squared errors share a fixed-time unrestricted minimizer. The factor $(1-t)^2$ changes time weighting, so this does not establish equal time-averaged shared-network optima or equivalence to the self-normalized loss in Eq. (4).

B Sampler Details

Initialization and empirical prior. Protein and ligand coordinates are centered at the protein-atom centroid. The ligand atom count N_M is drawn from an empirical pocket-size-dependent prior estimated from the training set. Initial coordinates follow $x^{(0)} \sim \mathcal{N}(0, I)$, and initial atom types are drawn uniformly from $\{1, \ldots, C\}$.

Coupled discretization and evaluation convention. We use a log-time grid of $K = 100$ intervals with $t_k = 1 - 0.01^{k/K}$ for $k = 0, \ldots, K$ ($t_0 = 0$, $t_K = 0.99$) and step size $h_k = t_k - t_{k-1}$. At interval k, the network is evaluated once at the pre-update state using the right-grid time t_k:

$$s_k^- = (x^{(k-1)}, v^{(k-1)}, p, t_k), \qquad (\hat{x}_1^{[k]}, \ell^{[k]}) = f_\theta(s_k^-). \tag{12}$$

Coordinates evolve via the Euler ODE step

$$x^{(k)} = x^{(k-1)} + \frac{h_k}{1-t_k}\big(\hat{x}_1^{[k]} - x^{(k-1)}\big). \tag{13}$$

Both modalities share this forward pass: categorical transitions use $\ell^{[k]}$ without re-evaluating the network after the coordinate update.

Categorical transition correction. For atom a with current type $z = v^{(k-1),a}$, define the temperature-scaled clean-type probabilities $q^{[k]} = \mathrm{softmax}(\ell^{[k],a}/T)$. With noise parameter $\eta \geq 0$, the unclipped off-diagonal weights are h_k times the uniform-source CTMC rates:

$$\widetilde{P}_{zj} = h_k \left[\frac{1 + \eta + \eta(C-1)t_k}{1 - t_k} \, q_j^{[k]} + \eta q_z^{[k]} \right], \qquad j \neq z. \tag{14}$$

The diagonal rates make each row sum to zero. We obtain nonnegative weights W and normalized transition probabilities P by

$$\begin{aligned} W_{zj} &= \mathrm{clip}(\widetilde{P}_{zj}, 0, 1) \quad (j \neq z), \qquad W_{zz} = \max\Big(0, 1 - \sum_{j \neq z} W_{zj}\Big), \\ P_{zj} &= \frac{W_{zj}}{\sum_{c=1}^{C} W_{zc}}, \qquad j \in \{1, \ldots, C\}. \end{aligned} \tag{15}$$

The next types are drawn conditionally independently across atoms as $v^{(k),a} \sim \mathrm{Categorical}(P_{z,:})$. As noted in the main text, this gives a numerical CTMC approximation rather than an exact finite-time transition kernel.

Terminal prediction and parameters. Sampling uses $T = 0.01$ and $\eta = 1.0$; training cross-entropy retains unit temperature $T = 1$. After the 100th interval, one final forward pass at $t = 1$ gives

$$\big(\hat{x}_1^{\mathrm{out}}, \ell^{\mathrm{out}}\big) = f_\theta\big(x^{(K)}, v^{(K)}, p, 1\big), \qquad v_{\mathrm{out}}^a = \arg\max_c \ell_c^{\mathrm{out},a}. \tag{16}$$

Each sampling trajectory therefore uses $100+1 = 101$ network evaluations. The terminal prediction supplies clean endpoints without evaluating division by $1 - t$ at $t = 1$. Finally, the pocket centroid is restored to the coordinates for reconstruction, validity checks, and reward scoring.

C MOLFRA ALGORITHM

Algorithm 1 summarizes MOLFRA, with $N_{\mathrm{upd}} = 1{,}300$ updates in the main experiments. Training settings and the collection protocol are specified in Appendices F and H.

The fixed denominator $N_{\mathrm{round}} = |\mathcal{B}|$ counts all actual collected candidates, including masked candidates, but excludes artificial padding and uncollected groups. A local microbatch $\mathcal{M}$ contributes

$$\mathcal{L}_{\mathcal{M}} = \frac{1}{N_{\mathrm{round}}} \sum_{i \in \mathcal{M}} w_i \left[r_i \mathcal{L}_i^{+} + (1 - r_i) \mathcal{L}_i^{-} \right].$$

Each learner rank multiplies this contribution by the number of ranks to offset distributed gradient averaging, without additional microbatch averaging.

Algorithm 1 Joint endpoint–logit reward alignment of a pocket-conditioned multimodal flow

Require: pretrained parameters θ; training pockets $\mathcal{P}$; raw reward R_{raw}; B pockets and G candidates per pocket; minimum returned groups B_{min}; N_{upd} updates; corruption-time distribution p_{train}; $\beta_{\mathrm{x}}, \beta_{\mathrm{d}}$; λ_v; EMA decay τ

$\theta_{\mathrm{old}} \leftarrow \theta$
for $u = 1, \ldots, N_{\mathrm{upd}}$ **do**
  freeze θ_{old}; sample B pockets from $\mathcal{P}$; initialize $\mathcal{B} \leftarrow \varnothing$
  launch generation and scoring of G endpoints per pocket from $\pi_{\theta_{\mathrm{old}}}(\cdot \mid p)$
  wait until all B groups return scored payloads or are marked failed; fail the round on timeout or fewer than B_{min} returned payloads
  for each collected pocket group p **do**
    set $w_i = 0$ for chemical invalidity or unresolved infrastructure errors; $w_i = 1$ otherwise
    add all collected tuples (p, M_i, R_i, w_i), including masked candidates, to $\mathcal{B}$
    if any $w_i = 1$, compute the unmasked reward mean and population standard deviation, then obtain r_i by Eq. (2)
    set $r_i = 1/2$ where $w_i = 0$; fully masked groups need no statistics and contribute zero loss and gradient
  end for
  zero gradients; fix $N_{\mathrm{round}} = |\mathcal{B}|$ before padding; assign padding copies $w_i = 0$
  for each local microbatch $\mathcal{M}$ of the padded candidate list **do**
    center protein and ligand coordinates by the protein-atom centroid
    draw one $t_i \sim p_{\mathrm{train}}$ per molecule; corrupt coordinates and types jointly
    evaluate trainable and stop-gradient anchor predictions at the same corrupted states
    form endpoint and logit branches by Eq. (3)
    compute $\mathcal{L}_i^{\pm}$ by Eqs. (4)–(5), using a separate, detached coordinate-normalization denominator for each branch
    accumulate gradients of $\mathcal{L}_{\mathcal{M}}$ with fixed N_{round} and the distributed scaling above
  end for
  clip the accumulated gradient and take one Adam step on θ, also for an all-masked round
  $\theta_{\mathrm{old}} \leftarrow \tau \theta_{\mathrm{old}} + (1 - \tau)\theta$
end for

Return: θ_{old} for sampling and evaluation

D Reward Weights and Coordinate Update Details

This section records the effective reward weights used in the conditional analysis and the implemented coordinate update. The three categorical results and their proofs are collected in Appendix E. Throughout, $\pi^0(\cdot \mid p) = \pi_{\theta_{\text{old}}}(\cdot \mid p)$ denotes the anchor rollout distribution, frozen within an update.

D.1 Effective group-normalized weights

For the unmasked population analysis, fix a pocket p and draw $M_{1:G} \overset{\text{i.i.d.}}{\sim} \pi^0(\cdot \mid p)$ with finite raw scores $S_i = R_{\text{raw}}(M_i, p)$. Let $\bar{S}_G$ and σ_G be their sample mean and population-convention standard deviation. The normalization in Eq. (2) gives

$$W_i = \frac{1}{2} + \frac{1}{2}\,\text{clip}\left(\frac{S_i - \bar{S}_G}{\max(\sigma_G, \epsilon_r)}, -1, 1\right), \qquad \epsilon_r > 0. \tag{17}$$

Averaging over the other group members defines the effective molecule weight

$$r_G(M, p) = \mathbb{E}[W_i \mid M_i = M, p] \in [0, 1]. \tag{18}$$

Let ξ_i be identically sampled corruption randomness, independent of the other candidates conditional on M_i, p. Exchangeability and the tower property then give, for any integrable single-candidate function f,

$$\mathbb{E}\left[\frac{1}{G}\sum_i W_i f(M_i, \xi_i)\right] = \mathbb{E}_{M\sim\pi^0,\,\xi|M,p}[r_G(M,p) f(M,\xi)]. \tag{19}$$

The same identity holds for $1 - W_i$ and $1 - r_G$, even though each group statistic includes the focal candidate. The conditional surrogate in Appendix E.3 uses r_G as its reward weight, rather than the raw Vina/SA score. Clipping need not preserve a group-average weight of $1/2$. The masked extension retains the actual group-dependent weights directly.

D.2 Implemented self-normalized coordinate update

Fix a corrupted context s with ligand size $n = N_M$, flatten coordinates into $\mathbb{R}^{3n}$, and write $X = x_1$, $m^0 = \hat{x}^0$, and m for the trainable endpoint prediction. The rollout law, anchor, mask w, and group-normalized weight r are fixed during differentiation; expectations include the other candidates used to compute r and w. Define

$$\begin{aligned} B^{\pm}(m) &= m^0 \pm \beta_{\text{x}}(m - m^0), \\ D_{\pm}(m, X) &= \max\left(\epsilon_x, \frac{\|B^{\pm}(m) - X\|_1}{3n}\right), \\ L^{\text{SN}}_{x,s}(m) &= \mathbb{E}\left[\frac{w}{n}\left\{r\frac{\|B^{+}(m) - X\|_2^2}{\text{sg}[D_{+}(m,X)]} + (1-r)\frac{\|B^{-}(m) - X\|_2^2}{\text{sg}[D_{-}(m,X)]}\right\} \,\middle|\, s\right]. \end{aligned} \tag{20}$$

Assume the loss expectation is finite and $\epsilon_x > 0$. With detached denominators and branch Jacobians $\pm\beta_{\text{x}} I$, the implemented negative semi-gradient is

$$\begin{aligned} u_x(m) &:= -\widetilde{\nabla}_m L^{\text{SN}}_{x,s}(m) \\ &= \frac{2\beta_{\text{x}}}{n}\mathbb{E}\left[w\left\{r\frac{X - B^{+}(m)}{D_{+}(m,X)} - (1-r)\frac{X - B^{-}(m)}{D_{-}(m,X)}\right\} \,\middle|\, s\right]. \end{aligned} \tag{21}$$

At anchor agreement, the branches and denominators coincide:

$$u_x(m^0) = \frac{2\beta_{\text{x}}}{n}\mathbb{E}\left[w(2r-1)\frac{X - m^0}{D_0(X)} \,\middle|\, s\right], \quad D_0(X) = \max\left(\epsilon_x, \frac{\|X - m^0\|_1}{3n}\right). \tag{22}$$

These expressions also apply to finite empirical averages. The symbol $\widetilde{\nabla}$ denotes the gradient used with detached denominators, rather than the total derivative of the numerical loss after recomputing them. Stopping denominator gradients does not make the denominators constant across samples or predictions. In particular, ordinary mean calibration does not imply cancellation of the normalized residual:

$$\mathbb{E}[X - m^0 \mid s, w = 1] = 0 \quad \not\Rightarrow \quad \mathbb{E}\left[\frac{X - m^0}{D_0(X)} \,\middle|\, s, w = 1\right] = 0.$$

Thus an ordinary-MSE optimizer cannot replace the implemented update in Eq. (21).

E Categorical Branch Geometry, Regularization, and Alignment

The following three results describe the categorical extension: symmetric exponential-tilt geometry, exact anchor-relative regularization, and conditional reward-surrogate alignment. The first two hold for finite logits without calibration or trainable–anchor agreement. The third requires both conditions under a fixed rollout–corruption law and includes aggregation with the actual mask and molecule reduction.

Fix one atom and a corrupted context $s = (x_t, v_t, p, t)$, including the ligand size. Let $Y \in \{1, \dots, C\}$ be the clean atom type, $\ell^0, \ell \in \mathbb{R}^C$ the finite anchor and trainable logits, and $q^0 = \mathrm{softmax}(\ell^0)$, $q = \mathrm{softmax}(\ell)$. The anchor is fixed during differentiation. For $\beta_{\mathrm{d}} > 0$, the branches are

$$\ell^{\pm} = \ell^0 \pm \beta_{\mathrm{d}}(\ell - \ell^0), \qquad q^{\pm} = \mathrm{softmax}(\ell^{\pm}). \tag{23}$$

E.1 Exact categorical branch geometry

Proposition 1 (Validity and natural-parameter geometry). For any finite ℓ^0, ℓ and $\beta_{\mathrm{d}} > 0$, both branches lie in the interior of the categorical simplex. Writing $d = \ell - \ell^0$,

$$q_k^{\pm} = \frac{q_k^0 \exp(\pm\beta_{\mathrm{d}} d_k)}{\sum_j q_j^0 \exp(\pm\beta_{\mathrm{d}} d_j)}. \tag{24}$$

Thus the branches are symmetric exponential tilts of the anchor, or equivalently normalized geometric combinations $q_k^{\pm} \propto (q_k^0)^{1\mp\beta_{\mathrm{d}}} q_k^{\pm\beta_{\mathrm{d}}}$. Their pairwise log-odds satisfy

$$\log \frac{q_k^{\pm}}{q_j^{\pm}} = \log \frac{q_k^0}{q_j^0} \pm \beta_{\mathrm{d}} \left(\log \frac{q_k}{q_j} - \log \frac{q_k^0}{q_j^0} \right). \tag{25}$$

For $0 < \beta_{\mathrm{d}} \le 1$, the positive branch interpolates in log-odds coordinates, while the negative branch extrapolates in the opposite direction. Validity holds for every $\beta_{\mathrm{d}} > 0$.

Proof. Substituting the branch logits into softmax and cancelling the anchor's normalizing constant gives

$$q_k^{\pm} = \frac{\exp(\ell_k^0)\exp(\pm\beta_{\mathrm{d}} d_k)}{\sum_j \exp(\ell_j^0)\exp(\pm\beta_{\mathrm{d}} d_j)} = \frac{q_k^0 \exp(\pm\beta_{\mathrm{d}} d_k)}{\sum_j q_j^0 \exp(\pm\beta_{\mathrm{d}} d_j)}.$$

Finite logits give strict positivity and unit total mass. Writing $\log q_k = \ell_k - \log \sum_j \exp(\ell_j)$ yields the geometric form after cancelling class-independent factors. Taking the ratio of two classes gives Eq. (25); additive logit constants cancel as well. □

These identities describe denoising distributions at a fixed context; they do not identify logit branching with affine CTMC-rate or terminal-policy mixtures.

E.2 Exact cross-entropy decomposition and output-space regularization

Proposition 2 (Exact categorical cross-entropy decomposition). For one atom with target type y, fix $r \in [0, 1]$, let $\delta_\ell = \ell - \ell^0$, and write e_y for the one-hot target. Then

$$\begin{aligned} L_d &= r\,\mathrm{CE}(\ell^0 + \beta_{\mathrm{d}}\delta_\ell, y) + (1-r)\,\mathrm{CE}(\ell^0 - \beta_{\mathrm{d}}\delta_\ell, y) \\ &= \mathrm{CE}(\ell^0, y) + \beta_{\mathrm{d}}(2r-1)\langle q^0 - e_y, \delta_\ell\rangle \\ &\quad + r D_{\mathrm{KL}}(q^0 \| q^+) + (1-r) D_{\mathrm{KL}}(q^0 \| q^-). \end{aligned} \tag{26}$$

The KL directions are from the anchor to each branch, and both weighted KL terms are nonnegative. The identity requires neither posterior calibration nor trainable–anchor agreement.

Proof. Let $A(z) = \log \sum_k \exp(z_k)$ and $q_h = \mathrm{softmax}(\ell^0 + h)$. Expanding the KL divergence gives

$$D_{\mathrm{KL}}(q^0 \| q_h) = A(\ell^0 + h) - A(\ell^0) - \langle q^0, h\rangle. \tag{27}$$

Since $\mathrm{CE}(z, y) = A(z) - z_y$,

$$\mathrm{CE}(\ell^0 + h, y) - \mathrm{CE}(\ell^0, y) = \langle q^0 - e_y, h\rangle + D_{\mathrm{KL}}(q^0 \| q_h). \tag{28}$$

Substitute $h = \pm\beta_{\mathrm{d}}\delta_\ell$ and weight the two identities by r and $1 - r$. The linear coefficients combine into $\beta_{\mathrm{d}}(2r - 1)$, proving Eq. (26). KL nonnegativity and $r \in [0, 1]$ give the stated nonnegative penalties. □

Remark (Logit-space curvature). Fix the context, anchor, target, and $r \in [0, 1]$. With $F(q) = \mathrm{Diag}(q) - qq^\top$, the affine branch Jacobians give

$$\nabla^2_\ell L_d = \beta^2_{\mathrm{d}}\big[rF(q^+) + (1 - r)F(q^-)\big] \succeq 0.$$

Here $a^\top F(q)a = \mathrm{Var}_{K\sim q}(a_K) \geq 0$. At $\ell = \ell^0$, $\nabla_\ell L_d = \beta_{\mathrm{d}}(2r - 1)(q^0 - e_y)$, while the Hessian $\beta^2_{\mathrm{d}}F(q^0)$ is independent of r. These output-space properties imply neither network-parameter convexity nor optimizer stability.

Interpretation and relation to the implemented objective. Neutral feedback ($r = 1/2$) retains both KL penalties but cancels the signed linear term; $w = 0$ removes the entire sample contribution. Summing Eq. (26) over candidates i and atoms a with weights $\lambda_v w_i/(|\mathcal{B}|N_{M_i})$ recovers the categorical term in Eq. (5). These penalties use the fixed EMA anchor and provide neither terminal-policy KL bounds nor molecular-diversity guarantees.

E.3 Conditional local and average surrogate alignment

Proposition 3 (Conditional reward-surrogate alignment). Use the fixed, unmasked rollout–corruption law of Appendix D.1 and set $R = r_G(M, p) \in [0, 1]$. At context s, define $p^0_k = \Pr(Y = k \mid s)$ and $\rho_k = \mathbb{E}[R \mid Y = k, s]$. Assume posterior calibration, $q^0 = p^0$, and trainable–anchor agreement, $\ell = \ell^0$. Finite anchor logits imply $p^0_k > 0$, so every conditional reward is defined. Let

$$J_s(\ell) = \sum_k [\mathrm{softmax}(\ell)]_k \rho_k, \qquad \mu = \mathbb{E}[R \mid s] = \sum_k p^0_k \rho_k, \tag{29}$$

and

$$\mathcal{L}^{\mathrm{disc}}_s(\ell) = \mathbb{E}[R\,\mathrm{CE}(\ell^+, Y) + (1 - R)\,\mathrm{CE}(\ell^-, Y) \mid s]. \tag{30}$$

The rollout law and ρ remain fixed during differentiation. Then

$$-\nabla_\ell \mathcal{L}^{\mathrm{disc}}_s(\ell^0) = 2\beta_{\mathrm{d}}\nabla_\ell J_s(\ell^0) = 2\beta_{\mathrm{d}}p^0 \odot (\rho - \mu\mathbf{1}). \tag{31}$$

For $\ell_\eta = \ell^0 - \eta\nabla_\ell \mathcal{L}^{\mathrm{disc}}_s(\ell^0)$,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} J_s(\ell_\eta)\bigg|_{\eta=0} = 2\beta_{\mathrm{d}}\|\nabla_\ell J_s(\ell^0)\|^2_2 \geq 0. \tag{32}$$

If the conditional rewards are not all equal, the derivative is strictly positive and sufficiently small positive steps improve J_s.

Masked population extension. Let ν^0 be the fixed distribution of corrupted contexts from all collected candidates before masking, and set $a_s = \Pr(w = 1 \mid s)$. For $a_s > 0$ and atom a, define

$$p^{w,a}_k(s) = \Pr(Y^a = k \mid s, w = 1), \qquad \rho^{w,a}_k(s) = \mathbb{E}[r \mid Y^a = k, s, w = 1],$$

where r is the actual group-normalized weight and expectations include the other rollout candidates. Assume $p^{w,a}_k(s) > 0$ for each class. Write

$$J^w_{s,a}(\ell) = \sum_k [\mathrm{softmax}(\ell)]_k \rho^{w,a}_k(s),$$
$$L^w_{d,s,a}(\ell) = \mathbb{E}[w\{r\,\mathrm{CE}(\ell^+, Y^a) + (1 - r)\,\mathrm{CE}(\ell^-, Y^a)\} \mid s].$$

Set both quantities to zero when $a_s = 0$. Retaining the implemented molecule reduction, define

$$\begin{aligned} \overline{J}^w_d(\theta) &= \mathbb{E}_{s\sim\nu^0}\left[\frac{a_s}{N_M(s)}\sum_a J^w_{s,a}(\ell^a_\theta(s))\right], \\ L^w_d(\theta) &= \mathbb{E}_{s\sim\nu^0}\left[\frac{1}{N_M(s)}\sum_a L^w_{d,s,a}(\ell^a_\theta(s))\right]. \end{aligned} \tag{33}$$

Here each sum runs over the molecule's atoms. The loss L_d^w is the categorical component of Eq. (5), excluding λ_v; its masked expectation already includes a_s, without dividing by the retained fraction. The reference law, conditional rewards, inclusion rates, and anchor are fixed during differentiation.

If $\ell^a_{\theta_0}(s) = \ell^{0,a}(s)$ and $q^{0,a}(s) = p^{w,a}(s)$ for ν^0-almost every context with $a_s > 0$ and every atom, then, assuming a differentiable network and interchange of expectation and differentiation,

$$-\nabla_\theta L_d^w(\theta_0) = 2\beta_{\mathrm{d}} \nabla_\theta \overline{J}_d^w(\theta_0). \tag{34}$$

Calibration here is with respect to the retained-candidate posterior; calibration before filtering does not imply calibration after filtering.

Proof. For one atom, $\nabla_z \mathrm{CE}(z, Y) = \mathrm{softmax}(z) - e_Y$ and the branch Jacobians are $\pm\beta_{\mathrm{d}} I$, giving

$$\nabla_\ell \mathcal{L}_s^{\mathrm{disc}}(\ell) = \beta_{\mathrm{d}} \mathbb{E}[R(q^+ - e_Y) - (1-R)(q^- - e_Y) \mid s].$$

At anchor agreement, $q^+ = q^- = q^0$. Using $\mathbb{E}[Re_Y \mid s] = p^0 \odot \rho$ and calibration yields

$$\nabla_\ell \mathcal{L}_s^{\mathrm{disc}}(\ell^0) = \beta_{\mathrm{d}}[(2\mu - 1)q^0 - 2p^0 \odot \rho + p^0] = -2\beta_{\mathrm{d}} p^0 \odot (\rho - \mu \mathbf{1}).$$

The softmax derivative gives $\nabla_\ell J_s(\ell) = q \odot (\rho - J_s(\ell)\mathbf{1})$, proving Eq. (31). Taking the inner product with $\nabla_\ell J_s(\ell^0)$ proves Eq. (32). Since $p_k^0 > 0$, this gradient vanishes exactly when all conditional rewards are equal; otherwise differentiability gives improvement for sufficiently small steps.

For the masked population extension, apply the same calculation under the law conditional on $w = 1$, with reward weight r, and multiply by a_s. This gives $u_{s,a} := -\nabla_\ell L^w_{d,s,a} = 2\beta_{\mathrm{d}} a_s \nabla_\ell J^w_{s,a}$ at the calibrated anchor. With $G_{s,a} = \partial \ell^a_\theta(s)/\partial\theta|_{\theta_0}$, the chain rule and the fixed context law give

$$\begin{aligned} -\nabla_\theta L_d^w(\theta_0) &= \mathbb{E}_{\nu^0}\left[\frac{1}{N_M(s)} \sum_a G_{s,a}^\top u_{s,a}\right] \\ &= 2\beta_{\mathrm{d}} \mathbb{E}_{\nu^0}\left[\frac{a_s}{N_M(s)} \sum_a G_{s,a}^\top \nabla_\ell J^w_{s,a}\right] = 2\beta_{\mathrm{d}} \nabla_\theta \overline{J}_d^w(\theta_0). \end{aligned}$$

This proves Eq. (34). The average surrogate's directional derivative is $2\beta_{\mathrm{d}} \|\nabla_\theta \overline{J}_d^w(\theta_0)\|_2^2 \geq 0$, although individual contexts may worsen. □

Remark (Joint update and scope). For $g = \nabla_\theta \overline{J}_d^w(\theta_0)$ and $U_x = -\widetilde{\nabla}_\theta L_x^{\mathrm{SN}}(\theta_0)$, including the implemented masks and molecule reductions, the raw joint direction is $U = U_x + 2\lambda_v \beta_{\mathrm{d}} g$. Hence

$$D\overline{J}_d^w(\theta_0)[U] = 2\lambda_v \beta_{\mathrm{d}} \|g\|_2^2 + \langle g, U_x \rangle. \tag{35}$$

The coordinate cross term may reinforce or offset the categorical contribution. Neither EMA averaging nor sampling temperature ensures posterior calibration or trainable–anchor agreement. These results concern population directions under a fixed rollout–corruption law; they do not establish terminal-reward improvement for finite-sample, gradient-clipped Adam steps followed by EMA updates.

F Training and Evaluation Details

Reward formulation and scaling. For ligand M and pocket p, the reward components are

$$\begin{aligned} \bar{R}_{\mathrm{Vina}}(M,p) &= -\frac{\mathrm{Clip}(\text{Vina Score}(M,p), -16, -1) + 1}{15}, \\ \bar{R}_{\mathrm{SA}}(M) &= \frac{\mathrm{SA}(M) - 0.17}{0.83}, \qquad \bar{R}_{\mathrm{QED}}(M) = \frac{\mathrm{QED}(M) - 0.17}{0.83}. \end{aligned} \tag{36}$$

Here, $\mathrm{Clip}(x, a, b) = \max(a, \min(x, b))$, and $\mathrm{SA} = (10 - s)/9$ for raw SAScore $s \in [1, 10]$. The SA and QED reward components are unclipped and become negative when the corresponding property value is below 0.17. The four variants in Table 3 use unit coefficients:

$$R_{\mathrm{raw}}(M,p) = \begin{cases} \bar{R}_{\mathrm{Vina}}, & \text{Vina,} \\ \bar{R}_{\mathrm{Vina}} + \bar{R}_{\mathrm{SA}}, & \text{Vina + SA (default),} \\ \bar{R}_{\mathrm{Vina}} + \bar{R}_{\mathrm{QED}}, & \text{Vina + QED,} \\ \bar{R}_{\mathrm{Vina}} + \bar{R}_{\mathrm{SA}} + \bar{R}_{\mathrm{QED}}, & \text{Vina + SA + QED.} \end{cases}$$

Equation (2) then maps R_{raw} to pocket-relative training weights.

Forward corruption and loss coefficients. Training draws one time independently per molecule from

$$p_{\text{train}}(t) = 0.98\,\text{Beta}(1.9, 1) + 0.02\,\mathcal{U}[0, 1]. \tag{37}$$

The same time is shared by all atoms and by both modalities. Coordinates interpolate toward independent standard Gaussian noise; each atom type is replaced by a uniform class with probability $1 - t$. Before corruption, the same protein-atom centroid is subtracted from protein and clean ligand coordinates (`center_pos_mode=protein`). Clean-endpoint prediction is often called x_0-prediction in diffusion notation; our noise-to-data convention denotes the clean coordinate target by x_1. The MOLFRA objective uses $\lambda_v = 1$ and $\epsilon_x = 10^{-3}$ in Eqs. (4)–(5).

Reward statistics and failures. The reward statistics use only unmasked, non-padding candidates in the current pocket group. The standard deviation uses the population convention (division by the number of these candidates) with floor $\epsilon_r = 10^{-3}$. Identical unmasked rewards, including a single unmasked candidate, give $r_i = 1/2$. Clipping means the group average of r_i need not equal $1/2$. Chemical invalidity, including reconstruction and chemically invalid ligand-preparation failures, sets $w_i = 0$. Infrastructure scoring failures are retried and, if unresolved, also receive $w_i = 0$. These candidates are excluded from reward statistics and the loss numerator. Their value $r_i = 1/2$ is only for bookkeeping; equal branch weights do not mask a sample. A group with no unmasked candidates, including an all-chemical-failure group, requires no reward statistics and contributes no loss or gradient.

Batch aggregation and round closure. Equation (5) divides by the number of actual collected candidates, including masked candidates, rather than $\sum_i w_i$. Artificial padding copies have $w_i = 0$ and are excluded from reward statistics, the loss numerator, and the denominator. Microbatch losses use this fixed full-round denominator and compensate for distributed gradient averaging, yielding one global loss and one optimizer step. Collection waits until all 12 groups return scored payloads or are marked failed; at least 10 returned payloads are required. Insufficient returns or a collection timeout cause the round to fail. Only returned payloads enter $\mathcal{B}$. The anchor is updated after the optimizer step; collection details are in Appendix H.

Pretraining. The reused base checkpoint was pretrained with Adam (Kingma & Ba, 2015) at an initial learning rate of 5×10^{-5}, betas $(0.95, 0.999)$, batch size 16, and gradient-norm clipping at 8. Pretraining augmented the data with Gaussian noise (standard deviation 0.1) on protein atom coordinates and used exponential learning-rate decay with factor 0.6 and a floor of 1×10^{-8}. This pretraining ran on two NVIDIA GH200 GPUs and converged within roughly 12 hours and 100k steps.

Alignment stage. We use a learning rate two orders of magnitude below pretraining: 5×10^{-7} for the online RL comparisons, against 5×10^{-5} for the base model. The default coefficients are $\beta_{\text{x}} = 0.3$ and $\beta_{\text{d}} = 0.03$ for the final configuration; the branch-scale study is reported in Section 4.4. These coefficients scale the displacement of each implicit branch from its anchor prediction; they do not bound the displacement of the trained model from the anchor. We use an exponential moving average with decay 0.995 for the anchor network update and gradient clipping at maximum norm 8.0. The main results use the composite reward over **Vina Score and SA**; the four reward-ablation settings are described in Section 4.3.

Each iteration allocates 16 candidate generations for each of 12 training-set pockets and uses the collected batch for one parameter update. The main comparisons, including the reward and alignment-objective ablations, are trained for 1,300 steps. Each step is one optimizer update. Test-time sampling uses EMA weights (Zheng et al., 2026). The main run allocates $12 \times 16 \times 1{,}300 = 249{,}600$ candidates across rollout batches; this is not a count of successful reward evaluations or docking calls. Rollout and test generation both use the 100-step log-time sampler with discrete temperature 0.01 and noise parameter 1.0 described in Appendix B.

Reward-weighted regression baseline. Reward-weighted regression (RWR) removes the implicit branches and directly fits the trainable model to sampled molecules after forward corruption; the EMA anchor is retained for rollout sampling. It uses exponential weights $\alpha_i \propto \exp(z_i / T_{\text{RWR}})$ with $T_{\text{RWR}} = 1$, where z_i is the within-pocket standardized *raw* reward, with a standard-deviation floor

of 10^{-3}. Weights are renormalized to preserve each pocket's input sample-weight sum, without the clipped-affine mapping in Eq. (2).

Branch-scale sensitivity study. Figure 2 compares three shared-scale settings $\beta_{\mathrm{x}} = \beta_{\mathrm{d}} \in \{0.03, 0.3, 3.0\}$ and two asymmetric pairs $(\beta_{\mathrm{x}}, \beta_{\mathrm{d}}) = (0.03, 0.3)$ and $(0.3, 0.03)$. All runs use learning rate 10^{-5} and the same pretrained initialization. The curves use the common first $1{,}111$ rollout rounds (indices 0–1,110), with a trailing 100-round moving average; round r records the reward before optimizer update $r + 1$. The reward is measured before pocket-wise normalization.

Evaluation protocol. The pretrained base and MOLFRA use identical sampling settings, atom-count rules, bond reconstruction, validity checks, and evaluation procedures. Each test pocket receives 100 generation attempts. Means over valid, successfully scored molecules are conditional means; they differ from a per-attempt expected reward that assigns an explicit value to invalid or unscorable outcomes.

Evaluation metrics. AutoDock Vina (Eberhardt et al., 2021) supplies Vina Score for the generated pose, Vina Min after local energy minimization, and Vina Dock for the lowest-energy pose returned by redocking. These affinity proxies are measured in kcal/mol, with lower values preferred. QED measures drug-likeness, and normalized SA measures synthetic accessibility. Diversity is the mean pairwise Tanimoto distance between molecular fingerprints within each pocket, averaged over pockets; higher QED, SA, and diversity are preferred. Atom-level and molecule-level protein–ligand clash ratios ($\mathrm{Ratio}_{\mathrm{cca}}$ and $\mathrm{Ratio}_{\mathrm{cm}}$) use a van der Waals overlap threshold of 0.4Å (Lin et al., 2025b). Lower clash ratios are preferred. Both clash-ratio metrics are computed on the original generated conformations, without additional coordinate optimization or redocking.

Conformation metrics. Table 2 follows the conventions of Qiu et al. (2025): Energy measures molecular strain and Clash counts protein–ligand steric clashes, both evaluated with PoseCheck (Harris et al., 2023). The reported Clash mean is distinct from the atom-level and molecule-level clash ratios above. The RMSD column gives the percentage of generated poses within 2Å of their Vina-redocked poses, using the source's non-symmetry-corrected RMSD. Connected measures the yield of valid molecules with a single connected component; DecompOpt includes all optimization rounds in its denominator.

Aggregation. For our models, Vina, QED, and SA means and medians are computed by pooling per-molecule results across the 100 test pockets, whereas diversity uses equal-pocket averaging.

G Additional Structural Evaluation

We compare the empirical distributions of all-atom distances and carbon–carbon bond distances, together with ring-size proportions, against the reference set, and present qualitative examples. These analyses characterize selected structural statistics rather than establish complete geometric or chemical validity.

Intramolecular distance distributions. Figure 3 shows the empirical distributions of all-atom distances (top) and carbon–carbon bond distances (bottom) for reference molecules, MOLFRA, and baseline methods. Distances are pooled across successfully evaluated molecules from all test pockets. For normalized histograms P and Q, we report the Jensen–Shannon distance

$$\mathrm{JSD}(P, Q) = \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P \| H) + \tfrac{1}{2} D_{\mathrm{KL}}(Q \| H)}, \qquad H = \tfrac{1}{2}(P + Q),$$

using natural logarithms. The MOLFRA values are 0.045 and 0.196, respectively; lower values indicate closer agreement with the reference.

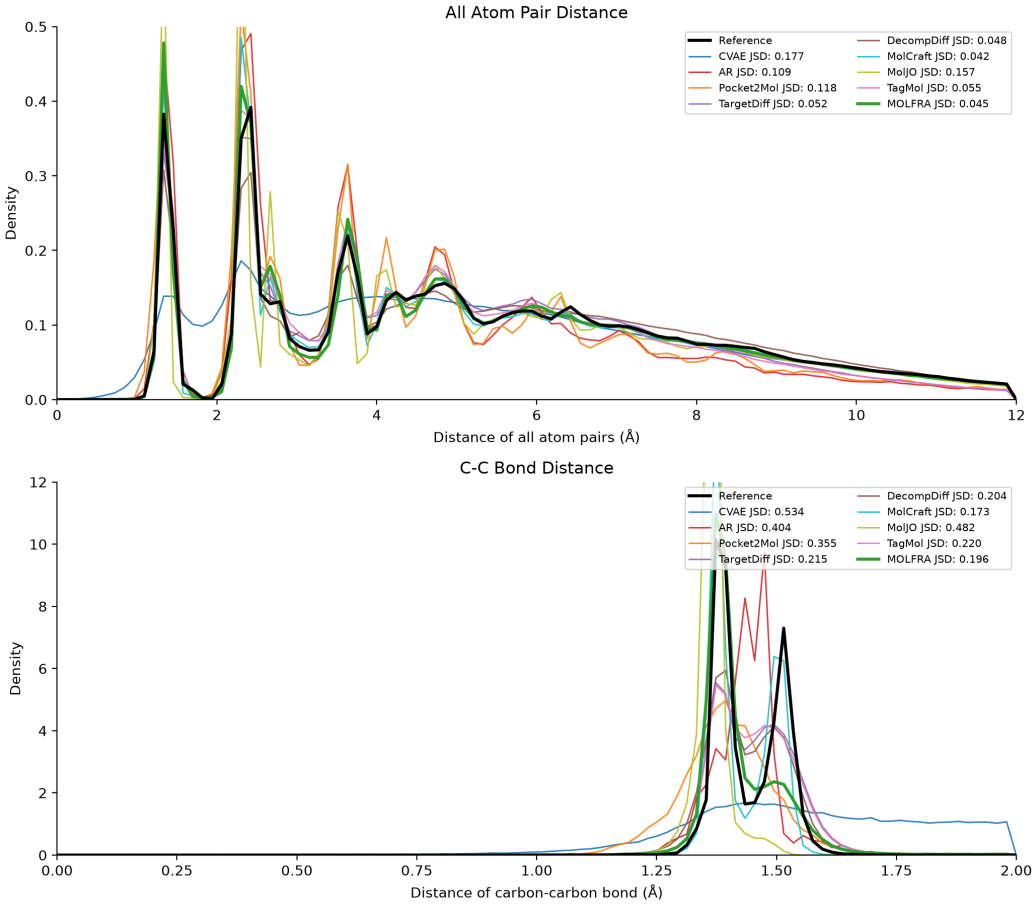

Figure 3: Empirical molecular distance distributions. All-atom distances (top) and carbon–carbon bond distances (bottom) for reference molecules (black), MOLFRA (green), and baseline methods. Legends report the Jensen–Shannon distance (JSD) relative to the reference distribution.

Qualitative examples. Figure 4 compares reference and generated ligands for three pockets.

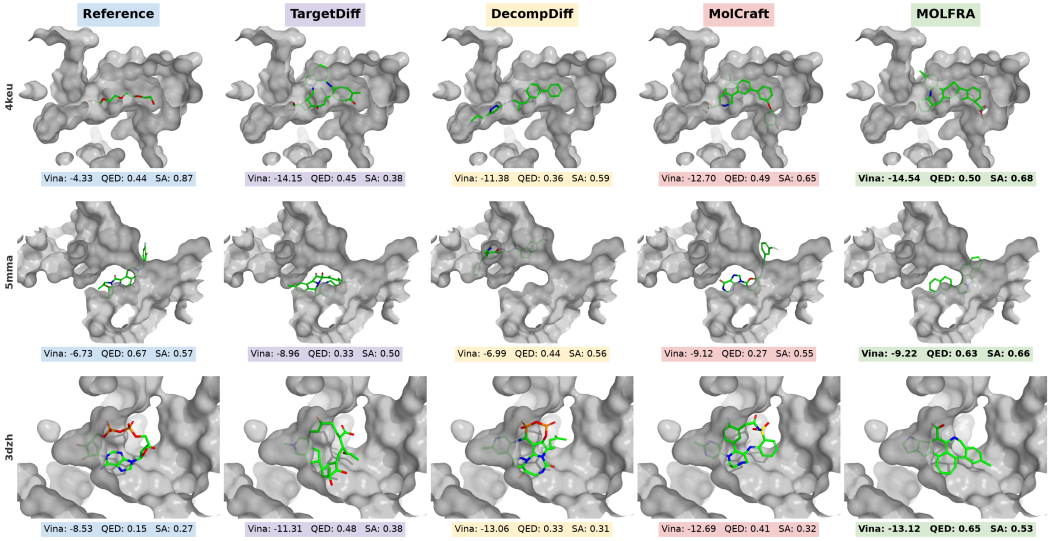

Figure 4: Reference and generated ligands for protein pockets 4keu (top), 5mma (middle), and 3dzh (bottom). Columns show reference ligands and samples from TargetDiff, DecompDiff, MolCraft, and MOLFRA. Vina denotes generated-pose Vina Score (kcal/mol); QED and normalized SA are also shown. These selected examples illustrate binding poses across different pocket shapes; Table 1 reports aggregate test-set results.

Ring sizes. Table 5 compares the proportions of rings with 3–8 atoms among all rings identified with RDKit in reference and generated molecules. MOLFRA has a similar six-membered ring proportion to the reference (68.5% vs. 66.0%), but fewer five-membered rings (18.5% vs. 29.6%) and more seven-membered rings (9.2% vs. 0.8%), indicating remaining differences in ring-size composition.

Table 5: Ring-size proportions (%) among all rings in reference and generated molecules, where 3-Ring denotes three-membered rings and so on. Each percentage is normalized by the total number of rings for that method, including sizes not shown.

Method	3-Ring	4-Ring	5-Ring	6-Ring	7-Ring	8-Ring
Reference	1.6	0.0	29.6	66.0	0.8	0.0
CVAE	28.1	15.6	29.7	22.6	2.6	0.8
AR	30.9	0.3	15.6	49.3	1.9	0.8
Pocket2Mol	0.1	0.0	16.3	79.8	2.6	0.3
TargetDiff	0.0	2.7	29.7	49.0	11.7	2.6
DecompDiff	2.9	3.7	31.0	46.0	11.9	2.3
MolCraft	0.0	0.2	23.1	69.9	5.4	0.6
MolJO	0.0	0.0	11.6	83.1	4.7	0.3
TAGMol	0.0	2.4	26.9	48.6	15.0	3.7
MOLFRA	0.0	0.6	18.5	68.5	9.2	1.9

H Co-located Execution and Computational Cost

Hardware and end-to-end cost. Online post-training co-locates rollout generation, molecular preparation, reward scoring, and learner updates on a single compute node. The node has four NVIDIA GH200 Grace Hopper Superchips, each with 96 GB HBM3 GPU memory, 72 Grace CPU cores, and 480 GB LPDDR5X host memory. The 288 CPU cores form four NUMA domains. The main run performs 1,300 optimizer updates with 12 pockets and 16 candidates per pocket per round, a nominal generation budget of $249{,}600$. It takes approximately 20 wall-clock hours, averaging about 55 s per rollout–update cycle. This timing includes generation, molecular preparation, AutoDock Vina reward scoring, and learner updates. Online rewards use Vina Score; the candidate budget is not a count of valid molecules, successful reward evaluations, or Vina calls.

Execution schedule and within-round overlap. Rollout actors and the PyTorch distributed-data-parallel learner share the same four GPUs and execute in alternating phases. Each round uses a fixed EMA anchor for generation and subsequent loss computation. Actors emit completed pocket groups of 16 candidates, allowing CPU preparation and scoring to begin while the GPUs generate the remaining groups. Generation–scoring overlap occurs within a round; GPU rollout and learner optimization do not overlap. Sampling uses 100 log-time intervals followed by one terminal prediction, totaling 101 network evaluations per trajectory (Appendix B). After collection, the learner accumulates gradients over the candidates, takes one optimizer step, and updates the EMA anchor.

CPU scoring and task recovery. Four supervisor daemons, bound to separate NUMA domains using `numactl`, manage a configured pool of 384 scoring worker threads over the 288 CPU cores. Workers dynamically claim scoring shards through POSIX atomic file renames in a shared filesystem queue, without a fixed worker-to-pocket assignment. Supervisors requeue retryable failed or interrupted tasks; unresolved candidate-level scoring failures follow the masking rules in Appendix F. The recorded Vina timing profile has a median of 2.17 s, a 99th percentile of 23.96 s, and a maximum above 45 s. These individual Vina timings are distinct from end-to-end round time.

Round closure and sample accounting. Collection waits until all 12 pocket groups have either returned scored payloads or been marked failed. A round proceeds only if at least 10 payloads have returned; reaching this threshold alone does not terminate collection. Insufficient returns or a collection timeout cause the round to fail. All actual candidates in returned payloads form $\mathcal{B}$ in Eq. (5), including candidates masked for chemical invalidity or unresolved scoring errors. Masked candidates contribute zero to the loss numerator but remain in its actual-candidate denominator. Failed or missing whole-pocket payloads and artificial padding are excluded from that denominator. Appendix C specifies distributed loss aggregation, and Appendix F details the masking rules.