Steering Flexible-Size Molecular Generation via RL on Trans-Dimensional Flows

Adrien Goldszal[1,2†] **Malte Franke**[3,4] **Andreas Krause**[3,4]
[1]Ecole Polytechnique [2]Université Paris Sciences et Lettres [3]ETH Zürich [4]NCCR Catalysis

Abstract

Flexible-size molecular generators can change cardinality during sampling through insertion and deletion, but steering such trans-dimensional processes with reinforcement learning requires accounting for actions whose dimension changes along the trajectory. We formulate flexible-size generation as a parameterized-action Markov decision process, in which discrete structural decisions select the relevant action-space component and continuous and discrete content is sampled conditionally within it. We derive the corresponding per-step policy density and use it for policy-gradient fine-tuning of Morph, a trans-dimensional 3D molecular generator. Across three settings, reward optimization steers both molecular properties and the model's structural decisions. On QM9, the model generates molecules far beyond the pretraining size support, reaching a mean of 69.9 atoms. On GEOM-Drugs, a sparse validity reward improves strict validity from 92.9% to 98.0%, while PoseBusters validity and strain also improve. For pocket-conditioned ligand generation, fine-tuning improves binding efficiency while ligand size changes in both directions across targets.

1 Introduction

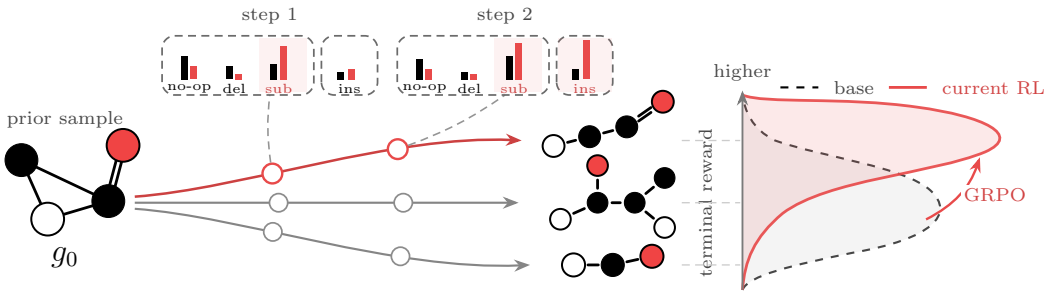


Figure 1: Reinforcement learning guided trans-dimensional generation of geometric graphs.

Generative models are playing a key role in accelerating molecular and materials design, but their practical value lies beyond matching the data distribution: most applications require steering the generation toward downstream objectives such as binding affinities, structural validity, or synthesizability. Reinforcement Learning (RL) provides a principled framework to optimize such objectives, with recent work showing how to cast diffusion and flow-matching trajectories as Markov Decision Processes (MDPs) amenable to policy-gradient methods [Black et al., 2024, Liu et al., 2025].

Recent approaches to molecular generation operate on multimodal states that combine continuous geometry with discrete attributes (e.g. protein structure and sequence, molecular geometry and atom

[†]Corresponding author: adrien.goldszal@polytechnique.edu. Work conducted during an internship at ETH Zürich.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: AI for Accelerated Materials Design.

types). Allowing cardinality to vary introduces additional expressivity: insertion and deletion events move the trajectory between state-space components of different dimensionality. Morph [Franke et al., 2026] realizes this construction by coupling a coordinate flow, discrete graph edits, and dimension-changing jumps to generate variable-size 3D molecular graphs.

Existing reward-guided methods address fixed-cardinality continuous, discrete, or multimodal generators, as well as variable-length categorical processes [Zhu et al., 2026, Wang et al., 2026, Tang et al., 2026], but RL approaches and policy constructions covering geometric generators that couple continuous coordinates and discrete attributes with dimension-changing jumps remain relatively underexplored. Steering these generators is however particularly valuable for out-of-distribution scientific discovery: by controlling insertion and deletion, RL can adapt the molecular-size distribution to the target property and shift probability mass beyond the size support observed during pretraining.

We study the use of policy-gradient finetuning for steering flexible-size geometric models in which dimension-changing decisions are sampled by the policy at each step, and instantiate this formulation on Morph. Our contributions are the following:

1. We cast the trans-dimensional multimodal generative trajectory as an MDP using the standard parameterized-action-space formalism. The sampled topology decision indexes the relevant action-space component and is included in the policy score.
2. We derive the corresponding per-step policy density for a generator coupling a coordinate flow, a discrete CTMC, insertion-deletion jumps, and apply group-relative policy optimization.
3. We show that reward-guided RL can steer flexible-size geometric generators through their insertion and deletion decisions and study three objectives: molecular-size generation beyond the pretraining support, verifier-based molecular validity, and ligand-pocket binding.

2 Trans-dimensional MDPs for Flexible-Size Geometric Graph Generation

Flexible-size generation makes cardinality an outcome of the sampling trajectory rather than an input fixed in advance, allowing molecular size to respond to the target objective. For policy-gradient fine-tuning, however, the action may therefore vary in dimension and composition across steps: at each step the policy samples a structural decision that determines both the successor cardinality and which continuous and discrete content variables are present in the action. The per-step density must therefore be defined on a state-dependent action space. We formalize this structure for a general trans-dimensional process before instantiating it for molecular generation with Morph [Franke et al., 2026].

2.1 Trans-Dimensional Generative MDPs

A time-discretized generative process can be cast as a finite-horizon MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \rho_0, P, R)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, ρ_0 is the initial-state distribution, P is the transition kernel, and R is the reward. We call the current object being generated the generative state and distinguish it from the complete MDP state, which also records the integration time. We denote its state space by $\mathcal{X}$. Since its cardinality may vary:

$$\mathcal{X} = \bigsqcup_{n \geq 1} \mathcal{X}_n, \tag{1}$$

where $\mathcal{X}_n$ contains states with n elements. At integration step k, the generative state is $x_k \in \mathcal{X}_{n_k}$, and the corresponding MDP state is $s_k = (x_k, t_k) \in \mathcal{S}$. Dimension-changing events may move the process between state-space components, such that $n_{k+1} \neq n_k$ and the dimensions of the generative state and action may vary along a trajectory.

We model the state-dependent action space using the parametrized-action-space formalism [Masson et al., 2016, Hausknecht and Stone, 2016]. We define $\mathcal{A}(s) \subseteq \mathcal{A}$ as the set of actions admissible at state s and $\Gamma(s)$ the set of admissible discrete structural decisions. Then let $\gamma \in \Gamma(s)$ be a discrete structural decision, and $\mathcal{A}_\gamma(s)$ be the space of content variables available under that decision. An action is a pair $a = (\gamma, \eta)$ with $\eta \in \mathcal{A}_\gamma(s)$, so

$$\mathcal{A}(s) = \bigsqcup_{\gamma \in \Gamma(s)} \left(\{\gamma\} \times \mathcal{A}_\gamma(s)\right). \tag{2}$$

The resulting action-space components may differ both in dimension and in the continuous and discrete variables they contain. A policy on this space can be viewed as a mixture over these components [Eisenach et al., 2019, App. B.5]. Its density at $a = (\gamma, \eta)$ factorizes as

$$\pi_\theta(a \mid s) = \pi_\theta(\gamma \mid s)\, \pi_\theta(\eta \mid s, \gamma). \tag{3}$$

Here, the first factor is the probability of selecting the structural decision and hence the corresponding action-space component, and the second is the conditional density of its content variables. In our trans-dimensional setting, γ specifies which elements of x_k persist, are deleted, or are created, while η contains the continuous and discrete content defined under that structural decision. Because γ is sampled by the policy rather than supplied by the transition kernel, its probability contributes to $\log \pi_\theta(a_k \mid s_k)$. Thus, every realized structural decision, including choosing not to perform an action (no-op), contributes to the policy score, whereas a content variable is scored only when it is present in $\mathcal{A}_\gamma(s)$.

We consider the case of terminal rewards which we write $r(x_T)$, and assume ρ_0 and P do not depend on θ. Writing T for the number of integration steps, let $\tau = (s_0, a_0, \ldots, s_T)$ denote a trajectory with induced distribution $p_\theta(\tau)$, the objective and its policy gradient are:

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta}\big[r(x_T)\big], \qquad \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta}\Big[r(x_T) \sum_{k=0}^{T-1} \nabla_\theta \log \pi_\theta(a_k \mid s_k)\Big] \tag{4}$$

The action enters the estimator only through the scalar $\log \pi_\theta(a_k \mid s_k)$ so neither its varying dimension nor its mix of continuous and discrete channels need estimator-specific treatment.

2.2 Generator and Trajectory MDP on Molecular Graphs

Molecular state space. Let A, B, and C be the sets of atom types, bond types and formal charges. A molecule with n atoms is represented as an indexed geometric graph $g = (n, V, E)$ where $V = (p_i)_{i=1}^n$, $p_i = (a_i, x_i) \in \mathsf{A} \times \mathbb{R}^3$, and $E = [e_{ij}]_{i,j=1}^n \in \mathcal{S}_n(\mathsf{B})$, with $\mathcal{S}_n(\mathsf{B})$ the set of symmetric $n \times n$ bond-type matrices. Writing $\mathbb{R}_0^{n\times 3}$ for the zero-centre-of-mass coordinate space and $n_{\max}$ for the maximum atom count used in practice, the indexed, centered MDP generative state space is:

$$g \in \mathcal{G} = \bigsqcup_{n=1}^{n_{\max}} \mathsf{A}^n \times \mathbb{R}_0^{n\times 3} \times \mathcal{S}_n(\mathsf{B}). \tag{5}$$

This is an instance of Eq. (1), with $\mathcal{X}_n = \mathsf{A}^n \times \mathbb{R}_0^{n\times 3} \times \mathcal{S}_n(\mathsf{B})$. At integration step k, let $g_k \in \mathcal{G}$ denote the generative state and n_k its number of atoms with the corresponding MDP state $s_k = (g_k, t_k)$. Formal charges are terminal readouts and are therefore not part of the evolving state space.

Action. At each step, Morph may substitute or delete existing atoms, substitute bonds, and insert new atoms. Let $d = (d_i)_{i=1}^{n_k}$ denote the node-edit decisions, with $d_i \in \{\text{NO-OP}, \text{SUB}, \text{DEL}\}$, and let $z = (z_i)_{i=1}^{n_k}$ denote the insertion decisions, with $z_i \in \{0, 1\}$. We define S the set of surviving atoms, I the set of sites inserting new atoms, and $E_S(d)$ the set of unordered pairs of surviving atoms:

$$S = \{i : d_i \neq \text{DEL}\}, \qquad I = \{i : z_i = 1\} \qquad E_S(d) = \{(i, j) : i < j,\ i, j \in S\}$$

Conditional on d, let $b = (b_{ij})_{(i,j)\in E_S}$ denote the bond-edit decisions, with $b_{ij} \in \{\text{NO-OP}, \text{SUB}\}$. The structural decision $\gamma = (d, b, z)$ is therefore:

$$\gamma = (d, b, z) \in \Gamma_{n_k} := \bigsqcup_{d \in \{\text{NO-OP,SUB,DEL}\}^{n_k}} \{d\} \times \{\text{NO-OP}, \text{SUB}\}^{E_S(d)} \times \{0, 1\}^{n_k}. \tag{6}$$

A pre-step node may simultaneously satisfy $d_i = \text{DEL}$ and $z_i = 1$: the existing atom is deleted, while the new atom inserted at that site remains in the successor. The structural decision γ then selects the action-space component:

$$\begin{aligned} \mathcal{A}_\gamma(s_k) = \quad & \underbrace{\mathbb{R}^{3|S|}}_{\text{surviving positions}} \times \underbrace{\mathsf{A}^{|\{i:d_i=\text{SUB}\}|}}_{\text{substituted atom types}} \times \underbrace{\mathsf{B}^{|\{(i,j)\in E_S : b_{ij}=\text{SUB}\}|}}_{\text{substituted bond types}} \\ & \times \underbrace{(\mathbb{R}^3 \times \mathsf{A})^{|I|}}_{\text{inserted atoms}} \times \underbrace{\mathsf{B}^{|I||S|}}_{\text{inserted–existing edges}} \times \underbrace{\mathsf{B}^{\binom{|I|}{2}}}_{\text{inserted–inserted edges}}. \end{aligned} \tag{7}$$

Here B includes the no-bond type, so every new atom pair carries one bond-type variable. The successor contains $n_{k+1} = |S| + |I|$ atoms. At the terminal step, the component additionally contains $\mathsf{C}^{n_{k+1}}$ for the final formal charges; intermediate charge predictions are not action variables because they are neither observed by the model nor retained in the generated output.

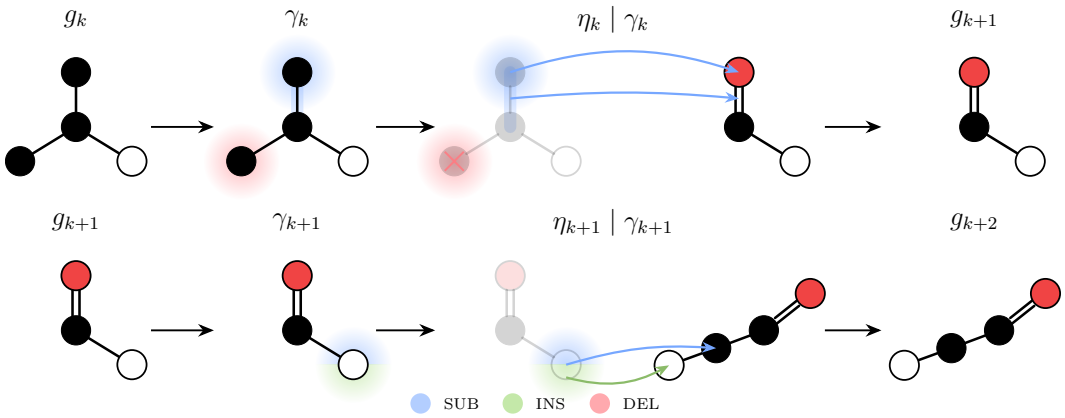


Figure 2: Two consecutive trans-dimensional actions $a_j = (\gamma_j, \eta_j)$ in Morph. At each step, the structural decision γ_j selects the action-space component, while the conditional content η_j specifies updated coordinates, substituted atom and bond types, and the position, type, and incident bonds of inserted atoms, yielding g_{j+1}. At step k, an atom and a bond are substituted and one atom is deleted, so $n_{k+1} = n_k - 1$. At step $k+1$, an existing atom is substituted and a new atom is inserted from the same site. Unhighlighted structural decisions are no-ops.

2.3 Per-Step Policy Density

For a realized action $a_k = (\gamma_k, \eta_k)$, we separate the conditional content η_k into the coordinate updates of surviving atoms and the remaining content variables. The policy factorization in Eq. (3) then yields:

$$\log \pi_\theta(a_k \mid s_k) = \log \pi_\theta^{\text{struct}} + \log \pi_\theta^{\text{coord}} + \log \pi_\theta^{\text{content}} \tag{8}$$

where $\log \pi_\theta^{\text{struct}}$ scores γ_k, $\log \pi_\theta^{\text{coord}}$ scores the stochastic coordinate updates of surviving atoms, and $\log \pi_\theta^{\text{content}}$ scores the remaining variables sampled conditionally on γ_k, including substitutions, inserted atoms and edges, and, at the terminal step, charges. Per-channel scoring domains and explicit distributions are given in Appendix D.

Positions. Continuous flow-based policies are deterministic and therefore carry no score-function density. Following RL finetuning of flow models [Liu et al., 2025, Zhang et al., 2025] we add stochastic noise to the coordinate update with a fixed variance Gaussian kernel $\sigma_{explore}$. Using $\Delta t = t_{k+1} - t_k$, for each surviving atom $i \in S$:

$$x_{i,t+\Delta t} \sim \mathcal{N}\big(x_{i,t} + \hat{v}_i\, \Delta t,\ \sigma^2_{\text{explore}} I_3\big), \qquad \hat{v}_i = \frac{\hat{x}_{1,i} - x_{i,t}}{1-t}, \tag{9}$$

with $\hat{x}_{1,i}$ the predicted clean position. We use fixed variance because Morph's log-spaced schedule drives $\Delta t \to 0$ near t=1: a marginal-preserving Δt-scaled variance collapses and produces exploding coordinate score gradients that dominate the discrete terms.

Structural decisions and conditional content. Morph's continuous-time edit rates are discretized into categorical or Bernoulli distributions over the structural decisions d, b, and z (Appendix D). Every realized decision contributes to $\log \pi_\theta^{\text{struct}}$, including no-op outcomes and non-selected insertions. The conditional-content term contains only variables present in the selected component $\mathcal{A}_\gamma(s_k)$.

Most conditional content variables are categorical, but two channels are mixtures and therefore require marginal likelihoods. In particular, inserting an atom requires sampling the joint of position and type from a typed Gaussian mixture (K components indexed by ℓ, weights w_ℓ):

$$\log \pi_\theta^{\text{ins}}(x, a) = \text{logsumexp}_\ell \Big[\log w_\ell + \log \mathcal{N}(x \mid \mu_\ell, \sigma_\ell^2 I) + \log p_{\text{atom},\ell}[a]\Big], \tag{10}$$

with an additional term $\log p_{\text{charge},\ell}[c]$ at the terminal step, where charges enter g_T. The mixture-component label is latent and is not part of the policy action, so the corresponding policy density is the marginal mixture density in Eq. (10). Bonds involving inserted atoms are scored analogously under their marginal mixture distributions; explicit forms are given in Appendix D.

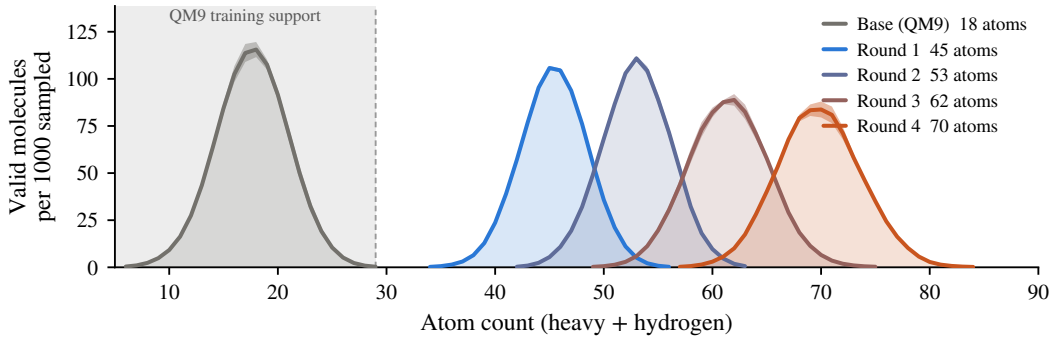


Figure 3: Atom-count distributions across the size curriculum. RDKit-valid molecules only, 1000 generated per seed. Line is the mean density over 3 inference seeds, band ± 1 std. Unit-width bins are convolved with a Gaussian of $\sigma = 1.6$ atoms.

2.4 RL Algorithm Instantiation

We use Group-Relative Policy Optimization (GRPO) [Shao et al., 2024] with the standard clipped surrogate objective and a reference-policy penalty toward the pretrained model π_{ref}. At each update, we sample a batch of B trajectories under the rollout policy π_{old}, partitioned into groups of G trajectories sharing the same initial graph $g_0 \sim \rho_0$. Trajectories are sampled independently conditional on g_0. To remove between-initial-state variation, terminal rewards are standardized within each group. For each realized action a_k^i, we define the group-relative advantage and per-step importance ratio

$$\hat{A}^i = \frac{r(g_T^i) - \mu_i}{\sigma_i}, \qquad (\mu_i, \sigma_i) = \text{mean}, \text{std}\big\{r(g_T^{i'})\big\}_{i' \in \text{grp}(i)}, \qquad \rho_k^i(\theta) = \frac{\pi_\theta(a_k^i \mid s_k^i)}{\pi_{\text{old}}(a_k^i \mid s_k^i)}. \quad (11)$$

We optimize the clipped surrogate objective

$$\mathcal{L}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{T}\sum_{k=0}^{T-1}\min\Big(\rho_k^i(\theta)\,\hat{A}^i,\ \text{clip}\big(\rho_k^i(\theta), 1-\epsilon, 1+\epsilon\big)\,\hat{A}^i\Big) + \beta\,\widehat{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\,, \quad (12)$$

where ϵ is the clipping threshold and β controls the strength of reference regularization. The log densities entering the importance ratio in Eq. (11) are computed using the factorization in Eq. (8). Re-scoring evaluates the numerator and denominator at the same realized action $a_k = (\gamma, \eta_k)$. Both densities therefore lie on the same action-space component selected by γ. The construction of the reference-policy penalty is detailed in Appendix E.

3 Experiments

We evaluate three aspects of trans-dimensional policy optimization: whether it can (1) steer cardinality beyond the pretraining support, (2) improve molecular quality under a sparse verifier reward while preserving the pretrained distribution, and (3) exploit adaptive cardinality in pocket-conditioned ligand optimization. The first two isolate controllability and reward alignment, while the third tests the approach in a multi-objective structure-based design setting.

3.1 Steering Toward Out-of-distribution Molecules

Setup. We validate our method first by maximizing atom count, a scalar objective that exercises the trans-dimensional machinery and tests out-of-distribution discovery. We pretrain Morph on QM9 [Ramakrishnan et al., 2014], a dataset over more than 100k small organic molecules, containing at most 32 atoms, and optimize final atom count with zero reward for RDKit invalid molecules. We apply a canonical-SMILES diversity penalty by adopting the occurrence-bucketed penalisation strategy introduced by REINVENT [Loeffler et al., 2024] to discourage mode collapse onto a single large scaffold. We repeat this finetuning process multiple times in a curriculum inspired fashion, each model being KL-anchored to its previously trained version.

Results. Figure 3 shows the atom-count distribution of the QM9-pretrained base model and of each curriculum round, and Table 1 reports the corresponding metrics. The pretrained model is concentrated between 13 and 23 atoms, and the curriculum rounds move the distribution to a mean of 69.9 atoms, $3.96\times$ the training mean and entirely outside the training support. This confirms that the trans-dimensional policy gradient propagates reward through insertion and deletion decisions, enabling extrapolation to molecular sizes beyond those represented in the training data.

Chemical validity remains stable across rounds (0.828–0.873) even as size quadruples, and diversity stays high despite the model steering towards large alkanes. The degradation at the largest sizes is concentrated in geometry rather than chemical validity. We score PoseBusters [Buttenschoen et al., 2024] after a single force-field minimization of the generated coordinates, because its bond-length and bond-angle checks require every bond and angle in a molecule to pass, making the molecule-level criterion increasingly strict as molecule size grows. Minimization corrects most small raw geometric violations without changing the topological checks (Appendix Table 8) with the relaxed PoseBusters score nearly matching RDKit validity up until round 4 (0.650 against a ceiling of 0.828).

Table 1: **Steering QM9 out of distribution by molecule size.** Mean over 3 inference seeds $\times$ 1000 generated molecules. Detailed results and PoseBusters per-check breakdown are in Table 8.

	Mean Atoms	Max Atoms	Validity ↑	POSEBUSTERS ↑	Diversity ↑
Base (QM9)	17.6	27	0.976	0.919	0.898
Round 1	45.5	54	0.828	0.828	0.837
Round 2	53.0	62	0.873	0.872	0.787
Round 3	61.6	75	0.834	0.798	0.786
Round 4	**69.9**	82	0.828	0.650	0.756

3.2 RLVR-style Molecule Validity Steering on GEOM

Setup. We test whether Morph can be steered by a sparse verifier-style reward without changing the target distribution. We pretrain a Morph model on GEOM-Drugs [Axelrod and Gómez-Bombarelli, 2022], then fine-tune it using a strict binary validity reward (RDKit sanitizable, single fragment, radical-free) on the final generated molecule. Fine-tuning proceeds over five curriculum rounds, each KL-anchored to the previous checkpoint (Table 10). As in Section 3.1, we include an occurrence-bucketed canonical-SMILES diversity wrapper.

Results. RL-steering successfully increases the rewarded strict-validity rate from 92.9% to 98.0%, as well as the broader RDKit-sanitizable validity ($96.0\% \rightarrow 99.1\%$). PoseBusters validity, despite not being optimized directly through the reward, rises from 87.8% to 94.3%, matching the GEOM-Drugs training set itself (0.94). Similarly PoseCheck [Harris et al., 2023] strain energy falls from 17.9 ± 0.4 to 14.8 ± 0.2 kcal/mol. Among the geometric PoseBusters checks, the gain is concentrated in steric clash and radical removal (Appendix Table 12).

Table 2: GEOM-Drugs. Literature numbers from Vonessen et al. [2026]; ours are mean±std over 3 seeds $\times$ 1000 molecules. Strain is aggregated as the median, matching their protocol.

Method	#Params	Validity ↑	Novelty ↑	Diversity ↑	POSEBUSTERS ↑	Strain ↓
TABASCO-spicy	59M	0.97	0.90	0.89	0.92	15.07
+ guidance	59M	0.97	0.92	0.89	0.94	19.23
Morph (base)	40.8M	0.96	0.99	0.89	0.88	17.91
Morph + GRPO	40.8M	**0.99**	0.99	0.88	**0.94**	**14.83**

Comparison to test-time guidance. Figure 4 places the curriculum trajectory on the PoseBusters-strain plane against TABASCO [Vonessen et al., 2026] and its distance-bounds guidance. Both methods reach the 0.94 data ceiling. While TABASCO's guidance is a per-sample correction, raising PoseBusters by 0.02 at a $7\times$ increase in sampling cost and converging to higher strain ($15.07 \rightarrow 19.23$), our reward fine-tuning is amortized and delivers a larger $+0.065$ PoseBusters gain and moves strain in the opposite direction ($17.91 \rightarrow 14.83$). Therefore guidance satisfies a distance constraint at the price of strain, whereas validity RL improves both.

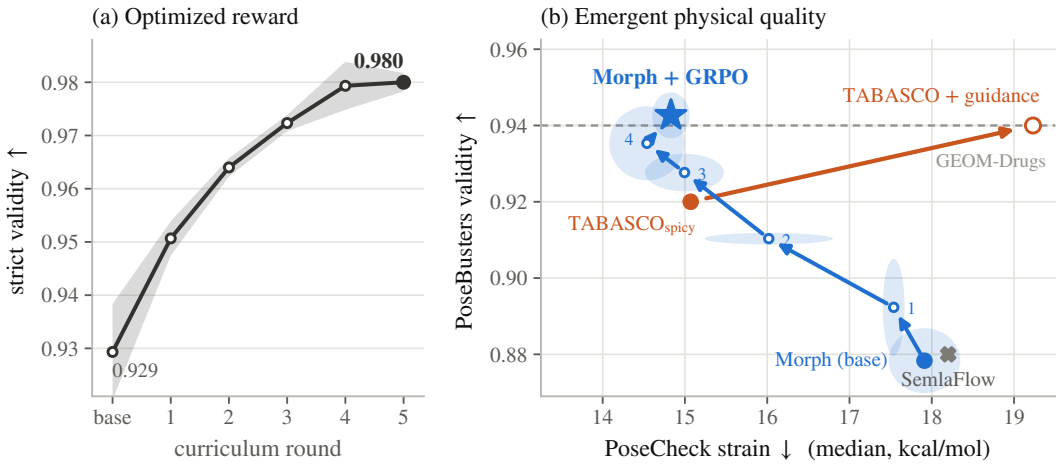


Figure 4: **Validity optimization induces emergent physical quality. (a)** Strict validity, the optimized RL reward, across five GEOM curriculum rounds (base = pretrained Morph). **(b)** The corresponding curriculum in the PoseBusters–PoseCheck strain plane. Morph points are means over 3 seeds; bands and shaded regions indicate ± 1 std. TABASCO and SemlaFlow are literature results [Vonessen et al., 2026]; the dashed line marks the GEOM-Drugs PoseBusters level.

Distributional fidelity. The gains are accompanied by minimal change in molecular coverage. Uniqueness remains at 100%, novelty is unchanged (99.6% $\rightarrow$ 99.4%), and mean pairwise ECFP6 Tanimoto distance is flat (0.886 $\rightarrow$ 0.884). Mean molecule size falls slightly, from 44.6 to 42.6 atoms, raising the atom-count KL to GEOM from 0.072 to 0.122 (Appendix Fig. 9); atom-, bond- and charge-type distributions are unchanged to three decimal places.

3.3 Adaptive-Size Pocket-Conditioned Ligand Generation

We finally study the trans-dimensional policy in the setting of pocket-conditioned ligand generation. Given a protein pocket, the model generates a ligand and its 3D pose while balancing binding affinity, drug-likeness, synthesizability, and geometric validity. Existing pocket-conditioned generators such as DiffSBDD [Schneuing et al., 2024] and MolPilot [Qiu et al., 2025], as well as DePPA [Xue et al., 2026] which similarly fine-tunes a pocket-conditioned diffusion model with reinforcement learning, sample the ligand atom count before denoising and hold it fixed throughout the trajectory. Morph instead allows the per-pocket reward optimization to adapt ligand size to the target.

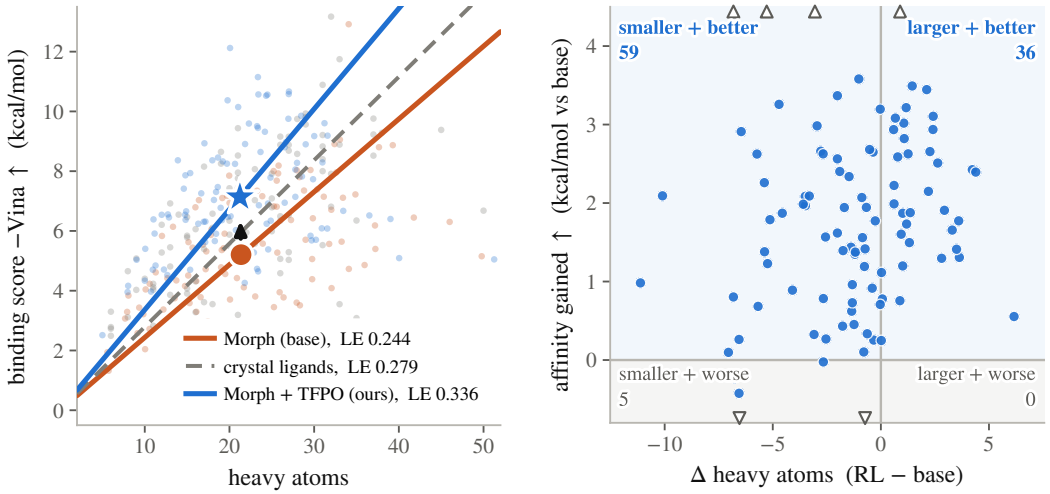
Setup. We pretrain a pocket-conditional Morph model on CrossDocked2020 [Francoeur et al., 2020] and evaluate on the standard 100-pocket test split. Unlike Section 3.2, we perform per-pocket test-time optimization and run a fixed amount of GRPO updates against each target, yielding one policy per pocket. The reward Gaussian-rank-transforms affinity (cube root normalized by size, see Shen et al. [2024]), QED, SA, geometry and diversity within each rollout group, behind a hard gate on PoseBusters bond validity and steric clashes. Molecules are conditioned on the 10 Å pocket and scored with AutoDock Vina 1.2 [Eberhardt et al., 2021] `score_only` on the generated pose in the full receptor, as in MolCRAFT [Qu et al., 2024], whose Vina protocol MolPilot also adopts, ensuring our rows and the literature rows of Table 3 are on a common scale. PoseBusters validity follows MolPilot. More details in Appendix H.3.

Results. Every optimized objective improves (Table 3): Vina Score $-5.21 \rightarrow -7.14$ ($+1.93$ kcal/mol), QED $0.508 \rightarrow 0.598$, normalized SA $0.694 \rightarrow 0.746$, and the fraction of valid, connected molecules $0.835 \rightarrow 0.943$. PoseBusters validity also rises from 0.606 to 0.794, although it is not optimized directly (Appendix Table 15). The fine-tuned policy clears the crystal reference (-6.36) and every generative baseline, including MolPilot (-6.88), while generating ligands smaller than either. It remains 1.4 kcal/mol short of DePPA, which is finetuned on top of a different base. Pocket-level diversity is lower than the baselines' (0.584, falling to 0.411 after fine-tuning), but so is CrossDocked's own within target diversity (0.597), which the base model matches (Appendix H.5).

Ligand size adaptation across pockets. While rewarding raw Vina encourages ligand growth, our reward structure ensures gains appear as efficiency (Figure 5a). Affinity improves in $95/100$

Table 3: **Pocket-conditioned ligand generation on CrossDocked2020**, 100 test pockets × 100 ligands. DiffSBDD and MolPilot are from Qiu et al. [2025], Table 1; DePPA from Xue et al. [2026].

	Vina Score ↓	Vina Min ↓	QED ↑	SA ↑	Div ↑	Size	Connected ↑	PB-Valid ↑
Reference ligands	−6.36	−6.71	0.48	0.73	—	22.8	1.000	0.950
Generative — Amortized pocket-conditioned generation								
DiffSBDD	−1.44	−4.52	0.47	0.58	0.73	24.4	0.932	0.376
MolPilot	−6.88	−7.23	0.56	0.74	0.69	22.6	**0.979**	**0.959**
Morph (base)	−5.21	−6.01	0.508	0.694	0.584	21.37	0.835	0.606
Optimisation — per-pocket reward optimization								
DePPA	**−8.50**	**−8.91**	0.57	0.69	**0.78**	21.2	0.953	—
Morph + GRPO (ours)	−7.14	−7.65	**0.598**	**0.746**	0.411	21.23	0.943	0.794



(a) Each ray is a line of constant −Vina/HAC (Heavy Atom Count). Steeper is more binding per heavy atom. Points are per-pocket means. The arrow marks the population mean moving −1.31 heavy atoms and +1.80 kcal/mol. Rays are ratios of means.

(b) Change in per-pocket mean heavy-atom count against affinity gained over the base generator, all 100 pockets. Six pockets fall outside the plotted range and are clipped to the nearest edge as hollow triangles.

Figure 5: **Binding affinity and ligand size after fine-tuning.** a The fine-tuned policy moves up and to the left: more binding from fewer atoms. b It gains affinity as ligand size shifts in both directions.

pockets by 1.8 kcal/mol while the mean heavy-atom count falls by 1.31, taking the average efficiency from 0.230 to 0.326, past the crystal ligands' 0.292. The policy achieves this improvement through bidirectional adaptation in size, growing the ligand in 36 pockets and shrinking it in 59 (Figure 5b), an action a fixed-cardinality generator does not have.

4 Related Work

Multimodal and trans-dimensional generators. Multimodal molecular generators couple continuous coordinates with discrete attributes, for example through coordinate flows and CTMCs over atom and bond types [Campbell et al., 2024, Dunn and Koes, 2026]. Separately, trans-dimensional generative models allow cardinality to change during sampling through operations such as insertion, deletion, or splitting [Campbell et al., 2023, Havasi et al., 2025, Billera et al., 2026]. Morph [Franke et al., 2026] combines these two directions, coupling coordinate flow, discrete graph edits, and insertion-deletion jumps over variable-size 3D molecular graphs. We give a more detailed comparison of these generators under the parametrized-action-space formalism in Appendix A.

Reward-guided steering. Policy-gradient methods have been used to fine-tune continuous diffusion and flow models [Black et al., 2024, Liu et al., 2025, Zhang et al., 2025], as well as discrete CTMC and graph generators [Zekri and Boullé, 2025, Wang et al., 2026, Zhu et al., 2026]. Existing approaches, however, largely operate at fixed cardinality. In Graph-GRPO [Zhu et al., 2026], graph size is sampled

before denoising and remains fixed within a trajectory; an adaptive prior can reweight sizes across rollouts, but size is not itself a scored policy decision. Fixed-dimensional multimodal generators have likewise been steered using guidance or preference optimization [Lin et al., 2025, Huang and Zhang, 2025]. A2D2 [Tang et al., 2026] addresses variable-length discrete diffusion, but uses a detached trajectory weighting rather than a per-step action density, and inserted positions are initially masked rather than generated jointly with their content. Our setting instead requires scoring decisions that change dimensionality and determine the continuous and discrete variables in the resulting action.

5 Limitations & Further Work

Empirical scope. Despite shifting cardinality beyond the pretraining support, RL remains constrained by the pretrained generator. QM9 extrapolation reduces validity, GEOM-Drugs gains saturate near reference-data levels, and pocket-conditioned performance trails stronger baselines. Applying the same optimization to stronger or higher-capacity pretrained models is therefore the most direct path to further improvement. Broader experimentation should also target distribution-level adaptation, for instance steering to extend the reachable chemical space such as from a QM9 generator toward GEOM-like cardinality and diversity. This broader adaptation perspective also extends to the pocket-conditioned setting, with the goal of amortizing optimization across targets to eliminate the need for per-pocket policies. Replacing scalar rewards with explicit multi-objective or Pareto formulations is also a natural next step, to help characterize the trade-offs among reward, validity, diversity, and cardinality.

Policy design and efficiency. The fixed-variance stochastic noise used to convert the deterministic coordinate flow into a policy requires more systematic analysis and comparison to marginal preserving alternatives [Liu et al., 2025] or methods learning timestep-dependent exploration noise [Zhang et al., 2025]. Moreover, long sampling trajectories make reinforcement learning computationally expensive, particularly when optimization requires costly oracle evaluations. Coarser discretizations, trajectory distillation, and few-step samplers could reduce this cost, while ensuring they preserve the fidelity of dimension-changing decisions. For example, the allocation of sampling steps could also be learned as part of the policy through adaptive timestep selection or stopping.

6 Conclusion

In this work, we cast flexible-size molecular generation as a parameterized-action Markov Decision Process and derive the corresponding per-step policy density for the Morph jump-flow process, making the dimension-changing decisions of a trans-dimensional generator directly amenable to policy-gradient optimization. Across three settings, we show how trans-dimensional policy optimization enables reward-driven control over molecular structure and size: (1) it drives molecular size far beyond the pretraining support on QM9, (2) turns a sparse validity signal into improved validity and emergent physical quality on GEOM-Drugs without shifting the distribution, (3) and adapts ligand size in both directions to improve binding efficiency in pocket-conditioned generation. These results show that reinforcement learning can steer cardinality alongside molecular structure, introducing an axis of control unavailable to fixed-cardinality models and therefore expanding the potential for out-of-distribution scientific discovery.

References

Simon Axelrod and Rafael Gómez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://doi.org/10.1038/s41597-022-01288-4.

Lukas Billera, Hedwig Nora Nordlinder, Jack Collier Ryder, Anton Oresten, Aron Stålmarck, Theodor Mosetti Björk, and Ben Murrell. Branching flows: Discrete, continuous, and manifold flow matching with splits and deletions. In *Integrating Generative and Experimental Platforms for Biomolecular Design (GEM) Workshop at ICLR 2026*, 2026. URL https://openreview.net/forum?id=xsghez038X. arXiv:2511.09465.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.13301.

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. doi: 10.1039/D3SC04185A. URL https://doi.org/10.1039/D3SC04185A.

Andrew Campbell, William Harvey, Christian Weilbach, Valentin De Bortoli, Tom Rainforth, and Arnaud Doucet. Trans-dimensional generative modeling via jump diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.16261.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5453–5512. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/campbell24a.html.

Ian Dunn and David R. Koes. FlowMol3: flow matching for 3d *de novo* small-molecule generation. *Digital Discovery*, 5(5):2052–2066, 2026. doi: 10.1039/D5DD00363F. URL https://doi.org/10.1039/D5DD00363F.

Jérôme Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi: 10.1021/acs.jcim.1c00203.

Carson Eisenach, Haichuan Yang, Ji Liu, and Han Liu. Marginal policy gradients: A unified family of estimators for bounded action spaces with applications. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019. OpenReview.net. URL https://openreview.net/forum?id=HkgqFiAcFm.

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, September 2020. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411.

Malte Franke, Stefan P. Schmid, Žarko Ivković, Kjell Jorner, and Andreas Krause. Generative molecular morphing for flexible-size design via unbalanced optimal transport. *arXiv preprint arXiv:2606.07239*, 2026. URL https://arxiv.org/abs/2606.07239.

Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Liò, and Tom L. Blundell. PoseCheck: Generative models for 3D structure-based drug design produce unrealistic poses. In *Machine Learning for Structural Biology Workshop at NeurIPS 2023*, 2023. URL https://www.mlsb.io/papers_2023/PoseCheck_Generative_Models_for_3D_Structure-based_Drug_Design_Produce_Unrealistic_Poses.pdf.

Matthew Hausknecht and Peter Stone. Deep reinforcement learning in parameterized action space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016. URL https://arxiv.org/abs/1511.04143.

Marton Havasi, Brian Karrer, Itai Gat, and Ricky T. Q. Chen. Edit flows: Variable length discrete flow matching with sequence-level edit operations. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, volume 38, 2025.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.20587.

Jie Huang and Daiheng Zhang. MolFORM: Multi-modal flow matching for structure-based drug design. In *2nd Generative AI for Biology Workshop at the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kylXbRzdD6.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – efficient 3d molecular generation with latent attention and equivariant flow matching. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3772–3780. PMLR, 2025. URL https://proceedings.mlr.press/v258/irwin25a.html.

Haowei Lin, Shanda Li, Haotian Ye, Yiming Yang, Stefano Ermon, Yitao Liang, and Jianzhu Ma. TFG-Flow: Training-free guidance in multimodal generative flow. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GK5ni7tIHp.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online rl. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, volume 38, 2025. URL https://arxiv.org/abs/2505.05470.

Hannes H. Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H. Mervin, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024. doi: 10.1186/s13321-024-00812-5. URL https://doi.org/10.1186/s13321-024-00812-5.

Warwick Masson, Pravesh Ranchod, and George Konidaris. Reinforcement learning with parameterized actions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1934–1940. AAAI Press, 2016. doi: 10.1609/aaai.v30i1.10226. URL https://doi.org/10.1609/aaai.v30i1.10226.

Keyue Qiu, Yuxuan Song, Zhehuan Fan, Peidong Liu, Zhe Zhang, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. Piloting structure-based drug design via modality-specific optimal schedule. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50619–50644. PMLR, 2025. URL https://proceedings.mlr.press/v267/qiu25g.html.

Yanru Qu, Keyue Qiu, Yuxuan Song, Jingjing Gong, Jiawei Han, Mingyue Zheng, Hao Zhou, and Wei-Ying Ma. MolCRAFT: Structure-based drug design in continuous parameter space. In *International Conference on Machine Learning (ICML)*, 2024.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L. Blundell, Pietro Liò, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, December 2024. doi: 10.1038/s43588-024-00737-x. URL https://doi.org/10.1038/s43588-024-00737-x.

John Schulman. Approximating KL divergence. http://joschu.net/blog/kl-approx.html, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Tony Shen, Seonghwan Seo, Grayson Lee, Mohit Pandey, Jason R. Smith, Artem Cherkasov, Woo Youn Kim, and Martin Ester. TacoGFN: Target-conditioned GFlowNet for structure-based drug design. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=N8cPv95zOU.

Sophia Tang, Yuchen Zhu, Molei Tao, and Pranam Chatterjee. A2D2: Fine-tuning any-length discrete diffusion for adaptive decoding. *arXiv preprint arXiv:2606.13565*, 2026. URL https://arxiv.org/abs/2606.13565.

Carlos Vonessen, Charles Harris, Miruna Cretu, and Pietro Liò. TABASCO: A fast, simplified model for molecular generation with improved physical quality. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Kg6CSrbXl4.

Guanghan Wang, Gilad Turok, Yair Schiff, Marianne Arriola, and Volodymyr Kuleshov. d2: Improving reasoning in diffusion language models via trajectory likelihood estimation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2509.21474.

Yuan Xue, Daniel Kudenko, and Megha Khosla. Fine-tuning pocket-aware diffusion models via denoising policy optimization. In *Multi-Objective Decision Making Workshop (MODeM 2026), held in conjunction with IJCAI–ECAI 2026*, Bremen, Germany, August 2026. URL https://modem2026.vub.ac.be/papers/MODeM_2026_paper_10.pdf.

Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, volume 38, 2025. URL https://arxiv.org/abs/2502.01384. arXiv:2502.01384.

Tonghe Zhang, Chao Yu, Sichang Su, and Yu Wang. Reinflow: Fine-tuning flow matching policy with online reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025. doi: 10.48550/arXiv.2505.22094. URL https://arxiv.org/abs/2505.22094. arXiv:2505.22094.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, volume 38, pages 56729–56762, 2025. doi: 10.52202/085713-1898.

Baoheng Zhu, Deyu Bo, Delvin Ce Zhang, and Xiao Wang. Graph-GRPO: Training graph flow models with reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.10395.

A Generators in the Common Formalism

We present here additional details and examples on parametrized-action control problems in the formalism of Section 2.1, as well as recasting representative generators in this setting. In Table 4 we present examples from the literature. In Table 5, we give examples of trans-dimensional generative models who fit the parametrized action setting. For each we give the state space $\mathcal{X}$, the discrete decision element $\gamma \in \Gamma(s)$ of the action and the content variables available under that decision $\mathcal{A}_\gamma(s)$ it selects.

Table 4: Classical parametrized action spaces

Setting	$\boldsymbol{\mathcal{X}}$	$\boldsymbol{\Gamma(s)}$	$\boldsymbol{\mathcal{A}_\gamma(s)}$
HFO [Hausknecht and Stone, 2016]	$\subseteq \mathbb{R}^{58}$	{DASH, TURN, TACKLE, KICK}	$[0, 100] \times [-180, 180]$ for DASH, KICK; $[-180, 180]$ otherwise
King of Glory [Eisenach et al., 2019]	$\mathbb{R}^{2775}$	{NOOP, MOVE, ATTACK, SKILL1, SKILL2, SKILL3, RECOVERY}	$\mathbb{R}^2$ for MOVE and SKILL1–3; singleton otherwise

In HFO, γ selects one of four high-level actions, with the selected action requiring either a power-direction pair (DASH, KICK) or only a direction (TURN, TACKLE). In King of Glory, γ selects one of seven high-level actions, four of which require a two-dimensional directional parameter while the remaining three have no additional parameter and therefore a singleton content space.

Table 5: Trans-dimensional generative models in the parametrized action space formalism. Spaces are written for one integration step of the discretized sampler at a state with n elements.

Setting	$\boldsymbol{\mathcal{X}}$	$\boldsymbol{\Gamma(s)}$	$\boldsymbol{\mathcal{A}_\gamma(s)}$
Edit Flows [Havasi et al., 2025]	$\bigsqcup_{n=0}^{N} \mathcal{T}^n$	$\{\text{NO-OP}, \text{SUB}, \text{DEL}\}^n \times \{0,1\}^n$	$\mathcal{T}^{\#\{\text{SUB}\}+\#\{\text{INS}\}}$
Jump diffusion [Campbell et al., 2023]	$\bigsqcup_{n=1}^{N} \{n\} \times \mathbb{R}^{nd}$	$\{0, 1\}$	$\mathbb{R}^{(n+\gamma)d} \times \left(\mathbb{R}^d \times [n{+}1]\right)^\gamma$
Morph	Eq. (5)	Eq. (6)	Eq. (7)

In Edit Flows, γ specifies the substitution/deletion decision and insertion gate at every current token, and the selected content space contains one vocabulary-valued content variable for every substitution and every firing insertion. In jump diffusion, γ specifies whether a new component is inserted; the content space contains the diffusion update of all $n + \gamma$ components and, when $\gamma = 1$, additionally the new d-dimensional component and its insertion index.

Beyond the examples in Table 5, several neighbouring constructions can also be viewed through this lens. We mention a few informally to illustrate related action structures rather than provide an exhaustive classification.

Fixed-cardinality continuous and discrete denoisers [Black et al., 2024, Liu et al., 2025, Zekri and Boullé, 2025, Zhao et al., 2025, Wang et al., 2026, Zhu et al., 2026] do not move between cardinality-indexed state-space components. At the coarsest MDP level, they can be represented by taking $\Gamma(s)$ to be a singleton and the action to be the next iterate, although particular samplers may admit finer action decompositions.

Other trans-dimensional models place the dimension-changing decision and the generation of new content differently. In A2D2 [Tang et al., 2026], insertion adds masked slots whose token values are subsequently resolved by the unmasking process. In Branching Flows [Billera et al., 2026], a split increases cardinality by duplicating an existing element in place; the two copies are subsequently evolved by the base process, rather than sampling new element content at the split itself.

B The Morph Trans-Dimensional Jump-Flow Process

B.1 Morph Rate–Mark Process

Morph couples a coordinate flow with atom insertion, deletion and substitution, and bond substitution to evolve the indexed, centred graph state $g_t \in \mathcal{G}$ defined in Eq. (5). Within a fixed-cardinality component $\mathcal{G}_{n_t}$, atom coordinates evolve according to the learned ODE

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t(g_t), \qquad v_t : \mathcal{G}_{n_t} \to \mathbb{R}^{n_t \times 3}. \tag{13}$$

The discrete process comprises atom insertion, atom deletion, atom substitution, and bond substitution: $u_t = u_t^{\mathrm{ins}} + u_t^{\mathrm{del}} + u_t^{\mathrm{sub},a} + u_t^{\mathrm{sub},e}$. Each discrete edit family admits a rate-mark decomposition

$$u_t^{\alpha}(g' \mid g) = \sum_i \lambda_{t,i}^{\alpha}(g)\, Q_{t,i}^{\alpha}(g' \mid g), \qquad \lambda_{t,i}^{\alpha}(g) = \Delta_i^{\alpha}(g)\, \frac{\dot{\kappa}_t^{\alpha}}{1-\kappa_t^{\alpha}}. \tag{14}$$

The rate determines whether an edit occurs, while the mark specifies its conditional content. This decomposition induces the gate and mark channels of the policy in Section 2.3; their discrete-time probabilities are given in Appendix D.

The underlying continuous-time construction and pretraining objective are unchanged from Morph; we refer to Franke et al. [2026] and Holderrieth et al. [2025] for their full derivation.

B.2 Policy-Channel Domains

Using the notation of Section 2.2, let $S = S(d)$ be the surviving pre-step atoms and $I = I(z)$ the inserted atoms, indexed by their spawn sites. Existing-atom pairs are indexed by $\binom{S}{2}$, insertion-to-existing pairs by $I \times S$, and insertion-to-insertion pairs by $\binom{I}{2}$. Table 6 gives the scoring domain of every policy channel.

Table 6: Policy channels and their scoring domains. Gates are scored whenever they are present in the conditional action; their marks are scored only when the corresponding gate fires.

Channel	**Domain**	**Distribution**	**Scored when**
Existing positions	S	$\mathcal{N}(\mu_i, \sigma^2_{\mathrm{explore}} I)$	atom survives
Node gate d_i	$[n_k]$	$\mathrm{Cat}(\textsc{no-op}, \textsc{sub}, \textsc{del})$	every pre-step atom
atom-type mark	$\{i : d_i = \textsc{sub}\}$	$\mathrm{Cat}(\boldsymbol{\pi}^i_{\mathrm{atom}})$	$d_i = \textsc{sub}$
Bond gate b_{ij}	$\binom{S}{2}$	$\mathrm{Bern}(p^{ij}_{\mathrm{sub},e})$	both endpoints survive
bond-type mark	$\{(i,j) \in \binom{S}{2} : b_{ij} = \textsc{sub}\}$	$\mathrm{Cat}(\boldsymbol{\pi}^{ij}_{\mathrm{edge}})$	$b_{ij} = \textsc{sub}$
Existing-atom charge	S	$\mathrm{Cat}(\boldsymbol{\pi}^i_{\mathrm{charge}})$	terminal step only
Insertion gate z_i	$[n_k]$	$\mathrm{Bern}(p^i_{\mathrm{ins}})$	every pre-step atom
Inserted atom	I	typed GMM over (x, a)	$z_i = 1$
Insertion $\to$ existing edge	$I \times S$	edge-type mixture	insertion exists and target survives
Insertion $\to$ insertion edge	$\binom{I}{2}$	edge-type mixture	both atoms are inserted

C Action Illustration

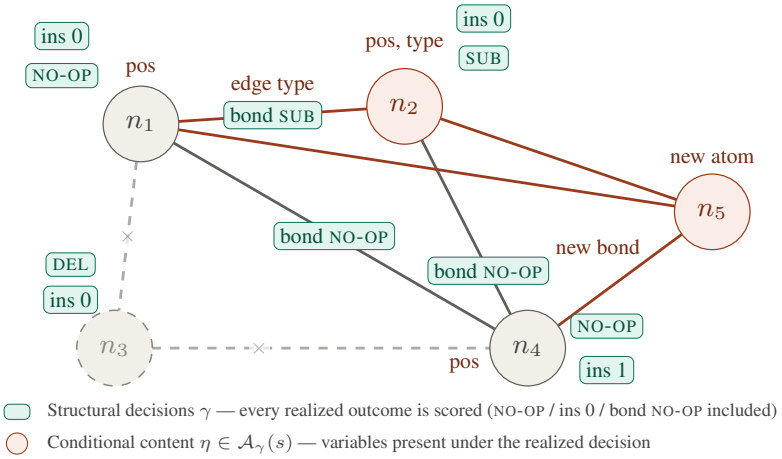


Figure 6: Structural decisions (green) are sampled at every applicable site and are scored for every realized outcome, including no-ops; conditional content variables (coral) are present only in the action space component selected by those decisions.

D Discrete Channel Details

We give the mapping from Morph's model outputs to the discrete per-step probabilities and conditional mark densities used in Section 2.3.

Model outputs. At each step, the model outputs sitewise gate magnitudes

$$\Delta_i^{\mathrm{sub},a},\ \Delta_i^{\mathrm{DEL}},\ \Delta_{ij}^{\mathrm{sub},e} \in (0,1), \qquad \Delta_i^{\mathrm{ins}} \geq 0,$$

together with categorical mark distributions

$$\boldsymbol{\pi}^i_{\mathrm{atom}} \in \triangle^{|\mathsf{A}|-1}, \qquad \boldsymbol{\pi}^i_{\mathrm{charge}} \in \triangle^{|\mathsf{C}|-1}, \qquad \boldsymbol{\pi}^{ij}_{\mathrm{edge}} \in \triangle^{|\mathsf{B}|-1}.$$

The insertion head additionally outputs, for each spawning node i, mixture weights $\boldsymbol{\pi}^i \in \triangle^{K-1}$, means $\mu^i_\ell \in \mathbb{R}^3$, scales σ^i_ℓ, and component-wise atom-type and charge distributions $p^i_{\mathrm{atom},\ell}$ and $p^i_{\mathrm{charge},\ell}$. A separate insertion-edge predictor supplies categorical edge-type distributions for inserted–existing and inserted–inserted pairs. Charge distributions are produced throughout sampling, but charge draws are scored only at the terminal step; intermediate values are not observed by the model and are overwritten.

Gate probabilities. For an Euler step Δt, each rate magnitude is converted to a per-step probability by

$$p^\alpha = \Delta^\alpha \frac{\dot{\kappa}^\alpha_t}{1-\kappa^\alpha_t}\,\Delta t, \tag{15}$$

with the appropriate node or edge index understood. Probabilities are clipped to the valid numerical range before sampling and scoring. Substitution and deletion compete at each existing atom. Writing

$$m_i = p_i^{\mathrm{sub}} + p_i^{\mathrm{DEL}}, \qquad q_i = \frac{p_i^{\mathrm{DEL}}}{m_i},$$

and $\bar{m}_i$ for the corresponding clipped edit probability, the realised node decision has distribution

$$\pi^{\mathrm{act}}_\theta(d_i) = \big(1-\bar{m}_i,\ \bar{m}_i(1-q_i),\ \bar{m}_i q_i\big), \qquad d_i \in \{\textsc{no-op}, \textsc{sub}, \textsc{del}\}. \tag{16}$$

When $d_i = \textsc{sub}$, the sampled atom type is additionally scored under $\boldsymbol{\pi}^i_{\mathrm{atom}}$. Existing-edge substitutions are Bernoulli with probability $p_{ij}^{\mathrm{sub},e}$, with the new bond type scored under $\boldsymbol{\pi}^{ij}_{\mathrm{edge}}$ when the gate fires. Insertion is likewise a Bernoulli gate with probability p_i^{ins}. Every realised gate outcome is scored, including no-ops and non-firing insertion gates. At the terminal step, surviving-atom charges are additionally scored under $\boldsymbol{\pi}^i_{\mathrm{charge}}$.

Insertion marks. Conditional on a firing insertion gate at node i, the inserted position and atom type are sampled jointly from the typed Gaussian mixture

$$\log \pi_\theta^{\mathrm{ins},i}(x,a) = \mathrm{logsumexp}_\ell \Big[\log \pi_\ell^i + \log \mathcal{N}(x \mid \mu_\ell^i, (\sigma_\ell^i)^2 I) + \log p^i_{\mathrm{atom},\ell}[a] \Big]. \tag{17}$$

At the terminal step, the term $\log p^i_{\mathrm{charge},\ell}[c]$ is included inside the logsumexp. The mixture-component label is latent and is therefore marginalised rather than treated as part of the action.

For each bond involving an inserted atom, the insertion-edge predictor gives a clean edge-type distribution $\boldsymbol{\pi}_{\mathrm{new}}$. The realised bond type at time $t + \Delta t$ is scored under the marginal distribution

$$\boldsymbol{p}_{\mathrm{new}} = (1 - \kappa^{\mathrm{sub},e}_{t+\Delta t})\, p^e_{\mathrm{prior}} + \kappa^{\mathrm{sub},e}_{t+\Delta t}\, \boldsymbol{\pi}_{\mathrm{new}}, \tag{18}$$

for both inserted–existing and inserted–inserted pairs. Since B contains the no-bond type, this is a categorical mark for every such pair and requires no additional edge-presence gate.

E GRPO Objective and Reference Regularization

Reference-policy regularization. We regularize the policy toward the frozen pretrained model π_{ref} using the non-negative k_3 estimator [Schulman, 2020], as in GRPO [Shao et al., 2024]. For policy channel c at step k of trajectory i, with corresponding log densities $\ell^i_{\theta,k,c}$ and $\ell^i_{\mathrm{ref},k,c}$, define

$$r^i_{\mathrm{ref},k,c} = \frac{\pi_{\mathrm{ref},c}(a^i_k \mid s^i_k)}{\pi_{\theta,c}(a^i_k \mid s^i_k)} = \exp\big(\ell^i_{\mathrm{ref},k,c} - \ell^i_{\theta,k,c}\big)\,, \tag{19}$$

and

$$k_3(r) = (r-1) - \log r. \tag{20}$$

The sample-based reference penalty used in Eq. (12) is

$$\widehat{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}) = \frac{1}{B}\sum_{i=1}^{B} \frac{1}{T}\sum_{k=0}^{T-1} \sum_{c \in \mathcal{C}^i_k} k_3\big(r^i_{\mathrm{ref},k,c}\big)\,, \tag{21}$$

where $\mathcal{C}^i_k$ denotes the policy-density contributions defined for the realized action at that step. Thus, contributions that are absent from the realized action-space component are not included. This keeps the reference regularization well defined when the action dimension varies across steps and trajectories.

Optimization. Rollouts are generated once under π_{old} and may be reused for multiple clipped optimization passes. During re-scoring, the realized action $a_k = (\gamma, \eta_k)$ is held fixed while its density is evaluated under the updated policy. The reference policy remains frozen throughout each fine-tuning run.

F QM9 Experiment Details

F.1 Training Details and Parameters

Reward. The reward is the final molecule's total atom count, gated on chemical validity and penalised for repetition. Writing g_T for the terminal graph of a rollout, the base reward is $r_{\text{atoms}}(g_T) = N_{atoms}(g_T)$, the number of nodes, which *includes explicit hydrogens* since the QM9 vocabulary carries `H` as its own atom token.

A validity gate multiplies it by $\mathbf{1}[g_T \text{ is RDKit-valid}]$, where validity requires successful sanitization, a single fragment, no radical electrons and neutral formal charges, so an invalid molecule scores exactly zero regardless of size.

Finally we apply the occurrence-bucketed diversity penalty of REINVENT [Loeffler et al., 2024]: each valid molecule is keyed by its canonical SMILES, and a rolling window of the last $W = 50$ batches records how many times each key has been accepted. Within a batch, up to $B_{\max} = 10$ occurrences of a key are accepted; any further occurrence has its reward multiplied by $\rho = 0.5$. The complete per-rollout reward is therefore

$$r(g_T) \;=\; N_{atoms}(g_T) \cdot \mathbf{1}[\text{valid}] \cdot \big(\rho\,\mathbf{1}[\text{bucket over-full}] + \mathbf{1}[\text{otherwise}]\big).$$

Curriculum procedure. Each curriculum round is a separate GRPO run, and rounds are chained through checkpoint selection. Within a round we track an exponential moving average of the batch mean reward with coefficient 0.9 and write a checkpoint whenever that average exceeds its own running maximum. Because reward degrades late in every round once the policy drifts far from its reference, the retained model is this best-EMA snapshot rather than the final iterate. Round $k+1$ is then initialised from round k's snapshot, and the KL reference is re-anchored to those same weights.

Table 7: Hyperparameters for the QM9 atom-count experiment.

Hyperparameter	**Value**
Batch size B	128
Group size G	8
Integration steps T	100
σ_{explore}	0.05
Learning rate	10^{-4}
Clipped passes μ	2
KL coefficient β	0.02
Clip threshold ϵ	0.2
Training rounds	3000
Diversity bucketing	SMILES
Maximum bucket size $B_{\max}$	10
Bucketing threshold ρ	0.5
Window size W	50
Curriculum rounds	4

Table 8: **Per-check POSEBUSTERS results across the QM9 size curriculum.** *Raw* and *MMFF-relaxed* scores. Check rows are pass rates over the molecules POSEBUSTERS evaluated, i.e. those RDKit could build and sanitise (row 'Parseable'); the bold row instead divides by all 1000 generated molecules, counting unparseable outputs as failures.

Check	Base	Round 1	Round 2	Round 3	Round 4
Raw (as generated)					
All atoms connected	1.000	0.978	0.980	0.970	0.977
No radicals	0.997	0.951	0.962	0.963	0.962
Bond lengths	1.000	0.430	0.731	0.344	0.186
Bond angles	0.999	0.684	0.768	0.428	0.173
Internal steric clash	0.999	0.995	0.960	0.854	0.939
Aromatic ring flatness	1.000	1.000	1.000	1.000	1.000
Double-bond flatness	0.997	1.000	1.000	1.000	1.000
Non-aromatic ring non-flatness	0.999	1.000	1.000	1.000	1.000
Internal energy	1.000	0.996	0.982	0.920	0.944
All checks (of parseable)	0.940±0.001	0.309±0.016	0.538±0.031	0.155±0.016	0.033±0.003
All checks (of generated)	**0.920±0.002**	**0.270±0.016**	**0.489±0.030**	**0.134±0.014**	**0.028±0.003**
MMFF-relaxed					
All atoms connected	1.000	0.978	0.980	0.970	0.977
No radicals	0.997	0.951	0.962	0.963	0.962
Bond lengths	1.000	0.992	0.997	0.984	0.985
Bond angles	1.000	0.996	0.998	0.947	0.780
Internal steric clash	0.999	0.993	0.990	0.992	0.992
Aromatic ring flatness	1.000	1.000	1.000	1.000	1.000
Double-bond flatness	0.997	1.000	1.000	1.000	1.000
Non-aromatic ring non-flatness	0.998	1.000	1.000	1.000	1.000
Internal energy	1.000	1.000	1.000	0.997	0.999
All checks (of parseable)	0.939±0.003	0.947±0.004	0.959±0.004	0.920±0.007	0.754±0.008
All checks (of generated)	**0.919±0.002**	**0.828±0.006**	**0.872±0.009**	**0.798±0.011**	**0.650±0.004**
Parseable / 1000	979	874	910	867	862
RDKit validity (stricter)	0.976	0.828	0.873	0.834	0.828
Mean atom count	17.6	45.5	53.0	61.6	69.9

F.2 Results

Force-field relaxation. Generated coordinates are scored both as emitted and after a single local minimisation, reported as the *raw* and *MMFF-relaxed* columns of Table 8. The relaxation is RDKit's `AllChem.MMFFOptimizeMolecule` under the MMFF94 force field, applied to a copy of the molecule with `maxIters`$= 1000$, acting on the first (and only) conformer. The optimiser's convergence flag is not used as a filter, so structures that fail to converge within 1000 iterations are still scored; molecules for which the force field cannot be constructed are treated as failures and retain their place in the denominator. POSEBUSTERS is then re-run on the relaxed coordinates with the identical configuration and the identical set of molecules, so the raw and relaxed columns differ only in the coordinates supplied. We report both because a force field can repair local geometry.

Remark: POSEBUSTERS' `mol_pred_loaded`, `sanitization` and `inchi_convertible` checks are omitted: the pipeline only submits molecules that already sanitised, so they are 1.000 by construction. 'RDKit validity' is a stricter criterion than parseability and is reported for reference. While they are kept in the appendix of the other experiments, the same situation arises.

Uniqueness, novelty, diversity. Uniqueness is the fraction of distinct canonical SMILES among molecules convertible to SMILES. Novelty is the fraction of those canonical SMILES absent from the training set. Diversity is the mean pairwise Tanimoto *distance* ($1-$ similarity) over ECFP fingerprints with radius 3 (ECFP6) and 2048 bits.

Table 9: **Seed variability for Table 1.** The same metrics with standard deviations over the 3 inference seeds (1000 generated molecules each). Novelty is 1.000 ± 0.000 for every RL round and is omitted.

	Atoms	Max	Validity ↑	POSEBUSTERS ↑	Diversity ↑
Base (QM9)	17.6±0.1	27	0.976±0.001	0.919±0.002	0.898±0.001
Round 1	45.5±0.1	54	0.828±0.006	0.828±0.006	0.837±0.001
Round 2	53.0±0.1	62	0.873±0.009	0.872±0.009	0.787±0.003
Round 3	61.6±0.1	75	0.834±0.006	0.798±0.011	0.786±0.005
Round 4	69.9±0.0	82	0.828±0.017	0.650±0.004	0.756±0.003

Examples.

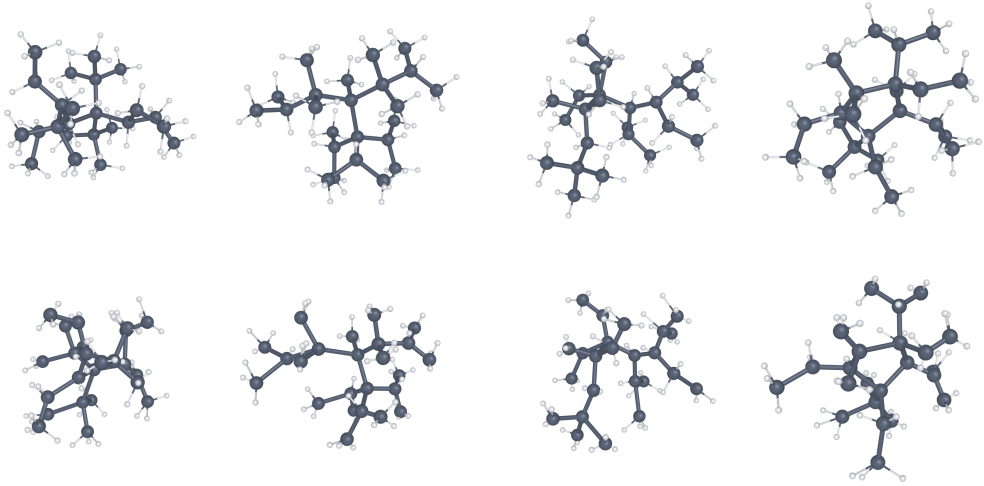


Figure 7: **Round 4 molecules before (bottom) and after (top) force-field relaxation.** Four Round 4 samples (71–74 atoms, 24 heavy atoms each), shown MMFF-relaxed on the top row and exactly as generated on the bottom row. All four pass every POSEBUSTERS check after relaxation.

G GEOM Experiment Details

G.1 Hyperparameter Details

Table 10: Hyperparameters for the GEOM validity experiment.

Hyperparameter	Value
Batch size B	140
Group size G	32
Integration steps T	100
σ_{explore}	0.02
Learning rate	3×10^{-6}
Clipped passes μ	1
KL coefficient β	0.1
Clip threshold ϵ	0.2
Training rounds	3000
Diversity bucketing	SMILES
Maximum bucket size $B_{\max}$	10
Bucketing threshold ρ	0.5
Window size W	50
Curriculum rounds	5

G.2 Metrics

Several metrics in Table 11 share a name with a metric in prior work but not a definition, and two rows both called "strain" are computed with different force fields, different reference states and different aggregations. We therefore state each one exactly.

Strict validity. The RL reward as defined in Section 3.2: RDKit sanitisation succeeds on g_T, the molecule is a single connected fragment, it carries no radical electrons, and its net formal charge is 0.

RDKit-parseable. Sanitisation alone, dropping the other three conjuncts. This is the quantity Vonessen et al. [2026] report as "Validity", and is what we use in Table 2 for comparability. It is strictly weaker than the reward.

POSEBUSTERS. The 12 checks of Buttenschoen et al. [2024] in `mol` mode; a molecule passes only if it passes all of them. We report two denominators, which differ by the molecules POSEBUSTERS cannot evaluate at all. *Of generated* divides by all 1000 sampled molecules and so compounds parseability with geometric quality; *of parseable* divides by the molecules that could be loaded, and is therefore the quality a rejection-sampling baseline would obtain. Per-check rates are given in Table 12.

PoseCheck strain energy. The metric of Harris et al. [2023], reproduced from their implementation, and the one Vonessen et al. [2026] tabulate. It is a UFF energy difference $E_{\text{local}} - E_{\text{global}}$. E_{local} relaxes the generated pose under per-atom position restraints (maximum displacement 0.1 Å, force constant 10^5), so that clashes and stretched bonds relax out while the conformational hypothesis is preserved; the quantity is therefore conformational strain rather than local geometric noise. E_{global} is the minimum UFF energy over 50 ETKDGv3-embedded, UFF-relaxed conformers together with the freely relaxed pose — the latter is required, since ETKDG can miss the basin the model generated and its omission produces spurious negative strains. We aggregate with the median, matching their default; the per-molecule distribution has a heavy right tail and its mean is larger by more than an order of magnitude, so the aggregation must be stated with any published number. Following Harris et al. [2023], force-field minimisation uses RDKit's default 200 iterations and does not fully converge; raising it lowers the local term by ~ 2 kcal/mol. We inherit this deliberately so that the comparison to published numbers is like-for-like, which makes absolute values soft and cross-model differences and within-curriculum trends the supported claims.

MMFF strain energy. SemlaFlow's `AverageStrainEnergy` [Irwin et al., 2025]: the MMFF94 energy of the raw generated pose minus the energy of that pose after unconstrained MMFF optimization, aggregated as a **mean**. Both terms of the difference and the aggregation differ from PoseCheck strain so the two strain rows are not on a common scale and must not be compared to each other. *MMFF strain per atom* and *MMFF energy per atom* divide the same quantities by the atom count of each molecule before averaging; the per-atom strain is what controls for changes in molecule size.

Optimization RMSD. Heavy-atom RMSD between the generated pose and that pose after MMFF optimization, averaged over molecules where optimization succeeds. Larger values mean the generated geometry sits further from the nearest force-field minimum.

Molecule stability. SemlaFlow's valence criterion: the fraction of molecules in which every atom has an allowed valence for its element and formal charge. Unparseable molecules count as unstable.

Uniqueness, novelty, diversity. Uniqueness is the fraction of distinct canonical SMILES among molecules convertible to SMILES. Novelty is the fraction of those canonical SMILES absent from the training set. Diversity is the mean pairwise Tanimoto *distance* ($1-$ similarity) over ECFP fingerprints with radius 3 (ECFP6) and 2048 bits, formed across the whole accumulated sample rather than averaged per batch. Radius, bit count and the distance convention all follow Vonessen et al. [2026].

Distributional KLs. For atom counts, atom types, bond types and formal charges we bin the generated sample into the same histogram as the training marginal and report $D_{\mathrm{KL}}(\text{generated} \,\|\, \text{target})$.

G.3 Detailed Results

Table 11: Full GEOM-Drugs results across the validity curriculum. Mean $\pm$ standard deviation over 3 seeds $\times$ 1000 molecules. Round 0 ('base') is the pretrained model.

		curriculum round					
Metric		base	1	2	3	4	5
Chemical validity							
Strict validity (RL reward)	↑	0.929 ±0.009	0.951 ±0.003	0.964 ±0.002	0.972 ±0.002	0.979 ±0.005	0.980 ±0.002
RDKit-parseable	↑	0.960 ±0.008	0.974 ±0.001	0.983 ±0.006	0.989 ±0.003	0.991 ±0.003	0.991 ±0.002
PoseBusters (of generated)	↑	0.878 ±0.009	0.892 ±0.013	0.910 ±0.002	0.928 ±0.005	0.935 ±0.010	0.943 ±0.006
PoseBusters (of parseable)	↑	0.915 ±0.004	0.916 ±0.014	0.926 ±0.007	0.938 ±0.007	0.944 ±0.011	0.952 ±0.005
Molecule stability	↑	0.956 ±0.006	0.971 ±0.001	0.980 ±0.006	0.986 ±0.004	0.989 ±0.004	0.989 ±0.002
Geometry and strain							
PoseCheck strain (median)	↓	17.91 ±0.44	17.54 ±0.13	16.02 ±0.78	14.99 ±0.48	14.54 ±0.46	14.83 ±0.23
MMFF strain (mean)	↓	38.69 ±0.36	36.52 ±0.50	35.35 ±0.87	33.50 ±3.00	30.52 ±1.71	31.54 ±0.43
MMFF strain per atom	↓	0.856 ±0.011	0.827 ±0.011	0.795 ±0.032	0.768 ±0.054	0.702 ±0.040	0.732 ±0.013
MMFF energy per atom	↓	1.687 ±0.013	1.632 ±0.038	1.574 ±0.056	1.520 ±0.084	1.531 ±0.056	1.504 ±0.011
Optimization RMSD	↓	0.914 ±0.024	0.889 ±0.015	0.901 ±0.038	0.862 ±0.019	0.861 ±0.022	0.878 ±0.024
Diversity							
Uniqueness	↑	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
Novelty	↑	0.995 ±0.001	0.997 ±0.002	0.995 ±0.002	0.994 ±0.002	0.997 ±0.002	0.994 ±0.001
Diversity (ECFP6)	↑	0.886 ±0.000	0.885 ±0.000	0.885 ±0.002	0.884 ±0.001	0.884 ±0.001	0.884 ±0.000
Distributional fidelity to GEOM-Drugs							
Mean size (atoms)	–	44.57 ±0.10	43.96 ±0.14	43.50 ±0.30	42.85 ±0.21	42.72 ±0.19	42.64 ±0.25
Atom-count KL	↓	0.072 ±0.007	0.084 ±0.013	0.095 ±0.016	0.110 ±0.007	0.117 ±0.007	0.122 ±0.014
Atom-type KL	↓	0.0044 ±0.0006	0.0048 ±0.0006	0.0056 ±0.0003	0.0059 ±0.0003	0.0065 ±0.0005	0.0065 ±0.0004
Bond-type KL	↓	0.0001 ±0.0000	0.0001 ±0.0000	0.0001 ±0.0000	0.0002 ±0.0000	0.0002 ±0.0000	0.0002 ±0.0000
Charge-type KL	↓	0.0001 ±0.0000	0.0000 ±0.0000	0.0001 ±0.0000	0.0001 ±0.0000	0.0002 ±0.0001	0.0001 ±0.0001

Table 12: POSEBUSTERS checks across the validity curriculum on GEOM-Drugs. Each entry is the pass rate among RDKit-parseable molecules, averaged over 3 seeds × 1000 molecules. Checks in the first block are already implied by the RL reward (sanitisable, single-fragment, neutral, radical-free); those in the second are geometric and never enter it.

	curriculum round						
Check	base	1	2	3	4	5	Δ
Implied by the RL reward							
loadable	1.000	1.000	1.000	1.000	1.000	1.000	+0.000
sanitization	1.000	1.000	1.000	1.000	1.000	1.000	+0.000
InChI convertible	1.000	0.999	1.000	1.000	1.000	1.000	+0.000
all atoms connected	0.991	0.995	0.994	0.997	0.995	0.996	+0.005
no radicals	0.975	0.980	0.986	0.986	0.993	0.993	+0.018
Geometry only — never rewarded							
bond lengths	0.997	0.996	0.998	0.999	1.000	0.998	+0.001
bond angles	0.999	0.999	0.998	0.999	1.000	0.999	−0.001
internal steric clash	0.955	0.950	0.957	0.962	0.961	0.967	+0.012
aromatic ring flatness	1.000	1.000	1.000	1.000	1.000	1.000	+0.000
non-aromatic ring non-flatness	0.995	0.996	0.994	0.995	0.995	0.997	+0.001
double bond flatness	0.995	0.995	0.994	0.995	0.999	0.998	+0.003
internal energy	1.000	1.000	1.000	1.000	1.000	1.000	+0.000
All checks pass (of parseable)	0.915	0.916	0.926	0.938	0.944	0.952	+0.037
All checks pass (of generated)	0.878	0.892	0.910	0.928	0.935	0.943	+0.064

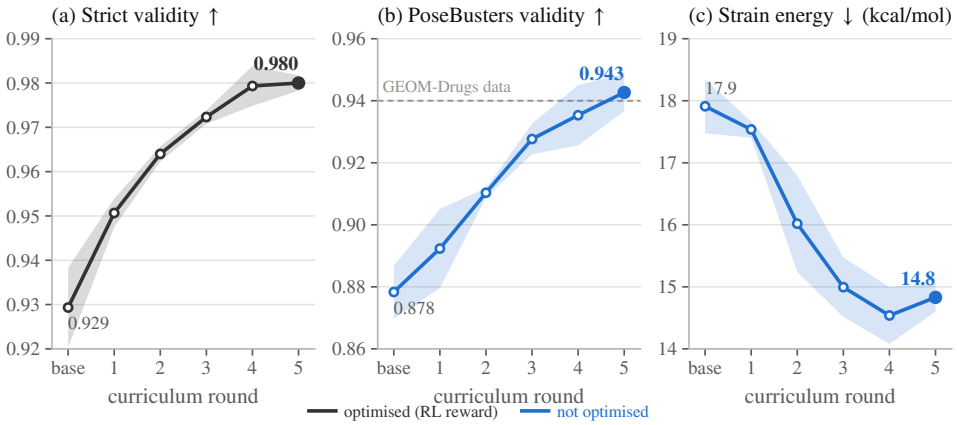


Figure 8: **Emergent physical quality under a validity-only reward.** Metrics across five GEOM curriculum rounds (base = pretrained Morph). Bands are ±std over 3 seeds.

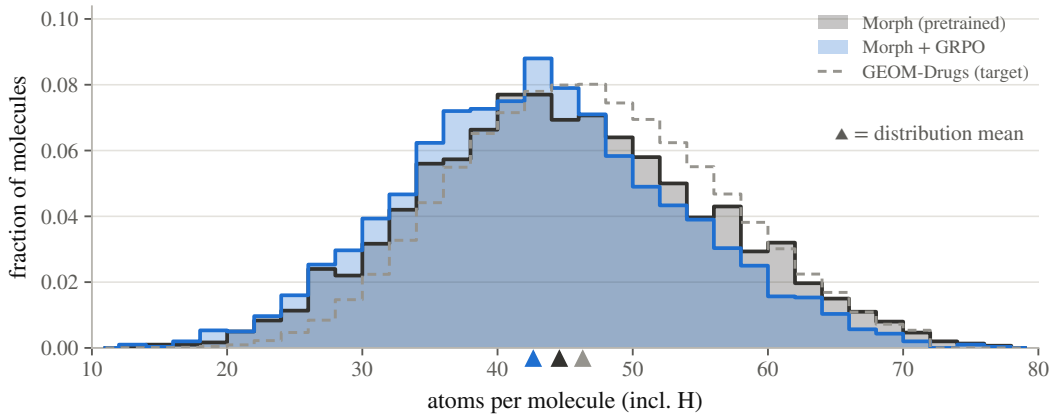


Figure 9: **Atom-count distributions before and after RL.** Morph pretrained (grey) vs. Morph+GRPO (blue), with the GEOM-Drugs target (dashed); triangles mark distribution means. Validity steering shifts mass slightly toward smaller molecules ($44.6 \rightarrow 42.6$) but leaves the overall shape close to the target, in contrast to the deliberate out-of-distribution shift of Section 3.1.

H CrossDocked Experiment Details

H.1 Hyperparameters

Table 13: Hyperparameters for the CrossDocked pocket experiment.

Hyperparameter	Value
Pretraining	
Dataset	CrossDocked2020
Split	`split_by_name.pt`, verbatim
Train / validation complexes	96,874 / 3,107
Test pockets	100
Atom vocabulary	7 tokens (C, Cl, F, N, O, P, S)
Coordinate standard deviation	2.4741
Parameters	54.0M
Epochs (optimiser steps)	676 (242,612)
Reinforcement learning	
Updates per pocket	250
Group size G (= batch size)	32
Integration steps T	100
σ_{explore}	0.01
σ_{ins}	0.01
Learning rate	3×10^{-6}
Clipped passes μ	1
Clip threshold ϵ	0.2
KL coefficient β	0
Maximum atoms	80
Reward	
Weights (affinity / QED / SA / diversity / geometry)	0.55 / 0.15 / 0.15 / 0.10 / 0.05
QED target	0.50
Normalised SA target	0.70
Geometry threshold τ	1.0
Maximum internal clashes	0
Gated-out offset	$\mu_B - 3\sigma_B$

H.2 Metrics & Pocket Scoring

Binding Affinity The model is conditioned on the 10 Å pocket carve distributed with CrossDocked2020, while every Vina number is computed against the full, uncropped receptor from the official test set, as in TargetDiff and MolCRAFT. The carve is what the generator sees; it is never what the ligand is scored against. We follow the MolCRAFT protocol throughout: AutoDock Vina 1.2 python bindings, exhaustiveness 8, and a per-molecule box sized as the generated pose's per-axis extent plus 5 Å and centred on that pose's bounding-box centre. Vina Score is `score_only` on the generated pose and Vina Min is `minimize` followed by rescoring. As reference implementations, we prepare receptors with `pdb2pqr30 --ff=AMBER` followed by AutoDockTools `prepare_receptor4.py`.

We validate the pipeline against the crystal reference ligands, with agreement to 0.01 and 0.02 kcal/-mol on both columns,

Table 14: Reference-ligand scores under our pipeline against the published values. Both are means over all 100 test pockets, one crystal ligand each.

	Vina Score	Vina Min	Pockets scored
Published [Qu et al., 2024]	−6.36	−6.71	100/100
Ours	−6.350	−6.694	100/100

Other metrics. Diversity is mean pairwise ECFP Tanimoto distance within a pocket; Size is the mean heavy-atom count; Complete is the fraction of samples that are valid and connected.

H.3 Reward for pocket fine-tuning

Rank based reward. We follow DePPA [Xue et al., 2026] in replacing each property by its within-group rank passed through the inverse normal CDF: this discards magnitude, but bounds every term's contribution to its weight times one rank range, so no property can be traded away for more than it is worth. We have $z_k = \Phi^{-1}\big(\mathrm{rank}/(G+1)\big)$, and the reward is the weighted sum $R = \sum_k w_k z_k$. Ties take the average rank, so a term on which every rollout is saturated gives $z_k = 0$ to all of them and drops out of the advantage entirely.

Geometry gate. A molecule reaches the ranking only if it is graph-valid, Vina-scorable, and passes a geometry gate built from the POSEBUSTERS `mol_fast` counts: every bond length and every bond angle valid ($\tau = 1.0$) and zero internal steric clashes. We gate rather than penalize to avoid reward hacking. Already when setting $\tau = 0.95$ the policy spends the tolerance rather than the sampler's jitter using it (all-valid $0.894 \to 0.411$ over $\sim$900 updates), while at $\tau = 1$ the mean valid bond fraction holds at 0.994 and the gate's acceptance rate is higher (0.848 vs. 0.781).

Terms and weights. The five ranked terms are affinity $-\mathrm{Vina}/\sqrt[3]{\mathrm{HAC}}$ [Shen et al., 2024], $\min(\mathrm{QED}/0.5,\, 1)$, $\min(\mathrm{SA}_{\mathrm{norm}}/0.7,\, 1)$, geometry $\min(f_{\mathrm{bonds}}, f_{\mathrm{angles}})$, and diversity 1 minus the mean Tanimoto similarity to the rest of the gated group (RDKit fingerprints, as in MolCRAFT and DePPA, so the term optimises the quantity that is reported). Weights are $0.55/0.15/0.15/0.05/0.10$; they sum to one and affinity exceeds the sum of the rest, to ensure that abandoning affinity to maximize the auxiliaries is arithmetically unprofitable. Geometry takes little weight because the gate already does that work. The QED and SA targets are benchmark levels: above them the term ties and goes inert.

The failure band. Following DePPA, rollouts that fail the gate are placed at $\mu - 3\sigma$ below the valid band, where μ and σ are that band's own mean and standard deviation, which keeps the penalty's relative strength constant across groups. We add one term: failures are additionally ranked among themselves by geometry, weighted by w_{geometry}. A flat penalty leaves an all-failing group with zero advantage and no direction to improve in, while the added spread of w_{geometry} times one rank range is far inside the offset, so a failure can never overtake a survivor.

H.4 Results.

Posebusters. Nothing is redocked, the intermolecular checks read the generated pose against its receptor. **PB-Valid** follows Qiu et al. [2025] and excludes the internal-energy check, which is reported separately over the molecules where it could be evaluated. Our crystal-ligand row reproduces their published 95.0% and 98.0%.

H.5 Pocket-level diversity

Diversity is a distribution-matching quantity rather than one to maximise. As Qiu et al. [2025] observe, the compounds that bind a given target occupy a comparatively narrow region of chemical space, since they must satisfy the same steric and pharmacophoric constraints; a generator that scores high on diversity may simply be straying from that region. We therefore measure the quantity on CrossDocked itself rather than treating larger values as better.

Each pocket carve holds exactly one ligand, so within-pocket diversity is undefined in the data and the finest observable grouping is the protein target. Real ligands sharing a target reach 0.597 over the 643 targets with at least four distinct ligands (Fig. 10a). Our base model sits at 0.584, inside that range and tracking the data's dependence on pocket size (Fig. 10b)

Every amortised baseline exceeds the data's upper end, and DePPA's 0.78 matches what 100 real ligands drawn at random ignoring the pocket entirely obtain (0.771). Two protocol differences could explain this discrepancy:

Table 15: POSEBUSTERS `dock` checks on the CrossDocked test set: 12 intramolecular plus 9 intermolecular, as pass rates over *generated* molecules (100 crystal ligands, 10,000 each for base and RL), with un-evaluable checks counted as failures. Δ is RL minus base.

Check	Reference	Morph (base)	Morph + GRPO	Δ
Implied by graph validity				
loadable	1.000	0.848	0.955	+0.107
sanitization	1.000	0.848	0.955	+0.107
InChI convertible	1.000	0.848	0.955	+0.107
all atoms connected	1.000	0.835	0.943	+0.108
no radicals	1.000	0.848	0.955	+0.107
Gated by the reward				
bond lengths	1.000	0.742	0.904	+0.162
bond angles	1.000	0.799	0.925	+0.126
internal steric clash	0.990	0.822	0.941	+0.119
Intramolecular geometry — not rewarded				
aromatic ring flatness	1.000	0.848	0.955	+0.107
non-aromatic ring non-flatness	0.980	0.828	0.935	+0.108
double bond flatness	1.000	0.810	0.911	+0.101
Intermolecular — not rewarded, needs the receptor				
protein–ligand maximum distance	1.000	0.848	0.955	+0.107
minimum distance to protein	0.980	0.763	0.932	+0.170
minimum distance to organic cofactors	1.000	0.846	0.945	+0.100
minimum distance to inorganic cofactors	1.000	0.845	0.943	+0.098
minimum distance to waters	1.000	0.848	0.955	+0.107
volume overlap with protein	1.000	0.847	0.952	+0.105
volume overlap with organic cofactors	1.000	0.848	0.955	+0.107
volume overlap with inorganic cofactors	1.000	0.848	0.955	+0.107
volume overlap with waters	1.000	0.848	0.955	+0.107
PB-Valid	0.950	0.606	0.794	**+0.187**
Energy passed (of evaluated)	0.989	0.923	0.922	−0.001
energy evaluated / generated	95/100	8122/10000	9273/10000	

1. First, MolCRAFT and MolPilot reconstruct bonds after generation from interatomic distances using OpenBabel, and draw the atom count based on the reference ligand. Morph emits the bond graph explicitly and infers cardinality from the pocket, so disconnection is a model failure rather than a reconstruction artifact, which also makes bond length and bond-angle checks substantially harder to pass.
2. Second, the models train for 15 epochs where our checkpoint is at 676. Over our own training run, pocket-level diversity falls monotonically from 0.803 at epoch 9 to 0.517 at epoch 1069. At a 15–30 epoch budget our diversity is 0.665–0.691, within the baselines' band, at connectivity too low to be usable. The diversity we give up is largely bought as validity, and our operating point sits far later on that curve.

Fine-tuning lowers diversity to 0.411, mostly due to the concentration of per-pocket policy optimization towards high reward samples, which a larger diversity weight or a KL anchor would counteract.

H.6 The size action in two pockets

Figures 11 and 12 show what the bidirectional size change of Section 3.3 looks like in structure. Under one reward and one fine-tuning procedure the policy adds four heavy atoms in one pocket and removes five in the other, improving Vina in both. These are single illustrative molecules: the lowest-Vina sample per method, PoseBusters-passing for RL and not the population means the quadrant figure reports.

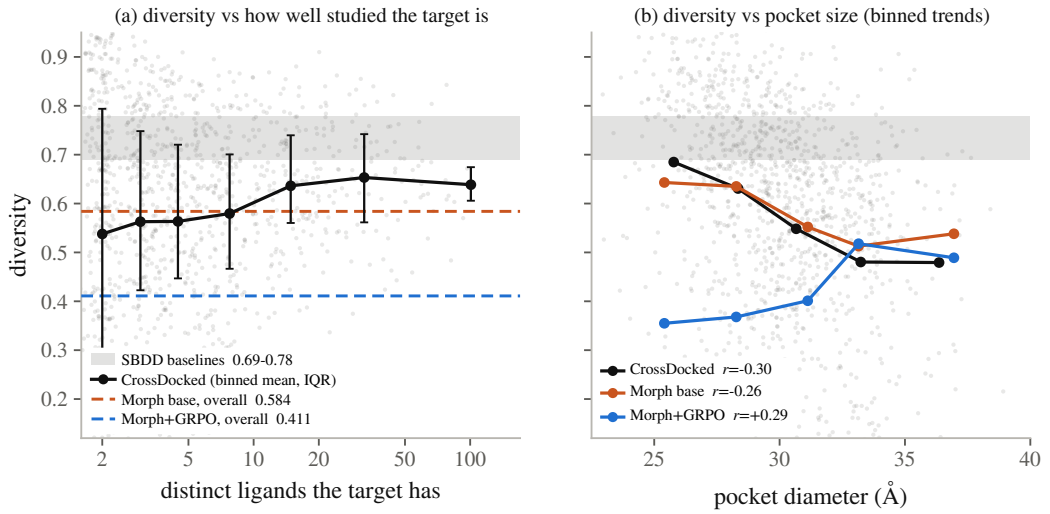
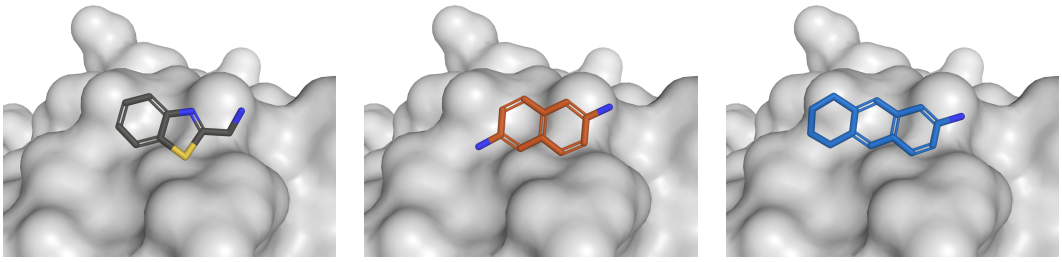
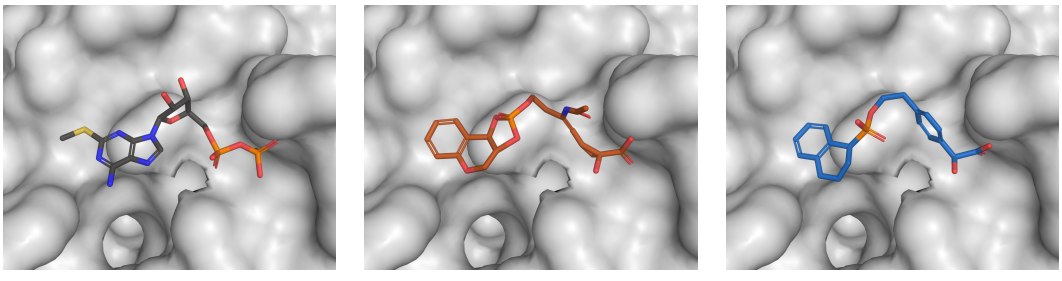


Figure 10: **Diversity in CrossDocked and in what we generate.** Grey dots are individual protein targets (1,205 carrying ≥ 2 distinct ligands), scored with the benchmark's own metric and fingerprints. **(a)** Diversity against how many distinct ligands a target carries: black is the binned mean with interquartile range, dashed lines are our two models' overall values, and the band spans the amortised SBDD baselines (0.69–0.78). **(b)** Binned trends against pocket diameter.



(a) crystal ligand, 11 heavy atoms (b) Morph (base), 12 (c) Morph + GRPO (ours), 15

Figure 11: **UBE2T: the base generator makes small ligands and the policy grows them.** Crystal ligand (11 heavy atoms, -4.23), best base molecule (12, -5.50) and best RL molecule (15, -7.13 kcal/mol Vina `score_only`). All three panels share one camera; the pocket surface is cut away in front of the ligand.



(a) crystal ligand, 29 heavy atoms (b) Morph (base), 31 (c) Morph + GRPO (ours), 27

Figure 12: **P2Y12: the base generator makes large ligands and the policy shrinks them.** Crystal ligand (29 heavy atoms, -9.81), best base molecule (31, -8.76) and best RL molecule (27, -9.80 kcal/mol). The RL molecule matches the crystal ligand's score with two fewer heavy atoms. All three panels share one camera; the pocket surface is cut away in front of the ligand.

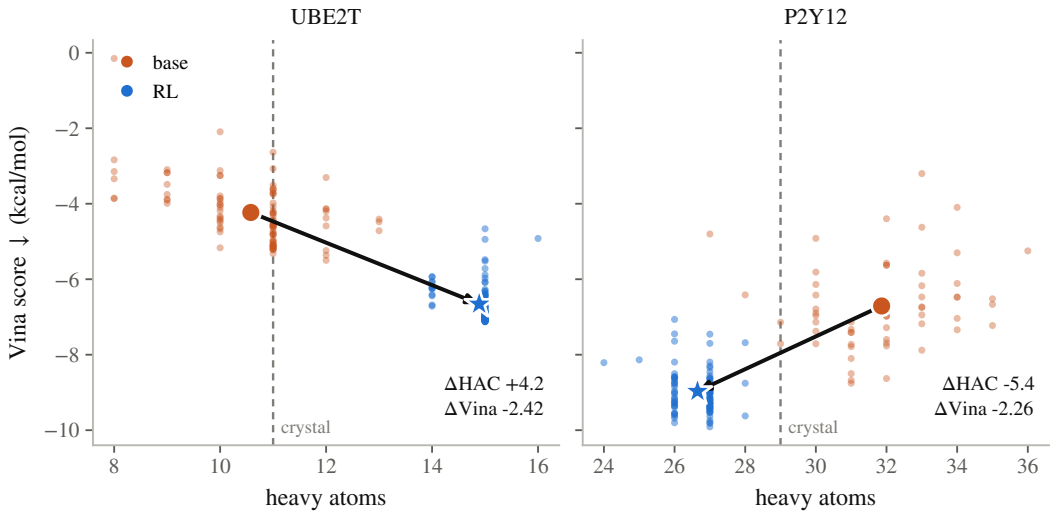


Figure 13: **Cardinality moves in both directions under one reward and one procedure.** Every point is a single generated molecule, all 100 samples per pocket that were valid and Vina-scorable; the filled circle and star are the base and fine-tuned means, joined by an arrow, and the dashed line marks the crystal ligand's heavy-atom count. *Left,* UBE2T (pocket 43): the base generator averages 10.6 ± 1.1 heavy atoms at -4.23 kcal/mol and the policy grows it to 14.9 ± 0.4 at -6.66. *Right,* P2Y12 (pocket 60): the base averages 31.9 ± 1.8 at -6.71 and the policy shrinks it to 26.6 ± 0.6 at -8.97. Ligand efficiency improves in both.