CONTROLLING PROPERTY TRADE-OFFS IN REWARD-ALIGNMENT OF 3D MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based drug design requires balancing predicted binding affinity, synthetic accessibility, and molecular validity. However, models trained to reproduce the distribution of training ligands do not necessarily satisfy these objectives jointly. We adapt Flow-GRPO to a pretrained, pocket-conditioned 3D ligand generative model to jointly optimize multiple molecular rewards such as Vina docking and synthetic accessibility scores. Optimizing either objective in isolation improves the target metric at the expense of the other. By varying their respective reward weights, we identify the trade-offs between these conflicting objectives and relate them to the composition, topology, and flexibility of generated ligands. We find that, depending on the reward composition, the aligned model learns to balance these ligand descriptors to maximize the reward. This insight allows us to control the aligned models' trade-off between Vina docking score and synthetic accessibility, with multiple configurations improving both over the baseline model, while also increasing drug-likeness and maintaining PoseBusters validity. Our results connect the trade-offs between objectives to interpretable changes in generated ligands' chemistry, providing a basis for understanding and selecting useful compromises across protein targets.

1 INTRODUCTION

Designing de novo molecules that exhibit drug-like properties has the potential to transform the field of drug development. Structure-based drug design (SBDD) offers a powerful strategy for accelerating drug discovery. By using the three-dimensional structure of a protein binding pocket to propose ligands, it enables the in-silico design of small molecules that bind to protein pockets of interest. Recently, diffusion (Guan et al., 2023; Schneuing et al., 2024) and flow-matching (Schneuing et al., 2025; Cremer et al., 2025a;b) models have made substantial progress toward the goal of generating pocket-conditioned 3D ligands. However, these models primarily learn to reproduce the training data distributions, whereas practical molecule design requires optimizing properties such as binding affinity, synthetic accessibility (SA) (Ertl & Schuffenhauer, 2009), quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012) and physical plausibility (Buttenschoen et al., 2024).

To address this issue and steer the generation toward ligands with desired properties, inference-time and reward- or preference-based post-training methods have been proposed. Among these approaches, online reinforcement learning (RL) offers a direct way to optimize molecular properties using feedback from newly generated samples. Two considerations are particularly important in this setting: Firstly, some generated samples cannot be decoded into chemically valid molecules and therefore cannot be evaluated by molecular-property oracles. How these invalid rollouts enter the training objective may affect both optimization stability and the yield of valid ligands. Improvements in scores among valid outputs must therefore be assessed alongside changes in validity. Secondly, some molecular properties interact and compete with each other, hence optimising for a certain objective can degrade other features of the generated samples. Prior work has observed reduced SA when optimizing docking scores (Cremer et al., 2024) and reduced molecular validity following preference alignment (Schneuing et al., 2025). Multi-objective alignment therefore requires balancing competing properties and assessing its consequences for molecular quality. However, changes in objective values alone do not reveal how alignment alters the generated molecules. Molecular features such as size, ring topology, and flexibility can help interpret the observed trade-offs (Cremer et al.,

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.

2024). This motivates a systematic analysis of the interaction between reward compositions, objective trade-offs and ligand structure and physicochemical properties during online post-training.

In this work, we adapt Flow-GRPO (Liu et al., 2026), an online post-training approach based on group relative policy optimization (GRPO) (Shao et al., 2024), to FLOWR.ROOT (Cremer et al., 2025b), a state-of-the-art flow matching model for pocket-conditioned 3D molecular generation. For each pocket, our model samples a group of mixed continuous-discrete trajectories, computes rewards on the final molecules, and updates the policy according to relative performance within a group. This provides a common framework for investigating invalid-rollout handling, varying the balance between Vina and SA rewards, and characterizing the resulting changes in generated ligands (Figure 1). We evaluate each reward composition across three training seeds on SPINDR and assess transfer to CrossDocked2020 without additional fine-tuning or checkpoint selection.

Our key contributions are as follows:

- We extend Flow-GRPO to online post-training of pocket-conditioned 3D molecular flow-matching models and show that penalizing invalid rollouts preserves the pretrained model's valid-molecule yield during alignment.
- We characterize the Vina–SA trade-off during alignment and demonstrate that the resulting compromises can be controlled by varying reward weights, with multiple model configurations improving both objectives over the pretrained baseline, while also improving QED and maintaining PoseBusters validity.
- We quantify changes in ligand topology, flexibility, and physicochemical properties across reward compositions and relate these changes to docking and synthetic accessibility scores, providing a better understanding of the observed trade-offs.

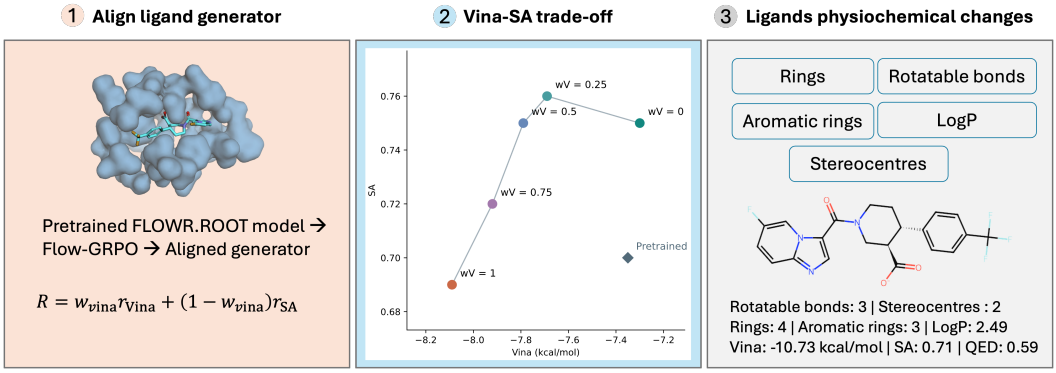


Figure 1: **Overview of our proposed Flow-GRPO based fine-tuning and evaluation workflow. Left:** We adapted Flow-GRPO to post-train FLOWR.ROOT, a pretrained pocket conditioned 3D ligand generator, using weighted combinations of Vina and synthetic accessibility (SA) rewards (r), with w_{Vina} controlling their relative emphasis. **Middle:** Varying the reward weights explores trade-offs between Vina docking scores and SA relative to the pretrained baseline. The w_V indicates the Vina reward weight (w_{Vina}), and lower Vina and higher SA are preferred. **Right:** Generated ligands were characterized by their physicochemical properties to understand how changes in ligand chemistry accompany the observed Vina–SA trade-offs.

2 Related Work

2.1 Pocket-conditioned 3D molecular generation

Generative SBDD models sample ligands conditioned on the three-dimensional structure of a protein pocket. A central challenge is coupling continuous atomic coordinates with categorical atom identities and, when modeled explicitly, bond types and formal charges. These variables must jointly describe a chemically valid molecule with a physically plausible geometry and a compatible placement

within the pocket. Pocket2Mol (Peng et al., 2022) generates molecules autoregressively, while TargetDiff (Guan et al., 2023) and DiffSBDD (Schneuing et al., 2024) use equivariant diffusion to generate coordinates and atom types by learning to reverse a prescribed noising process. More recently, models such as DrugFlow (Schneuing et al., 2025), MolFORM (Zhang & Zhang, 2025), and FlowR (Cremer et al., 2025a) use flow-matching-based approaches to directly learn transport dynamics along chosen probability paths from a simple prior to the data distribution (Lipman et al., 2022). FLOWR.ROOT (Cremer et al., 2025b) improves on FlowR by large-scale pretraining and an added affinity prediction head, providing a strong pretrained baseline for our study.

2.2 Inference-time guidance and post-training of molecular generators

Inference-time guidance. Inference-time guidance methods steer the sampling toward high-reward regions without modifying the parameters of the generator model. Gradient-based property guidance, including classifier guidance (Dhariwal & Nichol, 2021), modifies sampling dynamics using gradients of a classifier, property predictor, or energy function. MolJO (Qiu et al., 2025) applies gradient guidance within a differentiable Bayesian-flow representation to pocket-conditioned molecular generation. Classifier-free guidance (Ho & Salimans, 2021) avoids the need to train a separate classifier or property predictor by combining the generator's predictions with and without conditioning on the desired property. PILOT (Cremer et al., 2024) uses importance-weighted resampling by scoring intermediate states and resampling trajectories toward docking and SA objectives.

Preference-based fine-tuning. Offline preference-based methods update generator parameters using comparisons between ranked candidates in a fixed dataset. Building on direct preference optimization (DPO) (Rafailov et al., 2024) and DiffusionDPO (Wallace et al., 2024), molecular variants align diffusion and flow models for pocket-conditioned ligand generation using preference datasets constructed from computationally scored molecular candidates (Gu et al., 2024; Cheng et al., 2024; Zhang & Zhang, 2025; Schneuing et al., 2025; Lee et al., 2025).

Online reward-based fine-tuning. Online methods repeatedly collect samples from the evolving generator and evaluate their rewards to guide subsequent updates. Denoising diffusion policy optimization (DDPO) (Black et al., 2023) formulates denoising as a sequential decision-making process and applies policy gradients using rewards assigned to completed samples. Group relative policy optimization (GRPO) (Shao et al., 2024) estimates advantages from rewards within groups of samples. Flow-GRPO (Liu et al., 2026) adapts this approach to flow-matching models by introducing stochastic sampling. Graph-GRPO (Zhu et al., 2026) applies GRPO to discrete graph flows for molecular optimization using analytical transition probabilities. Additional molecular applications include online reward-based fine-tuning of diffusion (Zhou et al., 2025; Zhang et al., 2026) and flow-matching (Wang et al., 2025) models. In this work, we adapt Flow-GRPO to post-train FLOWR.ROOT using multiple molecular objectives.

2.3 Structural and physicochemical interpretation of molecular trade-offs

Previous work has already investigated certain aspects of structural and physicochemical properties of ligands during alignment. FLOWR.ROOT (Cremer et al., 2025b) reports the number of rotatable bonds and aromatic rings of its generated ligands, while DrugFlow (Schneuing et al., 2025) analyzes the distribution of bond and atom types as well as ring systems of different ligand generative models. PILOT (Cremer et al., 2024) relates ligand size and number of rings and rotatable bonds to docking scores, SA, and QED.

These studies help interpret molecular scores, but how ligand chemistry changes as the balance between competing rewards is varied during online post-training remains less well characterized. We investigate this question through a controlled Vina–SA reward-weight sweep, relating shifts in topology, flexibility, and physicochemical properties to the observed objective trade-offs.

3 METHOD

3.1 FLOW-GRPO FOR MOLECULAR ALIGNMENT

In this work, we consider a flow-based model for pocket-conditioned 3D molecular generation (*e.g.* Cremer et al. (2025b)) as the base model. Following Flow-GRPO (Liu et al., 2026), we formulate the reverse process of the base model as a finite-horizon Markov Decision Process (MDP). More specifically, the ligand state at flow time t is $z_t = (\mathbf{x}_t, \mathbf{a}_t, \mathbf{q}_t, \mathbf{b}_t)$, where $\mathbf{x}_t$ represents the continuous atomic coordinates and $\mathbf{a}_t$, $\mathbf{q}_t$, and $\mathbf{b}_t$ represent the categorical atom types, formal charges, and bond types, respectively. We denote by π_θ, π_{old} and π_{ref} the current model policy, the rollout policy and the frozen pretrained policy, respectively.

Given a pocket (P), rollout trajectories $\{\tau_i\}_i = \{z_{i,t}\}_{i,t} \sim \pi_{\text{old}}(\cdot|P)$ and the corresponding terminal rewards $\{R_i\}_i$, the Flow-GRPO objective is given by:

$$\mathcal{J}_{\text{GRPO}}(\{\tau_i\}_i, \theta, P) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathcal{L}_{i,t}^{\text{clip}}(P) - \beta D_{\text{KL}}\big(\pi_\theta(\cdot \mid z_{i,t}, P) \,\|\, \pi_{\text{ref}}(\cdot \mid z_{i,t}, P)\big)\Big) \quad (1)$$

where the clipped loss $\mathcal{L}_{i,t}^{\text{clip}}$ is the clipped weighted advantage:

$$\mathcal{L}_{i,t}^{\text{clip}}(P) = \min\left[\frac{\pi_\theta(z_{i,t+1} \mid z_{i,t}, P)}{\pi_{\text{old}}(z_{i,t+1} \mid z_{i,t}, P)}A_i,\ \text{clip}\big(\frac{\pi_\theta(z_{i,t+1} \mid z_{i,t}, P)}{\pi_{\text{old}}(z_{i,t+1} \mid z_{i,t}, P)}, 1-\epsilon, 1+\epsilon\big)A_i\right] \quad (2)$$

with ϵ the PPO clipping threshold and $A_i = \frac{R_i - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}}}$ the group-relative advantage for the given pocket P. Because the molecular-property oracles evaluate only the final decoded molecule, the same advantage is assigned to all T transitions in trajectory i. β in equation 1 is a hyperparameter that controls the regularization toward the pretrained policy.

Algorithm 1 describes the Flow-GRPO based fine-tuning algorithm. Following Zhou et al. (2025), we apply a size-invariant reduction so that likelihood magnitudes do not increase solely with molecular size. See Appendices A and B for more details.

3.2 MOLECULAR REWARDS AND INVALID ROLLOUTS

We consider the following molecular scores: Uni-Dock Vina score (Yu et al., 2023) as a proxy for binding affinity, Synthetic Accessibility (SA), and Quantitative Estimate of Drug-likeness (QED).

We map the raw scores to the interval $[0, 1]$, with higher values corresponding to better scores. We transform Vina so that it saturates at scores of -15 and 0 kcal/mol using $r_{\text{Vina}} = \text{clip}\left(-\text{Vina}/15, 0, 1\right)$, with thresholds selected empirically. SA is also normalized so that larger values indicate easier synthesis: $r_{\text{SA}} = \text{clip}\left((10 - s_{\text{SA}})/9, 0, 1\right)$. QED is already between 0 and 1 with higher values corresponding to better scores: $r_{\text{QED}} = \text{QED}$. The final reward ($R$) is the weighted sum of the individual components:

$$R = w_{\text{Vina}} r_{\text{Vina}} + w_{\text{QED}} r_{\text{QED}} + w_{\text{SA}} r_{\text{SA}} \quad (3)$$

where w_i is the corresponding weight. See Appendix C for the different reward weight hyperparameter choices of our experiments.

Because some samples do not result in valid molecules (*e.g.* invalid RDKit atom, charge, or bond assignments), no reward can be assigned to the sampling trajectories. During fine-tuning, we investigated two ways of handling invalid molecules. The first approach is simple masking of the invalid trajectories in both the group-relative advantage and the loss. Another approach is hard-assignment to the group-relative minimal reward given by:

$$R_{\text{invalid}} = \min\left(0, \min_{j \in \mathcal{V}_g} R_j\right) \quad (4)$$

where $\mathcal{V}_{\mathcal{G}}$ denotes the valid rollouts in group $\mathcal{G}$ and R_j denotes the reward of the jth valid sample. Note that when masking a trajectory, the effective group size G in equation 1 is reduced accordingly.

Algorithm 1 Flow-GRPO with invalid-molecule penalization

Require: Pretrained policy π_{θ_0}; pocket distribution $\mathcal{P}$; reward function R; group size G; sampling steps T; rollout iterations N; replay passes K; clipping threshold ϵ; KL weight β; reward stabilizer ε_R.

Initialize $\theta \leftarrow \theta_0$ and frozen reference $\pi_{\text{ref}} \leftarrow \pi_{\theta_0}$
for $n = 1, \ldots, N$ **do**
  // Rollout collection
  Set $\pi_{\text{old}} \leftarrow \pi_\theta$; sample a batch of pockets from $\mathcal{P}$
  Collect G trajectories $\tau_i = (z_{i,0}, \ldots, z_{i,T})$ per pocket using π_{old}
  Store group IDs, old transition log-likelihoods $\ell_{\text{old},it}$
  RDKit decode molecule, M_i from $z_{i,T}$
  // Reward and advantage estimation
  Score valid decoded molecules with $R(M_i, P)$
  Set invalid molecule reward to $min(0, \text{group valid rewards})$
  // Replay and policy update
  for $k = 1, \ldots, K$ **do**
    for each replay batch $\mathcal{B}$ **do**
      $\mathbf{A} \leftarrow (A_i)_{i \in \mathcal{B}}$; $\mathbf{P} \leftarrow (P_i)_{i \in \mathcal{B}}$; $\mathbf{J} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{B}|}$
      for $t = 0, \ldots, T-1$ **do**
        Evaluate batched log-likelihoods $\boldsymbol{\ell}_{\theta,t}$ at the stored transitions, $(\mathbf{z}_t, \mathbf{z}_{t+1}, t/T, \mathbf{P})$
        $\boldsymbol{\rho}_t \leftarrow \exp(\boldsymbol{\ell}_{\theta,t} - \boldsymbol{\ell}_{\text{old},t})$
        $\mathbf{L}_t^{\text{clip}} \leftarrow \min(\boldsymbol{\rho}_t \odot \mathbf{A}, \text{clip}(\boldsymbol{\rho}_t, 1-\epsilon, 1+\epsilon) \odot \mathbf{A})$
        $\mathbf{D}_t^{\text{KL}} \leftarrow D_{\text{KL}}(\pi_\theta(\cdot \mid \mathbf{z}_t, \mathbf{P}) \,\|\, \pi_{\text{ref}}(\cdot \mid \mathbf{z}_t, \mathbf{P}))$
        $\mathbf{J} \leftarrow \mathbf{J} + \mathbf{L}_t^{\text{clip}} - \beta \mathbf{D}_t^{\text{KL}}$
      end for
      $\mathcal{L} \leftarrow -\frac{1}{|\mathcal{B}|T} \sum_{i \in \mathcal{B}} J_i$
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
    end for
  end for
end for
return π_θ

4 RESULTS

4.1 EXPERIMENTAL SETUP

Datasets, baselines and model training. Using our adaptation of Flow-GRPO, we post-train FLOWR.ROOT on SPINDR (Cremer et al., 2025a), a dataset of 35,666 curated protein–ligand complexes derived from PLINDER (Durairaj et al., 2024). Throughout this study, we use Vina score-only without pose optimization in the reward when optimizing for Vina docking score. We use the original SPINDR splits for training, validation and test, using the validation split for checkpoint selection and the 225 held-out test pockets for final model evaluations. We further assess the transfer of the SPINDR-selected checkpoints to the CrossDocked2020 test set containing 100 pockets without additional fine-tuning or checkpoint selection; dataset and baseline details are provided in Appendix F.1, details about training compute and hyperparameters can be found in Appendix C.

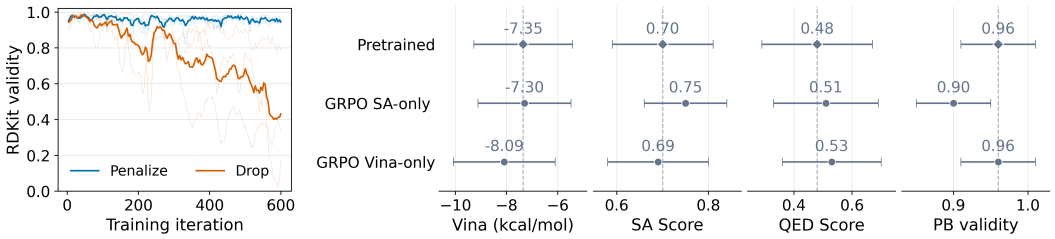
Evaluation metrics. We report Vina, QED, SA, PB-validity (PB-val) (Buttenschoen et al., 2024), bond-angle and bond-length Wasserstein distances, diversity, and uniqueness. We generate 100 ligands per pocket and aggregate three training seeds at the pocket level. Aggregation details are provided in Appendix D.1 and seed-level metrics are also shared in Appendix D.2. All metrics except for PB validity are computed on the subset of RDKit-valid samples.

4.2 SINGLE-OBJECTIVE ALIGNMENT AND TRADE-OFFS

Using Vina-only alignment experiments, we determined how invalid rollouts should be handled during training. Both, dropping invalid rollouts from the reward and penalizing them, improves the

Vina score among valid molecules and the validation reward (Figures 6B–C). However, dropping invalid rollouts reduces RDKit-valid yield to about 40% after approximately 600 iterations, whereas penalizing them preserves yield near the pretrained level (Figure 2A). Consequently, we penalize invalid rollouts in all subsequent experiments.

We next contrast Vina-only and SA-only alignment, revealing an opposing trade-off between the two objectives (Figure 2B and Appendix Figure 9): single-objective Flow-GRPO improved its targeted metric on the SPINDR test set. SA-only training increased mean normalized SA from 0.70 to 0.75 ($p = 3.2 \times 10^{-35}$ in a pocket-matched two-sided paired t-test), while mean Vina became slightly less favorable (-7.35 to -7.30 kcal/mol, $p = 0.044$) and PB validity decreased from 0.96 to 0.90. Conversely, Vina-only training improved mean Vina to -8.09 kcal/mol ($p = 5.4 \times 10^{-45}$) with a small decrease in normalized SA to 0.69 ($p = 0.041$), while maintaining PB validity at 0.96. Both settings increased mean QED, from 0.48 to 0.51 and 0.53, respectively, despite not directly optimizing it. The decrease in PB validity under SA-only optimization may reflect the absence of explicit 3D geometry constraints in the SA reward (Ertl & Schuffenhauer, 2009), whereas Vina evaluates geometry-dependent protein–ligand interactions and penalizes steric clashes (Yu et al., 2023). These results show that GRPO shifts the generated sample distribution toward the rewarded objective, but evaluation metrics do not improve uniformly, and in some cases even degrade, motivating a joint Vina-SA optimization.



(A) RDKit validity **(B)** Molecular metrics

Figure 2: **Invalid-rollout handling and single-objective alignment. (A)** RDKit-valid rollout fraction during fine-tuning when invalid molecules are penalized or dropped. Thin curves show individual seeds and bold curves show their mean. All runs were aligned using a Vina-only reward. **(B)** Vina, SA, QED, and PB validity for pretrained FLOWR.ROOT and SA-only and Vina-only Flow-GRPO on the SPINDR test set. Dashed vertical lines mark pretrained values. Lower Vina and higher normalized SA are preferred.

4.3 VINA-SA TRADE-OFF ACROSS DATASETS

Motivated by the Vina–SA trade-off observed under single-objective alignment, we investigate whether intermediate reward compositions can balance these objectives. Moving from SA-only ($w_{Vina} = 0$) to Vina-only ($w_{Vina} = 1$) improves mean Vina from -7.30 to -8.09 kcal/mol while reducing SA from 0.75 to 0.69 on the SPINDR test set (Figure 3; see Figures 7 and 8 for validation reward curves and the RDKit validity success rates). Intermediate weights provide useful compromises: $w_{Vina} = 0.75$, for example, improves the baseline results to a normalized SA value of 0.72, a Vina score of -7.92 kcal/mol, a pose-optimized Vina score of -8.22 kcal/mol, and a QED of 0.53 while maintaining the PoseBusters validity of 0.96 (Table 1). Although individual seeds vary, the aggregate values from three independent seeds show a consistent pattern across datasets: changing the reward weights moves the resulting model along a controllable Vina-SA frontier rather than producing a single uniformly optimal model. Evaluating the same checkpoints on CrossDocked2020 without additional fine-tuning reveals a similar trade-off (Figure 3 right panel), and the pattern remains after Vina's pose optimization (Figure 15).

Similarly, the three-objective Vina-QED-SA configuration improves the average Vina, SA, and QED scores simultaneously over the pretrained baseline, again preserving PoseBusters validity. Interestingly, diversity and uniqueness of the sampled molecules improve over the FLOWR.ROOT baseline across all tested Flow-GRPO configurations. Only Wasserstein distances of bond-lengths and bond-angles increase compared to the baseline (Table 1).

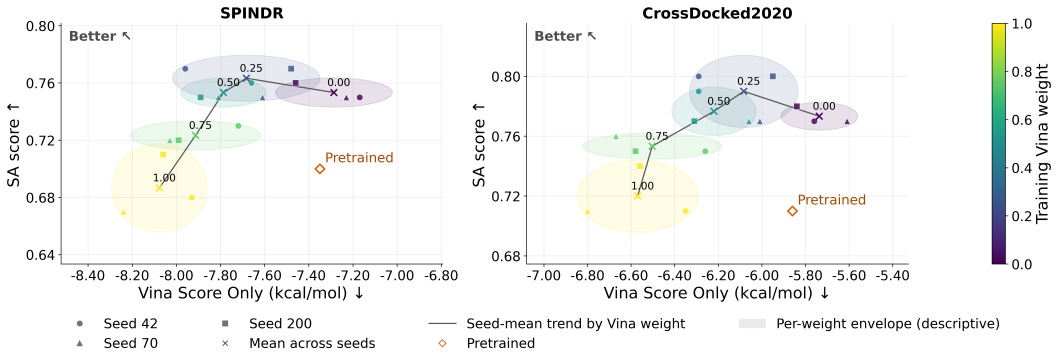


Figure 3: **Score-only Vina–SA trade-offs on SPINDR (left) and CrossDocked2020 (right) datasets.** Vina is evaluated directly on generated poses without pose optimization. All Flow-GRPO checkpoints are selected on SPINDR validation data and evaluated without further fine-tuning. Colors and numerical annotations indicate the Vina reward weight w_{Vina} and the SA weight is $1 - w_{\mathrm{V}}$. Circles, triangles, and squares show training seeds 42, 70, and 200, respectively. Crosses denote means across seeds, joined in increasing Vina-weight order. Shaded ellipses descriptively enclose the seed-level points for each weight, and orange diamonds denote pretrained FLOWR.ROOT. Lower Vina and higher SA are preferred.

In agreement with our observations from the single-objective alignment case, we find an improvement in PoseBusters validity with increasing Vina reward weight that can be explained in terms of internal energies, steric clashes and RDKit validity (Figure 16).

Table 1: **Metrics evaluation on the SPINDR test set.** In the Flow-GRPO labels, w_{Vina} denotes the Vina weight: Vina–SA uses $w_{\mathrm{SA}} = 1 - w_{\mathrm{V}}$, Vina–QED uses weights $(0.75, 0.25)$, and Vina–QED–SA uses $(0.50, 0.25, 0.25)$. Vina is score-only, whereas Vina min. is pose-optimized. Arrows indicate the preferred direction. The best result is shown in bold, and the second- and third-best distinct results are underlined. We generated molecules and computed metrics for FLOWR.ROOT using our evaluation pipeline. Values are means ± standard deviations across available pockets, except BA W1 and BL W1, which summarize distribution distances.

Model	PB-val. ↑	BA W1 ↓	BL W1 (10^{-2}) ↓	Vina ↓	Vina min. ↓	SA ↑	QED ↑	Diversity ↑	Uniqueness ↑
Test set	$0.99_{\pm 0.09}$	–	–	$-7.63_{\pm 2.18}$	$-7.88_{\pm 2.10}$	$0.67_{\pm 0.12}$	$0.49_{\pm 0.22}$	–	–
FLOWR.ROOT	$\underline{0.96_{\pm 0.05}}$	**0.68**	**0.077**	$-7.35_{\pm 1.93}$	$-7.73_{\pm 1.74}$	$0.70_{\pm 0.11}$	$0.48_{\pm 0.19}$	$\underline{0.53_{\pm 0.18}}$	$0.91_{\pm 0.16}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0$)	$0.90_{\pm 0.07}$	1.84	$\underline{0.116}$	$-7.30_{\pm 1.82}$	$-7.67_{\pm 1.68}$	$\underline{0.75_{\pm 0.09}}$	$0.51_{\pm 0.18}$	$\underline{0.58_{\pm 0.15}}$	$\underline{0.93_{\pm 0.15}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.25$)	$0.93_{\pm 0.08}$	2.03	$\underline{0.116}$	$-7.69_{\pm 1.90}$	$-8.03_{\pm 1.77}$	$\mathbf{0.76_{\pm 0.09}}$	$0.52_{\pm 0.18}$	$\mathbf{0.60_{\pm 0.14}}$	$\underline{0.93_{\pm 0.14}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.50$)	$\underline{0.95_{\pm 0.07}}$	1.71	0.121	$-7.79_{\pm 1.91}$	$-8.10_{\pm 1.80}$	$\underline{0.75_{\pm 0.10}}$	$0.52_{\pm 0.18}$	$\underline{0.58_{\pm 0.15}}$	$\underline{0.93_{\pm 0.15}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.75$)	$\underline{0.96_{\pm 0.04}}$	$\underline{1.21}$	$\underline{0.097}$	$\underline{-7.92_{\pm 2.00}}$	$\underline{-8.22_{\pm 1.88}}$	$0.72_{\pm 0.10}$	$\underline{0.53_{\pm 0.17}}$	$0.56_{\pm 0.15}$	$\underline{0.92_{\pm 0.15}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 1$)	$\underline{0.96_{\pm 0.05}}$	1.44	0.122	$\mathbf{-8.09_{\pm 1.99}}$	$\mathbf{-8.39_{\pm 1.89}}$	$0.69_{\pm 0.11}$	$\underline{0.53_{\pm 0.17}}$	$0.56_{\pm 0.14}$	$\mathbf{0.94_{\pm 0.13}}$
Flow-GRPO (Vina–QED)	$\mathbf{0.97_{\pm 0.04}}$	$\underline{1.06}$	0.124	$\underline{-7.88_{\pm 1.97}}$	$\underline{-8.21_{\pm 1.85}}$	$0.69_{\pm 0.11}$	$\mathbf{0.59_{\pm 0.17}}$	$0.57_{\pm 0.14}$	$\mathbf{0.94_{\pm 0.13}}$
Flow-GRPO (Vina–QED–SA)	$\underline{0.96_{\pm 0.04}}$	1.43	0.136	$-7.79_{\pm 1.93}$	$-8.09_{\pm 1.80}$	$\underline{0.73_{\pm 0.10}}$	$0.56_{\pm 0.17}$	$\underline{0.57_{\pm 0.15}}$	$\mathbf{0.94_{\pm 0.13}}$

Additional SPINDR evaluations of MolJO and MolForm-DPO, together with PILOT results reported by FLOWR.ROOT, are provided in Appendix Table 3. MolJO and MolForm-DPO are evaluated using checkpoints trained on CrossDocked2020, whereas Flow-GRPO is fine-tuned on SPINDR. We therefore include these results as reference points and we assess the performance of GRPO alignment through comparisons with pretrained FLOWR.ROOT.

Without additional fine-tuning, we benchmarked the SPINDR-trained and selected checkpoints on CrossDocked2020 and compared results against a range of models that have been trained or reward-aligned on CrossDocked2020 (Table 5). The $w_{Vina} = 0.75$ model improves pretrained PB-validity from 0.90 to 0.93, score-only Vina from -5.86 to -6.57 kcal/mol, pose-optimized Vina from -7.33 to -7.71 kcal/mol, SA from 0.71 to 0.75, and QED from 0.50 to 0.55. It obtains the best mean PB-validity, while the bond-length distance remains near baseline, and the bond-angle distance degrades. MolJO and AliDiff achieve better docking scores, and DrugFlow-DPO has higher SA and QED.

4.4 WHICH MOLECULAR DESCRIPTOR CHANGES ACCOMPANY THE VINA–SA TRADE-OFFS?

We next examine how ligand flexibility, ring topology, stereochemical complexity, and Crippen logP vary across reward compositions on SPINDR (Figure 4). During sampling, ligand sizes are drawn uniformly within $\pm 10\%$ of the reference ligand size, the mean heavy-atom count therefore remains nearly constant across models. Increasing the Vina weight from 0 to 1 reduces the mean rotatable-bond count from approximately 5.1 to 3.7 and increases the mean total ring count from approximately 3.2 to 4.0. Aromatic-ring count and logP exceed the pretrained baseline throughout the sweep but peak at intermediate weights. Potential stereocenter count is lowest at $w_{\text{Vina}} = 0.25$ and then increases with the Vina reward component, while remaining below the pretrained baseline. Appendix Figure 12 shows example generated molecules with property and metric annotations.

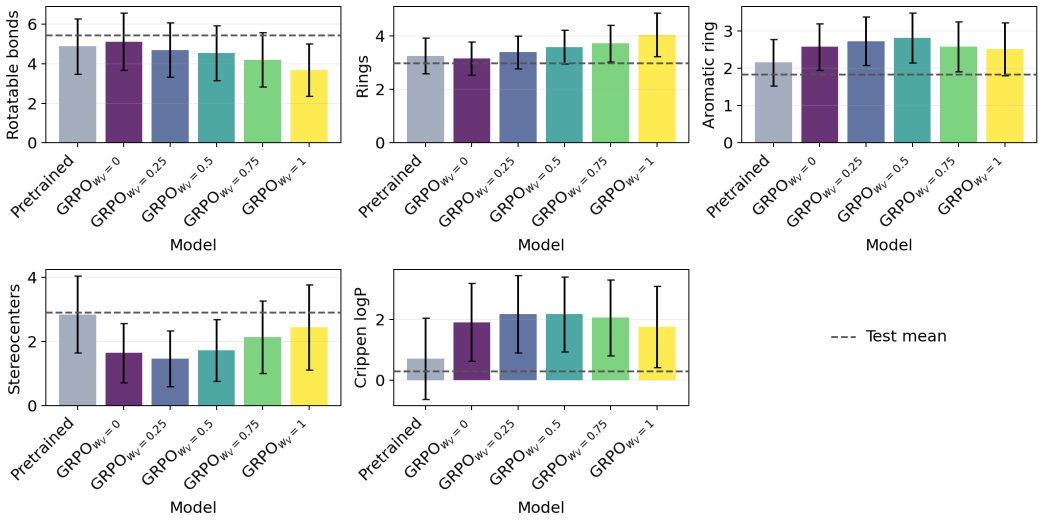


Figure 4: **Molecular descriptors across reward compositions on SPINDR.** Bars show rotatable-bond, total-ring, aromatic-ring, and potential stereocenter counts, and Crippen logP for pretrained FLOWR.ROOT and Flow-GRPO variants, where w_V is the Vina reward weight and the SA weight is $1 - w_{\text{V}}$. GRPO subscripts indicate the Vina reward weight. Descriptors are averaged over molecules within each pocket, then over seeds within pockets, and finally over pockets. Error bars show molecule-level standard deviations averaged over seeds and pockets. Dashed horizontal lines mark the matched test-ligand means.

Consistent with the ligand descriptor–score relationships examined in prior work such as PILOT (Cremer et al., 2024), our within-pocket analysis associates more rings with more favorable Vina scores (Appendix Figure 10). Moreover, fewer rotatable bonds correlate with favorable Vina since Vina penalizes ligand flexibility. Below, we focus on whether changes in these descriptors relative to pretrained FLOWR.ROOT accompany changes in scores after alignment.

To assess which descriptor changes accompany alignment, we subtract pretrained pocket means from aligned pocket means for each descriptor and score, and correlate these changes across matched pockets. For $w_{\text{Vina}} = 0.75$ (Figure 5), pockets with larger reductions in rotatable-bond count tend to show larger Vina improvements (reduction in Vina score, since lower values are better) ($\rho = 0.64$), while increases in total rings, aromatic rings, and logP associate with more favorable Vina changes ($\rho = -0.66$, -0.47, and -0.57, respectively). SA improvements are most strongly associated with reductions in potential stereocenter count ($\rho = -0.72$), and also with increases in aromatic-ring count ($\rho = 0.67$) and logP ($\rho = 0.65$). Results for the other reward weights are provided in Appendix Figure 11.

The $w_{\text{Vina}} = 0.75$ model shows similar associations on CrossDocked2020, although those involving Vina are weaker (Appendix Figure 14). Rotatable-bond and total-ring changes correlate with Vina

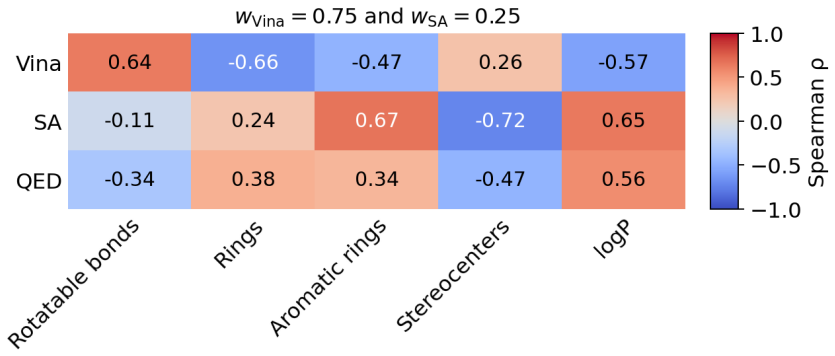


Figure 5: **Correlations of descriptor and score changes for $w_{\text{Vina}} = 0.75$ GRPO model on SPINDR.** Changes are aligned minus pretrained pocket means. Spearman correlations are computed across matched pockets separately for each training seed, then averaged across seeds. Rows show score-only Vina, SA, and QED; columns show molecular descriptors. Negative Vina changes and positive SA or QED changes indicate improvements. Cell annotations report mean correlation coefficients.

changes at $\rho = 0.33$ and -0.42, respectively, compared with 0.64 and -0.66 on SPINDR. SA changes correlate negatively with stereocenter changes ($\rho = -0.66$) and positively with aromatic-ring and logP changes ($\rho = 0.64$ and 0.60). The appendix figure also provides results for the other reward weights. Thus, the selected balanced model exhibits similar descriptor–score relationships across the two datasets, with differing strengths of association.

We have also analysed protein-ligand interaction features extracted using the ProLIF library (Bouysset & Fiorucci, 2021) (Appendix G.3). Across both datasets, alignment also increases the number of hydrophobic contacts relative to pretrained FLOWR.ROOT, while hydrogen-bond counts remain lower. Hydrophobic contacts peak at $w_{Vina} = 0.25$ rather than increasing with Vina weight, showing a less consistent relationship with the docking-score trend than the ligand-feature shifts.

5 Limitations

Although Flow-GRPO can be applied to diffusion- and flow-based generative models, we evaluate only FLOWR.ROOT (Cremer et al., 2025b). Future work will evaluate trade-offs between different optimization objectives and their impact on molecular descriptors using additional generative backbones.

6 Conclusions

We adapted Flow-GRPO to mixed continuous-discrete trajectories of pocket-conditioned 3D ligand generation, enabling online post-training using black-box rewards. Our experiments show that handling invalid molecules is crucial for fine-tuning: penalizing invalid rollout molecules preserves valid molecule yield, whereas excluding them leads to validity collapse. Single-objective alignment improves the optimized property but exposes a trade-off between Vina and SA. While multi-objective rewards provide controllable operating points, no single multi-objective configuration dominates all metrics. This study further provides a clear understanding of the intramolecular changes that accompany the optimization for Vina and SA: Vina improvements are strongly associated with a reduction in rotatable bonds and an increase in ring structures, whereas improvements in SA are linked to reduced stereocentres and increases in aromatic rings. Vina–SA multi-reward optimization learns to trade off these molecular descriptors to optimize the specific reward configuration.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the prose of this manuscript, for literature research, and for aiding code development. We have manually reviewed all LLM-generated code and text and verified results of any LLM-assisted literature review. We have not used generative AI tools for drafting full paper sections, and generating synthetic datasets as well as proving mathematical claims are not applicable to this work.

REPRODUCIBILITY STATEMENT

Code and model checkpoints used for this study will be made publicly available upon publication.

REFERENCES

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Cédric Bouysset and Sébastien Fiorucci. ProLIF: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1):72, 2021. doi: 10.1186/s13321-021-00548-6. URL https://doi.org/10.1186/s13321-021-00548-6.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical science*, 15(9):3130–3139, 2024.

Xiwei Cheng, Xiangxin Zhou, Yuwei Yang, Yu Bao, and Quanquan Gu. Decomposed direct preference optimization for structure-based drug design. *arXiv preprint arXiv:2407.13981*, 2024. URL https://arxiv.org/abs/2407.13981.

Julian Cremer, Tuan Le, Frank Noé, Djork-Arné Clevert, and Kristof T Schütt. Pilot: equivariant diffusion for pocket-conditioned de novo ligand generation with multi-objective guidance via importance sampling. *Chemical Science*, 15(36):14954–14967, 2024.

Julian Cremer, Ross Irwin, Alessandro Tibo, Jon Paul Janet, Simon Olsson, and Djork-Arné Clevert. Flowr: Flow matching for structure-aware de novo, interaction-and fragment-based ligand generation. *arXiv preprint arXiv:2504.10564*, 2025a.

Julian Cremer, Tuan Le, Mohammad M Ghahremanpour, Emilia Sługocka, Filipe Menezes, and Djork-Arné Clevert. Flowr. root: A flow matching based foundation model for joint multi-purpose structure-aware 3d ligand generation and affinity prediction. *arXiv preprint arXiv:2510.02578*, 2025b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009. doi: 10.1186/1758-2946-1-8.

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. doi: 10.1021/acs.jcim.0c00411.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in neural information processing systems*, 37:44040–44063, 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. *arXiv preprint arXiv:2303.03543*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Seungbeom Lee, Munsun Jo, Jungseul Ok, and Dongwoo Kim. Enhancing ligand validity and affinity in structure-based drug design with multi-reward optimization. In *ICML*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di ZHANG, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17644–17655. PMLR, 2022.

Keyue Qiu, Yuxuan Song, Jie Yu, Hongbo Ma, Ziyao Cao, Zhilong Zhang, Yushuai Wu, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Empower structure-based molecule optimization with gradient guided bayesian flow networks. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Arne Schneuing, Ilia Igashov, Adrian W Dobbelstein, Thomas Castiglione, Michael Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. *arXiv preprint arXiv:2508.17815*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Kunyu Wang, Jon Paul Janet, and Alessandro Tibo. Controllable molecular generation with fine-tuned flow-matching model. *ChemRxiv*, 2025(1125), 2025. doi: 10.26434/chemrxiv-2025-db0m9. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2025-db0m9.

Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-dock: Gpu-accelerated docking enables ultralarge virtual screening. *Journal of Chemical Theory and Computation*, 19(11):3336–3345, 2023.

Daiheng Zhang and Zhao Zhang. MolFORM: Preference-aligned multimodal flow matching for structure-based drug design. *arXiv preprint arXiv:2507.05503*, 2025. URL `https://arxiv.org/abs/2507.05503`.

Xudong Zhang, Sanqing Qu, Fan Lu, Jianmin Wang, Zhixin Tian, Shangding Gu, Yanping Zhang, Alois Knoll, Shaorong Gao, Guang Chen, et al. Steering semi-flexible molecular diffusion model for structure-based drug design with reinforcement learning. *Science Advances*, 12(16):eady9955, 2026.

Zhijian Zhou, Junyi An, Zongkai Liu, Yunfei Shi, Xuan Zhang, Fenglei Cao, Chao Qu, and Yuan Qi. Guiding diffusion models with reinforcement learning for stable molecule generation. *arXiv preprint arXiv:2508.16521*, 2025.

Baoheng Zhu, Deyu Bo, Delvin Ce Zhang, and Xiao Wang. Graph-grpo: Training graph flow models with reinforcement learning. *arXiv preprint arXiv:2603.10395*, 2026.

A FLOWR.ROOT SAMPLING DYNAMICS

FLOWR.ROOT is a flow-based generator for pocket-conditioned 3D molecular generation (Cremer et al., 2025b). Given a protein pocket P, the ligand state at flow time t is $z_t = (\mathbf{x}_t, \mathbf{a}_t, \mathbf{q}_t, \mathbf{b}_t)$, where $\mathbf{x}_t$ contains continuous atomic coordinates and $\mathbf{a}_t$, $\mathbf{q}_t$, and $\mathbf{b}_t$ represent categorical atom types, formal charges, and bond types. Conditioned on (z_t, t, P), FLOWR.ROOT predicts clean-state coordinates and categorical endpoint distributions, which define the transition from z_t to z_{t+1}. The final mixed state is decoded into an RDKit molecule.

Sampler state and endpoint prediction. For a padded ligand representation with at most N atoms, FLOWR.ROOT represents the state at flow time $t \in [0, 1]$ as

$$z_t = (\mathbf{X}_t, \mathbf{A}_t, \mathbf{Q}_t, \mathbf{B}_t), \tag{5}$$

where $\mathbf{X}_t \in \mathbb{R}^{N\times 3}$ contains coordinates, $\mathbf{A}_t \in \{0,1\}^{N\times C_a}$ atom types, $\mathbf{Q}_t \in \{0,1\}^{N\times C_q}$ formal charges, and $\mathbf{B}_t \in \{0,1\}^{N\times N\times C_b}$ bond types. The C_a, C_q, and C_b dimensions are the corresponding category counts; bond tensors are symmetric and include the no-bond category. Padded sites are excluded by an atom mask $\mathbf{m} \in \{0,1\}^N$, with $n = \sum_{u=1}^{N} m_u$. Conditioned on the current ligand state, flow time, and protein pocket P, the pretrained generator predicts the data endpoint,

$$f_\theta(z_t, t; P) = \left(\widehat{\mathbf{X}}_{1,\theta}, \widehat{\boldsymbol{p}}^{a}_{1,\theta}, \widehat{\boldsymbol{p}}^{q}_{1,\theta}, \widehat{\boldsymbol{p}}^{b}_{1,\theta}\right), \tag{6}$$

where the first output is a coordinate estimate and the remaining outputs are softmax probability vectors over endpoint categories. These outputs are the only FLOWR.ROOT interface required by Flow-GRPO.

Stochastic sampling dynamics. Sampling follows Algorithm 2 of Cremer et al. (2025b), specialized to de novo pocket-conditioned generation. Let $t_k = k/T$ and $h = 1/T$ be the time and the constant step size, respectively. Sampling starts with Gaussian coordinate noise and independent uniform categorical states. At step k, the endpoint-coordinate prediction induces the velocity and score-like correction

$$\mathbf{v}_{\theta,k} = \frac{\widehat{\mathbf{X}}_{1,\theta,k} - \mathbf{X}_k}{1 - t_k + \varepsilon_x}, \qquad \mathbf{s}_{\theta,k} = \frac{t_k \mathbf{v}_{\theta,k} - \mathbf{X}_k}{1 - t_k + \varepsilon_x}, \tag{7}$$

where $\varepsilon_x = 10^{-6}$ is a numerical stabilizer. The Flow-GRPO rollouts use the stochastic update

$$g(t) = \frac{\mathbb{1}[t < t_c]}{t + \varepsilon_g}, \qquad t_c = 0.9, \quad \varepsilon_g = 10^{-2}, \tag{8}$$

$$\mathbf{X}_{k+1} = \mathbf{X}_k + h\left(\mathbf{v}_{\theta,k} + g(t_k)\mathbf{s}_{\theta,k}\right) + h\sqrt{2g(t_k)\sigma_c}\,\boldsymbol{\xi}_k, \qquad \boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

with coordinate-noise scale σ_c. Thus, while $g(t_k) > 0$, the conditional coordinate transition is Gaussian with

$$\boldsymbol{\mu}^{x}_{\theta,k} = \mathbf{X}_k + h(\mathbf{v}_{\theta,k} + g(t_k)\mathbf{s}_{\theta,k}), \qquad \sigma^{x}_{k} = h\sqrt{2g(t_k)\sigma_c}. \tag{10}$$

After the cutoff, $g(t_k) = 0$ and the coordinate update reduces to the deterministic flow drift.

For a categorical feature $r \in \{a, q, b\}$ with C_r classes, let j be the current category at a site and let $\widehat{p}^{\,r}_{1,\theta,k,s}$ be the predicted endpoint probability of class s. The uniform discrete-flow kernel in FLOWR.ROOT is

$$\begin{aligned} K^{r}_{\theta,k}(s \mid j) &= \frac{h}{1 - t_k}\widehat{p}^{\,r}_{1,\theta,k,s}, && s \neq j, \\ K^{r}_{\theta,k}(j \mid j) &= 1 - \sum_{s\neq j} K^{r}_{\theta,k}(s \mid j). \end{aligned} \tag{11}$$

More details on FLOWR.ROOT sampling strategy can be found in Cremer et al. (2025b).

B Implementation of Flow-GRPO objective for FLOWR.ROOT model

Policies. The three policies in Equation 1 have different roles:

- π_{old} is frozen while its stored trajectories are replayed
- π_θ is the current trainable policy and
- π_{ref} is the frozen pretrained FLOWR.ROOT policy used by the KL penalty

At the beginning of collection, π_{old} is the current policy; it becomes off-policy only as replay updates θ.

Trajectory collection. For each conditioning pocket P_g, the rollout policy π_{old} samples a same-pocket group of G trajectories,

$$\{\tau_i\}_{i=1}^{G}, \qquad \tau_i = (z_{i,0}, \dots, z_{i,T}). \tag{12}$$

The state is $(z_{i,k}, t_k, P_g)$, the action is the sampled next mixed state $z_{i,k+1}$, and the terminal molecular reward R_i is computed after the final state is decoded into an RDKit molecule and scored. Its group-relative advantage is then used for every denoising transition in τ_i.

Group-relative advantage. Let $m_i \in \{0, 1\}$ be the rollout-inclusion mask. We set $m_i = 1$ for a rollout that decodes to an RDKit-valid molecule. For a decode-failed rollout, the policy in Section 3.2 sets $m_i = 0$ under *drop* and $m_i = 1$ under *penalize*. Define $\mathcal{I}_g = \{i : g(i) = g,\, m_i = 1\}$. For groups with at least two included members, our implementation uses the population moments

$$\mu_g = \frac{1}{|\mathcal{I}_g|} \sum_{j \in \mathcal{I}_g} R_j, \qquad \sigma_g = \left[\frac{1}{|\mathcal{I}_g|} \sum_{j \in \mathcal{I}_g} (R_j - \mu_g)^2 \right]^{1/2}, \tag{13}$$

$$A_i = \frac{R_i - \mu_{g(i)}}{\max(\sigma_{g(i)}, \varepsilon_R)}, \qquad i \in \mathcal{I}_{g(i)}. \tag{14}$$

Where ε_R is for numerical stability. Masked samples and all members of a group with fewer than two included members receive zero advantage and are discarded from the objective. Because the reward is terminal, A_i is constant across all T denoising transitions of trajectory i.

Clipped policy objective. For denoising transition $k \in \{0, \dots, T-1\}$, let $\ell_{\theta,ik}$ be the size-invariant mixed-state log-likelihood defined in Appendix B.2, and let $\ell_{\text{old},ik}$ be the value stored during rollout for the ith sample. We first clip the log-ratio for numerical stability and then apply PPO style clipping:

$$\Delta_{ik} = \ell_{\theta,ik} - \ell_{\text{old},ik}, \qquad \widetilde{\Delta}_{ik} = \text{clip}(\Delta_{ik}, -c_{\log}, c_{\log}), \tag{15}$$

$$\rho_{ik} = \exp(\widetilde{\Delta}_{ik}), \qquad \rho_{ik} = \text{clip}(\rho_{ik}, 1-\epsilon, 1+\epsilon), \tag{16}$$

$$J_{ik}^{\text{pol}}(\theta) = \min(\rho_{ik} A_i, \rho_{ik} A_i). \tag{17}$$

Here, J^{pol} is a policy term, $c_{\log}$ is the log-ratio bound and ϵ is the PPO clipping threshold. For a logical replay group $\mathcal{B}$, let $G_{\text{eff}} = \sum_{i \in \mathcal{B}} m_i$, then $J_{\mathcal{B}}$ is given by

$$J_{\mathcal{B}}(\theta) = \frac{1}{T G_{\text{eff}}} \sum_{k=0}^{T-1} \sum_{i \in \mathcal{B}} m_i \left[J_{ik}^{\text{pol}}(\theta) - \beta D_{ik}^{\text{KL}} \right]. \tag{18}$$

D_{ik}^{KL} is the KL term defined in Appendix B.3, and $\beta > 0$ controls regularization.

The two nested sums in Equation 18 correspond to a nested sample-by-denoising-step evaluation. For every $i \in \mathcal{B}$ with $m_i = 1$ and every $k \in \{0, \dots, T-1\}$, replay recomputes the current transition distribution at $(z_{i,k}, t_k, P)$, evaluates the policy term against the stored rollout likelihood, and evaluates the KL to the reference transition distribution at that same state. The terminal advantage

A_i is constant across trajectory i, but both transition-level terms vary with k. The resulting pairwise contributions are summed over all eligible (i, k) pairs and divided by TG_{eff}.

The clipped objective increases the likelihood of transitions associated with above-average molecules and decreases it for transitions associated with below-average molecules, while limiting excessively large policy updates. The KL term further constrains the fine-tuned generator from drifting too far from the pretrained molecular distribution.

B.1 Mixed continuous-discrete transition likelihood

For rollout i at denoising step k, the molecular state is $z_{i,k} = (\mathbf{X}_{i,k}, \mathbf{A}_{i,k}, \mathbf{Q}_{i,k}, \mathbf{B}_{i,k})$. Here, $a_{i,k,u}$ and $q_{i,k,u}$ denote the atom-type and formal-charge categories at atom site u, respectively, and $b_{i,k,uv}$ denotes the bond category for the undirected atom pair (u, v). The atom mask $m_{iu} \in \{0, 1\}$ selects the $n_i = \sum_{u=1}^{N} m_{iu}$ valid atoms from the N padded sites, and $\mathcal{E}_i = \{(u, v) : u < v,\, m_{iu} m_{iv} = 1\}$ contains all valid undirected bonded atom pairs.

Let $\mu^x_{\theta,ik,ud}$ be coordinate dimension $d \in \{1, 2, 3\}$ of the transition mean for atom u in Equation 10, evaluated at rollout state $(z_{i,k}, t_k, P)$, and let σ^x_{ik} be the corresponding coordinate standard deviation. We use $p^x_{\theta,ik} := p^x_\theta(\cdot \mid z_{i,k}, t_k, P)$ for this conditional coordinate distribution. Because its covariance is diagonal, its density at the next coordinate matrix factorizes as

$$p^x_{\theta,ik}(\mathbf{X}_{i,k+1}) = \prod_{u:m_{iu}=1} \prod_{d=1}^{3} \mathcal{N}\big(x_{i,k+1,ud} \mid \mu^x_{\theta,ik,ud}, (\sigma^x_{ik})^2\big)\,. \tag{19}$$

Thus, $p^x_{\theta,ik}$ is a joint density over all valid atomic coordinates, which is factored as a product over atoms and Cartesian dimensions.

For a categorical feature $r \in \{a, q, b\}$, let $K^r_{\theta,ik,\omega}(c' \mid c)$ denote its one-step Markov transition probability from the current category c to the next category c' at site ω. The site is an atom index $\omega = u$ for atom types and charges, and an undirected pair $\omega = (u, v)$ for bonds. Each K^r is the categorical transition kernel in Equation 11, evaluated using the endpoint probabilities predicted by the policy.

Conditioned on the complete current state $(z_{i,k}, t_k, P)$, the sampler uses independent coordinate noise and independent categorical draws across feature types and valid sites. This conditional-independence property gives the product factorization

$$\begin{aligned} p_\theta(z_{i,k+1} \mid z_{i,k}, t_k, P) = p^x_\theta(\mathbf{X}_{i,k+1} \mid z_{i,k}, t_k, P) &\prod_{u:m_{iu}=1} K^a_{\theta,ik,u}(a_{i,k+1,u} \mid a_{i,k,u}) \\ &\times \prod_{u:m_{iu}=1} K^q_{\theta,ik,u}(q_{i,k+1,u} \mid q_{i,k,u}) \\ &\times \prod_{(u,v)\in\mathcal{E}_i} K^b_{\theta,ik,(u,v)}(b_{i,k+1,uv} \mid b_{i,k,uv}), \end{aligned} \tag{20}$$

where P is the protein pocket and θ parameterizes the current policy. The coordinate factor contains the two products in Equation 19. The three products shown explicitly in Equation 20 are the joint conditional probabilities of all valid atom-type, formal-charge, and undirected bond updates, respectively. The factorization does not assume that the policy predictions are marginally independent: every factor is produced from the same complete ligand state and pocket context. It states that, after conditioning on that shared context, the transition draws are independent, so their joint probability is the product of their individual probabilities.

Taking the logarithm of Equation 19 for the stored sampled coordinates, and omitting the policy-independent Gaussian constant during replay, gives the raw coordinate log-likelihood

$$\ell^{x,\text{raw}}_{\theta,ik} = -\frac{1}{2} \sum_{u=1}^{N} \sum_{d=1}^{3} m_{iu} \left[\left(\frac{x_{i,k+1,ud} - \mu^x_{\theta,ik,ud}}{\sigma^x_{ik} + 10^{-6}} \right)^2 + 2 \log(\sigma^x_{ik}) \right]. \tag{21}$$

For the categorical terms, let $c^r_{i,k,\omega}$ denote the current category of feature r at site ω, and define the shorthand $K^r_{\theta,ik,\omega}(c) := K^r_{\theta,ik,\omega}(c \mid c^r_{i,k,\omega})$ for the probability of next category $c \in \{1, \ldots, C_r\}$. Let

$y^r_{i,k+1,\omega,c}$ be the one-hot encoding of the stored next category. The per-site categorical log-likelihood is

$$\lambda^r_{\theta,ik,\omega} = \sum_{c=1}^{C_r} y^r_{i,k+1,\omega,c} \log \max(K^r_{\theta,ik,\omega}(c)), \tag{22}$$

where C_r is the number of categories for feature r. Because the one-hot vector selects the observed next category, this expression is simply the log of its transition probability. The raw categorical log-likelihoods are therefore obtained by summing over the appropriate valid sites:

$$\ell^{a,\mathrm{raw}}_{\theta,ik} = \sum_{u=1}^{N} m_{iu} \lambda^a_{\theta,ik,u}, \qquad \ell^{q,\mathrm{raw}}_{\theta,ik} = \sum_{u=1}^{N} m_{iu} \lambda^q_{\theta,ik,u}, \tag{23}$$

$$\ell^{b,\mathrm{raw}}_{\theta,ik} = \sum_{(u,v)\in\mathcal{E}_i} \lambda^b_{\theta,ik,(u,v)}. \tag{24}$$

The superscript raw indicates these are unnormalized log-likelihood values. The same uniform-sample transition kernel is used during sampling, old-policy log-likelihood storage, and current/reference replay. All reported settings include the atom-type, formal-charge, and bond terms; hence reward-conditioned credit is assigned to both molecular geometry and graph identity.

B.2 Size-invariant mixed-state likelihood

Molecules with more atoms and candidate bonds contain more stochastic sites. An unnormalized sum would therefore produce systematically larger-magnitude log-ratios for larger ligands. Following the size-invariant motivation of Zhou et al. (2025), but using the reductions implemented for our mixed molecular state, we average each modality over its valid molecular sites. With n_i and $\mathcal{E}_i$ defined above, let $e_i = |\mathcal{E}_i| = \sum_{u<v} m_{iu} m_{iv}$ be the number of valid undirected atom pairs. Evaluated ligand states satisfy $n_i > 0$ and $e_i > 0$, so no zero-denominator guard is required. The four contributions are

$$\ell^x_{\theta,ik} = \frac{\ell^{x,\mathrm{raw}}_{\theta,ik}}{n_i}, \qquad \ell^a_{\theta,ik} = \frac{\ell^{a,\mathrm{raw}}_{\theta,ik}}{n_i}, \tag{25}$$

$$\ell^q_{\theta,ik} = \frac{\ell^{q,\mathrm{raw}}_{\theta,ik}}{n_i}, \qquad \ell^b_{\theta,ik} = \frac{\ell^{b,\mathrm{raw}}_{\theta,ik}}{e_i}, \tag{26}$$

$$\ell_{\theta,ik} = \ell^x_{\theta,ik} + \ell^a_{\theta,ik} + \ell^q_{\theta,ik} + \ell^b_{\theta,ik}. \tag{27}$$

Thus $\ell_{\theta,ik}$ is a size-normalized composite transition log-likelihood. For coordinates, the three Cartesian dimensions are first summed within each atom in Equation 21, and these per-atom contributions are then averaged over n_i. Atom-type and charge terms are also averaged over n_i, whereas bond terms are averaged over their natural sites, the e_i valid undirected pairs. Consequently, every modality contributes a site average rather than a graph-size-dependent sum. The same reduction is applied to π_θ and π_{old} before forming the importance ratio.

B.3 Mixed continuous-discrete KL computation

Decomposition from the transition factorization. We reuse the mixed-state factorization, masks, and valid-site sets defined in Appendix B.1. At transition (i,k), let $\pi_{\theta,ik}$ be the current policy distribution and $\pi_{\mathrm{ref},ik}$ the frozen reference distribution, both conditioned on the same $(z_{i,k}, t_k, P)$. Let $p^x_{\theta,ik}$ and $p^x_{\mathrm{ref},ik}$ be the coordinate factors. For product measures, the additivity of KL divergence gives $D_{\mathrm{KL}}(\prod_j p_j \| \prod_j q_j) = \sum_j D_{\mathrm{KL}}(p_j \| q_j)$. Applying this identity to the conditional factorization in Equation 20 yields the exact unnormalized decomposition

$$\begin{aligned}\widetilde{D}^{\mathrm{KL}}_{ik} = {} & D_{\mathrm{KL}}(p^x_{\theta,ik} \| p^x_{\mathrm{ref},ik}) + \sum_{u=1}^{N} m_{iu} D_{\mathrm{KL}}(K^a_{\theta,ik,u} \| K^a_{\mathrm{ref},ik,u}) \\ & + \sum_{u=1}^{N} m_{iu} D_{\mathrm{KL}}(K^q_{\theta,ik,u} \| K^q_{\mathrm{ref},ik,u}) + \sum_{(u,v)\in\mathcal{E}_i} D_{\mathrm{KL}}(K^b_{\theta,ik,(u,v)} \| K^b_{\mathrm{ref},ik,(u,v)}).\end{aligned} \tag{28}$$

In each categorical term, both kernels are evaluated at the same current category and the divergence compares their complete probability vectors over the next category. Thus no cross-modality terms are needed: the joint divergence is the sum of one continuous-coordinate term and categorical terms for every valid atom-type, charge, and undirected bond site. This exact decomposition relies on the same conditional-independence property used for the transition likelihood. We evaluate the factorwise KLs analytically, apply the size-invariant reductions below, and retain their additive aggregation.

Continuous coordinate KL. The current and reference coordinate transitions are diagonal Gaussians with means $\boldsymbol{\mu}^x_{\theta,ik}$ and $\boldsymbol{\mu}^x_{\mathrm{ref},ik}$ and standard deviations $\boldsymbol{\sigma}^x_{\theta,ik}$ and $\boldsymbol{\sigma}^x_{\mathrm{ref},ik}$. For valid atom u, the exact Gaussian KL summed over its three coordinate dimensions is

$$\kappa^x_{ik,u} = \sum_{d=1}^{3} \left[-\frac{1}{2} + \log \frac{\sigma^x_{\mathrm{ref},ik,ud}}{\sigma^x_{\theta,ik,ud}} + \frac{(\sigma^x_{\theta,ik,ud})^2}{2(\sigma^x_{\mathrm{ref},ik,ud})^2} + \frac{(\mu^x_{\theta,ik,ud} - \mu^x_{\mathrm{ref},ik,ud})^2}{2(\sigma^x_{\mathrm{ref},ik,ud})^2} \right]. \tag{29}$$

The coordinate contribution is

$$D^x_{ik} = \frac{1}{n_i} \sum_{u=1}^{N} m_{iu} \kappa^x_{ik,u}. \tag{30}$$

As in the likelihood calculation, the three Cartesian dimensions are summed within each atom and the resulting per-atom divergences are averaged over n_i.

Categorical KL and final aggregation. For feature $r \in \{a, q, b\}$ and valid site ω, the exact divergence between the current and reference next-category distributions from Appendix B.1 is

$$\kappa^r_{ik,\omega} = \sum_{c=1}^{C_r} K^r_{\theta,ik,\omega}(c) \log \frac{K^r_{\theta,ik,\omega}(c)}{K^r_{\mathrm{ref},ik,\omega}(c)}. \tag{31}$$

Here, $\omega = u$ for an atom-type or charge site and $\omega = (u, v)$ for a bond site; C_r and the shorthand $K^r_{\cdot,ik,\omega}(c)$ are defined in Appendix B.1. Both kernels condition on the same stored current category $c^r_{i,k,\omega}$. Atom-type and charge terms are averaged over valid atoms; bond terms are averaged over the valid undirected pairs $\mathcal{E}_i$:

$$D^a_{ik} = \frac{1}{n_i} \sum_{u=1}^{N} m_{iu} \kappa^a_{ik,u}, \tag{32}$$

$$D^q_{ik} = \frac{1}{n_i} \sum_{u=1}^{N} m_{iu} \kappa^q_{ik,u}, \tag{33}$$

$$D^b_{ik} = \frac{1}{e_i} \sum_{(u,v)\in\mathcal{E}_i} \kappa^b_{ik,(u,v)}. \tag{34}$$

The size-invariant mixed-state regularizer is their additive sum,

$$D^{\mathrm{KL}}_{ik} = D^x_{ik} + D^a_{ik} + D^q_{ik} + D^b_{ik}. \tag{35}$$

C Experiment setup and hyperparameters

FLOWR.ROOT is post-trained on the SPINDR training split and evaluated on the SPINDR and CrossDocked2020 test sets.

Training and compute. Each run uses four NVIDIA GB200 GPUs. A rollout-collection samples a global batch of 512 molecular trajectories using 32 pockets, organized into same-pocket groups of 16. Flow-GRPO stores the sampled trajectory states and old-policy likelihoods, then replays the states through the current and reference policies at every denoising step to evaluate the policy ratio

and KL regularizer. Each rollout collection is followed by reward computation and two parameter updates. Unless stated otherwise, all experiments use 450 rollout collection rounds, corresponding to 900 optimization iterations, and each fine-tuning takes approximately 3.5 days. The dominant costs are sampling complete trajectories, replaying every denoising step through the current and reference policies, and estimating Vina scores. Although Uni-Dock is GPU accelerated and supports batch mode, it supports batching ligands for only one pocket at a time rather than processing multiple pockets concurrently, limiting GPU utilization.

Experiment suite. We evaluate the following configurations:

- **Invalid-molecule handling:** Vina-only Flow-GRPO with either *penalize* or *drop*, evaluated using paired seeds $\{40, 70, 200\}$. Longitudinal comparisons use the shared 0–600 optimization-iteration range.
- **Vina only:** $(w_{\mathrm{Vina}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (1, 0, 0)$.
- **SA only:** $(w_{\mathrm{Vina}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (0, 0, 1)$.
- **Vina–SA sweep:** $(w_{\mathrm{Vina}}, w_{\mathrm{SA}}) \in \{(0, 1), (0.25, 0.75), (0.5, 0.5), (0.75, 0.25), (1, 0)\}$, with $w_{\mathrm{QED}} = 0$.
- **Vina–QED:** $(w_{\mathrm{Vina}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (0.75, 0.25, 0)$.
- **Vina–QED–SA:** $(w_{\mathrm{Vina}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (0.50, 0.25, 0.25)$.

The reward-composition experiments use seeds $\{42, 70, 200\}$, the invalid-molecule penalization policy, and otherwise identical training hyperparameters. Table 2 reports the settings shared by all experiments.

Table 2: **Flow-GRPO training and sampling settings shared by all experiments.** Experiment-specific reward compositions, seed sets, and training horizons are described in the text.

Setting	Value
Pretrained initialization	FLOWR.ROOT v2.1
Fine-tuning data	SPINDR training split
Hardware per run	4 NVIDIA GB200 GPUs
Group size / global rollout size	16 / 512
Optimizer / learning rate	Adam / 10^{-4}
Updates / replay passes per collection round	2 / 1
Training / validation integration steps	50 / 100
Time grid / solver	linear / Euler
Coordinate sampler	SDE, $\sigma_c = 0.2$
Categorical sampler	uniform-sample, $\eta = 1$
Likelihood components	coordinates, atoms, charges, bonds
Size-invariant likelihood / KL	enabled / enabled
KL regularization	exact, $\beta = 0.005$
PPO / log-ratio clipping	$\epsilon = 0.2$ / $c_{\log} = 10$
Validation cadence	every 2 collection rounds

Evaluation and checkpoint selection. When Vina contributes to the reward, we use score-only Uni-Dock without pose optimization and a $20\,\text{Å} \times 20\,\text{Å} \times 20\,\text{Å}$ box centered on the generated ligand's center of mass. We convert the Vina score to a bounded reward using $r_{\mathrm{Vina}} = \mathrm{clip}(-\mathrm{Vina}/15, 0, 1)$, which saturates at 1 for scores at or below -15 kcal/mol and at 0 for scores at or above 0 kcal/mol. For each run, we select a checkpoint attaining the maximum scalarized validation reward and report every property from that same checkpoint.

D Evaluation methods and aggregation

D.1 Seed-level and aggregated results

Pocket-level aggregation. We first compute each molecular metric for every generated molecule. For each training seed and target pocket, we average the corresponding molecule-level values to obtain one pocket-level value.

Single-seed results. For each training seed, we compute the mean and sample standard deviation of its pocket-level values across all available target pockets. The seed-level results in this appendix therefore report a pocket-macro mean and the variation across pockets. Tables 4 and 6 report seed level results.

Multi-seed aggregated results. We first average the pocket-level values across the available training seeds for each pocket, giving every seed equal weight. We then compute the mean and sample standard deviation of these seed-aggregated values across pockets. Tables 1 and 5 report seed aggregated results.

BA W1 and BL W1 are distribution-level metrics and do not follow the same pocket-level decomposition. The appendix tables report their individual seed-level values, while the main tables report their arithmetic mean across the three seeds.

D.2 Seed-level test-set results

We train each Flow-GRPO configuration with three independent seeds (42, 70, and 200). While the main analysis reports metrics aggregated across seeds as described in Appendix D.1, Tables 4 and 6 report results separately for each seed. Each seed uses its SPINDR validation-selected checkpoint for evaluation on both SPINDR and CrossDocked2020 test sets.

D.3 Molecular-feature extraction and correlation methodology

We compute heavy-atom, rotatable-bond, total-ring, aromatic-ring, and potential stereocenter counts, and Crippen logP using RDKit on sanitized ligands after hydrogen removal, including unassigned potential stereocenters. Within-pocket Spearman correlations are computed among molecules separately for each seed, then averaged equally across pockets and seeds. For change correlations, we subtract pretrained pocket means from aligned pocket means, correlate feature and score changes across matched pockets separately for each seed and reward composition, and average coefficients across seeds.

E Additional results on SPINDR

E.1 Baseline models

Table 3 reports the performance evaluation of PILOT (Cremer et al., 2024), MolJO (Qiu et al., 2025) and MolForm (Zhang & Zhang, 2025) models on SPINDR test targets. MolJo and PILOT were aligned for binding affinity and SA, while MolForm is aligned for binding affinity only. For PILOT, metrics were taken from FLOWR.ROOT (Cremer et al., 2025b) paper. For the other model, we used checkpoints and sampling pipelines shared by the corresponding authors to sample 100 molecules per SPINDR target and evaluated the models using our evaluation pipeline. These models were trained on CrossDocked2020 training set.

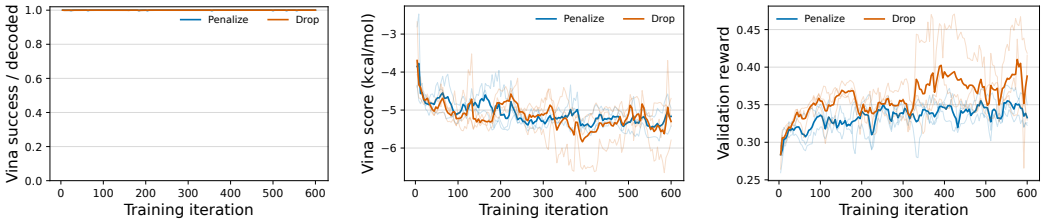
E.2 Validation logs for RDKit invalid molecules handling policies

Figure 6 complements the RDKit-validity trajectory in Figure 2A. Both invalid-decoding policies improve the optimized Vina-based quantities during post-training, but the improvement is not equivalent to preserving valid generation yield. In particular, Figure 6A shows that Vina evaluation succeeds for nearly every molecule that has already decoded, while the Vina-score curve in Figure 6B is explicitly validation-only and success-conditioned: it averages only attempts with a successful Vina evaluation

Table 3: **Baseline models results on SPINDR.** Using the sampling code shared by the authors, we sampled 100 molecules per SPINDR test pocket. Vina is score-only and Vina min. is pose-optimized. For PILOT, metrics were taken from FLOWR.ROOT (Cremer et al., 2025b) paper. MolJo and PILOT were aligned for binding affinity and SA, while MolForm is aligned for binding affinity only. Values are means ± standard deviations across available pockets, except BA W1 and BL W1, which summarize distribution distances. Arrows indicate the preferred direction. NA:note available

Model/Method	PB-val. ↑	BA W1 ↓	BL W1 (10^{-2}) ↓	Vina ↓	Vina min. ↓	SA ↑	QED ↑	Diversity ↑	Uniqueness ↑
MolJO	$0.96_{\pm 0.06}$	3.98	0.432	$-6.57_{\pm 2.09}$	$-7.57_{\pm 2.11}$	$0.83_{\pm 0.06}$	$0.54_{\pm 0.06}$	$0.81_{\pm 0.07}$	$0.82_{\pm 0.14}$
MolForm-DPO	$0.43_{\pm 0.16}$	7.40	0.387	$-5.25_{\pm 2.43}$	$-6.97_{\pm 2.14}$	$0.70_{\pm 0.09}$	$0.53_{\pm 0.05}$	$0.81_{\pm 0.06}$	$0.98_{\pm 0.04}$
PILOT	$0.79_{\pm 0.21}$	1.82	0.420	$-6.30_{\pm 0.96}$	$-6.68_{\pm 1.07}$	NA	NA	NA	NA

and therefore excludes failed decodes and docking failures from its denominator. Each logged point in all three panels summarizes the complete set of validation pockets for that seed.



(A) Vina success among decoded molecules

(B) Vina score among successful evaluations

(C) Mean validation reward

Figure 6: **Validation logs for Vina-only Flow-GRPO under *penalize* (blue) and *drop* (orange) assignment of tensor-to-RDKit decode failures.** **(A)** Fraction of decoded molecules for which Vina succeeds. **(B)** Success-conditioned mean score-only Vina score. **(C)** Mean validation reward. Each point summarizes the complete validation-pocket set for one seed. Thin curves show individual seed runs and thick curves are their arithmetic mean. Near-unity conditional Vina success in panel A indicates that the loss of valid yield in Figure 2A arises primarily during RDKit decoding rather than Vina evaluation.

E.3 Validation logs for the Vina–SA reward weights sweep

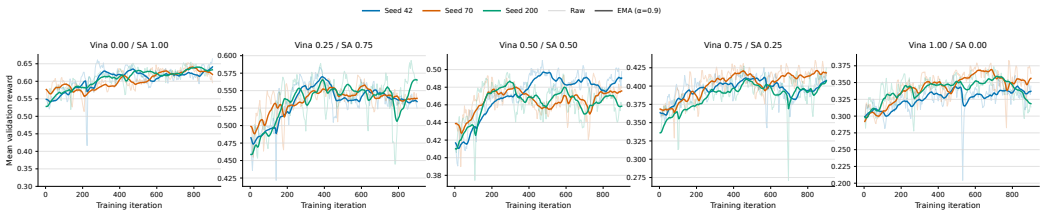


Figure 7: **Validation-reward logs for the Vina–SA weight sweep.** From left to right, panels use $(w_{\text{Vina}}, w_{\text{SA}}) \in \{(0,1), (0.25, 0.75), (0.5, 0.5), (0.75, 0.25), (1,0)\}$. Blue, orange, and green denote seeds 42, 70, and 200. Thin curves are the logged rewards. Thick curves show an exponential moving average $s_t = 0.9s_{t-1} + 0.1x_t$. Because the scalarized objective differs among panels, absolute reward levels should be interpreted with the corresponding weights.

To complement the endpoint comparison in Figure 3, we report the complete validation trajectories for the Vina–synthetic-accessibility (SA) reward-weight sweep. The sweep contains five complementary reward-weight pairs and three independent random seeds per pair. Figures 7 and 8 show the scalarized validation reward and the percentage of generation attempts that decode to molecules, respectively.

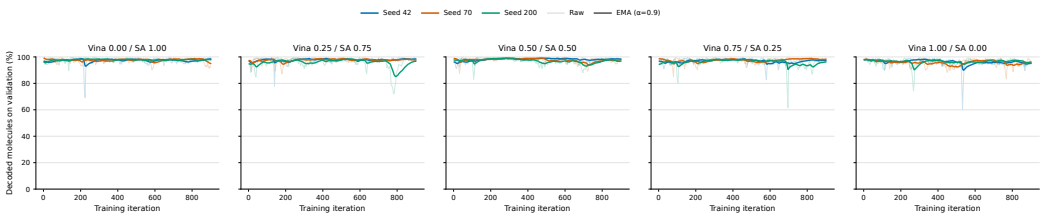


Figure 8: **RDKit decoding success on the SPINDR validation set during the Vina–SA weight sweep.** Panels and colors follow Figure 7. The plotted quantity is the percentage of generation attempts that decode to RDKit molecules; Thin curves are the logged values, and thick curves show the exponential moving average $s_t = 0.9s_{t-1} + 0.1x_t$.

E.4 Matched-pocket comparisons of single-objective alignment

Figure 9 compares SA-only and Vina-only alignment with pretrained FLOWR.ROOT. We average scores within each pocket, and then average scores across seeds. At the 0.05 significance level, both models improve their targeted score while worsening the untargeted score relative to pretrained FLOWR.ROOT.

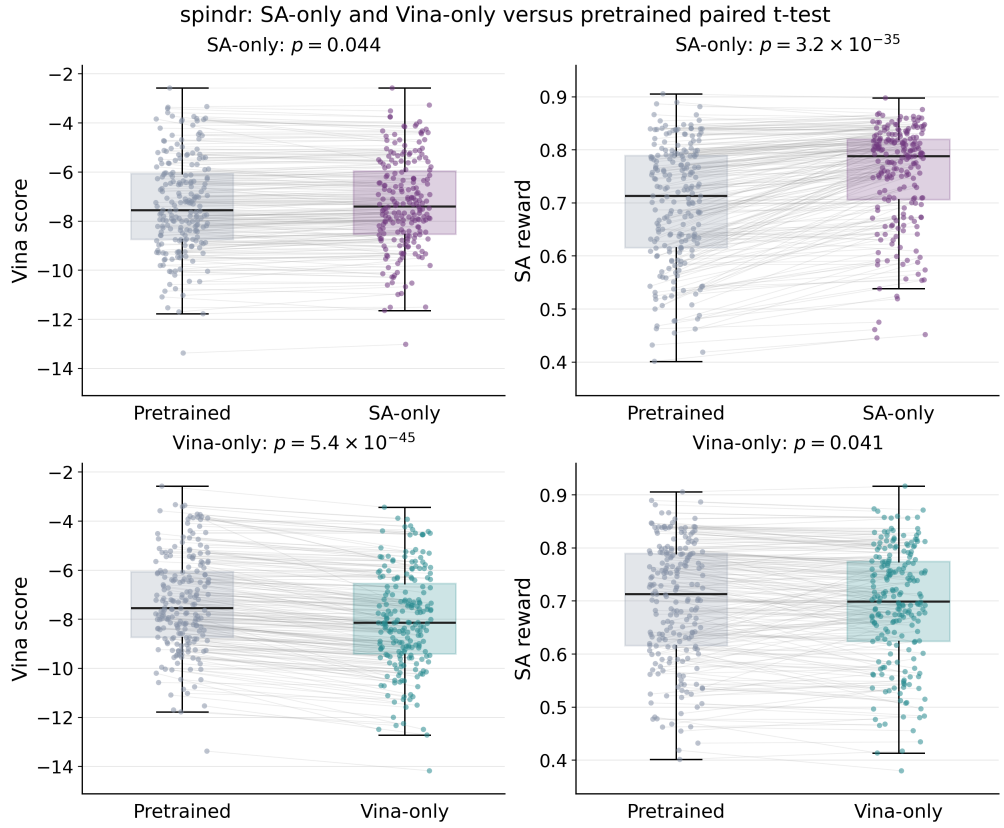


Figure 9: **Matched-pocket comparisons of single-objective Flow-GRPO with pretrained FLOWR.ROOT on SPINDR.** Rows show SA-only and Vina-only alignment. Columns show score-only Vina and SA comparisons. Points are seed-averaged pocket scores, and lines connect the same pocket across models. Boxes show the median and interquartile range, with whiskers extending to the most extreme observations within 1.5 interquartile ranges. Titles show two-sided paired t-test p-values for matched pockets. Lower Vina and higher SA are preferred.

E.5 Molecular-feature analyses on SPINDR

Figure 10 reports molecule-level associations within pockets for the $w_{\text{Vina}} = 0.75$ model discussed in the main text. These correlations describe variation among ligands generated for the same pocket, whereas Figure 5 relates changes in pocket means after alignment. Figure 11 provides the latter analysis for the remaining reward weights.

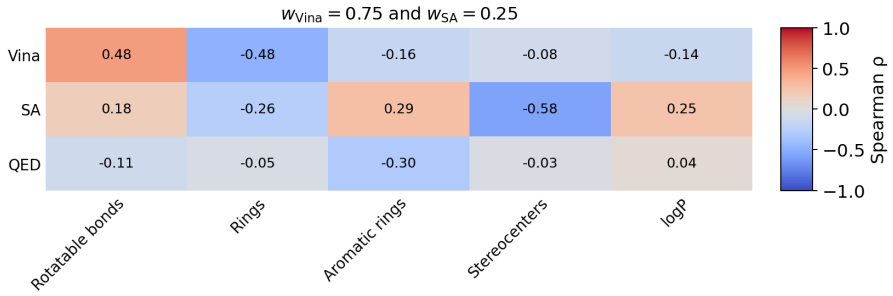


Figure 10: **Within-pocket descriptor–score correlations for $w_{\text{Vina}} = 0.75$ on SPINDR.** Molecule-level Spearman correlations are computed within each pocket separately for each seed, averaged equally across pockets, and then averaged across seeds. Rows show score-only Vina, SA, and QED, and columns show molecular descriptors. Annotations report mean coefficients. Lower Vina and higher SA or QED are preferred.

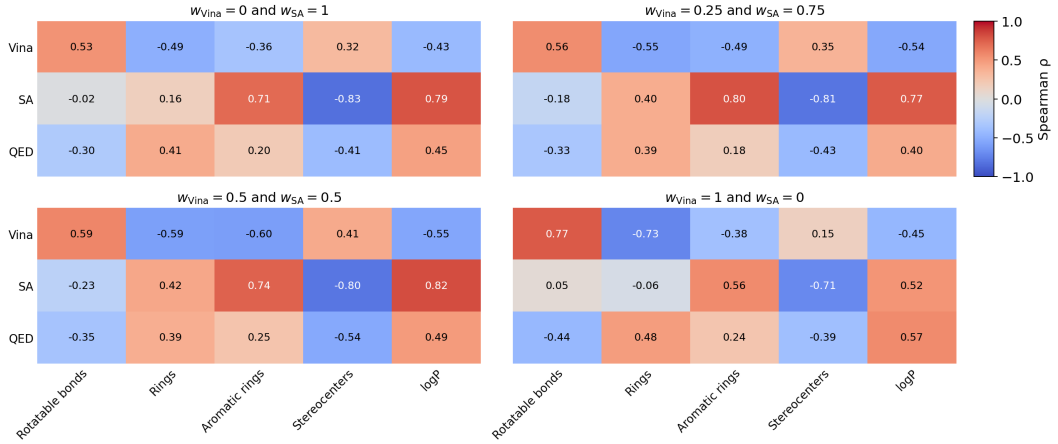


Figure 11: **Correlations of descriptor and score changes for the remaining SPINDR reward weights.** Panels show $w_{\text{Vina}} \in \{0, 0.25, 0.50, 1\}$, with $w_{\text{SA}} = 1 - w_{\text{Vina}}$. Changes are aligned minus pretrained pocket means; Spearman correlations are computed across matched pockets separately for each seed and then averaged across seeds. Rows show score-only Vina, SA, and QED; columns show molecular descriptors. The $w_{\text{Vina}} = 0.75$ result is shown in Figure 5.

E.6 Seed-level results on SPINDR

Seed level results on SPINDR test set are shown in Table 4.

F Additional results on CrossDocked2020

F.1 CrossDocked2020 evaluation and baselines

Dataset and baselines. CrossDocked2020 (Francoeur et al., 2020) contains ligand poses cross-docked into structurally similar protein pockets. We evaluate the SPINDR-selected Flow-GRPO

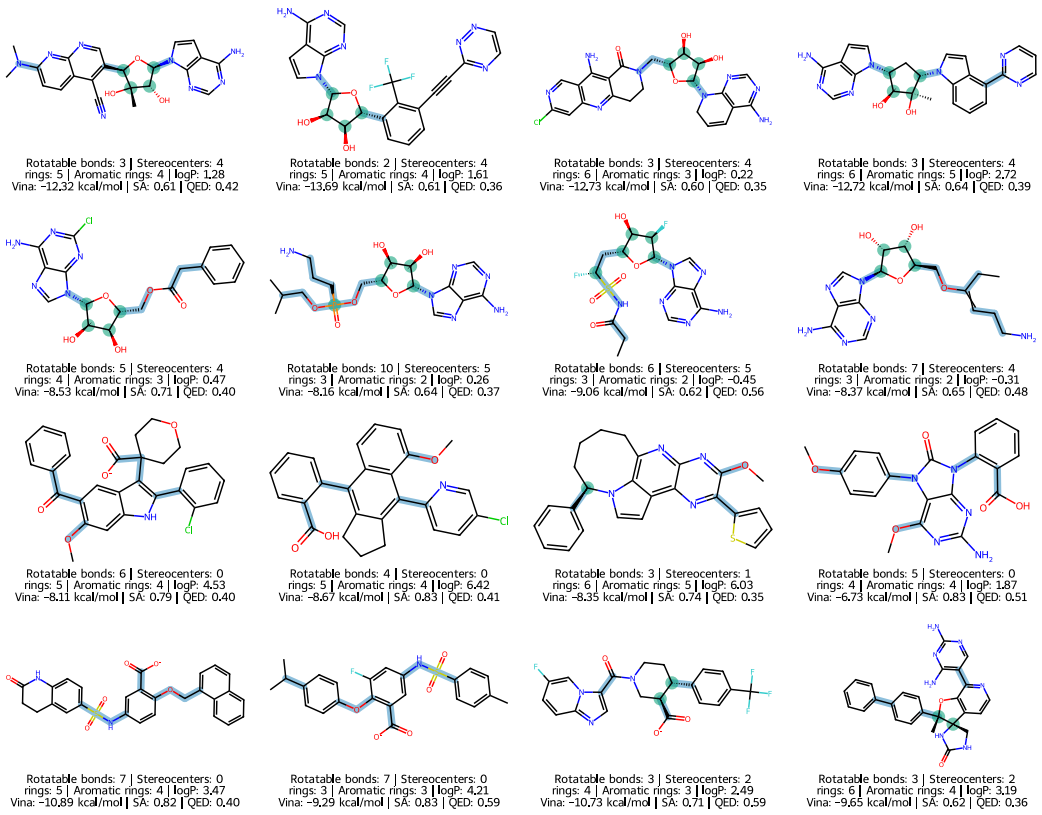


Figure 12: **Illustrative generated ligands conditioned on SPINDR test targets.** The ligands were generated using GRPO model aligned for Vina and SA with $w_{Vina} = 0.75$ and $w_{SA} = 0.25$. Each molecule is annotated with its rotatable-bond count, stereocenter count, total and aromatic ring counts, Crippen logP, Vina score (kcal/mol), SA, and QED. Blue highlights mark rotatable bonds, and green highlights mark stereocenters.

Table 4: **Seed-level performance on the SPINDR test set.** Rows with `Seed=None` are the raw test-set and pretrained references; the raw test set is excluded from ranking. Each Flow-GRPO row uses the SPINDR validation-selected checkpoint from one independent training seed. RDKit-val., PB-val., Vina, Vina min., SA, QED, uniqueness, and novelty are reported as means with sample SD across 225 pocket means. BA W1 and diversity are seed-level distribution summaries; BL W1 is the seed-integrated summary in units of 10^{-2}. In the method labels, w_{Vina} is the Vina weight: Vina–SA uses $w_{\mathrm{SA}} = 1 - w_{\mathrm{V}}$, Vina–QED uses $(w_{\mathrm{V}}, w_{\mathrm{QED}}) = (0.75, 0.25)$, and Vina–QED–SA uses $(w_{\mathrm{V}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (0.50, 0.25, 0.25)$. Vina is score-only; Vina min. includes pose optimization. Arrows indicate the preferred direction. Bold marks the best central value, and underlining marks the second- and third-best distinct values; ties share formatting. Values are reported as the mean $\pm$ standard deviation across pockets.

Model	Metrics Aligned	Seed	RDKit-val. ↑	PB-val. ↑	BA W1 ↓	BL W1 (10^{-2}) ↓	Vina ↓	Vina min. ↓	SA ↑	QED ↑	Diversity ↑	Uniqueness ↑	Novelty ↑
Test-set reference													
Test set	–	None	$1.00_{\pm 0.00}$	$0.99_{\pm 0.09}$	–	–	$-7.63_{\pm 2.18}$	$-7.88_{\pm 2.10}$	$0.67_{\pm 0.12}$	$0.49_{\pm 0.22}$	–	–	–
Pretrained baseline													
FLOWR.ROOT Pretrained		None	$0.98_{\pm 0.04}$	$0.96_{\pm 0.05}$	**0.68**	0.077	$-7.35_{\pm 1.93}$	$-7.73_{\pm 1.74}$	$0.70_{\pm 0.11}$	$0.48_{\pm 0.19}$	$0.53_{\pm 0.18}$	$0.91_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO													
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	42	$0.92_{\pm 0.08}$	$0.83_{\pm 0.14}$	1.90	0.125	$-7.17_{\pm 1.88}$	$-7.57_{\pm 1.68}$	$0.75_{\pm 0.09}$	$0.51_{\pm 0.19}$	$0.59_{\pm 0.14}$	$0.93_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	70	$0.97_{\pm 0.04}$	$0.93_{\pm 0.08}$	1.92	0.127	$-7.23_{\pm 1.81}$	$-7.66_{\pm 1.68}$	$0.75_{\pm 0.10}$	$0.49_{\pm 0.18}$	$0.56_{\pm 0.15}$	$0.93_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	200	$0.97_{\pm 0.05}$	$0.95_{\pm 0.06}$	1.70	0.097	$-7.46_{\pm 1.81}$	$-7.79_{\pm 1.70}$	$0.76_{\pm 0.10}$	$0.52_{\pm 0.18}$	$0.59_{\pm 0.15}$	$0.93_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	42	$0.97_{\pm 0.05}$	$0.95_{\pm 0.06}$	2.32	0.120	$-7.96_{\pm 2.04}$	$-8.35_{\pm 1.90}$	$\mathbf{0.77_{\pm 0.11}}$	$0.50_{\pm 0.17}$	$\mathbf{0.61_{\pm 0.15}}$	$0.92_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	70	$0.97_{\pm 0.04}$	$0.93_{\pm 0.09}$	1.64	0.104	$-7.61_{\pm 1.87}$	$-7.92_{\pm 1.75}$	$0.75_{\pm 0.10}$	$0.51_{\pm 0.19}$	$0.58_{\pm 0.15}$	$0.92_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	200	$0.94_{\pm 0.09}$	$0.92_{\pm 0.12}$	2.14	0.125	$-7.48_{\pm 1.83}$	$-7.82_{\pm 1.69}$	$\mathbf{0.77_{\pm 0.07}}$	$0.54_{\pm 0.18}$	$0.60_{\pm 0.14}$	$0.94_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	42	$0.98_{\pm 0.04}$	$\mathbf{0.97_{\pm 0.05}}$	1.61	0.094	$-7.66_{\pm 1.86}$	$-7.94_{\pm 1.73}$	$0.76_{\pm 0.10}$	$0.52_{\pm 0.18}$	$0.59_{\pm 0.15}$	$0.92_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	70	$0.98_{\pm 0.04}$	$0.93_{\pm 0.12}$	1.83	0.120	$-7.81_{\pm 1.93}$	$-8.16_{\pm 1.82}$	$0.75_{\pm 0.10}$	$0.51_{\pm 0.18}$	$0.58_{\pm 0.14}$	$0.93_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	200	$0.96_{\pm 0.08}$	$0.95_{\pm 0.09}$	1.69	0.148	$-7.89_{\pm 1.99}$	$-8.21_{\pm 1.86}$	$0.75_{\pm 0.10}$	$0.52_{\pm 0.17}$	$0.57_{\pm 0.16}$	$0.92_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	42	$\mathbf{0.99_{\pm 0.02}}$	$0.96_{\pm 0.06}$	1.39	0.099	$-7.72_{\pm 2.00}$	$-8.05_{\pm 1.86}$	$0.73_{\pm 0.11}$	$0.51_{\pm 0.19}$	$0.56_{\pm 0.15}$	$0.92_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	70	$0.98_{\pm 0.02}$	$\mathbf{0.97_{\pm 0.04}}$	1.26	0.118	$-8.03_{\pm 2.01}$	$-8.34_{\pm 1.90}$	$0.72_{\pm 0.10}$	$0.55_{\pm 0.17}$	$0.56_{\pm 0.15}$	$0.94_{\pm 0.13}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	200	$\mathbf{0.99_{\pm 0.02}}$	$\mathbf{0.97_{\pm 0.04}}$	0.99	**0.073**	$-7.99_{\pm 2.03}$	$-8.27_{\pm 1.92}$	$0.72_{\pm 0.11}$	$0.52_{\pm 0.17}$	$0.55_{\pm 0.15}$	$0.92_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	42	$0.98_{\pm 0.03}$	$0.95_{\pm 0.05}$	1.21	0.111	$-7.93_{\pm 1.99}$	$-8.26_{\pm 1.88}$	$0.68_{\pm 0.11}$	$0.51_{\pm 0.17}$	$0.54_{\pm 0.15}$	$0.93_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	70	$\mathbf{0.99_{\pm 0.01}}$	$\mathbf{0.97_{\pm 0.06}}$	1.28	0.136	$\mathbf{-8.24_{\pm 2.02}}$	$\mathbf{-8.52_{\pm 1.95}}$	$0.67_{\pm 0.12}$	$0.56_{\pm 0.18}$	$0.56_{\pm 0.15}$	$0.94_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	200	$0.97_{\pm 0.03}$	$0.94_{\pm 0.06}$	1.82	0.118	$-8.06_{\pm 1.99}$	$-8.39_{\pm 1.88}$	$0.71_{\pm 0.10}$	$0.51_{\pm 0.17}$	$0.57_{\pm 0.14}$	$\mathbf{0.95_{\pm 0.13}}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED	42	$0.98_{\pm 0.04}$	$0.96_{\pm 0.06}$	0.84	0.105	$-7.84_{\pm 1.95}$	$-8.13_{\pm 1.84}$	$0.70_{\pm 0.11}$	$0.58_{\pm 0.18}$	$0.56_{\pm 0.16}$	$0.92_{\pm 0.15}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED	70	$\mathbf{0.99_{\pm 0.02}}$	$\mathbf{0.97_{\pm 0.04}}$	1.15	0.130	$-7.81_{\pm 1.93}$	$-8.17_{\pm 1.81}$	$0.69_{\pm 0.11}$	$0.58_{\pm 0.17}$	$0.57_{\pm 0.15}$	$\mathbf{0.95_{\pm 0.13}}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED	200	$\mathbf{0.99_{\pm 0.02}}$	$\mathbf{0.97_{\pm 0.04}}$	1.18	0.136	$-7.99_{\pm 2.05}$	$-8.34_{\pm 1.92}$	$0.66_{\pm 0.10}$	$\mathbf{0.61_{\pm 0.16}}$	$0.59_{\pm 0.13}$	$\mathbf{0.95_{\pm 0.12}}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED–SA	42	$0.98_{\pm 0.03}$	$0.96_{\pm 0.05}$	1.47	0.136	$-7.70_{\pm 1.92}$	$-8.04_{\pm 1.78}$	$0.73_{\pm 0.09}$	$0.56_{\pm 0.17}$	$0.58_{\pm 0.15}$	$0.94_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED–SA	70	$0.97_{\pm 0.04}$	$0.95_{\pm 0.05}$	1.36	0.118	$-8.00_{\pm 2.03}$	$-8.30_{\pm 1.91}$	$0.73_{\pm 0.11}$	$0.56_{\pm 0.18}$	$0.57_{\pm 0.15}$	$0.94_{\pm 0.12}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO	Vina–QED–SA	200	$0.97_{\pm 0.04}$	$0.96_{\pm 0.05}$	1.45	0.153	$-7.65_{\pm 1.85}$	$-7.93_{\pm 1.74}$	$0.74_{\pm 0.10}$	$0.56_{\pm 0.18}$	$0.57_{\pm 0.17}$	$0.93_{\pm 0.14}$	$\mathbf{1.00_{\pm 0.00}}$

checkpoints on the standard 100 test pockets without additional fine-tuning or checkpoint selection. The comparison covers the unguided generators Pocket2Mol (Peng et al., 2022), DiffSBDD (Schneuing et al., 2024), TargetDiff (Guan et al., 2023), DrugFlow (Schneuing et al., 2025), FLOWR.ROOT (Cremer et al., 2025b), MolJO (Qiu et al., 2025), DrugFlow with DPO (Schneuing et al., 2025), and AliDiff (Gu et al., 2024).

F.2 Molecular-feature analyses on CrossDocked2020

We apply the SPINDR feature and correlation analyses to CrossDocked2020 using the same selected checkpoints. Mean heavy-atom count remains nearly constant across weights; Figure 13 shows the other molecular descriptors. From SA-only to Vina-only alignment, mean rotatable-bond count decreases from approximately 4.3 to 3.0, while mean ring count increases from approximately 2.7 to 3.5. Aromatic-ring count and logP exceed the pretrained baseline and peak at intermediate weights. Stereocenter count reaches its minimum at $w_{Vina} = 0.25$ and then increases, remaining below the pretrained value. These directions agree with SPINDR.

Figure 14 reports correlations of changes from pretrained for all reward weights. For $w_{\mathrm{Vina}} = 0.75$, reductions in rotatable bonds and increases in total rings accompany Vina improvements ($\rho = 0.33$ and -0.42). Aromatic-ring and logP changes also correlate negatively with Vina changes ($\rho = -0.26$ and -0.30). SA improvements associate with fewer potential stereocenters ($\rho = -0.66$) and increases in aromatic rings and logP ($\rho = 0.64$ and 0.60). The other panels provide the corresponding results for the remaining reward weights.

F.3 Seed-level results on CrossDocked2020

Seed level results on CrossDocked2020 test set are shown in Table 6.

Table 5: **Metrics evaluation on CrossDocked2020 test set.** Flow-GRPO models were fine-tuned on SPINDR. All other models except FLOWR.ROOT were either trained and/or finetuned on CrossDocked2020 training set. Other preference-aligned methods optimize different objectives: MolJO targets synthetic accessibility and affinity, DrugFlow-DPO jointly optimizes QED, SA, Vina, and REOS, and AliDiff optimizes Vina. To ensure a consistent comparison, we recompute all metrics using our evaluation pipeline. We evaluate released samples for the prior methods, except for FLOWR.ROOT and DrugFlow-DPO, for which we generate samples ourselves. PB-validity is not reported for Pocket2Mol because its sampling procedure generates molecules until 100 RDKit valid outputs are obtained, biasing PB-validity comparison with other methods. Values are reported as the mean ± standard deviation across pockets. Arrows indicate the preferred direction. Bold marks the best value, and underlining marks the second- and third-best distinct values; the test-set reference is excluded from ranking.

Model/Method	PB-val. ↑	BA W1 ↓	BL W1 (10^{-2}) ↓	Vina ↓	Vina min. ↓	SA ↑	QED ↑	Diversity ↑	Uniqueness ↑	Novelty ↑
Test-set reference										
Test set	$0.94_{\pm 0.24}$	–	–	$-6.53_{\pm 2.74}$	$-7.18_{\pm 2.20}$	$0.73_{\pm 0.14}$	$0.48_{\pm 0.21}$	–	–	–
Generative baselines										
Pocket2Mol	–	2.28	0.47	$-4.24_{\pm 1.30}$	$-6.11_{\pm 1.33}$	$0.81_{\pm 0.05}$	$0.59_{\pm 0.09}$	$\mathbf{0.84_{\pm 0.05}}$	$\mathbf{1.00_{\pm 0.00}}$	$\mathbf{1.00_{\pm 0.00}}$
DiffSBDD	$0.38_{\pm 0.13}$	6.99	0.33	$-2.31_{\pm 5.84}$	$-7.24_{\pm 1.43}$	$0.58_{\pm 0.05}$	$0.47_{\pm 0.09}$	$0.73_{\pm 0.07}$	$\mathbf{1.00_{\pm 0.00}}$	$\mathbf{1.00_{\pm 0.00}}$
TargetDiff	$0.45_{\pm 0.15}$	3.19	0.21	$-6.13_{\pm 2.17}$	$-7.66_{\pm 1.75}$	$0.58_{\pm 0.07}$	$0.48_{\pm 0.11}$	$0.72_{\pm 0.09}$	$\mathbf{1.00_{\pm 0.01}}$	$\mathbf{1.00_{\pm 0.00}}$
DrugFlow	$0.64_{\pm 0.19}$	2.06	$\underline{0.13}$	$-5.02_{\pm 2.50}$	$-6.85_{\pm 1.73}$	$0.71_{\pm 0.09}$	$0.51_{\pm 0.15}$	$0.73_{\pm 0.12}$	$\underline{0.96_{\pm 0.09}}$	$\underline{0.98_{\pm 0.06}}$
FLOWR.ROOT	$\underline{0.90_{\pm 0.13}}$	**0.62**	**0.105**	$-5.86_{\pm 2.25}$	$-7.33_{\pm 2.01}$	$0.71_{\pm 0.11}$	$0.50_{\pm 0.16}$	$0.60_{\pm 0.18}$	$0.88_{\pm 0.18}$	$\underline{0.99_{\pm 0.04}}$
MolJo	$0.85_{\pm 0.12}$	3.79	0.66	$\mathbf{-8.25_{\pm 2.45}}$	$\mathbf{-9.07_{\pm 2.40}}$	$0.77_{\pm 0.06}$	$0.55_{\pm 0.08}$	$\underline{0.75_{\pm 0.09}}$	$0.92_{\pm 0.12}$	$\mathbf{1.00_{\pm 0.00}}$
Post-training baselines										
DrugFlow-DPO	$0.50_{\pm 0.15}$	4.12	0.54	$-5.77_{\pm 2.36}$	$-7.51_{\pm 1.82}$	$\mathbf{0.85_{\pm 0.05}}$	$\mathbf{0.66_{\pm 0.07}}$	$\underline{0.79_{\pm 0.05}}$	$0.95_{\pm 0.08}$	$\mathbf{1.00_{\pm 0.00}}$
AliDiff-E2PO	$0.24_{\pm 0.18}$	6.46	$\underline{0.13}$	$\underline{-7.51_{\pm 2.65}}$	$\underline{-8.42_{\pm 2.13}}$	$0.57_{\pm 0.08}$	$0.50_{\pm 0.10}$	$0.74_{\pm 0.07}$	$\underline{0.99_{\pm 0.03}}$	$\mathbf{1.00_{\pm 0.00}}$
Flow-GRPO (ours)										
Flow-GRPO (Vina–SA, $w_{Vina} = 0$)	$0.83_{\pm 0.15}$	1.99	0.131	$-5.77_{\pm 2.16}$	$-7.29_{\pm 1.98}$	$0.78_{\pm 0.08}$	$0.54_{\pm 0.15}$	$0.66_{\pm 0.14}$	$0.90_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.03}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.25$)	$0.87_{\pm 0.14}$	2.20	0.144	$-6.16_{\pm 2.31}$	$-7.62_{\pm 2.06}$	$\underline{0.79_{\pm 0.08}}$	$0.55_{\pm 0.15}$	$0.65_{\pm 0.14}$	$0.88_{\pm 0.19}$	$\underline{0.99_{\pm 0.04}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.50$)	$\underline{0.90_{\pm 0.11}}$	1.79	0.143	$-6.25_{\pm 2.27}$	$-7.63_{\pm 2.08}$	$0.78_{\pm 0.09}$	$0.54_{\pm 0.15}$	$0.65_{\pm 0.14}$	$0.88_{\pm 0.19}$	$\mathbf{1.00_{\pm 0.02}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 0.75$)	$\mathbf{0.93_{\pm 0.10}}$	$\underline{1.35}$	$\underline{0.111}$	$-6.57_{\pm 2.40}$	$-7.71_{\pm 2.18}$	$0.75_{\pm 0.11}$	$0.55_{\pm 0.14}$	$0.63_{\pm 0.15}$	$0.88_{\pm 0.18}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO (Vina–SA, $w_{Vina} = 1$)	$0.91_{\pm 0.11}$	1.39	0.132	$\underline{-6.61_{\pm 2.40}}$	$\underline{-7.85_{\pm 2.23}}$	$0.72_{\pm 0.11}$	$0.55_{\pm 0.14}$	$0.64_{\pm 0.14}$	$0.90_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO (Vina–QED)	$\underline{0.91_{\pm 0.13}}$	$\underline{1.30}$	0.176	$-6.30_{\pm 2.35}$	$-7.68_{\pm 2.16}$	$0.72_{\pm 0.11}$	$\underline{0.60_{\pm 0.14}}$	$0.65_{\pm 0.14}$	$0.91_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO (Vina–QED–SA)	$\underline{0.91_{\pm 0.11}}$	1.50	0.155	$-6.34_{\pm 2.36}$	$-7.62_{\pm 2.14}$	$0.76_{\pm 0.10}$	$0.57_{\pm 0.14}$	$0.65_{\pm 0.15}$	$0.90_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.02}}$

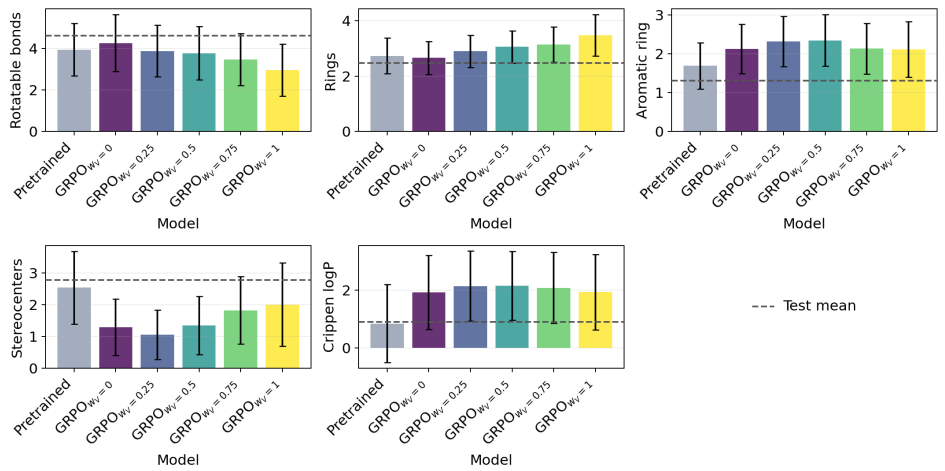


Figure 13: **Molecular features on CrossDocked2020.** Bars show rotatable-bond, total-ring, aromatic-ring, and potential stereocenter counts, and Crippen logP for pretrained FLOWR.ROOT and Flow-GRPO. GRPO subscripts denote Vina reward weights. Bars average molecule-level values within seed and pocket, seeds within pocket, and then pockets. Error bars show molecule-level standard deviations averaged over seeds and pockets. Dashed horizontal lines mark matched test-ligand means.

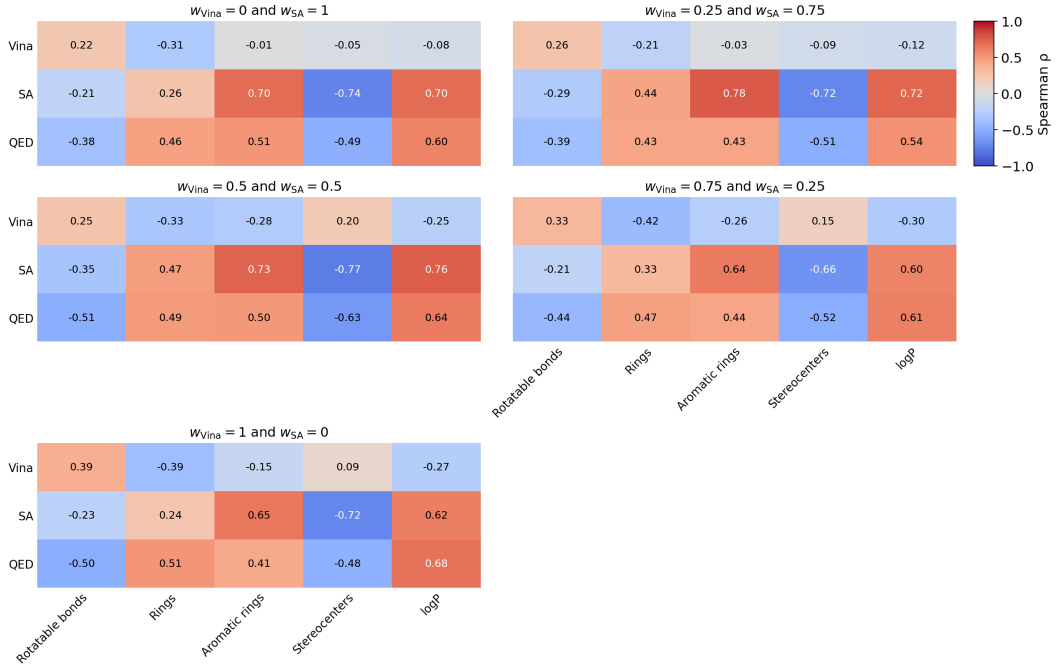


Figure 14: **Correlations of descriptor and score changes across reward weights on CrossDocked2020.** Panels identify the Vina and SA reward weights. Changes are aligned minus pretrained pocket means; Spearman correlations are computed across matched pockets separately for each training seed, then averaged across seeds. Within each panel, rows show score-only Vina, SA, and QED, and columns show molecular descriptors. Negative Vina changes and positive SA or QED changes indicate improvements. Annotations report mean coefficients.

Table 6: **Seed-level transfer performance on the CrossDocked2020 test set.** Rows with `Seed=None` are the raw test-set and pretrained references; the raw test set is excluded from ranking. Each Flow-GRPO row uses the SPINDR validation-selected checkpoint from one independent training seed, without additional CrossDocked fine-tuning or selection. RDKit-val., PB-val., Vina, Vina min., SA, QED, diversity, uniqueness, and novelty are reported as means with sample SD across 100 pocket means. BA W1 and BL W1 are seed-level distribution summaries. In the method labels, w_{Vina} is the Vina weight: Vina–SA uses $w_{\mathrm{SA}} = 1 - w_{\mathrm{V}}$, Vina–QED uses $(w_{\mathrm{V}}, w_{\mathrm{QED}}) = (0.75, 0.25)$, and Vina–QED–SA uses $(w_{\mathrm{V}}, w_{\mathrm{QED}}, w_{\mathrm{SA}}) = (0.50, 0.25, 0.25)$. Vina is score-only; Vina min. includes pose optimization. Arrows indicate the preferred direction. Bold marks the best central value, and underlining marks the second- and third-best distinct values; ties share formatting. Values are reported as the mean ± standard deviation across pockets.

Model	Metrics Aligned	Seed	RDKit-val. ↑	PB-val. ↑	BA W1 ↓	BL W1 (10^{-2}) ↓	Vina ↓	Vina min. ↓	SA ↑	QED ↑	Diversity ↑	Uniqueness ↑	Novelty ↑
Test-set reference													
Test set	–	None	$1.00_{\pm 0.00}$	$0.94_{\pm 0.24}$	–	–	$-6.53_{\pm 2.74}$	$-7.18_{\pm 2.20}$	$0.73_{\pm 0.14}$	$0.48_{\pm 0.21}$	––	––	––
Pretrained baseline													
FLOWR.ROOT	Pretrained	None	$0.98_{\pm 0.03}$	$0.90_{\pm 0.13}$	**0.62**	0.105	$-5.86_{\pm 2.25}$	$-7.33_{\pm 2.01}$	$0.71_{\pm 0.11}$	$0.50_{\pm 0.16}$	$0.60_{\pm 0.18}$	$0.88_{\pm 0.18}$	$0.99_{\pm 0.04}$
Flow-GRPO													
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	42	$0.93_{\pm 0.08}$	$0.78_{\pm 0.19}$	1.93	0.113	$-5.76_{\pm 2.00}$	$-7.23_{\pm 1.97}$	$0.77_{\pm 0.09}$	$0.54_{\pm 0.16}$	$0.66_{\pm 0.15}$	$0.91_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	70	$0.97_{\pm 0.03}$	$0.84_{\pm 0.17}$	2.18	0.164	$-5.61_{\pm 2.19}$	$-7.28_{\pm 2.00}$	$0.77_{\pm 0.08}$	$0.52_{\pm 0.15}$	$0.65_{\pm 0.15}$	$0.91_{\pm 0.18}$	$0.99_{\pm 0.04}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0$)	200	$0.97_{\pm 0.03}$	$0.87_{\pm 0.14}$	1.85	0.115	$-5.84_{\pm 2.19}$	$-7.36_{\pm 1.98}$	$0.78_{\pm 0.09}$	$0.55_{\pm 0.15}$	$\mathbf{0.67_{\pm 0.15}}$	$0.90_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.03}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	42	$0.97_{\pm 0.03}$	$0.87_{\pm 0.17}$	2.45	0.165	$-6.29_{\pm 2.41}$	$-7.87_{\pm 2.12}$	$\mathbf{0.80_{\pm 0.10}}$	$0.53_{\pm 0.14}$	$0.65_{\pm 0.14}$	$0.84_{\pm 0.22}$	$0.99_{\pm 0.04}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	70	$0.97_{\pm 0.03}$	$0.86_{\pm 0.16}$	1.81	0.113	$-6.01_{\pm 2.37}$	$-7.51_{\pm 2.12}$	$0.77_{\pm 0.09}$	$0.55_{\pm 0.16}$	$0.64_{\pm 0.17}$	$0.89_{\pm 0.19}$	$0.99_{\pm 0.04}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.25$)	200	$0.94_{\pm 0.08}$	$0.87_{\pm 0.14}$	2.35	0.154	$-5.95_{\pm 1.91}$	$-7.47_{\pm 1.98}$	$\mathbf{0.80_{\pm 0.07}}$	$0.56_{\pm 0.15}$	$\mathbf{0.67_{\pm 0.14}}$	$0.90_{\pm 0.18}$	$\mathbf{1.00_{\pm 0.03}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	42	$0.98_{\pm 0.03}$	$\mathbf{0.94_{\pm 0.07}}$	1.75	0.111	$-6.29_{\pm 2.03}$	$-7.55_{\pm 2.04}$	$0.79_{\pm 0.09}$	$0.55_{\pm 0.16}$	$0.64_{\pm 0.16}$	$0.86_{\pm 0.21}$	$0.99_{\pm 0.04}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	70	$0.98_{\pm 0.02}$	$0.85_{\pm 0.16}$	2.03	0.159	$-6.06_{\pm 2.40}$	$-7.69_{\pm 2.13}$	$0.77_{\pm 0.09}$	$0.54_{\pm 0.15}$	$0.65_{\pm 0.14}$	$0.89_{\pm 0.18}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.50$)	200	$0.97_{\pm 0.06}$	$0.91_{\pm 0.13}$	1.61	0.160	$-6.31_{\pm 2.27}$	$-7.66_{\pm 2.10}$	$0.77_{\pm 0.10}$	$0.55_{\pm 0.14}$	$0.65_{\pm 0.14}$	$0.89_{\pm 0.19}$	$\mathbf{1.00_{\pm 0.03}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	42	$\mathbf{0.99_{\pm 0.02}}$	$0.92_{\pm 0.10}$	1.43	0.109	$-6.26_{\pm 2.13}$	$-7.59_{\pm 2.12}$	$0.75_{\pm 0.10}$	$0.54_{\pm 0.16}$	$0.63_{\pm 0.15}$	$0.88_{\pm 0.19}$	$\mathbf{1.00_{\pm 0.02}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	70	$0.98_{\pm 0.02}$	$0.93_{\pm 0.11}$	1.49	0.136	$-6.67_{\pm 2.49}$	$-7.83_{\pm 2.24}$	$0.76_{\pm 0.11}$	$0.56_{\pm 0.14}$	$0.64_{\pm 0.14}$	$0.89_{\pm 0.18}$	$\mathbf{1.00_{\pm 0.03}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 0.75$)	200	$\mathbf{0.99_{\pm 0.02}}$	$0.93_{\pm 0.10}$	1.12	**0.089**	$-6.58_{\pm 2.37}$	$-7.72_{\pm 2.20}$	$0.75_{\pm 0.11}$	$0.54_{\pm 0.14}$	$0.62_{\pm 0.16}$	$0.88_{\pm 0.19}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	42	$0.98_{\pm 0.02}$	$0.90_{\pm 0.13}$	1.40	0.149	$-6.35_{\pm 2.42}$	$-7.78_{\pm 2.20}$	$0.71_{\pm 0.11}$	$0.53_{\pm 0.14}$	$0.64_{\pm 0.15}$	$0.90_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	70	$\mathbf{0.99_{\pm 0.01}}$	$0.93_{\pm 0.09}$	0.98	0.132	$\mathbf{-6.80_{\pm 2.38}}$	$-7.86_{\pm 2.23}$	$0.71_{\pm 0.13}$	$0.57_{\pm 0.14}$	$0.64_{\pm 0.15}$	$0.89_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.02}}$
Flow-GRPO	Vina–SA ($w_{Vina} = 1$)	200	$0.97_{\pm 0.03}$	$0.89_{\pm 0.13}$	1.79	0.115	$-6.56_{\pm 2.41}$	$\mathbf{-7.92_{\pm 2.28}}$	$0.74_{\pm 0.11}$	$0.54_{\pm 0.14}$	$0.64_{\pm 0.15}$	$0.91_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–QED	42	$0.98_{\pm 0.02}$	$0.92_{\pm 0.11}$	1.02	0.128	$-6.34_{\pm 2.29}$	$-7.59_{\pm 2.14}$	$0.73_{\pm 0.12}$	$0.59_{\pm 0.15}$	$0.64_{\pm 0.15}$	$0.89_{\pm 0.18}$	$\mathbf{1.00_{\pm 0.02}}$
Flow-GRPO	Vina–QED	70	$\mathbf{0.99_{\pm 0.02}}$	$0.90_{\pm 0.14}$	1.28	0.178	$-6.20_{\pm 2.38}$	$-7.69_{\pm 2.13}$	$0.73_{\pm 0.11}$	$0.59_{\pm 0.14}$	$0.65_{\pm 0.14}$	$0.91_{\pm 0.16}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–QED	200	$\mathbf{0.99_{\pm 0.02}}$	$0.90_{\pm 0.15}$	1.58	0.220	$-6.31_{\pm 2.41}$	$-7.76_{\pm 2.23}$	$0.70_{\pm 0.10}$	$\mathbf{0.61_{\pm 0.13}}$	$0.66_{\pm 0.13}$	$\mathbf{0.92_{\pm 0.16}}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–QED–SA	42	$0.98_{\pm 0.03}$	$0.89_{\pm 0.14}$	1.49	0.163	$-6.04_{\pm 2.30}$	$-7.56_{\pm 2.07}$	$0.76_{\pm 0.10}$	$0.58_{\pm 0.14}$	$0.66_{\pm 0.14}$	$0.91_{\pm 0.16}$	$0.99_{\pm 0.03}$
Flow-GRPO	Vina–QED–SA	70	$0.98_{\pm 0.02}$	$0.92_{\pm 0.10}$	1.35	0.131	$-6.56_{\pm 2.49}$	$-7.77_{\pm 2.24}$	$0.76_{\pm 0.11}$	$0.58_{\pm 0.14}$	$0.64_{\pm 0.15}$	$0.90_{\pm 0.17}$	$\mathbf{1.00_{\pm 0.01}}$
Flow-GRPO	Vina–QED–SA	200	$0.97_{\pm 0.03}$	$0.92_{\pm 0.11}$	1.67	0.173	$-6.31_{\pm 2.28}$	$-7.53_{\pm 2.12}$	$0.77_{\pm 0.10}$	$0.57_{\pm 0.15}$	$0.64_{\pm 0.18}$	$0.88_{\pm 0.19}$	$\mathbf{1.00_{\pm 0.03}}$

G Comparisons across datasets

G.1 Pose-optimized Vina–SA trade-offs

Figure 15 show the Vina–SA trade-off on the SPINDR validation set and on the SPINDR and CrossDocked2020 test sets, respectively. Vina scores are UniDock's pose-optimized Vina values obtained by evaluating the selected checkpoints on the two test benchmarks.

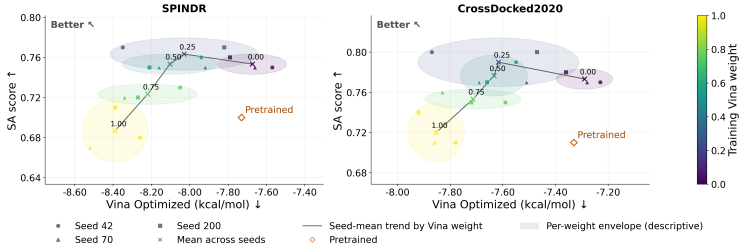


Figure 15: **Pose-optimized Vina–SA trade-offs on SPINDR (left) and CrossDocked2020 (right).** Vina is evaluated after pose optimization. All Flow-GRPO checkpoints are selected on SPINDR validation data and evaluated without further fine-tuning. Colors and numerical annotations indicate the Vina reward weight w_{Vina} and the SA weight is $1 - w_{\mathrm{V}}$. Circles, triangles, and squares show training seeds 42, 70, and 200, respectively. Crosses denote means across seeds, joined in increasing Vina-weight order. Shaded ellipses descriptively enclose the seed-level points for each weight, and orange diamonds denote pretrained FLOWR.ROOT. Lower Vina and higher SA are preferred.

G.2 POSEBUSTERS VALIDITY COMPONENTS

Figure 16 examines the component checks underlying PB validity across the Vina–SA reward sweep. On SPINDR, SA-only alignment yields the lowest PB validity among these configurations (90%, compared with 96% for pretrained FLOWR.ROOT (Table 1). The component breakdown links this to internal-energy failures: the SA-only pass rate is approximately 91%, compared with 99% for the pretrained model. SA-only alignment also reduces the internal-steric-clash pass rate to 97% from 99%, alongside lower RDKit validity. Incorporating Vina into the reward improves overall PB validity up to 96%.

On CrossDocked2020, SA-only alignment again has the lowest PB validity within the sweep (83%), compared with 90% for pretrained FLOWR.ROOT and 93% for the $w_{Vina} = 0.75$ model (Table 5). Similarly, internal energy and internal steric-clash pass rates are lowest under SA-only alignment. In contrast to SPINDR, the minimum-distance-to-protein check is the main source of failures.

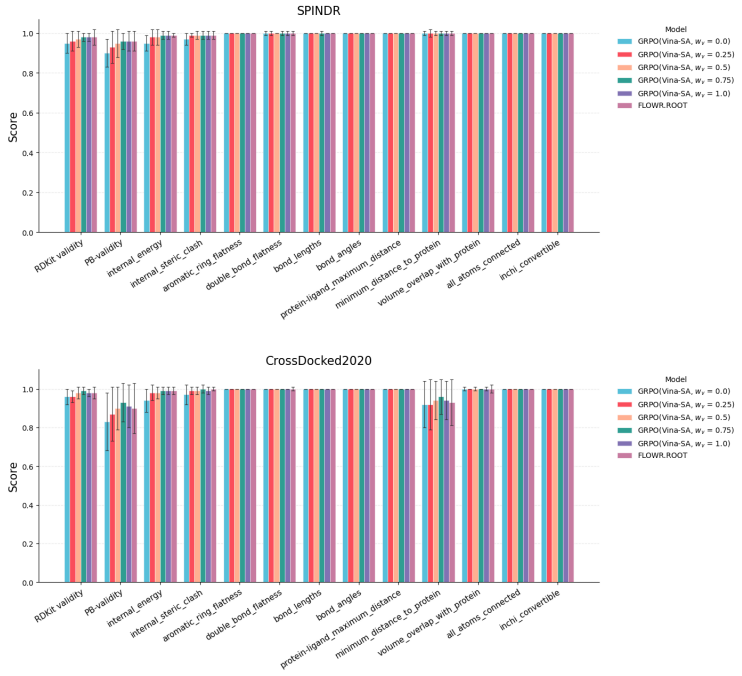


Figure 16: **PoseBusters validity and component pass rates on SPINDR (top) and CrossDocked2020 (bottom).** Bars compare pretrained FLOWR.ROOT with the Vina–SA reward sweep; w_{Vina} denotes the Vina weight and the SA weight is $1 - w_{\mathrm{V}}$. Values are means with sample standard deviations across pockets after averaging seeds within each pocket. RDKit validity is included as one of the criteria in PB validity and both of them are computed out of 100 attempted molecules. Higher values indicate higher pass rates.

G.3 POCKET–LIGAND INTERACTION ANALYSIS

We extract interaction counts from the generated ligand poses using ProLIF (Bouysset & Fiorucci, 2021), comparing the Vina–SA sweep with pretrained FLOWR.ROOT and test ligands. Proteins and ligands are prepared through PoseCheck complex preparation, adding hydrogens while retaining existing ligand coordinates. We use ProLIF's default parameters, except for the residue-neighborhood cutoff, which is set to 7 Å to match the pocket-cut threshold used during sampling. We count all matches across ligand–residue pairs, separating hydrogen bonds where the ligand accepts or donates. Counts are averaged over successfully evaluated ligands within each pocket, then equally over seeds and pockets.

Figure 17 shows similar patterns across datasets. All aligned models have more hydrophobic contacts than the pretrained model, with a maximum at $w_{Vina} = 0.25$ followed by a decline toward Vina-only

alignment. Hydrogen-bond acceptor and donor counts remain below the pretrained and test-ligand references. The interaction counts therefore do not increase uniformly as docking scores improve across reward weights.

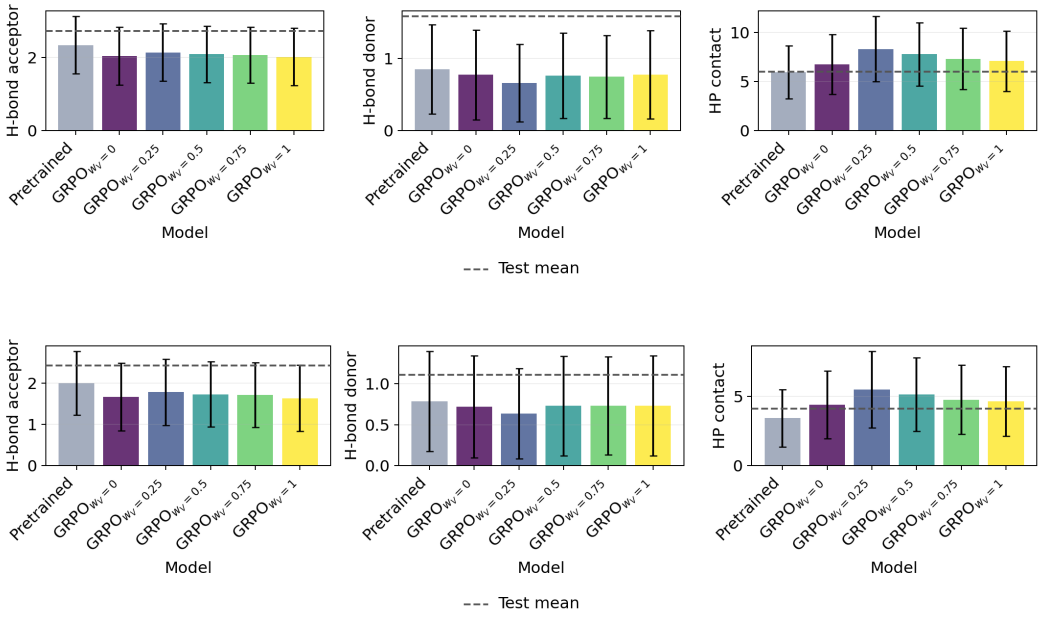


Figure 17: **Pocket–ligand interactions on SPINDR (top) and CrossDocked2020 (bottom).** The displayed values are mean hydrogen-bond acceptor, hydrogen-bond donor, and hydrophobic (HP) contact counts per ligand. The donor/acceptor roles refer to the ligand. GRPO subscripts indicate the Vina reward weight. Bars average ligand counts within each pocket and seeds. Error bars show the mean within-pocket standard deviations.