LEARNING DEPLOYABLE POLICY IMPROVEMENTS FROM MEASURED MOLECULAR PANELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A molecular policy can earn a higher reward yet retain little of that improvement once endpoint regression is controlled. We study this gap in finite ADMET panels: historical measurements cover every candidate, while deployment exposes only molecular structures and endpoint directions. DeltaADMET-R1 combines an exact, anchor-relative multi-endpoint target with a total-variation (TV) controller that bounds regression on each endpoint. Its direct learning algorithm, Contextwise Primal–Dual Policy Optimization with structured Reasoning (CPDPO-R), differentiates the expected utility of the controlled policy using complete action feedback. When the TV budget is active, this objective maximizes gain per unit of policy displacement. We establish target feasibility, contraction under identity-aligned permutation averaging, and a tight outcome-free endpoint bound; source calibration provides a less conservative statistical operating point. On scaffold-disjoint RLM clearance and solubility panels, five-seed Qwen3-8B CPDPO-R exceeds matched scalar projection by $.0000731$ (95% CI $[.0000065, .0001403]$) under the same $\epsilon = .005$ budget. A post hoc joint source-conformal analysis retains mean gain $.000769$, versus $.000641$ for scalar, at nominal 95% coverage. The target-based pipeline yields positive aggregate gains across six Biogen, TDC, and OpenADMET evaluations, with further analyses of larger panels, four endpoints, and missing measurements. Qwen4B and Mistral show that higher raw reward need not translate into a controlled advantage. These results support retained utility as a useful RL objective, with model-family transfer as the principal empirical boundary.

1 INTRODUCTION

Multi-endpoint molecular optimization must balance improvements in one assay against regressions in another. Scalar rewards offer a convenient training signal for solubility, clearance, efflux, and toxicity objectives (Olivecrona et al., 2017; Loeffler et al., 2024; Nguyen et al., 2025). Yet a higher average reward can conceal a weaker endpoint on an individual molecular panel. The practical decision is therefore not simply which policy earns more reward, but which policy update remains useful under a specified endpoint-regression tolerance.

We study this decision after candidate generation. Language-model editors and tool agents can propose molecular alternatives (Liu et al., 2024a; Wu et al., 2024; Wang et al., 2025b); measured assay collections allow us to learn how to rank them (Fang et al., 2023; Huang et al., 2021; Simpkins et al., 2026). Each context consists of a source molecule, a finite candidate set, and endpoint directions. Historical panels provide every candidate's ADMET vector, whereas deployment exposes only structures and directions. The existing action-first language model supplies an anchor against which to measure improvement.

Complete action feedback makes the reference update exactly solvable. For each context, we maximize mean endpoint utility with a reverse-KL penalty and a separate non-regression constraint for every endpoint. The anchor is feasible, and the optimum has full support. This instantiates the separation of target construction from parameter fitting (Kaddour, 2026; Shu et al., 2026) with complete measured actions and contextwise constraints (Hounie et al., 2026). A neural policy can then amortize the update across panels.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

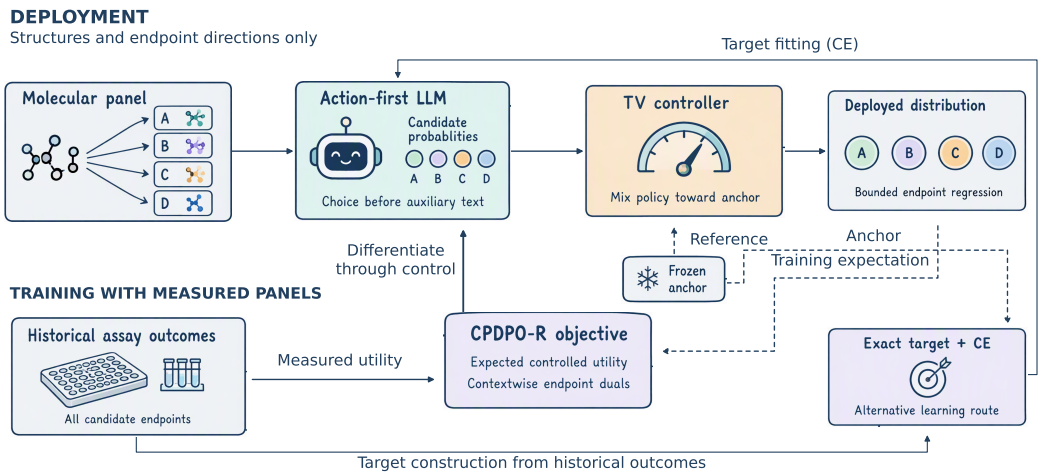

Figure 1: DeltaADMET-R1 separates historical feedback from deployment inputs. CPDPO-R differentiates the expected utility of the TV-controlled policy; target CE instead fits an exact measured update. Both routes use a frozen anchor. Optional permutation averaging aligns both policies by molecule identity before control. Current-panel assay outcomes are unavailable at deployment.

Target feasibility alone does not control the fitted policy. DeltaADMET-R1 therefore interpolates the learned distribution toward its anchor until their total variation (TV) meets a prescribed budget. Bounded utilities turn this observable distance into a tight endpoint-regression bound without current-panel measurements. Identity-aligned averaging across candidate permutations can reduce the distance consumed by serialization. Source calibration offers a less conservative alternative when a statistical, rather than worst-case, guarantee is appropriate.

Our principal algorithm, Contextwise Primal–Dual Policy Optimization with structured Reasoning (CPDPO-R), trains the policy that this controller will deploy. It differentiates exact expectations over all measured actions and updates a dual variable for each context and endpoint. When control is active, the objective is gain per unit TV: learning a productive update direction matters more than increasing raw displacement. The action-first format places the decision before auxiliary evidence and endpoint deltas, allowing structured supervision to shape shared parameters without conditioning the current action on later fields.

The contributions follow this single decision pipeline:

- An exact, unique, full-support target for anchor-relative improvement with contextwise endpoint constraints.
- CPDPO-R, a full-information contextual-bandit method that directly optimizes controlled utility. Its five-seed Qwen3-8B test improvement over scalar $\eta = 4$ is .0000731 $[.0000065, .0001403]$.
- Composable deployment operators: identity averaging preserves order-average endpoint gains and contracts TV; radial interpolation gives a tight outcome-free bound. Joint source-conformal control yields a paired Qwen3-8B gain of .000128 $[.000013, .000243]$ at nominal 95% coverage.
- Evaluation across Biogen, TDC, and OpenADMET, including model scale and family, larger panels, four endpoints, missing measurements, and structural baselines.

The scope is retrospective ranking on supplied candidate panels with bounded assay utilities; clinical safety, causal efficacy, and performance beyond those panels require separate evidence.

2 RELATED WORK

Molecular language models support property prediction, editing, generation, and tool use (Yu et al., 2024; Liu et al., 2024a; Wu et al., 2024; Yue et al., 2024; Xia et al., 2025; Wang et al., 2025a; Zheng et al., 2025). Recent systems combine these capabilities through multi-objective reasoning, agent coordination, and tool trajectories (Zeng et al., 2026; Li et al., 2026a; Liu et al., 2025; 2026; Li et al., 2026b; Le et al., 2026). Molecular RL shifts generators toward property rewards (Olivecrona et al., 2017; Bjerrum et al., 2023; Loeffler et al., 2024; Liu et al., 2024b), while IIFT uses importance-weighted fine-tuning for developability (Berlaga et al., 2026). RL-Pareto and uncertainty-aware diffusion address multiple objectives and uncertain predictors (Nguyen et al., 2025; Chen et al., 2025); POLO combines trajectory optimization with turn-level preferences in multi-turn lead optimization (Wang et al., 2025b). We instead fix the candidate panel and learn an anchor-relative ranking policy from complete measured alternatives.

Constrained policy improvement has been studied through trust regions, baseline bootstrapping, offline-bandit pessimism, and sequential CMDP primal–dual methods (Schulman et al., 2015; 2017; Laroche et al., 2019; Satija et al., 2021; Wang et al., 2024; Stradi et al., 2024). Multi-objective alignment introduces Pareto controls or global constraints (Agnihotri et al., 2025; He et al., 2026; Li et al., 2026c; 2025; He & Maghsudi, 2025; Niu et al., 2026), whereas pointwise learning imposes example-level conditions (Boero et al., 2026; Hounie et al., 2026). Target-projection methods separate distributional improvement from fitting, usually over sampled completions (Kaddour, 2026; Shu et al., 2026; Qu et al., 2026). Complete finite feedback allows us to solve the multi-endpoint target exactly and to retain exact action expectations in CPDPO-R's controlled objective and contextwise dual. Pareto-constrained DPO filters conflicting preference pairs (Xie et al., 2026); our preference variants compare that signal with direct constrained optimization.

Conformal prediction, tolerance limits, and risk-controlling prediction sets calibrate exchangeable scores or bounded losses (Vovk et al., 2005; Wilks, 1941; Bates et al., 2021). Conformal Policy Control clips likelihood ratios from safe-policy samples and uses rejection sampling to control expected risk (Prinster et al., 2026). Our complete feedback instead supports a source-level maximum of endpoint violation divided by TV, calibrated over a fixed finite policy family. Deep Sets, Set Transformers, Janossy pooling, and SetRank encode permutation structure through architecture or aggregation (Zaheer et al., 2017; Lee et al., 2019; Murphy et al., 2019; Pang et al., 2020); permutation self-consistency aggregates LLM rankings with a Kemeny objective (Tang et al., 2024). We average identity-aligned probabilities to preserve linear endpoint expectations and exploit TV convexity.

Measured Biogen, TDC, and OpenADMET assays provide the evaluation data (Fang et al., 2023; Huang et al., 2021; Simpkins et al., 2026). Scaffold partitions and structure-based exclusions address similarity leakage (Joeres et al., 2025). Validation-led development and held-out evaluation also feature in automated molecular property research (Ning et al., 2026). Here, these protocols assess retrospective ranking rather than prospective molecular efficacy.

3 MEASURED-FEEDBACK POLICY IMPROVEMENT

A context x contains a source molecule, a candidate set $\mathcal{A}_x$ of size K_x, and the desired directions of D endpoints. Historical data provide $U_x(a) \in [0,1]^D$ for every candidate. Structures and directions enter the prompt; measurements supervise and evaluate the update. A fixed action-first LLM supplies a full-support anchor $\pi_0(a \mid x) > 0$, and the decision is a distribution over the supplied candidates.

For each oriented assay value z_d, we compute the median m_d and robust scale $r_d = (Q_{.75} - Q_{.25})/1.349$ over unique training molecules, then fix

$$U_{x,d}(a) = \sigma\left(\frac{z_{x,d}(a) - m_d}{\max(r_d, 10^{-8})}\right). \tag{1}$$

This transformation bounds future utilities without current-panel outcomes. Panel-relative analyses are reported separately. With $s_x(a) = D^{-1}\sum_d U_{x,d}(a)$, define

$$G_{x,d}(p) = \langle p - \pi_{0,x}, U_{x,d}\rangle, \qquad g_x(p) = D^{-1}\sum_d G_{x,d}(p), \tag{2}$$

$$V_x(p) = \max_d[-G_{x,d}(p)]_+, \qquad H_x(p) = \mathbf{1}\{\min_d G_{x,d}(p) \geq 0\}. \tag{3}$$

We report mean gain, endpoint gains, mean worst-endpoint violation $\mathbb{E}V_x$, and no-regression rate $\mathbb{E}H_x$. All measure a policy update relative to the same anchor, not a selected molecule relative to the source molecule.

4 Learning the update that can be deployed

4.1 An exact reference update

Complete measured feedback defines a separate constrained improvement problem for every context:

$$q_x^\star = \arg\max_{q\in\Delta_{\mathcal{A}_x}} \{\langle q, s_x\rangle - \tau D_{\mathrm{KL}}(q\|\pi_{0,x})\} \quad \text{s.t. } G_{x,d}(q) \geq 0, \quad d = 1,\dots,D. \tag{4}$$

The reverse-KL penalty controls displacement from the anchor. Equal scalar weights value each endpoint, while separate inequalities prevent improvements on one endpoint from compensating for regression on another. The anchor provides a feasible starting point. The following result characterizes the exact update.

Proposition 1 (Exact measured improvement). *For $\tau > 0$ and a strictly positive anchor, Eq. (4) has a unique full-support optimum, with every $G_{x,d}(q_x^\star) \geq 0$ and $g_x(q_x^\star) \geq \tau D_{\mathrm{KL}}(q_x^\star\|\pi_{0,x}) \geq 0$.*

Strict concavity gives uniqueness. At a boundary zero, a feasible mixture toward the positive anchor increases entropy fast enough to improve the objective, establishing full support. The KKT distribution is

$$q_{x,\lambda}(a) \propto \pi_0(a \mid x)\exp\{[s_x(a) + \lambda^\top U_x(a)]/\tau\}, \qquad \lambda \geq 0. \tag{5}$$

We minimize the convex dual in float64. For two endpoints, monotone one-multiplier bisection handles a single active constraint; the full dual and a primal fallback handle coupled active sets. The solver checks feasibility before returning a target (Appendix B).

Proposition 2 (Constraint granularity). *For two or more contexts, $\mathbb{E}_x G_{x,d}(q) \geq 0$ for every endpoint does not imply $G_{x,d}(q) \geq 0$ for every context and endpoint.*

Global constraints allow gains on one panel to offset losses on another. Full action support also matters: restricting Eq. (4) to sampled actions can remove every feasible distribution. Appendix Eq. (26) gives the exact probability that N anchor samples retain feasible support.

4.2 Outcome-free deployment control

For a learned policy p, let $t_x = \mathrm{TV}(p, \pi_{0,x}) = \frac{1}{2}\|p - \pi_{0,x}\|_1$. Given $\epsilon \geq 0$, define

$$C_\epsilon(p; \pi_{0,x}) = \pi_{0,x} + \beta_x(p - \pi_{0,x}), \qquad \beta_x = \min\{1, \epsilon/t_x\}, \tag{6}$$

with $\beta_x = 1$ when $t_x = 0$. The controller needs only the learned and anchor distributions.

Proposition 3 (Tight outcome-free control). *For any $U_{x,d} : \mathcal{A}_x \to [0,1]$, $G_{x,d}(C_\epsilon(p;\pi_{0,x})) \geq -\mathrm{TV}(C_\epsilon(p;\pi_{0,x}), \pi_{0,x}) \geq -\epsilon$. The bound is tight, and Eq. (6) retains the largest coefficient among radial mixtures satisfying the worst-case TV budget. An outcome-free policy guaranteeing zero regression for every bounded endpoint must equal the anchor.*

Linearity gives $G_{x,d}(C_\epsilon(p)) = \beta_x G_{x,d}(p)$. In the active regime $t_x > \epsilon$, retained mean gain is $\epsilon g_x(p)/t_x$: the controller rewards gain per unit displacement. Away from absolute-value kinks,

$$\nabla_\theta g_x(C_\epsilon(\pi_{\theta,x})) = \frac{\epsilon}{t_x}\left[\nabla_\theta g_x(\pi_{\theta,x}) - \frac{g_x(\pi_{\theta,x})}{t_x}\nabla_\theta t_x\right]. \tag{7}$$

The second term removes outward movement that does not improve this ratio. Holding β_x constant drops that correction, as tested by the stop-gradient ablation. We compare policies at $\epsilon = .005$ and report tolerance curves; this budget is measured in normalized assay utility.

4.3 CPDPO-R: DIRECT OPTIMIZATION OF RETAINED UTILITY

Let $\widetilde{\pi}_{\theta,x} = C_\epsilon(\pi_{\theta,x}; \pi_{0,x})$, $v_{x,d} = -G_{x,d}(\widetilde{\pi}_{\theta,x})$, and $\lambda_{x,d} \geq 0$. CPDPO-R minimizes

$$\mathcal{L}_{\mathrm{RL}} = -w_q \mathbb{E}_x g_x(\widetilde{\pi}_{\theta,x}) + w_r \mathcal{L}_{\mathrm{LM}} + w_c \mathbb{E}_{x,d}\left[\alpha\left[v_{x,d} + \frac{\lambda_{x,d}}{2\alpha}\right]_+^2 - \frac{\lambda_{x,d}^2}{4\alpha}\right]. \tag{8}$$

The projected supergradient update is

$$\lambda_{x,d} \leftarrow \Pi_{[0,\lambda_{\max}]}\left(\lambda_{x,d} + \eta_\lambda \max\{v_{x,d}, -\lambda_{x,d}/(2\alpha)\}\right).$$

This adapts pressure to each context's active endpoint, following pointwise constrained learning (Hounie et al., 2026). Both reward and constraint expectations sum over every measured action, and differentiation includes the radial coefficient in Eq. (6). On RLM–solubility we use $w_q = w_c = 1000$, $w_r = .5$, $\alpha = 10$, $\eta_\lambda = .1$, and $\lambda_{\max} = 1000$. Dual state is updated for each visited training example during loss evaluation.

The action-first LoRA LLM (Hu et al., 2022) normalizes single-token candidate logits. Its supervised continuation contains structural evidence, endpoint deltas, and confidence; the choice token is masked from $\mathcal{L}_{\mathrm{LM}}$. These later fields shape shared parameters through training but cannot enter the current teacher-forced action's hidden state. CPDPO-R is thus one-step, full-information contextual-bandit RL with exact action expectations and minibatch/optimizer randomness.

The target-amortization comparison uses the same policy with

$$\mathcal{L}_{\mathrm{target}} = \alpha_q \mathbb{E}_x[-\langle q_x^\star, \log \pi_{\theta,x}\rangle] + \alpha_c \mathbb{E}_{x,d}[-G_{x,d}(\pi_{\theta,x})]_+ + \alpha_r \mathcal{L}_{\mathrm{LM}}. \tag{9}$$

The two routes share measured feedback: one fits the exact reference update, while the other optimizes its deployed realization.

4.4 IDENTITY-ALIGNED PERMUTATION AVERAGING

Candidate labels serialize a molecular set. Let $\mathcal{G}_x$ be its permutation group and R_g map probabilities back to molecule-identity order. For either learned or anchor policy, define

$$\bar{\rho}_x = \mathcal{S}[\rho]_x = |\mathcal{G}_x|^{-1} \sum_{g \in \mathcal{G}_x} R_g \rho(\cdot \mid x_g). \tag{10}$$

This group average (Murphy et al., 2019) preserves linear endpoint objectives. Write $V_x(r; p) = \max_d[-\langle r - p, U_{x,d}\rangle]_+$.

Proposition 4 (Identity averaging contracts deployment distance)**.** *Full-orbit averaging is invariant to candidate reserialization and satisfies*

$$\mathrm{TV}(\bar{\pi}_{\theta,x}, \bar{\pi}_{0,x}) \leq \mathbb{E}_g \,\mathrm{TV}(R_g \pi_\theta(\cdot \mid x_g), R_g \pi_0(\cdot \mid x_g)), \tag{11}$$

$$\langle \bar{\pi}_{\theta,x} - \bar{\pi}_{0,x}, U_{x,d}\rangle = \mathbb{E}_g \langle R_g(\pi_\theta - \pi_0)(\cdot \mid x_g), U_{x,d}\rangle, \tag{12}$$

$$V_x(\bar{\pi}_{\theta,x}; \bar{\pi}_{0,x}) \leq \mathbb{E}_g V_x(R_g \pi_\theta(\cdot \mid x_g); R_g \pi_0(\cdot \mid x_g)). \tag{13}$$

Applying C_ϵ to the averaged pair gives an invariant controlled policy. TV contraction can retain more displacement, although realized gain need not exceed every individual ordering. We enumerate all 24 orders at $K = 4$ and use shared sampled orders at larger K; Appendix B gives the approximation bound. The direct-RL comparison uses canonical-order control, while the target-based study evaluates orbit deployment.

4.5 SOURCE-CALIBRATED RETENTION

For a fixed finite policy family $\mathcal{F}$, define a source-cluster score

$$S_c = \max_{f \in \mathcal{F},\, x \in c} \frac{V_x(\pi_f)}{\mathrm{TV}(\pi_{f,x}, \pi_{0,f,x})} \in [0, 1], \tag{14}$$

with zero for a zero denominator. Replacing t_x by qt_x in Eq. (6) bounds regression by ϵ across all included panels, endpoints, and members on the event $S_c \leq q$. We use two calibration rules

with distinct guarantees. For n exchangeable calibration sources and an independently fixed family, the marginal conformal rule sets $q_{\mathrm{SC}} = S_{(\lceil (n+1)(1-\delta) \rceil)}$, or 1 when the rank exceeds n, giving $\Pr(S_{\mathrm{new}} \leq q_{\mathrm{SC}}) \geq 1 - \delta$.

The PAC tolerance rule instead specifies source coverage $1 - \gamma$ and calibration confidence $1 - \alpha$. It chooses the smallest k with $\Pr\{B_{k,n+1-k} \geq 1 - \gamma\} \geq 1 - \alpha$, where B is beta distributed, and sets $q_{\mathrm{PAC}} = S_{(k)}$ (Wilks, 1941). If no rank qualifies, it uses $q = 1$.

Proposition 5 (PAC source-family certificate)**.** *With probability at least $1 - \alpha$ over independent calibration sources, a new exchangeable source satisfies $G_{x,d}(\pi_f^{(\epsilon)}) \geq -\epsilon$ simultaneously for every included panel, endpoint, and fixed family member with probability at least $1 - \gamma$.*

The joint CPDPO-R/scalar analysis uses marginal conformal ranks; the HLM-based PAC experiments use calibration/confirmation splits. Both require source exchangeability and a fixed family. The deterministic $q = 1$ controller has neither distributional requirement (Appendix G.14).

5 Experimental design

The experiments examine exact target construction, direct optimization of the deployed update, and robustness to serialization and assay shift. Biogen contains 3,521 compounds with six in-vitro ADME assays (Fang et al., 2023). We canonicalize molecules, split Bemis–Murcko scaffolds (Bemis & Murcko, 1996), and construct candidate panels using radius-2 Morgan fingerprints (Rogers & Hahn, 2010), with source Tanimoto similarity at least .20. RLM–solubility provides the direct-RL comparison; HLM-based tasks evaluate target fitting and control.

Table 1: Four-candidate panels. External tasks use no model fine-tuning. HLM and RLM denote human and rat microsomal intrinsic clearance.

Data	oriented endpoints	Train	Valid.	Test
Biogen	RLM↓, solubility↑	1,696	265	295
Biogen	HLM↓, solubility↑	3,626	599	716
Biogen	HLM↓, MDR1↓	4,998	1,003	916
Biogen	HLM↓, RLM↓, Sol.↑, MDR1↓	1,995	330	391
TDC	microsome↓, hepatocyte↓	0	0	1,170
OpenADMET	CYP3A4↑, CYP2J2↑	0	0	2,168
OpenADMET	above, CYP3A4 pIC_{50} ↓	0	0	938

TDC joins AstraZeneca microsomal and hepatocyte clearance assays (Huang et al., 2021). OpenADMET uses the July 2026 CYP release and well-level replicates (Simpkins et al., 2026); its 3D task adds inhibition pIC50 to two remaining-fraction endpoints. External panels exclude exact or scaffold overlap between any source/candidate and Biogen. Both OpenADMET tasks and a 375-panel TDC subset require candidate–source similarity of at least .35.

Primary Biogen utilities use Eq. (1). External fixed-scale analyses calibrate a sigmoid on 30% of scaffolds and retain panels entirely within the remaining 70%, without retraining the policy. Twenty fixed partitions assess split sensitivity. The real 8/16-candidate HLM–solubility tasks contain 588/381 test panels; their panel-relative scales are identified separately.

We use Qwen3-4B, Qwen3-8B (Qwen Team, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), with a common single-endpoint HLM anchor within each family. RLM–solubility uses five optimizer seeds $(11, 37, 73, 101, 2026)$ at 8B and three at 4B. CPDPO-R and scalar $\eta = 4$ share 200 steps, rank/alpha 16/32 LoRA, learning rate 10^{-4}, batch size 4 with four-step accumulation, BF16, and masked language-loss weight .5. The 8B comparison is against scalar; 4B additionally includes target CE and exact PG.

Target-based controls include scalar mirror projection, Tchebycheff scalarization, exact PG with reverse KL, matched endpoint hinges, and global/pointwise augmented duals (Hounie et al., 2026). A Morgan panel MLP and a position-free ChemBERTa–Set Transformer test structural representations (Chithrananda et al., 2020; Lee et al., 2019). Molecular encoders use their own architectures and budgets; matched LLM-loss comparisons share architecture and measured feedback. Appendix E gives configurations and selection protocols.

Evaluation scores action probabilities, with learned and controlled policies reported separately. Direct RL uses canonical order and the same $\epsilon = .005$ controller for every method. The target-based order study evaluates all 24 permutations with probability, plurality, and Kemeny aggregation (Tang et al., 2024); larger panels use $M \in \{1, 2, 4, 8, 16, 32, 64\}$ shared orders. Current-panel outcomes enter neither action scoring nor deterministic control.

Five-seed intervals jointly resample optimizer seeds and source clusters over 50,000 draws. Two-member analyses condition on trained members; three-seed analyses include optimizer variation. We report paired two-sided 95% intervals, molecule-disjoint dependence sensitivity, and 5,000 replicate-error draws for OpenADMET measurement uncertainty. The 20 external partitions and 20 calibration splits reuse observations and policies and therefore measure sensitivity, not independent replication. Joint source-conformal results are post hoc; prospective coverage requires calibration sources unused in policy selection.

6 RESULTS

6.1 DIRECT RL IMPROVES RETAINED UTILITY AT 8B

The RLM–solubility results separate reward growth from deployable improvement (Table 2; Figure 2). At 4B, CPDPO-R raises raw gain from .013935 to .019132 relative to scalar $\eta = 4$, a paired difference of .005197 $[.002982, .007563]$, and exceeds target CE and exact PG. Endpoint violation also rises. Under the common controller, gains are .000431 and .000435, leaving the controlled comparison unresolved.

Table 2: RLM–solubility test results, multiplied by 10^3. V is mean worst-endpoint violation. The 4B/8B rows average three/five optimizer seeds; every controlled policy uses deterministic $\epsilon = .005$.

Model	Action objective	Raw gain	Raw V	Controlled gain	Controlled V
Qwen3-4B	Target CE	3.442	4.740	.300	.585
	Exact PG	3.304	4.316	.302	.571
	Scalar $\eta = 4$	13.935	12.209	.435	.522
	CPDPO-R	19.132	21.247	.431	.566
Qwen3-8B	Scalar $\eta = 4$	10.56	11.30	.3665	.520
	CPDPO-R	20.00	21.62	.4396	.530

At 8B, CPDPO-R retains .0004396 $[.0003362, .0005428]$, versus .0003665 $[.0002701, .0004621]$ for scalar. The paired improvement is .0000731 $[.0000065, .0001403]$. Both endpoint gains relative to the anchor are positive: .0004885 for RLM and .0003907 for solubility. The paired mean-violation difference is .0000098 $[-.0000527, .0000711]$. This interval does not establish equivalence, despite the shared deterministic worst-case tolerance. Raw gain improves by .00944 $[.00601, .01305]$; the controlled hypervolume difference is positive but its interval crosses zero (Appendix G.5).

Figure 3 separates model scale from endpoint-level evidence. Raw gain increases at both scales, but the retained advantage is resolved only at 8B. Its aggregate paired interval is positive; the separate endpoint intervals cross zero. Gains relative to the anchor therefore do not imply endpoint-specific superiority over scalar.

6.2 CALIBRATION RETAINS MORE OF THE UPDATE

A shared source score takes the maximum over all five CPDPO-R and five scalar members. At nominal 95% marginal coverage, $q_{\mathrm{SC}} = .571858$ yields CPDPO-R gain .000769 and paired advantage .000128 $[.000013, .000243]$ (Table 3). Across its members, 193 of 197 test sources meet the CPDPO-R coverage criterion. Both methods use the same calibrated coefficient. Across 20 source-hash calibration splits, the 95% coefficient's 5th/median/95th percentiles are $[.4931, .5719, .6001]$, and paired test-gain percentiles are $[.000122, .000128, .000148]$. These characterize recalibration sensitivity on reused policies and observations. In the fully contracted regime, a shared q rescales the same gain difference. On HLM-based tasks, the separate 90%-coverage/95%-confidence PAC rule also retains more gain than $q = 1$ for Qwen and Mistral (Appendix G.14).

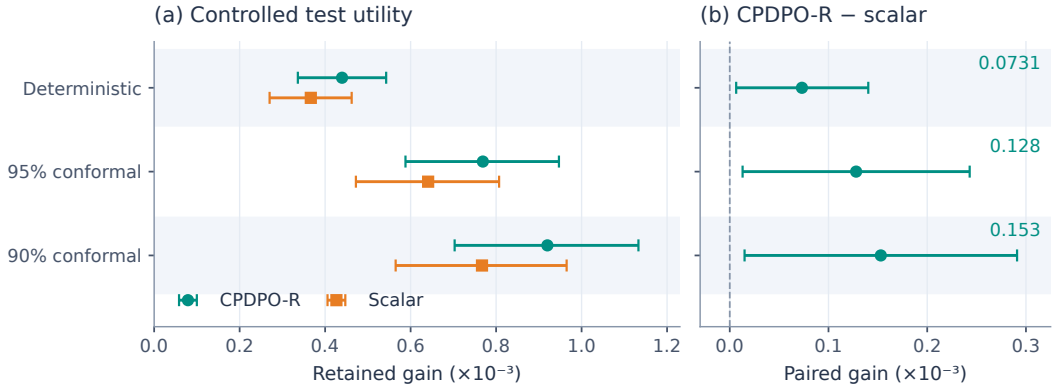

Figure 2: The training objective matters after control. Left: Qwen3-8B test gain under deterministic and joint source-conformal control. Right: paired CPDPO-R–scalar differences with reported 95% intervals. The conformal points are post hoc analyses of the same policies, not independent replications. Units are 10^{-3} normalized utility.

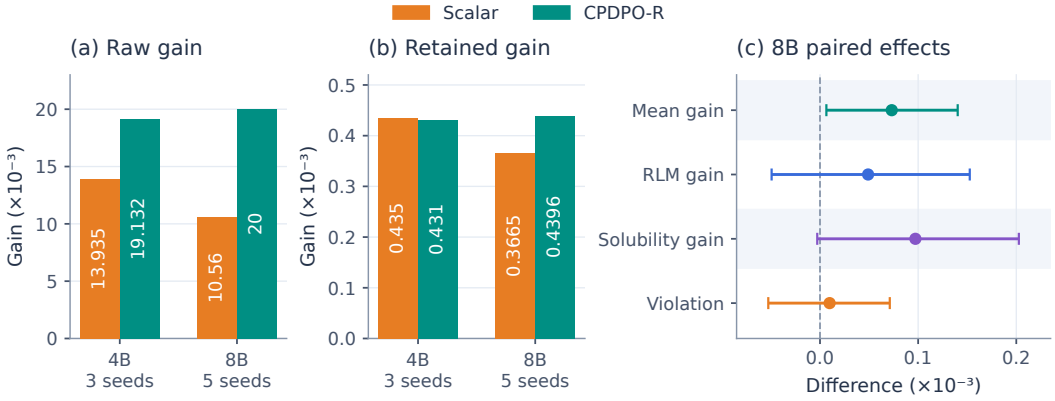

Figure 3: Model scale changes the controlled comparison. Left and center: raw and controlled RLM–solubility test means from Table 2, using three/five seeds at 4B/8B and the same $\epsilon = .005$ budget at both model scales. Right: five-seed 8B CPDPO-R–scalar differences with reported 95% intervals (Appendix G.5 and Section 6.1). Positive violation differences mean more violation. The three axes have different scales; bars show means without inferred uncertainty.

Table 3: Joint source-conformal control on Qwen3-8B. Gains and paired differences are multiplied by 10^3; brackets give paired 95% intervals. Covered sources meet $\epsilon = .005$ across all CPDPO-R members, panels, and endpoints. These marginal ranks differ from the appendix's PAC tolerance rule.

Nominal coverage	q_{SC}	Covered	CPDPO-R	Scalar	Paired gain
90%	.477972	187/197	.920	.767	.153 [.015, .291]
95%	.571858	193/197	.769	.641	.128 [.013, .243]

6.3 WHERE IMPROVEMENT IS LOST

Contextwise constraints protect panels that global targets sacrifice. The global-target witness rates are 3.14% on HLM–solubility, 6.99% on HLM–MDR1, 3.99% on OpenADMET-2D, and 44.99% on OpenADMET-3D. For the first two tasks, scalar values differ by less than 3×10^{-5}, making the lost protection nearly invisible to aggregate reward. Sampling only two actions preserves feasible support in 46.22% and 27.47% of OpenADMET-2D/3D contexts (Tables 22–23).

Neural fitting introduces a second gap. Under fixed-sigmoid HLM–solubility utilities, the exact target gains .03435, while the fitted 4B policy gains .00388 at target–policy TV .0984. A three-seed ChemBERTa–Set Transformer obtains raw gain/violation $.002178/.001710$ and controlled gain .000487, versus $.003878/.003651$ and .000393 for the canonical LLM target policy. Thus the LLM's larger raw gain does not imply a controlled advantage over the structural encoder. On the direct-RL 8B task, target–policy TV correlates with raw violation ($\rho = .519$), whereas minimum anchor support does not explain that distance. Equation (6) contracts each panel separately. Raw gains concentrated on heavily contracted updates can therefore contribute little to the final controlled comparison.

Serialization consumes deployment budget even when mean utility changes little. Identity averaging on fixed-sigmoid HLM–solubility changes gain from .003878 to .003875, lowers violation from .003651 to .002534, and increases controlled gain from .000393 to .000509 $[.000436, .000582]$. Probability averaging also retains more certified gain than plurality or exact Kemeny aggregation across the reported panels. This pattern accords with linear endpoint expectations and TV contraction (Appendix G.13).

6.4 TRANSFER ACROSS PANELS AND MODEL FAMILIES

The target pipeline has positive aggregate intervals on six retrospective evaluations. Under external scaffold-disjoint fixed scales, TDC and OpenADMET-2D retain positive aggregate and controlled intervals in 18/20 partitions, versus 14/20 for OpenADMET-3D. The 8/16-candidate controlled gains are $.001497/.001484$. A separate panel-relative four-endpoint test gains .004762 $[.003965, .005575]$, with positive endpoint intervals; fixed-scale four-endpoint results are less uniform. At 50% masked training cells, interval-based learning retains gain .000893 $[.000178, .001622]$.

Mistral reproduces HLM target fitting and PAC retention gains, but its fixed-anchor direct-RL validation gain is .0002121, versus .0003485 for scalar, with an unresolved paired interval; joint-calibrated test results also favor scalar. Stop-gradient TV, Pareto preferences, entropy ensembles, and a second RL round can increase raw gain without improving controlled utility. A Nash objective improves Qwen validation but does not transfer to Mistral or solubility–MDR1. Structured rationale supervision helps one target-based 4B comparison, whereas fourfold counterfactual continuations do not improve CPDPO-R (Appendix I). These results identify learning the update direction, rather than its raw magnitude, as the recurring challenge.

7 CONCLUSION

Measured molecular panels connect exact policy improvement to the update deployed in practice. Complete feedback defines a contextwise constrained target; TV control determines how much of a learned update is retained. CPDPO-R optimizes retained utility directly, improving five-seed Qwen3-8B performance over matched scalar projection. Identity averaging and source calibration operate on the same anchor-relative update, linking set representation to endpoint control. Together, these results support retained utility as a useful RL objective rather than raw reward alone.

Model-family transfer remains the principal empirical boundary: the Qwen4B and Mistral comparisons distinguish increasing raw reward from improving a controlled policy. The present results concern retrospective finite-panel ranking with modest bounded gains. Source reuse, adaptive model development, and assay shift limit population conclusions; prospective calibrated deployment requires independent calibration data. Learning productive update directions across model families and chemical series, evaluated at a common decision budget, is the next step.

AI USE STATEMENT

Generative AI tools assisted hypothesis development, methodology and experiment design, implementation, analysis, literature retrieval, and manuscript drafting and revision. Generative AI also assisted the conceptual overview illustration and language editing; quantitative figures were plotted from the reported values using Python. The authors are responsible for the final text, code, analyses, mathematical arguments, and citations.

ETHICS STATEMENT

This study uses public retrospective molecular assay data and involves no human participants. Its guarantees concern expectations of bounded assay utilities on supplied candidate panels. Clinical efficacy, patient toxicity, and synthesizability require separate experimental evidence.

REPRODUCIBILITY STATEMENT

Appendices B–H describe the optimization problem, proofs, numerical tolerances, data construction, hyperparameters, baselines, metrics, and statistical procedures; Appendix I details additional optimization studies. Dataset identifiers and licenses appear in Appendix D. The accompanying revision package includes the plotting data, Python scripts, and document build resources; the study's supplementary code covers preprocessing, training, evaluation, uncertainty analysis, and numerical validation.

REFERENCES

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*, 2025. URL https://arxiv.org/abs/2505.10892.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

Alex Berlaga, Maria Roman, Scott Snyder, and Andrew L. Ferguson. Shifting a molecular generator toward developability with iterative importance fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Esben Jannik Bjerrum, Christian Margreitter, Thomas Blaschke, Simona Kolarova, and Raquel López-Ríos de Castro. Faster and more diverse de novo molecular optimization with double-loop reinforcement learning using augmented SMILES. *Journal of Computer-Aided Molecular Design*, 37(8):373–394, 2023. doi: 10.1007/s10822-023-00512-6.

Ignacio Boero, Ignacio Hounie, Luiz Chamon, and Alejandro Ribeiro. Everywhere learning: Artificial intelligence with pointwise constraints. *arXiv preprint arXiv:2606.01557*, 2026.

Lianghong Chen, Dongkyu Eugene Kim, Mike Domaratzki, and Pingzhao Hu. Uncertainty-aware multi-objective reinforcement learning-guided diffusion models for 3d de novo molecular design. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b4a804ef8f138e301680f6dda6253be4-Abstract-Conference.html.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. In *NeurIPS Workshop on Machine Learning for Molecules*, 2020.

Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023. doi: 10.1021/acs.jcim.3c00160.

Qiang He and Setareh Maghsudi. Pareto multi-objective alignment for language models. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2025.

Qiang He, Yucheng Yang, Tianyi Zhou, Meng Fang, Mykola Pechenizkiy, and Setareh Maghsudi. One model for all: Multi-objective controllable language models. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2604.04497.

Ignacio Hounie, Ignacio Boero, and Alejandro Ribeiro. Every sample counts: Supervised fine-tuning of language models with pointwise constraints. *arXiv preprint arXiv:2607.08968*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2102.09548.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16(1), 2025. doi: 10.1038/s41467-025-58606-8.

Jean Kaddour. Target policy optimization. *arXiv preprint arXiv:2604.06159*, 2026. URL https://arxiv.org/abs/2604.06159.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3891–3898, 2019. doi: 10.1609/aaai.v33i01.33013867.

Khiem Le, Sreejata Dey, Marcos Martínez Galindo, Vanessa Lopez, Ting Hua, Nitesh V. Chawla, and Hoang Thanh Lam. Can decision trees teach large language models? distilling verbalized knowledge for molecular property prediction. *arXiv preprint arXiv:2603.12344*, 2026. URL https://arxiv.org/abs/2603.12344.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Lingxiao Li, Haobo Zhang, Ruohao Fan, Bin Chen, and Jiayu Zhou. Molecular lead optimization via agentic tool planning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026a. doi: 10.1145/3770855.3819017.

Xuan Li, Zhanke Zhou, Zongze Li, Jiangchao Yao, Yu Rong, Lu Zhang, and Bo Han. Reference-guided policy optimization for molecular optimization via LLM reasoning. *arXiv preprint arXiv:2603.05900*, 2026b. URL https://arxiv.org/abs/2603.05900.

Yining Li, Peizhong Ju, and Ness Shroff. Provable last-iterate convergence for multi-objective safe LLM alignment via optimistic primal-dual. *arXiv preprint arXiv:2602.22146*, 2026c. URL https://arxiv.org/abs/2602.22146.

Haoran Liu, Zheni Zeng, Yukun Yan, Yuxuan Chen, and Yunduo Xiao. DrugR: Optimizing molecular drugs through LLM-based explicit reasoning. *arXiv preprint arXiv:2602.08213*, 2026. URL https://arxiv.org/abs/2602.08213.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=yRrPfKyJQ2.

Xuefeng Liu, Songhao Jiang, Siyu Chen, Zhuoran Yang, Yuxin Chen, Ian Foster, and Rick Stevens. DrugImproverGPT: A large language model for drug optimization with fine-tuning via structured policy optimization. *arXiv preprint arXiv:2502.07237*, 2025. URL https://arxiv.org/abs/2502.07237.

Yijing Liu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Wei Chen. Graph diffusion policy optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9585–9611, 2024b. doi: 10.52202/079017-0305.

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

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJluy2RcFm.

Hoang-My Nguyen, Nguyet-Hang Vu, Hoang Thanh Lam, and Hoang D. Nguyen. Pareto-guided reinforcement learning for multi-objective ADMET optimization in generative drug design. In *NeurIPS 2025 Workshop on AI for Science*, 2025. URL https://openreview.net/forum?id=naWLNIblGq.

Jingjie Ning, Xiaochuan Li, Ji Zeng, Chenyan Xiong, and Guolin Ke. Closed-loop auto research for molecular property prediction: Discovering and certifying generalizable improvements. *arXiv preprint arXiv:2606.22731*, 2026.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. Mitigating the safety alignment tax with null-space constrained policy optimization. In *International Conference on Learning Representations*, 2026.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1), 2017. doi: 10.1186/s13321-017-0235-x.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 705–714, 2020. doi: 10.1145/3397271.3401104.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. *arXiv preprint arXiv:2603.02196*, 2026. URL https://arxiv.org/abs/2603.02196.

Yun Qu, Qi Wang, Yixiu Mao, Heming Zou, Yuhang Jiang, Yingyue Li, Wutong Xu, Lizhou Cai, Weijie Liu, Clive Bai, Kai Yang, Yangkun Chen, Saiyong Yang, and Xiangyang Ji. Listwise policy optimization: Group-based RLVR as target-projection on the LLM response simplex. *arXiv preprint arXiv:2605.06139*, 2026. URL https://arxiv.org/abs/2605.06139.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

RDKit contributors. RDKit: Open-source cheminformatics. https://www.rdkit.org/, 2026. Accessed 2026-08-29; archived releases at doi:10.5281/zenodo.591637.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Harsh Satija, Philip S. Thomas, Joelle Pineau, and Romain Laroche. Multi-objective SPIBB: Seldonian offline policy improvement with safety constraints in finite MDPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0f65caf0a7d00afd2b87c028e88fe931-Abstract.html.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Yao Shu, Chenxing Wei, Hongbin Lin, Shuang Qiu, and Hui Xiong. Reference-sampled boltzmann projection for KL-regularized RLVR: Target-matched weighted SFT, finite one-shot gaps, and policy mirror descent. *arXiv preprint arXiv:2605.02469*, 2026. URL https://arxiv.org/abs/2605.02469.

Scott Simpkins, Robert Warneford-Thomson, Steven Edgar, and Hugo MacDermott-Opeskin. Data release: Building the OpenADMET data engine. Zenodo dataset, 2026. URL https://zenodo.org/records/21450795.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46692–46721. PMLR, 2024. URL https://proceedings.mlr.press/v235/stradi24a.html.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2327–2340, 2024. doi: 10.18653/v1/2024.naacl-long.129.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. TxGemma: Efficient and agentic LLMs for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025a. URL https://arxiv.org/abs/2504.06196.

Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 766–774, 2024.

Ziqing Wang, Yibo Wen, William Pattie, Xiao Luo, Weimin Wu, Jerry Yao-Chieh Hu, Abhishek Pandey, Han Liu, and Kaize Ding. Polo: Preference-guided multi-turn reinforcement learning for lead optimization. *arXiv preprint arXiv:2509.21737*, 2025b. URL https://arxiv.org/abs/2509.21737. Accepted at ACM SIGKDD 2026 (Oral).

S. S. Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941. doi: 10.1214/aoms/1177731788.

Zhenxing Wu, Odin Zhang, Xiaorui Wang, Li Fu, Huifeng Zhao, Jike Wang, Hongyan Du, Dejun Jiang, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nature Machine Intelligence*, 6(11):1359–1369, 2024. doi: 10.1038/s42256-024-00916-5.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, et al. Nature language model: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*, 2025. URL https://arxiv.org/abs/2502.07527.

Hongyan Xie, Yikun Ban, Ruiyu Fang, Di Wu, Zixuan Huang, Deqing Wang, Jianxin Li, and Shuangyong Song. Direct preference optimization with pareto dominance constraint for online multi-objective alignment. *Neurocomputing*, 702:134604, 2026. doi: 10.1016/j.neucom.2026.134604.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=lY6XTF9tPv.

Jie Yue, Bingxin Peng, Yu Chen, Jieyu Jin, Xinda Zhao, Chao Shen, Xiangyang Ji, Chang-Yu Hsieh, Jianfei Song, Tingjun Hou, Yafeng Deng, and Jike Wang. Unlocking comprehensive molecular design across all scenarios with large language model and unordered chemical language. *Chemical Science*, 15(34):13727–13740, 2024. doi: 10.1039/d4sc03744h.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Daojian Zeng, Tianle Li, Jiahao Yang, Jiacai Yi, Xieping Gao, Lincheng Jiang, Tengfei Ma, and Xiangxiang Zeng. Expert-inspired multi-agent coordination for multi-objective molecular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34575–34583, 2026. doi: 10.1609/aaai.v40i41.40757.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447, 2025. doi: 10.1038/s42256-025-00994-z.

A CLOSEST-WORK ASSUMPTION MAP

Table 4 locates the contribution in assumption space. Prior work supplies Boltzmann or reverse-KL targets, multi-objective and baseline-relative constraints, pointwise learning, and conformal policy control as separate ingredients. DELTAADMET-R1 combines complete measured-action feedback with an exact contextwise target, an LLM anchor, full support, action-first amortization, identity-aligned orbit averaging, and a tight outcome-free ϵ-certificate. A source-level tolerance layer then calibrates conservatism for a fixed policy family.

Table 4: Assumption-level comparison with the closest method families. "Global" constraints average prompts or trajectories; "target" guarantees refer to the constructed distribution unless a separate policy certificate is provided.

Method family	feedback	constraint granularity	support / fitting path	guarantee location
TPO / BOLT / LPO	sampled scalar scores	no multi-endpoint lower bounds	sampled list; CE / weighted SFT	sampled target/objective
MOPO	pairwise preferences	global primary + secondary thresholds	reference policy; dual iterations	global optimization target
MOC / OPD	sampled multi-reward	global expectation / preference conditioned	PPO or primal–dual	expected constraint / convergence
Everywhere / Every Sample	sampled task losses	per sample, optionally relaxed	augmented Lagrangian; sample duals	parameterized empirical constraints
Multi-Objective SPIBB	offline MDP transitions	baseline-relative objective returns	bootstrapped MDP policy	high-probability return
CPC	sampled safe-policy actions / losses	expected bounded deployment risk	likelihood-ratio clipping; rejection sampling	selected-policy finite-sample risk
RL-Pareto / uncertainty RL diffusion	learned ADMET/property predictors	generative multi-objective rewards	chemical-space generation; Pareto or uncertainty shaping	predicted properties / generation metrics
Ours	measured vectors for all panel actions	per context, all endpoints vs. LLM anchor	exact target; action-first CE or direct primal–dual PG	exact target; deterministic / calibrated control

B PROOFS AND FORMAL DETAILS

B.1 PROOF OF PROPOSITION 1

Fix a context and suppress x. The probability simplex is compact. The anchor $p = \pi_0(\cdot \mid x)$ is feasible because every endpoint constraint holds with equality at $q = p$. The objective is continuous, so a maximizer exists. Because $p_a > 0$ for every retained action, $D_{\mathrm{KL}}(q\|p)$ is finite on the simplex and strictly convex in q on any nontrivial face. Therefore $q^\top s - \tau D_{\mathrm{KL}}(q\|p)$ is strictly concave and the maximizer is unique. The endpoint inequalities are the primal constraints. To establish full support, suppose an optimum has a zero coordinate. The feasible mixture $q_t = (1-t)q^\star + tp$ restores each zero coordinate. The contribution to $-\tau \mathrm{KL}(q_t\|p)$ from these coordinates is a positive term of order $-t \log t$, whereas changes on positive coordinates and the linear reward are of order t. For sufficiently small $t > 0$, the objective increases, contradicting optimality. Finally, comparison with the anchor gives

$$(q^\star)^\top s - \tau D_{\mathrm{KL}}(q^\star\|p) \geq p^\top s - \tau D_{\mathrm{KL}}(p\|p) = p^\top s, \tag{15}$$

$$(q^\star - p)^\top s \geq \tau D_{\mathrm{KL}}(q^\star\|p) \geq 0. \tag{16}$$

This proves feasibility, uniqueness, endpoint no-regression, and nonnegative scalar improvement for the exact target; Proposition 3 handles its parameterized approximation.

B.2 PROOF OF PROPOSITION 3

Let $r = \pi^{(\epsilon)}_{\theta,x}$, $p = \pi_{0,x}$, and let $S_- = \{a : r(a) - p(a) < 0\}$. For any $U : \mathcal{A} \to [0, 1]$,

$$(r-p)^\top U \geq \sum_{a \in S_-} (r(a) - p(a)) = -\frac{1}{2}\|r - p\|_1 = -\,\mathrm{TV}(r, p). \tag{17}$$

The inequality is attained by setting $U(a) = 1$ on S_- and zero elsewhere, so it is tight without assumptions on the endpoint. (6) gives $\mathrm{TV}(r, p) = \beta_x \,\mathrm{TV}(\pi_{\theta,x}, p) \le \epsilon$. Because p has full support and r is its convex mixture with a softmax policy, r has full support. If a rule uniformly guarantees exact non-regression for every $[0, 1]$-valued U, tightness forces $\mathrm{TV}(r, p) = 0$ and hence $r = p$. Finally, $(r - p)^\top U = \beta_x(\pi_{\theta,x} - p)^\top U$, proving the gain-scaling statement.

B.3 Proof of proposition 5

Write $R_{f,x} = V_x(\pi_f)$. Let F_S be the population CDF of the source score in Eq. (14). For n i.i.d. calibration source clusters, $F_S(S_{(k)})$ has a $\mathrm{Beta}(k, n + 1 - k)$ distribution when F_S is continuous; ties make the upper tolerance statement conservative (Wilks, 1941). The selected rank therefore gives

$$\Pr_{\mathrm{cal}}\{F_S(q_{\mathrm{PAC}}) \ge 1 - \gamma\} \ge 1 - \alpha. \tag{18}$$

On this event, a new exchangeable source cluster satisfies $S_c \le q_{\mathrm{PAC}}$ with probability at least $1 - \gamma$. Hence every included member and panel has $R_{f,x} \le q_{\mathrm{PAC}} \,\mathrm{TV}(\pi_{f,x}, \pi_{0,f,x})$. Interpolation scales each endpoint gain by $\beta_{f,x} = \min\{1, \epsilon/[q_{\mathrm{PAC}} \,\mathrm{TV}(\pi_{f,x}, \pi_{0,f,x})]\}$, so its violation is at most ϵ. The maxima in S_c make the statement simultaneous over the fixed policy family, panels, and endpoints. They do not extend it to a changed family or a nonexchangeable source distribution.

For the radial-retention statement, on the event $S_c \le q_{\mathrm{PAC}}$ we have $R_{f,x} \le q_{\mathrm{PAC}} \,\mathrm{TV}(\pi_{f,x}, \pi_{0,f,x})$. A mixture with coefficient β scales its endpoint gains and violation by β. The sufficient condition is therefore $q_{\mathrm{PAC}}\beta \,\mathrm{TV} \le \epsilon$, with $0 \le \beta \le 1$. The largest such coefficient is $\min\{1, \epsilon/(q_{\mathrm{PAC}} \,\mathrm{TV})\}$, proving pointwise optimality among radial mixtures.

B.4 Marginal source-conformal calibration

Fix the policy family independently of the calibration outcomes. Under exchangeability of n calibration source scores and one new source score, the new score's rank is uniform in the absence of ties; assigning tied ranks conservatively yields the same lower bound. Therefore the order statistic at $k = \lceil (n + 1)(1 - \delta) \rceil$ covers the new source with probability at least $1 - \delta$. If $k = n + 1$, the known upper score bound one supplies the threshold. The score maximum makes coverage simultaneous within the specified family and source. This is marginal over calibration and the new source. It differs from the high-confidence population statement of the PAC tolerance rule. Empirical recalibration of policies developed using the same validation outcomes does not itself provide that independence condition.

B.5 Proof of proposition 4

Write $\rho_g = R_g \rho(\cdot \mid x_g)$ in molecule-identity coordinates. Reserializing the input by any fixed $h \in \mathcal{G}_x$ only reindexes the complete set of products gh as g ranges over the group. Their average is therefore unchanged, which proves invariance. For learned and anchor members r_g and p_g, convexity of the ℓ_1 norm gives

$$\frac{1}{2}\left\| |\mathcal{G}_x|^{-1} \sum_g (r_g - p_g) \right\|_1 \le |\mathcal{G}_x|^{-1} \sum_g \frac{1}{2} \|r_g - p_g\|_1. \tag{19}$$

This is the stated total-variation contraction. For any endpoint vector in the same identity coordinates, linearity gives

$$\left\langle |\mathcal{G}_x|^{-1} \sum_g (r_g - p_g), U_{x,d} \right\rangle = |\mathcal{G}_x|^{-1} \sum_g \langle r_g - p_g, U_{x,d} \rangle. \tag{20}$$

The map from the endpoint-gain vector to violation, $z \mapsto \max_d [-z_d]_+$, is convex. Jensen's inequality applied to the order-specific gain vectors proves the violation inequality. If $r_g = q_g^\star$ and every summand is nonnegative by Proposition 1, the average target is feasible. Applying Proposition 3 to the two averaged distributions completes the certified-deployment statement.

Corollary 1 (Finite sampled orbit)**.** *Let $\widehat{\rho}_{x,M}$ average M i.i.d. uniformly sampled permutations. For a K-action panel and fixed policy ρ, with probability at least $1-\delta$,*

$$\|\widehat{\rho}_{x,M} - \bar{\rho}_x\|_1 \le K\sqrt{\frac{\log(2K/\delta)}{2M}}. \tag{21}$$

For learned and anchor policies simultaneously, replace $2K$ by $4K$; then both the sampled-versus-full-orbit endpoint-gain error and the error in their TV distance are at most $K\sqrt{\log(4K/\delta)/(2M)}$.

For each action coordinate, the mapped probability is an i.i.d. $[0,1]$-valued random variable. Hoeffding's inequality and a union bound give the coordinate error; summing over K coordinates gives the first statement. For two policies, another union bound and the triangle inequality give the TV claim. Since the difference between two probability vectors sums to zero, its inner product with any $[0,1]$ utility is bounded by half its ℓ_1 norm, which gives the endpoint-gain claim. The bound is distribution-free and conservative; empirical stabilization in M remains important.

B.6 Dual and exponential form

Let $b = p^\top U \in \mathbb{R}^D$ and write the constraints as $q^\top U - b \ge 0$. For multipliers $\lambda \ge 0$, the Lagrangian of the concave maximization is

$$\mathcal{J}(q,\lambda) = q^\top s - \tau D_{\mathrm{KL}}(q\|p) + \lambda^\top(q^\top U - b). \tag{22}$$

Maximizing over the simplex gives Eq. (5). Substitution yields the convex dual objective

$$g(\lambda) = \tau \log \sum_a p_a \exp\left(\frac{s_a + U_a^\top \lambda}{\tau}\right) - \lambda^\top b, \qquad \nabla g(\lambda) = q_\lambda^\top U - b. \tag{23}$$

Affine constraints and a nonempty relative-interior domain give the standard convex KKT characterization. In degenerate cases where a finite numerical dual solution is difficult to certify, our implementation solves the four-variable primal directly and otherwise returns the feasible anchor.

B.7 Why the common 2D path needs at most one initial multiplier

Let $D = 2$ and $s_a = w_1 U_{a,1} + w_2 U_{a,2}$ with $w_1, w_2 > 0$. The unconstrained Boltzmann target q^0 maximizes $q^\top s - \tau D_{\mathrm{KL}}(q\|p)$. By comparison with p, $(q^0 - p)^\top s \ge \tau D_{\mathrm{KL}}(q^0\|p) \ge 0$. Therefore

$$w_1(q^0 - p)^\top U_{:,1} + w_2(q^0 - p)^\top U_{:,2} \ge 0. \tag{24}$$

Both endpoint gains cannot be negative. If one constraint is violated, its nonnegative multiplier monotonically tilts probability toward higher values of that endpoint; one-dimensional bisection finds the first feasible boundary. The implementation still checks both constraints and uses the complete dual iteration in the rare coupled numerical case. For nonlinear scores such as a Tchebycheff minimum, the weighted-sum contradiction is unavailable and the implementation routes directly to the complete solver.

B.8 Proof of proposition 2

Consider two equally likely contexts, two actions, and one endpoint with utilities $(0,1)$. Use the full-support anchor $(1/2, 1/2)$ in both contexts. Policies $(1/4, 3/4)$ and $(3/4, 1/4)$ produce gains $G_{x_1} = 1/4$ and $G_{x_2} = -1/4$, respectively. Then $\mathbb{E}_x G_x = 0$, so the global constraint is satisfied, while context x_2 regresses. Additional endpoints can be held constant, and the construction can be embedded in a finite action simplex. Hence global endpoint constraints do not imply their contextwise counterparts.

B.9 EXACT SAMPLED-SUPPORT PROBABILITY

For a nonempty set $S \subseteq \mathcal{A}$, the probability that N i.i.d. anchor draws have support exactly S is the probability that all draws lie in S and every element of S appears. Inclusion–exclusion over missing elements gives

$$\Pr(\mathrm{supp}(A_{1:N}) = S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} \left(\sum_{a \in T} p_a \right)^N . \tag{25}$$

Summing this mass over the subsets for which the linear feasibility program $q^\top U \geq p^\top U$, $q \in \Delta_S$ has a solution yields $F_x(S) = 1$ and therefore

$$\Pr(F_x(S_N)) = \sum_{\emptyset \neq S \subseteq \mathcal{A}_x} F_x(S) \sum_{T \subseteq S} (-1)^{|S|-|T|} \left(\sum_{a \in T} \pi_0(a \mid x) \right)^N . \tag{26}$$

We enumerate all 15 nonempty supports because $|\mathcal{A}| = 4$.

C TARGET CONSTRUCTION ALGORITHM AND NUMERICAL VALIDATION

Algorithm 1: contextwise full-support target.
Input: positive anchor $p \in \Delta_{|\mathcal{A}|}$, measured utility $U \in [0,1]^{|\mathcal{A}| \times D}$, mirror step $\eta > 0$, tolerance $\epsilon = 10^{-7}$.
1. Set $s_a = D^{-1} \sum_d U_{a,d}$ and $\tau = 1/\eta$.
2. If $D = 2$, form the unconstrained target $q \propto p \exp(\eta s)$. If feasible, return it. Otherwise find the single initially violated endpoint and bisect its multiplier; verify all constraints.
3. If $D \geq 3$ (or the 2D verification fails), minimize the nonnegative smooth dual with float64 L-BFGS-B and reconstruct q from Eq. (5).
4. If any endpoint is below its anchor expectation by more than 10^{-8}, solve the $|\mathcal{A}|$-variable primal with SLSQP.
5. Validate after the final precision cast. If exponential underflow creates zero support, mix a vanishing amount of an explicit feasible full-support fallback; if feasibility still fails within ϵ, return p.
Output: a normalized full-support target $q^\star$ with certified endpoint inequalities at the stated tolerance.

We solved 1,000 random problems at each of D $\in \{2, 3, 6\}$. At tolerance 10^{-7}, no output violated an endpoint constraint, decreased scalar utility, or decreased the KL-regularized primal objective relative to the anchor. Primal fallback rates were 0%, 0%, and 0.2% for 2D, 3D, and 6D. On 100 independent 2D problems, a direct SLSQP optimum never exceeded the fast solver by more than 10^{-5}; the minimum signed objective gap was -3.8×10^{-8}, consistent with numerical tolerance. The larger-panel validation independently solves 1,000 problems in each of the nine cells |A| $\in \{4, 8, 16\}$ and D $\in \{2, 3, 6\}$ with anchor floor 10^{-8}. Across all 9,000 solves there are zero endpoint, scalar-improvement, or full-support failures. Per-cell median time is 0.06–1.73 ms; the 16-action, 6-objective p95 is 1.13 ms. Rare difficult active sets make the worst p95 164.4 ms (4 actions, 6 objectives) and the maximum below one second.

D DATA CONSTRUCTION AND INDEPENDENCE

Standardization and splits. We parse and sanitize SMILES with RDKit (RDKit contributors, 2026), canonicalize them, compute Bemis–Murcko scaffolds, and assign all members of a scaffold to one split. Candidate retrieval uses 2,048-bit radius-2 Morgan fingerprints and Tanimoto similarity, a maximum heavy-atom difference of ten, and only molecules within the same split. Candidate ordering and repeated panels use stable hashes rather than process-dependent randomness. Every action in an episode must have finite measurements for all requested endpoints.

Utilities. Raw measurements are first oriented so that larger is better; the primary map then applies Eq. (1) with statistics computed once from the Biogen training split. For HLM, the oriented median and robust scale are -1.2306 and 0.8467; for solubility they are 1.5820 and 0.2251. The fixed sigmoid has no finite tail cutoff. On the held-out split its utility range is [.0013, .8988], and no value exceeds .99.

The retrospective panel-relative sensitivity uses

$$U_{a,d} = \frac{\widetilde{y}_{a,d} - \min_b \widetilde{y}_{b,d}}{\max_b \widetilde{y}_{b,d} - \min_b \widetilde{y}_{b,d}}, \tag{27}$$

with a small denominator floor. It is never used for the primary deployment certificate because it requires outcomes from the current panel. We require panel spread to exceed a fixed fraction of the endpoint standard deviation. The hard structured-rationale label is the candidate maximizing the minimum normalized endpoint, with mean utility as a deterministic tie-breaker; the action loss itself uses the full distributional target.

Leakage checks. Biogen has zero exact-SMILES, source-scaffold, and candidate overlap among train/validation/test splits by construction. External all-clean analyses exclude a context if its source or any candidate exactly matches or shares a Murcko scaffold with any Biogen molecule. The TDC strict-local and both OpenADMET analyses additionally require every candidate to have source Tanimoto similarity at least 0.35. These filters enforce the stated structural separation; residual chemical-series and assay-domain relations remain possible.

OpenADMET provenance. We use the OpenADMET/Octant CYP July 20, 2026 release, DOI `https://doi.org/10.57967/hf/9647`, under CC BY 4.0. The 2D task uses compound-level CYP3A4 and CYP2J2 percentage remaining; the 3D task adds quality-control-passing CYP3A4 inhibition pIC50. Uncertainty analysis uses the accompanying well-level reactivity table with four biological replicates per condition. Data never enter the language-model prompt.

E Implementation and hyperparameters

Table 5: Training configuration.

Component	value
Base models	Qwen3-4B-Instruct-2507; Qwen3-8B; Mistral-7B-Instruct-v0.3
Action vocabulary	four single tokens, A–D
Sequence length	1,024
LoRA rank / scale / dropout	16 / 32 / 0.05
LoRA modules	q,k,v,o,gate,up,down projections
Optimizer	fused AdamW; cosine schedule; 3% warmup
Learning rate	10^{-4}
Steps	100 (main pipeline); 200 (CPDPO-R task)
Batch / accumulation	4 / 4 (4B); 2 / 8 (8B)
Precision	BF16 training; TF32 enabled
Mirror step η / temperature τ	2 / 0.5
Target CE weight α_q	1
Endpoint hinge weight α_c	20
Rationale LM weight α_r	0.5 (HLM+Sol/transfer); 0.1 (HLM+MDR1)
Bootstrap	50,000 source-cluster resamples; seed 2026
Assay uncertainty	5,000 draws; seed 2026

For the RLM–solubility CPDPO-R runs, the action weight is 1000, the masked structured reasoning weight is 0.5, and the augmented-context dual has weight 1000, α = 10, dual learning rate 0.1, and cap 1000. Scalar η=4, target CE, and exact PG use the same RLM–solubility seed set and LoRA configuration; only the specified action objective changes. The Qwen3-8B scale comparison uses five fixed seeds (11, 37, 73, 101, 2026); its training wall-clock is about ten minutes per member on one 72-GB

accelerator, and action scoring takes roughly 15–30 seconds per member after model loading.

For the exact-PG baseline, the action loss is

$$-\mathbb{E}_{a\sim\pi_{\theta,x}} s_x(a) + 0.5 D_{\mathrm{KL}}(\pi_{\theta,x} \| \pi_{0,x}), \tag{28}$$

plus the same rationale loss and, in the matched variant, exactly the same weight-20 endpoint hinge in Eq. (9). All action expectations are computed by summing the four logits; there is no Monte Carlo policy-gradient estimator.

F COMPLETE POLICY EFFECTS AND COMPONENT ABLATIONS

Table 6: Complete normalized stochastic-policy results. ΔV is mean endpoint-violation reduction relative to scalar projection. Parentheses give the source-cluster 95% CI for mean gain.

Data / endpoints	model	n	G_1	G_2	G_3	mean gain (95% CI)	ΔV
Biogen HLM+Sol	4B	716	.004784	.002513	–	.003649 (.003059,.004261)	.008180
Biogen HLM+Sol	8B	716	.007194	.003925	–	.005559 (.004779,.006352)	.010915
Biogen HLM+MDR1	4B	916	.005075	.006372	–	.005723 (.005123,.006350)	.010864
TDC clearance, all-clean	8B	1,170	.002079	.000739	–	.001409 (.000876,.001936)	.010058
OpenADMET CYP, 2D clean	8B	2,168	.000692	.001288	–	.000990 (.000579,.001395)	.008654
OpenADMET CYP, 3D clean	8B	938	.000959	.000469	.000677	.000702 (.000237,.001167)	.011416

Table 7: Policy gains in native dataset units, oriented so positive is favorable. HLM, MDR1, solubility, and pIC50 are on their released log or pIC50 scales; CYP remaining is in percentage points; TDC values retain dataset-native clearance units. Bold intervals exclude zero. The last column averages endpoint gains after division by a global endpoint SD.

Panel	raw endpoint 1 (95% CI)	raw endpoint 2 (95% CI)	raw endpoint 3 (95% CI)	standardized mean (95% CI)
Biogen HLM+Sol 4B	**.00647 (.00534,.00759)**	**.00228 (.00158,.00299)**	–	**.00814 (.00681,.00949)**
Biogen HLM+Sol 8B	**.00973 (.00827,.01124)**	**.00354 (.00268,.00441)**	–	**.01237 (.01060,.01422)**
Biogen HLM+MDR1 4B	**.00688 (.00573,.00802)**	**.00892 (.00774,.01013)**	–	**.01208 (.01077,.01343)**
TDC HLM+hepatocyte 8B	**.1460 (.0832,.2090)**	.0537 [−0.0042, 0.1117]	–	**.00202 (.00093,.00311)**
OpenADMET 2D 8B	**.0507 (.0026,.0983)**	**.1121 (.0732,.1510)**	–	**.00191 (.00117,.00264)**
OpenADMET 3D 8B	**.0783 (.0062,.1502)**	.0506 [−0.0045, 0.1054]	**.00134 (.00016,.00254)**	**.00146 (.00051,.00242)**

For the first two columns, endpoint order follows Table 6. OpenADMET-3D's third-endpoint SD is computed across unique candidate pIC50 values because the 3D extension stores this value separately from the 2D delta vector.

Table 8: Standardized raw mean gain for each available paired training member. Both members are positive on every panel; this is a robustness check, not an estimate of a training-seed confidence interval.

Panel	member 1	member 2
Biogen HLM+Sol 4B	.008114	.008157
Biogen HLM+Sol 8B	.012895	.011848
Biogen HLM+MDR1 4B	.012068	.012086
TDC clearance 8B	.002210	.001827
OpenADMET 2D 8B	.001627	.002194
OpenADMET 3D 8B	.001027	.001889

Table 9: Component and matched-objective ablations on 4B Biogen HLM+solubility. Rows share two paired anchors; the blocks isolate the learned hinge and matched loss, respectively.

Comparison	update	mean gain	violation	no-regression
Component	target CE, no learned hinge	.021767	.013535	.52584
Component	target CE + weight-20 hinge	.003908	.005078	.56006
Matched loss	exact PG + KL + weight-20 hinge	.003163	.005810	.52374
Matched loss	target CE + same hinge	.003649	.005239	.55517

G ADDITIONAL ROBUSTNESS RESULTS

G.1 HYPERPARAMETER, OBJECTIVE, AND ADAPTER SENSITIVITY

Tables 10–11 evaluate hyperparameter sensitivity on the Biogen HLM+solubility validation split, with test data excluded from selection. Increasing η generally raises both utility and violation;

increasing the hinge weight reverses that trade-off. Direct PG remains positive across the tested hinge weights. The relative performance of target projection therefore depends on the operating point.

Table 10: Validation sensitivity of target CE and direct exact PG.

Update	sweep	value	mean gain	violation	no-regression
Target CE	mirror step η	.5	.005737	.007584	.4825
		1	.008243	.007551	.5192
		2	.008101	.008318	.4808
		4	.011489	.010539	.4942
		8	.017299	.013897	.4825
Target CE	hinge α_c	5	.024078	.014216	.5109
		8	.016314	.010442	.5359
		10	.015260	.009804	.5259
		12	.014033	.009413	.5125
		15	.009148	.008531	.4825
		20	.008101	.008318	.4808
Exact PG	hinge α_c	5	.004532	.007099	.4725
		10	.003401	.006131	.5225
		20	.002483	.005984	.5125
		50	.003117	.005666	.5459

For the rank sweep, the fixed rank-16 anchor adapter is first merged into the base model and a new projection adapter is trained; evaluation reconstructs that merged base before loading the new adapter. Every rank comparison is therefore independent of residual weights from a fixed-rank adapter. Alternative scalarizations retain positive gain on both endpoints but do not dominate equal weighting.

Table 11: Validation LoRA-rank and scalarization sensitivity. Scalar weights are ordered HLM/solubility; "Tcheby." uses the minimum endpoint utility.

Sweep	setting	mean gain	G_1	G_2	violation
New LoRA rank	8	.004242	–	–	.006120
	16	.007368	–	–	.007220
	32	.010730	–	–	.008108
Scalarization	$(.1, .9)$	.007402	.011396	.003408	.008515
	$(.25, .75)$	.008371	.012529	.004214	.008300
	$(.5, .5)$	.009646	.014046	.005246	.008198
	$(.75, .25)$	.009727	.014234	.005220	.009578
	$(.9, .1)$	.009757	.014525	.004989	.008782
	Tcheby.	.010105	.015242	.004969	.009310

G.2 Low support, larger panels, and missing endpoints

No learned 4/8/16-action context needs an active target action with anchor probability at most .01. As the panel grows, median minimum active support is .233/.117/.058, target–anchor KL is .145/.108/.087, and learned target–policy TV is .221/.166/.146; mean maximum absolute target/anchor log ratio remains .985/1.075/1.142. Divergence remains stable over this range; neural-policy behavior near zero support is untested. Independently, over 9,000 synthetic targets spanning 4/8/16 actions, 2/3/6 objectives, and anchor floors down to 10^{-8}, there are zero feasibility, scalar-improvement, or support failures.

Table 12: End-to-end Biogen HLM+solubility scaling with real measured candidate panels. Four actions use the fixed sigmoid and complete 24-order average; 8/16 actions use panel-relative utilities and $M = 8$. Certified columns use $\epsilon = .005$.

actions	test n	learned gain	violation	no-reg.	certified gain	certified violation
4	716	.003875	.002534	.485	.000509	.000449
8	588	.026668	.004062	.719	.001497	.000264
16	381	.027860	.000865	.890	.001484	.000049

All three learned-gain cluster-bootstrap intervals exclude zero. Because the larger panels use separately constructed training sets, these results establish viability at larger K rather than a controlled scaling law. Their mean certified interpolation fractions are .064 and .057, because more action coordinates increase learned–anchor TV, yet the larger raw gain leaves certified gain above the four-action value.

A separate panel-relative four-endpoint pipeline averages three optimizer seeds on 391 test contexts. Mean gain is .004762 ([.003965, .005575]); endpoint CIs are HLM .007008 ([.005577, .008439]), rat microsomal clearance .006300 ([.004804, .007796]), solubility .002074 ([.000555, .003578]), and MDR1 .003664 ([.002015, .005300]). Mean violation is .012189 and no-regression .222. Its ϵ = .005 certificate retains gain .000654 ([.000546, .000762]), violation .001929, and $\bar{\beta}$ = .244. This supports four measured constraints end to end, while also showing that the direct parameterization gap grows with dimension.

Table 13: Interval-robust targets and end-to-end policies under simulated training missingness. Missing values use $[0, 1]$; tests are fully observed. The 25% learned row averages three optimizer seeds; 10/50% use seed 37.

missing	nontrivial targets	learned test gain (95% CI)	violation	no-reg.
.10	.974	.005426 (.004342, .006522)	.008548	.462
.25	.852	.004099 (.003310, .004891)	.007358	.471
.50	.496	.000893 (.000178, .001622)	.007338	.443

The exact robust targets retain 100% no-regression for every interval completion; their true-data gains decline from .1373 to .0552 as missingness rises (complete-information target: .1525). Learned policies remain significantly positive even at 50% missingness. Relative to the matched fully observed, no-rationale three-seed policy (gain .008912, violation .009616), 25% missingness trades gain for lower violation. Interval robustness supplies the guarantee without imputing missing outcomes.

G.3 Learned global and pointwise constrained baselines

The shared global augmented-dual baseline attains validation endpoint gains .05883/.02378 but context violation .02156, confirming that positive average constraints do not protect each molecule. For the pointwise augmented-dual baseline, three independently trained test members obtain mean gain .014261 and violation .011486. Target CE at a validation-matched hinge obtains .016444 and .012224: paired differences favor target CE in utility by .002182 ([.001579, .002774]) but favor the pointwise method in violation by .000738 ([.000252, .001222]), and the no-regression-rate CI crosses zero. The trade-off is Pareto: target CE raises utility, whereas the pointwise method lowers violation.

G.4 Tuning frontiers and a structural non-LLM control

Table 14 reports test points chosen through validation sweeps over steps, learning rates, hinge weights, and trust strengths. The sweeps contain 5 exact-PG, 7 target-CE, and 6 pointwise-dual configurations. Exact PG occupies the lowest-violation region, target CE the middle, and the pointwise dual a higher-gain, higher-violation region. No method dominates the full frontier.

Table 14: Representative points from Biogen test frontiers. Each listed gain has a 95% cluster-bootstrap interval above zero.

method	setting	gain	violation	no-reg.
Exact PG+hinge	50 steps	.004336	.005454	.553
Exact PG+hinge	200 steps	.004298	.005754	.543
Target CE+hinge	lr 3×10^{-5}	.006650	.007758	.478
Target CE+hinge	50 steps	.010328	.009131	.483
Target CE+hinge	200 steps	.013929	.012002	.503
Pointwise dual	multiplier cap 400	.006391	.008350	.451
Pointwise dual	50 steps	.013766	.009985	.494
Pointwise dual	weight 100, 200 steps	.018304	.016577	.468

An explicitly order-equivariant Morgan-fingerprint panel MLP controls for access to source/candidate structures without an LLM representation. Its validation-selected conservative point has test gain .016853 ([.012232, .021512]), violation .028287 ([.025210, .031511]), and no-regression .399. It therefore occupies a high-utility/high-risk region rather than matching the LLM's low-violation frontier on the present scaffold split.

A stronger pretrained control uses fixed ChemBERTa embeddings followed by a two-layer, position-free Set Transformer. The validation-selected anchor-residual configuration uses 300 steps and

endpoint-hinge weight 50. Training the set head takes 11–12 seconds after embedding extraction, compared with approximately 230 seconds for a 100-step Qwen-4B adapter on the same accelerator. Across three training seeds under the fixed sigmoid, its mean gain is .002178 ([.001822, .002548]) with endpoint gains .002078 and .002278; mean violation is .001710 ([.001318, .002105]), lower than the LLM policy by .001941 ([.001534, .002339]). Its ϵ = .005 certified gain is .000487 ([.000414, .000560]). The LLM thus provides more raw gain, while the pretrained set encoder provides lower violation and slightly more certified gain.

G.5 RLM–SOLUBILITY RL BENCHMARK

Table 15 gives the complete three-seed Qwen-4B comparison on the fixed-scale RLM-clint-decrease plus solubility-increase task. Raw columns use the learned action distribution; controlled columns apply the same outcome-free TV interpolation to every row. The shifted hypervolume is the context average of $\prod_d [G_{x,d} + 10^{-3}]_+$, with the shift fixed before test scoring. The Qwen-4B result is retained as a separate scale-specific replication.

Table 15: Fixed-scale RLM–solubility test results. Values are mean gains relative to the Qwen HLM anchor; intervals are reported for the raw and controlled aggregate gains in the main text.

Policy	raw mean	raw RLM	raw Sol.	raw V_x	$\epsilon = .005$ mean	controlled V_x
Scalar $\eta = 4$	.013935	.018997	.008873	.012209	.000435	.000522
Target CE	.003442	.004367	.002517	.004740	.000300	.000585
Exact PG	.003304	.003817	.002790	.004316	.000302	.000571
CPDPO-R	**.019132**	**.026805**	**.011459**	.021247	.000431	.000566

The RLM–solubility task contains 1,696/265/295 train/validation/test contexts and no cross-split scaffold overlap. CPDPO-R uses the exact action expectation in Eq. (29), a per-context augmented dual, and the masked structured-output loss. On the corresponding validation split its raw mean gain is .03099 [.02523, .03747] and its paired advantage over scalar is .01057 [.00637, .01560]; these validation results determined the configuration evaluated on test. A task-specific Mistral family did not reproduce this raw advantage on validation (CPDPO-R .00753 [−.00701, .02293] versus scalar .01336 [.01022, .01656]), so the raw-utility advantage is not reproduced in the task-specific-anchor family. A separate prespecified stability grid on the original Mistral anchors restored a positive raw paired advantage at lower learning rate (weight 100, 3×10^{-5}: .00862 [.00459, .01271]), but its paired raw violation increase was .01105 [.00749, .01472] and its controlled advantage crossed zero ([−.000163, .000049]). These settings were evaluated on validation only. A subsequent fixed-anchor Mistral study and its test-time entropy ensemble are reported in Appendix I.

Table 16: Five-seed Qwen3-8B RLM–solubility validation and test.

	Validation, controlled				Test, controlled			
Policy	mean	RLM	Sol.	violation	mean	RLM	Sol.	violation
Scalar $\eta = 4$	.000604	.000919	.000289	.000436	.000366	.000440	.000293	.000520
CPDPO-R	**.000701**	**.001004**	**.000399**	.000406	**.000440**	**.000488**	**.000391**	.000530

The five-seed Qwen3-8B validation paired mean difference is .0000978 ([.0000124, .0001793]); the test paired difference is .0000731 ([.0000065, .0001403]). Test endpoint differences are .0000490 ([−.0000493, .0001526]) for RLM and .0000972 ([−.0000028, .0002026]) for solubility. The aggregate difference is positive, violation does not significantly increase, and the hypervolume difference has a positive point estimate with a CI that narrowly crosses zero.

Eight-billion dual and support diagnostics. For the five Qwen3-8B CPDPO-R members, the final-minibatch context-dual active fractions were .75, .25, .25, .50, .50 for seeds (11, 37, 73, 101, 2026); the corresponding mean multipliers were 2.16×10^{-5}, 7.55×10^{-6}, 1.00×10^{-5}, 2.80×10^{-5}, 2.11×10^{-5}, and the largest observed final-step multiplier was 9.95×10^{-5}. The cap of 1000 was never approached and the constraint loss was below 4×10^{-8} at the final step.

The final-minibatch summaries describe the last optimization step, rather than constraint activity over the full training trajectory. The full-support anchor had no near-zero active target action in the measured panel range; synthetic solver tests down to an anchor floor of 10^{-8} remained feasible.

Fingerprint-nearest-neighbor stress test. To complement Murcko-scaffold splitting, we computed each test context's maximum Morgan-2048 Tanimoto similarity (source and candidates) to any of the 1,230 molecules in the training split. The 295 Qwen3-8B test contexts had nearest-similarity quantiles (.432, .536, .702); 2, 95, and 198 contexts fell in [.20, .35), [.35, .50), and [.50, 1), respectively. Qwen3-8B raw CPDPO-R-minus-scalar gains were .00436, .01134, .00858 in these bins; after TV=.005 control they were .000158, .000169, .000026, respectively. The high-similarity bin is therefore close to parity under the certificate, and the performance margin varies with the evaluated similarity stratum.

For an explicit amortization diagnostic, the Qwen3-8B validation target has mean target–anchor KL .0315 (median .0271, 95th percentile .0783) and mean target gain .0326, whereas the seed-37 learned policy has mean gain .0301 and mean violation .0202; target–policy TV averages .206. Across contexts, target–policy TV correlates with learned violation (ρ = .519, Spearman p < 10^{-12}), while minimum anchor probability has no detectable relation to TV (ρ = .009). These associations support interpreting the gap in terms of neural projection fidelity rather than tiny-support failure on the measured panels.

The direct action objective used in the implementation is

$$\mathcal{L}_{\text{CPDPO-R}} = w_q\,\mathbb{E}_x[-G_x(\widetilde{\pi_{\theta,x}})] + w_c\,\mathbb{E}_{x,d}\left[\alpha\left[v_{x,d} + \frac{\lambda_{x,d}}{2\alpha}\right]_+^2 - \frac{\lambda_{x,d}^2}{4\alpha}\right] + w_r\,\mathcal{L}_{\text{LM}}, \qquad (29)$$

where $G_x(p) = \sum_a [p(a \mid x) - \pi_0(a \mid x)]s_x(a)$, $v_{x,d} = \mathbb{E}_{\pi_{0,x}} U_{x,d} - \mathbb{E}_{\widetilde{\pi_{\theta,x}}} U_{x,d}$, and $\widetilde{\pi_{\theta,x}}$ is the TV-controlled policy. After each training loss evaluation we project $\lambda_{x,d}$ to $[0, 1000]$ using the constraint supergradient; $\alpha = 10$, $w_q = 1000$, $w_c = 1000$, and $w_r = 0.5$ on this task. Thus the RL term uses every measured action directly, while the auxiliary loss excludes the choice token.

G.6 Normalization, rationale, and divergence analyses

Table 17 places the primary fixed-sigmoid model beside earlier re-evaluations of a common learned/scalar pair. Target feasibility failures are zero for every transform. Absolute magnitudes vary with scale, but every bounded transformation retains positive gain.

Table 17: Biogen test re-evaluation under alternative endpoint scales. The certificate column is shown only for bounded transforms.

normalization	learned gain (95% CI)	learned viol.	scalar viol.	certified gain
training-only sigmoid	.003878 (.003006, .004850)	.003651	.004925	.000393
panel min–max	.003649 (.003054, .004264)	.005239	.013418	.000654
training quantile	.002798 (.002346, .003252)	.003062	.007347	.000501
robust training min–max	.002102 (.001769, .002444)	.002207	.004865	.000379
training z-score	.008196 (.006859, .009545)	.009194	.020628	–

Training new policies under the fixed sigmoid, quantile, robust-min–max, and z-score utilities yields positive test gains (.00388, .00937, .00768, and .03031) with violations .00365, .00559, .00383, and .01737. Z-scores are unbounded, so Proposition 3 requires an explicit assay range and is therefore not applied to this unbounded transform.

Across three seeds, removing the structured-rationale loss (α_r = 0) gives gain .008912 ([.007616, .010221]), violation .009616, and no-regression .4697. With α_r = .5 these are .010609 ([.009207, .012061]), .009445, and .4967. The paired rationale contribution to gain is .001697 ([.001165, .002236]) and to no-regression is .0270 ([.00749, .04675]), with no significant violation change. The rationale channel provides useful auxiliary supervision, but this benefit does not establish a faithful trace of the decision process.

A Pearson-χ^2-regularized target remains positive (gain .009912, violation .009617) but worsens target–policy KL/TV to .192/.229 versus .141/.221 for reverse KL. The better projection fidelity supports the reverse-KL default.

G.7 ANCHOR AND OPTIMIZER SEEDS

On Biogen HLM+MDR1, all six combinations of two anchor seeds and three optimizer seeds have positive validation and test expected mean gains. A three-member 4B ensemble yields test mean gain 0.004277 (95% CI [0.003701, 0.004869]) and mean violation 0.004713 ([0.004311, 0.005117]). The main table reports the selected single-policy result; this analysis quantifies seed variation.

A separate anchor-strength experiment starts directly from the unadapted Qwen model. Across three optimizer seeds, mean gain is .001289 with interval [−.003530, .006224], and mean violation rises to .032586. The ϵ = .005 interpolation retains a small positive gain of .000137 ([.000036, .000238]). Reliable raw improvement therefore requires a task-trained anchor, while the distribution-free certificate remains valid for either anchor.

G.8 INDEPENDENT MISTRAL MODEL FAMILY

To test architecture transfer, the Mistral study used three paired anchor/projection seeds, the same training-only sigmoid as Qwen, and an action projection loss without the structured-language term. We compared the four configurations below on validation. The fixed rule selected the row with the largest ϵ = .005 certified gain, breaking differences below 10^{-4} by lower violation; only that configuration was scored on test.

Mistral fixed-scale validation results.

configuration	steps	mean gain	violation	certified gain
target only, hinge 0	100	.008249	.004891	.000539
target only, hinge 20	100	.005499	.004688	.000416
target only, hinge 100	100	.004943	.005146	.000403
target only, hinge 20	200	.007534	.004419	.000559

The selected 200-step configuration produced direct mean, HLM, and solubility gains of .006621, .006077, and .007166, with intervals [.005524, .007752], [.004685, .007417], and [.005640, .008746]. Compared with the matched scalar target, mean gain was indistinguishable while violation fell from .005982 to .005088 (paired reduction CI [.000006, .001924]). Complete orbit averaging raised gain to .006953 ([.005858, .008087]), lowered violation to .004206, and retained certified gain .000558 ([.000486, .000631]). Zero-shot aggregate and every endpoint interval were positive on TDC and both OpenADMET panels. Omitting the language loss isolates action-policy transfer from rationale generation.

G.9 EXTERNAL FIXED-SCALE DEPLOYMENT SIMULATION

We partitioned complete external scaffold groups by a stable hash before examining policy effects: 30% calibrated the endpoint sigmoid and 70% were eligible for evaluation. A panel was retained only if its source and every candidate belonged to evaluation scaffolds, and calibration and retained evaluation scaffolds have empty intersection. Acyclic molecules use canonical SMILES as their split key. No model was retrained or rescored; the analysis reuses the same Mistral predictions.

Table 18: External fixed-scale deployment simulation. "Endpoint +" counts endpoint-gain intervals above zero. Certificate columns use $\epsilon = .005$.

panel	test n	calibration molecules	mean gain (95% CI)	endpoint +	certified gain (95% CI)
TDC	203	184	.003425 [.001450, .005373]	2 / 2	.000217 [.000099, .000328]
OpenADMET 2D	430	283	.002790 [.001386, .004266]	1 / 2	.000150 [.000073, .000228]
OpenADMET 3D	233	202†	.000989 [−.000647, .002531]	1 / 3	.000044 [−.000050, .000137]

†OpenADMET-3D has 233 calibration molecules, of which 202 have complete three-endpoint values; only complete vectors determine the map. Retained evaluation counts are 203/430/233 from 1,170/2,168/938 original panels. Within-molecule raw reconstruction ranges are below 3×10^{-14}, and the shared-scaffold count is zero for every dataset. TDC has both endpoint intervals above zero. OpenADMET-2D has positive CYP2J2 but inconclusive CYP3A4; OpenADMET-3D remains aggregate-inconclusive.

We next varied the scaffold hash while leaving the three paired policies and all statistical choices unchanged. Table 19 summarizes the 20 partitions. A positive count means that the per-partition joint training-seed/source-cluster 95% interval was above zero. None of the 60 aggregate intervals was significantly negative.

Table 19: External fixed-scale sensitivity across 20 predeclared scaffold partitions. Point entries are median [minimum, maximum] across overlapping partitions; CI+ is the number of per-partition 95% intervals above zero. The certificate uses $\epsilon = .005$.

panel	retained n	mean gain	gain CI+	certified gain	cert. CI+	endpoint CI+
TDC	254.5 [123, 403]	.003257 [.002095, .005097]	18/20	.000205 [.000124, .000303]	18/20	19/20, 15/20
OpenADMET 2D	549 [346, 753]	.002544 [.000569, .004437]	18/20	.000150 [.000031, .000269]	18/20	19/20, 3/20
OpenADMET 3D	264.5 [143, 404]	.001535 [−.000835, .002927]	14/20	.000112 [−.000057, .000217]	14/20	14/20, 1/20, 6/20

Endpoint counts follow the dataset order in Table 1; for OpenADMET-3D the order is CYP2J2 remaining, CYP3A4 remaining, and CYP3A4 pIC50. TDC and OpenADMET-2D have positive aggregate and controlled intervals in 18/20 partitions; OpenADMET-3D reaches 14/20. The overlapping partitions characterize split sensitivity on the same observations and cannot be pooled as independent tests.

Table 20: All external split point estimates. Each dataset block reports retained panels / aggregate gain / $\epsilon = .005$ certified gain.

seed	TDC			OpenADMET 2D			OpenADMET 3D		
	n	gain	cert.	n	gain	cert.	n	gain	cert.
20260830	203	.003425	.000217	430	.002790	.000150	233	.000989	.000044
20260831	188	.002132	.000124	540	.003913	.000269	231	.001597	.000121
20260832	193	.005097	.000291	541	.001742	.000119	247	.001353	.000122
20260833	249	.003606	.000239	467	.003119	.000189	216	.002927	.000217
20260834	284	.002864	.000181	449	.002887	.000189	244	.002346	.000159
20260835	354	.004310	.000303	528	.002556	.000168	234	.001631	.000133
20260836	275	.002329	.000161	612	.000933	.000083	264	.001116	.000088
20260837	175	.003221	.000197	698	.002737	.000179	374	.001425	.000105
20260838	290	.004013	.000258	517	.001871	.000130	293	-.000026	.000024
20260839	262	.003143	.000199	346	.000569	.000031	199	.001207	.000074
20260840	254	.002485	.000145	546	.004437	.000263	265	.002023	.000113
20260841	293	.003293	.000204	704	.002848	.000198	404	.001490	.000112
20260842	136	.002095	.000132	619	.001937	.000145	282	.001283	.000110
20260843	383	.004093	.000266	697	.004004	.000249	323	.001845	.000141
20260844	210	.003672	.000225	638	.002062	.000121	304	.000581	.000040
20260845	255	.002491	.000166	552	.002162	.000121	264	.002403	.000165
20260846	266	.003222	.000206	656	.001870	.000110	335	.001580	.000110
20260847	403	.004122	.000262	604	.002817	.000190	276	.001781	.000159
20260848	123	.004795	.000285	435	.002171	.000120	143	-.000835	-.000057
20260849	190	.002334	.000127	753	.002532	.000151	385	.002237	.000141

G.10 Locality, greedy decisions, and replicate uncertainty

Table 21: Panel dependence and molecule-disjoint sensitivity. Candidate reuse reports the 95th percentile and maximum episode count. Disjoint n is the median across 200 maximal greedy subsets. Biogen HLM+Sol uses the fixed sigmoid; the remaining rows use their retrospective panel-relative scales.

panel	episodes	sources	candidates	cand. reuse	graph comp.	disjoint n	disjoint gain CI
Biogen HLM+Sol	716	241	255	27 / 40	1	30	[.001109, .007952]
Biogen HLM+MDR1	916	307	323	26.9 / 51	1	37	[.003128, .008160]
TDC	1,170	481	543	21 / 46	1	62	[−.000406, .002703]
OpenADMET 2D	2,168	764	876	19 / 30	7	107	[−.000163, .002655]
OpenADMET 3D	938	448	610	14.6 / 26	9	80	[.000389, .002845]

The strict-local TDC subset (every candidate Tanimoto $\geq$ .35) retains mean gain .002528 (95% CI [.001720, .003310]) and scalar-violation reduction .010769 ([.008772, .012859]); its second-endpoint CI crosses zero. In contrast, OpenADMET-3D argmax mean gain is $-.002293$ with a CI crossing zero, so the 3D objective is distributional rather than greedy. Across 5,000 draws from replicate-derived OpenADMET-2D measurement distributions, endpoint gains lie in [.000509, .000828] and [.001054, .001612], with all draws positive. Mean violation is .006009 ([.005849, .006166]) for DeltaADMET-R1 versus .015148 ([.014695, .015590]) for scalar projection, an approximately 60% reduction. This analysis propagates the measured assay repeatability.

G.11 Global versus contextwise targets

The global solver shares the same anchor, utilities, scalar objective, mirror step, and exact exponential family as the contextwise solver; the only change is replacing each x-specific endpoint constraint by its dataset average. All reported global averages satisfy every endpoint inequality to 10^{-9}, and all contextwise targets satisfy every per-context inequality to 10^{-7}. This isolates constraint granularity while holding reward and parameterization fixed.

Table 22: Exact target analysis: global-average versus contextwise constraints. The witness rate is harmed by the global optimum but protected by the contextwise optimum; contextwise target harm is zero within 10^{-7}.

Target utilities	witness rate	gain, global	gain, contextwise
Biogen HLM+Sol (2D)	3.1425%	.156094	.156082
Biogen HLM+MDR1 (2D)	6.9869%	.118227	.118202
OpenADMET CYP (2D)	3.9899%	.172757	.172743
OpenADMET CYP+inhibition (3D)	44.9893%	.104013	.099789

Table 23: Exact probability that N anchor draws contain a support on which all endpoint constraints are feasible.

Panel	$N = 2$	$N = 4$	$N = 8$	$N = 16$
OpenADMET 2D	.4622	.7166	.9053	.9842
OpenADMET 3D	.2747	.5511	.8336	.9735

G.12 Parameterization gap and certificate analysis

Let $q^\star$ be an exact target and π_θ its neural projection. For any endpoint in $[0, 1]$,

$$|\mathbb{E}_{\pi_\theta} U_d - \mathbb{E}_{q^\star} U_d| \leq \mathrm{TV}(\pi_\theta, q^\star) \leq \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(q^\star \| \pi_\theta)}. \tag{30}$$

The sufficient certificate $\mathbb{E}_{q^\star} U_d - \mathbb{E}_{\pi_0} U_d \geq \mathrm{TV}$ is conservative. Under the fixed sigmoid, the diagnostic test-time oracle target has mean gain .03435, whereas the fitted 4B policy retains .00388. Mean target-to-policy KL and TV are .0315 and .0984; the TV 95th percentile is .1808.

Target-to-policy TV correlates with learned violation (ρ = .252, p < 10^{-11}). Target slack alone therefore does not certify an unconstrained neural approximation. Proposition 3 instead compares a deployable interpolation directly with the anchor and is tight without endpoint observations.

Table 24: Certificate curve on 4B Biogen HLM+solubility. Every context has worst-case endpoint regression at most ϵ.

ϵ	realized mean gain	realized violation	mean β
.0025	.000327	.000490	.299
.0050	.000654	.000981	.375
.0100	.001308	.001962	.527
.0200	.002564	.003889	.820
.0500	.003629	.005227	.999

G.13 CANDIDATE-ORDER ANALYSIS AND ORBIT COMPUTE CURVE

Across all 24 orders of each four-candidate Biogen test panel, the fixed anchor chooses the same molecule only 40.5% of the time and the learned policy only 41.6%; their mean identity-aligned TV from the identity-order prediction is .0978 and .1005. Candidate serialization therefore changes the predicted distributions substantially. Full orbit averaging is exactly invariant by construction, and the sampled-orbit curve in Table 25 shows that most of the violation and certificate benefit appears before exact enumeration.

Table 25: Four-action Biogen order averaging for one trained member. $M = 24$ enumerates the complete orbit; the certificate uses $\epsilon = .005$.

orders M	learned gain	learned violation	certified gain	mean β
1	.003421	.005501	.000632	.382
4	.003581	.003153	.001046	.421
8	.003583	.002735	.001203	.463
24	.003629	.002314	.001308	.499

Table 26: Sampled-orbit curves for real 8- and 16-candidate panels. Each entry is learned gain / violation / $\epsilon = .005$ certified gain, averaged over two training members.

M	$K = 8$	$K = 16$
1	.026473/.005303/.001352	.028151/.001279/.001335
2	.026309/.004379/.001426	.027746/.001068/.001404
4	.026733/.004118/.001483	.027804/.000917/.001459
8	.026668/.004062/.001497	.027860/.000865/.001484
16	.026698/.003954/.001508	.027835/.000752/.001498
32	.026784/.003902/.001514	.027763/.000764/.001499
64	.026794/.003853/.001517	.027714/.000758/.001504

Table 27: Probability, plurality-PSC, and exact-Kemeny aggregation over 24 orders. Entries are learned gain / violation / $\epsilon = .005$ certified gain for one matched member.

panel	probability mean	plurality PSC	exact Kemeny
Biogen HLM+Sol 4B	.003629/.002314/.001308	.018518/.024778/.000773	.030462/.042716/.000152
Biogen HLM+Sol 8B	.004633/.002923/.001250	.014167/.021160/.000643	.005059/.040730/.000025
Biogen HLM+MDR1	.005895/.002860/.001495	.029545/.028685/.001025	.033067/.058292/.000165
TDC	.001665/.003184/.000613	.009169/.020584/.000398	.014564/.026848/.000073
OpenADMET 2D	.000692/.003343/.000344	.002391/.021515/.000114	.005761/.048841/.000029
OpenADMET 3D	.000617/.004289/.000282	.002154/.032614/.000133	.008527/.080548/.000043

Table 28: Direct versus complete 24-order deployment on all primary panels. Entries are learned gain / violation; certified gains use $\epsilon = .005$. Every row averages the same two matched members before source-cluster resampling.

panel	direct gain / viol.	sym. gain / viol.	direct cert. gain	sym. cert. gain	sym. $\bar{\beta}$
Biogen HLM+Sol 4B	.003603 / .005169	.003827 / .002290	.000660	.001350	.492
Biogen HLM+Sol Mistral 7B	.006621 / .005088	.006953 / .004206	.000477	.000558	.111
Biogen HLM+Sol 8B	.005559 / .005295	.005011 / .003062	.000888	.001278	.390
Biogen HLM+MDR1 4B	.005723 / .005318	.005844 / .002668	.000911	.001554	.386
TDC clearance 8B	.001409 / .006173	.001460 / .002987	.000229	.000533	.559
OpenADMET 2D 8B	.000990 / .007312	.000968 / .004102	.000183	.000366	.520
OpenADMET 3D 8B	.000702 / .009579	.000775 / .004926	.000137	.000293	.557

Every symmetrized learned-gain and certified-gain source-cluster interval is above zero. Compared with the direct matched rows, mean violation falls 42–58% and certified gain rises 44–132%. Symmetrization need not dominate every single ordering in utility; Proposition 4 preserves the order average and bounds violation by its order average.

Two-order training augmentation reduces learned identity TV to .0659 and raises choice consistency to .637, but produces an aggressive policy (full-orbit gain about .020 with violation .013). We therefore use inference-time orbit averaging of the selected policy rather than treating augmentation as a uniform reduction in endpoint regression. For K > 4, exact K! enumeration is unnecessary for the theorem's Monte Carlo approximation: report M, use the same sampled orders for learned and anchor policies, use Corollary 1 as a conservative error bound, and increase M until empirical TV and endpoint estimates stabilize.

In the batched 716-panel benchmark (batch size 128), scoring all 17,184 ordered prompts takes 594/597 seconds for the 4B learned/anchor models and 942/941 seconds for 8B, approximately 1.66 and 2.63 accelerator-seconds per original panel for both distributions at M = 24; M = 8 incurs about one third of that throughput cost. These are hardware- and batching-specific throughput measurements, not single-request latency guarantees.

G.14 PAC FAMILY-UNIFORM CERTIFICATE

We hashed the 203 Biogen validation sources before calibration: 95 sources formed the calibration set and 108 formed a confirmation set. For each calibration source, Eq. (14) takes the maximum score across all panels, endpoints, three Qwen members, and three Mistral members. The validation and test sets share zero source SMILES. The 90%-coverage, 95%-confidence tolerance rule selected rank 91/95, q_{PAC} = .636539, and achieved confidence 96.66%. Confirmation coverage was 104/108 for Qwen and 106/108 for Mistral. Confirmation gains were .000744 ([.000564, .000926]) and .000856 ([.000660, .001060]); their paired improvements over q = 1 were [.000205, .000337] and [.000240, .000385].

An independent HLM+MDR1 task used training-only sigmoids from 1,711 unique training molecules, with no validation/test utility saturation below .01 or above .99. Its 166/174 calibration/confirmation sources gave q_{PAC} = .406812 and 97.40% tolerance confidence. Confirmation coverage was 169/174 (97.13%); gain was .000807 ([.000666, .000953]), and the paired improvement interval was [.000395, .000566]. The unchanged Mistral configuration produced direct test gain .005391 ([.004568, .006204]), HLM/MDR1 gains .004891/.005890, violation .002740, and deterministic gain .000490 ([.000444, .000534]). Its validation-only PAC calibration gave q_{PAC} = .402289, 97.40% tolerance confidence, and 163/174 (93.68%) confirmation coverage.

Table 29: Fixed PAC rules on Biogen test. Coverage requires every panel, endpoint, and member for a source to have violation at most .005. Binomial intervals describe the realized test coverage, not the PAC confidence.

task / family	covered sources	coverage (95% CI)	uncovered	q95 source V	max source V
HLM+Sol / Qwen-4B	234 / 241	97.10% [94.11, 98.82]	7	.004594	.006742
HLM+Sol / Mistral-7B	238 / 241	98.76% [96.41, 99.74]	3	.004547	.005559
HLM+MDR1 / Qwen-4B	300 / 307	97.72% [95.36, 99.08]	7	.004587	.005774
HLM+MDR1 / Mistral-7B	288 / 307	93.81% [90.50, 96.23]	19	.005157	.006277

Table 30: Efficiency of the PAC and deterministic TV rules. Endpoint columns are PAC gains; "secondary" is solubility or MDR1. All intervals jointly resample training members and sources.

task / family	$q = 1$ gain	PAC gain (95% CI)	HLM gain (95% CI)	secondary gain (95% CI)	PAC–$q = 1$ (95% CI)
HLM+Sol / Qwen	.000393	.000618 [.000488, .000746]	.000641 [.000518, .000767]	.000595 [.000380, .000801]	.000225 [.000177, .000271]
HLM+Sol / Mistral	.000477	.000750 [.000637, .000862]	.000682 [.000545, .000813]	.000818 [.000644, .000995]	.000273 [.000232, .000313]
HLM+MDR1 / Qwen	.000347	.000852 [.000759, .000947]	.000745 [.000605, .000884]	.000959 [.000824, .001096]	.000505 [.000450, .000562]
HLM+MDR1 / Mistral	.000490	.001216 [.001102, .001327]	.001125 [.000900, .001361]	.001308 [.001098, .001512]	.000727 [.000658, .000793]

Mean $\bar{\beta}$ for $q = 1$/PAC was .178/.256, .103/.161, .330/.552, and .135/.329 across the four rows; corresponding PAC mean violations were .000765, .000715, .000739, and .000817.

An earlier marginal-quantile analysis of these same policies yielded 90% nominal-coverage test rates of 89.21% for Qwen and 83.82% for Mistral. The PAC-tolerance results reuse those test observations after revising the calibration protocol, so they characterize empirical retention and coverage on Biogen sources rather than a new independent prospective coverage trial; no source-calibrated guarantee is evaluated on TDC or OpenADMET.

Joint source-conformal analysis on the Qwen3-8B RLM–solubility task. The rank rule is the marginal conformal rank $\lceil (n+1)(1-\delta) \rceil$ without the PAC tolerance confidence parameter. This post hoc deployment analysis reuses evaluated policies; prospective coverage requires calibration independent of policy selection. We calibrated one q_{SC} over all five CPDPO-R and five scalar members on the 168 validation source clusters, then applied it to the 197 test source clusters. At nominal 95% coverage (δ = .05), q_{SC} = .571858, 193 sources were covered, and empirical coverage was 97.97%. CPDPO-R retained .000769 mean gain [.000588, .000947] versus .000641 [.000472, .000807] for scalar; the paired difference was .000128 [.000013, .000243]. At nominal 90% coverage, q_{SC} = .477972, 187 sources were covered, and the paired difference was .000153 [.000015,

.000291]. The same joint rule on the Mistral fixed-anchor family did not improve over scalar, so the calibrated gain is reported as Qwen3-8B-specific.

Table 31: Joint source-conformal versus deterministic TV control on the Qwen3-8B RLM– solubility test. One q_{SC} is calibrated jointly over both fixed method families; intervals resample source clusters and members.

nominal coverage	q_{SC}	CPDPO covered	CPDPO gain	scalar gain	paired CPDPO–scalar
90%	.477972	187 / 197	.000920 [.000703, .001133]	.000767 [.000565, .000965]	.000153 [.000015, .000291]
95%	.571858	193 / 197	.000769 [.000588, .000947]	.000641 [.000472, .000807]	.000128 [.000013, .000243]

Source-conformal split stability. To test sensitivity to which validation sources calibrate the threshold, we repeated a 50/50 source-hash calibration/confirmation split for 20 fixed hash seeds, with the shared q always calibrated over both method families. The Qwen paired test gain was positive in all 20 splits at both nominal coverage levels, whereas Mistral's was negative in every split. The splits reuse trained policies and test observations. When every policy is contracted, the shared positive coefficient rescales the same difference; sign stability therefore measures calibration sensitivity, not independent replication.

Table 32: Stability of joint source-conformal calibration over 20 fixed validation source splits. Bracketed entries are the 5th, 50th, and 95th percentiles over splits; paired gain is CPDPO-R minus scalar.

family	nominal	q_{SC}	confirmation coverage	test paired gain	positive splits
Qwen3-8B	90%	[.4186,.4705,.4978]	[.878,.921,.976]	[.000147,.000155,.000175]	20/20
Qwen3-8B	95%	[.4931,.5719,.6001]	[.903,.958,.989]	[.000122,.000128,.000148]	20/20
Mistral-7B	90%	[.5607,.6002,.6081]	[.848,.933,.966]	[−.000354,−.000331,−.000327]	0/20
Mistral-7B	95%	[.6170,.6489,.6643]	[.933,.973,.989]	[−.000322,−.000306,−.000299]	0/20

An initial four-endpoint fixed-sigmoid study evaluated policy re-scaling, scale-matched retraining, dimension-normalized mean hinge, and max-endpoint hinge. Aggregate, HLM, RLM, and MDR1 intervals were positive, but solubility lower bounds ranged from −.000370 to −.000056; those variants were evaluated only on validation. The retraining variants used two-endpoint completions, whereas the policy-iteration study in Appendix I provides all four endpoints. These results therefore do not establish an intrinsic failure of four-dimensional optimization.

G.15 Violation tails and conservative certificate baselines

Mean violation can hide rare failures, so Table 33 reports the upper tail for every primary test. Unconstrained policies have nonzero tails; only the controlled rows carry the stated ϵ bound. The ϵ = .005 policy stays at or below .005 up to floating-point tolerance, whereas the scalar projection has substantially heavier tails.

Table 33: Contextwise violation tails: 95th percentile / maximum.

panel	learned	scalar projection	certified $\epsilon = .005$
Biogen HLM+Sol 4B	.01699 / .02791	.05793 / .10888	.00321 / .00448
Biogen HLM+Sol Mistral 7B	.01921 / .04675	.02218 / .05584	.00153 / .00291
Biogen HLM+Sol 8B	.01776 / .03810	.06810 / .14072	.00316 / .00500
Biogen HLM+MDR1 4B	.01794 / .03130	.07333 / .16193	.00317 / .00500
TDC clearance 8B	.01892 / .03335	.05811 / .12402	.00353 / .00495
OpenADMET 2D 8B	.02174 / .05472	.05165 / .11717	.00362 / .00490
OpenADMET 3D 8B	.02234 / .04310	.05589 / .11482	.00402 / .00500

On the 4B Biogen test, the adaptive TV rule retains gain .000654 at violation .000981 and $\bar{\beta}$ = .375; fixed-β and Pinsker variants retain .000294 and .000568. Outcome-aware safe-ray and oracle projections reach .004682/.001992 and .006424/zero violation, quantifying the opportunity for future uncertainty-aware certificates.

On the retrospective Biogen scale, ϵ = .005 corresponds to median HLM and solubility tolerances .00608 and .00353 in dataset-native log units, while the OpenADMET replicate SEM is larger. Applications should report the full decision-tolerance curve and predeclare ϵ from assay repeatability.

H STRUCTURED-GENERATION METRICS

Each model emits at most 128 tokens of compact JSON: a choice, up to two descriptor-change evidence tags, one signed delta per endpoint, and confidence in $[0, 1]$. We score measured choice utility, balanced hard-label accuracy, scale-normalized delta MAE, confidence Brier score, format validity, and evidence F1; invalid or missing fields receive conservative defaults.

Table 34: Paired improvement over each model's action-first SFT checkpoint on 512 Biogen HLM+solubility generations. Positive is favorable after orienting error metrics; † marks a 95% CI crossing zero.

Metric	4B improvement	8B improvement
Choice utility	.07545	.01113
Choice accuracy	.04785	.01660
Normalized delta MAE	.02910	-.03537
Brier score	.06177	-.00295
Format validity	.07813	.00000†
Evidence F1	.00195†	.00195†

These metrics diagnose retrospective structured-output fidelity; causal mediation by language or tags remains an open question.

I ADDITIONAL OPTIMIZATION ANALYSES

These experiments compare alternative targets, optimization objectives, and structured supervision under the same deterministic TV controller. Unless otherwise specified, ϵ = .005 and gains are relative to the original task anchor. "Raw" denotes the uncontracted policy on bounded utility, rather than physical assay units. Validation-only analyses characterize optimization behavior; they do not provide additional test replications.

I.1 FOUR-ENDPOINT ITERATION AND THE EFFECTIVE-STEP CONTROL

The four-endpoint serializer provides HLM, RLM, solubility, and MDR1 deltas in every completion and retains 1,995/330/391 train/validation/test panels. A 100-step Pareto-target round followed by another 100-step round around the learned reference defines measured mirror policy iteration (CMPI); its matched fixed-target control continues the same initial round without recentering. Both use target CE, hinge 20, rank-16 LoRA, learning rate 10^{-4}, three optimizer seeds, and no auxiliary language loss.

On test, CMPI improves controlled gain over the fixed target by .000114 [.000066, .000160], isolating a benefit of target recentering relative to that control. Table 35, however, shows that scalar and pointwise-dual baselines have higher controlled gain: CMPI-minus-scalar is $-.000081$ [$-.000140$, $-.000031$] and CMPI-minus-dual is $-.000071$ [$-.000124$, $-.000022$]. All displayed method/endpoint test intervals are positive; validation solubility intervals for PG and the set encoder were inconclusive.

Table 35: Fixed-scale four-endpoint test. Gains and violation are multiplied by 10^3. Every row uses the same original-anchor TV controller and three optimizer seeds.

Method	Mean	HLM	RLM	Sol.	MDR1	V
Fixed-target CE	.304	.374	.437	.178	.224	.747
Scalar projection	.498	.671	.722	.423	.178	.655
Exact PG	.422	.508	.595	.316	.269	.668
Pointwise dual	.488	.633	.692	.404	.223	.637
ChemBERTa–Set	.289	.434	.315	.176	.230	.508
CMPI	.417	.517	.564	.299	.288	.718

Two scalar mirror rounds (CSPI) give controlled test gain .000537 versus .000498 for one scalar η = 2 round, but the paired interval [$-.000003$, .000079] crosses zero. A one-step η = 4 control gives .000540 and is statistically indistinguishable from CSPI ([$-.000042$, .000032] for CSPI minus the

control). The observed improvement over a smaller mirror step therefore does not isolate iterative correction of neural error. A third round changes validation gain by only .000004 [−.000012, .000020], and orbit averaging also leaves CSPI versus scalar unresolved. Mistral CSPI validation gain .000387 is close to scalar .000391, and both have inconclusive solubility intervals.

Other target constructions reinforce the role of the learning signal. TV-constrained target imitation at its selected radius .10 gives controlled validation gain .000360, compared with .000559 for scalar $\eta = 4$. A teacher already contracted to TV=.005 gives .000024 [−.000023, .000073]; a two-stage .05-to-.005 teacher schedule gives .000027 [−.000023, .000077]. Scalar-to-Pareto iteration retains positive four-endpoint gain .000528 but does not exceed scalar, pointwise dual, or CSPI. The targets remain mathematically feasible; the small retained gains expose the difficulty of fitting them with a neural policy.

I.2 Direct gradients and context constraints on four endpoints

On the four-endpoint validation task, certificate-aware PG with objective weight 1000 (CAPG) obtains .000579 [.000518, .000640]. Contextwise duals (CPDPO) give .000601 [.000540, .000660] and exceed scalar $\eta = 4$ by .000042 [.000014, .000070], but CPDPO-minus-CAPG is .000022 [−.000001, .000043]; CPDPO's MDR1 gain is .000101 [−.000025, .000231]. Full four-endpoint superiority and all-endpoint conclusions remain unresolved, as does the incremental dual contribution. The direct 8B result concerns the complete algorithm and cannot be attributed to the dual component alone.

Adding the same structured loss to four-endpoint CPDPO gives .00060044, compared with .00060053 without it; the paired difference is -9.0×10^{-8} [−.0000291, .0000274]. Within-panel endpoint-span balancing gives .00056582 and leaves MDR1 inconclusive. A learned certified expert router obtains .000561 with positive endpoint intervals and exceeds uniform mixing but not the strongest static or direct-PG alternatives. Four specialist actors combined by a geometric product give .000307, and a training-selected linear mixture gives .000216. An active-assay analysis reaches .000960 with fitted Q versus .000946 for a fixed three-assay subset, but its MDR1 mean is negative; the outcome-aware three-assay upper bound is .001029, close to the four-assay value .001033.

I.3 Raw reward and retained utility can move in opposite directions

The following RLM–solubility analyses use the Qwen4B setup with three optimizer seeds. Removing the radial derivative while keeping the same forward controller increases raw gain but reduces controlled gain (Table 36); the direct comparison with scalar yields a controlled difference of −.000257 [−.000365, −.000150] for the stop-gradient CPDPO variant. This supports differentiating the deployed objective rather than treating its radial scale as constant.

Table 36: Qwen4B RLM–solubility validation ablations. Values are multiplied by 10^3. Preference weights and gradient choices are fixed within each row; all deployment rows use the same original-anchor TV=.005 rule.

Variant	Raw mean	Controlled mean	Controlled Sol.
CPDPO-R	30.99	.719	.378
Stop-gradient CPDPO	63.37	.445	.140
Stop-gradient, endpoint balanced	47.83	.335	.122
Pareto preference (.005), stop-gradient	72.07	.604	.191
Pareto preference (.010), stop-gradient	69.64	.607	.199
Pareto preference (.020), stop-gradient	67.90	.628	.230

Restoring the radial derivative for smaller preference weights .0005/.001/.002 gives controlled means .000670/.000693/.000692, each with a paired interval against scalar crossing zero; the solubility trade-off remains. Adding a scalar-target prior yields .000725, and preference additions to that prior give .000694/.000712/.000689. Endpoint reweighting toward solubility gives .000706/.000689/.000627. Eight learning-rate, step-budget, language-loss, and penalty variants do not establish an improvement over scalar.

Initializing CPDPO-R from the scalar adapter gives .000642, versus .000710 for its matched cold-start policy. A two-policy molecular router gives .000714 versus .000719 for the training-selected

static route. A validation-selected 20% residual blend evaluated on the existing test gives .00044874 versus .00043532 for scalar, with paired interval [−.00000700, .00003693]; the paired comparison remains unresolved.

I.4 Model-family and endpoint-pair transfer

Five fixed-anchor Mistral seeds give controlled validation CPDPO-R gain .00021209 [.00001133, .00041203] and scalar gain .00034849 [.00023033, .00046470]; the paired difference is −.00013640 [−.00038314, .00010319]. A lower-learning-rate grid improves raw utility but increases raw violation. A raw-TV quadratic barrier with coefficients .25/.5/1/2 gives negative controlled validation gains in the seed-37 analysis. These results concern optimization transfer across model families, not failure of the TV guarantee.

Training-entropy weighting of the five Mistral members gives test raw mean .021547 versus .008434 for scalar, with paired difference .013113 [.006952, .019113]. Its controlled mean is .00033840 versus .00032723, with paired difference .00001117 [−.00008834, .00010992]; the RLM endpoint difference is negative, −.00019925 [−.00032197, −.00007722]. Joint calibration of the unensembled Mistral family also favors scalar.

The Nash objective uses $cD^{-1}\sum_d \log[(c+G_{x,d})/c]$ with $c = .05 > \epsilon$, retaining the same controller and dual. Five-seed Qwen8B validation gain is .00069828 $[.00058404, .00080957]$ and exceeds scalar by .00009471 $[.00000711, .00018078]$. Mistral gives .00020847 versus .00034849 for scalar, with an unresolved negative paired estimate. On a separate solubility–MDR1 task (1,947/336/391 panels), Nash validation gain .00037525 is statistically unresolved from scalar .00036896. Equal-mean CPDPO-R evaluated on that task's test gives .00045178 versus .00044883 for scalar; its solubility difference is $-.00007849$ $[-.00015096, -.00000517]$. Re-evaluating Nash on the previously used RLM–solubility test gives .00040146 versus .00043957 for mean-CPDPO-R. These findings favor retaining the equal-mean objective on the main 8B task, while limiting transfer conclusions.

On Qwen4B, the existing three-seed joint source-conformal analysis gives q_{SC} = .514905 at nominal 95% coverage. Test CPDPO-R gain is .00083696 versus .00084544 for scalar; CPDPO-R minus scalar is −.00000848 [−.00013015, .00012643]. A shared calibration threshold increases retained gains without resolving the underlying controlled tie.

I.5 Structured supervision and additional recentering

Each RLM–solubility training panel was expanded into four candidate-conditioned evidence and delta continuations, with the choice token masked and the CPDPO-R objective unchanged. The three-seed Qwen8B controlled validation mean is .00071099 versus .00071582 for the matched original policy; the paired difference is −.00000483 [−.00010227, .00010145]. Fixed-anchor Mistral gives −.00002631 [−.00011012, .00005743], with negative RLM gain. The additional continuations yield no resolved improvement in controlled utility and do not establish an additional reasoning contribution.

A separate Qwen8B seed-37 analysis scores each first-round actor on training panels, attaches its predictions as a new reference, and runs 100 additional steps of CPDPO-R or scalar fitting. Relative to the original HLM anchor, second-round raw gains are .06378 and .03346; after original-anchor TV control they are .00062352 and .00067379, with CPDPO-R-minus-scalar interval [−.00013624, .00003356]. This one-seed validation comparison increases raw displacement without demonstrating a better retained update. Both arms update their reference, but their first-round objectives differ, so the comparison does not isolate recentering alone.

I.6 Alternative parameterizations and task construction

We also evaluate KKT log-ratio heads and encoder-adaptive variants, longer ratio fitting, interval missingness, alternative divergences, scalarizations, and support/solver analyses; the preceding sections report these comparisons, with the remaining results below. Head-only KKT fitting is below the action-token policy. A 200-step adaptive linear ratio model raises controlled validation gain from .000500 to .000608, but target-policy TV changes only from .100373 to .097053, and longer 400-step training does not establish a better target approximation. A prospective HLM–RLM task

construction yields only 167 validation and 195 test panels under the fixed trade-off filter and was not used for policy training. The experiments collectively identify direction learning under a tight deployment budget as the recurring challenge; they do not establish superiority of every proposed variant.

J VISUAL SUMMARIES OF THE REPORTED EXPERIMENTS

The following figures complement the model-scale comparison in Figure 3 with views of transfer, constraint granularity, and data availability. The constraints act at target construction, while order averaging and missing measurements affect different parts of the learning and deployment pipeline. Each plot uses the corresponding study's reported normalization and evaluation split. Error bars denote the original 95% confidence intervals wherever those intervals were reported.

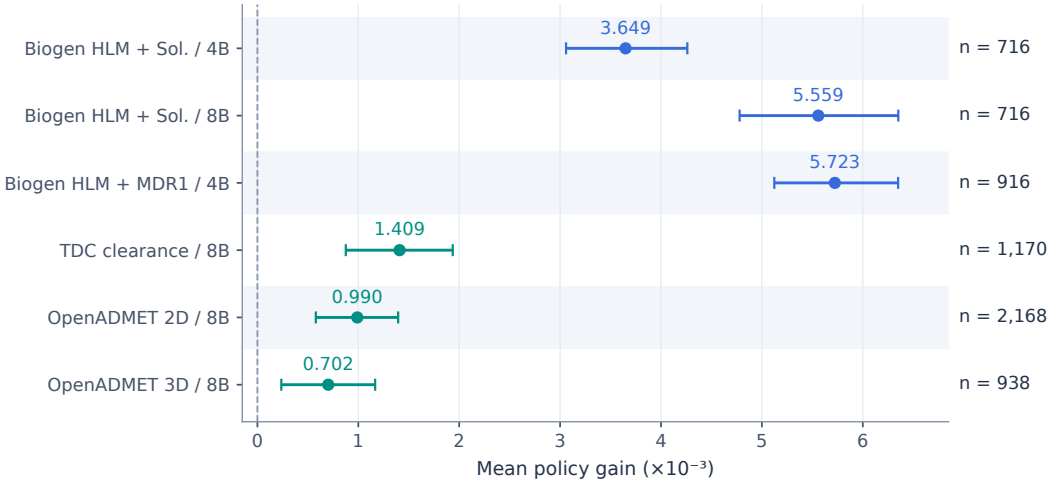

Figure 4: Target-based policy gains across the six evaluations in Table 6. Points and bars are the reported means and source-cluster 95% intervals on retrospective normalized utilities. Sample sizes count panels. These are the table's panel-relative evaluations, distinct from the fixed-scale external analyses in Tables 18–20.

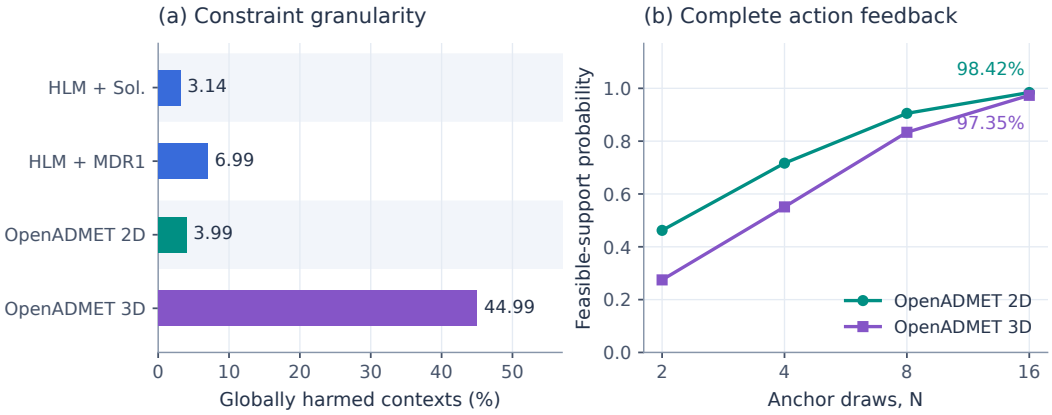

Figure 5: Why constraint granularity and complete feedback matter. Left: percentage of contexts harmed by the exact global target but protected by the contextwise target (Table 22). Right: exact probability that N anchor draws contain a feasible action support (Table 23). These are target-construction quantities, not learned-policy outcomes; connecting lines guide the eye.

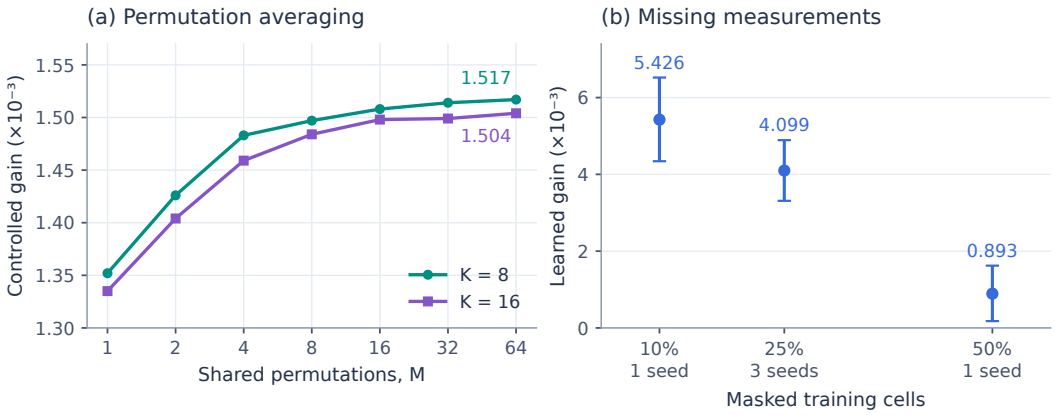

Figure 6: Two separate robustness studies. Left: controlled gains as shared permutation count increases for real 8/16-candidate panels, using panel-relative utilities and two training members (Table 26). Right: learned-policy test gains under masked training measurements, with original 95% intervals (Table 13); tests are fully observed. The 25% masking row averages three seeds, while 10/50% use seed 37. The panels have different objectives and scales and do not form a joint frontier.