Budgeted Quotient-Residual Guidance for Frozen Pocket-Conditioned Molecular Diffusion

Xinyu Wang[1]* **Jinbo Bi**[1] **Minghu Song**[2]
[1]Department of Computer Science and Engineering
University of Connecticut, Storrs, CT 06269, USA
[2]Institute of Health and Medicine
Hefei Comprehensive National Science Center, Hefei 230601, China

Abstract

Pocket-conditioned molecular diffusion updates ambient atom coordinates, but many lead-optimization objectives are expressed on quotient features such as distances, contacts, and anchored substructures. We introduce *budgeted quotient-residual guidance* (QRG), an inference-time correction that makes these objectives active without retraining the molecular generator. QRG lifts quotient covectors to metric-horizontal directions and bounds their delivery using the frozen sampler's own step norm: quotient geometry chooses the direction, while sampler motion sets the scale. We derive an ideal sampler-scaled update, its geometric and kinetic interpretation, and equivariance conditions, with separately budgeted section and residual controls for context compatibility and relational geometry. Controlled quotient tasks show that budgeted delivery activates signals that raw local guidance leaves dormant. On frozen TARGETDIFF, LOCAL-QRG improves fragment reconstruction from 81.9% to 86.8%. QRG also transfers to the frozen DiffSBDD sampler, improving reconstruction-and-connectivity yield from 38.8% to 43.8% and demonstrating applicability beyond the original molecular backbone. Fragment RDKit validity and QED also improve, with a diversity trade-off and task-dependent budget requirements. QRG offers a lightweight route to quotient-aware molecular inference, with no retraining or rollout denoiser calls in its deployed variants and explicit runtime accounting.

1 Introduction

Pocket-conditioned molecular diffusion has two coordinate systems. The sampler evolves ambient atom coordinates, but many scientific objectives are defined only after quotienting away nuisance choices such as global rotation, translation, atom-index representatives, or irrelevant local coordinates. Fragment geometry, scaffold preservation, linker closure, and pocket-relative contacts are naturally described by distances, anchors, and substructure features rather than by one absolute coordinate representative. This ambient–quotient mismatch is present even when the backbone itself is equivariant: the model can move coordinates correctly under symmetry, while the guidance objective still lives on features that identify an equivalence class.

There are two broad ways to handle the mismatch. One can build the structure into training through equivariant architectures or symmetry-aware objectives [Thomas et al., 2018, Fuchs et al., 2020, Satorras et al., 2021, Hoogeboom et al., 2022, Tong et al., 2025]. This is principled, but it requires changing or retraining the generator. Alternatively, one can keep a pretrained sampler fixed and add plug-and-play guidance at inference time [Dhariwal and Nichol, 2021, Ho and Salimans, 2022, Chung et al., 2023, Bansal et al., 2023]. This second regime is attractive in structure-based drug

*Corresponding author: xinyu.wang@uconn.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

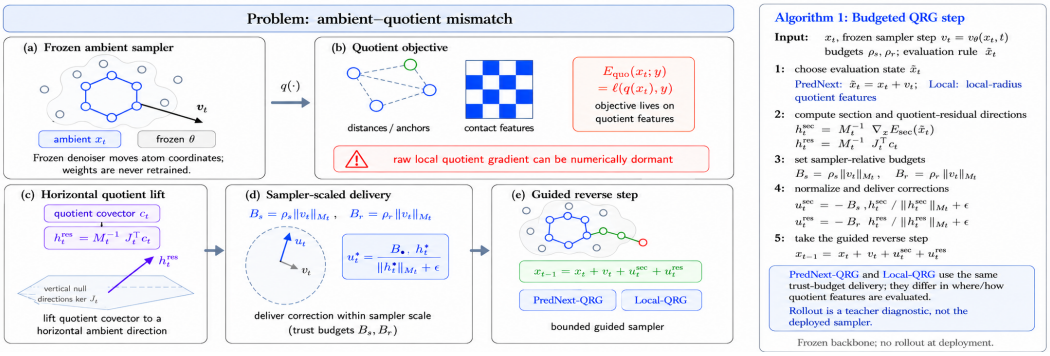

Figure 1: Budgeted quotient-residual guidance. Quotient geometry selects a relational correction direction; sampler motion sets its scale. The schematic shows a numerically stabilized illustration of the ideal normalized rule. PredNext and Local both use the predicted-next state, with different residual energies. Molecular runs use branch clipping followed by total and absolute caps (equation (27)). Both deployed variants avoid rollout denoiser calls; rollout is a diagnostic reference.

design because strong pretrained backbones and benchmark code are already available, while retraining for every lead-optimization objective can be expensive or impractical [Guan et al., 2023a, Lin et al., 2025].

Inference-time quotient guidance, however, has a failure mode that is easy to miss. A quotient energy removes nuisance directions, but the frozen sampler still takes finite ambient denoising steps. Direct local gradients of an invariant energy can be directionally meaningful yet numerically dormant: their scale is set by the quotient feature map rather than by the sampler motion. Rollout lookahead can activate such signals by repeatedly querying the denoiser, but then guidance becomes an expensive oracle rather than a practical inference rule. Sampler-relative delivery uses the scale already chosen by the sampler, avoiding separate calibration of an absolute coordinate radius. The correction adapts to changing denoising-step magnitudes through a dimensionless motion budget.

We propose BUDGETED QUOTIENT-RESIDUAL GUIDANCE, a quotient-lifted trust-region formulation for frozen molecular diffusion samplers. The quotient differential first lifts a quotient covector into a metric-horizontal ambient direction, i.e., the ambient motion that changes quotient features most efficiently while ignoring vertical representative motion. The allowable correction length is then set by the frozen sampler's own update norm, which is also the local kinetic scale of the frozen reverse step. In one sentence, *quotient geometry chooses direction; sampler motion sets scale*. This converts a dormant invariant signal into an active, bounded correction while keeping the method compatible with pretrained backbones. The resulting plug-and-play formulation lets task-specific relational objectives steer an existing molecular generator while its learned parameters remain fixed. Expensive rollout serves as a teacher or diagnostic, while deployed instantiations reuse either the sampler's predicted next state or a cheap local quotient surrogate.

Our empirical focus is the frozen-backbone setting. We integrate QRG with TARGETDIFF [Guan et al., 2023a] on CBGBENCH-style lead-optimization tasks [Lin et al., 2025], and evaluate transfer to the frozen DiffSBDD de novo sampler [Schneuing et al., 2024]. This choice makes the runtime question explicit: a useful inference-time method should improve generation quality without hiding a many-rollout cost. We therefore report validity, novelty, diversity, control budget, wall-clock runtime, paired target-level statistics, and corrected docking diagnostics. The main ligand-generation/editing slice covers fragment growing, scaffold hopping, and linker design. Side-chain decoration tests budget transfer to a different ligand decomposition and fixed-context configuration; the controller does not update protein side chains.

Our contributions are:

1. **A quotient-lifted trust-region formulation.** We derive a metric-horizontal lift of quotient covectors and a closed-form sampler-relative correction scaled by the frozen sampler's motion. In a conditional Gaussian mean-shift model, the ideal update solves a local KL-budgeted quotient-descent problem; deterministic samplers use the corresponding kinetic trust region. PredNext and

Local instantiate this geometry through full pairwise and local-neighbor energies with clipped delivery.

2. **Theory that matches the implementation questions.** We prove the horizontal lift, sampler-scaled descent update, conditional KL/kinetic interpretation, and equivariance under compatible isometric actions, and characterize how independent branch budgets determine the effective weighting of section and residual directions.
3. **Generation quality, transfer, and mechanism.** Controlled tasks isolate the activation of dormant quotient signals. On frozen TARGETDIFF, LOCAL-QRG improves fragment reconstruction from 81.9% to 86.8%, alongside higher RDKit validity and QED. Integration with DiffSBDD extends the method to a second frozen sampler and increases reconstruction-and-connectivity yield by 5.00 percentage points. Direction controls identify the role of the quotient-residual direction, while parameter studies characterize budget sensitivity and the trade-off between guidance frequency and sampling cost.

2 Background

Groups, invariants, and quotient coordinates. Let a group G act on a data space Ω by $g \circ x$. A function $f : \Omega \to \mathbb{R}$ is G-invariant if $f(g \circ x) = f(x)$ for every $g \in G$, and a map $F : \Omega \to \Omega$ is G-equivariant if $F(g \circ x) = g \circ F(x)$. Molecular coordinates are typically considered up to rigid motions, and molecular graphs may have atom-index automorphisms. A quotient representation $q : \Omega \to \mathcal{Q}$ identifies states within the same equivalence class: $q(g \circ x) = q(x)$. Pairwise distances are the simplest example for rigid-motion invariance; pocket-relative or substructure-restricted distances give task-specific quotients.

Diffusion samplers and frozen-backbone guidance. A denoising diffusion model constructs noisy states x_t and learns a reverse-time denoising update [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020, Song et al., 2021, Nichol and Dhariwal, 2021, Kingma et al., 2021, Karras et al., 2022]. We write one sampler step abstractly as

$$x_{t-1} = x_t + v_\theta(x_t, t), \tag{1}$$

where v_θ includes the neural denoiser and the numerical update rule. Plug-and-play guidance modifies this step by adding a correction u_t derived from an auxiliary objective, giving $x_{t-1} = x_t + v_\theta(x_t, t) + u_t$. Classifier guidance and classifier-free guidance popularized this idea in image generation [Dhariwal and Nichol, 2021, Ho and Salimans, 2022]; later methods use diffusion guidance for inverse problems, restoration, inpainting, and general differentiable objectives [Chung et al., 2023, Kawar et al., 2022, Lugmayr et al., 2022, Bansal et al., 2023]. Our setting differs in that the objective is often invariant while the correction must be delivered in ambient molecular coordinates.

Equivariant molecular diffusion. Equivariant neural networks encode symmetry by construction [Cohen and Welling, 2016, Kondor and Trivedi, 2018, Thomas et al., 2018, Weiler et al., 2018, Anderson et al., 2019, Fuchs et al., 2020, Finzi et al., 2020, Satorras et al., 2021, Batzner et al., 2022]; broader surveys frame these architectures as part of geometric deep learning [Bronstein et al., 2021]. Molecular diffusion models use these ideas to generate 3D conformers, molecules, docking poses, linkers, and ligands in protein pockets [Gebauer et al., 2019, Hoogeboom et al., 2022, Xu et al., 2022, Jing et al., 2022, Corso et al., 2023, Guan et al., 2023a, Igashov et al., 2024, Schneuing et al., 2024]. Recent work also revisits symmetry-aware training itself: Orbit Diffusion treats data augmentation as a Monte Carlo gradient estimator and Rao-Blackwellizes over group orbits to reduce variance without extra neural function evaluations [Tong et al., 2025]. Our paper is complementary. Rather than changing the training objective, we ask how to use quotient information at sampling time when the backbone is frozen.

Structure-based generation and lead optimization. Pocket-conditioned 3D molecular generation aims to produce valid molecules that satisfy geometric, chemical, and interaction constraints inside a protein binding pocket. Early models used 3D density grids [Ragoza et al., 2022]; later models introduced equivariant graph sampling [Peng et al., 2022] and diffusion backbones [Guan et al., 2023a, Schneuing et al., 2024, Igashov et al., 2024, Guan et al., 2023b]. Linker and fragment-editing settings also have task-specific graph and 3D generative models [Imrie et al., 2020]. CBGBENCH

unifies de novo generation and lead-optimization tasks as completion of a protein-molecule binding graph, including linker design, fragment growing, scaffold hopping, and side-chain decoration [Lin et al., 2025]. This multi-task protocol lets us evaluate quotient guidance across different preserved substructures, generated ligand regions, and pocket-conditioned editing objectives.

Molecular quality and docking metrics. Generation validity, novelty, diversity, and substructure consistency measure different aspects of molecular quality, and benchmark suites for molecular generation standardize related distributional and optimization metrics [Brown et al., 2019, Polykovskiy et al., 2020]. QED summarizes drug-likeness through physicochemical descriptors [Bickerton et al., 2012]. Docking scores such as AutoDock Vina are useful diagnostics, but raw scores can be affected by molecular size, strain, coordinate-frame choices, and post-processing conventions [Wang et al., 2005, Trott and Olson, 2010, Eberhardt et al., 2021]. Recent structure-based benchmarks and datasets emphasize that raw pocket-based 3D generators can produce invalid conformations and that local relaxation or dataset construction can substantially change apparent affinity estimates [Francoeur et al., 2020, Baillif et al., 2024]. Our metric pipeline relies on standard cheminformatics toolkits for molecular parsing and conversion [RDKit contributors, 2026, O'Boyle et al., 2011]. We therefore separate the main generation claim from docking diagnostics and use corrected raw-pocket docking only as supporting evidence.

3 Budgeted quotient-residual guidance

We describe a local, first-order quotient-lifted trust-region update for a frozen sampler. Figure 1 illustrates the ideal construction; below we relate it to the clipped molecular implementation. Pred-Next and Local use full pairwise and local-neighbor energies on the predicted-next state, leaving generator weights and the denoising schedule fixed.

A concrete relation. For generated atom x_i and pocket atom p_j, let $q_{ij} = \|x_i - p_j\|_2$ and $E = (d_{\text{clash}} - q_{ij})_+^2$. Below the clash distance, quotient descent increases q_{ij}, moving the ligand atom away from the pocket. The trust budget sets displacement relative to sampler motion. Joint ligand–pocket rigid motions preserve this distance; ligand-only motions with a fixed pocket need not.

3.1 Quotient differential and horizontal lift

Let $x \in \Omega \subset \mathbb{R}^d$ denote ambient molecular coordinates and let $q : \Omega \to \mathcal{Q} \subset \mathbb{R}^m$ be a local quotient chart, such as pocket-relative distances, anchor distances, or substructure features. A task energy has the form

$$E_{\text{quo}}(x; y) = \ell(q(x), y), \tag{2}$$

where y is specified in quotient space. Write $J_x = Dq_x \in \mathbb{R}^{m\times d}$ for the quotient differential. Ambient motions in $\ker J_x$ are vertical: to first order they move the representative without moving the quotient features. A quotient covector $c_x = \nabla_z \ell(z, y)|_{z=q(x)}$ pulls back to the ambient covector $J_x^\top c_x$, but this pullback alone does not say how far a frozen sampler should move.

To separate quotient geometry from delivery scale, choose a positive definite sampler metric M_x on ambient motions. Official molecular runs use the Euclidean metric; sensitivity screens also test diagonal coordinate preconditioners. We work locally where q is differentiable, with neighbor identities fixed for local features. The metric-horizontal lift of the quotient covector is

$$h_x = M_x^{-1} J_x^\top c_x. \tag{3}$$

It is horizontal because $\langle h_x, w\rangle_{M_x} = 0$ for every $w \in \ker J_x$, where $\langle a, b\rangle_M = a^\top M b$. Thus equation (3) is a metric-horizontal ambient representative of the local quotient gradient; the normalized direction $-h_x/\|h_x\|_{M_x}$ is the corresponding steepest quotient descent direction under the metric.

Proposition 1 (Horizontal steepest lift)**.** *Let $J \in \mathbb{R}^{m\times d}$, $c \in \mathbb{R}^m$, and $M \succ 0$. If $h = M^{-1}J^\top c \neq 0$, then $-h/\|h\|_M$ solves*

$$\min_{\|u\|_M \le 1} \langle c, Ju\rangle . \tag{4}$$

Moreover h is M-orthogonal to $\ker J$.

The lift chooses the steepest quotient-descent direction; sampler-relative delivery below sets its length.

3.2 Sampler-relative trust-region delivery

At denoising step t, the frozen sampler proposes $v_t = v_\theta(x_t, t)$. We use the sampler's own step norm to define the trust radius

$$B_t = \rho_t \|v_t\|_{M_t}, \tag{5}$$

where $\rho_t \geq 0$, $M_t = M_{\tilde{x}_t}$ and $\tilde{x}_t$ is the state where the quotient direction is evaluated. The ideal quotient correction solves the local problem

$$\min_{u \in \mathbb{R}^d} \langle c_t, J_t u \rangle \quad \textit{s.t.} \quad \|u\|_{M_t} \leq B_t. \tag{6}$$

For $h_t \neq 0$, the solution is a sampler-scaled horizontal lift:

$$u_t^{\mathrm{quo}} = -B_t \frac{h_t}{\|h_t\|_{M_t}}, \qquad h_t = M_t^{-1} J_t^\top c_t. \tag{7}$$

We set $u_t^{\mathrm{quo}} = 0$ when $h_t = 0$. This ideal rule delivers radius B_t for a nonzero lift. The molecular implementation clips each direction to its branch budget, then applies total and absolute caps (equation (27)); its budgets are upper bounds rather than mandatory step lengths.

Proposition 2 (Sampler-scaled quotient descent)**.** *Assume $h_t \neq 0$. The update in equation* (7) *is the unique optimizer of equation* (6)*. Its first-order quotient-energy change at $\tilde{x}_t$ is*

$$\langle c_t, J_t u_t^{\mathrm{quo}} \rangle = -\rho_t \|v_t\|_{M_t} \|h_t\|_{M_t}. \tag{8}$$

Thus the correction norm is set by the sampler-relative budget; the linearized energy decrease also depends on $\|h_t\|_{M_t}$. The statement concerns the local correction, not the full denoising step.

KL/kinetic interpretation. Consider a Gaussian reference kernel with $\tau_t > 0$ and $v_t, M_t, \tilde{x}_t$ fixed conditional on x_t:

$$P_t(\mathrm{d}x_{t-1} \mid x_t) = \mathcal{N}\left(\mu_t, \tau_t M_t^{-1}\right), \qquad \mu_t = x_t + v_t, \tag{9}$$

a control u fixed conditional on x_t shifts the mean while preserving covariance, giving

$$\mathrm{KL}\left(P_t^u(\cdot \mid x_t) \,\|\, P_t(\cdot \mid x_t)\right) = \frac{1}{2\tau_t} \|u\|_{M_t}^2. \tag{10}$$

Thus equation (6) is a conditional KL budget in this Gaussian model. Deterministic samplers use the corresponding quadratic kinetic cost directly [Kappen, 2005, Todorov, 2009]. Setting $B_t = \rho_t \|v_t\|_{M_t}$ ties the correction scale to sampler motion. The experimental adapter computes control after sampling the proposal; its noise-dependent update is specified in appendix C.

Proposition 3 (KL-budgeted quotient descent)**.** *Assume the Gaussian reference kernel above, $h_t \neq 0$ and $\kappa_t \geq 0$. Among conditional mean-shift controls satisfying*

$$\mathrm{KL}\left(P_t^u(\cdot \mid x_t) \,\|\, P_t(\cdot \mid x_t)\right) \leq \kappa_t, \tag{11}$$

the update that minimizes the linearized quotient loss $\langle c_t, J_t u \rangle$ is

$$u_t^\star = -\sqrt{2\tau_t \kappa_t} \frac{h_t}{\|h_t\|_{M_t}}. \tag{12}$$

The QRG radius $B_t = \rho_t \|v_t\|_{M_t}$ corresponds to the conditional budget $\kappa_t = \rho_t^2 \|v_t\|_{M_t}^2 / (2\tau_t)$.

3.3 Section-residual product budget

Some molecular tasks require preserving a concrete representative as well as improving quotient features. We therefore use a product trust region with independently budgeted section and residual directions. The section branch uses $h_t^{\mathrm{sec}} = M_t^{-1} \nabla_x E_{\mathrm{sec}}(\tilde{x}_t)$ to stabilize anchors or preserved atoms. The residual branch uses the quotient-horizontal lift $h_t^{\mathrm{res}} = M_t^{-1} J_t^\top c_t$ from equation (3). The ideal branchwise update is

$$u_t = -\rho_s \|v_t\|_{M_t} \frac{h_t^{\mathrm{sec}}}{\|h_t^{\mathrm{sec}}\|_{M_t}} - \rho_r \|v_t\|_{M_t} \frac{h_t^{\mathrm{res}}}{\|h_t^{\mathrm{res}}\|_{M_t}}. \tag{13}$$

Zero-direction branches contribute zero. The endpoint $\rho_r = 0$ is SECTION-ONLY. Active variants allocate $\rho_r > 0$. The branchwise problems separately minimize their linearized objectives under their norm budgets. Their sum allocates motion to representative preservation and quotient-residual descent; the deployed clipping and final caps are given in equation (27).

Proposition 4 (Branch allocation and effective weighting)**.** *Let $a, b \in \mathbb{R}^d$ be non-collinear and $B_a, B_b > 0$. The split $s = -B_a a/\|a\| - B_b b/\|b\|$ equals $-\|s\|\,(a + \lambda_* b)/\|a + \lambda_* b\|$ for*

$$\lambda_* = \frac{B_b \|a\|}{B_a \|b\|}.$$

A fixed scalarization weight reproduces one branch-budget ratio; varying that ratio requires an adaptive weight.

Relation to normalized guidance. Normalized guidance also divides by a gradient norm. QRG additionally ties delivery to the frozen sampler's current motion and normalizes section and residual directions separately. Its dimensionless branch budgets specify how much sampler-relative motion is allocated to each objective, separating delivery scale from the relative magnitudes of the raw gradients. Section 4.4 evaluates direction dependence and compares this step-adaptive delivery with a calibrated absolute radius.

3.4 Instantiating the quotient direction

The trust-region mechanism requires a relational energy and an evaluation state $\tilde{x}_t$, without requiring rollout. PredNext and Local use the same predicted-next state and clipped delivery, but choose different residual energies: full pairwise clash penalties and local neighbor-radius penalties, respectively (appendix C). The rollout teacher changes the evaluation state.

Predicted-next quotient-residual guidance. PREDNEXT-QRG reuses the frozen sampler's own one-step prediction, $\tilde{x}_t = x_t + v_t$, and applies the residual correction every two denoising steps in the reported configuration. It is useful when the predicted next state supplies the relevant relation without engineering a local surrogate; it does not require a separate rollout trajectory.

Local-radius quotient surrogate. LOCAL-QRG constructs the quotient residual from local neighbor distances, using four neighbors per generated atom in the default configuration. This focuses the correction on short-range relations when they capture the task objective. Like PREDNEXT-QRG, it evaluates features on the sampler's predicted next state and requires no rollout denoiser calls. The local energy and feature construction are specified in appendix C; a development ablation in table 22 compares the surrogate with sampled-pair alternatives.

Shared-rollout teacher. ROLLOUT-QRG TEACHER evaluates the quotient direction after one explicit short rollout and is a higher-cost diagnostic reference. It is not the deployed method or a performance upper bound. Reporting it separately makes the additional rollout cost explicit.

3.5 Equivariance of the lift and delivery

Let G be an isometry group acting linearly by R_g on ambient coordinates. The quotient chart is invariant if $q(R_g x) = q(x)$, and the metric is equivariant if $M_{R_g x} = R_g M_x R_g^\top$.

Proposition 5 (Equivariant quotient-lifted correction)**.** *Assume $q(R_g x) = q(x)$, $M_{R_g x} = R_g M_x R_g^\top$, and $v_\theta(R_g x, t) = R_g v_\theta(x, t)$. Also assume the evaluation state transforms as $\tilde{x}_t(R_g x_t) = R_g \tilde{x}_t(x_t)$. Then the horizontal lift and ideal residual correction satisfy*

$$h_{R_g x} = R_g h_x, \qquad u_t^{\mathrm{quo}}(R_g x_t) = R_g u_t^{\mathrm{quo}}(x_t). \tag{14}$$

The correction preserves symmetry under these compatibility conditions. Pocket and anchor coordinates transform jointly for pocket-conditioned features.

Proofs and implementation details are provided in appendices B and C.

Table 1: Controlled quotient tasks under matched denoising budgets. Errors are quotient MMD for Gaussian 2D and invariant template error for the geometry tasks.

Task	Method	Success ↑	Key error ↓	NFE
Gaussian 2D	BASE	0.159	0.528	64
	Dormant local	0.159	0.529	64
	Budgeted residual	**0.246**	**0.451**	64
Orbit points	BASE	0.117	0.406	72
	Dormant local	0.117	0.406	72
	Budgeted residual	**0.310**	**0.299**	72
Toy molecules	BASE	0.079	0.514	80
	Dormant local	0.079	0.514	80
	Budgeted residual	**0.371**	**0.356**	80

4 Experimental setup and results

We ask whether sampler-relative delivery activates otherwise dormant quotient signals, whether the direction matters beyond added motion, and how the method transfers across tasks and frozen samplers.

Evidence map. The principal TARGETDIFF comparison uses three seeds and target-paired reconstruction rates. An independent eight-target DiffSBDD experiment evaluates transfer to a second frozen sampler. Twelve-target screens study parameters and scale selection. Full follow-up results and earlier seed-0/matched-slice diagnostics are distinguished in appendix D; their cohorts are not pooled. Protocols, configurations and statistical procedures are specified in appendix C. Docking is a supporting diagnostic, not biological validation.

4.1 Controlled quotient tasks

Three controlled tasks isolate objectives defined in quotient space: Gaussian 2D with nuisance rotations, orbit-point matching, and a toy 3D molecule task with invariant distance templates. Base, raw local quotient guidance, and budgeted residual guidance share the same denoising schedule. The tasks provide controlled tests of how quotient geometry and delivery scale affect guidance.

Raw local guidance remains close to Base, whereas budgeted delivery increases success from 0.159 to 0.246, 0.117 to 0.310, and 0.079 to 0.371 respectively (table 1). With the denoising schedule held fixed, these gains isolate the activation of quotient signals through budgeted delivery.

4.2 Three-seed TargetDiff evaluation

The frozen TARGETDIFF evaluation covers fragment growing, scaffold hopping, and linker design in the CBGBENCH lead-optimization protocol. Base, section-only, PredNext and Local use identical target sets and seeds 0/1/2, with 100 generations per target per seed. There are 61 fragment, 43 linker and 64 scaffold targets. The headline benchmark-validity metric is the reconstruction rate; RDKit validity is evaluated separately. Absolute per-seed results, including section-only, appear in table 5.

LOCAL-QRG improves fragment reconstruction from 81.91% to 86.80%, a gain of 4.89 percentage points with corrected $p = 0.00147$ (table 2). It also has positive mean differences on scaffold and linker, of 4.37 and 1.52 percentage points, respectively. PREDNEXT-QRG gives positive mean differences on fragment and scaffold and a negative difference on linker. The Local fragment gain is the statistically significant result after correction across the six comparisons; uncertainty in the remaining comparisons is reported in table 4.

Chemistry and diversity trade-off. The fragment chemistry follow-up reports increased RDKit validity and QED together with decreased uniqueness and diversity (figure 2a); novelty is unchanged. Reconstruction counts saved outputs, whereas RDKit validity additionally requires successful sanitization, explaining the distinct gains $+0.04891$ and $+0.05000$. Full quality results for all three editing tasks and section-only are in tables 7 to 9.

Table 2: Three-seed reconstruction rates and target-paired differences versus Base. W/L/T counts seed-averaged target wins/losses/ties. Exact target-cluster sign-flip tests receive Holm correction across the six headline comparisons. Confidence intervals are reported in table 4.

Task	Method	Base	Guided	Δ	W/L/T	Holm p
Fragment	LOCAL-QRG	0.81907	0.86798	+0.04891	46/13/2	0.00147
Fragment	PREDNEXT-QRG	0.81907	0.83776	+0.01869	37/21/3	0.07732
Linker	LOCAL-QRG	0.68938	0.70457	+0.01519	28/15/0	1.00000
Linker	PREDNEXT-QRG	0.68938	0.67915	-0.01023	26/17/0	1.00000
Scaffold	LOCAL-QRG	0.66484	0.70849	+0.04365	47/17/0	0.59886
Scaffold	PREDNEXT-QRG	0.66484	0.68557	+0.02073	47/17/0	1.00000

4.3 Independent frozen DiffSBDD transfer

We also integrate QRG into the official CBGBENCH DiffSBDD de novo sampler [Schneuing et al., 2024], holding the backbone frozen. The same eight targets receive 20 samples each per method. With guidance disabled, the adapter delegates directly to the native sampler; implementation checks are detailed in appendix C.

LOCAL-QRG raises yield by 5.00 percentage points (12.9% relatively) and improves accepted counts on 5/8 targets (figure 2b). All methods cover every target. All accepted SDFs independently re-sanitize in RDKit, and canonical-SMILES uniqueness is 100% within each accepted set. Mean/median accepted heavy-atom counts are 20.48/19 for unguided and 20.09/19 for Local. These accepted-set checks do not imply 100% validity among all 160 requested generations.

Together, the increased yield and target-level gains extend quotient-residual guidance beyond TARGETDIFF to a second frozen molecular sampler. Mechanism telemetry characterizes the delivered corrections in table 11.

4.4 Direction and delivery-scale controls

LOCAL-QRG outperforms the perpendicular control by 3.31 percentage points on fragment (61 targets; Monte Carlo $p = 0.00010$) and 5.94 points on scaffold (64 targets; $p = 0.00001$) in the formal seed-0 comparison. The control replaces the reference correction with an orthogonal direction after rigid-motion projection, targeting the same magnitude. These gains show that the choice of correction direction matters; figure 3a compares absolute rates with the same-cohort Base. The unadjusted tests are separate from the six headline tests. Full paired results and additional random/mismatched/opposite-direction controls appear in appendix D.5. Orthogonality to the correction is distinct from membership in the full Jacobian nullspace.

A fixed-absolute baseline calibrates its coordinate radii on four targets per task, freezes them, and evaluates 12 disjoint targets. Fragment/scaffold reconstruction rates are $0.853/0.763$, versus $0.863/0.760$ for sampler-relative Local; neither paired difference is statistically significant. Sampler-relative delivery avoids this separate absolute-radius calibration by scaling the correction to each reverse step. Calibration matches mean delivered norms within 5%, with stepwise norms free to differ on held-out targets. Frozen radii and paired intervals are in appendix D.5.

Parameter selection and sensitivity. Matched 12-target fragment/scaffold screens use 25 samples per target; full neighborhood, metric, budget, allocation and frequency results are in appendix D.4. Across 2/4/8 neighbors, reconstruction ranges from 86.3% to 89.0% on fragment and 76.0% to 80.3% on scaffold. Identity, contact-density and motion-diagonal metrics give 86.3–88.7% and 76.0–77.3%, respectively (table 12). The default neighborhood and metric are therefore not isolated favorable settings.

For a new task, first encode the invariant relation: prefer Local when a small neighborhood captures it, and compare PredNext when full pairwise interactions are relevant. On validation targets, compare with Base and section-only, starting from the identity metric and a small residual budget. Check reconstruction, chemical quality and the delivered correction/base ratio before increasing the budget; test smaller section budgets if section-only is detrimental. Choose the least frequent guidance that

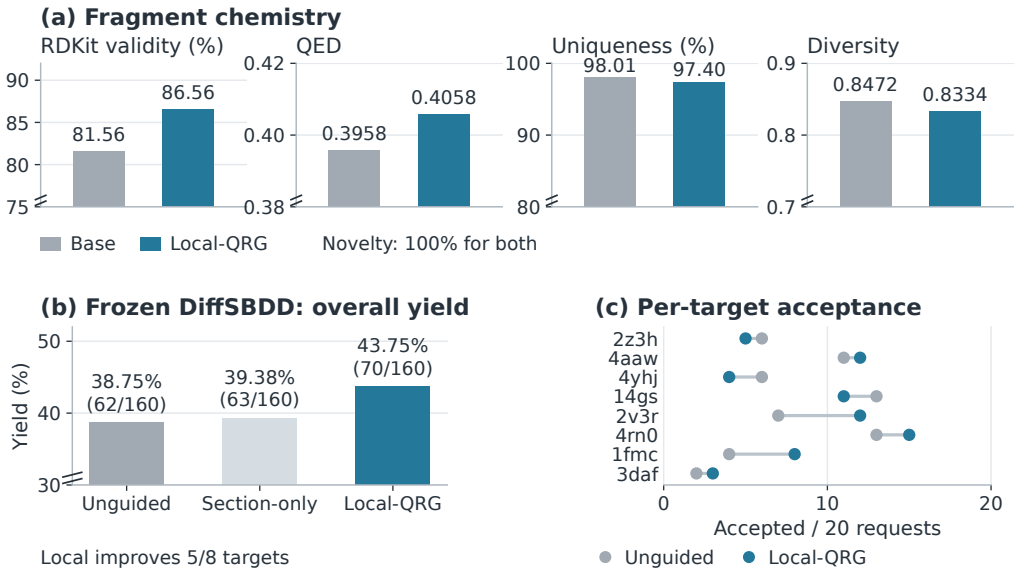

Figure 2: Chemistry trade-offs and frozen-sampler transfer. (a) Fragment chemistry: target-balanced means across seeds 0/1/2, with metric-specific denominators (appendix C). Higher is preferable for the four displayed metrics. (b) DiffSBDD reconstruction-and-connectivity yield on eight targets, 20 requests each; counts are accepted/requested. (c) Paired acceptance counts, labeled by pocket PDB ID; gray is unguided and blue is Local. Yield requires a connected fragment retaining at least 80% of generated atoms. Full intervals and telemetry are in tables 7 and 11.

meets the validation quality–cost requirement, then freeze the variant, budgets and frequency before test evaluation. This is a recommended selection procedure, not a universal setting.

Figure 3b motivates testing Local every four or eight steps when sampling cost matters. Moving from every step to every eight steps reduces sampling time from 22.07 to 17.13 seconds/request on fragment and from 21.93 to 16.73 on scaffold, with reconstruction changes of 86.33% to 86.00% and 76.00% to 76.67%. On these matched screens, less frequent guidance therefore reduces sampling cost with reconstruction changes below one percentage point. Budget choice remains task-specific (figure 3c): nonzero Local budgets all fall below the fragment screen's 92.00% Base rate, while several exceed scaffold's 74.33%. The allocation screen varies both branch radii, not their ratio at fixed total budget. These screens are separate from the full three-seed benchmark. Controller-only timings and memory use, including the cost of dense feature construction, are reported in table 31.

4.5 Task transfer and docking evaluation

Side-chain decoration evaluates a different ligand decomposition with fixed protein and ligand-context coordinates. In the earlier default-setting sweep, PredNext/Local/section-only have validity $0.848/0.837/0.828$, below the unguided rate of 0.873. A lower-budget Local configuration uses residual budget 0.01 every four steps on 64 targets across three seeds. Base/Local reconstruction rates are $0.87286/0.86865$, with paired difference -0.00422 ($[-0.01516, 0.00620]$; appendix D.6). This lower-budget comparison shows no statistically significant gain.

The docking audit covers four tasks with eight targets each and 288 requested calls. Scaffold PredNext has guided-minus-Base mean -0.328 kcal/mol, target-bootstrap interval $[-0.667, -0.023]$, and 6/8 target wins; Local has mean -0.190. Linker PredNext is unfavorable ($+2.786$, interval $[-0.074, 8.132]$). All eight comparisons are in table 21. Earlier docking results are reported separately in table 26. Neither analysis establishes binding activity or a universal docking winner.

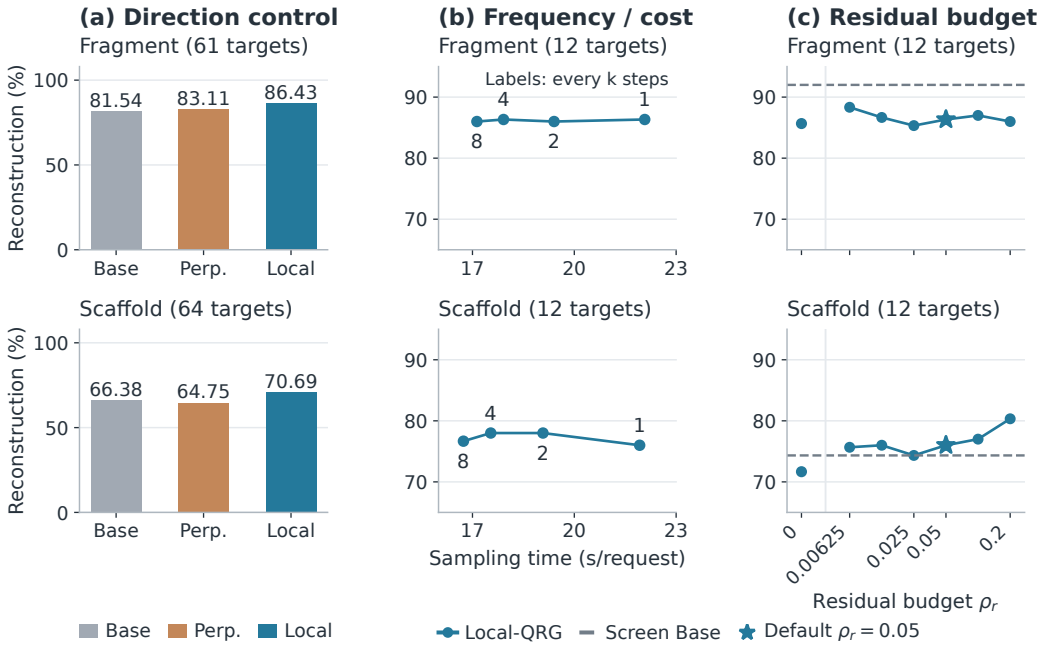

Figure 3: Direction and sensitivity analyses (seed 0). (a) Formal direction controls: 61/64 fragment/scaffold targets, 100 requests each. (b,c) Local frequency–cost and budget screens: 12 targets/task, 25 requests each. Positive budgets are log-spaced; zero is section-only. Bar axes start at zero; line panels show 65–95%. Full statistics, settings and magnitude-matching diagnostics: appendices D.4 and D.5.

5 Conclusion and limitations

QRG turns relational objectives into active, bounded corrections for frozen molecular diffusion by separating quotient-derived direction from sampler-relative scale. Its deployed variants require neither retraining nor rollout denoiser calls. Experiments demonstrate signal activation, improved fragment reconstruction and chemistry on TARGETDIFF, and transfer to DiffSBDD. Direction controls show that the correction direction matters, while frequency screens reveal a practical quality–cost trade-off.

Benefits depend on the task and configuration: higher fragment validity and QED come with reduced diversity, and side-chain and docking outcomes are mixed. Validation-based feature and budget selection, sparse feature construction, and broader task evaluation are priorities for extending the method.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using GenBench3D: Ligand conformation quality matters. arXiv preprint arXiv:2407.04424, 2024.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. arXiv preprint arXiv:2302.07121, 2023.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022.

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.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations*, 2024.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

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

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, 2020.

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.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020.

Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*, 2019.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023a.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In *International Conference on Machine Learning*, 2023b.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, 2022.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6:417–427, 2024. doi: 10.1038/s42256-024-00815-9.

Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995, 2020.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems*, 2022.

Hilbert J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, 2021.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. In *Advances in Neural Information Processing Systems*, 2023.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, 2020.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, 2018.

Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems - Series A*, 34(4):1533–1574, 2014.

Haitao Lin, Guojiang Zhao, Odin Zhang, Yufei Huang, Lirong Wu, Cheng Tan, Zicheng Liu, Zhifeng Gao, and Stan Z. Li. CBGBench: Fill in the blank of protein-molecule complex binding graph. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mOpNrrV2zH.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11451–11461, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, 2022.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecular structures conditional on a receptor binding site with deep generative models. *Chemical Science*, 2022.

RDKit contributors. RDKit: Open-source cheminformatics software, version 2025.09.6. https://www.rdkit.org/docs/index.html, 2026.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, 2021.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Rokkum Jamasb, Ilia Igashov, Weitao Du, Carla P. Gomes, Tom L. Blundell, Pietro Liò, Max Welling, Michael M. Bronstein,

and Bruno E. Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024. doi: 10.1038/s43588-024-00737-x.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d Infomax improves GNNs for molecular property prediction. In *International Conference on Machine Learning*, 2022.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. In *Advances in Neural Information Processing Systems*, 2018.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

Vinh Tong, Trung-Dung Hoang, Anji Liu, Guy Van den Broeck, and Mathias Niepert. Rao-blackwell gradient estimators for equivariant denoising diffusion. arXiv preprint arXiv:2502.09890, 2025.

Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

Renxiao Wang, Xiang Fang, Yipin Lu, Chen-Yang Yang, and Shaomeng Wang. The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119, 2005.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable CNNs: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems*, 2018.

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

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning*, 2023.

A Expanded related work

Equivariant architectures. Group-equivariant convolutional networks formalized convolutional symmetry beyond translations [Cohen and Welling, 2016], and compact-group equivariance results clarified when convolutional structure is forced by symmetry [Kondor and Trivedi, 2018]. Tensor Field Networks, 3D steerable CNNs, SE(3)-Transformers, and LieConv introduced rotation-aware or Lie-group-aware message passing for 3D geometric data [Thomas et al., 2018, Weiler et al., 2018, Fuchs et al., 2020, Finzi et al., 2020]; EGNNs gave a simpler coordinate-update architecture for Euclidean equivariance [Satorras et al., 2021]. Molecular and atomistic models such as Cormorant and NequIP show the value of covariant features in molecular property prediction and interatomic potentials [Anderson et al., 2019, Batzner et al., 2022]. These models are powerful but can be more complex or expensive than non-equivariant alternatives, motivating complementary approaches that preserve or exploit symmetry without changing the entire architecture.

Equivariant diffusion and flows. Equivariant normalizing flows and flow matching models learn densities or transports under known group actions [Köhler et al., 2020, Klein et al., 2023]. Symmetry-adapted 3D point-set generation and equivariant diffusion for 3D molecules showed how to generate atom types and coordinates with Euclidean symmetries [Gebauer et al., 2019, Hoogeboom et al., 2022]; GeoDiff and Torsional Diffusion focus on conformer generation [Xu et al., 2022, Jing et al., 2022]. In proteins, SE(3) diffusion and flow matching have been used for backbone generation [Yim et al., 2023, Bose et al., 2024]. Our method is not an architectural contribution in this line; it is an inference-time correction applied to a frozen model.

Rao-Blackwellization and orbit averaging. Orbit Diffusion interprets data augmentation as a Monte Carlo gradient estimator of a symmetrized loss and reduces variance by estimating an orbit-conditional expectation [Tong et al., 2025]. Its practical implementation requires only one neural forward/backward pass per training example, because the additional work occurs in the target construction rather than through repeated model evaluations. This observation strongly informs our runtime goal: symmetry-aware or quotient-aware improvements are most convincing when they do not hide large neural evaluation overhead.

KL-budgeted control. Path-integral and linearly solvable optimal-control formulations make relative-entropy or quadratic control effort central to the control objective [Kappen, 2005, Todorov, 2009]. The Schrödinger problem gives a broader path-space entropy-regularized transport viewpoint [Léonard, 2014], and diffusion Schrödinger bridges bring this path-space KL viewpoint into score-based generative modeling [De Bortoli et al., 2021]. QRG uses a local Gaussian reference model with a control fixed conditional on the current state: its mean shift has a closed-form conditional KL, relating the sampler-relative radius to a one-step entropy budget. Deterministic samplers use the corresponding quadratic kinetic cost.

SBDD and benchmark design. Structure-based generation has progressed from grid-based latent models [Ragoza et al., 2022] to pocket-conditioned graph samplers [Peng et al., 2022] and 3D diffusion models [Guan et al., 2023a, Schneuing et al., 2024, Guan et al., 2023b]. Task-specific models address docking, linker design, scaffold-oriented generation, and molecular property prediction with 3D pretraining [Corso et al., 2023, Imrie et al., 2020, Igashov et al., 2024, Stärk et al., 2022]. CBG-BENCH unifies these tasks through binding-graph completion and emphasizes interaction, chemical, geometric, and substructure metrics [Lin et al., 2025]. GENBENCH3D and CrossDocked-style datasets highlight the importance of conformation validity, dataset construction, and the risk of over-interpreting raw docking scores from invalid geometries [Francoeur et al., 2020, Baillif et al., 2024]. General molecular-generation benchmark suites provide complementary distributional metrics such as validity, novelty, and diversity [Brown et al., 2019, Polykovskiy et al., 2020]. These benchmarks shape our claim boundary.

B Proofs

B.1 Proof of proposition 1

Let $h = M^{-1}J^\top c$. For any feasible u in equation (4),

$$\langle c, Ju\rangle = \left\langle J^\top c, u\right\rangle = \langle h, u\rangle_M \geq -\|h\|_M \|u\|_M \geq -\|h\|_M . \tag{15}$$

The lower bound is attained by $u^\star = -h/\|h\|_M$, proving optimality. If $w \in \ker J$, then

$$\langle h, w\rangle_M = h^\top M w = c^\top J w = 0, \tag{16}$$

so h is M-orthogonal to the vertical subspace $\ker J$.

B.2 Proof of proposition 2

The objective in equation (6) can be written as $\langle h_t, u\rangle_{M_t}$, where $h_t = M_t^{-1}J_t^\top c_t$. For any feasible u,

$$\langle h_t, u\rangle_{M_t} \geq -\|h_t\|_{M_t} \|u\|_{M_t} \geq -B_t \|h_t\|_{M_t} . \tag{17}$$

For $B_t > 0$, equality requires u to oppose h_t and satisfy $\|u\|_{M_t} = B_t$, giving the unique optimizer $u_t^{\mathrm{quo}} = -B_t h_t/\|h_t\|_{M_t}$. For $B_t = 0$, the only feasible point is zero. Substituting $B_t = \rho_t \|v_t\|_{M_t}$ gives equation (7). The linearized quotient-energy change at $\tilde{x}_t$ is

$$\langle c_t, J_t u_t^{\mathrm{quo}}\rangle = \langle h_t, u_t^{\mathrm{quo}}\rangle_{M_t} = -\rho_t \|v_t\|_{M_t} \|h_t\|_{M_t} . \tag{18}$$

The ideal correction norm is B_t, independent of the nonzero gradient magnitude. The energy change still scales with $\|h_t\|_{M_t}$. This is a first-order statement at the evaluation state; finite corrections and the full sampler displacement also involve curvature and sampler motion.

B.3 Proof of proposition 3

For Gaussian kernels with the same covariance, the conditional KL divergence is the quadratic mean-shift cost

$$\mathrm{KL}\left(P_t^u(\cdot \mid x_t) \,\|\, P_t(\cdot \mid x_t)\right) = \frac{1}{2} u^\top (\tau_t M_t^{-1})^{-1} u = \frac{1}{2\tau_t} \|u\|_{M_t}^2 . \tag{19}$$

The constraint $\mathrm{KL}(P_t^u \| P_t) \leq \kappa_t$ is therefore equivalent to

$$\|u\|_{M_t} \leq \sqrt{2\tau_t\kappa_t}. \tag{20}$$

The linearized quotient loss satisfies

$$\langle c_t, J_t u\rangle = \left\langle J_t^\top c_t, u\right\rangle = \langle h_t, u\rangle_{M_t} . \tag{21}$$

By Cauchy–Schwarz, every feasible u obeys

$$\langle h_t, u\rangle_{M_t} \geq -\|h_t\|_{M_t} \|u\|_{M_t} \geq -\sqrt{2\tau_t\kappa_t}\, \|h_t\|_{M_t} . \tag{22}$$

The lower bound is attained uniquely, for $h_t \neq 0$, by

$$u_t^\star = -\sqrt{2\tau_t\kappa_t}\, \frac{h_t}{\|h_t\|_{M_t}}. \tag{23}$$

Choosing $\kappa_t = \rho_t^2 \|v_t\|_{M_t}^2 /(2\tau_t)$ gives $\|u_t^\star\|_{M_t} = \rho_t \|v_t\|_{M_t}$, the QRG trust radius. The calculation assumes that the covariance and mean shift are fixed conditional on x_t. A correction depending on the sampled proposal generally changes the kernel beyond a fixed Gaussian mean shift. Deterministic samplers instead use the quadratic kinetic cost $\frac{1}{2}\|u\|_{M_t}^2$ directly.

B.4 Proof of proposition 4

Let a, b be non-collinear and define $\gamma = B_a/\|a\| > 0$. Then

$$s = -\gamma(a + \lambda_* b), \qquad \lambda_* = \frac{B_b \|a\|}{B_a \|b\|}. \tag{24}$$

Since $\|s\| = \gamma \|a + \lambda_* b\|$, normalizing gives the representation in proposition 4. Linear independence makes λ_* unique. For fixed a, b, a fixed weight therefore reproduces only one ratio B_b/B_a; adaptive scalarization can reproduce the full split family. The construction specifies independent branch budgets, rather than an impossibility of scalarized representation. Branchwise optimality applies to the separate linear objectives on the product of their norm balls; it does not imply that their sum decreases each energy.

B.5 Proof of proposition 5

Because $q(R_g x) = q(x)$, differentiating gives

$$J_{R_g x} R_g = J_x, \qquad \text{so} \qquad J_{R_g x}^\top = R_g J_x^\top. \tag{25}$$

The quotient covector $c_x = \nabla_z \ell(q(x), y)$ is unchanged because $q(R_g x) = q(x)$. Using $M_{R_g x} = R_g M_x R_g^\top$,

$$h_{R_g x} = M_{R_g x}^{-1} J_{R_g x}^\top c_x = (R_g M_x R_g^\top)^{-1} R_g J_x^\top c_x = R_g M_x^{-1} J_x^\top c_x = R_g h_x. \tag{26}$$

The metric norm is preserved: $\|R_g w\|_{M_{R_g x}} = \|w\|_{M_x}$ for every w. Equivariance of the evaluation state allows the same identities to be applied at $\tilde{x}_t$. Sampler equivariance preserves the trust radius, yielding the correction identity in proposition 5, including the zero-lift case. Stochastic evaluations require the noise to transform jointly with the state. Euclidean branch clipping and scalar total caps also commute with orthogonal actions when their directions, masks and evaluation states transform consistently.

B.6 Why the trust-region update addresses dormancy

Dormancy is a mismatch between raw quotient-gradient magnitude and sampler-step scale. A raw update $-\eta h_t$ has norm $\eta \|h_t\|_{M_t}$, whereas the ideal normalized update has norm B_t for every nonzero lift. In the molecular controller, a clipped branch instead has norm $\min\{\|g\|_2, B_t\}$ away from the numerical floor, before final caps. It reaches the budget boundary only when its raw direction is large enough. The toy tasks test signal activation; the molecular rates and telemetry evaluate the clipped implementation. Neither rule by itself guarantees improved molecular outcomes.

C Experimental details

C.1 Toy quotient tasks

The Gaussian 2D task uses nuisance rotations and quotient MMD; orbit-point matching and the toy 3D molecule task use invariant template errors. Base, raw local guidance and budgeted residual guidance share the frozen denoising schedule. These controlled mechanism tests are separate from the molecular experiments.

C.2 TargetDiff protocol and recorded configurations

The three-seed editing study uses the frozen TARGETDIFF checkpoint within CBGBENCH (upstream revision `983fca2a066e1ba9c9f06b2a61ef207ff3c86264`, with the recorded working-tree patch). Seeds are 0/1/2. Fragment, linker and scaffold have 61/43/64 targets. Each task–method–seed target set is partitioned into five shards; each target receives 100 generations in total, with batch size 2. Base, section-only, PredNext and Local account for 180 completed runs and 2,016 target–seed–method records, or 201,600 requested generations within this four-method matrix. This is not an inventory total for all development, sidechain, transfer and control experiments.

Table 3: Recorded official controller settings. All guided rows operate on the predicted-next generated ligand coordinates, with trust ratio 1 and absolute correction cap 4. The original frozen controller clips gradients rather than always normalizing them to the budget boundary. Branch radii marked by † are traced through the matching frozen constructor/factory and recorded overrides; not all were separately serialized in historical run summaries.

Method	α	$\rho_s^\dagger$	$\rho_r^\dagger$	Every k steps	Residual
SECTION-ONLY	0	1	0	1	Disabled
PREDNEXT-QRG	0.05	1	0.05	2	Full pairwise
LOCAL-QRG	0.05	1	0.05	1	Local, $k_{\mathrm{nn}} = 4$
Sidechain rescue	0.01	1	0.01	4	Local, $k_{\mathrm{nn}} = 4$

The full follow-up sampler uses 1,000 reverse steps, giving 1,000 Local/section-only and 500 PredNext delivery opportunities per sample; the older 500-step slice in table 30 is separately scoped.

Official guided configurations do not enable explicit SE(3) residual projection, residual shuffling or section-residual orthogonalization. The ambient metric is Euclidean. The historical seed-0 invocations lack individual source hashes. The July source snapshot and configuration records identify the archived implementation; they do not independently identify the source executed in each May run.

One-step implementation. The runner first samples position and atom-type proposals, packs the generated ligand atoms, and forms the realized position displacement $v = x^{\mathrm{proposal}} - x$. It detaches the selected evaluation state before differentiating the batch-mean section and residual energies. Only the generated ligand coordinates receive the added control; pocket and preserved ligand-context coordinates are fixed. Let g_s, g_r denote the masked negative gradients and

$$C_B(g) = g \min\{1, B/\max(\|g\|_2, 10^{-8})\}.$$

With $b = \max\{\|v\|_2, 10^{-8}\}$, the official delivered update is

$$u = C_4(C_b[C_{\rho_s b}(g_s) + C_{\rho_r b}(g_r)]) \,. \tag{27}$$

The residual term is omitted for section-only. Thus branch budgets are upper bounds; small gradients need not saturate them. This differs from the ideal boundary-normalized local optimizer in equation (7). Batch-mean reduction affects unsaturated gradient magnitudes. Since control is computed after the random proposal, the same-covariance deterministic mean-shift KL identity is an interpretation under its stated assumptions, not an established exact divergence for this postprocessed stochastic kernel.

Concrete energies. For generated ligand atoms L, fixed pocket atoms P and preserved ligand context C, let $\langle\cdot\rangle$ denote a masked mean over the relevant valid atoms or pairs, and $(a)_+ = \max(a, 0)$. The frozen task implementation uses

$$E_s = \left\langle (1.6 - d_{LP})_+^2 \right\rangle + \tfrac{1}{2} \left\langle (1.2 - d_{LC})_+^2 \right\rangle + w_s \left\langle (\min_{c \in C} d_{ic} - 1.8)^2 \right\rangle_{i \in L} .$$

Absent context terms return zero. The full residual is $w_q \langle (1.2 - d_{ij})_+^2 \rangle_{i<j \in L}$. The Local residual instead uses $w_q \langle (d_{ij} - 1.2)^2 \rangle_{j \in \mathrm{kNN}(i)}$, with four neighbors in the default configuration. These are different local and full energies, not merely two algorithms for evaluating the same sum. The matching frozen task constructor sets $w_s = 0.5$ and $w_q = 0.25$; these implementation constants are source-derived rather than separately recorded per-run measurements. The full run/configuration index preserves explicit overrides and missing-field labels.

Differentiation and preconditioning. Reverse-mode differentiation computes coordinate covectors without constructing a dense Jacobian. Masked Euclidean norms use $\sqrt{\sum \mathrm{mask}\, x^2}$ clamped below at 10^{-8}. Neighbor selection is discrete and locally piecewise constant; gradients flow through selected distances. The sensitivity overlay's contact-density and motion-diagonal settings apply an inverse diagonal metric to directions, while delivery still uses Euclidean norm caps. Motion diagonals use per-atom squared proposal displacements normalized by their masked mean and clipped to $[0.25, 4]$; these are coordinate preconditioners, distinct from feature-loss weights.

C.3 Statistics and chemical metrics

For a method comparison, form paired rate differences on the seed-by-target matrix. Each of 100,000 bootstrap replicates samples seeds with replacement and targets with replacement, then averages the resulting crossed matrix. The 2.5/97.5 percentiles give pointwise 95% intervals. The fixed analysis stream is `numpy.random.default_rng(20260723)`; iteration order is preserved in the archived script. W/L/T uses target effects averaged over seeds, with tolerance 10^{-12}. Exact two-sided target-cluster sign-flip probabilities are computed by dynamic programming. The six Local/PredNext comparisons across three tasks form the headline Holm family. Section-only contrasts use a separate three-comparison family, not nine-comparison correction.

Reconstruction rate is the number of saved reconstructed SDFs divided by requested generations. RDKit validity additionally requires parsing/sanitization (`sanitize=True`, `removeHs=False`). Uniqueness is unique canonical isomeric SMILES divided by valid molecules; novelty is the fraction of unique molecules absent from the frozen training-reference SMILES cache. Diversity is mean

pairwise $1 - \mathrm{Tanimoto}$ over unique molecules using Morgan fingerprints of radius 3 and 2,048 bits. QED, synthetic-accessibility (SA), LogP and Lipinski descriptors use valid molecules before target/seed aggregation. Metric-specific missingness can change paired target counts. Quality intervals use their separately recorded analysis stream; even reconstruction intervals can differ slightly from the main table because the stream/iteration order differs.

C.4 DiffSBDD transfer and semantic checks

The frozen de novo DiffSBDD adapter uses seed 2022, batch size 2, 1,000 reverse steps, eight targets and 20 requested generations per target/method. Local uses $\alpha = 0.05$, predicted-next evaluation and guidance every step; section-only uses $\alpha = 0$. The zero-center-of-mass adapter and checkpoint identity are preserved with the run manifests. Acceptance requires successful coordinate-to-bond reconstruction and retention of a connected fragment containing at least 80% of generated atoms. Accepted counts are 62/63/70 for Base/section-only/Local, each out of 160. Per-target counts appear in table 10.

Controller-free execution delegates to the native sampler. Preserved single-step tests check position/type noise draw order and RNG-state restoration under the recorded CUDA tolerance; post-telemetry regression checks are also retained. An earlier independent full-trajectory equivalence attempt failed, so these checks are not evidence that separately executed stochastic trajectories are identical. Accepted files are consecutively renumbered after acceptance. Original requested IDs, rejected coordinates and mutually exclusive rejection categories were not serialized; they cannot be recovered from accepted filenames. Accepted-set sanitization and uniqueness do not use all requested generations as their denominator.

C.5 Direction controls, fixed radius and selection

The formal perpendicular comparison uses official seed-0 Local as reference on 61/64 fragment/scaffold targets, each with 100 generations. Its target-level bootstrap uses 100,000 replicates; the two-sided sign-flip test uses 100,000 Monte Carlo draws with an add-one correction and seed 20260723. These raw p values are not Holm-adjusted headline values. Smaller screens use 12 targets per task with 25 generations each, and preserve unfavorable and neutral outcomes.

Fixed-radius calibration uses four targets per task and 25 samples per target. The frozen configuration is evaluated on 12 disjoint targets; calibration/evaluation identity intersections are zero. Calibration iteratively matches mean delivered control norm to relative guidance within 5%, without imposing held-out stepwise equality. Frozen radii and results are given in appendix D.5. The four task-specific val16 target lists are disjoint from the official test target lists. These checks verify the named split memberships; a complete chronology of exploratory parameter selection is unavailable. The branch-allocation screen varies ρ_s, ρ_r; a ratio-only experiment at fixed total budget was not performed.

C.6 Docking and sidechain

The new audit uses seed 0, four tasks, eight targets per task, three methods and three samples per target-method: 288 requested calls. Scores, including zeros, are averaged per target-method before paired guided-minus-Base differences; negative values are favorable. The audit uses target bootstrap intervals (100,000 replicates, seed 20260723) and exact sign tests. The earlier table instead bootstraps 30/44/42 target–seed pairs (10/16/14 unique targets) with 10,000 replicates and `random.Random(20260430)`. Its target sets have zero intersection with the new eight-target sets. For the historical scaffold experiment, the request status of 27 nominal slots is unrecorded. Subsequent completeness and aggregation checks are reported separately from the original estimates.

Sidechain is ligand decomposition with fixed protein/context coordinates. The low-budget rescue is `sidechain_rescue_a001_k4`: $\alpha = 0.01$, every four steps, seeds 0/1/2 and 64 targets. It is evaluated separately from default-budget variants. Docking scores and molecular descriptors do not establish biological activity.

D Additional completed results

D.1 Complete three-seed reconstruction results

Table 4: Complete paired reconstruction results corresponding to table 2, including 95% confidence intervals from 100,000 crossed seed/target bootstrap replicates. Intervals are pointwise, not simultaneous. W/L/T counts seed-averaged target wins/losses/ties; exact target-cluster sign-flip tests receive Holm correction across the six comparisons.

Task	Method	Base	Guided	Δ	95% interval	W/L/T	Holm p
Fragment	LOCAL-QRG	0.81907	0.86798	+0.04891	[0.02279, 0.07601]	46/13/2	0.00147
Fragment	PREDNEXT-QRG	0.81907	0.83776	+0.01869	[0.00235, 0.03459]	37/21/3	0.07732
Linker	LOCAL-QRG	0.68938	0.70457	+0.01519	[-0.02767, 0.05752]	28/15/0	1.00000
Linker	PREDNEXT-QRG	0.68938	0.67915	-0.01023	[-0.04527, 0.02287]	26/17/0	1.00000
Scaffold	LOCAL-QRG	0.66484	0.70849	+0.04365	[-0.01677, 0.10036]	47/17/0	0.59886
Scaffold	PREDNEXT-QRG	0.66484	0.68557	+0.02073	[-0.02781, 0.06286]	47/17/0	1.00000

Table 5: Absolute reconstruction rates for all four methods on the same seeds 0/1/2 and 61/43/64 fragment/linker/scaffold targets. Seed SD is the sample standard deviation across the three seed rates, not a confidence interval. Control/base is the recorded mean across seeds.

Task	Method	Seed 0	Seed 1	Seed 2	Mean	Seed SD	Control/base
Fragment	BASE	0.81541	0.82262	0.81918	0.81907	0.00361	0.00000
Fragment	SECTION-ONLY	0.83508	0.84279	0.84049	0.83945	0.00396	0.25019
Fragment	PREDNEXT-QRG	0.82852	0.84410	0.84066	0.83776	0.00818	0.19210
Fragment	LOCAL-QRG	0.86426	0.86607	0.87361	0.86798	0.00496	0.26532
Linker	BASE	0.68140	0.69860	0.68814	0.68938	0.00867	0.00000
Linker	SECTION-ONLY	0.67558	0.66628	0.66744	0.66977	0.00507	0.23892
Linker	PREDNEXT-QRG	0.68023	0.68023	0.67698	0.67915	0.00188	0.18325
Linker	LOCAL-QRG	0.71163	0.70442	0.69767	0.70457	0.00698	0.25144
Scaffold	BASE	0.66375	0.66859	0.66219	0.66484	0.00334	0.00000
Scaffold	SECTION-ONLY	0.65625	0.66312	0.66406	0.66115	0.00427	0.25607
Scaffold	PREDNEXT-QRG	0.68266	0.68469	0.68937	0.68557	0.00345	0.19340
Scaffold	LOCAL-QRG	0.70688	0.70328	0.71531	0.70849	0.00618	0.26773

Table 6: Section-only minus Base. These three secondary comparisons have a separate three-test Holm family; their adjusted values are not part of the six-test Local/PredNext family in the main text. Intervals use the same crossed seed/target construction.

Task	Δ vs. Base	95% interval	W/L/T	Raw p	Secondary Holm p
Fragment	+0.02038	[0.00224, 0.03863]	41/19/1	0.02577	0.07732
Linker	-0.01961	[-0.05876, 0.01860]	25/17/1	0.30434	0.60869
Scaffold	-0.00370	[-0.05885, 0.04620]	39/25/0	0.89402	0.89402

The fragment section-only gain is $+0.02038$, compared with Local's $+0.04891$; the section-only comparison remains unresolved after its own family correction. Linker and scaffold section-only means are below Base. The tables report absolute rates for all variants alongside their corresponding paired statistical comparisons.

D.2 Complete molecular-quality follow-up

The following tables contain all 81 recorded task–method–metric contrasts: Local, PredNext and section-only across three tasks and nine metrics. Each is guided minus Base. W/L/T follows the numerical sign of the metric, not desirability; for example, a lower SA score is conventionally favorable. These metric-specific intervals are not multiplicity-adjusted significance claims. Reconstruction and RDKit validity use all requested generations; uniqueness, novelty, diversity and descriptors use the denominators defined in appendix C.

Table 7: Fragment completion: complete three-seed molecular-quality contrasts. Each row uses seeds 0/1/2. The n column is the number of paired targets retained for that metric.

Method	Metric	n	Δ vs. Base	95% interval	W/L/T
LOCAL-QRG	Reconstruction	61	+0.04891	[0.02279, 0.07601]	46/13/2
LOCAL-QRG	RDKit validity	61	+0.05000	[0.02393, 0.07678]	45/15/1
LOCAL-QRG	Uniqueness	61	-0.00608	[-0.01109, -0.00114]	17/33/11
LOCAL-QRG	Novelty	61	+0.00000	[0.00000, 0.00000]	0/0/61
LOCAL-QRG	Diversity	61	-0.01386	[-0.01755, -0.01037]	6/55/0
LOCAL-QRG	QED	61	+0.00991	[0.00219, 0.01799]	42/19/0
LOCAL-QRG	SA	61	-0.02224	[-0.02832, -0.01644]	10/51/0
LOCAL-QRG	LogP	61	-0.19450	[-0.25937, -0.12898]	12/49/0
LOCAL-QRG	Lipinski	61	-0.02140	[-0.05954, 0.01459]	30/31/0
PREDNEXT-QRG	Reconstruction	61	+0.01869	[0.00235, 0.03448]	37/21/3
PREDNEXT-QRG	RDKit validity	61	+0.01836	[0.00180, 0.03410]	37/21/3
PREDNEXT-QRG	Uniqueness	61	-0.00317	[-0.00726, 0.00039]	17/32/12
PREDNEXT-QRG	Novelty	61	+0.00000	[0.00000, 0.00000]	0/0/61
PREDNEXT-QRG	Diversity	61	-0.00203	[-0.00426, 0.00002]	23/38/0
PREDNEXT-QRG	QED	61	-0.00178	[-0.00872, 0.00516]	28/33/0
PREDNEXT-QRG	SA	61	-0.01608	[-0.02097, -0.01137]	10/51/0
PREDNEXT-QRG	LogP	61	-0.09321	[-0.14608, -0.03907]	19/42/0
PREDNEXT-QRG	Lipinski	61	-0.03500	[-0.06760, -0.00734]	25/36/0
SECTION-ONLY	Reconstruction	61	+0.02038	[0.00230, 0.03869]	41/19/1
SECTION-ONLY	RDKit validity	61	+0.02066	[0.00246, 0.03918]	41/17/3
SECTION-ONLY	Uniqueness	61	-0.00371	[-0.00766, 0.00015]	19/30/12
SECTION-ONLY	Novelty	61	+0.00000	[0.00000, 0.00000]	0/0/61
SECTION-ONLY	Diversity	61	-0.00314	[-0.00540, -0.00097]	20/41/0
SECTION-ONLY	QED	61	-0.00364	[-0.01154, 0.00449]	27/34/0
SECTION-ONLY	SA	61	-0.01984	[-0.02537, -0.01458]	9/52/0
SECTION-ONLY	LogP	61	-0.17648	[-0.23535, -0.11853]	11/50/0
SECTION-ONLY	Lipinski	61	-0.03623	[-0.07396, -0.00077]	30/31/0

Table 8: Linker design: complete three-seed molecular-quality contrasts, with the same metric definitions as table 7.

Method	Metric	n	Δ vs. Base	95% interval	W/L/T
LOCAL-QRG	Reconstruction	43	+0.01519	[-0.02752, 0.05775]	28/15/0
LOCAL-QRG	RDKit validity	43	+0.01566	[-0.02729, 0.05806]	28/15/0
LOCAL-QRG	Uniqueness	43	-0.00954	[-0.01613, -0.00370]	6/25/12
LOCAL-QRG	Novelty	43	+0.00001	[-0.00045, 0.00045]	1/1/41
LOCAL-QRG	Diversity	43	-0.01192	[-0.01897, -0.00448]	9/34/0
LOCAL-QRG	QED	43	+0.01510	[0.00183, 0.02681]	36/7/0
LOCAL-QRG	SA	43	-0.02348	[-0.03082, -0.01636]	8/35/0
LOCAL-QRG	LogP	43	-0.16943	[-0.24751, -0.08574]	8/35/0
LOCAL-QRG	Lipinski	43	+0.00925	[-0.05063, 0.06424]	29/14/0
PREDNEXT-QRG	Reconstruction	43	-0.01023	[-0.04543, 0.02302]	26/17/0
PREDNEXT-QRG	RDKit validity	43	-0.01023	[-0.04504, 0.02302]	26/16/1
PREDNEXT-QRG	Uniqueness	43	-0.00258	[-0.00832, 0.00237]	13/17/13
PREDNEXT-QRG	Novelty	43	+0.00019	[0.00000, 0.00064]	2/0/41
PREDNEXT-QRG	Diversity	43	+0.00618	[0.00096, 0.01229]	26/17/0
PREDNEXT-QRG	QED	43	+0.00413	[-0.00483, 0.01223]	28/15/0
PREDNEXT-QRG	SA	43	-0.02315	[-0.03140, -0.01654]	4/39/0
PREDNEXT-QRG	LogP	43	-0.07748	[-0.16938, 0.01049]	17/26/0
PREDNEXT-QRG	Lipinski	43	-0.01083	[-0.07796, 0.04379]	27/16/0
SECTION-ONLY	Reconstruction	43	-0.01961	[-0.05884, 0.01837]	25/17/1
SECTION-ONLY	RDKit validity	43	-0.01930	[-0.05845, 0.01891]	26/17/0
SECTION-ONLY	Uniqueness	43	-0.00265	[-0.00705, 0.00159]	10/19/14
SECTION-ONLY	Novelty	43	+0.00010	[-0.00036, 0.00060]	2/0/41
SECTION-ONLY	Diversity	41	+0.00388	[-0.00258, 0.01275]	16/25/0
SECTION-ONLY	QED	43	+0.00445	[-0.00590, 0.01441]	24/19/0
SECTION-ONLY	SA	43	-0.02388	[-0.03172, -0.01648]	6/37/0
SECTION-ONLY	LogP	43	-0.07054	[-0.17304, 0.03478]	16/27/0
SECTION-ONLY	Lipinski	43	-0.01615	[-0.09214, 0.04931]	23/20/0

Table 9: Scaffold hopping: complete three-seed molecular-quality contrasts. Changes in reconstruction, descriptors and diversity are distinct outcomes; no universal improvement is inferred from a positive rate difference.

Method	Metric	n	Δ vs. Base	95% interval	W/L/T
LOCAL-QRG	Reconstruction	64	+0.04365	[-0.01734, 0.10021]	47/17/0
LOCAL-QRG	RDKit validity	64	+0.04297	[-0.01776, 0.10016]	47/17/0
LOCAL-QRG	Uniqueness	64	-0.00013	[-0.00599, 0.00517]	14/12/38
LOCAL-QRG	Novelty	64	+0.00011	[0.00000, 0.00048]	2/0/62
LOCAL-QRG	Diversity	63	-0.00097	[-0.00451, 0.00212]	34/29/0
LOCAL-QRG	QED	64	+0.01430	[-0.00020, 0.02856]	41/23/0
LOCAL-QRG	SA	64	-0.02397	[-0.03064, -0.01758]	7/57/0
LOCAL-QRG	LogP	64	-0.23591	[-0.39514, -0.06665]	20/44/0
LOCAL-QRG	Lipinski	64	-0.02219	[-0.08888, 0.03956]	35/29/0
PREDNEXT-QRG	Reconstruction	64	+0.02073	[-0.02760, 0.06286]	47/17/0
PREDNEXT-QRG	RDKit validity	64	+0.02068	[-0.02740, 0.06281]	47/16/1
PREDNEXT-QRG	Uniqueness	64	+0.00353	[-0.00033, 0.00844]	17/10/37
PREDNEXT-QRG	Novelty	64	+0.00011	[0.00000, 0.00048]	2/0/62
PREDNEXT-QRG	Diversity	63	+0.00206	[-0.00052, 0.00444]	36/27/0
PREDNEXT-QRG	QED	64	+0.00465	[-0.00875, 0.01808]	33/31/0
PREDNEXT-QRG	SA	64	-0.02059	[-0.02706, -0.01425]	10/54/0
PREDNEXT-QRG	LogP	64	-0.18887	[-0.33006, -0.03949]	20/44/0
PREDNEXT-QRG	Lipinski	64	-0.06092	[-0.12373, -0.00419]	30/34/0
SECTION-ONLY	Reconstruction	64	-0.00370	[-0.05823, 0.04646]	39/25/0
SECTION-ONLY	RDKit validity	64	-0.00422	[-0.05911, 0.04578]	39/25/0
SECTION-ONLY	Uniqueness	63	+0.00454	[-0.00092, 0.01049]	18/9/36
SECTION-ONLY	Novelty	63	+0.00012	[0.00000, 0.00048]	2/0/61
SECTION-ONLY	Diversity	63	+0.00260	[-0.00005, 0.00526]	43/20/0
SECTION-ONLY	QED	63	-0.00272	[-0.01624, 0.01053]	31/32/0
SECTION-ONLY	SA	63	-0.02012	[-0.02587, -0.01429]	10/53/0
SECTION-ONLY	LogP	63	-0.22721	[-0.37945, -0.06032]	19/44/0
SECTION-ONLY	Lipinski	63	-0.03163	[-0.08939, 0.02641]	30/33/0

D.3 DiffSBDD target-level acceptance

Table 10: Accepted generations out of 20 per DiffSBDD target and method, seed 2022. Pocket identifiers are shown; full target identities are preserved in the manifest. Column totals are 62/63/70 out of 160. Files are indexed after acceptance, not by original request ID.

Target (PDB pocket identifier)	Base /20	Section /20	Local /20
2z3h_A_rec_1wn6_bst_lig_tt_docked_3_pocket10	6	6	5
4aaw_A_rec_4ac3_r83_lig_tt_min_0_pocket10	11	11	12
4yhj_A_rec_4yhj_an2_lig_tt_min_0_pocket10	6	6	4
14gs_A_rec_20gs_cbd_lig_tt_min_0_pocket10	13	13	11
2v3r_A_rec_1dy4_snp_lig_tt_docked_1_pocket10	7	7	12
4rn0_B_rec_4rn1_l8g_lig_tt_min_0_pocket10	13	13	15
1fmc_B_rec_1fmc_cho_lig_tt_docked_1_pocket10	4	5	8
3daf_A_rec_3daf_feg_lig_tt_docked_0_pocket10	2	2	3

Table 11: DiffSBDD yield and mean final control/base ratio for the same eight-target experiment as figure 2b. Yield is the fraction of 160 requests passing the reconstruction-and-connectivity test, not unconditional RDKit validity.

Method	Reconstruction-and-connectivity yield	Final control/base
Unguided	0.3875	0
SECTION-ONLY	0.3938	0.0000242
LOCAL-QRG	0.4375	0.0351

Mechanism telemetry gives a section-only control/base ratio of 2.42×10^{-5}. Local's pre-budget residual is 10.32% of the base step; 68.30% of that direction is delivered; final control/base is 3.51%, with nonzero control on all 80,000 reverse steps. These are distinct telemetry summaries, not factors to multiply to reconstruct the final ratio.

The Local accepted count improves over Base on five of eight targets. The retained-fragment acceptance test, accepted-set chemical checks and frozen-sampler semantic checks answer different questions and are not combined into an unconditional validity guarantee.

D.4 Complete neighborhood, metric, budget and frequency screens

Each setting below uses seed 0, 12 targets per task and 25 generations per target (300 per task-setting). C/B is the recorded mean correction/base ratio. Sampling seconds/sample divides recorded sampling time by requested generations; it excludes separately recorded postprocessing/writing time and is not controller-only timing. Conditions are preserved under their recorded names; decimal suffixes such as `0p05` denote 0.05. The exploratory screens do not establish universal optimal settings.

Table 12: Complete neighborhood and diagonal-metric screens. Contact-density and motion-diagonal settings precondition the coordinate direction; official main runs use the identity metric.

Screen	Condition	Frag. rate	C/B	s/sample	Scaf. rate	C/B	s/sample
Neighbors	`local_k_2`	0.890	0.266	21.04	0.783	0.266	21.64
Neighbors	`local_k_4`	0.863	0.273	22.25	0.760	0.275	21.69
Neighbors	`local_k_8`	0.870	0.277	21.79	0.803	0.279	22.33
Metric	`contact_density`	0.877	0.274	21.93	0.770	0.278	22.11
Metric	`identity`	0.863	0.273	22.24	0.760	0.275	22.05
Metric	`motion_diagonal`	0.887	0.329	21.08	0.773	0.329	22.57

Table 13: Complete Local and PredNext residual-budget screens. `base` is unguided; `rho_r_0` has no residual. The six nonzero Local radii are 0.00625, 0.0125, 0.025, 0.05, 0.1 and 0.2; the PredNext screen stops at 0.1. Both favorable and unfavorable settings are included.

Screen	Condition	Frag. rate	C/B	s/sample	Scaf. rate	C/B	s/sample
Local budget	`base`	0.920	0.000	16.45	0.743	0.000	16.09
Local budget	`rho_r_0`	0.857	0.258	17.62	0.717	0.261	17.18
Local budget	`rho_r_0p00625`	0.883	0.259	22.23	0.757	0.261	22.00
Local budget	`rho_r_0p0125`	0.867	0.260	22.33	0.760	0.263	21.70
Local budget	`rho_r_0p025`	0.853	0.262	22.26	0.743	0.266	21.69
Local budget	`rho_r_0p05`	0.863	0.273	21.93	0.760	0.275	21.51
Local budget	`rho_r_0p1`	0.870	0.294	21.63	0.770	0.290	21.89
Local budget	`rho_r_0p2`	0.860	0.319	20.54	0.803	0.306	21.30
PredNext budget	`rho_r_0p00625`	0.867	0.198	19.61	0.750	0.197	18.86
PredNext budget	`rho_r_0p0125`	0.857	0.198	18.40	0.753	0.196	18.89
PredNext budget	`rho_r_0p025`	0.837	0.198	19.12	0.740	0.195	18.83
PredNext budget	`rho_r_0p05`	0.837	0.198	19.60	0.750	0.195	18.78
PredNext budget	`rho_r_0p1`	0.833	0.198	19.54	0.747	0.195	46.12

Table 14: Branch-allocation screen. A condition `s_a_r_b` specifies section/residual radius ratios a, b. Since total radius varies, this is not a ratio-only experiment at fixed total budget.

Screen	Condition	Frag. rate	C/B	s/sample	Scaf. rate	C/B	s/sample
Branch allocation	`s_0_r_0p05`	0.903	0.048	22.00	0.820	0.049	21.09
Branch allocation	`s_0p1_r_0p05`	0.900	0.108	22.23	0.823	0.107	20.85
Branch allocation	`s_0p25_r_0p0125`	0.867	0.177	21.06	0.770	0.171	21.76
Branch allocation	`s_0p25_r_0p05`	0.887	0.186	21.09	0.790	0.180	20.71
Branch allocation	`s_0p25_r_0p1`	0.897	0.207	22.21	0.807	0.196	21.72
Branch allocation	`s_0p5_r_0p025`	0.863	0.234	22.15	0.793	0.235	21.88
Branch allocation	`s_0p5_r_0p05`	0.880	0.244	21.58	0.797	0.243	20.58
Branch allocation	`s_0p5_r_0p1`	0.890	0.265	21.68	0.797	0.259	21.87

Table 15: Complete frequency screens. `every_k` delivers guidance every k reverse steps. The official PredNext every-two configuration is reported in the three-seed study; the additional frequency screen contains every 1/4/8.

Screen	Condition	Frag. rate	C/B	s/sample	Scaf. rate	C/B	s/sample
Local frequency	`local_every_1`	0.863	0.273	22.07	0.760	0.275	21.93
Local frequency	`local_every_2`	0.860	0.202	19.40	0.780	0.201	19.08
Local frequency	`local_every_4`	0.863	0.132	17.91	0.780	0.126	17.54
Local frequency	`local_every_8`	0.860	0.080	17.13	0.767	0.073	16.73
PredNext frequency	`every_1`	0.840	0.256	21.32	0.713	0.258	21.82
PredNext frequency	`every_4`	0.863	0.131	18.06	0.743	0.123	17.37
PredNext frequency	`every_8`	0.873	0.080	17.42	0.740	0.072	16.74

D.5 Direction and absolute-scale controls

Table 16: Small direction-control screens and held-out fixed-radius evaluation. `anti` reverses the residual, `covector_permuted` uses a mismatched signal, and `random_horizontal` uses a random projected direction. Each task-setting has 12 targets and 300 requests.

Screen	Condition	Frag. rate	C/B	s/sample	Scaf. rate	C/B	s/sample
Direction	`anti`	0.857	0.255	22.19	0.667	0.262	22.62
Direction	`covector_permuted`	0.870	0.271	21.24	0.780	0.276	22.03
Direction	`random_horizontal`	0.853	0.265	22.52	0.720	0.268	22.55
Fixed radius	`evaluation`	0.853	0.278	22.30	0.763	0.275	21.23

Table 17: Small-screen paired controls minus Local on the same 12 targets, including neutral and adverse results. Intervals are target bootstrap intervals; raw sign-flip p values are exploratory and are not the main-table Holm values. Positive rate differences favor the listed control.

Task	Control	Δ vs. Local	95% interval	W/L/T	n	Raw p
Fragment	evaluation	-0.01000	[-0.04333, 0.02667]	3/4/5	12	0.71875
Scaffold	evaluation	+0.00333	[-0.06333, 0.06333]	6/4/2	12	1.00000
Fragment	anti	-0.00667	[-0.07667, 0.06333]	4/4/4	12	0.92188
Fragment	covector_permuted	+0.00667	[-0.02333, 0.04000]	3/3/6	12	0.87500
Fragment	random_horizontal	-0.01000	[-0.04667, 0.02667]	2/3/7	12	0.81250
Scaffold	anti	-0.09333	[-0.15667, -0.02333]	1/10/1	12	0.03320
Scaffold	covector_permuted	+0.02000	[-0.02667, 0.07333]	5/4/3	12	0.55078
Scaffold	random_horizontal	-0.04000	[-0.10000, 0.02000]	5/6/1	12	0.28516

Fixed-radius calibration uses four targets and freezes separate section/residual/total caps. For fragment these are 0.03380869/0.00929346/0.03459311; for scaffold, 0.04219308/0.01310632/0.04393459. The final calibration mean-delivery errors are 3.779% and 3.555%, respectively. The 12 held-out target identities do not overlap the four calibration identities. Held-out fixed-minus-relative differences are -0.01000 ($[-0.04333, 0.02667]$, $p = 0.71875$) and $+0.00333$ ($[-0.06333, 0.06333]$, $p = 1$). Neither proves equivalence or superiority. The

Table 18: Formal seed-0 perpendicular control versus official Local. W/L/T and differences are control minus Local; negative differences favor Local. Tests are separate unadjusted Monte Carlo sign-flip tests, with 100,000 draws and add-one correction.

Task	Local rate	Control rate	Δ control–Local	95% interval	W/L/T	Raw p
Fragment	0.86426	0.83115	-0.03311	[-0.04885, -0.01754]	13/40/8	0.0001
Scaffold	0.70688	0.64750	-0.05937	[-0.07687, -0.04344]	9/51/4	1e-05

reported 6.93×10^{-9} absolute-cosine field and 3.07×10^{-8} matching-error field originate from a one-target, one-sample, 1,000-step debug run. They are run-level averages of telemetry fields, not global maxima across the formal campaign. Table 19 gives all ten formal shard records. Matching precision in the formal experiment is characterized by these shard records, separately from the debug check. Paired reconstruction effects and matching-error diagnostics measure distinct outcomes.

Table 19: Formal-run telemetry, seed 0. All displayed fields are the stored `telemetry_means` (averages of per-step controller statistics). The abs-cos and match-error fields have a `max` suffix in code, which does not make their run-level averages global maxima. Cap columns are recorded activation fractions; shard target counts vary.

Task	Shard	Targets	Abs-cos field	Match-error field	Residual cap	Total cap
Fragment	shard_00	14	1.58e-05	0.1562	0.8421	0.0041
Fragment	shard_01	13	7.73e-06	0.1153	0.8692	0.0061
Fragment	shard_02	13	2.47e-05	0.0999	0.8977	0.0015
Fragment	shard_03	10	3.01e-05	0.1483	0.8450	0.0035
Fragment	shard_04	11	2.92e-05	0.1875	0.8055	0.0040
Scaffold	shard_00	15	2.69e-06	0.0314	0.9640	0.0014
Scaffold	shard_01	12	2.34e-05	0.0706	0.9225	0.0023
Scaffold	shard_02	15	1.34e-05	0.0418	0.9667	0.0008
Scaffold	shard_03	12	6.69e-06	0.0457	0.9517	0.0018
Scaffold	shard_04	10	1.2e-05	0.0603	0.9270	0.0014

D.6 Sidechain low-budget rescue

The default-budget sidechain results below remain a negative transfer test. The separate three-seed rescue uses $\rho_r = 0.01$ every four steps, with $\rho_s = 1$, on 64 targets. Base and rescue reconstruction means are 0.87286458 and 0.86864583; the paired difference is -0.00421875, crossed 95% interval $[-0.01515625, 0.00619792]$, W/L/T 32/27/5, and exact target-cluster sign-flip $p = 0.40125$. This is one secondary comparison. Its small mean deficit is not evidence that default-budget transfer succeeds.

Table 20: Complete three-seed quality follow-up for the sidechain low-budget rescue, guided minus Base. W/L/T follows numerical sign, not chemical desirability. Metric definitions match the three editing-task tables.

Metric	n	Δ vs. Base	95% interval	W/L/T
Reconstruction	64	-0.00422	[-0.01516, 0.00620]	32/27/5
RDKit validity	64	-0.00391	[-0.01500, 0.00661]	33/26/5
Uniqueness	64	-0.00240	[-0.00552, 0.00053]	16/19/29
Novelty	64	+0.00000	[0.00000, 0.00000]	0/0/64
Diversity	64	-0.00218	[-0.00415, -0.00056]	24/40/0
QED	64	+0.00085	[-0.00552, 0.00725]	30/34/0
SA	64	-0.01565	[-0.01996, -0.01133]	9/55/0
LogP	64	-0.04789	[-0.10304, 0.00670]	29/35/0
Lipinski	64	+0.00107	[-0.02393, 0.02590]	28/36/0

D.7 Complete follow-up docking audit

The linker PredNext mean is unfavorable and imprecise, while scaffold PredNext is favorable on this small cohort. These eight-target cohorts are disjoint from the unique targets in the historical table. The sidechain docking labels refer to the audit configuration; the three-seed lower-budget configuration is a separate experiment.

Table 21: All eight guided-minus-Base docking comparisons from the 32-target, 288-call audit, seed 0. Each task has eight targets; each method has three scores per target. Negative ΔVina is favorable. Sign-test p values are unadjusted; intervals bootstrap targets. No rows are selected by favorable direction.

Task	Method	Δ Vina	95% interval	Wins/8	Sign p
Fragment	PREDNEXT-QRG	-0.132	[-0.527, 0.300]	5/8	0.7266
Fragment	LOCAL-QRG	-0.227	[-0.652, 0.229]	5/8	0.7266
Linker	PREDNEXT-QRG	+2.786	[-0.074, 8.132]	2/8	0.2891
Linker	LOCAL-QRG	+0.736	[-0.464, 2.800]	4/8	1.0000
Scaffold	PREDNEXT-QRG	-0.328	[-0.667, -0.023]	6/8	0.2891
Scaffold	LOCAL-QRG	-0.190	[-0.477, 0.136]	6/8	0.2891
Sidechain	PREDNEXT-QRG	-0.272	[-0.670, 0.169]	6/8	0.2891
Sidechain	LOCAL-QRG	+0.379	[-0.076, 0.859]	2/8	0.2891

D.8 Generation metrics

The earlier summaries below remain separate from the complete follow-up tables above. They record development and historical seed-0/matched-slice cohorts rather than additional samples to pool into the three-seed headline estimate.

D.9 Development ablation for the cheap surrogate

Table 22: Development-slice fragment seed-0 ablation for cheap quotient surrogates. The local-radius surrogate is not introduced as an arbitrary extra branch: on this slice it outperforms sampled-pair residual surrogates at comparable runtime while preserving diversity.

Cheap surrogate	Validity ↑	Unique	Diversity	sec/sample
Sampled pairs, 16	0.888	854	0.896	17.84
Sampled pairs, 128	0.886	856	0.897	18.15
Local radius	**0.902**	**869**	0.896	16.88

The local-radius variant was selected as the practical cheap surrogate before the official full sweeps. This ablation is intentionally scoped as development evidence rather than an additional official benchmark claim.

D.10 Paired target-level validity

Table 23: Paired official seed-0 target-level validity deltas. Each row compares a guided method to the unguided sampler on the same targets; bootstrap intervals resample targets. These statistics support the aggregate table while showing target heterogeneity.

Task	Method	n	Mean Δ	Median Δ	95% bootstrap CI	Wins/losses/ties
Fragment	PREDNEXT-QRG	61	+0.013	+0.010	$[-0.006, 0.032]$	32/27/2
Fragment	LOCAL-QRG	61	**+0.049**	+0.030	$[0.019, 0.078]$	42/14/5
Scaffold	PREDNEXT-QRG	64	+0.019	+0.040	$[-0.029, 0.061]$	44/17/3
Scaffold	LOCAL-QRG	64	**+0.043**	+0.060	$[-0.016, 0.099]$	43/19/2
Linker	PREDNEXT-QRG	43	-0.001	+0.020	$[-0.037, 0.033]$	22/19/2
Linker	LOCAL-QRG	**43**	**+0.030**	+0.040	$[-0.012, 0.073]$	25/14/4

This older seed-0 paired analysis is separate from the three-seed headline table. LOCAL-QRG has the clearest fragment signal, with mean target-level validity delta $+0.049$ and bootstrap interval $[0.019, 0.078]$. Scaffold has positive mean deltas for both deployed variants and positive win/loss counts, but wider target-level intervals. Linker is near-neutral for PREDNEXT-QRG and positive in aggregate for LOCAL-QRG, with a mean paired delta of $+0.030$ and a bootstrap interval crossing zero.

D.11 Target-level heterogeneity

Table 24: Distribution of per-target official seed-0 validity deltas versus the unguided sampler. Each target contributes one target-level validity difference for the same task and method. The quantiles show that aggregate gains coexist with heterogeneous pockets, including a small number of large negative and large positive target-level shifts.

Task	Method	n	Mean Δ	Median	P10	P25	P75	P90	Min/Max
Fragment	PREDNEXT-QRG	61	+0.013	+0.010	-0.060	-0.020	+0.050	+0.100	-0.160/+0.210
Fragment	LOCAL-QRG	61	+0.049	+0.030	-0.060	+0.000	+0.100	+0.190	-0.280/+0.400
Scaffold	PREDNEXT-QRG	64	+0.019	+0.040	-0.097	-0.010	+0.103	+0.167	-0.860/+0.330
Scaffold	LOCAL-QRG	64	+0.043	+0.060	-0.246	-0.023	+0.190	+0.271	-0.860/+0.520
Linker	PREDNEXT-QRG	43	-0.001	+0.020	-0.158	-0.035	+0.060	+0.130	-0.340/+0.270
Linker	LOCAL-QRG	43	+0.030	+0.040	-0.148	-0.050	+0.130	+0.156	-0.260/+0.440

Table 24 summarizes all 504 task–method–target rows using target-level delta distributions. The quantiles and win/loss counts describe the consistency and spread of effects across targets. The pattern is consistent with the paired table: LOCAL-QRG has the strongest fragment distribution, scaffold has positive central tendency with wider tails, and linker remains task-dependent.

D.12 Summary across evaluations

Across the completed evaluations, Local fragment validity improves across three seeds after six-comparison correction; the other five headline tests remain unresolved. The independent eight-target DiffSBDD experiment demonstrates transfer to a second frozen sampler. Perpendicular controls support direction dependence, but do not rule out all scalarized guidance; fixed-radius and relative delivery are not statistically distinguished. Fragment chemistry gains include a diversity decrease. Earlier seed-0 and matched-slice results describe their own cohorts, not new three-seed absolute means. Side-chain failures and mixed docking bound task transfer. Local aggregation retains dense construction and does not consistently reduce controller time.

D.13 Anti-reparameterization ablation

Table 25: Fragment ablation frontier. ROLLOUT-QRG TEACHER is useful as a diagnostic upper-cost point but is not the deployed method. The comparison separates the core trust-region delivery rule from rollout-heavy or unstructured guidance controls.

Fragment variant	Seeds	Validity ↑	Runtime	Control/base
BASE	3	0.842	1.00×	0.000
SECTION-ONLY	1	0.896	1.07×	–
SHAM-QRG	3	0.888	1.22×	0.255
PREDNEXT-QRG	3	0.896	1.10×	0.193
LOCAL-QRG	3	**0.905**	1.22×	0.271
ROLLOUT-QRG TEACHER	3	0.898	2.15×	0.253

Table 26: Earlier raw-pocket docking diagnostic: 30/44/42 fragment/scaffold/linker target–seed pairs, corresponding to 10/16/14 unique targets, separate from the later 32-target, 288-call audit. Intervals bootstrap target–seed pairs. Negative guided-minus-Base ΔVina is favorable. The historical ledger is not a complete requested-call audit (appendix C).

Task / method	Target–seed pairs	Mean ΔVina ↓	Better pairs
Fragment / PREDNEXT-QRG	30	$0.061\ [-0.049, 0.174]$	14/30
Fragment / LOCAL-QRG	30	$-0.101\ [-0.189, -0.023]$	17/30
Scaffold / PREDNEXT-QRG	44	$-0.132\ [-0.234, -0.035]$	28/44
Scaffold / LOCAL-QRG	44	$\mathbf{-0.285}\ [-0.397, -0.178]$	35/44
Linker / PREDNEXT-QRG	42	$-0.001\ [-0.087, 0.081]$	20/42
Linker / LOCAL-QRG	42	$0.038\ [-0.055, 0.130]$	17/42

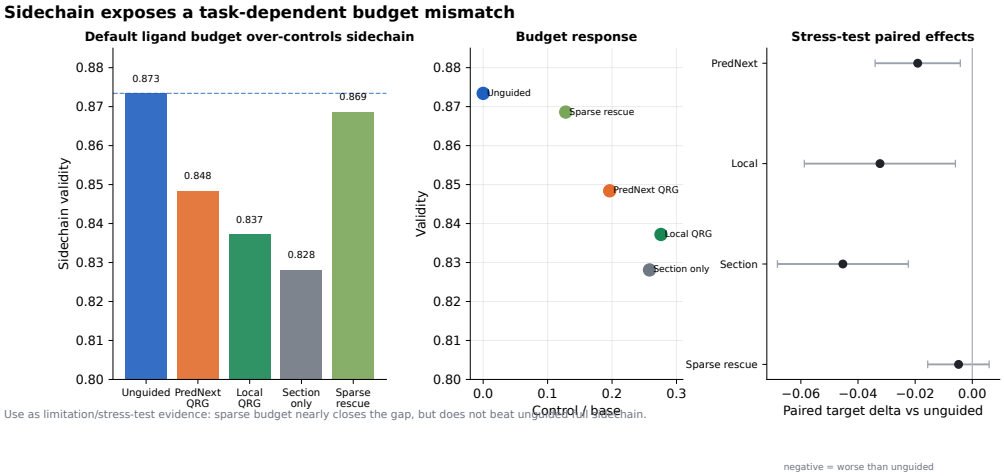

Figure 4: Side-chain decoration exposes a task-dependent budget mismatch. Default ligand-generation budgets reduce side-chain validity relative to the unguided sampler. A sparse low-budget rescue nearly closes the gap, but does not beat unguided in the full sweep.

D.14 Earlier seed-0 and matched-slice results

The following table and figures retain the earlier cohorts for comparison and runtime context. Their captions identify the scope so they are not mistaken for the added three-seed or DiffSBDD analyses.

Table 27: Earlier study: validity, time, and control/base are official seed-0 aggregates; novelty/diversity come from a separate matched multi-seed slice. This mixed-cohort table is retained for historical runtime context, not the three-seed headline estimate. The small slice's near-constant diversity does not negate the decrease in the later fragment chemistry analysis.

Task	Method	Official parts	Validity ↑	Δ val.	sec/sample	Control/base	Novelty ↑	Diversity ↑
Fragment	BASE	5/5	0.815	–	16.68	0.000	1.000	0.898
	PREDNEXT-QRG	5/5	0.829	+0.014	18.28	0.192	1.000	0.899
	LOCAL-QRG	5/5	**0.864**	**+0.049**	19.91	0.265	1.000	0.896
Scaffold	BASE	5/5	0.664	–	17.99	0.000	1.000	0.913
	PREDNEXT-QRG	5/5	0.683	+0.019	20.36	0.194	1.000	0.911
	LOCAL-QRG	5/5	**0.707**	**+0.043**	20.18	0.268	1.000	0.912
Linker	BASE	5/5	0.681	–	18.82	0.000	1.000	0.888
	PREDNEXT-QRG	5/5	0.680	-0.001	19.72	0.183	1.000	0.892
	LOCAL-QRG	5/5	**0.712**	**+0.031**	21.38	0.251	1.000	0.891

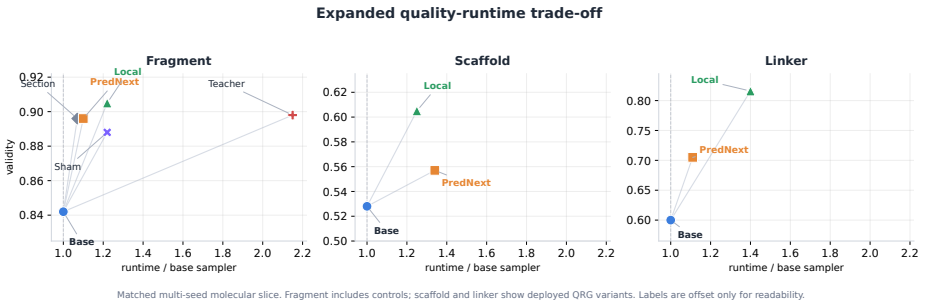

Figure 5: Earlier matched-slice frontier. It illustrates that slice's runtime and molecular metrics; it is not the completed three-seed headline analysis or evidence of general diversity preservation.

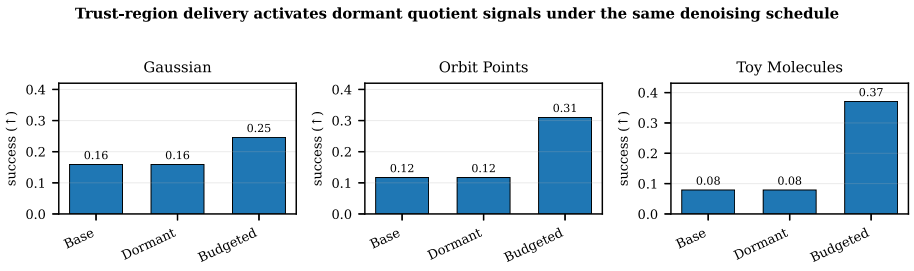

Figure 6: Controlled-task activation under the same denoising schedules. These mechanism tests are separate from molecular benchmarks.

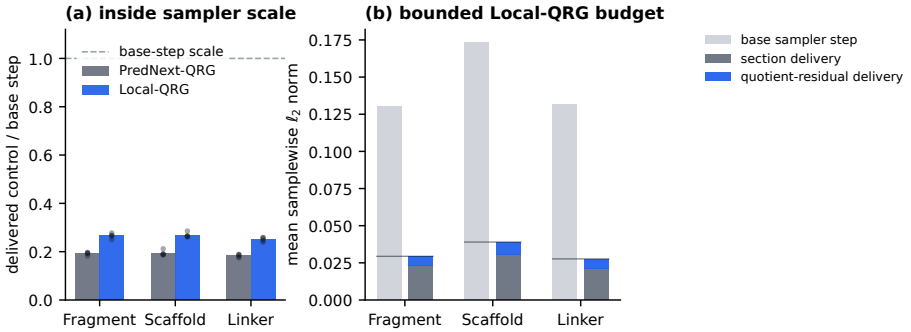

Figure 7: Earlier official seed-0 norm accounting. Delivered controls are compared to sampler-step norms; this is a kinetic diagnostic, not an empirical whole-path KL measurement or a norm summary for the new DiffSBDD experiment.

E Broader impact and compute resources

E.1 Broader impact

QRG is a sampling-time method for scientific molecular generation. Its intended use is methodological: evaluating how quotient-aware guidance can improve frozen generative samplers under explicit runtime budgets. The generated molecules in this paper are computational candidates produced inside benchmark pockets; they are not validated binders, drugs, or therapeutic leads. We therefore separate generation validity from docking diagnostics and avoid claims of biological activity. This boundary is important because over-interpreting raw docking scores or visually plausible structures could mislead downstream drug-design decisions.

The positive impact of this work is improved evaluation discipline for inference-time molecular guidance: the method reports wall-clock cost, delivered control budget, validity, novelty, diversity, and task-dependent failures rather than only best-case samples. Potential negative impact comes from the general dual-use nature of molecular generation. The work does not release a new biological assay, trained wet-lab model, or actionable synthesis pipeline; nonetheless, any use of generated molecules should require independent chemistry review, safety screening, and experimental validation.

E.2 Compute resources

The original editing experiments use a frozen TARGETDIFF backbone inside the CBGBENCH evaluation code; the independent DiffSBDD test is separate. Official seed-0 sweeps are split into 20-target shards, with 100 generated samples per target, batch size 2, and seed 0. Fragment, scaffold, linker, and side-chain official tasks use the same sharding contract. Runtime is reported as wall-clock seconds per generated sample from each run's `run_summary.json`; aggregate rows weight by the number of generated samples. Contaminated local co-runs are excluded from fair runtime comparisons when clean single-run measurements are available.

The official ligand-generation/editing shards run on cluster GPU workers. Completed official shard durations for linker are typically about 3.6–5.2 hours per shard, with occasional slower shards around 7 hours. Fragment and scaffold official rows report aggregate seconds/sample in table 27; these values include denoising, guidance, postprocessing, and file writing as captured by the benchmark runner. The multi-seed molecular slice and sham/teacher controls required additional runs, with rollout-teacher variants intentionally treated as diagnostic high-cost points rather than deployed methods.

Table 28: Earlier evaluation inventory, excluding the added full three-seed analysis, DiffSBDD test, and follow-up screens. Its 92,800 samples are not a total for the revised study. The table records the original assets. It is an accounting table, not an additional performance ranking.

Evaluation asset	Completed rows	Generated samples	Role in the paper
Official ligand-generation/editing sweeps	9	50,400	Main validity, runtime, and control-budget evidence for fragment growing, scaffold hopping, and linker design in table 27.
Official section-only controls	3	16,800	Residual-off controls used to separate section delivery from quotient-residual guidance.
Official side-chain sweep	4	25,600	Out-of-family budget-transfer stress test; not counted as a ligand-generation win claim.
Matched multi-seed molecular slice	40	–	Novelty/diversity preservation, development ablations, rollout-teacher diagnostics, and sham-control comparisons.
Target-balanced docking diagnostic	6	–	Supporting raw-pocket AutoDock Vina diagnostic with paired target comparisons; not a binding-affinity claim.
Official full aggregate total	16	92,800	Completed official seed-0 aggregate rows across all four molecular tasks and four reported method families.

Table 29: Compute accounting for the earlier official seed-0 ligand-generation/editing rows. Approximate GPU hours are computed from generated samples times aggregate wall-clock seconds/sample divided by 3600, and therefore include denoising, guidance, postprocessing, and file writing as recorded by the runner.

Task	Method	Generated	sec/sample	Approx. GPU h	Control/base
Fragment	BASE	6,100	16.68	28.3	0.000
Fragment	PREDNEXT-QRG	6,100	18.28	31.0	0.192
Fragment	LOCAL-QRG	6,100	19.91	33.7	0.265
Scaffold	BASE	6,400	17.99	32.0	0.000
Scaffold	PREDNEXT-QRG	6,400	20.36	36.2	0.194
Scaffold	LOCAL-QRG	6,400	20.18	35.9	0.268
Linker	BASE	4,300	18.82	22.5	0.000
Linker	PREDNEXT-QRG	4,300	19.72	23.6	0.183
Linker	LOCAL-QRG	4,300	21.38	25.5	0.251

E.3 Denoiser-query accounting

The deployed methods are budgeted inference-time corrections rather than many-rollout optimizers. Table 30 records the earlier 500-step diagnostic schedule, not the 1,000-step follow-up configurations in table 3. The counts are delivery opportunities, not hardware-level FLOPs. PredNext and Local add no rollout denoiser calls; the teacher is a separate higher-cost reference.

E.4 Existing assets and licenses

The experiments reuse the public TARGETDIFF model family and the CBGBENCH benchmark/task definitions, both cited in the main text. The upstream TARGETDIFF repository identifies its code license as MIT. The upstream CBGBENCH repository identifies its code license as GPL-3.0, and the CBGBENCH OpenReview paper page identifies the paper license as CC BY 4.0. The original submission code package contains the QRG implementation, runner wrapper, aggregation scripts and configuration description. It excludes upstream benchmark data, checkpoints, generated molecules, logs, raw structures and rendered figures; upstream assets retain their original licenses. Access to the complete follow-up materials is described in appendix F. Users should obtain any upstream data and checkpoints from the original sources under those original terms.

Table 30: Earlier 500-step slice: denoiser-query and guidance-call accounting. The full follow-up uses 1,000 steps (table 3); counts from these schedules are not interchangeable.

Method family	Guidance schedule	Guidance deliveries/sample	Denoiser-query interpretation
BASE	Frozen reverse sampler only	0	No guidance branch and no extra denoiser rollout.
SECTION-ONLY	Section branch, residual budget zero	500	Tests delivery without quotient-residual correction.
PREDNEXT-QRG	Predicted-next residual every two denoising steps	250	Reuses the sampler's predicted next state; deployed point avoids rollout-teacher queries.
LOCAL-QRG	Local-radius quotient surrogate with trust-region delivery	500	Uses a cheap local surrogate rather than querying a rollout teacher.
ROLLOUT-QRG TEACHER	One shared rollout before residual evaluation	500	Diagnostic high-cost teacher/oracle point, not the deployed method.

E.5 Follow-up compute and implementation scaling

The added DiffSBDD comparison requests $3 \times 8 \times 20 = 480$ generations in total. The four-method three-seed TargetDiff matrix requests $4 \times 3 \times (61+43+64) \times 100 = 201{,}600$ generations. The seed-0 records overlap with the earlier 92,800-generation inventory, so the two counts are not additive. A full campaign GPU-allocation-hour total remains unavailable; recorded elapsed times do not by themselves establish scheduler allocation hours.

The current distance construction costs $O(n_l n_p + n_l n_{\mathrm{ctx}} + n_l^2)$ for ligand, pocket and context sizes $n_l, n_p, n_{\mathrm{ctx}}$. Local reduces subsequent residual aggregation to $O(n_l k)$, but still forms dense distance matrices. The same-A100 benchmark uses batch 2 and 20 warmups, followed by five blocks of 20 measured repeats. Timing synchronizes CUDA at measurement boundaries; reported time is the median of block mean times. Incremental peak memory subtracts allocated memory after warmup from the reset peak allocation. The complete 36 method–size records are grouped below into 12 size rows, each showing all three methods.

Table 31: Complete controller-only scaling: no-controller baseline, full pairwise controller, and Local ($k = 4$). Time is ms per step; memory is incremental peak MiB. These are not end-to-end generation costs.

Ligand/pocket	None ms	MiB	Full ms	MiB	Local ms	MiB
16/128	0.048	0.003	10.808	0.262	10.679	0.257
16/256	0.048	0.002	10.832	0.496	10.766	0.491
16/512	0.048	0.002	10.830	0.965	10.725	0.960
32/128	0.048	0.003	13.206	0.545	13.919	0.516
32/256	0.048	0.003	13.154	1.014	13.909	0.985
32/512	0.048	0.003	13.164	1.952	13.918	1.922
64/128	0.048	0.004	13.204	1.207	13.938	1.083
64/256	0.048	0.004	13.213	2.145	13.929	2.021
64/512	0.047	0.004	13.152	4.020	13.923	3.896
128/128	0.048	0.008	13.261	2.916	13.965	2.403
128/256	0.048	0.007	13.284	4.791	13.946	4.278
128/512	0.048	0.007	13.294	8.541	13.962	8.028

Local modestly reduces incremental peak memory at the larger checkpoints but does not consistently reduce controller time. Sparse neighbor kernels remain a future optimization, not an implemented scaling claim.

F Reproducibility and availability

Materials and reported results. The experimental archive contains run summaries, target/seed manifests, resolved configurations, generated SDF/tensor outputs, the frozen controller implementation, the DiffSBDD adapter and sampler-check records, control/calibration outputs, analysis scripts, environment specifications and checkpoint identities. Tables are derived from these recorded outputs.

This appendix reports the complete three-seed rates, chemical-quality contrasts, sensitivity points, direction/scale controls, side-chain lower-budget results, docking comparisons and 36 implementation-scaling records.

Software versions and dependencies. The TargetDiff/CBGBench environment is specified by its upstream revision and working-tree patch, with controller overlays versioned separately. Reproduction on another system requires mapping historical absolute paths to a local working copy while preserving the relative input structure. Frozen training-SMILES caches define the novelty reference. Upstream benchmark data and pretrained checkpoints are identified by source and hash; they must be obtained from the original distributors under their respective terms.

Analysis workflow and verification. The analysis scripts separately aggregate reconstruction summaries, compute chemical descriptors from saved SDFs, and summarize existing docking scores. These analyses use saved inputs without requiring new molecule generation or docking. Absolute rates, paired mean gains and W/L/T have been cross-checked against the 2,016 target records. The reported bootstrap intervals, randomization probabilities, chemical descriptors and sampler checks are the original experimental outputs; they have not been independently recomputed in a clean environment. File-integrity and count checks verify the archived inputs but do not constitute an independent rerun of these analyses.

Record coverage. The DiffSBDD outputs contain accepted molecules but lack a complete mapping to requested IDs and rejection categories. Formal main-run dynamic traces were not saved, and legacy seed-0 invocations lack individual source hashes. The available metadata do not establish a complete parameter-selection chronology or total campaign GPU-allocation hours. The branch-allocation screens vary both branch radii; a ratio-only sweep at fixed total budget was not performed. These distinctions determine which analyses can be reproduced from the saved records.

Access to materials. The complete follow-up archive is currently private, and no public download URL is provided. The original submission code package covers the controller, runner wrapper and analysis utilities, but is not a complete release of the follow-up experiments. Appendices C–E provide the experimental protocols, recorded settings, metric definitions, statistical procedures and full result tables. Reproducing the follow-up analyses additionally requires access to the archived inputs and the software dependencies described above; a complete public package has not been released.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract, introduction and results distinguish the corrected three-seed fragment gain, limited DiffSBDD transfer, diversity decrease, unresolved comparisons and secondary docking evidence. Earlier runtime cohorts are labeled separately.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion and limitations section explicitly states the current scope, the task-dependent linker and side-chain behavior, and the limits of raw docking.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states five formal propositions in the method section and provides full proofs in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: Appendices C–F describe the protocols, configurations, analysis inputs and statistical procedures. Independent clean-environment reproduction is incomplete. Appendix F specifies the private availability of the complete archive and gaps in request-level, source-version and parameter-selection records.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The original submission code package covers part of the study. The complete follow-up archive is private, and no public download URL is provided. Appendix F describes the available documentation and the materials required for reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Section 4 and Appendix C specify seeds, sampling counts, budgets, calibration radii, metric denominators and recorded split checks. Some historical defaults are source-derived and a complete parameter-selection chronology is unavailable, so full historical settings disclosure is not asserted.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4 and Appendix C define the crossed seed/target bootstrap, replicate count, target-level W/L/T, exact cluster sign-flip test and six-test Holm family. Appendix D separates exploratory tests and secondary families; Appendix F distinguishes historical estimates from independent recomputation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix E provides the complete 36-record controller scaling study and separates historical inventory from the follow-up sample matrix. A complete campaign GPU-allocation-hour total remains unavailable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses standard scientific benchmarks, frozen backbones, and diagnostic docking protocols.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a Broader Impact discussion that covers both positive scientific-use cases and the risk of overclaiming or misuse in drug-design settings.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The submission does not release a high-risk model or dataset beyond standard scientific benchmarking assets, so no additional safeguards are required beyond the stated non-release of raw benchmark data/checkpoints.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix E credits the official TARGETDIFF and CBGBENCH repositories and states the identified upstream licenses: TARGETDIFF code is MIT, CBGBENCH code is GPL-3.0, and the CBGBENCH OpenReview paper is CC BY 4.0. The submission does not redistribute upstream data, checkpoints, or generated molecules.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: Appendices C–F document the adapters, controls, configurations and analysis records. The complete follow-up archive remains private; a complete public release with distribution documentation is not available.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, scientific claims, or experimental pipeline of the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.