SCORED: SELECTIVE OSDA GENERATION WITH CONTRASTIVE REWARDS FROM DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

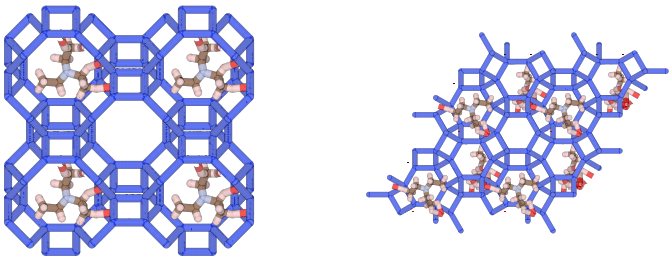
Scientific design often requires candidates that favor a target condition over plausible alternatives (competitors). Conditional generative models do not directly optimize this preference, while evaluating every candidate across multiple conditions is often too costly for online optimization. We introduce SCORED, a model-intrinsic selectivity reward that uses shared noise to contrast a conditional diffusion generator's denoising errors for the same candidate under target and competitor conditions. Scoring needs only forward passes: a six-condition panel takes 7.86 ms, against 8.3 s for a learned surrogate. We apply SCORED to generate organic structure-directing agents (OSDAs), molecules that guide zeolite framework formation. Mean-centered reward-weighted regression incorporates offline binding information, after which multi-objective reinforcement learning (RL) combines selectivity and molecular-quality rewards without physical evaluation in the online optimization loop. Across three runs, SCORED reaches a mean competition score of 5.03 ± 0.33 kJ/mol Si over six held-out frameworks under frozen-pose force-field evaluation, where a strong LLM-based generator scores 9.72 (lower is better). Matched ablations that drop the competitor contrast (6.34) or the denoising reward entirely (6.50) are both worse, isolating the contribution of the target–competitor comparison. Extensions to structure-based drug design and metal–organic framework generation give uneven gains across models and domains, which we examine through pre-RL denoising loss–property alignment. These results support conditional denoising comparisons as inexpensive online feedback for selectivity optimization.

1 INTRODUCTION

Generative models support scientific design in applications such as organic structure-directing agent discovery for zeolites (Jensen et al., 2021; Hu et al., 2025), structure-based drug design (Guan et al., 2023; Gu et al., 2024), and metal–organic framework generation (Fu et al., 2024). Many such tasks require candidates that favor a requested condition over plausible alternatives, yet optimizing target performance alone does not ensure such selectivity. Molecular optimization can exploit weaknesses in scoring functions (reward hacking) (Ghugare et al., 2024) and yield superficially "universal" candidates with favorable docking scores across many protein targets (Gao et al., 2024). Evaluating candidates across competing conditions can provide the missing feedback, but simulations or experiments are often too costly for online optimization even for a single condition, let alone multiple alternatives.

Zeolites make this problem concrete. These crystalline microporous materials support catalysis and separations (Ozin et al., 1989; Li et al., 2025; Wen et al., 2025; Zeng et al., 2025; Wu et al., 2025). OSDAs occupy developing pores to guide framework formation, yet the same OSDA can fit several topologies and show high affinity for multiple frameworks (Figure 1). Among 1,743 OSDAs reported in the literature, 844 direct multiple zeolite frameworks (Schwalbe-Koda et al., 2021), and $1{,}094$ of $3{,}553$ candidates from a prior generator bind multiple frameworks with binding energy below -10 kJ/mol Si (Jensen et al., 2021). An effective design pipeline must therefore favor OSDAs that are selective for the target *over competing frameworks*. This selectivity can be evaluated by the *competition score*, computed from binding energies across frameworks (Schwalbe-Koda et al., 2021; Schwalbe-Koda & Gómez-Bombarelli, 2021a). More negative binding energy means stronger affinity; a lower competition score means greater target selectivity.

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.



AEI: $E_{\text{bind}} = -10.76$ kJ/mol Si CHA: $E_{\text{bind}} = -11.10$ kJ/mol Si

Figure 1: Strong binding need not imply selectivity: optimizing affinity alone is not enough. The same OSDA, `CCN(CC)CC`, found by a previous generator that optimizes for affinity and other molecular metrics (Hu et al., 2025), fits both AEI and CHA (two distinct zeolite frameworks) with comparable binding energies, which may complicate phase-pure crystallization of either framework.

Beyond affinity and selectivity, zeolite-OSDA discovery objectives also include chemical validity, low synthetic complexity, and novelty. These objectives involve discrete checks and scoring functions that need not be differentiable, so we optimize them jointly with RL through reward feedback. The bottleneck here is obtaining a selectivity reward for every newly generated molecule. We therefore ask how to supply an inexpensive selectivity signal for RL.

Prior work used diffusion-model denoising error as a confidence signal (Kim & Cho, 2026), a post-hoc screening criterion (Yu & Gao, 2025) and the basis of a zero-shot classifier (Li et al., 2023). Inspired by this, we introduce SCORED (**S**elective OSDA Generation with **CO**ntrastive **RE**wards from **D**enoising), which contrasts conditional denoising errors to obtain a model-intrinsic signal for target-relative preference and uses this signal as a selectivity reward for RL fine-tuning. In our implementation, we use shared noise to calculate such denoising errors for stability, and use a SoftMin aggregate to emphasize plausible competitors. We initialize the generator with reward-weighted regression (RWR) on offline binding-energy labels after OSDA-wise mean-centering, then optimize the selectivity reward together with molecular-quality rewards. The same target–competitor scoring principle applies to pretrained diffusion generators in other domains.

We make the following contributions:

- **A model-intrinsic selectivity reward.** SCORED turns a target–competitor contrast of conditional denoising error into a selectivity reward. It requires only forward passes: scoring a six-condition panel takes 7.86 ms against 8.3 s for a docking-based evaluator, a 1056× reduction on the reward path. Our shared-noise implementation further lowers the estimator standard deviation by a geometric-mean factor of 56.3×, making the reward practical for online optimization.
- **A binding-informed offline-to-online pipeline.** Offline OSDA-wise mean-centered RWR biases model updates toward molecule–framework pairs with stronger relative binding, after which SCORED reuses the learned conditional model as the source of inexpensive online feedback. Removing either the offline initialization or the online updates degrades selectivity (Table 2).
- **Frozen-pose evaluation and generalization analysis.** Across three runs on six held-out frameworks, SCORED attains a mean competition score of 5.03 ± 0.33 kJ/mol Si under a frozen-pose force-field protocol, against 9.72 for OSDA Agent, a recent LLM-based OSDA generator. Matched ablations that drop the competitor contrast (6.34) or the denoising reward entirely (6.50) are both worse, isolating SCORED's contribution. Drug-design and MOF experiments give uneven gains (Section 5.4), which we examine through pre-RL denoising loss–property alignment.

2 Preliminaries

Denoising Diffusion Probabilistic Models (Ho et al., 2020) learn a reverse process $p_\theta(z_{t-1}|z_t)$ that inverts a T-step Gaussian forward process $q(z_t|z_{t-1})$ or $q(z_t|z_{t-1}, c)$ if conditioned on c. We parametrize the reverse process using a conditional noise predictor $\hat{\epsilon}_\theta(z_t, t, c)$. Following Ho et al. (2020), the diffusion model's training objective can be simplified as minimizing the denoising loss through supervised fine-tuning (SFT):

$$\mathcal{L}_{\text{SFT}}(z_0, c) = \mathbb{E}_{t,\epsilon}[||\epsilon - \hat{\epsilon}_\theta(z_t, t, c)||^2] \text{ where } z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I). \quad (1)$$

Here $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative signal retention ratio controlled by a scheduler. We adopt this simplified objective without the timestep-dependent weighting. Following the manifold hypothesis (Cayton, 2005), latent diffusion models (LDMs) (Rombach et al., 2022) perform denoising in a lower-dimensional latent space, which is typically learned via a pretrained autoencoder. In molecular generation, a molecular encoder h maps a molecule x to its clean latent $z_0 = h(x)$, which is corrupted to z_t at diffusion step t. A target framework y is encoded into its condition $c = e(y)$ with condition encoder e. The conditional denoiser $\hat{\epsilon}_\theta(z_t, t, c)$ is trained on a dataset containing $\{(x, y)\}$ tuples using Equation (1). At generation, denoising starts from $z_T \sim \mathcal{N}(0, I)$ and produces a clean latent z_0 conditioned on c; the corresponding molecular decoder g then maps z_0 back to a molecule x.

3 Related Work

OSDA discovery and target-relative design. OSDA discovery has progressed from physically scored heuristic and evolutionary search (Lewis et al., 1996; Daeyaert & Deem, 2019; Muraoka et al., 2020) to generators conditioned on zeolite topology or gel chemistry (Jensen et al., 2021; Xu et al., 2023; Ito et al., 2024; Hu et al., 2025). These approaches propose candidates for a requested framework, but only a few account for competition across alternative frameworks using physics-based evaluators (Daeyaert & Deem, 2019; Schwalbe-Koda et al., 2021). Such evaluators require repeating pore-based docking and atomistic simulations at every online update, which is computationally expensive (Schwalbe-Koda & Gómez-Bombarelli, 2021a). Learned binding-energy models accelerate screening (Daeyaert et al., 2019; Hu et al., 2025; Xie et al., 2025), but remain external, docking-dependent evaluators rather than model-intrinsic selectivity signals.

Molecular diffusion post-training. Pocket-conditioned diffusion models generate ligands for a specified protein (Schneuing et al., 2024; Guan et al., 2023; Huang et al., 2024; Gu et al., 2024); selectivity-oriented variants explicitly optimize target–off-target affinity gaps (Gao et al., 2024; Park et al., 2026). These methods estimate affinity gaps by docking generated ligands to target and off-target pockets or by using separately trained affinity predictors. RL, reward backpropagation, and preference-based post-training then update diffusion generators using external evaluations or labeled pairs (Black et al., 2024; Fan et al., 2023; Prabhudesai et al., 2023; Wallace et al., 2024; Cheng et al., 2025; Sun et al., 2026). SCORED instead obtains its selectivity signal from the generator's conditional denoising errors, retaining a standard group-relative RL optimizer (Ding & Ye, 2025; Liu et al., 2025; Xue et al., 2025; Liu et al., 2026).

4 Methodology

Diffusion-based classification uses denoising errors to identify which condition best matches a given sample (Li et al., 2023; Yu & Gao, 2025; Kim & Cho, 2026; Clark & Jaini, 2023). Selectivity asks something stronger: not which framework wins, but whether the target is preferred over plausible competitors by a meaningful margin. We therefore turn the comparison of denoising errors into a scalar margin and optimize it directly. In this section, we first construct such a denoising comparison, then describe how OSDA-wise mean-centered RWR incorporates offline binding information into the denoiser, and finally explain how the comparison is used to refine generation.

4.1 Denoising-Based Selectivity Reward

Conditional compatibility. For molecule x encoded as $z_0 = h(x)$ and target framework y represented by $c = e(y)$, use Equation (1) with $t \sim \mathrm{U}\{1, \ldots, T\}$ and $\epsilon \sim \mathcal{N}(0, I)$ to define

$$d_\theta(z_0, c; t, \epsilon) := \|\epsilon - \hat{\epsilon}_\theta(z_t, t, c)\|_2^2, \tag{2}$$

$$D_\theta(z_0, c) := \mathbb{E}_{t,\epsilon}[d_\theta(z_0, c; t, \epsilon)] = \mathcal{L}_{\mathrm{SFT}}(z_0, c). \tag{3}$$

For an encoded molecule, we write $d_\theta(x, c; t, \epsilon) := d_\theta(h(x), c; t, \epsilon)$ and $D_\theta(x, c) := D_\theta(h(x), c)$. During online rollouts, d_θ is evaluated directly on the generated latent z_0, without decoding and re-encoding. Lower denoising error serves as a model-based indication of conditional compatibility.

From target confidence to selectivity. Physical selectivity is inherently relative: whether a molecule favors a target framework depends on its preference over plausible competitors, rather than

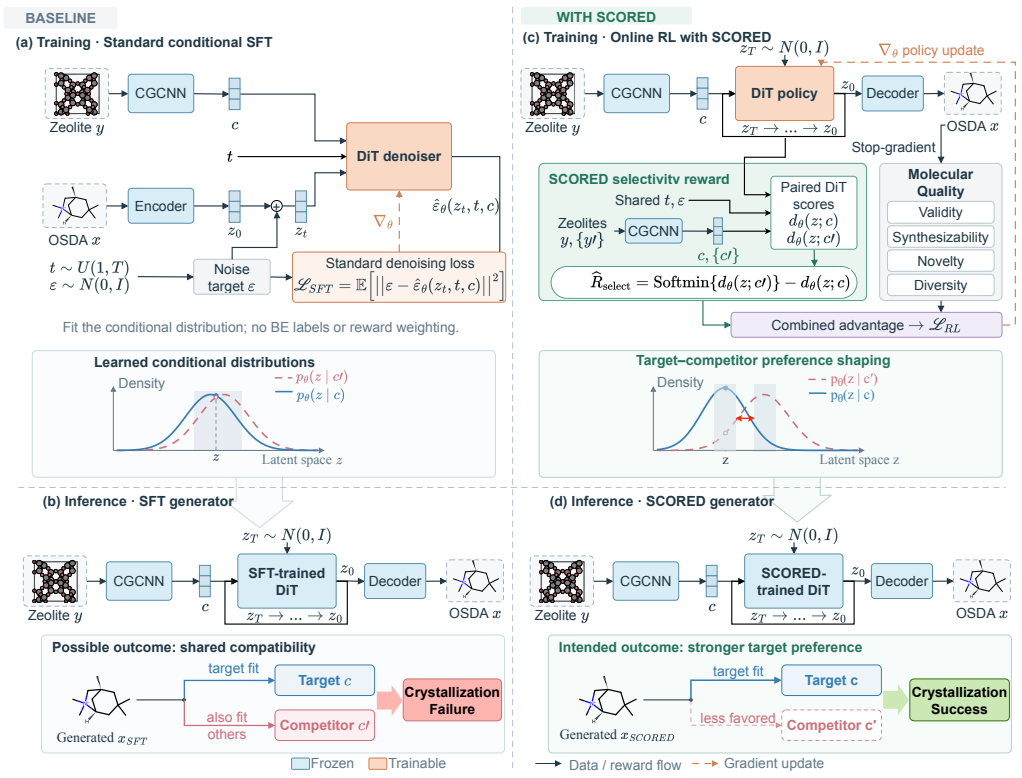


Figure 2: From target affinity to selectivity. (a,b) Supervised fine-tuning (SFT) permits molecules compatible with multiple frameworks. (c,d) SCORED uses shared-noise target–competitor scores to guide online RL alongside molecular-quality rewards. Mean-centered RWR initializes the policy before stage (c); Section 4.2 develops its binding-based rationale. Crystallization outcomes depict the intended design motivation, not validated synthesis results. OSDA figures are illustrative.

on the target interaction alone. This relative structure is also reflected in the Boltzmann-normalized competition metric used for physical evaluation (Equation (A.3)), which normalizes the target binding energy against those of competing frameworks. Inspired by this relative normalization, we construct a model-intrinsic analogue using conditional denoising errors, without assuming that denoising errors are physical energies or calibrated likelihoods. Given a fixed corruption ϵ at timestep t, a nonempty competitor set $\mathcal{C}_{\text{off}}(c) = \{c'_1, \ldots, c'_m\}$ and temperature $\tau > 0$, we define the normalized target-confidence score

$$s_\theta(c \mid x; \mathcal{C}_{\text{off}}, t, \epsilon) := \frac{\exp[-d_\theta(x, c; t, \epsilon)/\tau]}{\exp[-d_\theta(x, c; t, \epsilon)/\tau] + \sum_{j=1}^{m} \exp[-d_\theta(x, c'_j; t, \epsilon)/\tau]}. \tag{4}$$

To express this relative support as a zero-centered reward in denoising-error units, we define

$$\begin{aligned} M_\theta(x, c; t, \epsilon) &:= \tau \log \frac{s_\theta}{1 - s_\theta} + \tau \log m \\ &= \text{SoftMin}_\tau\big(\{d_\theta(x, c'_j; t, \epsilon)\}_{j=1}^{m}\big) - d_\theta(x, c; t, \epsilon), \end{aligned} \tag{5}$$

$$\text{SoftMin}_\tau(a_{1:m}) := -\tau \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{-a_j/\tau} \right). \tag{6}$$

The centering sets $M_\theta = 0$ when all conditions have equal errors, for any competitor-set size m. SoftMin emphasizes competitors with lower denoising errors, which provide stronger alternative fits to the same molecule. As $\tau \to 0$, the margin approaches the error gap between the best-denoising competitor and the target. Thus, the reward measures a target advantage relative to competing conditions, rather than absolute target fit. Note that this is a model-level preference score, not a calibrated probability of physical selection. The algebraic characterization is given in Appendix D.1.

Shared-corruption scoring. For each candidate we draw $\Omega = \{(t_r, \epsilon_r)\}_{r=1}^{K_t}$ across K_t respaced denoising steps, reuse each corruption across the target and all competitors (Li et al., 2023; Clark & Jaini, 2023), and average the resulting margins,

$$\widehat{\mathcal{R}}_{\text{select},\theta}(x, c; \Omega) = \frac{1}{K_t} \sum_{r=1}^{K_t} M_\theta(x, c; t_r, \epsilon_r). \tag{7}$$

Aggregating competitors within each corruption and averaging afterwards keeps the comparison paired at every (t_r, ϵ_r). The population score is $\mathcal{R}^{\text{pop}}_{\text{select},\theta}(x, c) = \mathbb{E}_{t\sim\nu,\epsilon}\left[M_\theta(x, c; t, \epsilon)\right]$ where ν denotes the scoring-timestep distribution. In the main configuration $\mathcal{C}_{\text{off}}(c) = \mathcal{C}_{\mathcal{B}} \setminus \{c\}$ contains the other conditions in the sampled minibatch $\mathcal{B}$.

4.2 Binding-Informed Initialization

The intrinsic comparison above operates on the conditional behavior already encoded by the denoiser; it does not itself introduce physical binding information. We therefore use offline binding data during initialization through reward-weighted regression (RWR). We use OSDA-wise mean-centered RWR since we observe OSDAs that bind well to various frameworks in the dataset. For each molecule, RWR assigns greater training weight to molecule–framework pairs whose binding energy is more favorable than that molecule's observed framework average.

Let $E_{\text{off}}(x, y)$ be the offline binding energy label from the literature for molecule x and framework y, in kJ/mol Si, with lower values preferred. For the set $\mathcal{O}_x$ of frameworks with available labels for molecule x, define

$$\mu_{\text{off}}(x) = \frac{1}{|\mathcal{O}_x|} \sum_{\tilde{y}\in\mathcal{O}_x} E_{\text{off}}(x, \tilde{y}), \qquad R_{\text{off}}(x, y) = \mu_{\text{off}}(x) - E_{\text{off}}(x, y). \tag{8}$$

Thus $R_{\text{off}}(x, y) > 0$ when the molecule x binds framework y more favorably than its average across the observed frameworks. RWR then increases the contribution of these favorable pairs to the denoising objective:

$$\mathcal{L}_{\text{RWR}}(\theta) = \mathbb{E}_{y\sim p_0}\mathbb{E}_{x\sim q_0(\cdot|y)}\left[\exp\{\beta R_{\text{off}}(x, y)\}\, D_\theta(x, e(y))\right]. \tag{9}$$

Here $p_0(y)$ and $q_0(x \mid y)$ describe the unweighted offline sampling distributions, and $\beta > 0$ controls the reweighting strength and has units of inverse energy. A positive utility suggests that the OSDA binds more favorably to the paired framework and increases the pair's training weight.

RWR and SCORED thus play complementary roles: RWR introduces binding-informed offline training emphasis, while SCORED compares the resulting conditional behavior online on generated candidates. The reweighted-distribution characterization in Appendix E.1 describes the training emphasis introduced by the offline objective. Thus, offline binding labels inform the model, while online scoring reuses its conditional denoising behavior on new candidates.

4.3 Turning Comparisons into Generative Feedback

We now use the comparison to change what the model generates. The denoiser and frozen decoder define the conditional policy $\pi_\theta(x \mid c)$. At each RL iteration, we sample a minibatch of target frameworks, generate several candidates per target, and compute SCORED and molecular-quality rewards. Candidates with higher combined advantages are favored by the policy update.

Balance relative preference with molecular quality. To discourage reward hacking as observed by Ghugare et al. (2024), we include molecular-quality rewards for validity, synthesizability, novelty, and diversity alongside selectivity; their definitions are in Appendix B.2.

Update generation, not the computed reward. The denoiser both generates and scores, so we freeze the reward inside each update. At outer iteration s the rollout policy is π_{θ_s}, and every channel reward is evaluated with θ_s and detached, $r_{j,k} = \text{sg}[\, r_j(x_k, c)\,]$, so no gradient flows through Equation (7). The candidates generated for one target form a rollout group $\mathcal{G}$; we standardize each channel within the group and sum the channels,

$$A_k = \sum_{j\in\mathcal{J}} \frac{r_{j,k} - \mu_{j,\mathcal{G}}}{\sigma_{j,\mathcal{G}} + \varepsilon}, \tag{10}$$

where $\mathcal{J}$ collects the SCORED selectivity reward and the molecular-quality channels, and $\mu_{j,\mathcal{G}}, \sigma_{j,\mathcal{G}}$ are the group mean and standard deviation of channel j. Sampling branches from a shared prefix, so candidate k is reached by a set of stochastic transitions $\mathcal{T}_k$, and its advantage credits all of them through one dimension-normalized trajectory ratio $\widetilde{\rho}_k$ (defined in Equation (A.4)),

$$\mathcal{L}_{\mathrm{RL}}(\theta) = -\,\mathbb{E}_k\big[\min\big(\widetilde{\rho}_k A_k,\ \mathrm{clip}(\widetilde{\rho}_k, 1-\delta, 1+\delta)A_k\big)\big]\,, \tag{11}$$

where δ is the clip range. Proximal policy optimization (PPO) (Schulman et al., 2017) takes several inner steps on Equation (11) per rollout, after which the updated denoiser generates and scores the next one. Hyperparameters are in Appendix B.2.

5 Experiments and Results

5.1 Experimental Setting

We set up three research questions (RQs) that ask: (1) whether SCORED improves OSDA selectivity, (2) which design choices matter, and (3) how and when SCORED generalizes to other domains.

Dataset and architecture. Our offline corpus covers 188 zeolite framework types, with molecules represented as Simplified Molecular-Input Line-Entry System (SMILES) strings (Jensen et al., 2021; Schwalbe-Koda et al., 2021; Xie et al., 2025). We exclude pairs with binding energy $E_{\mathrm{bind}} \leq -35$ kJ/mol Si, which can indicate unphysical framework bonding, and pairs with no dockable conformer (Xie et al., 2025), leaving 267,300 OSDA–zeolite pairs. Following Hu et al. (2025), we hold out six framework types for evaluation: AEI, AFI, AFX, ITE, LTA, and MOR. The six evaluation frameworks are excluded from SFT, RWR, and RL training. The generator is a 12-layer Diffusion Transformer (DiT; hidden size 384) in LDMol's pretrained variational autoencoder (VAE) latent space (Peebles & Xie, 2023; Chang & Ye, 2025; Higgins et al., 2017). The frozen VAE encodes SMILES as $h(x)$ and decodes sampled latents with a 98% reconstruction rate. A frozen crystal graph convolutional neural network (CGCNN) embeds framework y as $c = e(y)$ (Xie & Grossman, 2018).

Training pipeline. We first warm-start the DiT with 500,000 RWR steps on the offline corpus. We then perform 500 RL steps at a constant learning rate of 2×10^{-6}, optimizing selectivity alongside molecular-quality rewards. For $\widehat{R}_{select}$, each generated candidate uses the other frameworks in the same minibatch as competitors with one shared noise draw per respaced diffusion step (K_t=10). Appendix B.2 provides detailed training hyperparameters. We also provide the ablations of competitor-pool size, competitor sampling, and aggregation in Section 5.3.

Evaluating selectivity. We use competition scores to measure selectivity. Lower scores indicate greater selectivity (Equation (A.3)). We compute binding energies using the frozen-pose protocol with the DREIDING force field as implemented in GULP (Schwalbe-Koda & Gómez-Bombarelli, 2021a; Gale, 1997; Schwalbe-Koda & Gómez-Bombarelli, 2021b). We generate at most 20 Merck Molecular Force Field (MMFF) conformers and dock them into each framework using Voronoi docking. We optimize each host–guest complex, then separate host (zeolite) and guest (OSDA) at that optimized geometry and compute their single-point energies; subtraction from the complex energy gives binding energy (Halgren, 1996; Schwalbe-Koda & Gómez-Bombarelli, 2021b; Mayo et al., 1990). Candidate counts at each filtering stage are reported in Table A.3 to distinguish evaluation coverage from competition scores computed on successfully evaluated candidates. For all methods, competition scores are computed over the same six-framework panel; online reward scoring during RL uses only training frameworks as competitors. We report competitor docking coverage and discuss its implications for competition-score comparisons in Figure A.1.

Baselines and molecular quality. We compare SCORED with OSDA Agent (Hu et al., 2025) and with same-corpus SFT and RWR baselines before RL. We also evaluate OSDA Generator (Jensen et al., 2021) in Table A.4 but exclude it from the main analysis because of severe mode collapse. For OSDA Agent, we use generated samples provided by Hu et al. (2025); all other methods generate 500 samples per framework. Validity is the RDKit-parseable fraction. On all valid outputs, novelty is one minus maximum Tanimoto similarity to target-framework literature OSDAs, and diversity is mean pairwise Tanimoto distance among valid generations. SCScore measures synthetic complexity (Coley et al., 2018) with 1 being the easiest and 5 being the hardest.

Harmonic trade-off summary. To summarize trade-offs across different units and directions, we convert each metric to utility u, with larger values indicating better performance: reference/value

Table 1: Single-seed AFI/LTA comparison. Comp.: competition score (kJ/mol Si); Valid.: RDKit validity (%). Red subscripts give $\Delta = \text{row} - \text{SCORED}$ for Comp. and both H scores. Arrows indicate preferred directions; bold marks the best value, including ties.

Fw.	Method	Selectivity	Molecular quality				Trade-off	
		Comp.↓	Valid.↑	Nov.↑	Div.↑	SCScore↓	H_{sel} ↑	H_{eq} ↑
AFI	OSDA Agent	$11.44_{+8.92}$	**100.0**	0.68	0.75	**1.86**	$0.356_{-0.595}$	$0.565_{-0.359}$
	SFT	$8.77_{+6.25}$	98.4	0.66	0.85	2.46	$0.434_{-0.518}$	$0.624_{-0.301}$
	RWR	$8.37_{+5.85}$	90.4	0.75	**0.89**	3.09	$0.443_{-0.508}$	$0.617_{-0.307}$
	SCORED	**2.52**	94.4	**0.76**	**0.89**	2.51	**0.951**	**0.924**
LTA	OSDA Agent	$11.29_{+8.91}$	**100.0**	0.75	0.78	**1.75**	$0.345_{-0.597}$	$0.558_{-0.353}$
	SFT	$5.22_{+2.84}$	98.8	0.78	0.84	2.47	$0.602_{-0.340}$	$0.745_{-0.166}$
	RWR	$4.56_{+2.18}$	92.6	**0.81**	0.88	3.17	$0.637_{-0.305}$	$0.734_{-0.177}$
	SCORED	**2.38**	94.2	**0.81**	**0.89**	2.50	**0.942**	**0.911**

for competition score and SCScore, and value/reference otherwise. References are provided in Appendix G.2. Let u_{comp} denote selectivity utility and Q the harmonic mean of the others:

$$H(\omega) = \left[\frac{\omega}{u_{\text{comp}}} + \frac{1-\omega}{Q}\right]^{-1}, \qquad 0 \le \omega \le 1. \tag{12}$$

The selectivity-weighted $H_{\text{sel}} = H(1/2)$ assigns 50% to competition score and 12.5% to each quality metric; the metric-equal $H_{\text{eq}} = H(1/5)$ assigns 20% to each metric. Harmonic averaging penalizes weak utilities, so strength on one metric cannot freely compensate for weakness on another.

5.2 RQ1: Does SCORED Improve OSDA Selectivity?

Does SCORED improve selectivity on individual frameworks? Following OSDA Agent (Hu et al., 2025), we compare one SCORED run against it and against the same-corpus SFT and RWR stages trained with the same seed on two representative frameworks AFI and LTA in Table 1, where SCORED lowers competition score relative to every baseline on both. Across three runs, SCORED achieves competition scores of 2.92 ± 0.69 and 2.44 ± 0.45 kJ/mol Si on AFI and LTA, respectively (Table A.1; mean $\pm$ sample SD). Lower competition scores are not explained by reduced competitor coverage: Full SCORED has slightly higher mean competitor docking coverage than OSDA Agent (4.26 vs. 4.14; Appendix C.3).

Does improved selectivity come at the expense of molecular quality? While SCORED achieves better selectivity with higher novelty and diversity, OSDA Agent still has higher validity and lower synthetic complexity. We summarize this trade-off using the harmonic weighting conventions H_{sel} and H_{eq}. SCORED attains the highest value under both conventions on AFI and LTA (Table 1), and the same pattern still holds relative to its SFT and RWR training-stage ablations.

Does the target–competitor contrast matter? To isolate the contribution of SCORED, we conduct ablation studies on the necessity of the selectivity reward in Table 2 with architecture, data, RWR initialization, and quality rewards fixed. Full SCORED uses the complete pipeline with the SCORED reward. − competitor contrast replaces the SCORED reward with only the negative target-condition denoising error appearing in Equation (5), whereas − denoising reward removes the denoising-error reward entirely and retains only the molecular-quality rewards. Full SCORED achieves 5.026 ± 0.334 versus 6.503 ± 0.110 for − denoising reward and 6.340 ± 0.031 for − competitor contrast. Full SCORED reduces competition score by 20.7% relative to − competitor contrast and leads in both H_{sel} (0.964 versus 0.874) and H_{eq} (0.944 versus 0.932). Full SCORED exceeds − competitor contrast under H(ω) for selectivity weights $\omega > 0.156$, including both reported weighting conventions (Figure A.4). Table A.3 reports candidate counts at each stage: validity checking, deduplication, conformer generation, target docking, and competition-score calculation, with Full SCORED yielding the largest final count. An additional experiment using a binding-energy surrogate to provide a selectivity reward is reported in Table A.6; none of its generated candidates pass the physical-validity filter so we omit it from the main comparison. Together, these controls isolate the benefit of the target–competitor contrast: denoising-based feedback improves over molecular-quality optimization alone, and explicit competitor comparison improves further over target-only scoring.

Table 2: Ablation experiments on SCORED, averaged over six frameworks and multiple runs. Units follow Table 1; red deltas are relative to Full SCORED. H anchors are fixed (Appendix G.2); bold marks the best value. Sample SDs appear in Table A.2.

Ablation	Setting	Selectivity	Molecular quality				Trade-off	
		Comp.↓	Valid.↑	Nov.↑	Div.↑	SCScore↓	H_{sel} ↑	H_{eq} ↑
Reference	Full SCORED	**5.026**	**92.34**	0.720	0.889	1.908	**0.964**	**0.944**
Reward	− competitor contrast	$6.340_{+1.314}$	91.18	**0.764**	**0.895**	**1.683**	$0.874_{-0.090}$	$0.932_{-0.012}$
	− denoising reward	$6.503_{+1.477}$	91.26	0.712	0.884	1.894	$0.843_{-0.122}$	$0.891_{-0.053}$
Pipeline	− RWR initialization	$8.157_{+3.131}$	**95.61**	**0.763**	**0.920**	**1.637**	$0.762_{-0.202}$	$0.889_{-0.055}$
	− scorer updates	$7.149_{+2.123}$	89.60	0.710	0.883	1.957	$0.794_{-0.170}$	$0.862_{-0.082}$

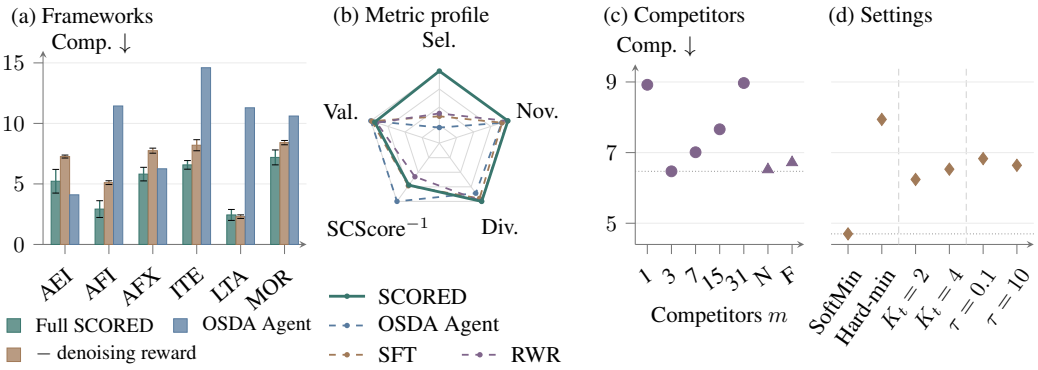


Figure 3: Selectivity and its trade-offs. (a) Per-framework mean ± sample SD for Full SCORED and − denoising reward. (b) Five best-relative utilities averaged over AFI/LTA (Sel.: selectivity; Val.: validity; Nov.: novelty; Div.: diversity; SCScore^{-1}: inverse synthetic complexity); outward is better (outer ring 1, spacing 0.25). (c) Controlled competitor sweep (32 frameworks/batch; N/F: nearest/farthest, $m = 3$). (d) Single-run reward diagnostics in three dashed-separated groups: aggregation at $(K_t, \tau)=(10, 1)$ (SoftMin versus hard-min), scoring draws $K_t \in \{2, 4\}$ at $\tau=1$, and temperature $\tau \in \{0.1, 10\}$ at $K_t=10$; (c,d) share the competition score axis. Dotted lines mark (c) random $m=3$ (6.47) and (d) a single-seed Full SCORED six-framework mean (4.70), which uses SoftMin aggregation. Comp. is in kJ/mol Si; details are in Appendices C.3, F and G.2.

5.3 RQ2: Which Design Choices Matter, and at What Cost?

Is SCORED cheap and stable enough for online optimization? On an NVIDIA A100, SCORED scoring on six conditions takes 7.86 ms, compared with approximately 8.3 s for a learned surrogate (Xie et al., 2025), which is 1056× faster (Table A.8). Shared noise lowers estimator standard deviation in all 16 diagnostic cells, with a geometric mean reduction of 56.3× (Appendix F.4).

Which reward-design choices matter? More comparisons do not guarantee better selectivity (Figure 3(c,d)). In the controlled competitor sweep, random $m = 3$ gives the lowest competition score, while $m = 7$ and 31 favor quality. Nearest selection modestly improves on farthest selection, but neither beats random $m = 3$. These results show that more competitors do not necessarily provide better training feedback: a small, randomly sampled competitor set performs best for selectivity in this sweep. SoftMin yields a lower competition score than hard-min, whereas hard-min gives higher validity (Appendix F). This comparison favors SoftMin for selectivity, suggesting that smoothly weighting plausible competitors is more effective in line with our theoretical analysis in Appendix D.1.

Do binding-informed initialization and scorer updates matter? Table 2 tests whether binding-informed initialization and scorer updates matter. − RWR initialization replaces the RWR warm start with SFT, testing whether selectivity optimization can recover without binding-informed initialization. − scorer updates retains the RWR initialization but fixes the scorer during RL, testing whether the initial scorer remains sufficient throughout policy optimization. Their higher competition scores support retaining both RWR initialization and scorer updates in the evaluated setting. Appendix E.3 reports a case study on LTA and MOR: the RWR-initialized scorer ranks candidates in agreement with

both physical selectivity endpoints, an SFT-initialized one does not, and the ordering survives across all three Full SCORED checkpoints. This suggests that updating the scorer alongside the policy does not degrade it into a self-reinforcing signal. Together, these ablations suggest that selectivity benefits from both binding-informed initialization and adapting the scorer as the policy distribution shifts.

5.4 RQ3: Does SCORED generalize beyond zeolites, and when does it work?

We test SCORED beyond OSDA design through structure-based drug design (SBDD) on unseen pockets and multiple backbones, and metal–organic framework (MOF) generation with MOFFUSION-H_2. Outcomes are Vina selectivity gap for SBDD and predicted hydrogen working capacity for MOF, both with larger values preferred. Detailed setup can be found in Appendix H.

Does SCORED transfer beyond zeolites? Using 30 training pockets and 30 disjoint evaluation pockets, we find that adding SCORED to TargetDiff (Guan et al., 2023) increases mean Vina selectivity gap from 2.21 to 2.42 kcal/mol (Table A.11). The effect is not uniform across SBDD backbones: IPDiff (Huang et al., 2024) changes by -0.02 kcal/mol in Vina gap, whereas AliDiff (Gu et al., 2024) changes by $+0.16$ kcal/mol (Table A.13). This backbone dependence motivates examining whether the pretrained model already contains score–property alignment before RL. For MOFFUSION-H_2, the matched reward toggle increases mean predicted hydrogen working capacity from 32.03 to 34.28 g/L, although neither arm uses the hydrogen-capacity reward (Table A.14).

Table 3: Cross-domain summary with or without SCORED. Alignment is measured before RL. Outcome units are kJ/mol Si for OSDA, kcal/mol for SBDD, and g/L for MOFFUSION-H_2.

Setting	τ_b	Without	With
OSDA	.209	6.50	**5.03**
TargetDiff	-.063	2.21	**2.42**
AliDiff	.126	2.41	**2.57**
IPDiff	-.021	**2.55**	2.53
MOFFUSION	.224	32.03	**34.28**

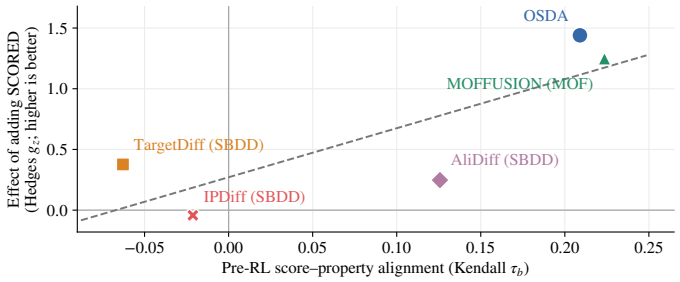


Figure 4: Cross-setting relation between pre-RL score–property alignment and the standardized downstream effect of SCORED. The dashed fit is descriptive; definitions are in Equation (A.18).

When does SCORED help? As shown in Figure 4, the OSDA and MOF settings combine the strongest pre-RL alignment with the largest downstream gains, while the three SBDD backbones cluster at weak alignment and small effects. The IPDiff–AliDiff pair is the cleanest controlled comparison since AliDiff finetunes IPDiff for target awareness through Energy Preference Optimization, and it shows both higher pre-RL alignment and a larger gain than IPDiff. This pattern is consistent with target-aware pretraining or fine-tuning, including AliDiff and binding-informed RWR, helping establish useful pre-RL score–property alignment. With only five settings, we treat the overall pattern as a descriptive generalization hypothesis rather than a mechanistic or predictive law.

6 Conclusion

SCORED replaces expensive external selectivity evaluation with a cheap intrinsic signal: how much better a conditional denoiser matches a candidate to the target than to competitors, with shared noise keeping the comparison paired. Following mean-centered RWR initialization, optimizing the generator with this reward alongside molecular-quality rewards lowers competition scores relative to both a strong LLM-based generator and matched ablations that remove the competitor contrast.

However, competition scores rank computational target preference within a fixed six-framework panel; whether the selected OSDAs crystallize also depends on gel chemistry and crystallization conditions that we do not model. Cross-domain experiments further show that the benefit of SCORED is not uniform across backbones, with uneven performance across different models.

More broadly, our results suggest that when physical evaluation is too expensive for online RL, conditional denoising comparisons can provide useful target-relative signal when the underlying model already encodes property alignment, without the need for physical evaluation during optimization.

AI USE STATEMENT

In this work, we used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments and implementing methods. We have not used generative AI tools for assisting in translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, interpreting results and generating synthetic data sets. Additionally, we used generative AI tools for proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. LLM-generated code and claims have been verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops molecular-generation methods for zeolite templates and pocket-binding scaffolds. Generated candidates require standard safety assessment before synthesis or experimental use, including toxicity and off-target testing where relevant.

REPRODUCIBILITY STATEMENT

The appendix documents OSDA training and evaluation for zeolites, reward wall-clock time profiling, ablations, and frozen-pose method details. The dataset for RWR training is obtained by merging several public datasets for OSDA-zeolite discovery; the merged corpus will be released upon acceptance. The SBDD experiments use the public CrossDocked2020 (Francoeur et al., 2020), pretrained TargetDiff, IPDiff, and AliDiff checkpoints (Guan et al., 2023; Huang et al., 2024; Gu et al., 2024), a 30%-sequence-identity cluster split, and AutoDock Vina v1.2.5 at exhaustiveness 8 (Appendix H). Recentering, zeolite and TargetDiff RL hyperparameters, held-out evaluation, and the MOF toggle are specified in Section 4.2, Appendices B.2 and H.2, and Table A.14. We will release source code, trained checkpoints, and generated structures upon acceptance.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12 (1-17):1, 2005.

Jinho Chang and Jong Chul Ye. LDMol: A text-to-molecule diffusion model with structurally informative latent space surpasses AR models. 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. 7374–7392. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chang25d.html.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=dwSpo5DRk8.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=fxNQJVMwK2.

Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. SCScore: Synthetic complexity learned from a reaction corpus. *Journal of Chemical Information and Modeling*, 58(2): 252–261, 2018. doi: 10.1021/acs.jcim.7b00622. URL https://doi.org/10.1021/acs.jcim.7b00622. PMID: 29309147.

Frits Daeyaert and Michael W. Deem. Design of organic structure directing agents for polymorph A zeolite beta. *Journal of Materials Chemistry A*, 7(16):9854–9866, 2019. doi: 10.1039/C8TA11913A. URL https://doi.org/10.1039/C8TA11913A.

Frits Daeyaert, Fengdan Ye, and Michael W. Deem. Machine-learning approach to the design of OSDAs for zeolite beta. *Proceedings of the National Academy of Sciences*, 116(9):3413–3418, February 2019. doi: 10.1073/pnas.1818763116. URL https://www.pnas.org/doi/10.1073/pnas.1818763116. Publisher: Proceedings of the National Academy of Sciences.

Zheng Ding and Weirui Ye. TreeGRPO: Tree-advantage grpo for online rl post-training of diffusion models. *arXiv preprint arXiv:2512.08153*, 2025.

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 *NeurIPS*, 2023.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Xiang Fu, Tian Xie, Andrew S. Rosen, Tommi Jaakkola, and Jake Smith. MOFDiff: Coarse-grained diffusion for metal-organic framework design. In *International Conference on Learning Representations (ICLR)*, 2024.

Julian D. Gale. GULP: A computer program for the symmetry-adapted simulation of solids. *J. Chem. Soc., Faraday Trans.*, 93(4):629–637, 1997. doi: 10.1039/A606455H.

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

Raj Ghugare, Santiago Miret, Adriana Hugessen, Mariano Phielipp, and Glen Berseth. Searching for high-value molecules using reinforcement learning and transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nqlymMx42E.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 *ICLR*, 2023.

Thomas A. Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996. doi: https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C490%3A%3AAID-JCC1%3E3.0.CO%3B2-P.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. OSDA agent: Leveraging large language models for de novo design of organic structure directing agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9YNyiCJE3k.

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

Shusuke Ito, Koki Muraoka, and Akira Nakayama. De Novo Design of Organic Structure-Directing Agents for Zeolites Using a General-Purpose Large Language Model, March 2024. URL https://chemrxiv.org/engage/chemrxiv/article-details/65f9481766c13817294a415e.

Zach Jensen, Soonhyoung Kwon, Daniel Schwalbe-Koda, Cecilia Paris, Rafael Gómez-Bombarelli, Yuriy Román-Leshkov, Avelino Corma, Manuel Moliner, and Elsa A. Olivetti. Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks. *ACS Cent. Sci.*, 7(5):858–867, May 2021. ISSN 2374-7943. doi: 10.1021/acscentsci.1c00024. URL https://doi.org/10.1021/acscentsci.1c00024. Publisher: American Chemical Society.

Seungwook Kim and Minsu Cho. Improving text-to-image generation with intrinsic self-confidence rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dewi W Lewis, David J Willock, C Richard A Catlow, John Meurig Thomas, and Graham J Hutchings. De novo design of structure-directing agents for the synthesis of microporous solids. *Nature*, 382 (6592):604–606, 1996.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023.

Ruijin Li, Danlian Huang, Zhen Wei, Yashi Chen, Guangfu Wang, Wei Zhou, Ruihao Xiao, and Wenbo Xu. Synthesis of biomass-mediated hierarchical zeolite socony mobil–5 and their selectivity for lignin conversion to bio-oils: Review and perspective. *Renewable and Sustainable Energy Reviews*, 207:114977, 2025.

Jie Liu et al. Flow-GRPO: Training flow matching models via online RL, 2025. arXiv preprint.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Stephen L. Mayo, Barry D. Olafson, and William A. Goddard. Dreiding: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, 94(26):8897–8909, 1990. doi: 10.1021/j100389a010. URL https://doi.org/10.1021/j100389a010.

Koki Muraoka, Watcharop Chaikittisilp, and Tatsuya Okubo. Multi-objective de novo molecular design of organic structure-directing agents for zeolites using nature-inspired ant colony optimization. *Chemical Science*, 11:8214–8223, 2020. doi: 10.1039/D0SC03075A. URL https://doi.org/10.1039/D0SC03075A.

Geoffrey A Ozin, Alex Kuperman, and Andreas Stein. Advanced zeolite, materials science. *Angewandte Chemie International Edition in English*, 28(3):359–376, 1989.

Hyoungjoon Park, Hwanhee Kim, Seungyeon Choi, Seungyong Lee, Yoonju Kim, and Sanghyun Park. TheSelective: Dual affinity-guided diffusion for selective molecular generation. In *Advances in Knowledge Discovery and Data Mining*, volume 16598 of *Lecture Notes in Computer Science*, pp. 16–27. Springer, 2026. doi: 10.1007/978-981-92-1462-4_2. URL https://doi.org/10.1007/978-981-92-1462-4_2.

Junkil Park, Youhan Lee, and Jihan Kim. Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation. 2024.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 745–750, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023. arXiv:2310.03739.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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, Dec 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00737-x. URL https://doi.org/10.1038/s43588-024-00737-x.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Daniel Schwalbe-Koda and Rafael Gómez-Bombarelli. Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites. *The Journal of Chemical Physics*, 154 (17), 2021a.

Daniel Schwalbe-Koda and Rafael Gómez-Bombarelli. Supramolecular recognition in crystalline nanocavities through monte carlo and voronoi network algorithms. *The Journal of Physical Chemistry C*, 125(5):3009–3017, 2021b.

Daniel Schwalbe-Koda, Soonhyoung Kwon, Cecilia Paris, Estefania Bello-Jurado, Zach Jensen, Elsa Olivetti, Tom Willhammar, Avelino Corma, Yuriy Román-Leshkov, Manuel Moliner, and Rafael Gómez-Bombarelli. A priori control of zeolite phase competition and intergrowth with high-throughput simulations. *Science*, 374(6565):308–315, October 2021. doi: 10.1126/science.abh3350. URL https://www.science.org/doi/10.1126/science.abh3350. Publisher: American Association for the Advancement of Science.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Heqi Sun, Hong Tan, Yanyi Chu, Jiayi Li, Ruixuan Wang, and Dong-Qing Wei. Design of permeability-optimized target-binding macrocycles via direct preference optimization. *Chemical Science*, 17:10223–10236, 2026. doi: 10.1039/D6SC01722C. URL https://doi.org/10.1039/D6SC01722C.

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 *CVPR*, 2024.

Yan Wen, Changhong Zhan, Jiacheng Liu, Xinxin Zhuang, Siyu Liu, Tang Yang, Wenqiang Liu, Xiaozhi Liu, Cheng-Wei Kao, Yu-Cheng Huang, et al. Zeolite-confined Cu single-atom clusters stably catalyse CO to acetate at 1 A cm^{-2} beyond 1,000 h. *Nature Nanotechnology*, 2025. doi: 10.1038/s41565-025-01892-6.

Huixin Wu, Hai Wang, Yating Lv, Yuexin Wu, Yike Wang, Qingsong Luo, Yu Hui, Lujie Liu, Mengting Zhang, Kunming Hou, Lina Li, Jianrong Zeng, Weili Dai, Liang Wang, and Feng-Shou Xiao. Ultra-small metallic nickel nanoparticles on dealuminated zeolite for active and durable catalytic dehydrogenation. *Angewandte Chemie International Edition*, 64(8):e202420306, 2025. doi: https://doi.org/10.1002/anie.202420306. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202420306.

Mingrou Xie, Daniel Schwalbe-Koda, Yolanda Marcela Semanate-Esquivel, Estefanía Bello-Jurado, Alexander Hoffman, Omar Santiago-Reyes, Cecilia Paris, Manuel Moliner, and Rafael Gómez-Bombarelli. A comprehensive mapping of zeolite–template chemical space. *Nat Comput Sci*, 5 (8):661–674, August 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00842-5. URL https://www.nature.com/articles/s43588-025-00842-5. Publisher: Nature Publishing Group.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

Liukou Xu, Xin Peng, Zhenhao Xi, Zhiqing Yuan, and Weimin Zhong. Predicting organic structures directing agents for zeolites with conditional deep learning generative model. *Chemical Engineering Science*, 282:119188, December 2023. ISSN 00092509. doi: 10.1016/j.ces.2023.119188. URL https://linkinghub.elsevier.com/retrieve/pii/S0009250923007443.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Chenning Yu and Sicun Gao. Improving compositional generation with diffusion models using lift scores. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Bo-Jun Zeng, Si-Ming Wu, Markus Antonietti, Christoph Janiak, Osamu Terasaki, and Xiao-Yu Yang. Hierarchy design of zeolite mesocrystals. *Chemical Engineering Journal*, pp. 159282, 2025.

A Evaluation Definitions

A.1 Binding Energies and Competition Score

For OSDA molecule x and zeolite framework y, the binding energy is

$$E_{\mathrm{bind,Si}}(x,y) = \frac{E_{\mathrm{h+g}}(x,y) - \big(E_{\mathrm{host}}(y) + nE_{\mathrm{guest}}(x)\big)}{N_{\mathrm{Si}}(y)}, \tag{A.1}$$

where n is the number of guest molecules per cell, $N_{\mathrm{Si}}(y)$ is the number of Si atoms in the framework cell. We use E_{bind} as a shorthand for $E_{\mathrm{bind,Si}}$ throughout this manuscript. Several conformers of x are first generated and then docked individually to y. Upon successful docking, the docked complex with both host (zeolite y) and guest (OSDA x) is optimized with the DREIDING force field to obtain $E_{\mathrm{h+g}}(x,y)$. Host and guest single-point energies are then calculated at the geometries extracted from the optimized complex, without further relaxation. More negative binding energies indicate stronger host–guest stabilization.

Given a set of competing frameworks $\mathcal{Y}$, the competition energy defined by Schwalbe-Koda et al. (2021) is

$$E_c(x,y) = E_{\mathrm{bind}}(x,y) - \min_{y'\in\mathcal{Y}} E_{\mathrm{bind}}(x,y'). \tag{A.2}$$

We adopt the Boltzmann-normalized competition metric introduced by Schwalbe-Koda et al. (2021), restricting it to the competition component and evaluating it as

$$\begin{aligned}
C_{BN}(x,y;\mathcal{Y}) &= \frac{\exp[-E_c(x,y)/(R_{\mathrm{gas}}\mathcal{T})]}{\sum_{y'\in\mathcal{Y}}\exp[-E_c(x,y')/(R_{\mathrm{gas}}\mathcal{T})]} \\
&= \frac{\exp[-(E_{\mathrm{bind}}(x,y) - \min_{y'\in\mathcal{Y}} E_{\mathrm{bind}}(x,y'))/(R_{\mathrm{gas}}\mathcal{T})]}{\sum_{y'\in\mathcal{Y}}\exp[-(E_{\mathrm{bind}}(x,y') - \min_{y'\in\mathcal{Y}} E_{\mathrm{bind}}(x,y'))/(R_{\mathrm{gas}}\mathcal{T})]} \\
&= \frac{\exp[\min_{y'\in\mathcal{Y}} E_{\mathrm{bind}(x,y')}/R_{\mathrm{gas}}\mathcal{T}]\exp[-E_{\mathrm{bind}}(x,y)/R_{\mathrm{gas}}\mathcal{T}]}{\sum_{y'\in\mathcal{Y}}\{\exp[\min_{y'\in\mathcal{Y}} E_{\mathrm{bind}}(x,y')/R_{\mathrm{gas}}\mathcal{T}]\exp[-E_{\mathrm{bind}}(x,y')/R_{\mathrm{gas}}\mathcal{T}]\}} \\
&= \frac{\exp[-E_{\mathrm{bind}}(x,y)/(R_{\mathrm{gas}}\mathcal{T})]}{\sum_{y'\in\mathcal{Y}}\exp[-E_{\mathrm{bind}}(x,y')/(R_{\mathrm{gas}}\mathcal{T})]} \\
S_{\mathrm{comp}}(x,y;\mathcal{Y}) &= -R_{\mathrm{gas}}\mathcal{T}\log C_{BN}(x,y;\mathcal{Y}),
\end{aligned} \tag{A.3}$$

where C_{BN} is Boltzmann-normalized competition and S_{comp} is the *competition score* (in kJ/mol Si), with lower values preferred. We use $\mathcal{T} = 400$ K and $R_{\mathrm{gas}} = 8.314462618 \times 10^{-3}$ kJ mol^{-1} K^{-1} with the per-Si binding convention. Here $\mathcal{Y}$ contains the eligible frameworks for x from the common evaluation set $\{\mathrm{AEI}, \mathrm{AFI}, \mathrm{AFX}, \mathrm{ITE}, \mathrm{LTA}, \mathrm{MOR}\}$ following Hu et al. (2025), including the requested target. If no valid docking position exists for a competing framework, we exclude that framework from the S_{comp} calculation. Six-framework results are unweighted means of S_{comp} over those framework averages.

B Detailed Training Procedure

For the offline warm-start, RWR (Peters & Schaal, 2007) weights each denoising loss by $\exp\{\beta R_{\mathrm{off}}(x,y)\}$ using the mean-centered binding reward in Equation (8). A theoretical analysis of the impact of RWR can be found in Appendix E. Following denoising diffusion policy optimization (DDPO) (Black et al., 2024), we treat reverse denoising as a Markov decision process for RL. Our TreeGRPO-based implementation (Ding & Ye, 2025) reuses trajectory prefixes through branching; each candidate's standardized advantage credits the stochastic transitions on its own path. The novelty and diversity rewards use exact matches to known target OSDAs and within-rollout fingerprint dissimilarity for unique, mutually dissimilar SMILES.

B.1 Policy Update and Reward Detachment

At outer iteration s, the current denoiser θ_s samples trajectories and computes the SCORED selectivity reward using Equation (7). Detached rewards yield channel-normalized leaf advantages, which are

Algorithm 1 SCORED: mean-centered RWR and online selectivity optimization

Require: Offline binding records, denoiser $\hat{\epsilon}_\theta$, frozen encoders/decoder
1: Compute $R_{\text{off}}(x, y) = \mu_{\text{off}}(x) - E_{\text{off}}(x, y)$ and warm-start θ using Equation (9)
2: **for** each RL iteration **do**
3: Sample a condition minibatch $\mathcal{C}_\mathcal{B}$ and branching latent rollouts $z_{0,k}$ for each target c
4: For each generated latent $z_{0,k}$, sample shared scoring corruptions $\Omega_k = \{(t_r, \epsilon_r)\}_{r=1}^{K_t}$
5: For each r, compute $d_\theta(z_{0,k}, c'; t_r, \epsilon_r)$ for the target and every competitor $c' \in \mathcal{C}_\mathcal{B}$, using the same (t_r, ϵ_r) across conditions
6: Compute $\widehat{\mathcal{R}}_{\text{select},\theta}$ using Equation (7): apply SoftMin across competitor errors at each timestep, subtract the corresponding target error, then average over timesteps
7: Decode $x_k = g(z_{0,k})$ and evaluate validity and molecular-quality rewards; retain invalid decodings in the rollout group, applying the invalid-output conventions in Appendix B.2
8: Detach rewards, standardize each channel within the rollout group, sum channels, and update θ with clipped PPO using the dimension-normalized trajectory ratio in Equation (A.4)
9: **end for**
Ensure: Fine-tuned conditional generator π_θ

combined and propagated to tree edges. For each tracked stochastic transition e, define the mean Gaussian log-density over the d_z latent dimensions:

$$\bar{\ell}_\theta(e) = \frac{1}{d_z} \sum_{i=1}^{d_z} \log p_\theta(z_{t-1,i} \mid z_t, t, c),$$
$$\widetilde{\rho}_k(\theta) = \exp\left[\sum_{e \in \mathcal{T}_k} \left(\bar{\ell}_\theta(e) - \bar{\ell}_{\text{old}}(e) \right) \right]. \tag{A.4}$$

Here $\mathcal{T}_k$ contains the two stochastic branching transitions on the path to candidate k. The remaining deterministic DDIM transitions are excluded from this calculation. We apply clipping once to the resulting dimension-normalized trajectory ratio in the PPO-style objective. The full training process can be found in Algorithm 1. During rollout scoring, denoising errors are evaluated directly on the generated latent $z_{0,k}$, without decoding and re-encoding. Consequently, the selectivity reward remains defined even when $x_k = g(z_{0,k})$ is not a valid molecule. The parameters are updated by minimizing the loss $L_{\text{RL}}(\theta)$ defined in Equation (11) with dimension-normalized trajectory ratios provided in Equation (A.4) and the advantages defined in Equation (10) by $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{RL}}(\theta)$ over M inner epochs per rollout, where θ_{old} is the parameter used to collect and score this rollout, after which the scorer becomes the updated denoiser.

B.2 Hyperparameters

Offline warm start. For RWR we set the weighting coefficient β to $0.2\,(\text{kJ mol}^{-1}\,\text{Si})^{-1}$ and train for $500{,}000$ steps with a batch size 512 on one NVIDIA A100, using a cosine-decayed learning rate initialized at 2×10^{-5} with the first 10% of steps ($50{,}000$ steps) as warmup. AdamW optimizer with 0 weight decay is used. The forward diffusion horizon is $1{,}000$ steps.

Rollouts. For RL the diffusion is respaced to 10 steps. Each rollout branches at the first two reverse steps with branching factor 3, giving 9 candidates per target; these two transitions are sampled ancestrally and their Gaussian log-probabilities are tracked, while the remaining eight steps use deterministic DDIM (Song et al., 2021) updates and are excluded from the ratio calculation. Log-probabilities use the model's learned variance and are averaged over latent dimensions, so $\delta = 10^{-4}$ in our setting is not comparable to a token-level clip range that is normally set at the scale of 10^{-1}. Each step samples four framework conditions, and classifier-free guidance uses scale 2.5.

Selectivity reward. Each candidate is scored with one shared corruption at every respaced timestep (K_t=10), aggregated by SoftMin with τ=1; competitors are the other conditions present in the step except for those in competitor sampling ablations. In the default configuration the scorer is the current policy; the frozen variant in Table 2 instead uses a copy taken immediately after RWR.

Reward channels. Five channels are active: the SCORED selectivity reward; validity, which contributes $+1$ for an RDKit-parseable decoding and -10 otherwise; synthesizability $5 - \mathrm{SCScore}$; novelty against known OSDAs for the target, combining an unseen-SMILES term with $1 - \max$ Tanimoto similarity; and diversity, $1-$ mean pairwise Tanimoto similarity with other candidates in the step. Channels are standardized within the rollout group (unbiased standard deviation, $\varepsilon{=}10^{-8}$) and summed with unit weights. A candidate whose decoding fails keeps its place in the group with a zero value for the channels that require a parsed molecule, while its selectivity reward is still computed from the latent.

Optimization. We use clip range $\delta{=}10^{-4}$, 16 inner epochs per rollout, gradient-norm clipping at 1.0, and AdamW with learning rate $2{\times}10^{-6}$ and weight decay 10^{-2}, for 500 RL steps.

C OSDA Performance and Reward Attribution

C.1 AFI/LTA Repeat-Run Robustness

SCORED consistently achieves lower competition scores than OSDA Agent on both AFI and LTA (Table A.1). The same runs are summarized across six frameworks in Figure 3(a) and Table A.2.

Table A.1: Repeat-run force-field competition score on AFI and LTA. OSDA Agent is a point reference, with red deltas relative to the SCORED mean. Each run generates 500 candidates per framework and uses the same six-framework evaluator. Values are in kJ/mol Si (lower is better); $\pm$ denotes sample standard deviation.

Framework	OSDA Agent	SCORED mean$\pm$SD
AFI	$11.44_{+8.52}$	2.92 ± 0.69
LTA	$11.29_{+8.85}$	2.44 ± 0.45

C.2 Matched Three-Arm Reward Ablation

This ablation asks whether the competitor comparison adds value beyond target-only scoring ($-$ competitor contrast). Full SCORED uses the complete selectivity reward, $-$ competitor contrast uses $-\frac{1}{K_t}\sum_r d_\theta(x, c; t_r, \epsilon_r)$, and $-$ denoising reward removes the selectivity reward (Table 2). Full SCORED and $-$ competitor contrast both use the current generator as the scorer.

All arms retain validity, synthesizability, novelty, and diversity rewards with matched RWR initialization, data, condition schedule, and generation settings. They use the hyperparameters in Appendix B.2. Selectivity is evaluated after generation with the competition score using the common six-framework protocol below.

C.3 Repeated-Run Scorer and Molecular-Quality Summaries

For each setting, we evaluate three runs with different random seeds. The primary run reported in Table 1 is one of these three runs, rather than an additional run. Primary-run tables report results from this designated run, whereas repeated-run tables report the mean and sample standard deviation across all three runs. Table A.2 summarizes repeated-run competition scores and molecular quality for the reward and initialization ablations with three runs per setting. In the frozen variant, a separate copy of the initial RWR denoiser supplies rewards while the generator continues to learn. Each run generates 500 outputs per target; run-level metrics average the six target means equally, and uncertainty is the sample standard deviation across runs.

Per-framework repeated-run visualization. For Figure 3(a), Full SCORED's framework means (sample SD) are AEI 5.22 (0.98), AFI 2.92 (0.69), AFX 5.81 (0.56), ITE 6.57 (0.36), LTA 2.44 (0.45), and MOR 7.19 (0.62) kJ/mol Si. Averaging the six frameworks within each run and then summarizing the three runs gives 5.026 ± 0.334 kJ/mol Si. All three Full SCORED run-level means are below those of both $-$ denoising reward and $-$ competitor contrast, with mean reductions of 1.48 and 1.31 kJ/mol Si, respectively.

Table A.2: Repeated-run six-framework summaries. Raw metrics are mean ± sample SD; H scores use method-level mean metrics with fixed anchors (Appendix G.2), without an H standard deviation. Red deltas use the corresponding Full SCORED summary; bold marks the best H score. Units follow Table 1.

Setting	Comp.↓	Valid.↑	Nov.↑	Div.↑	SCScore↓	H_{sel} ↑	H_{eq} ↑
Full SCORED	5.026 ±0.334	92.34 ±1.76	0.7201 ±0.0153	0.8892 ±0.0114	1.908 ±0.080	**0.964**	**0.944**
− competitor contrast	$6.340_{+1.314}$ ±0.031	91.18 ±1.23	0.7640 ±0.0103	0.8954 ±0.0023	1.683 ±0.083	$0.874_{-0.090}$	$0.932_{-0.012}$
− denoising reward	$6.503_{+1.477}$ ±0.110	91.26 ±2.26	0.7118 ±0.0054	0.8838 ±0.0068	1.894 ±0.075	$0.843_{-0.122}$	$0.891_{-0.053}$
− scorer updates	$7.149_{+2.123}$ ±0.248	89.60 ±1.74	0.7102 ±0.0121	0.8833 ±0.0158	1.957 ±0.018	$0.794_{-0.170}$	$0.862_{-0.082}$
− RWR initialization	$8.157_{+3.131}$ ±0.270	95.61 ±1.76	0.7630 ±0.0131	0.9203 ±0.0108	1.637 ±0.161	$0.762_{-0.202}$	$0.889_{-0.055}$

Removing RWR initialization. − RWR initialization starts from SFT rather than RWR and yields a six-framework competition score of 8.16 ± 0.27 kJ/mol Si. − RWR initialization has a within-framework unique-among-valid rate of $62.61 \pm 8.58\%$, mean MW of 113.4 ± 16.7 Da. These diagnostics reveal concentration on small, repeated molecules despite high validity. Frequent outputs include `C[NH+]C`, `C[N+](C)CO`, and `[NH3+]O`. Low SCScore thus accompanies a small-molecule generation pattern rather than establishing an overall quality advantage.

Generation-to-evaluation filtering. Table A.3 reports candidate counts through the generation and physical-evaluation pipeline across six frameworks. Each setting generates 500 candidates per framework, giving 3,000 outputs in total. Deduplication is performed on all valid SMILES across different frameworks; hence the final number of unique SMILES may differ from those in other tables.

Table A.3: Candidate counts through the generation and physical-evaluation pipeline across six frameworks. Counts are different from the designated runs included in Table 1 and Table 2. Valid: RDKit-valid outputs; Unique: distinct valid structures; Conf.: successful conformer generation; Target: successful target docking; Joint: successful target and competitor evaluations. Joint/Gen. is the fraction of generated outputs reaching joint success. Counts describe evaluation coverage rather than selectivity.

Setting	Generated	Valid	Unique	Conf.	Target	Joint	Joint/Gen.
SFT	3000	2961	2461	2244	644	552	18.4%
RWR	3000	2721	2713	2453	332	255	8.5%
− RWR initialization	3000	2872	1940	1876	901	840	28.0%
Full SCORED	3000	2813	2198	2115	1208	1150	38.3%
− denoising reward	3000	2788	2201	2187	1228	1126	37.5%
− competitor contrast	3000	2733	2259	2248	635	468	15.6%

Full SCORED achieves a higher target-docking success rate among conformer-successful candidates than − denoising reward (57.1% versus 56.1%). These counts characterize evaluation coverage; selectivity is assessed separately using competition scores, which can be affected by the number of competitor energies available, so we provide the following analysis.

Competitor docking coverage. Omitting unavailable competitors can lower the competition score for a fixed molecule, making evaluation coverage a potential confounder. Figure A.1 therefore reports competitor docking coverage. Full SCORED has a mean of 4.26 successfully docked competitors, compared with 4.14 for OSDA Agent, and a larger fraction of molecules with all five competitors successfully docked (66% versus 56%). Thus, Full SCORED does not achieve lower competition scores on these aggregate measures through reduced competitor coverage.

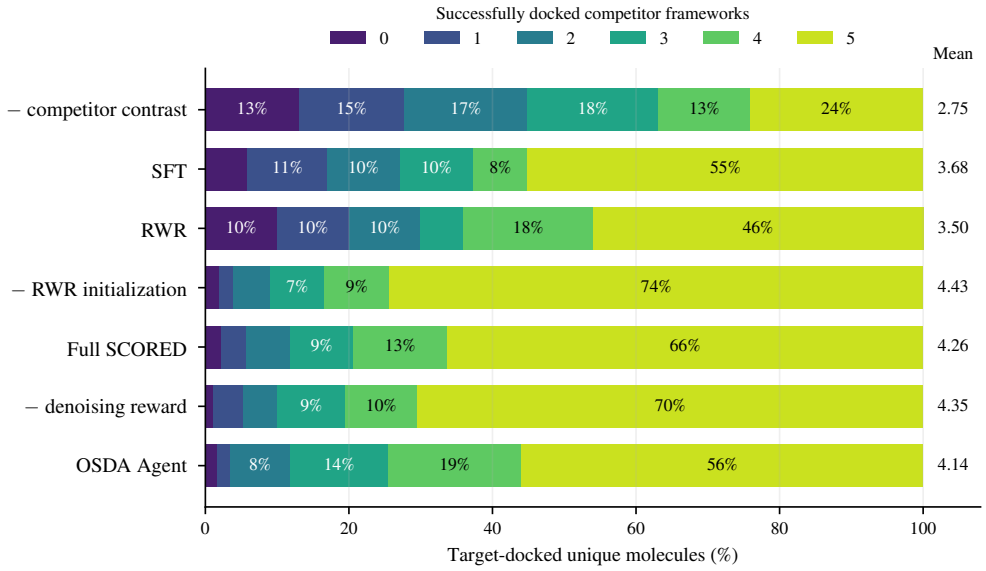


Figure A.1: Competitor docking coverage among target-docked unique molecules. Stacked bars show the percentage with zero to five successfully docked competitor frameworks; the right column reports the mean count.

Molecular-quality population. All molecular-quality statistics originate from the 500 outputs generated per target, independently of GULP candidate selection. Validity is the RDKit-parseable fraction of these outputs. Novelty and diversity use all RDKit-valid outputs, including repeats, without binding-based filtering or deduplication.

C.4 OSDA Generator Baseline

The OSDA Generator (Jensen et al., 2021) provides a complementary molecular-quality comparison at the same 500-generation budget (Table A.4). Its novelty and diversity on both frameworks are lower than those of SCORED in Table 1, with only $\sim$10% of generated OSDAs being unique. Moreover, across 3,000 OSDA candidates generated for 6 different frameworks, only 150 unique candidates are found, suggesting severe mode collapse. We therefore exclude it from further competition comparison.

Table A.4: Molecular-quality statistics for OSDA Generator at a budget of 500 generations per framework.

Framework	Valid. (%)↑	Unique (%)↑	Novelty↑	Diversity↑	SCScore↓
AFI	93.80	13.80	0.33	0.67	1.21
LTA	84.60	10.40	0.57	0.23	1.05

C.5 Six-Framework Selectivity Evaluation

Common physical evaluator. A candidate passes the physical-validity filter when both the requested target and at least one competing framework have binding energies within $-35 \leq E_{\mathrm{bind,Si}} \leq 0$ kJ/mol Si, following the energy range of prior frozen-pose evaluations (Schwalbe-Koda et al., 2021; Schwalbe-Koda & Gómez-Bombarelli, 2021a; Xie et al., 2025).

Table A.5: Single-run force-field competition scores on all six frameworks, in kJ/mol Si. Lower is better. Bold marks the lowest mean in each row. The final row is the unweighted six-framework mean.

Framework	OSDA Agent	Full SCORED	− denoising reward
AEI	**4.10**	4.25	7.21
AFI	11.44	**2.52**	5.11
AFX	6.25	**5.17**	7.75
ITE	14.60	**6.17**	8.44
LTA	11.29	**2.38**	2.46
MOR	10.61	**7.70**	8.49
Six-framework mean	9.72	**4.70**	6.58

D Theoretical Analysis of SCORED

D.1 Derivation of the Confidence–Margin Identity

Fix a corruption (t, ϵ), a nonempty competitor set of size m, and a temperature $\tau > 0$. Write $a_c = d_\theta(x, c; t, \epsilon)$ and $a_j = d_\theta(x, c'_j; t, \epsilon)$. From the normalized confidence in Equation (4),

$$\begin{aligned} \frac{s_\theta}{1 - s_\theta} &= \frac{e^{-a_c/\tau}}{\sum_{j=1}^{m} e^{-a_j/\tau}}, \\ \tau \log \frac{s_\theta}{1 - s_\theta} + \tau \log m &= -a_c - \tau \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{-a_j/\tau} \right) \\ &= \mathrm{SoftMin}_\tau(a_{1:m}) - a_c. \end{aligned} \tag{A.5}$$

This proves Equation (5) at each corruption. The centering term $\tau \log m$ makes the margin zero when the target and all competitor errors are equal. The identity follows from the chosen confidence normalization; it does not require denoising errors to be likelihoods or physical energies.

D.2 Aggregation Order and Estimation

The reward in Equation (7) applies SoftMin across competitors within each scoring evaluation, subtracts the target error, and then averages across timesteps. For a fixed candidate and competitor set, let ν denote the scoring timestep distribution. Its population counterpart is

$$\mathcal{R}^{\mathrm{pop}}_{\mathrm{select},\theta}(x, c) := \mathbb{E}_{t \sim \nu, \epsilon} \left[M_\theta(x, c; t, \epsilon) \right], \qquad \epsilon \sim \mathcal{N}(0, I). \tag{A.6}$$

Thus competitor aggregation is inside the expectation. Averaging each condition's error before applying SoftMin generally gives a different score because SoftMin is nonlinear.

For independent scoring corruptions drawn from this distribution and an integrable margin, Equation (7) is an unbiased sample mean and converges to Equation (A.6) as the number of draws increases. There is no nonlinear plug-in bias for this estimand: the quantity being averaged is the already-aggregated margin.

For a fixed schedule of K_t timesteps, such as the respaced scoring schedule, the corresponding expectation is instead

$$\frac{1}{K_t} \sum_{r=1}^{K_t} \mathbb{E}_\epsilon [M_\theta(x, c; t_r, \epsilon)].$$

One noise draw with the stated marginal distribution per scheduled timestep gives an unbiased estimate of this finite-schedule quantity. It equals Equation (A.6) when ν is uniform over the scheduled timesteps, not necessarily over all training timesteps. Changing the scoring schedule can therefore change the estimand as well as its sampling variability.

Shared corruptions pair the target and competitors within each evaluation. For an individual competitor, positive covariance between its error and the target error reduces the variance of their

difference relative to independent scoring with the same marginals. This is not an unconditional variance-reduction guarantee for the aggregated reward. The hard-min variant uses the same order of operations, replacing SoftMin with $\min_j$ within each evaluation before timestep averaging (Table A.6).

D.3 Competitor Sensitivity and Interpretation

At a fixed corruption, write $M = M_\theta(x, c; t, \epsilon)$ and use the errors a_c, a_j defined above. Differentiating the margin gives

$$\frac{\partial M}{\partial a_j} = \frac{e^{-a_j/\tau}}{\sum_{k=1}^{m} e^{-a_k/\tau}}, \qquad \frac{\partial M}{\partial a_c} = -1. \tag{A.7}$$

Lower-error competitors have greater influence at that timestep, and the most influential competitor can change across timesteps. These derivatives describe the scalar score; rewards are detached during the policy update.

By Equation (5), $M > 0$ precisely when $s_\theta > 1/(m+1)$. This does not require the target to have a lower error than every competitor: normalized SoftMin is a smooth reference, not an exact minimum. Likewise, a positive averaged reward does not imply a positive margin at every timestep. Equal errors across all conditions at every scoring evaluation do imply a zero averaged reward.

Writing $s_{\theta,r}$ for confidence at (t_r, ϵ_r), the implemented reward averages centered log-odds:

$$\widehat{\mathcal{R}}_{\text{select},\theta}(x, c; \Omega) = \frac{\tau}{K_t} \sum_{r=1}^{K_t} \log \frac{s_{\theta,r}}{1 - s_{\theta,r}} + \tau \log m.$$

Consequently, confidence and margin induce the same candidate ranking at a fixed corruption, but mean confidence and mean margin need not induce the same ranking. The reward is not generally the centered log-odds of mean confidence. Its interpretation remains relative to the competitor set, temperature, and scoring schedule, rather than a calibrated probability of physical framework selection.

E RWR Weighting and Empirical Score Alignment

The following analysis characterizes the offline weighting used to initialize our denoiser. For each molecule, we center its available binding energy labels across the observed training frameworks and use the negative centered cost as the RWR utility, without standard-deviation normalization. For notational convenience, this appendix uses c to index the framework associated with condition $c = e(y)$.

We first characterize the resulting weighted denoising risk and its population-optimal predictor, then analyze the properties of the molecule-centered binding utility. Finally, we assess the empirical rank association between denoising errors and centered binding costs. This diagnostic evaluates score alignment empirically; it is not a theoretical guarantee of cross-framework selectivity.

E.1 Weighted Denoising Risk and the Population Predictor

Let $q_0(x \mid c)$ and $p_0(c)$ denote the unweighted sampling distributions, and define $Z_\beta(c) = \mathbb{E}_{q_0(\cdot|c)}[e^{\beta R_{\text{off}}(x,c)}]$ and $q_\beta(x \mid c) = q_0(x \mid c) e^{\beta R_{\text{off}}(x,c)} / Z_\beta(c)$. For $0 < Z_\beta(c) < \infty$, direct substitution into Equation (9) gives

$$\begin{aligned} \mathcal{L}_{\text{RWR}}(\theta) &= \mathbb{E}_{c\sim p_0} \mathbb{E}_{x\sim q_0(\cdot|c)}[e^{\beta R_{\text{off}}(x,c)} D_\theta(x; c)] \\ &= \mathbb{E}_{c\sim p_0}\left[Z_\beta(c) \mathbb{E}_{x\sim q_\beta(\cdot|c)} D_\theta(x; c)\right]. \end{aligned} \tag{A.8}$$

Thus the effective joint distribution is proportional to $p_0(c) Z_\beta(c) q_\beta(x \mid c)$. The condition factors affect capacity allocation in a shared network; within-condition reweighting alone does not describe the entire objective.

Proposition E.1 (Training-distribution enrichment)**.** *For fixed c and $0 < Z_\beta(c) < \infty$, define*

$$q_\beta(x \mid c) = \frac{q_0(x \mid c) e^{\beta R_{\text{off}}(x,c)}}{Z_\beta(c)}, \qquad Z_\beta(c) = \mathbb{E}_{q_0(\cdot|c)} e^{\beta R_{\text{off}}(x,c)}. \tag{A.9}$$

RWR minimizes denoising risk under q_β, with condition weights $p_0(c)Z_\beta(c)$. Where differentiation is valid, $\frac{d}{d\beta}\mathbb{E}_{q_\beta}[R_{\text{off}}] = \text{Var}_{q_\beta}(R_{\text{off}}) \geq 0$.

Proof of Proposition E.1. Equation (A.8) proves the risk identity. On positive support, the relative odds satisfy

$$\frac{q_\beta(x_i \mid c)}{q_\beta(x_j \mid c)} = \frac{q_0(x_i \mid c)}{q_0(x_j \mid c)} e^{\beta[R_{\text{off}}(x_i,c) - R_{\text{off}}(x_j,c)]}. \tag{A.10}$$

For fixed labels and q_0, assume the first two exponentially weighted moments are finite in a neighborhood of β, allowing differentiation under the integral. Differentiating the normalized expectation yields

$$\frac{d}{d\beta}\mathbb{E}_{q_\beta}[R_{\text{off}}] = \frac{\mathbb{E}_{q_0}[R_{\text{off}}^2 e^{\beta R_{\text{off}}}]}{Z_\beta} - \left(\frac{\mathbb{E}_{q_0}[R_{\text{off}} e^{\beta R_{\text{off}}}]}{Z_\beta}\right)^2 = \text{Var}_{q_\beta}(R_{\text{off}}) \geq 0. \tag{A.11}$$

□

RWR therefore tilts the offline training distribution toward higher-utility examples within each framework, with expected utility increasing in β whenever its variance is nonzero. This binding-informed distribution defines the denoising risk optimized before online selectivity training. The corresponding population-optimal predictor is characterized below.

Proposition E.2 (Population-optimal noise predictor)**.** *For squared noise-prediction loss and an unrestricted predictor, the minimizer satisfies*

$$f^\star(z_t, t, c) = \mathbb{E}_{q_\beta}[\epsilon \mid z_t, t, c], \tag{A.12}$$

almost surely under the weighted training law, which includes the fixed encoder and forward corruption, with a condition-independent scoring law for (t, ϵ).

Proof. Conditioning on (z_t, t, c) decomposes the risk as

$$\mathbb{E}[\|\epsilon - f\|^2 \mid z_t, t, c] = \mathbb{E}[\|\epsilon - f^\star\|^2 \mid z_t, t, c] + \|f - f^\star\|^2. \tag{A.13}$$

The first term does not depend on f. A fixed positive normalization of the squared norm preserves the minimizer. □

This statement concerns the squared-error minimizer under the weighted training law. Finite capacity, optimization, and the reverse sampler still determine the generated outputs. It supplies no guarantee about their physical properties.

E.2 Mean-Centered Binding Utility

For molecule x, let $\mathcal{O}_x$ contain its available training-framework labels and let $m_x = |\mathcal{O}_x| > 0$. Rewrite $E(x, c) = E_{\text{off}}(x, c)$ for the `Binding/Si` label in kJ/mol Si and define

$$\mu(x) = \frac{1}{m_x} \sum_{j \in \mathcal{O}_x} E(x, j), \qquad \widetilde{E}(x, c) = E(x, c) - \mu(x), \qquad R_{\text{off}}(x, c) = -\widetilde{E}(x, c). \tag{A.14}$$

Expanding the mean gives

$$R_{\text{off}}(x, c) = \frac{1}{m_x} \sum_{j \in \mathcal{O}_x} [E(x, j) - E(x, c)]. \tag{A.15}$$

For $c \in \mathcal{O}_x$ and $m_x > 1$, this equals $(m_x - 1)/m_x$ times the mean gap to the other observed frameworks. If the target is the only available label, the reward is zero. The mean uses the availability mask without standard-deviation scaling; it need not represent all possible competing frameworks. Here μ, $\widetilde{E}$, and R_{off} retain units of kJ/mol Si, so β has units $(\text{kJ/mol Si})^{-1}$.

Proposition E.3 (Invariance to a molecule-only offset)**.** *For a fixed availability mask and $c \in \mathcal{O}_x$, adding the same $a(x)$ to all available energies leaves $R_{\text{off}}(x, c)$ and hence the RWR weights unchanged.*

Proof. Both $E(x, c)$ and $\mu(x)$ increase by $a(x)$, so their difference is unchanged. □

This invariance removes a common energy offset, not framework-dependent bias or bias in which labels are available.

E.3 Empirical Rank Alignment

We measure Kendall's τ_b of denoising loss against mean-centered binding cost $\widetilde{E}(x, c)$, with lower values preferred in both quantities. The reported average is 0.209 (used in Figure 4). This is a modest ranking association and it does not measure the cross-framework correlation between a denoising margin and an energy gap, nor establish calibration or its preservation during RL. We do not attribute the observed gains to strong score–property correlation; instead, we assess those gains directly through physical evaluation.

Held-out OSDA case study. We also examine LTA and MOR as two illustrative held-out frameworks. On a fixed candidate cohort, the RWR-initialized reward attains a mean Kendall correlation of $\tau_b = 0.283$ with the negative competition score (to make it higher the better) and $\tau_b = 0.306$ with the hard competition energy gap (defined as $\min_{y' \in \mathcal{Y}}(E_{\mathrm{bind,Si}}(x, y')) - E_{\mathrm{bind,Si}}(x, y)$ also to make it higher the better), compared with -0.028 and -0.055, respectively, for SFT. This positive ordering is preserved across all three Full SCORED checkpoints, with $\tau_b = 0.269$–0.276 for the competition score and $\tau_b = 0.239$–0.255 for the hard gap. These results show that binding-informed RWR can establish a physically relevant target–competitor denoising-score ordering on some unseen frameworks, and that subsequent online optimization can preserve this ordering despite updating the denoiser used for both generation and scoring.

F Reward Design and Sensitivity

F.1 Reward-Evaluation Cost

We compared SCORED with a binding-energy surrogate model (Xie et al., 2025) as a selectivity reward. The binding-energy surrogate takes the docked complex as input and outputs the predicted binding energy, so it requires a dedicated docking process (Xie et al., 2025). It is reported to have root-mean-square error (RMSE) 1.07 kJ/mol Si and Spearman correlation 0.96 against frozen-pose evaluation. To compute a competition score with this surrogate, we generate several conformers for each OSDA candidate and dock them to the target and competing frameworks. The resulting complexes are evaluated by the surrogate, and the competition score is computed from the predicted binding energies. We compare the wall-clock cost of this pipeline with SCORED in Table A.8 and report the performance of the model trained with this reward in Table A.6.

F.2 Reward Aggregation

Table A.6 compares SoftMin, hard-min, and an external surrogate-based selectivity reward. SoftMin offers a lower competition score and higher diversity than hard-min, while hard-min increases validity. Molecular weight analysis further characterizes the chemistry produced by each reward. The surrogate-reward arm drifts toward large, highly flexible molecules: its mean molecular weight is 424 Da, against 195 Da for SoftMin and 207 Da for hard-min, with 7.0 rotatable bonds on average, whereas prior OSDA design work favors fewer than five rotatable bonds (Ito et al., 2024; Hu et al., 2025). Such molecules are less favorable OSDA candidates because conformational rigidity is generally desirable for maintaining stable pore interactions during crystallization (Hu et al., 2025; Ito et al., 2024). This shows up in the physical evaluation. Of the $3{,}000$ molecules generated by this arm (500 candidates for each of the six frameworks), only six molecules produced a dockable pose, all of them in LTA, and each had $E_{\mathrm{bind,Si}} < -35$ kJ/mol Si, which indicates unphysical framework bonding. No candidate therefore passed the physical-validity filter of Appendix C.5, and we report no competition score for this arm. Candidates without a dockable pose receive only molecular quality rewards during training, so RL proceeds even when most generated molecules cannot be scored physically.

Table A.6: Six-framework reward-design diagnostic. Competition score (Comp.) is in kJ/mol Si. The surrogate-reward arm has no competition score because none of its candidates passed the physical-validity filter. MW is mean molecular weight in Da.

Reward	**Comp.↓**	**Valid.↑ (%)**	**Div.↑**	**MW (Da)**
Selectivity (SoftMin)	**4.70**	93.90	0.876	195
Selectivity (hard-min)	$7.94_{+3.24}$	**95.58**	0.820	207
Surrogate-based	–	94.03	**0.880**	424

F.3 Competitor-Set Size and Composition

This sweep in Table A.7 uses a separate configuration: 32 framework conditions per step sampled without replacement, 63 RL steps, and m competitors drawn per target. For each target, random sampling selects m competitors from the other 31 frameworks in the batch; $m = 31$ uses all of them. Nearest/farthest selection uses $m = 3$, ranked by cosine distance between CGCNN embeddings. The random $m = 3$ row is the reference for this controlled sweep and has the lowest competition score. The $m = 7$ and $m = 31$ settings improve molecular-quality metrics at the cost of higher competition scores. Nearest selection (6.52) slightly improves on farthest selection (6.72), but neither improves on random $m = 3$ (6.47). All resulting candidates use the same six-framework evaluation protocol; m changes only online reward scoring.

Table A.7: Controlled competitor-set sweep: 32 frameworks per batch, eight generated candidates per framework, and a common training-step budget below the main setting. Nearest/farthest use $m = 3$. Single-run, six-framework means; units follow Table 1. Red deltas are relative to random $m = 3$. H anchors remain fixed (Appendix G.2); bold marks the best value within this sweep.

Setting	Comp.↓	Valid.↑	Nov.↑	Div.↑	SCScore↓	H_{sel} ↑	H_{eq} ↑
random ($m = 1$)	$8.92_{+2.45}$	90.80	0.710	0.820	2.620	$0.669_{-0.148}$	$0.754_{-0.090}$
random ($m = 3$, reference)	**6.47**	92.60	0.730	0.870	2.450	0.818	0.844
random ($m = 7$)	$7.01_{+0.54}$	93.60	0.816	**0.937**	1.799	$\mathbf{0.832}_{+0.014}$	$\mathbf{0.920}_{+0.076}$
random ($m = 15$)	$7.66_{+1.19}$	**99.45**	0.720	0.780	1.760	$0.771_{-0.047}$	$0.862_{+0.017}$
random ($m = 31$)	$8.97_{+2.50}$	93.53	**0.817**	0.936	**1.741**	$0.718_{-0.100}$	$0.864_{+0.019}$
nearest ($m = 3$)	$6.52_{+0.05}$	86.80	0.600	0.910	1.910	$0.821_{+0.003}$	$0.855_{+0.010}$
farthest ($m = 3$)	$6.72_{+0.25}$	83.80	0.630	0.870	2.080	$0.798_{-0.020}$	$0.831_{-0.014}$

F.4 Shared-Noise Variance Diagnostic

When the target and competitor conditional scores share the same diffusion timestep t and noise ϵ inside each Monte Carlo sample, common random components can cancel in the difference. We measure this effect on the OSDA DiT backbone using target AFI, five competitors (LTA/AEI/AFX/ITE/MOR), four generated SMILES, eight Monte Carlo (MC) repeats per cell, and $t \in \{50, 350, 650, 950\}$. For each (candidate, t) cell we compute the empirical standard deviation of the selectivity reward estimator under two regimes. In SHARED, the same `torch` random number generator (RNG) seed is used for the target and all off-targets within an MC sample. In INDEP, each condition draws fresh noise. Across all 16 (candidate, t) cells, shared noise yields lower estimator standard deviation. The geometric-mean and median ratios $\sigma_{\text{indep}}/\sigma_{\text{shared}}$ are $56.3\times$ and $50.5\times$, respectively.

F.5 Scoring Draws and Temperature

For each candidate, K_t is the number of scored diffusion timesteps with one shared noise per timestep across the target and competitor conditions. At each timestep, we apply SoftMin to the competitor errors, subtract the target error, and then average the resulting margins over the K_t timesteps. The temperature τ controls the competitor aggregation. The default $(K_t, \tau) = (10, 1)$ has the lowest competition score among the reported settings (Table A.9). Decreasing K_t to four raises validity by 2.07 percentage points and diversity by 0.037, with a competition score increase of 1.83 kJ/mol Si. At $K_t = 10$, increasing τ from 0.1 to 10 reduces the competition score by 0.19 while slightly reducing

Table A.8: Per-sample reward-path timing on one A100, excluding policy optimization. Relative cost is normalized to a six-condition, $K_t = 4$ SCORED panel (24 denoiser forwards). Main training uses four conditions and $K_t = 10$. The pose-based multilayer perceptron (MLP) pipeline includes conformer generation, docking and MLP forwarding (Xie et al., 2025).

Reward path	Time	Relative cost
DiT forward (one condition/noise)	0.327 ms	0.04×
SCORED panel (6 conditions, $K_t = 4$)	7.86 ms	1×
MLP network forward only	$\sim$ 1 ms	0.13×
Conformer + dock + MLP	$\sim$ 8.3 s	$\sim$ 1056×

Table A.9: Ablation on K_t (number of scored timesteps with one shared corruption per timestep) and τ (SoftMin temperature).

Setting	K_t	τ	**Comp.↓**	**Valid. (%)↑**	**Nov.↑**	**Div.↑**
Main	10	1	4.70	93.90	0.702	0.876
Diagnostic	2	1	$6.24_{+1.54}$	97.27	0.746	0.898
	4	1	$6.53_{+1.83}$	95.97	0.763	0.913
	10	0.1	$6.83_{+2.13}$	96.30	0.743	0.894
	10	10	$6.64_{+1.94}$	96.83	0.734	0.884

novelty and diversity. Using the measured 0.327 ms per condition-noise forward, a four-condition scoring panel costs approximately 13.08 ms per candidate at $K_t = 10$ and 5.23 ms at $K_t = 4$.

G Molecular Diversity and Selectivity–Quality Trade-offs

G.1 Descriptor Coverage

Figures A.2 and A.3 compare generated OSDAs with literature examples and the full dataset on AFI and LTA. These projections describe molecular coverage alongside the scalar diversity metrics.

G.2 Harmonic Trade-off Analysis

We use fixed-reference utilities: $u = v/v^\star$ for validity, novelty, and diversity, and $u = v^\star/v$ for competition score and SCScore. AFI and LTA each use the best values among the four methods in Table 1. All six-framework tables share the fixed anchors Comp. 5.026, Valid. 95.61%, Nov. 0.764, Div. 0.920, and SCScore 1.637, each the best value for that metric among the repeated runs reported in Table A.2. A utility above one denotes a result beyond its fixed reference; scores are not clipped. Define

$$Q = \frac{4}{u_{\text{valid}}^{-1} + u_{\text{nov}}^{-1} + u_{\text{div}}^{-1} + u_{\text{synth}}^{-1}}, \qquad H_{\text{sel}} = \frac{2}{u_{\text{comp}}^{-1} + Q^{-1}}. \tag{A.16}$$

Q is the harmonic mean of the four molecular-quality utilities. The selectivity-weighted H_{sel} assigns 50% weight to competition score and 12.5% to each quality metric. The complexity utility u_{synth} is based on SCScore, rather than measured synthesis success.

Using the weighted family in Equation (12), the selectivity-weighted summary is $H_{\text{sel}} = H(1/2)$. The metric-equal summary assigns 20% weight to each of the five metrics:

$$H_{\text{eq}} = H(1/5) = \frac{5}{u_{\text{comp}}^{-1} + u_{\text{valid}}^{-1} + u_{\text{nov}}^{-1} + u_{\text{div}}^{-1} + u_{\text{synth}}^{-1}}. \tag{A.17}$$

The harmonic mean penalizes low utilities through their reciprocal contributions. H scores and weight curves use method-level mean metrics, rather than averaging per-run H scores; no H standard deviation is reported. Harmonic scores and their deltas are rounded only after calculation.

SCORED ranks first under both weighting conventions on AFI and LTA. In the repeated-run six-framework ablation, Full SCORED also leads in both H_{sel} and H_{eq}. − competitor contrast has the

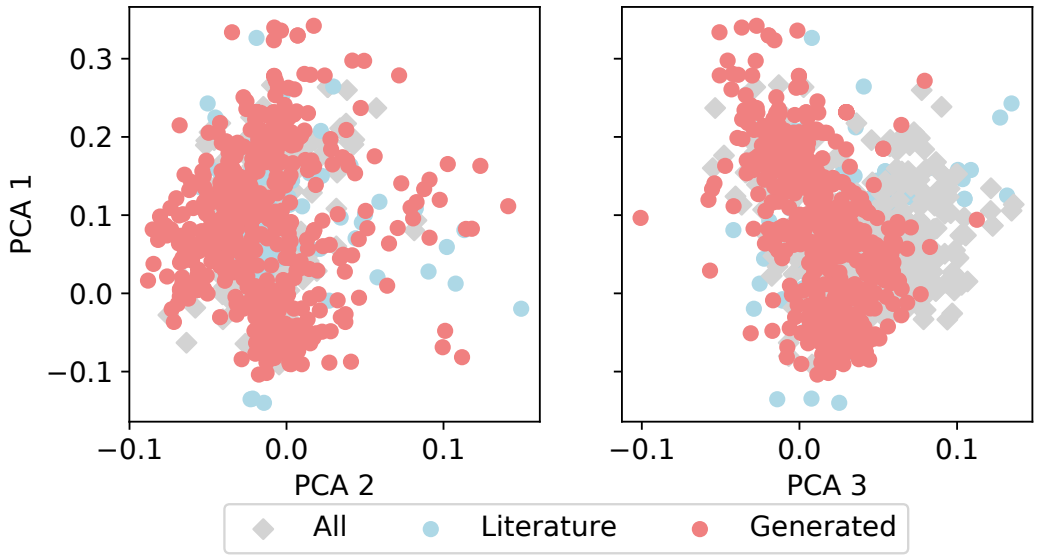


Figure A.2: Descriptor coverage on AFI. Principal-component analysis (PCA) of weighted holistic invariant molecular (WHIM) descriptors compares generated OSDAs with literature examples and the full dataset in two projections. Generated structures occupy both shared and additional descriptor regions.

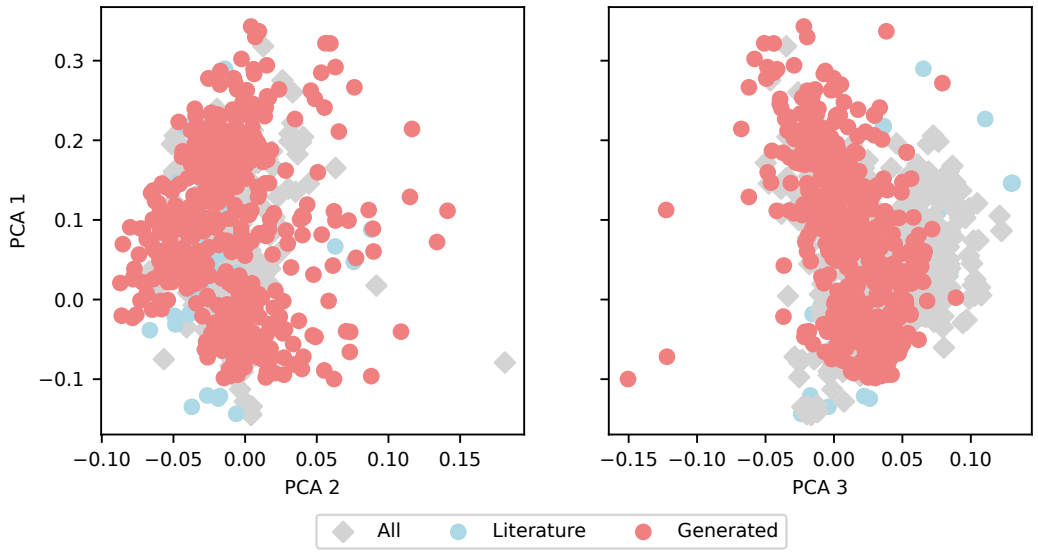


Figure A.3: PCA of WHIM descriptors on LTA. In both displayed projections, generated OSDAs overlap the literature set and extend beyond its observed range.

higher quality-only aggregate Q, whereas Full SCORED achieves greater selectivity. Figure A.4 makes this preference-dependent trade-off explicit without changing the inputs or training.

Table A.10: Trade-off summaries from Tables 1 and 2. Q summarizes molecular quality. H_{sel} weights competition score at 50% and each quality metric at 12.5%; H_{eq} weights all five metrics at 20% each. Bold marks the best value within each comparison.

Comparison	Method	u_{comp}	Q	H_{sel}	H_{eq}
AFI	OSDA Agent	0.220	0.929	$0.356_{-0.595}$	$0.565_{-0.359}$
	SFT	0.287	0.882	$0.434_{-0.518}$	$0.624_{-0.301}$
	RWR	0.301	0.837	$0.443_{-0.508}$	$0.617_{-0.307}$
	SCORED	1.000	0.907	**0.951**	**0.924**
LTA	OSDA Agent	0.211	0.948	$0.345_{-0.597}$	$0.558_{-0.353}$
	SFT	0.457	0.885	$0.602_{-0.340}$	$0.745_{-0.166}$
	RWR	0.523	0.816	$0.637_{-0.305}$	$0.734_{-0.177}$
	SCORED	1.000	0.891	**0.942**	**0.911**
Six-framework	Full SCORED	1.000	0.931	**0.964**	**0.944**
	− competitor contrast	0.793	0.974	$0.874_{-0.090}$	$0.932_{-0.012}$
	− denoising reward	0.773	0.926	$0.843_{-0.122}$	$0.891_{-0.053}$

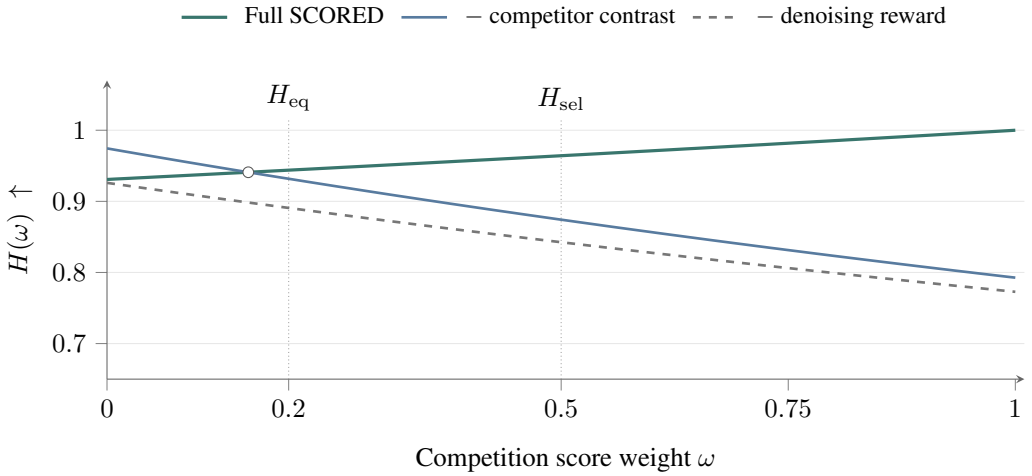


Figure A.4: Sensitivity to the competition score weight ω using the repeated-run means. Full SCORED exceeds − competitor contrast above $\omega \approx 0.156$. Both the selectivity-weighted H_{sel} ($\omega = 0.5$) and metric-equal H_{eq} ($\omega = 0.2$) favor Full SCORED.

H Cross-Domain Generalization Experiments

H.1 Evaluation and Training Setup

Dataset and split. We use CrossDocked2020 (Francoeur et al., 2020) preprocessed by Schneuing et al. (2024) (10 Å pocket extraction, atom-level representations). We construct a 30% sequence-identity cluster split via MMseqs2 yielding 1,947 train / 516 test clusters. The split is locked across all SBDD experiments in Section 5.4.

RL setup for TargetDiff. We use a tree-branched sampler with branching factor $K{=}2$ and depth $L{=}2$, giving four leaves per pocket. Among the 1,000 diffusion steps, the last 50 steps are branched and tracked for transition log-probabilities. The per-step PPO loss uses a clip range of 0.2. Gradients are accumulated across branched transitions. We use the AdamW optimizer with learning rate 10^{-5}, weight decay 10^{-10}, and gradient norm clip 1.0. We use three off-target pockets per training step, chosen once at the start from independent clusters in the train split (`TPIS_MYCTU_1_261_0`, `CSF1R_HUMAN_539_921_0`, `TRPF_THEMA_1_205_0`). The matched arms within each experiment use the same budget.

Docking protocol. We use AutoDock Vina v1.2.5 with exhaustiveness 8, a $20 \times 20 \times 20$ Å box centered on the reference ligand centroid for each pocket, and 10 poses per docking. The most negative pose score gives the docking score $V(x, c)$ of ligand x in pocket c. For the requested target c_{tgt}, the selectivity gap is $\Delta\mathrm{Vina}_{\mathrm{gap}} = \min_{c \in \mathcal{C}_{\mathrm{off}}} V(x, c) - V(x, c_{\mathrm{tgt}})$. Positive values indicate stronger predicted binding to the target than to any off-target in the panel.

H.2 Matched Multi-Pocket Evaluation

Using the TargetDiff setup above, we train one policy per arm for 300 outer steps with 16 inner epochs. The full arm uses SCORED, while the $-$SCORED arm removes the selectivity reward completely. Training rotates over 30 pockets, and evaluation uses 30 disjoint pockets. For each evaluation pocket, both arms share a panel of three off-target pockets. We sample 10 ligands per pocket and arm and dock each ligand against the target and the shared off-target panel. Tables A.11 and A.12 report the aggregate results and per-pocket breakdown, respectively.

Table A.11: Matched multi-pocket evaluation on the 30-pocket evaluation set. Scores and differences are in kcal/mol, and paired Δ is $+$SCORED minus $-$SCORED. Top: selectivity gap $\Delta\mathrm{Vina}_{\mathrm{gap}}$ (higher is better). The favorable-pocket count records pockets whose mean favors $+$SCORED in the metric's indicated direction. The 95% CI uses 5,000 paired bootstrap resamples of the 30 pocket-level mean differences. Bottom: target docking score V_{tgt} (more negative indicates stronger predicted binding).

Metric	$-$SCORED	$+$SCORED	Paired Δ	95% CI	Pockets fav. $+$SCORED
$\Delta\mathrm{Vina}_{\mathrm{gap}}$ ($\uparrow$)	2.21	2.42	**0.21**	[0.01, 0.40]	**23/30**
V_{tgt} ($\downarrow$, tighter)	-6.73	-6.97	-0.24	–	19/30

H.3 Cross-Backbone Evaluations

Table A.13 compares SCORED and baseline policies across three SBDD backbones, including molecular quality alongside selectivity.

Table A.12: Per-pocket mean $\Delta\mathrm{Vina}_{\mathrm{gap}}$ (kcal/mol) for the matched multi-pocket policy on the 30-pocket evaluation set. Differences are computed before rounding. Each arm generates 10 ligands per pocket. Rows are sorted in descending order by Δ, the $+$SCORED minus $-$SCORED difference in mean gap.

Target cluster	$-$SCORED	$+$SCORED	Δ	Target cluster	$-$SCORED	$+$SCORED	Δ
ERBB2_HUMAN_695_1029	4.50	6.05	$+1.55$	FOLB_MYCTU_1_121	1.50	1.61	$+0.11$
PVDQ_PSEAE_26_217	3.55	4.86	$+1.30$	PRDX5_HUMAN_53_214	0.22	0.33	$+0.10$
ODP1_ECOLI_2_887	2.02	3.00	$+0.99$	KGD_MYCS2_360_1227	3.23	3.34	$+0.10$
HS90A_HUMAN_7_233	2.71	3.63	$+0.92$	PRFA_LISMO_1_237	2.42	2.52	$+0.10$
MAP12_MYCTU_1_285	2.79	3.38	$+0.59$	ACRB_ECOLI_1_1048_amp	-0.21	-0.17	$+0.04$
OGA_HUMAN_58_704	2.12	2.69	$+0.57$	SYQ_ECOLI_2_552	3.06	3.07	$+0.01$
NOS2_HUMAN_78_505	1.33	1.86	$+0.53$	ENO_CHLAA_1_426	1.89	1.66	-0.24
KKA3_ENTFL_1_264	1.45	1.98	$+0.53$	CARP_CRYPA_90_419_cat	1.70	1.46	-0.25
FNTB_HUMAN_8_432	3.48	3.79	$+0.31$	RON_HUMAN_1060_1357_ATP	2.30	2.03	-0.26
CBH_CLOPE_1_329	3.42	3.70	$+0.28$	TETR8_PASMD_2_207	3.76	3.45	-0.31
METE_CANAL_1_767	2.37	2.66	$+0.28$	BRD1_HUMAN_924_1049	2.20	1.69	-0.51
MURI_HELPJ_1_255_allosteric2	3.48	3.75	$+0.27$	CPT2_RAT_28_658	2.03	1.27	-0.77
THM1_THADA_1_207	-0.18	0.06	$+0.23$	PA21B_BOVIN_23_145	2.41	1.49	-0.93
FPPS_HUMAN_69_419_cat	2.70	2.91	$+0.21$				
HMUO_CORDI_1_215	2.36	2.55	$+0.19$				
P53_HUMAN_94_306	0.76	0.92	$+0.16$				
PLMN_HUMAN_101_183	0.87	0.99	$+0.13$				

Table A.13: SBDD policy comparisons on 30 evaluation pockets with 10 ligands per pocket and arm. Validity, quantitative estimate of drug-likeness (QED), and synthetic accessibility (SA) are reported as $+$SCORED / $-$SCORED. Bold marks the better arm. Gap change is the paired mean difference in $\Delta\mathrm{Vina}_{\mathrm{gap}}$ between $+$SCORED and $-$SCORED, in kcal/mol.

Backbone	Valid.%↑	QED↑	SA↓	Gap change↑
TargetDiff (Guan et al., 2023)	$\mathbf{90.0}/89.0$	$\mathbf{0.474}/0.473$	$4.63/\mathbf{4.58}$	$\mathbf{+0.21}$
AliDiff (Gu et al., 2024)	$\mathbf{85.3}/84.7$	$\mathbf{0.498}/0.493$	$\mathbf{4.60}/4.73$	$+0.16$
IPDiff (Huang et al., 2024)	$\mathbf{87.0}/86.0$	$0.486/\mathbf{0.504}$	$\mathbf{4.50}/4.61$	-0.02

H.4 MOFFUSION-H_2 Matched Reward Toggle

We fine-tune MOFFUSION-H_2 for 100 stochastic-DDIM RL steps. The matched arms use the same 50-step sampler and evaluators and generate 80 outputs each. The $+$SCORED arm retains the SCORED selectivity reward and generic quality rewards, whereas $-$SCORED removes the selectivity reward. Neither arm uses the H_2 surrogate reward.

Table A.14: Matched MOFFUSION-H_2 reward toggle with 80 outputs generated per arm. H_2 is surrogate-predicted working capacity in g/L. Top-5 is the mean capacity of the five highest-scoring outputs. CIF-built is the crystallographic information file (CIF) construction rate, and Diversity is computed over generated structures. Both arms omit the H_2 reward.

Arm	Mean H_2 ↑	Max H_2 ↑	Top-5 H_2 ↑	CIF-built (%)↑	Diversity↑
$-$SCORED	32.03	37.26	36.76	$\mathbf{97.50}$	0.861
$+$SCORED	$\mathbf{34.28}$	$\mathbf{44.19}$	$\mathbf{42.34}$	93.75	$\mathbf{0.883}$

H.5 Cross-Domain Alignment and Standardized Effects

The diagnostic separates two questions: whether pre-RL scores rank the desired property, and whether adding SCORED improves downstream outcomes. For pre-RL alignment, we orient both variables so that larger values are favorable and compute Kendall's τ_b between negative denoising error and the corresponding property: mean-centered binding utility for OSDA, experimental $\mathrm{p}K_d$ for SBDD (which is the negative base-ten logarithm of the molar dissociation constant), and surrogate-predicted H_2 working capacity for MOF. For OSDA, we report the macro-average of within-framework correlations.

To compare improvements in different physical units, orient each evaluation outcome Y so that larger is better: negative competition score in kJ/mol Si for OSDA, Vina selectivity gap in kcal/mol for SBDD, and H_2 working capacity in g/L predicted by a surrogate model following Park et al. (2024) for MOF. For n paired evaluation units, define $\Delta_i = Y_{+\mathrm{SCORED},i} - Y_{-\mathrm{SCORED},i}$ and the corrected paired effect

$$g_z = \left(1 - \frac{3}{4(n-1)-1}\right)\frac{\overline{\Delta}}{s_\Delta}, \qquad s_\Delta = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta_i - \overline{\Delta})^2}. \tag{A.18}$$

Standardization removes physical units, while s_Δ retains the variability of the domain's pairing unit. We pair 6 framework means for OSDA, 30 pocket means for SBDD, and generation-batch means for MOF. The latter covers 80 generated outputs per arm in 10 paired batches. For OSDA, each paired evaluation unit is a framework: we use the three-seed average of the $-$ denoising reward and Full SCORED arms, then compute the six framework-level differences. Details are provided in Table A.15.

Table A.15: Inputs to the cross-setting diagnostic in Figure 4. Alignment is measured before RL. Outcome means are shown in their original metric orientation, whereas Δ and g_z are oriented so that positive values favor SCORED. The paired effect g_z is computed using Equation (A.18).

Setting	τ_b	Without	With	Δ	Paired units	g_z
OSDA	0.2090	6.5035	5.0260	1.4775	6 frameworks	1.4388
TargetDiff	-0.0628	2.2086	2.4165	0.2080	30 pockets	0.3766
AliDiff	0.1257	2.4086	2.5702	0.1616	30 pockets	0.2474
IPDiff	-0.0212	2.5494	2.5259	-0.0235	30 pockets	-0.0422
MOFFUSION	0.2236	32.0337	34.2817	2.2480	10 batches	1.2485