BEYOND ACTION ENTROPY: QUOTIENT-SPACE EXPLORATION FOR GENOME-SCALE METABOLIC MODEL REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Repairing scientific models from functional observations differs fundamentally from supervised prediction: feedback may certify a solution without revealing which structural correction is responsible. We study this setting for genome-scale metabolic model (GEM) repair, where multiple reaction edits can explain the same phenotypes and many apparently distinct edits correspond to the same biological mechanism. This many-to-one structure creates a hidden failure mode for conventional exploration: diversity in the output space need not translate into diversity of scientific hypotheses. We introduce QUOTIENTPO, which collapses equivalent repairs into canonical mechanisms and optimizes exploration directly over the resulting quotient space. To make quotient exploration informative under finite rollouts, we derive a kernelized Rényi estimator that resolves graded crowding among distinct repair cores beyond coarse exact-match counts. On 2,212 held-out GEMs, QUOTIENTPO improves Success@32 from 17.93% to 20.10% (+12.1% relative) while consistently increasing distinct successful-core discovery under the same sampling budget. These results establish quotient-space exploration as a principled approach to mechanism-level discovery under verifier-induced equivalence.

1 INTRODUCTION

AI4Sci (Wang et al., 2023) is increasingly moving beyond predicting scientific observations toward reasoning over, generating, and modifying scientific models and structures (Abramson et al., 2024; Merchant et al., 2023). Unlike ordinary prediction targets, these scientific objects act as hypotheses about underlying mechanisms whose validity is assessed indirectly through physical laws (Karniadakis et al., 2021) or experimental feedback (Boiko et al., 2023). Evaluating candidates against functional constraints rather than gold targets characterizes *verifier-guided scientific AI*, posing the core challenge of learning structural corrections from indirect evidence alone. This challenge directly materializes in genome-scale metabolic models (GEMs) (Mardinoglu & Palsson, 2025), which operationalize genomic sequences into executable metabolic machinery. To overcome this barrier, we introduce *phenotype-supervised GEM repair*: an amortized framework that replaces costly manual curation to scale predictive mechanistic modeling to uncharacterized organisms.

This repair paradigm is fundamentally challenged by partial identification: phenotype observations validate models without uniquely determining underlying mechanisms (Raue et al., 2009; Benedict et al., 2014). This underdetermination creates a core tension: while multiple distinct mechanisms can satisfy identical phenotypes, combinatorial edits often implement the exact same mechanism, rendering surface diversity deceptive. Prior work addresses non-surface redundancy via semantic classes, kernels, or symmetries (Kuhn et al., 2023; Nikitin et al., 2024; Kim et al., 2025; Li et al., 2026). Yet in verifier-guided repair, a single functional intervention admits diverse raw realizations, meaning output similarity cannot define which candidates constitute the same scientific hypothesis. Here, identity is induced operationally by the verifier itself. This raises a key question: How should exploration be structured when diversification is defined up to verifier-preserving transformations?

Our answer is to *quotient before exploring*. Conceptually, we partition repair edits into verifier-preserving equivalence classes through canonicalization, establishing scientific identity instead of syntactic multiplicity as the unit of diversity. Computationally, exact identity provides only a coarse

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.

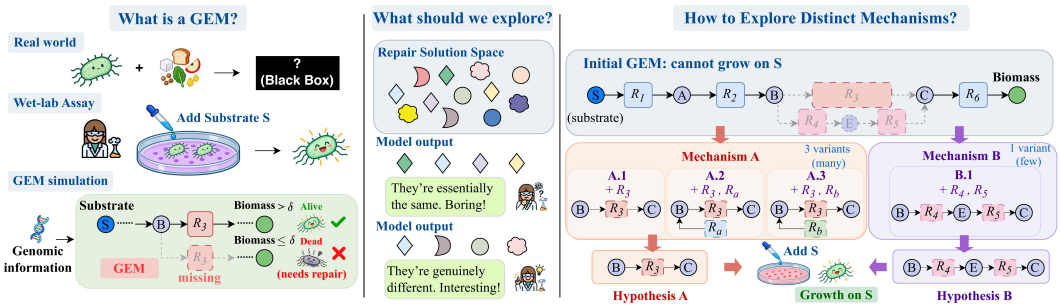


Figure 1: **Overview.** QUOTIENTPO canonicalizes equivalent surface edits to explore distinct mechanistic hypotheses rather than redundant syntactic variants for GEM repair.

finite-sample exploration signal, becoming especially degenerate when exact mechanism collisions are sparse. We therefore relax binary collision counting into a graded, kernelized Rényi entropy objective (Principe et al., 2000; Giraldo et al., 2014), allowing mechanistic crowding to be detected even among non-identical repairs. Based on these principles, we introduce QUOTIENTPO (Fig. 1), an algorithm combining group-relative policy optimization with quotient-space exploration. Together, our theory and experiments establish quotient-space exploration as a principled route to broader valid-mechanism coverage, advancing a Sci4AI perspective on verifier-guided learning.

Our contributions are threefold:

- **Formulation and Benchmark for Scientific Repair:** We formulate *phenotype-supervised GEM repair* to generalize edits across organisms without test-time phenotypes, and establish the first large-scale benchmark for amortized GEM correction.
- **Theoretical Foundation of Quotient Exploration:** We identify *representation multiplicity* as an intrinsic failure mode in which standard policy entropy rewards syntactic redundancy, and establish theoretical connections between quotient-space diversity and deployed hypothesis coverage.
- **Tractable Algorithm and Empirical Validation:** We develop QUOTIENTPO, making quotient exploration practical via kernelized Rényi estimation of graded mechanistic crowding, and demonstrate consistent gains in multi-budget repair and distinct mechanism recovery.

2 BACKGROUND AND RELATED WORK

Genome-scale metabolic reconstruction and model repair. GEMs support mechanistic phenotype simulation and metabolic engineering, yet incomplete biological knowledge and reconstruction choices can leave missing and spurious reactions (Thiele & Palsson, 2010; Machado et al., 2018; Zimmermann et al., 2021; Hsieh et al., 2024). Classical constraint-based methods repair each GEM independently through gap filling or phenotype reconciliation, without sharing repair knowledge across organisms. The same phenotypic constraints may admit multiple valid repairs (Satish Kumar et al., 2007; Kumar & Maranas, 2009; Thiele et al., 2014; Hartleb et al., 2016). Recent learning-based methods recover missing reactions through link or hypergraph prediction (Yadati et al., 2020; Chen et al., 2023; Huang et al., 2024; Liu et al., 2024; Zhao et al., 2026), but largely target synthetically removed reactions rather than functional bidirectional repair. We instead learn a shared repair policy across organisms, using phenotypes only as training supervision to predict functionally validated bidirectional edits without test-time phenotype access.

Diversity-aware reinforcement learning for structured prediction. Reinforcement learning (RL) optimizes structured outputs under non-differentiable functional rewards, including molecular design (You et al., 2018), executable code (Le et al., 2022), and outcome-based reasoning (Shao et al., 2024). Diversity has been pursued through maximum-entropy policies (Haarnoja et al., 2018), unsupervised skill discovery (Eysenbach et al., 2018; Sharma et al., 2020), quality-diversity optimization (Grillotti et al., 2024), and GFlowNets for reward-weighted object generation (Bengio et al., 2021; Jain et al., 2022). Recent work further considers semantic equivalence and kernelized set diversity (Kuhn et al., 2023; Li et al., 2026). However, these approaches do not resolve the verifier-induced many-to-one mapping central to GEM repair, where distinct edit sets can collapse to the same repair core. Raw-output diversity can therefore reward representation multiplicity without expanding distinct repair hypotheses, motivating quotienting before exploration.

3 PRELIMINARIES AND PROBLEM FORMULATION

Genome-scale metabolic models. A GEM represents an organism's metabolic network as a directed bipartite graph $M = (\mathcal{C}, \mathcal{R}, S)$ between metabolites $\mathcal{C}$ and reactions $\mathcal{R}$ (Klamt et al., 2009; Smart et al., 2008; Basler et al., 2011). The stoichiometric matrix $S \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{R}|}$ defines weighted edges between them, specifying the consumption and production coefficients of each reaction.

Formulation of GEM repair. We frame GEM correction as constrained graph editing through reaction additions and deletions. Given an initial model M_0 with reaction set $\mathcal{R}_0$ and a candidate reaction universe $\mathcal{U}$, a repair $E = (A, D)$ adds $A \subseteq \mathcal{U} \setminus \mathcal{R}_0$ and removes $D \subseteq \mathcal{R}_0 \cap \mathcal{U}$, yielding the repaired model M_E. For each tested phenotype i, flux balance analysis (FBA;Orth et al., 2010) over its nutrient-specific assay(s) computes optimal biomass flux $v_i^*(M_E)$, giving predicted phenotype $\widehat{y}_i(M_E) = \mathbb{I}[v_i^*(M_E) > \delta]$ under growth threshold $\delta > 0$. Given observed BacDive phenotypes $y \in \{0,1\}^N$, repair quality is measured by phenotype consistency $\Phi_y(M_E) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\widehat{y}_i(M_E) = y_i]$. FBA therefore serves directly as the functional verifier that maps a repaired GEM to phenotype predictions. The finite repair space $\mathcal{F}$ contains all legal reaction edits over candidate universe $\mathcal{U}$.

Learning problem. Let G denote the organism-specific protein representation derived from its genome-encoded proteins. We assume a training distribution $\mathcal{D}$ over organisms, where for each training instance $(M_0, G, y) \sim \mathcal{D}$, y contains the available wet-lab phenotypes. Neither a corrected GEM nor a gold reaction edit set is available, and the observed phenotypes generally do not uniquely identify the underlying repair. We therefore learn a stochastic repair policy $\pi_\theta(E \mid M_0, G)$ over valid repairs. Importantly, the phenotypes y provide training supervision through $\Phi_y(M_E)$ only and are never given as model input. The resulting learning objective is

$$\max_{\theta} \mathbb{E}_{(M_0,G,y)\sim\mathcal{D}} \mathbb{E}_{E\sim\pi_\theta(\cdot|M_0,G)} \left[\Phi_y(M_E)\right] \quad \text{s.t.} \quad E \in \mathcal{F}. \tag{1}$$

At test time, only (M_0, G) is available and the learned policy outputs a repair E and the corresponding M_E. Because no organism-specific corrected GEM or gold edit set is available, evaluation uses only held-out phenotype consistency target.

4 METHODOLOGY

The formulation in Eq. (1) directly motivates **policy-gradient RL**[1] **with explicit multi-solution exploration**: **(1)** *Bypassing non-differentiable verifiers and solver bias*: Repair quality is evaluated by non-differentiable FBA-based verifiers, preventing direct backpropagation. While offline mixed-integer linear programming (MILP) (Satish Kumar et al., 2007; Benedict et al., 2014; Schroeder & Saha, 2020) can provide demonstrations for supervised fine-tuning (SFT), it arbitrarily breaks ties across non-unique feasible sets, biasing SFT toward solver heuristics. We therefore adopt RL to optimize verifier rewards directly beyond offline demonstrations. **(2)** *Preventing collapse under partial identification*: Across a combinatorial graph edit space, observed phenotypes only partially identify the metabolic network, creating a landscape with multiple discrete optima where mechanistically distinct repairs achieve identical rewards. A policy collapsed onto any single valid repair is already reward-optimal, leaving standard reward maximization with zero gradient signal to discover alternative mechanisms. Under finite rollout budgets, this wastes samples on redundant variations around a single mode, motivating explicit exploration to preserve diverse scientific hypotheses.

4.1 REPRESENTATION-INVARIANT EXPLORATION WITH QUOTIENT ENTROPY

Action entropy rewards representation multiplicity, not mechanistic diversity. A natural exploration bonus is the Shannon entropy over raw repair actions $H(\pi_\theta)$, leading to the standard entropy-regularized objective $\mathcal{J}_{\text{act}}(\theta) = \mathbb{E}_{(M_0,G,y)\sim\mathcal{D}}[\mathbb{E}_{E\sim\pi_\theta}[\Phi_y(M_E)] + \beta H(\pi_\theta)]$, where $\beta > 0$ controls the exploration strength. The problem is that entropy counts representations, not repair hypotheses. If two equally rewarded hypotheses differ only in that one admits n equivalent action realizations, action entropy gives it an additional preference of $\beta \log n$. Thus, a hypothesis is favored

[1] We frame our learning under group-relative policy optimization (GRPO) (Shao et al., 2024), which baselines a critic-free rollout group to avoid critic variance under combinatorial edits (Bello et al., 2017; Kool et al., 2019) while providing the sample pool for quotient estimation.

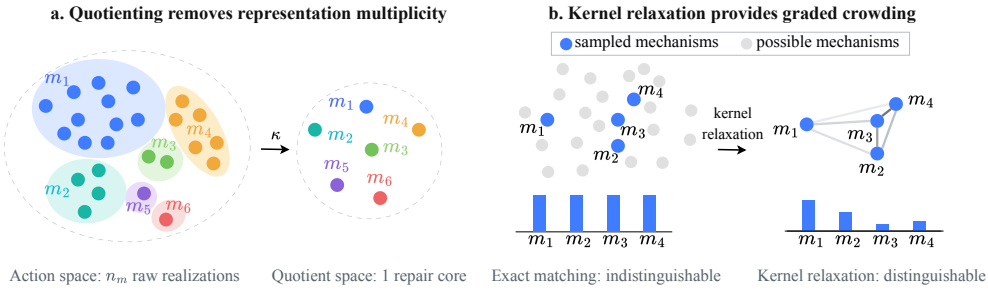


Figure 2: **Theoretical intuition.** **(a)** Each core m contains n_m equivalent raw edits contracted by κ into a single point, preventing large n_m from distorting exploration. **(b)** Bar heights show exploration credit. Exact matching gives identical credit to equal-count cores, whereas kernel similarity favors isolated discoveries (m_1) over clustered candidates (m_2–m_4).

purely by representation redundancy, rather than any distinct biological mechanism. We formalize this multiplicity bias in Appendix B.2.

Exploring distinct mechanisms requires quotienting equivalent repairs. To remove this bias, equivalent repairs should contribute only once to exploration. We define a deterministic, verifier-preserving map κ_y that removes redundant edits, contracting the n_m equivalent raw realizations into a canonical repair core $m = \kappa_y(E)$ (Fig. 2a). Let $\mathcal{M}_y$ denote the set of repair cores, serving as operational representations of underlying mechanisms. Repairs sharing the same core are treated as equivalent hypotheses, and the policy induces the quotient distribution $q = q_{\theta,y}$, where $q(m) = \sum_{E:\kappa_y(E)=m} \pi_\theta(E \mid M_0, G)$ aggregates probability mass across all equivalent raw realizations.

Theorem 4.1 (Within-core invariance characterizes quotient exploration)**.** *An exploration functional $\mathfrak{D}$ is invariant to within-core redistribution if and only if it depends on the raw policy π only through its induced core distribution q_π. Equivalently, there exists a functional $\overline{\mathfrak{D}}$ such that $\mathfrak{D}(\pi) = \overline{\mathfrak{D}}(q_\pi)$.*

Theorem 4.1 formalizes why exploration should depend only on the induced distribution over repair cores. The intuition is direct: splitting probability across equivalent realizations inflates action entropy, whereas the aggregated core probability $q(m)$ remains unchanged. Evaluating entropy over this aggregated distribution removes multiplicity bias by construction, leading to the quotient-regularized objective: $\mathcal{J}_{\text{quot}}(\theta) = \mathbb{E}_{(M_0,G,y)\sim\mathcal{D}}\left[\mathbb{E}_{E\sim\pi_\theta}\left[\Phi_y(M_E)\right] + \beta H(q)\right]$, where the quotient entropy is $H(q) = -\sum_{m\in\mathcal{M}_y} q(m)\log q(m)$. Crucially, optimizing $\mathcal{J}_{\text{quot}}$ weights distinct hypotheses strictly by task utility rather than raw representation multiplicity (Appendix B.2).

4.2 Graded Quotient Estimation with Rényi Kernels

Exact matching gives coarse finite-sample quotient estimates. Because the core probability $q(m)$ is an unobserved fiber mass rather than a closed-form likelihood, the quotient diversity must be estimated from sampled cores. Exact-frequency estimation, however, provides only binary crowding information: structurally different cores with the same empirical multiplicity receive identical credit. This limitation becomes extreme in the sparse-collision regime. With L sampled cores, the expected number of exact pairwise collisions is $\binom{L}{2}\sum_m q(m)^2$. When $\binom{L}{2}\sum_m q(m)^2 \ll 1$, a rollout group contains few or no repeated cores, making distinct samples indistinguishable under exact matching (Appendix B.3). Thus, exact-count estimation provides only a coarse finite-sample exploration signal, even before collisions vanish entirely.

Rényi kernels turn binary identity into graded mechanistic crowding. The key limitation is the knife-edge nature of exact matching: it distinguishes identical from non-identical cores, but provides no graded notion of crowding among distinct repairs. Indeed, the exact pairwise collision probability $\sum_m q(m)^2$ is precisely the quadratic Rényi concentration $\exp[-H_2(q)]$ (Principe et al., 2000; Giraldo et al., 2014). This suggests replacing quotient Shannon entropy with quadratic Rényi entropy, whose exact collision form admits a graded relaxation while preserving the quotient identity. Representing each repair core as a signed edit set, we replace the exact-match collision operator with the graded kernel $k(m, m') = \exp(-(1 - J(m, m'))/h)$, where J is Jaccard similarity and $h > 0$ controls smoothness. As $h \to 0$, k recovers exact matching, whereas finite h assigns progressively larger collision mass to increasingly overlapping repairs. This kernel relaxation exposes mechanistic

crowding before exact duplicates dominate (Figure 2b), discounting exploration credit for clustered repairs while rewarding genuinely distinct bypasses. Replacing exact-collision Rényi entropy with the kernelized surrogate $H_{2,k}(q) = -\log \mathbb{E}_{m,m' \sim q}[k(m, m')]$ gives the crowding-aware objective:

$$\mathcal{J}_{\text{renyi}}(\theta) = \mathbb{E}_{(M_0,G,y)\sim\mathcal{D}} \left[\mathbb{E}_{E\sim\pi_\theta} \left[\Phi_y(M_E) \right] + \beta H_{2,k}(q) \right]. \tag{2}$$

Kernelized Rényi estimation therefore preserves the quotient notion of identity while replacing coarse binary collisions with a graded finite-sample exploration signal.

4.3 QuotientPO: Quotient-Space Policy Optimization

QuotientPO operationalizes quotient exploration within GRPO by decomposing verifier feedback and quotient diversity into candidate-level credit. We optimize a verifier-shaped relaxation where each component provides a principled finite-sample realization of Eq. (2).

Verifier-shaped task reward. For a candidate repair E, we define the terminal task reward as $R_y(E) = \Phi_y(M_E) - \Phi_y(M_0) + 2\,\mathbb{I}[\text{all satisfied}]$. Since $\Phi_y(M_E) - \Phi_y(M_0) \leq 1$, the success weight of 2 enforces a strict margin ensuring that any fully verified repair ($R_y \geq 2$) strictly outranks any partial improvement ($R_y < 1$). This rules out reward-level satisficing on partial repairs while retaining dense shaping rewards during early exploration.

Group-relative quotient credit. For each organism, we sample L repairs and canonicalize them into repair cores $m_1, \ldots, m_L$. While the kernelized Rényi entropy in Section 4.2 evaluates diversity collectively across the rollout, policy optimization requires candidate-level credit. A leave-one-out linearization of this group entropy around the empirical mean gives the normalized crowding cost

$$c_i = 2L \frac{\sum_{j\neq i} k(m_i, m_j)}{\sum_{a\neq b} k(m_a, m_b)} - 2, \qquad \sum_{i=1}^{L} c_i = 0. \tag{3}$$

A positive c_i indicates that a repair core is crowded by similar candidates, while a negative value indicates that it expands quotient-space coverage. The zero-sum form aligns this diversity signal with group-relative optimization. Its derivation is given in Appendix B.4.

QuotientPO objective. Let $\rho_i(\theta) = \pi_\theta(E_i \mid M_0, G)/\pi_{\theta_{\text{old}}}(E_i \mid M_0, G)$, and let $A_i = (R_y(E_i) - \bar{R})/(\sigma_R + \varepsilon)$ denote the group-relative advantage with sample mean $\bar{R}$ and standard deviation σ_R. The complete objective is

$$\mathcal{J}(\theta) = \frac{1}{L}\sum_{i=1}^{L} \min\Big(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\Big) - \frac{\beta}{L}\sum_{i=1}^{L} \rho_i(\theta)c_i + \mathcal{R}_{\text{aux}}(\theta). \tag{4}$$

The GRPO term optimizes verifier quality, while the quotient term shifts probability away from crowded cores toward distinct repair hypotheses. The auxiliary term $\mathcal{R}_{\text{aux}}$ combines QPI credit for newly discovered successful cores and an optional redundancy penalty for edits outside the canonical core. More details are given in Appendix A.

5 Experiments

5.1 Experimental Setup

Datasets. We pair strain-level metabolic phenotypes from BacDive (Schober et al., 2025) with matched NCBI (Sayers et al., 2022) proteomes, yielding 11,275 organisms spanning 2,454 genera and 40 phyla. Each proteome is reconstructed under 4 CarveMe (Machado et al., 2018) metabolic universes in the BiGG (King et al., 2016) namespace. Mapping and filtering phenotype annotations onto 16 core BiGG metabolite families produces 3,978 phenotype-consistent candidate GEMs for synthetic pretraining and 28,842 phenotype-inconsistent GEMs for post-training repair and evaluation. We first partition the post-training pool into train, validation, and test splits with an 85:5:10 ratio. Subsequent phenotype-assay filtering and certified-teacher selection leave 2,504/294 training/validation GEMs for SFT, while the RL split retains 9,196/1,087 GEMs and the final sealed test set contains 2,212 GEMs. All splits are species-disjoint, with 5 held-out genera demanding out-of-distribution generalization rather than taxonomic interpolation. Throughout training and evaluation, BacDive phenotypes are never provided as policy inputs, but are used to generate expert repairs via phenotype-constrained MILP for SFT, define RL rewards, and evaluation (Appendix C).

Table 1: **Main results.** MILP is an oracle with test-time phenotype access. SFT methods use checkpoints selected on the validation split, while RL methods are evaluated at the final checkpoint after an identical 864-update training budget. All results are evaluated on the test set, reporting mean $\pm$ standard deviation across three seed runs.

Method	Greedy	Success@1	Success@8	Success@32	Improved (mean@32)	Success Cores@32
MILP (Oracle)	15.01	-	-	-	25.63	0.15
SFT-Hard	$1.72_{\pm 0.16}$	$0.51_{\pm 0.03}$	$1.96_{\pm 0.11}$	$4.49_{\pm 0.05}$	$1.64_{\pm 0.03}$	$0.09_{\pm 0.01}$
SFT-Soft	$2.65_{\pm 0.35}$	$2.40_{\pm 0.05}$	$5.27_{\pm 0.41}$	$7.52_{\pm 0.49}$	$9.92_{\pm 0.18}$	$0.16_{\pm 0.01}$
Vanilla GRPO	$7.97_{\pm 0.29}$	$7.43_{\pm 0.03}$	$13.79_{\pm 1.09}$	$17.93_{\pm 1.54}$	$30.21_{\pm 0.59}$	$0.50_{\pm 0.01}$
MaxEnt	$7.29_{\pm 1.16}$	$5.97_{\pm 0.86}$	$12.69_{\pm 1.27}$	$17.12_{\pm 1.41}$	$29.35_{\pm 2.03}$	$0.48_{\pm 0.05}$
GFlowNet	$3.20_{\pm 0.23}$	$3.23_{\pm 0.15}$	$8.35_{\pm 0.38}$	$13.10_{\pm 0.17}$	$14.91_{\pm 0.49}$	$0.34_{\pm 0.01}$
QuotientPO	$\mathbf{8.47}_{\pm 0.25}$	$\mathbf{7.76}_{\pm 0.23}$	$\mathbf{15.30}_{\pm 0.30}$	$\mathbf{20.10}_{\pm 0.52}$	$\mathbf{31.59}_{\pm 0.82}$	$\mathbf{0.57}_{\pm 0.02}$

Verifier. We define 18 standardized FBA assays across 16 BacDive nutrient families, where each tested nutrient serves as the sole elemental source against an element-depleted background to preclude metabolic bypass (Fig. 3). Nutrients and oxygen uptake bounds are 10 and 20, respectively, with growth defined by biomass flux $> 10^{-3}$. The terminal verifier rewards any repair reproducing the observed phenotypes.

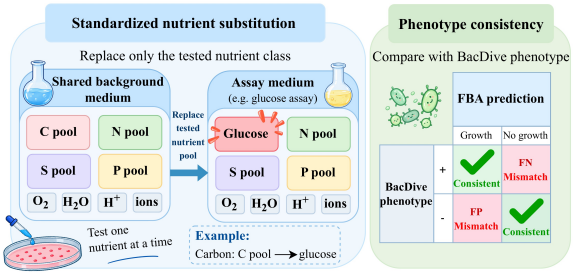


Figure 3: **Verifier design.**

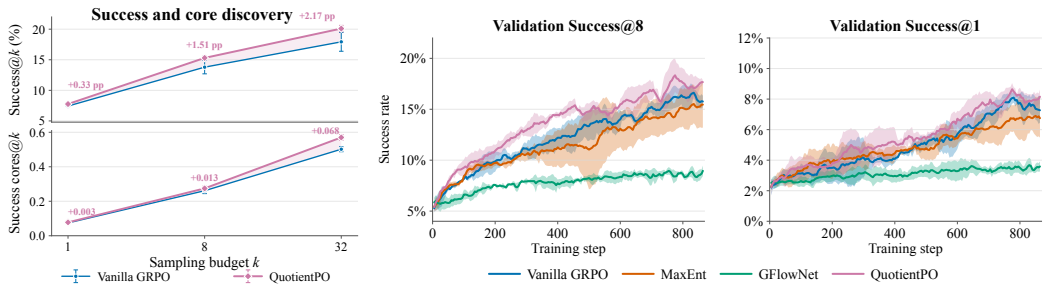
Evaluation Metrics. We use **Success@*k***(%) as our primary metric, defined as the probability that at least one of k sampled repairs achieves full phenotype recovery ($\Phi_y(M_E) = 1$), following multi-sample evaluation in RL and generative modeling (Hong et al., 2026; Chen et al., 2021; Feng et al., 2026). We report $k \in \{1, 8, 32\}$ alongside greedy decoding (Greedy), the rate of partial progress $\Phi_y(M_E) > \Phi_y(M_0)$ (Improved), and the macro-averaged number of distinct canonical repair cores recovered among successful repairs within the same 32 samples (Success Cores@32).

Baselines. We compare against 6 baselines across 3 paradigms: (1) *domain oracle*: MILP with test-time phenotype access; (2) *standard policy learning*: imitation across probabilistic and discrete-threshold architectures (SFT-Soft/Hard) alongside verifier RL (Vanilla GRPO); and (3) *diversity-oriented baselines*: action-entropy GRPO (MaxEnt) and GFlowNet (Bengio et al., 2021) (reward-proportional raw-space sampling without quotient equivalence). Appendix D details all baselines.

Implementation Details. Following prior methods of GEM GNN modeling (Hasibi et al., 2024; Huang et al., 2024), we train a 92.85M-parameter GNN policy that fuses frozen ESM-2 protein embeddings (Lin et al., 2023) into the global GEM representation, predicting one-shot repair probabilities over a candidate universe of 28,301 reactions. For RL training, we initialize from SFT-Soft and train for 3 epochs using AdamW with a fixed learning rate of $2e{-}6$, group size $L = 32$, bandwidth $h = 0.25$, and the quotient term weight $\beta = 0.02$. All FBA simulations use HiGHS (Huangfu & Hall, 2015). See Appendix A and E for more details.

5.2 Main Results

Quotient exploration improves repair coverage. Table 1 shows a consistent separation between mechanism-aware exploration and raw-space diversity objectives. QuotientPO achieves the strongest performance across all learned metrics, improving Success@32 from 17.93% (Vanilla GRPO) to 20.10% and core discovery from 0.50 to 0.57. These absolute success rates reflect task difficulty: even the phenotype-informed MILP oracle, despite direct test-label access, reaches 15.01% full repair within its fixed 600 s search budget. In contrast, MaxEnt does not improve over vanilla,



(a) Test-time scaling. (b) Training dynamics.

Figure 4: **Test-time scaling and training dynamics. (a)** Repair success (Success@k) and distinct successful-core discovery (Success cores@k). **(b)** Validation curves across different RL methods.

and GFlowNet underperforms despite explicitly encouraging diversity, confirming that syntactic variety fails to expand mechanistic diversity. The large gap over SFT further shows that supervised imitation alone cannot optimize verifier-defined outcomes.

The advantage emerges through test-time scaling. Figure 4a shows that quotient exploration improves both repair success and successful-core discovery as the sampling budget increases. Relative to Vanilla GRPO, the Success@k margin grows from $+0.33$ points at $k = 1$ to $+2.17$ at $k = 32$, while the gap in Success cores@k expands from $+0.003$ to $+0.068$ cores per GEM. Thus, the growing deployment advantage reflects genuine coverage of distinct successful repair cores rather than raw scaling alone. Training dynamics (Fig. 4b) mirror this pattern: differences in Success@1 remain modest, whereas QuotientPO separates clearly under Success@8 as training progresses. Together, these results show that quotient exploration becomes increasingly effective as more samples are allocated to uncover alternative valid repairs.

5.3 Empirical Analysis of QuotientPO's Mechanism

Verifier feedback moves repair beyond the supervised teacher prior. SFT is trained to amortize MILP-generated repairs, but a phenotype-consistent repair is not uniquely identified by its teacher edit set. Following Tokar et al. (2020), we project sampled repair sets via metric MDS on pairwise signed-Jaccard distances under shared coordinate axes (Fig. 5). While SFT-Soft clusters tightly around the teacher, QuotientPO discovers a well-separated region containing more successful solutions. This divergence also holds quantitatively across the test set: mean distance to the MILP teacher increases from 0.784 for SFT-Soft to 0.980 for QuotientPO (0.780 vs. 0.961 among successful repairs). Thus, the gain from verifier-guided optimization is not explained by more faithful imitation of the teacher, as successful QuotientPO repairs are substantially farther from it. This supports the partial-identification view of GEM repair, where the teacher provides one admissible correction but verifier feedback reveals alternative successful repairs beyond the supervised mode.

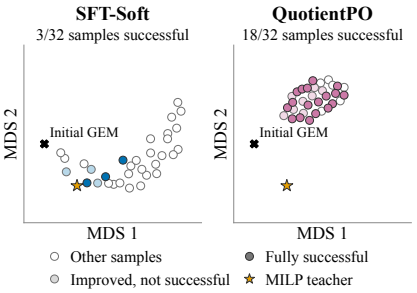
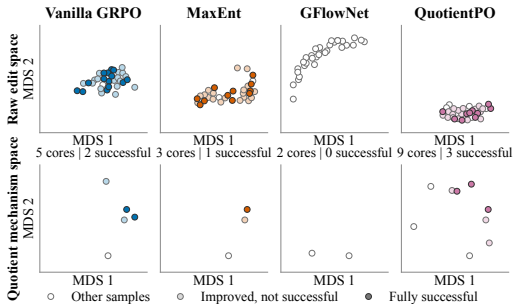
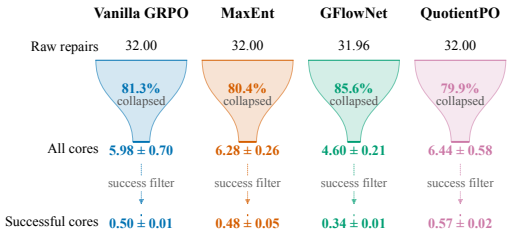


Figure 5: **Landscape of sampled repairs** (*Streptomyces axinellae*).

Quotient exploration converts diversity into useful mechanism coverage. Figure 6 exposes a key failure mode: although MaxEnt and GFlowNet explicitly diversify raw repair sets, over 80% of their sampled representations collapse after canonicalization, and neither improves successful-core yield over Vanilla GRPO. In contrast, QuotientPO applies diversity pressure directly in quotient space, reaching 6.44 distinct cores and 0.57 successful cores per GEM under the same 32-sample budget, compared with 5.98 and 0.50 for Vanilla GRPO. To disentangle mechanism coverage from the higher repair success rate itself, we further compare the two methods on paired GEMs where both produce at least r successful repairs and equalize the successful-sample budget. QuotientPO recovers more distinct successful cores at $r \in \{2, 4, 8\}$: 1.46 vs. 1.43, 2.05 vs. 1.96, and 2.78 vs.



(a) Raw diversity does not imply mechanism diversity.



(b) Raw-to-core collapse and success filtering.

Figure 6: **Exploration across raw and quotient spaces. (a)** MDS projections of sampled repairs before (top) and after (bottom) canonicalization on (*S. griseoflavus*). **(b)** Funnel transitions from distinct raw repairs ($k = 32$) to verifier-validated cores across the test set.

2.66, respectively (Appendix F.4). Thus, its coverage advantage persists after controlling for the number of successful samples, rather than arising solely from a higher repair-success frequency.

Kernel relaxation recovers graded diversity signals from coarse finite-sample counts. Under Exact Rényi-2, the exact-count tie rate rises from 7.34% to 10.23% over training (Table 2), while 78.2% of these final ties are resolved by kernel credit. In contrast, Kernel Rényi-2 reduces the exact-count tie rate of its rollout groups to 7.99%, while achieving the lowest successful-core collision (0.518) and largest mean edit Jaccard distance $d_J = 1 - J$ (0.301). Consistently, it also achieves the best Success@32 and successful-core (Table 12). These results show that kernel relaxation provides finer exploration credit precisely where exact multiplicity counts lose resolution.

Table 2: **Finite-sample credit resolution and successful-core diversity.** Training diagnostics use all rollouts.

	Training (200→864)		Successful Cores	
Method	**Exact ties (%)**	**Resolved (%)**	**Collision ↓**	**Mean d_J ↑**
Vanilla	8.19 → 8.59	60.55 → 84.86	0.5769	0.2408
Shannon	8.60 → 9.72	59.23 → 80.20	0.5687	0.2431
Exact R2	7.34 → 10.23	59.71 → 78.20	0.5285	0.2861
Kernel R2	8.57 → **7.99**	63.81 → 72.78	**0.5176**	**0.3008**

Structured initialization turns sparse verifier feedback into a learnable signal. Figure 7 separates the roles of representation learning and supervised policy formation. During single-edit imitation (SFT-A), which trains the model on individual MILP teacher operations, operation-level Hit@1 drops from 20.7% to 11.4% without structural pretraining and to 14.3% without ESM-2 conditioning, showing that both improve local repair prediction. Yet representation alone provides almost no terminal learning signal: before any RL update, only 0.4% and 0.8% of rollout groups exhibit within-group reward variation under random and pretrained policies. Set-level supervision changes this regime sharply, increasing the rate to 48.1% without SFT-A and to 51.6% with the complete supervised pipeline. Thus, while pretraining and genomic conditioning provide a local repair prior, set-level supervision is crucial to unlock group reward variance for RL.

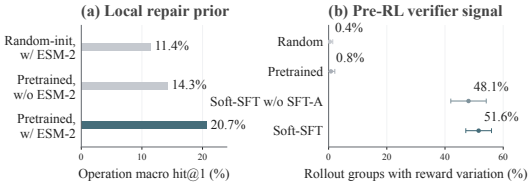


Figure 7: **Building an exploration-ready policy. (a)** Pretraining and ESM-2 enhance the local repair prior. **(b)** Set-level SFT unlocks rollout reward variance for RL. Error bars denote standard deviation across 8 rollout batches.

5.4 Robustness and Ablation Analysis

Robustness across RL objectives and policy architectures. As shown in Table 3(a,b), QuotientPO consistently outperforms baselines across RL formulations and architectures. Under RLOO (Ahmadian et al., 2024), it reaches 17.21 Success@32 (S@32, vs. 15.75 Vanilla), while under GRPO, it improves both Soft (20.10 vs. 17.93) and Hard (6.65 vs. 5.92) architectures. These results show that the benefit of quotient-space exploration is not tied to a specific RL formulation or architecture.

Table 3: **Robustness and ablation studies.** S@32 denotes Success@32 (%).

(a) RLOO Base

Method	S@32
Vanilla	15.75
MaxEnt	14.29
GFlowNet	11.57
QuotientPO	**17.21**

(b) Policy Architecture

Arch.	RL	S@32
Soft	Vanilla	17.93
	QuotientPO	20.10
Hard	Vanilla	5.92
	QuotientPO	6.65

(c) Design Ablation

Variant	S@32
Default	20.10
Raw-space kernel	18.16
w/o Quotient term	17.54
w/o QPI term	19.62
w/o Red. penalty	19.94

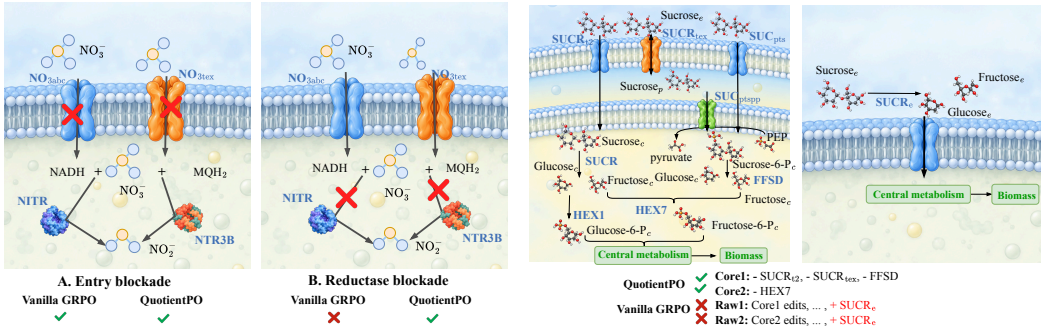
(d) Sensitivity

Param.	Setting	S@32
–	Default	20.10
$h(0.25)$	0.125	19.41
	0.5	18.57
$\beta(0.02)$	0.01	19.05
	0.04	20.00

Design ablation and sensitivity. QPI and the redundancy penalty yield smaller mean improvements, while the quotient term accounts for the dominant performance gain: removing it incurs the steepest drop in S@32 (from 20.10 to 17.54; Table 3c). Critically, with the same kernel and training budget, quotient-space exploration improves both Success@32 (20.10 vs. 18.16) and successful-core discovery (0.57 vs. 0.52) over its raw-space counterpart. Sensitivity analysis (Table 3d) shows broad robustness to the diversity weight β, but sensitivity to bandwidth h: a larger h over-smooths distant cores and blunts crowding selectivity, while leaving quotient identity unchanged.

6 Case Study

Figure 8 grounds the mechanistic failure modes in concrete biological repairs. On *S. mashuensis* (Fig. 8a), Vanilla GRPO collapses entirely onto boundary entry blockade (`NO3abc`, `NO3tex`), whereas QuotientPO discovers an orthogonal intracellular mechanism: jointly knocking out the NADH- and menaquinol-dependent reductases (`NITR`, `NTR3B`). On *S. palmae* (Fig. 8b), syntactically different Vanilla proposals repeatedly retain the same extracellular bypass (`SUCRe`), so surrounding edits vary without resolving the sucrose phenotype, whereas QuotientPO recovers clean successful cores. Counterfactual replays confirm that toggling `SUCRe` directly flips repair outcomes. These cases expose two failures of raw exploration: viable mechanisms can be missed, while surface variation around the same outcome-determining core can be mistaken for useful diversity. QuotientPO resolves both by directing exploration toward functionally distinct mechanisms.



(a) *S. mashuensis.* Nitrate utilization. (b) *S. palmae.* Sucrose utilization.

Figure 8: **Case studies of mechanistic repair.**

7 Conclusion

Repairing GEMs from indirect functional feedback faces a structural mismatch: combinatorial candidate edits can collapse to the same underlying repair mechanism, causing standard exploration to mistake syntactic multiplicity for scientific diversity. We address this mismatch by changing the unit of exploration from raw reaction-level actions to invariant repair mechanisms. By canonicalizing verifier-preserving edits, QuotientPO explores directly in quotient space, with a kernelized Rényi relaxation replacing coarse exact-match statistics with graded finite-sample crowding signals. This formulation removes representation-multiplicity bias while empirically improving repair success and the coverage of distinct valid mechanisms. More broadly, our results suggest that when scientific solutions admit equivalent representations, effective exploration should be defined over equivalence classes rather than surface representations (limitations and future directions in Appendix H).

AI USE STATEMENT

We used generative AI tools to assist with writing refinement and code generation. For writing, these tools were used to improve the clarity, grammar, conciseness, and presentation of author-written text. For code, generative AI was used to assist with generating and revising portions of the implementation. All AI-assisted code was manually reviewed, tested, and verified by the authors before use. All AI-assisted writing was likewise reviewed and edited by the authors. The authors take full responsibility for the correctness and final content of this work.

ETHICS STATEMENT

This work uses publicly available microbial genomic and phenotype data from NCBI (Sayers et al., 2022) and BacDive (Schober et al., 2025), together with established computational resources for genome-scale metabolic modeling (CarveMe (Machado et al., 2018), BiGG (King et al., 2016), etc.), in accordance with the terms and policies of their respective sources. All experiments are computational and involve no human participants, personally identifiable information, animal experiments, or experimental manipulation of living organisms. Our method operates only on in silico metabolic models and is intended for scientific modeling and computational research. While improved metabolic models may support downstream applications in metabolic engineering or synthetic biology, such uses should follow appropriate institutional biosafety procedures, regulatory requirements, and responsible-use practices.

REPRODUCIBILITY STATEMENT

We provide detailed materials to support reproduction of our results. The complete QuotientPO algorithm and optimization procedure are described in Appendix A, with the core implementation and usage instructions included in the supplemental materials. The benchmark construction and data-processing pipeline, based on publicly available databases, are documented in Appendix C; the benchmark construction pipeline and evaluation code are also provided in the supplemental materials. Model architecture and training details for all stages are specified in Appendix E, and the corresponding code is included in the supplemental materials. Implementation details for all baselines are provided in Appendix D, with their core implementations likewise included in the supplemental materials.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Georg Basler, Oliver Ebenhöh, Joachim Selbig, and Zoran Nikoloski. Mass-balanced randomization of metabolic networks. *Bioinformatics*, 27(10):1397–1403, 05 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr145. URL https://doi.org/10.1093/bioinformatics/btr145.

Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning, 2017. URL https://arxiv.org/abs/1611.09940.

Matthew N. Benedict, Michael B. Mundy, Christopher S. Henry, Nicholas Chia, and Nathan D. Price. Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. *PLOS Computational Biology*, 10(10):1–14, 10 2014. doi: 10.1371/journal.pcbi.1003882. URL https://doi.org/10.1371/journal.pcbi.1003882.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL https://doi.org/10.1038/s41586-023-06792-0.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, 1997. doi: 10.1109/SEQUEN.1997.666900.

Can Chen, Chen Liao, and Yang-Yu Liu. Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning. *Nature Communications*, 14(1):2375, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38110-7. URL https://doi.org/10.1038/s41467-023-38110-7.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function, 2018. URL https://arxiv.org/abs/1802.06070.

Bin Feng, Jiying Zhang, Xinni Zhang, Zijing Liu, and Yu Li. BioMD: All-atom generative model for biomolecular dynamics simulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LQDeJk6NOr.

Luis G. Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible kernels, 2014. URL https://arxiv.org/abs/1211.2459.

Luca Grillotti, Maxence Faldor, Borja G. León, and Antoine Cully. Quality-diversity actor-critic: Learning high-performing and diverse behaviors via value and successor features critics. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine*

Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 16416–16459. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/grillotti24a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Daniel Hartleb, Florian Jarre, and Martin J. Lercher. Improved metabolic models for e. coli and mycoplasma genitalium from globalfit, an algorithm that simultaneously matches growth and non-growth data sets. *PLOS Computational Biology*, 12(8):1–22, 08 2016. doi: 10.1371/journal.pcbi.1005036. URL https://doi.org/10.1371/journal.pcbi.1005036.

Ramin Hasibi, Tom Michoel, and Diego A. Oyarzún. Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. *npj Systems Biology and Applications*, 10(1):24, Mar 2024. ISSN 2056-7189. doi: 10.1038/s41540-024-00348-2. URL https://doi.org/10.1038/s41540-024-00348-2.

Matthew M. Hong, Jesse Zhang, Anusha Nagabandi, and Abhishek Gupta. TMRL: Diffusion Timestep-Modulated Pretraining Enables Exploration for Efficient Policy Finetuning. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.208.

Yunli Eric Hsieh, Kshitij Tandon, Heroen Verbruggen, and Zoran Nikoloski. Comparative analysis of metabolic models of microbial communities reconstructed from automated tools and consensus approaches. *npj Systems Biology and Applications*, 10(1):54, 2024. ISSN 2056-7189. doi: 10.1038/s41540-024-00384-y. URL https://doi.org/10.1038/s41540-024-00384-y.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks, 2020. URL https://arxiv.org/abs/1905.12265.

Weihong Huang, Feng Yang, Qiang Zhang, and Juan Liu. A dual-scale fused hypergraph convolution-based hyperedge prediction model for predicting missing reactions in genome-scale metabolic networks. *Briefings in Bioinformatics*, 25(5):bbae383, 09 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae383. URL https://doi.org/10.1093/bib/bbae383.

Q. Huangfu and J. A. J. Hall. Parallelizing the dual revised simplex method, 2015. URL https://arxiv.org/abs/1503.01889.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with GFlowNets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9786–9801. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jain22a.html.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, Jun 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. URL https://doi.org/10.1038/s42254-021-00314-5.

Hohyun Kim, Seunggeun Lee, and Min hwan Oh. Symmetry-aware gflownets, 2025. URL https://arxiv.org/abs/2506.02685.

Zachary A. King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson, and Nathan E. Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1):D515–D522, 01 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1049. URL https://doi.org/10.1093/nar/gkv1049.

Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLOS Computational Biology*, 5(5):1–6, 05 2009. doi: 10.1371/journal.pcbi.1000385. URL https://doi.org/10.1371/journal.pcbi.1000385.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems!, 2019. URL https://arxiv.org/abs/1803.08475.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Vinay Satish Kumar and Costas D. Maranas. Growmatch: An automated method for reconciling in silico/in vivo growth predictions. *PLOS Computational Biology*, 5(3):1–13, 03 2009. doi: 10.1371/journal.pcbi.1000308. URL https://doi.org/10.1371/journal.pcbi.1000308.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21314–21328. Curran Associates, Inc., 2022. doi: 10.52202/068431-1549. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8636419dea1aa9fbd25fc4248e702da4-Paper-Conference.pdf.

Chenyi Li, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. Setpo: Set-level policy optimization for diversity-preserving llm reasoning, 2026. URL https://arxiv.org/abs/2602.01062.

Ping Li, Anshumali Shrivastava, Joshua Moore, and Arnd König. Hashing algorithms for large-scale learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/0a1bf96b7165e962e90cb14648c9462d-Paper.pdf.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Xiaoyi Liu, Hongpeng Yang, Chengwei Ai, Ruihan Dong, Yijie Ding, Qianqian Yuan, Jijun Tang, and Fei Guo. A generalizable framework for unlocking missing reactions in genome-scale metabolic networks using deep learning, 2024. URL https://arxiv.org/abs/2409.13259.

Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, 46(15):7542–7553, 09 2018. ISSN 0305-1048. doi: 10.1093/nar/gky537. URL https://doi.org/10.1093/nar/gky537.

Adil Mardinoglu and Bernhard Ø. Palsson. Genome-scale models in human metabologenomics. *Nature Reviews Genetics*, 26(2):123–140, Feb 2025. ISSN 1471-0064. doi: 10.1038/s41576-024-00768-0. URL https://doi.org/10.1038/s41576-024-00768-0.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL https://doi.org/10.1038/s41586-023-06735-9.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities, 2024. URL https://arxiv.org/abs/2405.20003.

Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, Mar 2010. ISSN 1546-1696. doi: 10.1038/nbt.1614. URL https://doi.org/10.1038/nbt.1614.

Jose C. Principe, Dongxin Xu, Qun Zhao, and John W. Fisher. Learning from examples with information theoretic criteria. *Journal of VLSI signal processing systems for signal, image and video technology*, 26(1):61–77, Aug 2000. ISSN 0922-5773. doi: 10.1023/A:1008143417156. URL https://doi.org/10.1023/A:1008143417156.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 08 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp358. URL https://doi.org/10.1093/bioinformatics/btp358.

Vinay Satish Kumar, Madhukar S. Dasika, and Costas D. Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):212, 2007. doi: 10.1186/1471-2105-8-212. URL https://doi.org/10.1186/1471-2105-8-212.

Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Francoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D20–D26, 01 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1112. URL https://doi.org/10.1093/nar/gkab1112.

Isabel Schober, Julia Koblitz, Joaquim Sardà Carbasse, Christian Ebeling, Marvin Leon Schmidt, Adam Podstawka, Rohit Gupta, Vinodh Ilangovan, Javad Chamanara, Jörg Overmann, and Lorenz Christian Reimer. Bacdive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research*, 53(D1):D748–D756, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae959. URL https://doi.org/10.1093/nar/gkae959.

Wheaton L. Schroeder and Rajib Saha. Optfill: A tool for infeasible cycle-free gapfilling of stoichiometric metabolic models. *iScience*, 23(1):100783, 2020. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2019.100783. URL https://www.sciencedirect.com/science/article/pii/S2589004219305280.

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

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgLZR4KvH.

Ashley G. Smart, Luis A. N. Amaral, and Julio M. Ottino. Cascading failure and robustness in metabolic networks. *Proceedings of the National Academy of Sciences*, 105(36):13223–13228, 2008. doi: 10.1073/pnas.0803571105. URL https://www.pnas.org/doi/abs/10.1073/pnas.0803571105.

Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, Jan 2010. ISSN 1750-2799. doi: 10.1038/nprot.2009.203. URL https://doi.org/10.1038/nprot.2009.203.

Ines Thiele, Nikos Vlassis, and Ronan M. T. Fleming. fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17):2529–2531, 09 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu321. URL https://doi.org/10.1093/bioinformatics/btu321.

Tomas Tokar, Chiara Pastrello, and Igor Jurisica. Gsoap: a tool for visualization of gene set over-representation analysis. *Bioinformatics*, 36(9):2923–2925, 05 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa001. URL https://doi.org/10.1093/bioinformatics/btaa001.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, Aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06221-2. URL https://doi.org/10.1038/s41586-023-06221-2.

Naganand Yadati, Vikram Nitin, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Talukdar. Nhp: Neural hypergraph link prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 1705–1714, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411870. URL https://doi.org/10.1145/3340531.3411870.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d60678e8f2ba9c540798ebbde31177e8-Paper.pdf.

Yanlong Zhao, Yixiao Chen, Yi Yu, Xiang Liu, Jiawen Du, Jun Wen, Quan Sun, Ren Wang, and Can Chen. A multi-way SMILES-based hypergraph inference network for metabolic model reconstruction. *Communications Biology*, 9(1):531, 2026. ISSN 2399-3642. doi: 10.1038/s42003-026-09761-1. URL https://doi.org/10.1038/s42003-026-09761-1.

Johannes Zimmermann, Christoph Kaleta, and Silvio Waschina. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology*, 22(1):81, 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02295-1. URL https://doi.org/10.1186/s13059-021-02295-1.

A ALGORITHM DETAILS OF QUOTIENTPO

This section gives the implementation details of QuotientPO. For compactness, let $x = (M_0, G, y)$ denote a training instance, while emphasizing that the repair policy $\pi_\theta(E \mid M_0, G)$ conditions only on the initial GEM and genomic representation. The phenotype labels y are used exclusively by the verifier. QuotientPO fine-tunes the supervised one-shot repair policy using four signals: verifier reward, kernelized quotient crowding (Quotient term), finite-panel successful-core discovery credit (QPI term), and a redundancy penalty outside the canonical repair core.

A.1 ONE-SHOT REPAIR POLICY

Let $\mathcal{A}(M_0)$ denote the legal signed reaction edits for an initial GEM, where (ADD, r) and (REMOVE, r) are treated as distinct actions. The policy predicts the complete unordered repair set in one shot. For each legal edit $a \in \mathcal{A}(M_0)$, the actor produces a logit $z_{\theta,a}$ and an operation-specific threshold $\tau_{\theta,o(a)}$, giving the inclusion probability $p_{\theta,a} = \mathrm{sigmoid}\left(z_{\theta,a} - \tau_{\theta,o(a)}\right)$. The complete edit-set likelihood is the product of the corresponding Bernoulli decisions,

$$\log \pi_\theta(E \mid M_0, G) = \sum_{a \in \mathcal{A}(M_0)} \left[\mathbb{I}[a \in E] \log p_{\theta,a} + \mathbb{I}[a \notin E] \log(1 - p_{\theta,a})\right].$$

Algorithm 1 Verifier-preserving canonicalization κ_y

Require: Repair set E, phenotype targets y, deterministic verifier FBA_y, maximum parallelism $B \geq 1$
Ensure: Canonical repair core $m = \kappa_y(E)$

$\boldsymbol{\sigma}^\star \leftarrow \mathrm{FBA}_y(E)$
$C \leftarrow \mathrm{sort}_{(\mathrm{operation},\mathrm{reaction})}(E)$
while $|C| > 0$ **do**
  $(c_1, \ldots, c_n) \leftarrow C$
  removed $\leftarrow$ false
  $i \leftarrow 1$
  $w \leftarrow 1$
  while $i \leq n$ **do**
    $r \leftarrow \min(n, i + w - 1)$
    for $j = i, \ldots, r$ **do**
      $S_j \leftarrow C \setminus \{c_j\}$
      $h_j \leftarrow$ SUBMIT($\mathrm{FBA}_y(S_j) = \boldsymbol{\sigma}^\star$) ▷ Asynchronous verifier check
    end for
    for $j = i, \ldots, r$ **do**
      $a_j \leftarrow$ AWAIT(h_j) ▷ Consume results in the fixed order
      if $a_j =$ true **then**
        $C \leftarrow C \setminus \{c_j\}$
        removed $\leftarrow$ true
        CANCELORDISCARD($h_{j+1:r}$)
        break
      end if
    end for
    if removed **then**
      break ▷ Restart after the committed deletion
    end if
    $i \leftarrow r + 1$
    $w \leftarrow \min(2w, B)$ ▷ Expand only after the entire batch fails
  end while
  if not removed **then**
    break ▷ No verifier-preserving singleton deletion remains
  end if
end while
return C

A.2 VERIFIER-PRESERVING CANONICALIZATION

For a sampled repair E, let $\boldsymbol{\sigma}_y(E) = (\mathbb{I}[\widehat{y}_i(M_E) = y_i])_{i=1}^N$ denote the complete phenotype-satisfaction vector across all N conditions. The quotient map κ_y (Section 4.1) is implemented

as a deterministic, verifier-preserving reduction (Algorithm 1) that systematically removes redundant edits while preserving $\boldsymbol{\sigma}_y(E)$. The resulting deletion-irreducible *repair core* $m = \kappa_y(E) \subseteq E$ defines the canonical functional representation of E under the verifier and induces the redundancy cost $|E| - |\kappa_y(E)|$. Because κ_y is deterministic, every repair is assigned to exactly one canonical core, yielding a well-defined quotient partition over the repair space.

A.3 Kernelized quotient crowding (Quotient term)

For each organism, we sample a rollout group $E_{1:L} \sim \pi_{\theta_{\text{old}}}(\cdot \mid M_0, G)$ and obtain their canonical cores $m_i = \kappa_y(E_i)$. All cores participate in quotient exploration, irrespective of whether they achieve complete phenotype recovery. As in Section 4.2, signed repair cores are compared via the kernel

$$k(m_i, m_j) = \exp\left(-\frac{1 - J(m_i, m_j)}{h}\right), \qquad J(m_i, m_j) = \frac{|m_i \cap m_j|}{|m_i \cup m_j|},$$

with $J(\varnothing, \varnothing) = 1$, and with ADD and REMOVE edits treated as distinct signed elements.

For $L \geq 2$, the empirical off-diagonal kernel concentration is

$$\widehat{I} = \frac{1}{L(L-1)} \sum_{a \neq b} k(m_a, m_b).$$

The candidate-level crowding cost is directly given by Eq. (3):

$$c_i = 2L \frac{\sum_{j \neq i} k(m_i, m_j)}{\sum_{a \neq b} k(m_a, m_b)} - 2.$$

For numerical robustness, the implemented score is clipped: $\widetilde{c}_i = \text{clip}(c_i, -8, 8)$. The analysis in Section 4.3 characterizes the unclipped crowding score c_i; implementation clipping is used only for finite-sample numerical stability. A positive crowding cost suppresses probability assigned to locally redundant cores, while a negative cost favors isolated cores.

A.4 Finite-panel successful-core discovery credit (QPI term)

Kernelized Rényi credit encourages geometric diversity over the entire space of repair cores, but finite rollout groups provide sparse signals regarding which distinct cores are functionally useful. We therefore supplement it with a finite-panel *quotient policy iteration* (QPI) credit that explicitly rewards successful-core discovery.

For the reported setting, each rollout group of $L = 32$ samples is deterministically repartitioned $R = 4$ times into panels of size $K = 8$. For a panel P, let $S_P = \sum_{j \in P} s_y(E_j)$ denote the number of successful samples, and $N_P(m) = \sum_{j \in P} \mathbb{I}[m_j = m]$ denote the multiplicity of core m within the panel. Each training organism maintains a bounded archive $\mathcal{H}_x$ of previously discovered successful cores. Let $\mathcal{H}_x^-$ denote the archive before processing the current rollout group, and let $\nu_i = \mathbb{I}[s_y(E_i) = 1,\ m_i \notin \mathcal{H}_x^-]$ indicate that sample i discovers a previously unseen successful core. The multiplicity of novel samples realizing core m in panel P is $N_P^{\text{new}}(m) = \sum_{j \in P} \nu_j \, \mathbb{I}[m_j = m]$.

For sample i in panel P, the credit decomposes into three terms:

$$a_i(P) = \begin{cases} s_y(E_i)/S_P, & S_P > 0, \\ 0, & S_P = 0, \end{cases} \qquad \text{successful coverage,}$$

$$b_i(P) = \frac{1}{K} \mathbb{I}[N_P(m_i) = 1], \qquad \text{within-panel core diversity,}$$

$$n_i(P) = \begin{cases} \dfrac{1}{K \, N_P^{\text{new}}(m_i)}, & \nu_i = 1, \\ 0, & \nu_i = 0, \end{cases} \qquad \text{archive novelty.}$$

The first term shares credit among successful candidates in the panel; the second rewards cores that occur uniquely within the current panel; and the third rewards successful cores not previously archived. Averaging over the R deterministic repartitions gives the detached QPI credit

$$u_i = \frac{1}{R} \sum_{r=1}^{R} \left[a_i(P_i^{(r)}) + \alpha_t \left(b_i(P_i^{(r)}) + n_i(P_i^{(r)}) \right) \right],$$

Algorithm 2 QuotientPO: quotient-space policy optimization

Require: Training set $\mathcal{D}$; initialized policy π_{θ_0}; verifier FBA; rollout size L; panel size K
Ensure: Trained policy π_{θ_T}

$\theta \leftarrow \theta_0$
Initialize successful-core archives $\{\mathcal{H}_x\}_{x\in\mathcal{D}}$
for $t = 1, \ldots, T$ **do**
  Select a training batch, including archive-guided context replay
  $\theta_{\text{old}} \leftarrow \theta$
  Set coefficients $(\beta_t, \lambda_t, \alpha_t, \eta_t)$
  for each training instance $x = (M_0, G, y)$ in the batch **do**
    Sample $E_{1:L} \sim \pi_{\theta_{\text{old}}}(\cdot \mid M_0, G)$
    for $i = 1, \ldots, L$ **do**
      Verify M_{E_i} to obtain $\text{FBA}_y(E_i)$, $\boldsymbol{\sigma}_y(E_i)$, and $\Phi_y(M_{E_i})$
      Compute terminal reward $R_y(E_i)$
      $m_i \leftarrow \textsc{Canonicalize}(E_i, y, \text{FBA}_y)$ ▷ Algorithm 1
      Compute redundancy $(|E| - |\kappa_y(E)|)$
    end for
    Compute group-relative advantages $A_{1:L}$
    Compute kernel crowding costs $\widetilde{c}_{1:L}$ from the valid cores
    Compute finite-panel QPI credits $u_{1:L}$ against the pre-update archive $\mathcal{H}_x^-$
    Insert newly discovered successful cores into $\mathcal{H}_x$
  end for
  Rescore all sampled edit sets under π_θ and compute policy ratios $\rho_i(\theta)$
  Form $\mathcal{J}_t(\theta)$
  Take one gradient update on θ
end for
return π_{θ_T}

where $P_i^{(r)}$ is the panel containing sample i in repartition r. Newly discovered successful cores are subsequently inserted into $\mathcal{H}_x$ (bounded to eight distinct cores per organism). The archive is used strictly to evaluate novelty and prioritize under-explored instances; archived repairs are never replayed as actions, and no distillation loss is applied.

A.5 Complete QuotientPO objective

Within each rollout group, the verifier reward is normalized using the GRPO baseline, $A_i = \frac{R_y(E_i)-\overline{R}}{\sigma_R+\varepsilon}$, with $A_i = 0$ when the group reward variance vanishes. For a sampled repair, the policy ratio is $\rho_i(\theta) = \frac{\pi_\theta(E_i|M_0,G)}{\pi_{\theta_{\text{old}}}(E_i|M_0,G)}$. The implemented maximization objective is

$$\begin{aligned}\mathcal{J}_t(\theta) = \frac{1}{L}\sum_{i=1}^{L} \min\Big(\rho_i(\theta)A_i,\ \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)A_i\Big) \\ - \frac{\beta_t}{L}\sum_{i=1}^{L}\rho_i(\theta)\ \text{sg}(\widetilde{c}_i) \\ + \frac{\lambda_t K}{L}\sum_{i=1}^{L}\rho_i(\theta)\ \text{sg}(u_i) - \frac{\eta_t}{L}\sum_{i=1}^{L}\rho_i(\theta)\ \text{sg}(|E| - |\kappa_y(E)|),\end{aligned}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The first term is verifier-guided GRPO. The second implements the kernelized quotient-diversity objective by shifting probability away from crowded repair cores. The third is the finite-panel QPI term that promotes successful-core discovery and retention. The final term suppresses surface edits removed by canonicalization. All verifier outcomes, advantages, core identities, crowding costs, QPI credits, and redundancy costs are detached from the actor graph. Gradients therefore flow only through the complete edit-set likelihood $\pi_\theta(E \mid M_0, G)$. PPO clipping is applied to the GRPO task term, whereas the quotient and auxiliary terms use the unclipped likelihood ratio. The full algorithm process is Algorithm 2.

In all experiments, the hyperparameters balance verifier reward optimization against mechanism-level diversity: we use a rollout group size of $L = 32$ with PPO clipping threshold $\epsilon = 0.2$. For exploration, we set the kernel crowding weight to $\beta = 0.02$ to penalize mechanism collapse, the finite-panel QPI discovery weight to $\lambda = 0.10$ with panel size $K = 8$ and internal novelty scale $\alpha = 0.05$, and the canonical redundancy penalty to $\eta = 0.001$ to gently suppress extraneous surface edits without impeding phenotype recovery.

B Theoretical Details

B.1 Formulation of GEM Modeling

A genome-scale metabolic model (GEM) is a directed, weighted bipartite graph $M = (\mathcal{C}, \mathcal{R}, S)$, where metabolites $\mathcal{C}$ and reactions $\mathcal{R}$ form the two node types and $S \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{R}|}$ encodes their stoichiometric edges. For metabolite c and reaction r, $S_{cr} < 0$ denotes consumption ($c \to r$), $S_{cr} > 0$ denotes production ($r \to c$), and $|S_{cr}|$ gives the stoichiometric coefficient. Reaction addition or deletion therefore corresponds to adding or removing a reaction node and its associated column of S. Flux balance analysis (FBA) assigns reaction fluxes v and computes

$$v^*(M) = \max_v v_{\text{bio}} \qquad \text{s.t.} \qquad Sv = 0, \quad \ell \le v \le u,$$

where $Sv = 0$ enforces steady-state mass balance and $[\ell, u]$ specifies reaction directionality and the reference environment. Wet-lab phenotypes provide organism-specific functional constraints through $\widehat{y}_z(M)$, and $v^*(M) > \delta$ enforces a non-negligible biomass-supporting feasible flux.

B.2 Theory of Representation-Invariant Exploration

Notation and setup. Given an initial draft metabolic network M_0, genomic context G, and experimental phenotype target y, the repair policy generates a raw repair candidate $E \in \mathcal{F}$ (a set of reaction additions or deletions). The deterministic verifier-preserving reduction $\kappa_y : \mathcal{F} \to \mathcal{F}$ maps each raw repair E to an operational repair core $m = \kappa_y(E)$. We denote by $\mathcal{M}_y = \kappa_y(\mathcal{F}) \subseteq \mathcal{F}$ the quotient mechanism space of all such canonical repairs. Throughout this subsection, we fix an arbitrary training instance (M_0, G, y) and suppress this conditioning from the raw policy π_θ and the induced quotient policy $q_{\theta,y}$ to streamline notation, writing $\pi(E)$ and $q(m)$ respectively. For each mechanism $m \in \mathcal{M}_y$, we define:

- the *repair fiber* $\Omega_m = \{E \in \mathcal{F} : \kappa_y(E) = m\}$, which collects all syntactically distinct reaction edits mapped to the same canonical repair mechanism m.
- the *fiber multiplicity* $n_m = |\Omega_m|$, quantifying the syntactic degeneracy (the number of redundant representations) of mechanism m.
- the *quotient policy marginal* $q(m) = \sum_{E\in\Omega_m} \pi(E)$, representing the aggregate probability assigned by the policy to deploying mechanism m.

Because κ_y preserves the complete phenotype response, any raw edit $E \in \Omega_m$ achieves an identical post-repair phenotype evaluation $\Phi_y(M_E) = \Phi_y(M_m)$. Consequently, the mechanism-level class utility $U_y(m) = \Phi_y(M_E), E \in \Omega_m$, is well defined and invariant to the choice of raw representative within Ω_m. Since the candidate edit space $\mathcal{F}$ is formed by reaction additions and deletions over a finite reaction database, it is finite. Consequently, the quotient mechanism space $\mathcal{M}_y$ and every repair fiber Ω_m are finite. We denote by $\Delta(\mathcal{F})$ and $\Delta(\mathcal{M}_y)$ the probability simplices over candidate edits and repair mechanisms, respectively, and assume an exploration coefficient $\beta > 0$ throughout.

Lemma B.1 (Entropy decomposition over repair fibers)**.** *For any probability distribution $\pi \in \Delta(\mathcal{F})$, define the conditional distribution $\pi(E \mid m) = \pi(E)/q(m)$ for $E \in \Omega_m$ whenever $q(m) > 0$. Then*

$$H_{\text{act}}(\pi) = H(q) + \sum_{m\in\mathcal{M}_y} q(m) H(\pi(\cdot \mid m)), \tag{5}$$

where zero-mass fibers contribute zero. Moreover,

$$H(\pi(\cdot \mid m)) \le \log n_m, \tag{6}$$

with equality, for $q(m) > 0$, if and only if $\pi(\cdot \mid m)$ is uniform on Ω_m.

Proof. Because $\mathcal{F} = \bigsqcup_{m\in\mathcal{M}_y} \Omega_m$, for every $E \in \Omega_m$ with $q(m) > 0$ we may write $\pi(E) = q(m)\pi(E \mid m)$. Therefore,

$$\begin{aligned} H_{\mathrm{act}}(\pi) &= -\sum_{m\in\mathcal{M}_y}\sum_{E\in\Omega_m} q(m)\pi(E \mid m)\log\big(q(m)\pi(E \mid m)\big) \\ &= -\sum_m q(m)\log q(m)\sum_{E\in\Omega_m}\pi(E \mid m) - \sum_m q(m)\sum_{E\in\Omega_m}\pi(E \mid m)\log\pi(E \mid m) \\ &= H(q) + \sum_m q(m)H(\pi(\cdot \mid m))\,, \end{aligned}$$

which proves Eq. (5).

For the second claim, let u_m denote the uniform distribution on Ω_m, so $u_m(E) = 1/n_m$. Non-negativity of the Kullback–Leibler divergence gives

$$0 \le D_{\mathrm{KL}}(\pi(\cdot \mid m)\,\|\,u_m) = \sum_{E\in\Omega_m}\pi(E \mid m)\log\frac{\pi(E \mid m)}{1/n_m} = -H(\pi(\cdot \mid m)) + \log n_m.$$

Hence $H(\pi(\cdot \mid m)) \le \log n_m$. Equality in the KL inequality holds if and only if $\pi(\cdot \mid m) = u_m$, proving the equality condition. □

Proposition B.2 (Action-space multiplicity bias)**.** *For a fixed training instance, let $\{\Omega_m\}_{m\in\mathcal{M}_y}$ partition the feasible repair space into verifier-equivalent classes, with common utility $U_y(m) = \Phi_y(M_E)$ for $E \in \Omega_m$ and multiplicity $n_m = |\Omega_m|$. Optimizing $\mathcal{J}_{\mathrm{act}}$ over the unrestricted probability simplex induces $q^\star_{\mathrm{act}}(m) \propto n_m \exp(U_y(m)/\beta)$. Equivalently, action entropy introduces an implicit class-level utility $\beta \log n_m$ in addition to biological utility $U_y(m)$.*

Proof. For a fixed raw repair distribution π, phenotype utility depends only on its quotient marginal:

$$\mathbb{E}_{E\sim\pi}[\Phi_y(M_E)] = \sum_{m\in\mathcal{M}_y}\sum_{E\in\Omega_m}\pi(E)U_y(m) = \sum_{m\in\mathcal{M}_y} q(m)U_y(m). \tag{7}$$

Combining Eq. (7) with Lemma B.1, the fixed-instance action-entropy objective can be written as

$$\mathcal{J}_{\mathrm{act}}(\pi) = \sum_m q(m)U_y(m) + \beta H(q) + \beta\sum_m q(m)H(\pi(\cdot \mid m))\,. \tag{8}$$

Now fix an arbitrary quotient marginal q. The first two terms in Eq. (8) are independent of the conditional distributions inside the fibers. By Eq. (6), the final term is maximized independently in every positive-mass fiber by

$$\pi^\star(E \mid m) = \frac{1}{n_m}, \qquad E \in \Omega_m.$$

Therefore optimization over the raw simplex $\Delta(\mathcal{F})$ is equivalent to

$$\max_{\pi\in\Delta(\mathcal{F})}\mathcal{J}_{\mathrm{act}}(\pi) = \max_{q\in\Delta(\mathcal{M}_y)}\left\{\sum_m q(m)U_y(m) + \beta H(q) + \beta\sum_m q(m)\log n_m\right\}. \tag{9}$$

Define

$$a_m = n_m\exp\left(\frac{U_y(m)}{\beta}\right), \qquad Z_{\mathrm{act}} = \sum_{m\in\mathcal{M}_y} a_m,$$

and

$$q^\star_{\mathrm{act}}(m) = \frac{a_m}{Z_{\mathrm{act}}}.$$

Since every fiber is nonempty, $n_m \ge 1$, and all utilities are finite, $q^\star_{\mathrm{act}}(m) > 0$ for every m. Using $\log a_m = \log n_m + U_y(m)/\beta$, the objective in Eq. (9) becomes

$$\begin{aligned} &\sum_m q(m)U_y(m) + \beta\sum_m q(m)\log n_m - \beta\sum_m q(m)\log q(m) \\ &\quad= \beta\sum_m q(m)\log\frac{a_m}{q(m)} \\ &\quad= \beta\log Z_{\mathrm{act}} - \beta D_{\mathrm{KL}}(q\,\|\,q^\star_{\mathrm{act}})\,. \end{aligned} \tag{10}$$

Because KL divergence is nonnegative and vanishes only when its two arguments coincide, Eq. (10) proves that the unique quotient marginal maximizing the objective is

$$q^\star_{\mathrm{act}}(m) = \frac{n_m \exp(U_y(m)/\beta)}{\sum_{m' \in \mathcal{M}_y} n_{m'} \exp(U_y(m')/\beta)}. \tag{11}$$

Furthermore, because this optimum assigns positive mass to every fiber, Lemma B.1 requires the unique optimal within-fiber conditional to be uniform. Thus the corresponding raw-policy optimum satisfies

$$\pi^\star(E) = \frac{q^\star_{\mathrm{act}}(m)}{n_m} = \frac{\exp(U_y(m)/\beta)}{Z_{\mathrm{act}}}, \qquad E \in \Omega_m.$$

Hence the multiplicity factor arises only when raw action probabilities are aggregated at the class level: each representative receives the same Boltzmann weight for a fixed utility, and a class containing n_m representatives accumulates n_m such contributions. Equivalently, action entropy contributes the implicit class utility $\beta \log n_m$ claimed in the proposition. □

Lemma B.3 (Surjectivity of the quotient marginalization map)**.** *Let $T : \Delta(\mathcal{F}) \to \Delta(\mathcal{M}_y)$ be defined by*

$$[T\pi](m) = \sum_{E \in \Omega_m} \pi(E),$$

where $\Omega_m = \{E \in \mathcal{F} : \kappa_y(E) = m\}$. Then T is surjective.

Proof. Because $\mathcal{M}_y = \kappa_y(\mathcal{F})$, every $m \in \mathcal{M}_y$ has at least one representative $r(m) \in \mathcal{F}$ satisfying $\kappa_y(r(m)) = m$. For any $q \in \Delta(\mathcal{M}_y)$, define

$$\pi_q(E) = \begin{cases} q(m), & E = r(m) \text{ for some } m \in \mathcal{M}_y, \\ 0, & \text{otherwise.} \end{cases}$$

Since distinct fibers are disjoint, the selected representatives are distinct and $\pi_q \in \Delta(\mathcal{F})$. Moreover,

$$[T\pi_q](m) = \sum_{E \in \Omega_m} \pi_q(E) = \pi_q(r(m)) = q(m).$$

Hence $T\pi_q = q$, proving that T is surjective. □

Remark B.4 (Biological and probabilistic intuition). In probabilistic terms, the operator T simply formalizes marginalization over functionally equivalent repairs: it aggregates the probabilities of all distinct reaction edit sets mapped to the same canonical repair mechanism, yielding the total probability $q(m)$ of deploying mechanism m. Biologically, surjectivity ensures that the quotient mechanism space $\Delta(\mathcal{M}_y)$ has no "ghost states": any target distribution over metabolic repair mechanisms is physically realizable by an underlying reaction-level policy.

Proof of Theorem 4.1. Let $T : \Delta(\mathcal{F}) \to \Delta(\mathcal{M}_y)$ be the quotient marginalization operator from Lemma B.3.

Sufficiency. Suppose there exists $\overline{\mathfrak{D}} : \Delta(\mathcal{M}_y) \to \mathbb{R}$ such that $\mathfrak{D}(\pi) = \overline{\mathfrak{D}}(T\pi)$, for all $\pi \in \Delta(\mathcal{F})$. If two raw policies π, π' induce the same quotient marginal, $T\pi = T\pi'$, then

$$\mathfrak{D}(\pi) = \overline{\mathfrak{D}}(T\pi) = \overline{\mathfrak{D}}(T\pi') = \mathfrak{D}(\pi').$$

Hence $\mathfrak{D}$ is invariant to all probability redistributions that leave the quotient law unchanged.

Necessity. Conversely, assume that

$$T\pi = T\pi' \quad \Longrightarrow \quad \mathfrak{D}(\pi) = \mathfrak{D}(\pi')$$

for all $\pi, \pi' \in \Delta(\mathcal{F})$. By Lemma B.3, for every $q \in \Delta(\mathcal{M}_y)$ there exists at least one lift $\pi_q \in \Delta(\mathcal{F})$ satisfying $T\pi_q = q$. Define

$$\overline{\mathfrak{D}}(q) = \mathfrak{D}(\pi_q). \tag{12}$$

We first verify that this definition is independent of the chosen lift. If π_q and $\widetilde{\pi}_q$ both satisfy $T\pi_q = T\widetilde{\pi}_q = q$, then the assumed invariance gives $\mathfrak{D}(\pi_q) = \mathfrak{D}(\widetilde{\pi}_q)$. Thus Eq. (12) defines a well-defined functional on the entire quotient simplex.

Now take an arbitrary $\pi \in \Delta(\mathcal{F})$ and set $q = T\pi$. Both π and the selected lift π_q induce the same quotient law. Therefore,

$$\mathfrak{D}(\pi) = \mathfrak{D}(\pi_q) = \overline{\mathfrak{D}}(q) = \overline{\mathfrak{D}}(T\pi),$$

which proves the required factorization.

It remains to prove uniqueness. Suppose two functionals $\overline{\mathfrak{D}}_1$ and $\overline{\mathfrak{D}}_2$ satisfy $\mathfrak{D} = \overline{\mathfrak{D}}_1 \circ T = \overline{\mathfrak{D}}_2 \circ T$. For any $q \in \Delta(\mathcal{M}_y)$, surjectivity of T provides a lift π_q with $T\pi_q = q$. Hence

$$\overline{\mathfrak{D}}_1(q) = \overline{\mathfrak{D}}_1(T\pi_q) = \mathfrak{D}(\pi_q) = \overline{\mathfrak{D}}_2(T\pi_q) = \overline{\mathfrak{D}}_2(q).$$

Therefore $\overline{\mathfrak{D}}_1 = \overline{\mathfrak{D}}_2$ on $\Delta(\mathcal{M}_y)$. This establishes both existence and uniqueness. □

Remark B.5 (Preservation under a representative lift). Beyond existence and uniqueness, the quotient factorization preserves basic analytical properties of the exploration functional on the probability simplex. Fix, for each $m \in \mathcal{M}_y$, an arbitrary representative $r(m) \in \Omega_m$, so that $\kappa_y(r(m)) = m$, as in Lemma B.3. This representative selection induces a lift $L_r : \Delta(\mathcal{M}_y) \to \Delta(\mathcal{F})$ defined by

$$[L_r q](E) = \begin{cases} q(m), & E = r(m) \text{ for some } m \in \mathcal{M}_y, \\ 0, & \text{otherwise.} \end{cases}$$

By construction, $TL_r q = q$ for every $q \in \Delta(\mathcal{M}_y)$. Moreover, L_r is linear:

$$L_r(\alpha q_1 + (1-\alpha)q_2) = \alpha L_r q_1 + (1-\alpha) L_r q_2, \qquad \alpha \in [0,1].$$

Therefore, for the quotient functional $\bar{\mathcal{D}}$ characterized by Theorem 4.1,

$$\bar{\mathcal{D}}(q) = \mathcal{D}(L_r q),$$

and continuity and convexity or concavity of $\mathcal{D}$ restrict directly to $\bar{\mathcal{D}}$ through this linear lift.

Corollary B.6 (Quotient Shannon entropy removes the multiplicity factor)**.** *For a fixed training instance, unrestricted maximization of* $\sum_{m \in \mathcal{M}_y} q(m) U_y(m) + \beta H(q)$ *over* $q \in \Delta(\mathcal{M}_y)$ *has the unique solution*

$$q^\star_{\mathrm{quot}}(m) = \frac{\exp(U_y(m)/\beta)}{\sum_{m' \in \mathcal{M}_y} \exp(U_y(m')/\beta)}. \tag{13}$$

Proof. Define

$$Z_{\mathrm{quot}} = \sum_{m \in \mathcal{M}_y} \exp\left(\frac{U_y(m)}{\beta}\right), \qquad q^\star_{\mathrm{quot}}(m) = \frac{\exp(U_y(m)/\beta)}{Z_{\mathrm{quot}}}.$$

For arbitrary $q \in \Delta(\mathcal{M}_y)$,

$$\begin{aligned} \sum_m q(m) U_y(m) + \beta H(q) &= \beta \sum_m q(m) \log \frac{\exp(U_y(m)/\beta)}{q(m)} \\ &= \beta \log Z_{\mathrm{quot}} - \beta D_{\mathrm{KL}}\left(q \,\|\, q^\star_{\mathrm{quot}}\right). \end{aligned} \tag{14}$$

The KL divergence is nonnegative and equals zero if and only if $q = q^\star_{\mathrm{quot}}$. Therefore Eq. (13) is the unique maximizer. □

Comparing this expression with the action-space optimum, $q^\star_{\mathrm{act}}(m) \propto n_m \exp(U_y(m)/\beta)$, shows that quotient Shannon entropy removes the representation-volume factor n_m exactly.

Scope of the results. Lemma B.1 applies to any distribution on the finite feasible repair space under the partition induced by κ_y. The Gibbs solutions characterize unrestricted simplex optima and serve as variational references for the parameterized policy. The factorization theorem holds for any fixed verifier-preserving κ_y, establishing rigorous representation invariance directly over the operational mechanism classes induced by the verifier across diverse biological hypotheses.

B.3 Collision Sparsity and Kernel Rényi Estimation

Notation and setup. This subsection provides the formal results supporting **Sec. 4.2**, adopting the quotient setup from Appendix B.2. Unless stated otherwise, $m_1, \dots, m_L$ are drawn IID from the quotient distribution q on the finite space $\mathcal{M}_y$. We write

$$C(q) = \sum_{m \in \mathcal{M}_y} q(m)^2 = \exp[-H_2(q)]$$

for the collision concentration, where H_2 denotes categorical Rényi entropy of order two.

We first characterize the collision-free regime as an extreme case of the broader finite-sample coarseness of exact-count statistics.

Proposition B.7 (Sparse-collision barrier)**.** *For L independently sampled repair cores $m_1, \dots, m_L \sim q_{\theta,y}$, the pairwise collision count $N_{\mathrm{col}} = \sum_{i<j} \mathbf{1}[m_i = m_j]$ satisfies $\mathbb{P}(N_{\mathrm{col}} > 0) \le \mathbb{E}[N_{\mathrm{col}}] = \binom{L}{2} \sum_{m \in \mathcal{M}_y} q_{\theta,y}(m)^2$. Consequently, along any sequence of instances and policies for which $L^2 \sum_{m \in \mathcal{M}_y} q_{\theta,y}(m)^2 \to 0$, the collision probability vanishes ($\mathbb{P}(N_{\mathrm{col}} > 0) \to 0$), leaving all sampled candidates with identical zero exploration credit.*

Proof. Writing $N_{\mathrm{col}} = \sum_{1 \le i < j \le L} \mathbf{1}[m_i = m_j]$, independence gives

$$\mathbb{P}(m_i = m_j) = \sum_{m \in \mathcal{M}_y} q(m)^2 = C(q).$$

Hence, by linearity of expectation,

$$\mathbb{E}[N_{\mathrm{col}}] = \binom{L}{2} C(q).$$

Since N_{col} is nonnegative and integer valued, Markov's inequality yields

$$\mathbb{P}(N_{\mathrm{col}} > 0) \le \min\left\{1,\ \binom{L}{2} C(q)\right\}.$$

More formally, along any sequence of quotient laws q_n and rollout sizes L_n satisfying

$$L_n^2 C(q_n) \longrightarrow 0,$$

the preceding bound implies $\mathbb{P}(N_{\mathrm{col}} > 0) \to 0$.

If the group is collision-free, its empirical distribution assigns mass $1/L$ to each sampled core, so both its plug-in Shannon entropy and categorical Rényi-2 entropy equal $\log L$. After removing any candidate i, the corresponding entropy satisfies

$$H_{-i} = \log(L-1) \qquad \text{for all } i.$$

Since this holds for every i, the mean across the group is $\bar{H}_- = \frac{1}{L} \sum_{j=1}^{L} H_{-j} = \log(L-1)$, yielding a leave-one-out entropy contrast

$$\bar{H}_- - H_{-i} = \log(L-1) - \log(L-1) = 0 \tag{15}$$

identically for all $i \in \{1, \dots, L\}$. The same argument applies to any symmetric equality-only statistic whose leave-one-out value depends solely on collision patterns. Hence, under collision-free rollout groups, exact-match leave-one-out evaluation provides zero exploration credit to all sampled candidates. □

We now formalize the kernel relaxation used to replace exact collisions. Each repair core is identified with its canonical finite set of signed reaction edits, so equality of cores coincides with equality of their signed-edit sets. Define

$$J(m, m') = \frac{|m \cap m'|}{|m \cup m'|}, \qquad J(\varnothing, \varnothing) = 1,$$

and, for fixed bandwidth $h > 0$,

$$k_y(m, m') = \exp\left(-\frac{1 - J(m, m')}{h}\right).$$

The kernel rule is shared across instances and the subscript y indicates its restriction to the instance-specific quotient space $\mathcal{M}_y$.

Proposition B.8 (Validity and consistency of the signed-Jaccard relaxation)**.** *The kernel k_y is positive semidefinite and satisfies*

$$e^{-1/h} \le k_y(m, m') \le 1, \qquad k_y(m, m) = 1.$$

Moreover, on the finite quotient space $\mathcal{M}_y$, $k_y(m, m')$ converges uniformly to the exact-match kernel $\mathbf{1}[m = m']$ as $h \to 0$.

Define the population kernel concentration and its associated kernelized quadratic Rényi functional by

$$V_k(q) = \mathbb{E}_{m,m' \sim q}[k_y(m, m')], \qquad H_{2,k}(q) = -\log V_k(q). \tag{16}$$

Then $C(q) \le V_k(q) \le 1$, $0 \le H_{2,k}(q) \le H_2(q)$, and as $h \to 0$,

$$V_k(q) \longrightarrow C(q), \qquad H_{2,k}(q) \longrightarrow H_2(q).$$

Proof. We first establish positive semidefiniteness of the Jaccard similarity. Let $\mathcal{U}_\pm$ denote the finite universe of signed reaction edits, and let ζ be a uniformly random permutation of $\mathcal{U}_\pm$. For each nonempty core m, let $b_\zeta(m)$ be its minimum element under ζ, assigning the empty core a distinguished sentinel symbol.

For any finite collection $m_1, \dots, m_n$, the matrix

$$G^\zeta_{ab} = \mathbf{1}[b_\zeta(m_a) = b_\zeta(m_b)]$$

is a Gram matrix of one-hot vectors and is therefore positive semidefinite. For two nonempty cores, the standard minwise-hashing identity gives

$$\mathbb{E}_\zeta[G^\zeta_{ab}] = \frac{|m_a \cap m_b|}{|m_a \cup m_b|} = J(m_a, m_b),$$

with the sentinel convention giving the stated values involving the empty core (Broder, 1997; Li et al., 2011). Hence the Jaccard Gram matrix Q is positive semidefinite.

For the same collection, the Gram matrix K of k_y can be expressed via the Hadamard exponential as

$$K = e^{-1/h} \exp^\odot(Q/h) = e^{-1/h} \sum_{r=0}^{\infty} \frac{Q^{\odot r}}{h^r r!}.$$

By the Schur product theorem, every Hadamard power $Q^{\odot r}$ is positive semidefinite. Thus every finite partial sum is positive semidefinite; since the positive-semidefinite cone is closed and the series converges in finite dimension, $K \succeq 0$. The bounds follow directly from $0 \le J \le 1$.

For the small-bandwidth limit, the case $|\mathcal{M}_y| = 1$ is immediate. If $|\mathcal{M}_y| > 1$, finiteness of $\mathcal{M}_y$ and the canonical signed-edit representation guarantee

$$\gamma_y = \min_{m \ne m'} \big(1 - J(m, m')\big) > 0.$$

Therefore,

$$\sup_{m \ne m'} k_y(m, m') \le e^{-\gamma_y/h} \longrightarrow 0 \quad \text{as } h \to 0, \tag{17}$$

while $k_y(m, m) = 1$ for all $h > 0$, proving uniform convergence to the exact-match kernel.

Finally, expanding $V_k(q)$ into diagonal and off-diagonal terms gives

$$V_k(q) = \sum_m q(m)^2 + \sum_{m \ne m'} q(m) q(m') k_y(m, m') = C(q) + \sum_{m \ne m'} q(m) q(m') k_y(m, m') \ge C(q),$$

while $V_k(q) \le 1$. Taking negative logarithms yields $0 \le H_{2,k}(q) \le H_2(q)$.

Note that the total off-diagonal probability mass satisfies

$$\sum_{m \neq m'} q(m)q(m') = \left(\sum_m q(m)\right)^2 - \sum_m q(m)^2 = 1 - C(q).$$

Consequently, factoring out the off-diagonal supremum yields

$$\begin{aligned} 0 \leq V_k(q) - C(q) &= \sum_{m \neq m'} q(m)q(m')k_y(m, m') \\ &\leq \left(\sup_{m \neq m'} k_y(m, m')\right) \sum_{m \neq m'} q(m)q(m') \\ &= (1 - C(q)) \sup_{m \neq m'} k_y(m, m') \\ &\leq \sup_{m \neq m'} k_y(m, m') \leq e^{-\gamma_y/h}. \end{aligned} \tag{18}$$

As $h \to 0$, this upper bound vanishes by Eq. (17), so $V_k(q) \to C(q)$. Because $\mathcal{M}_y$ is finite and nonempty, Cauchy–Schwarz ensures $C(q) = \sum_m q(m)^2 \geq 1/|\mathcal{M}_y| > 0$; thus continuity of $-\log$ on $(0, 1]$ implies $H_{2,k}(q) \to H_2(q)$. □

Theorem B.9 (Finite-group estimation and consistency)**.** *For $L \geq 2$, define*

$$\widehat{V}_L = \frac{1}{L(L-1)} \sum_{i \neq j} k_y(m_i, m_j), \qquad \widehat{H}_{2,k} = -\log \widehat{V}_L, \qquad J_L(q) = \mathbb{E}[\widehat{H}_{2,k}].$$

Then $\widehat{V}_L$ is an unbiased estimator of the population kernel concentration,

$$\mathbb{E}[\widehat{V}_L] = V_k(q),$$

and, for every fixed bandwidth $h > 0$,

$$0 \leq J_L(q) - H_{2,k}(q) \leq B_L(h), \qquad B_L(h) = \frac{(e^{1/h} - 1)^2}{4L}.$$

Hence, for fixed h, $J_L(q)$ converges uniformly over quotient laws to $H_{2,k}(q)$ at rate $O(L^{-1})$.

Proof. The statistic $\widehat{V}_L$ is the order-two U-statistic

$$\widehat{V}_L = \frac{2}{L(L-1)} \sum_{i<j} k_y(m_i, m_j),$$

so independence immediately gives $\mathbb{E}[\widehat{V}_L] = V_k(q)$. Since $-\log$ is convex, Jensen's inequality implies

$$J_L(q) = \mathbb{E}[-\log \widehat{V}_L] \geq -\log V_k(q) = H_{2,k}(q).$$

Let $X = \widehat{V}_L$, $V = V_k(q)$, and recall that $X, V \in [e^{-1/h}, 1]$. For $f(v) = -\log v$, $f''(v) = 1/v^2$. By Taylor's theorem, for some ξ between X and V,

$$f(X) = f(V) + f'(V)(X - V) + \frac{f''(\xi)}{2}(X - V)^2.$$

Since $\xi \in [e^{-1/h}, 1]$, we have $f''(\xi) \leq e^{2/h}$, so

$$0 \leq f(X) - f(V) - f'(V)(X - V) \leq \frac{e^{2/h}}{2}(X - V)^2.$$

Taking expectations and using $\mathbb{E}[X] = V$ yields

$$0 \leq J_L(q) - H_{2,k}(q) \leq \frac{e^{2/h}}{2} \operatorname{Var}(\widehat{V}_L). \tag{19}$$

To bound the variance, let $g(m, m') = k_y(m, m')$, $g_1(m) = \mathbb{E}_{m'}[g(m, m')] - V$, and define the degenerate second-order projection $g_2(m, m') = g(m, m') - V - g_1(m) - g_1(m')$. Let $\zeta_1 = \mathbb{E}[g_1(m)^2]$, $\zeta_2 = \mathbb{E}[g_2(m, m')^2]$. The Hoeffding decomposition for an order-two U-statistic gives

$$\mathrm{Var}(\widehat{V}_L) = \frac{4}{L}\zeta_1 + \frac{2}{L(L-1)}\zeta_2,$$

while orthogonality of the projections implies $2\zeta_1 + \zeta_2 = \mathrm{Var}(g(m, m'))$. Therefore, for $L \geq 2$,

$$\mathrm{Var}(\widehat{V}_L) \leq \frac{2}{L}(2\zeta_1 + \zeta_2) = \frac{2}{L}\,\mathrm{Var}(k_y(m, m')) \leq \frac{(1 - e^{-1/h})^2}{2L}, \tag{20}$$

where the last inequality follows from Popoviciu's inequality since $k_y(m, m') \in [e^{-1/h}, 1]$. Substituting Eq. (20) into Eq. (19) gives

$$J_L(q) - H_{2,k}(q) \leq \frac{e^{2/h}(1 - e^{-1/h})^2}{4L} = \frac{(e^{1/h} - 1)^2}{4L}.$$

Therefore, for every fixed $h > 0$,

$$\sup_q |J_L(q) - H_{2,k}(q)| = O(L^{-1}),$$

which establishes uniform consistency over quotient laws. □

Scope of the result. Theorem B.9 establishes uniform $O(L^{-1})$ consistency for fixed bandwidth h. The distribution-free bound is a conservative worst-case guarantee and is not intended as a tight characterization of finite rollout groups. The limit $H_{2,k}(q) \to H_2(q)$ is a population result, while signed-Jaccard similarity quantifies structural overlap rather than biological equivalence.

B.4 QUOTIENTPO OPTIMIZATION THEORY

This subsection establishes the theoretical role of the quotient-diversity term in Eq. (4). We deliberately separate this analysis from the verifier-shaped reward, GRPO clipping, and auxiliary discovery and retention mechanisms, which are optimization choices detailed in Appendix A. Our goal here is twofold: **to characterize the candidate-level crowding cost in Eq. (3), and to relate its policy update to the population kernel Rényi functional introduced in Appendix B.3.**

Notation and setup. In this subsection, fix a training instance and suppress its conditioning. For a rollout group $m_{1:L}$, recall the group kernel concentration $\widehat{V}_L = \frac{1}{L(L-1)} \sum_{a \neq b} k_y(m_a, m_b)$ from Theorem B.9, and define the candidate-level average kernel crowding

$$\bar{k}_i = \frac{1}{L-1} \sum_{j \neq i} k_y(m_i, m_j), \qquad \text{so that} \quad \frac{1}{L} \sum_{i=1}^{L} \bar{k}_i = \widehat{V}_L. \tag{21}$$

Proposition B.10 (Candidate crowding from leave-one-out kernel entropy)**.** *For $L \geq 3$, let $\widehat{H}_{-i} = -\log \widehat{V}_{-i}$ be the kernel Rényi entropy of the group after removing candidate i, where*

$$\widehat{V}_{-i} = \frac{L\widehat{V}_L - 2\bar{k}_i}{L-2}.$$

Then the leave-one-out contrast $\Delta_i = \widehat{H}_{2,k} - \widehat{H}_{-i}$ satisfies

$$\Delta_i = \log\left(1 - \frac{2(\bar{k}_i - \widehat{V}_L)}{(L-2)\widehat{V}_L}\right). \tag{22}$$

As $L \to \infty$, the first-order Taylor expansion around the group mean $\widehat{V}_L$ yields

$$\Delta_i = -\frac{c_i}{L-2} + \mathcal{O}\left(\frac{1}{L^2}\right), \tag{23}$$

where $c_i = \frac{2\bar{k}_i}{\widehat{V}_L} - 2$ is the crowding cost in Eq. (3)*. Moreover,*

$$\sum_{i=1}^{L} c_i = 0, \quad c_i > 0 \iff \bar{k}_i > \widehat{V}_L, \quad c_i < 0 \iff \bar{k}_i < \widehat{V}_L.$$

Proof. By the kernel symmetry $k_y(m, m') = k_y(m', m)$, removing candidate i deletes $2\sum_{j\neq i} k_y(m_i, m_j) = 2(L-1)\bar{k}_i$ off-diagonal mass from the total group mass $L(L-1)\widehat{V}_L$. The remaining $(L-1)(L-2)$ ordered pairs have average concentration

$$\widehat{V}_{-i} = \frac{L(L-1)\widehat{V}_L - 2(L-1)\bar{k}_i}{(L-1)(L-2)} = \frac{L\widehat{V}_L - 2\bar{k}_i}{L-2} = \widehat{V}_L - \frac{2(\bar{k}_i - \widehat{V}_L)}{L-2}. \tag{24}$$

By Proposition B.8, $k_y(m, m') \geq e^{-1/h} > 0$ holds uniformly, guaranteeing that $\widehat{V}_{-i} \geq e^{-1/h} > 0$ and the logarithm is strictly well-defined. Dividing Eq. (24) by $\widehat{V}_L$ and evaluating the leave-one-out entropy contrast gives

$$\Delta_i = \widehat{H}_{2,k} - \widehat{H}_{-i} = -\log \widehat{V}_L - (-\log \widehat{V}_{-i}) = \log\left(\frac{\widehat{V}_{-i}}{\widehat{V}_L}\right) = \log\left(1 - \frac{2(\bar{k}_i - \widehat{V}_L)}{(L-2)\widehat{V}_L}\right),$$

which establishes Eq. (22).

Next, let $x_i = \frac{2(\bar{k}_i - \widehat{V}_L)}{(L-2)\widehat{V}_L}$. Since $|\bar{k}_i - \widehat{V}_L| \leq 1 - e^{-1/h}$ and $\widehat{V}_L \geq e^{-1/h}$, we have $x_i = \mathcal{O}(L^{-1})$. Applying the Taylor expansion $\log(1-x) = -x + \mathcal{O}(x^2)$ around $x = 0$ yields

$$\Delta_i = -x_i + \mathcal{O}(x_i^2) = -\frac{2(\bar{k}_i - \widehat{V}_L)}{(L-2)\widehat{V}_L} + \mathcal{O}(L^{-2}) = -\frac{c_i}{L-2} + \mathcal{O}(L^{-2}),$$

which establishes Eq. (23) and shows that the candidate crowding cost $c_i = \frac{2\bar{k}_i}{\widehat{V}_L} - 2$ governs the leading leave-one-out entropy rate.

Finally, using the definition $\sum_{i=1}^{L} \bar{k}_i = L\widehat{V}_L$ from Eq. (3), the group sum of the costs satisfies

$$\sum_{i=1}^{L} c_i = \frac{2}{\widehat{V}_L}\sum_{i=1}^{L} \bar{k}_i - 2L = \frac{2}{\widehat{V}_L}(L\widehat{V}_L) - 2L = 0. \tag{25}$$

Rewriting the cost as $c_i = \frac{2}{\widehat{V}_L}(\bar{k}_i - \widehat{V}_L)$ with strictly positive normalizer $\widehat{V}_L > 0$, the signs satisfy $c_i > 0 \iff \bar{k}_i > \widehat{V}_L$ and $c_i < 0 \iff \bar{k}_i < \widehat{V}_L$ directly by affine monotonicity. □

The preceding result gives a finite-group interpretation of the implemented cost. We next show that the same quantity converges to the score coefficient of the population kernel Rényi objective. Define the population kernel density of a repair core m by

$$d_q(m) = \mathbb{E}_{m'\sim q}[k_y(m, m')],$$

so that

$$V_k(q) = \mathbb{E}_{m\sim q}[d_q(m)].$$

Define the corresponding population crowding coefficient

$$c_q^k(m) = 2\frac{d_q(m)}{V_k(q)} - 2. \tag{26}$$

Theorem B.11 (Population score identity and local consistency)**.** *Assume that π_θ is differentiable with locally fixed support and integrable score $\mathbb{E}_{E\sim\pi_\theta}[\|g_\theta(E)\|] < \infty$, and that κ_y and k_y remain fixed during the policy update. Let $g_\theta(E) = \nabla_\theta \log \pi_\theta(E \mid M_0, G)$ and let $q_{\theta,y}$ denote the distribution of repair cores $m = \kappa_y(E)$ induced by $E \sim \pi_\theta$ (abbreviated as q). Then:*

(i) ***Score Identity:*** *The population kernel Rényi functional satisfies*

$$\nabla_\theta H_{2,k}(q_{\theta,y}) = -\mathbb{E}_{E\sim\pi_\theta}\left[c^k_{q_{\theta,y}}(\kappa_y(E))\, g_\theta(E)\right]. \tag{27}$$

(ii) ***Local Consistency:*** *For the empirical crowding cost c_i in Eq.* (23)*, we have $|c_i - c_q^k(m_i)| \to 0$ in probability as $L \to \infty$. Consequently, for the detached quotient-diversity surrogate $\mathcal{J}_{k,L}(\theta) = -\frac{\beta}{L}\sum_{i=1}^{L} \rho_i(\theta)\mathrm{sg}(c_i)$ in Eq.* (4)*, its expected gradient at the behavior policy satisfies*

$$\lim_{L\to\infty} \mathbb{E}\left[\nabla_\theta \mathcal{J}_{k,L}(\theta)\big|_{\theta=\theta_{\mathrm{old}}}\right] = \beta\nabla_\theta H_{2,k}(q_{\theta_{\mathrm{old}},y}). \tag{28}$$

Thus the QuotientPO kernel term is a locally consistent score surrogate for population quotient diversity as the rollout group grows.

Proof. Using the deterministic core map $m = \kappa_y(E)$, the population concentration can be written as

$$V_k(q_{\theta,y}) = \mathbb{E}_{E,E' \sim \pi_\theta}\left[k_y(\kappa_y(E), \kappa_y(E'))\right].$$

Differentiating the two IID policy factors gives

$$\nabla_\theta V_k(q_{\theta,y}) = 2\,\mathbb{E}_{E\sim\pi_\theta}\left[d_q(\kappa_y(E))\, g_\theta(E)\right].$$

Since $H_{2,k}(q) = -\log V_k(q)$, we have

$$\nabla_\theta H_{2,k}(q_{\theta,y}) = -\frac{2}{V_k(q_{\theta,y})}\mathbb{E}_{E\sim\pi_\theta}\left[d_q(\kappa_y(E))\, g_\theta(E)\right].$$

The score identity $\mathbb{E}_{E\sim\pi_\theta}[g_\theta(E)] = 0$ allows baseline subtraction without altering the expectation, yielding

$$\nabla_\theta H_{2,k}(q_{\theta,y}) = -\mathbb{E}_{E\sim\pi_\theta}\left[\left(2\frac{d_q(\kappa_y(E))}{V_k(q_{\theta,y})} - 2\right) g_\theta(E)\right] = -\mathbb{E}_{E\sim\pi_\theta}\left[c_q^k(\kappa_y(E))\, g_\theta(E)\right],$$

which proves Eq. (27).

Next, we establish the local consistency of the empirical surrogate. Conditional on m_i, the remaining $L-1$ cores are IID draws from q. By the weak law of large numbers, $\bar{k}_i \to d_q(m_i)$ in probability as $L \to \infty$. In addition, the bounded order-two U-statistic satisfies $\widehat{V}_L \to V_k(q)$ in probability. Because $V_k(q) \geq e^{-1/h} > 0$ is bounded away from zero, the continuous mapping theorem implies that $|c_i - c_q^k(m_i)| \to 0$ in probability as $L \to \infty$.

Finally, evaluating the surrogate gradient at the behavior policy gives $\nabla_\theta \rho_i(\theta)|_{\theta=\theta_{\text{old}}} = g_{\theta_{\text{old}}}(E_i)$. By the exchangeability of the IID rollout sample $(E_1, \ldots, E_L)$,

$$\mathbb{E}\left[\nabla_\theta \mathcal{J}_{k,L}(\theta)\big|_{\theta=\theta_{\text{old}}}\right] = -\beta\mathbb{E}\left[\frac{1}{L}\sum_{i=1}^{L} c_i g_{\theta_{\text{old}}}(E_i)\right] = -\beta\mathbb{E}\left[c_1^k g_{\theta_{\text{old}}}(E_1)\right].$$

Since $k_y \in [e^{-1/h}, 1]$, the statistics $\bar{k}_1$ and $\widehat{V}_L$ both lie in $[e^{-1/h}, 1]$, so c_1^k is uniformly bounded in L by $2(e^{1/h}-1)$. Combined with score integrability $\mathbb{E}[\|g_{\theta_{\text{old}}}(E_1)\|] < \infty$, the dominated convergence theorem implies

$$\lim_{L\to\infty} \mathbb{E}\left[c_1^k g_{\theta_{\text{old}}}(E_1)\right] = \mathbb{E}\left[c_q^k(\kappa_y(E_1))\, g_{\theta_{\text{old}}}(E_1)\right].$$

Substituting this limit back into the surrogate expectation and applying Eq. (27) directly yields Eq. (28). □

Relation to deployment-time success and mechanism coverage. Because k is already used for the similarity kernel, let K denote the number of independently generated repairs in the evaluation metric success@K. For a fixed evaluation instance x, let $s_x(E) \in \{0, 1\}$ denote the evaluation success predicate and define the single-draw success probability

$$Z_x(\theta) = \mathbb{P}_{E\sim\pi_\theta(\cdot|x)}[s_x(E) = 1].$$

Proposition B.12 (Successful quotient mass determines Success@K)**.** *For $K \geq 1$ IID repairs sampled from $\pi_\theta(\cdot \mid x)$, the probability that at least one repair succeeds is*

$$P_K(x;\theta) = 1 - \big(1 - Z_x(\theta)\big)^K. \tag{29}$$

Therefore $P_K(x;\theta)$ is monotonically increasing in $Z_x(\theta)$, and strictly increasing whenever $Z_x(\theta) < 1$. If the evaluation success predicate is invariant within repair cores, i.e., there exists s_x^{core} such that $s_x(E) = s_x^{\text{core}}(\kappa_y(E))$, then

$$Z_x(\theta) = \sum_{m\in\mathcal{M}_y} q_{\theta,y}(m)\, s_x^{\text{core}}(m). \tag{30}$$

Thus success@K is determined by the total quotient probability mass assigned to successful repair cores.

Proof. A single repair fails with probability $1 - Z_x(\theta)$. Under IID sampling, the probability that all K repairs fail is therefore

$$\big(1 - Z_x(\theta)\big)^K.$$

Taking the complement proves Eq. (29). Moreover,

$$\frac{\partial P_K(x;\theta)}{\partial Z_x} = K(1 - Z_x)^{K-1},$$

which is positive for $Z_x < 1$.

Under core invariance of the success predicate,

$$\begin{aligned} Z_x(\theta) &= \sum_E \pi_\theta(E \mid x) s_x^{\mathrm{core}}(\kappa_y(E)) \\ &= \sum_{m \in \mathcal{M}_y} s_x^{\mathrm{core}}(m) \sum_{E:\kappa_y(E)=m} \pi_\theta(E \mid x) \\ &= \sum_{m \in \mathcal{M}_y} q_{\theta,y}(m) s_x^{\mathrm{core}}(m), \end{aligned}$$

which proves Eq. (30). □

Proposition B.13 (Expected distinct successful-core coverage)**.** *Let*

$$\mathcal{M}_x^{\mathrm{succ}} = \{m \in \mathcal{M}_y : s_x^{\mathrm{core}}(m) = 1\}$$

denote the successful repair cores for instance x. For K IID repairs $E_1, \ldots, E_K \sim \pi_\theta(\cdot \mid x)$, let

$$D_K = |\{\kappa_y(E_t) : 1 \le t \le K,\ s_x(E_t) = 1\}|$$

be the number of distinct successful cores discovered within the K samples. Then

$$\mathbb{E}[D_K] = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} \Big[1 - (1 - q_{\theta,y}(m))^K\Big].$$

Thus, unlike Success@K, distinct successful-core coverage depends not only on the total successful mass

$$Z_x(\theta) = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} q_{\theta,y}(m),$$

but also on how that mass is distributed across successful cores.

Proof. For each successful core $m \in \mathcal{M}_x^{\mathrm{succ}}$, define

$$I_m = \mathbf{1}\left[m \in \{\kappa_y(E_1), \ldots, \kappa_y(E_K)\}\right].$$

Then

$$D_K = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} I_m.$$

Because each sampled repair induces core m with probability $q_{\theta,y}(m)$, independence gives

$$\Pr(I_m = 0) = (1 - q_{\theta,y}(m))^K.$$

Hence

$$\mathbb{E}[I_m] = 1 - (1 - q_{\theta,y}(m))^K.$$

Applying linearity of expectation yields

$$\mathbb{E}[D_K] = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} \Big[1 - (1 - q_{\theta,y}(m))^K\Big].$$

□

Corollary B.14 (Successful-mass equalization improves mechanism coverage)**.** *Fix an instance x, a sampling budget $K > 1$, and a set of S successful repair cores. For a fixed total successful mass*

$$\sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} q_{\theta,y}(m) = Z,$$

the expected distinct successful-core coverage

$$\mathbb{E}[D_K] = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} \left[1 - (1 - q_{\theta,y}(m))^K\right]$$

is Schur-concave in the successful-core probability vector. Consequently, any mean-preserving redistribution that equalizes probability mass across successful cores weakly increases expected coverage. The maximum is attained at the uniform allocation

$$q_{\theta,y}(m) = \frac{Z}{S},$$

yielding

$$\mathbb{E}[D_K] = S\left[1 - \left(1 - \frac{Z}{S}\right)^K\right].$$

Meanwhile, Success@K remains

$$1 - (1 - Z)^K,$$

and therefore depends only on the total successful mass Z.

Proof. For $K > 1$, define

$$f_K(p) = 1 - (1 - p)^K.$$

Since

$$f_K''(p) = -K(K-1)(1-p)^{K-2} \le 0,$$

f_K is concave on $[0, 1]$. Therefore,

$$\mathbb{E}[D_K] = \sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} f_K(q_{\theta,y}(m))$$

is a symmetric Schur-concave function of the successful-core probability vector. Hence, any mean-preserving equalization of this vector weakly increases expected coverage. Under the fixed-sum constraint

$$\sum_{m \in \mathcal{M}_x^{\mathrm{succ}}} q_{\theta,y}(m) = Z,$$

the maximum is attained at the uniform allocation

$$q_{\theta,y}(m) = \frac{Z}{S}.$$

Substituting this value gives

$$\mathbb{E}[D_K] = S\left[1 - \left(1 - \frac{Z}{S}\right)^K\right].$$

By Proposition B.12, Success@K is

$$1 - (1 - Z)^K,$$

which depends only on the total successful mass Z. □

Interpretation and scope. Propositions B.12–B.13 reveal two complementary dimensions of deployment-time performance. Success@K is governed solely by the total probability mass Z_x assigned to successful repair cores, whereas distinct-core coverage additionally rewards how broadly this mass is distributed across successful mechanisms. At fixed Z_x, equalizing probability across successful cores monotonically improves expected mechanism coverage, with uniform allocation achieving the optimum. Thus, successful-mass concentration and successful-mass allocation control different aspects of test-time scaling: increasing Z_x raises the probability of obtaining a valid repair, while distributing that mass across distinct successful cores increases the number of mechanisms recoverable under the same sampling budget.

This characterization motivates quotient-space exploration at the mechanism level. By shifting probability away from crowded regions of the repair-core distribution, the kernelized quotient objective promotes broader allocation across candidate mechanisms, while verifier-guided optimization concentrates probability on functionally successful regions. Their combination therefore aligns naturally with the two quantities identified above: successful probability mass and successful-core coverage. Empirically, QuotientPO improves both, yielding higher Success@K and greater distinct successful-core recovery under the same sampling budget.

C Dataset and Benchmark Construction

C.1 Construction Pipeline

Construction of initial GEM and proteome matching. BacDive phenotype records were linked to NCBI (Sayers et al., 2022) proteomes via prefix/suffix-normalized INSDC assembly accessions. Retaining matches with available `protein.faa(.gz)` files yielded 12,303 proteome inputs across 11,238 unique assemblies. We constructed candidate GEMs under four CarveMe settings: default, standard Gram-positive, standard Gram-negative, and a custom archaeal universe (augmenting CarveMe with archaeal biomass and ATP maintenance), producing 36,855 SBML Level 3/FBC2 GEMs as uncurated starting hypotheses. Concurrently, raw BacDive records were filtered for unambiguous binary phenotypes and mapped to BiGG identifiers. Retaining organisms with at least one mapped phenotype label produced 11,275 organisms and 37,864 organism–model entries (as multiple BacDive strains may share an assembly). This cohort was filtered and bifurcated into two downstream pipelines: 3,978 phenotype-consistent GEMs under the selected assays were identified as candidates for self-supervised pre-training, 28,842 canonical error-bearing GEMs formed the post-training cohort. Records sharing the same normalized accession were assigned to the same split to prevent genome-level leakage.

Table 4: **Nutrient-growth assay coverage.**

Assay category	BacDive families
Carbon source	glucose, maltose, sucrose, glycerol, acetate, lactate, pyruvate, succinate, citrate, glutamate
Nitrogen source	ammonium, urea, nitrate, nitrite
Sulfur source	sulfate
Phosphorus source	phosphate

Construction of BacDive targets. Binary BacDive nutrient-utilization phenotypes were mapped to BiGG metabolites across 16 nutrient families (Table 4), with stereospecific assays for lactate and glutamate yielding 18 total assay configurations. In each assay, the tested nutrient s serves as the sole elemental source for its macro-element class (C, N, S, or P) against an element-depleted minimal background to preclude metabolic bypass, with flux bounds set to 10 for the tested substrate, 20 for oxygen, and 1,000 for non-limiting element-free inorganics and water. To ensure consistent boundary conditions across draft GEMs, native exchanges were replaced by standardized boundary reactions (excluded from the editable action space), and all unsupplemented exchange uptakes were strictly disabled. Growth is predicted when optimal biomass flux exceeds 10^{-3}. For multi-assay labels, positive predictions require growth in at least one assay, while negative predictions require absence across all. Requiring at least one decidable target and one initial phenotype error yields 21,615 post-training eligible GEMs across all partitions.

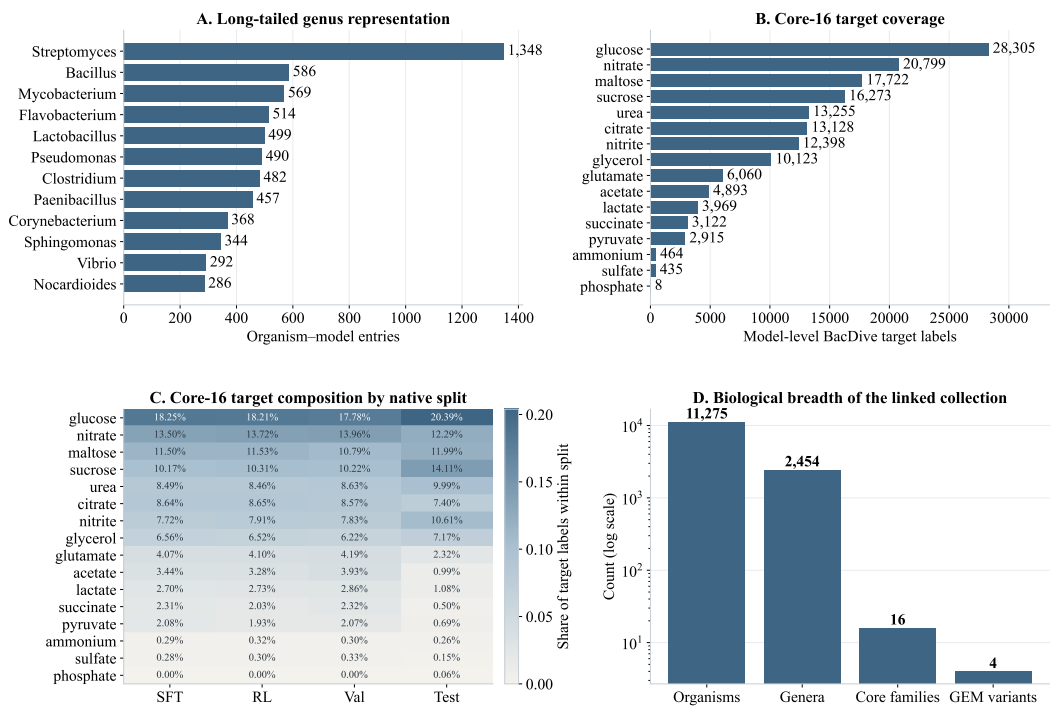


Figure 9: **Dataset characterization and split diagnostics.**

Certified repair supervision. Teacher repair sets were generated offline via mixed-integer optimization (Appendix D.1) and verified by FBA under identical assay conditions. We distinguish *complete repairs* (correcting all phenotype errors) from *partial repairs* (strictly reducing net errors while preserving initially correct phenotypes). Partial repairs provide action-level supervision for SFT-A (masking actions that temporarily degrade phenotype accuracy), whereas complete repairs supervise set-level prediction in SFT-B. Training states are constructed from unordered subsets of verified repairs to eliminate arbitrary edit-order assumptions, retaining multiple valid solutions with per-GEM weight normalization. This supervision covers 1,426 GEMs (10,837 states) for SFT-A and 1,078 GEMs (40,151 rows) for SFT-B after augmentation. Crucially, phenotype labels, error diagnostics, and teacher edits are never exposed as policy inputs.

Data partitioning and leakage prevention. Dataset partitions were fixed before post-training eligibility filtering. Entries connected by normalized organism identity, proteome identity, GEM provenance, or model checksum were merged by transitive closure into 8,568 indivisible identity components. A genus-held-out test set contains *Bacillus*, *Clostridium*, *Lactobacillus*, *Mycobacterium*, and *Streptomyces*. The remaining identity components were deterministically assigned to the SFT, RL, and validation pools: $28{,}842 = 12{,}223_{\text{SFT}} + 12{,}226_{\text{RL}} + 1{,}445_{\text{validation}} + 2{,}948_{\text{test}}$. After phenotype-assay filtering and certified-teacher selection, SFT-A retained 1,426/177 training/validation GEMs and SFT-B retained 1,078/117. The RL split retained 9,196/1,087 training/validation GEMs, while the final repair test set contained 2,212 GEMs. Eligible models in the SFT pool without certified supervision were left unused rather than reassigned to RL. All augmented states inherit the partition of their source GEM. Beyond the component-level partitioning of the post-training cohort, the 3,978 clean GEMs allocated to self-supervised pre-training were strictly quarantined from the outset. Pre-training explicitly excludes all five held-out evaluation genera and shares zero normalized INSDC accessions or proteomes with the test split. Auditing confirmed zero biological identity overlap across the entire lifecycle (Pre-training, SFT, RL, Validation and Test), preventing models from acquiring prior familiarity with test-split metabolic networks.

C.2 Dataset Characterization and Quality Control

Figure 9 summarizes the biological breadth and split-level properties of the post-training collection. The linked dataset is taxonomically long-tailed, with 11,275 model-linked organisms spanning 2,454

genera and four GEM reconstruction variants (Fig. 9A,D). The core-16 supervision covers 153,869 model-level BacDive target groups across the 28,842 retained GEMs (Fig. 9B). These counts measure phenotype-target coverage rather than repair errors: the same organism contributes separately through different GEM variants, and distinct compound-level annotations remain independent targets. Across the native SFT, RL, and validation pools, major phenotype partitions are broadly consistent, while the sealed genus-held-out test set exhibits the expected distribution shifts induced by taxonomic holdout (Fig. 9C). Together, these diagnostics establish broad biological coverage, explicit bidirectional phenotype supervision, and a leakage-controlled evaluation shift.

C.3 Benchmark Validation

The sealed benchmark (2,212 GEMs) evaluates generalization across biological taxa, reconstruction templates, and phenotype error modes (Figure 10). Taxonomically, it spans five held-out genera: *Streptomyces* (926), *Bacillus* (443), *Lactobacillus* (403), *Mycobacterium* (266), and *Clostridium* (174) (Fig. 10A). It balances four draft reconstruction templates: 697 default, 517 archaeal, 507 Gram-negative, and 491 Gram-positive models (Fig. 10B). Across the 16 core metabolites, initial errors display dominant false negatives alongside non-trivial false positives (Fig. 10C), requiring policies to coordinate reaction additions and deletions rather than relying on monotonic gap-filling.

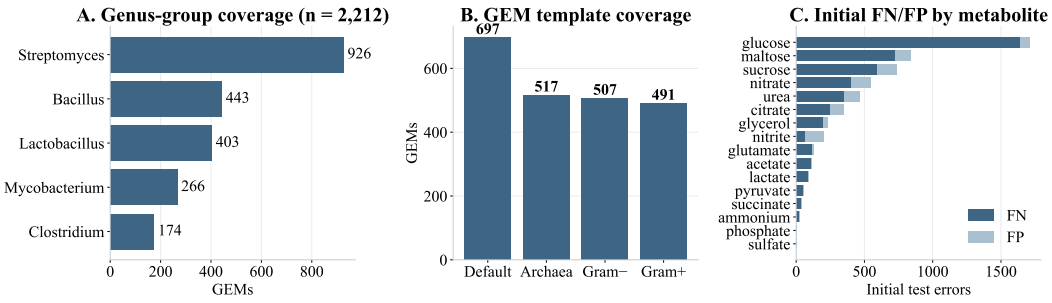


Figure 10: **Benchmark validation.**

D Baseline Details

This section details the baselines that require bespoke algorithmic reformulation to operate over metabolic networks: the phenotype-informed MILP oracle (§D.1) and the object-level GFlowNet (§D.2), both adapted from external paradigms to phenotype-supervised GEM repair. The remaining baselines build directly upon our neural architecture and training pipeline: imitation models (*SFT-Soft* and *SFT-Hard*) are described alongside architectural details in Appendix E.3, while reinforcement learning policies (Vanilla GRPO, MaxEnt, and QuotientPO) are detailed in Appendix A and E.4.

D.1 MILP

Mathematical formulation and phenotype-guided separation. We construct a phenotype-informed mixed-integer linear programming (MILP) oracle that couples discrete reaction edits with continuous intracellular fluxes. For an initial GEM M_0 with reaction set $\mathcal{R}_0$ and a candidate reaction universe $\mathcal{U}$, a binary decision vector $x \in \{0,1\}^{|\mathcal{U}|}$ specifies toggled reactions, defining the repair edit set $E(x) = (A(x), D(x))$ with additions $A(x) = \{r \in \mathcal{U} \setminus \mathcal{R}_0 : x_r = 1\}$ and deletions $D(x) = \{r \in \mathcal{R}_0 \cap \mathcal{U} : x_r = 1\}$, such that $\|x\|_1 = |E(x)|$. For each nutrient assay s, biomass flux is evaluated via flux balance analysis (FBA), formulated as a continuous linear programming (LP) problem:

$$v_s^*(x) = \max_v \left\{ c^\top v : Sv = 0,\ \ell_s(x) \le v \le u_s(x) \right\}.$$

Given target phenotype labels $y_i \in \{0,1\}$ and mapped assay sets $\mathcal{A}_i$, the predicted phenotype is $\widehat{y}_i(x) = \mathbf{1}[\max_{s\in\mathcal{A}_i} v_s^*(x) > 10^{-3}]$, yielding the total phenotype error $\mathrm{Err}(x) = \sum_i \mathbf{1}[\widehat{y}_i(x) \neq y_i]$. The oracle accesses test-time labels to lexicographically minimize phenotype errors before edit cardinality: $\min_x(\mathrm{Err}(x), \|x\|_1)$. We implement this hierarchy via binary violation indicators

$z_i \in \{0,1\}$:

$$\min_{x,z} \quad \sum_{r\in\mathcal{U}} x_r + (|\mathcal{U}|+1)\sum_i z_i,$$

where $z_i = 0$ requires $\widehat{y}_i(x) = y_i$. Because $\|x\|_1 \le |\mathcal{U}|$, the weight $(|\mathcal{U}|+1)$ guarantees that resolving any phenotype strictly dominates toggling candidate reactions. Rather than embedding continuous fluxes into a large big-M formulation, the discrete master optimizes (x, z) and generates phenotype-guided cuts through FBA separation. At iteration k, a violated phenotype induces

$$D_H(x, x^{(k)}) + z_i \ge 1, \qquad D_H(x, x^{(k)}) = \sum_{r\in H} \left| x_r - x_r^{(k)} \right|,$$

where $H \subseteq U$ concentrates the separation on phenotype-relevant reactions identified from LP dual information for failed positive phenotypes or active flux support for failed negative phenotypes. These phenotype-guided restricted-support cuts concentrate the finite-budget search on functionally relevant edit configurations, while every retained candidate is independently certified by complete FBA evaluation under the original assay conditions.

Solver implementation and verification pipeline. Each GEM receives a 600-second search budget using double-precision HiGHS (Huangfu & Hall, 2015) with one thread per GEM, LP feasibility tolerance $1e{-}8$, MIP integrality tolerance $1e{-}9$, and zero requested MIP gap. The search allocates 120 seconds to greedy subset repair (up to 1,000 separation iterations) and the remainder to lexicographic search (up to 4,000 iterations). Each candidate is independently evaluated by the complete runtime FBA verifier, which determines its phenotype error under the original assay conditions. Across the 2,212 sealed test GEMs, we retain the FBA-verified candidate with the best lexicographic objective $(\mathrm{Err}(x), \|x\|_1)$ encountered within the search budget, with unimproved cases retaining M_0. This phenotype-informed oracle therefore provides a strong label-access baseline for evaluating amortized repair policies.

D.2 GFlowNet

We adapt GFlowNet (Bengio et al., 2021) to one-shot GEM repair as an object-level diversity baseline. It is to test whether distributing probability across high-reward raw repair sets, without identifying equivalent repairs, is sufficient to obtain the multi-solution benefits targeted by QuotientPO.

Object-level GFlowNet objective. Let π_0 denote the frozen SFT policy used to initialize RL. For a complete repair set E, we define the reward-tilted target distribution

$$p^\star(E \mid M_0, G, y) \propto \pi_0(E \mid M_0, G)^\gamma \exp(R_y(E)/\tau)\,, \tag{31}$$

where τ controls reward temperature and γ controls the strength of the SFT base distribution. The target therefore favors high-reward repairs while retaining a structured prior over plausible edits. Crucially, its support is defined over raw edit sets: two repairs remain distinct GFlowNet objects whenever their edit sets differ, even if they canonicalize to the same repair core. This provides a direct object-space counterpart to the quotient-space exploration of QuotientPO.

One-shot adaptation. Unlike standard trajectory-based GFlowNets, our policy predicts the entire unordered repair set in a single forward pass via independent Bernoulli decoders, requiring no intermediate states, action orderings, or learned backward policies. Flow consistency under target (31) then requires policy probabilities to match target weights up to a partition constant. For a rollout group of L candidates, defining the log flow residual $z_i = R_y(E_i)/\tau + \gamma \log \pi_0(E_i) - \log \pi_\theta(E_i)$ and eliminating the unknown instance offset yields the log-partition variance (LPV) loss,

$$\mathcal{L}_{\mathrm{LPV}} = \frac{1}{L}\sum_{i=1}^{L}(z_i - \bar{z})^2,$$

which enforces reward-proportional probabilities without explicit partition-function estimation.

SFT anchor and task adaptation. A uniform base distribution over complete edit sets concentrates probability on unrealistically large cardinalities (averaging half of all available actions), causing unanchored exploration ($\gamma = 0$) to rapidly degenerate into thousands of spurious ADD edits (Table 5). The SFT anchor in Eq. (31) resolves this combinatorial mismatch by tempering the target distribution with the structured prior π_0^γ while keeping policy initialization matched. Comparing

completed candidates across $\gamma \in \{0.4, 0.6, 1.0\}$ under identical training recipes, the intermediate anchor $\gamma = 0.6$ achieves the highest validation Success@8 (9.38%) while maintaining sparse, plausible repairs (10.98 ADD edits); we therefore fix $\gamma = 0.6$ before running replicated seeds without consulting test data. In contrast, an identity-uniform diagnostic that matched expected edit counts but removed reaction-specific SFT preferences collapsed toward nearly empty repairs (0.065 edits, 0.276% validation success), confirming that GFlowNet requires biological reaction-level preferences rather than mere cardinality constraints.

Table 5: **SFT-anchor selection for the GFlowNet baseline.** Anchored variants use the same final training recipe and complete all 864 updates. The unanchored run is incomplete and is included only to show object-space cardinality degeneration. Success values are validation Success@8 (%).

Anchor γ	Best	Best step	Last	Last ADD
0 (no anchor)	7.91	15	5.80	7367.57
0.4	8.46	745	7.64	23.05
0.6	**9.38**	864	**9.38**	10.98
1.0	8.00	715	7.27	6.05

Implementation details. To isolate object-level exploration under matched task optimization, we train the baseline using the joint objective $\mathcal{L}_{\text{GFlowNet}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{LPV}}\mathcal{L}_{\text{LPV}}$, combining standard verifier-guided GRPO on group-normalized advantages with LPV distribution matching on raw rewards. GFlowNet uses no quotient-space mechanisms during training, with repair cores computed strictly post hoc for evaluation. The baseline shares the exact policy backbone, candidate action masks, FBA verifier, and frozen SFT initialization (π_0) with all other RL methods. We use rollout size $L = 32$, anchor strength $\gamma = 0.6$, reward temperature $\tau = 1$, LPV weight $\lambda_{\text{LPV}} = 0.02$, and PPO clipping threshold 0.2. Each update batches 32 organisms with 32 fresh on-policy samples for a fixed budget of 864 updates (matching three epochs) via AdamW (learning rate $2e{-}6$, weight decay $1e{-}4$, global gradient clipping at 1.0). Across seeds 42, 43, and 44, test metrics are reported from the final checkpoint matching Table 1.

E Experimental Details

Our training follows a three-stage *Pretrain–SFT–RL* pipeline. Since GEM repair is evaluated only through non-differentiable terminal feedback, direct RL from random initialization is ineffective in the large and sparse reaction space. We therefore first pretrain the model to learn GEM structure and reaction compatibility, then use phenotype-supervised SFT to learn a repair prior, and finally optimize complete edit sets with verifier-grounded RL. These stages correspond to *structural literacy*, *repair prior learning*, and *joint consequence optimization*.

E.1 Base model architecture.

All post-training methods use the same pretrained backbone (Fig. 11). Each GEM is represented as a metabolite–reaction bipartite graph with 19-dimensional node features and 3-dimensional stoichiometric edge features, and encoded by a 12-layer edge-aware GINE (Hu et al., 2020) with hidden dimension 768, residual connections, LayerNorm, and dropout. Reaction and metabolite nodes use type embeddings together with shared biochemical identity embeddings, while separate attention pooling over the two node types produces a global GEM representation. Candidate ADD reactions are encoded by a universal reaction encoder over 28,301 reactions, which combines a shared reaction-ID embedding, pooled substrate/product metabolite embeddings, and reaction-level structural features into the same 768-dimensional latent space. Organism-level proteomic context is extracted by a frozen ESM2-t12-35M [2] encoder, whose 480-dimensional protein embeddings are deterministically reduced to 256 dimensions, attention-pooled over at most 512 proteins, and mapped to 768 dimensions by a learnable gated MLP before fusion with the global GEM representation. Based on this shared representation, separate ADD and REMOVE heads score absent and existing reactions, respectively: the ADD head combines graph–reaction matching, pairwise interactions, and local metabolite compatibility, whereas the REMOVE head scores interactions between

[2] https://huggingface.co/facebook/esm2_t12_35M_UR50D

the global GEM state and reaction-node embeddings. During post-training, the ESM-2 backbone remains frozen, while the GINE encoder, universal reaction encoder, proteome conditioner, and ADD/REMOVE scoring heads are jointly fine-tuned.

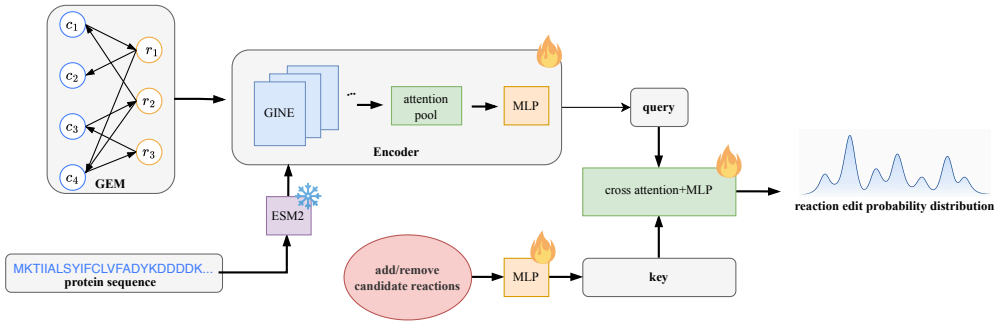


Figure 11: **Base model architecture.**

E.2 Pretraining Details

Dataset construction. Of the 3,978 candidate GEMs, 143 were excluded by the pretraining leakage-control criteria, leaving 3,806 training GEMs and 29 held-out validation GEMs. We pretrain the GEM backbone through single-reaction denoising on these 3,806 clean GEMs. Each training example contains exactly one corruption: either a valid reaction is removed and recovered by the ADD head, or an absent reaction is inserted and identified by the REMOVE head. The two tasks are sampled with equal probability. To prevent synthetic shortcuts, injected reactions must be absent from the original GEM, avoid duplicate stoichiometric signatures, and connect metabolites already represented in the graph; artifact-prone metadata features are masked during pretraining. Each epoch contains 125,000 examples, comprising 100,000 dynamically generated corruptions and 25,000 fixed atomic examples, for a total of five million training exposures over 40 epochs. This objective trains reaction-level compatibility without imposing a synthetic prior over downstream repair-set cardinality or co-occurrence.

Model training and evaluation. Each pre-training example contains exactly one corrupted reaction, corresponding to either an ADD or REMOVE target. For action type $a \in \{\mathrm{ADD}, \mathrm{REMOVE}\}$, let $\mathcal{C}_i^a$ denote the legal candidate set for example i and y_i^a its target reaction. ADD uses up to 4,096 sampled absent reactions, always including the positive target and frequency-based hard negatives, whereas REMOVE scores all reactions currently present in the corrupted GEM. Given logits s_{ir}^a, the reaction-identification loss is $\mathcal{L}_{\mathrm{CE}}^a = -\frac{1}{|\mathcal{B}_a|}\sum_{i\in\mathcal{B}_a} \log \frac{\exp(s_{iy_i^a}^a)}{\sum_{r\in\mathcal{C}_i^a}\exp(s_{ir}^a)}$. We further emphasize difficult confounders using the eight highest-scoring incorrect candidates, $\mathcal{H}_i^a = \mathrm{Top8}\{s_{ir}^a : r \in \mathcal{C}_i^a \setminus \{y_i^a\}\}$, with pairwise ranking loss $\mathcal{L}_{\mathrm{rank}}^a = \frac{1}{|\mathcal{B}_a|}\sum_{i\in\mathcal{B}_a}\frac{1}{|\mathcal{H}_i^a|}\sum_{r\in\mathcal{H}_i^a}\mathrm{softplus}\left(s_{ir}^a - s_{iy_i^a}^a\right)$. The complete objective is

$$\mathcal{L}_{\mathrm{pretrain}} = \mathcal{L}_{\mathrm{CE}}^{\mathrm{ADD}} + \mathcal{L}_{\mathrm{CE}}^{\mathrm{REMOVE}} + 0.2\left(\mathcal{L}_{\mathrm{rank}}^{\mathrm{ADD}} + \mathcal{L}_{\mathrm{rank}}^{\mathrm{REMOVE}}\right).$$

Models are trained with AdamW using a learning rate of $1e{-}4$, weight decay $1e{-}5$, global batch size 384 across eight GPUs[3], BF16 mixed precision, and gradient clipping at 1.0. Validation uses 580 fixed corruptions generated from 29 held-out GEMs, including multi-edit corruptions, and scores the full reaction vocabulary. Checkpoint selection uses the macro-average of ADD and REMOVE recall on the operation-specific single-target subsets.

Scaling of model size and data size. We study scaling along both model capacity and cumulative pretraining exposure. Specifically, we compare 256(hidden size)×4(layer), 384×6, and 768×12 GINE backbones containing 11.16M, 18.98M, and 58.45M parameters, respectively, at 0.25M, 0.5M, 1M, 2.5M, and 5M training examples. Scaling is evaluated using oracle-count edit recall

[3]All experiments across pretraining, SFT, RL, and evaluation in this work are conducted on NVIDIA RTX PRO 6000 Blackwell GPUs.

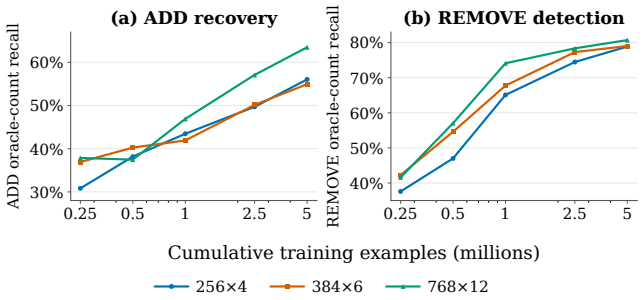


Figure 12: **Scaling behavior of GEM pretraining.**

on 580 fixed validation examples generated from 29 held-out GEMs. Unlike the single-reaction corruptions used for training, validation examples may contain multiple target edits. For each example and operation, reactions are ranked by model score and the top k_i predictions are retained, where k_i equals the number of ground-truth targets for that operation. Recall is then micro-averaged over target reactions separately for ADD and REMOVE. This metric isolates reaction-ranking quality from threshold calibration and cardinality prediction. As shown in Figure 12, both additional data and model capacity improve recovery, with the largest gains appearing for ADD. At five million examples, the 768×12 model reaches 63.49% ADD recall and 80.73% REMOVE recall. We therefore use the 768×12 checkpoint as the shared pretrained backbone for all downstream experiments.

E.3 Supervised Fine-tuning Details

Supervised fine-tuning bridges single-reaction pretraining and one-shot RL by establishing a directed biochemical repair prior. Because directly predicting complete repairs conflates reaction prioritization with combinatorial set sizing under underdetermined phenotypes, we divide SFT into two sequential stages. SFT-A learns local reaction rankings via multi-positive partial-label supervision, while SFT-B calibrates state-conditioned thresholds to decode complete edit sets in one shot, stably initializing downstream policy optimization.

E.3.1 SFT-A: Partial-Label Reaction-Edit Ranking

Algorithm design. SFT-A bridges pretrained metabolic representations and downstream complete-set decoding by learning which reaction edits are promising from partially edited states. Because phenotype supervision is partially identifying and admits multiple valid repair paths, treating a single teacher edit as unique penalizes alternative valid choices, while treating unlisted reactions as negatives overstates supervisory completeness. We therefore formulate SFT-A as multi-positive partial-label ranking. Evaluated by FBA verification on partially edited states, candidate edits are categorized into strictly improving actions (immediate phenotype gain, forming set P^+), required neutral actions (zero immediate gain but necessary for multi-step completions, forming set P^0), and regressive actions (degrading agreement, masked from gradients). Let C^- denote the regressive actions and define the effective candidate set $\widetilde{C} = C \setminus C^-$; all losses below are evaluated only over $\widetilde{C}$. By construction, each SFT-A state contains at least one certified positive ($P \neq \varnothing$). When $P^+ = \varnothing$, the strict-tier term is omitted ($L_{\text{strict}} = 0$), and when $H = \varnothing$, $L_{\text{tail}} = 0$. Improving and neutral edits jointly form the certified positive set $P = P^+ \cup P^0$. For candidate logits z over legal candidates $\mathcal{C}$, the complete training objective combines partial-label ranking, strict-tier enhancement, hard-tail margin separation against the top unlabeled candidates $\mathcal{H} \subseteq \mathcal{C} \setminus P$ ($|\mathcal{H}| \leq 32$), and a logit regularizer:

$$
\begin{aligned}
\mathcal{L}_{\text{SFT-A}} &= \mathcal{L}_{\text{union}} + 0.15\,\mathcal{L}_{\text{strict}} + 0.15\,\mathcal{L}_{\text{tail}} + 10^{-5}\mathcal{L}_{\text{logit}}, \\
\text{where} \quad \mathcal{L}_{\text{union}} &= \mathrm{LSE}_{b\in\widetilde{C}}\, z_b - \mathrm{LSE}_{a\in P}\, z_a, \\
\mathcal{L}_{\text{strict}} &= \mathrm{LSE}_{b\in\widetilde{C}}\, z_b - \mathrm{LSE}_{a\in P^+}\, z_a, \\
\mathcal{L}_{\text{tail}} &= \frac{1}{|\mathcal{H}|}\sum_{b\in\mathcal{H}} \mathrm{softplus}\big(z_b - \max_{a\in P} z_a + 0.5\big).
\end{aligned}
$$

Table 6: **Direct SFT-A development results.** Hit rates are operation-macro, model-equal retrieval scores on certified partial labels (seed 42).

Configuration	Hit@1↑	dev-select loss↓
Pretrained + ESM2	**0.207**	**3.248**
Random + ESM2	0.114	4.524
Pretrained, no ESM2	0.143	3.922

Here, $\mathrm{LSE}_{i\in\mathcal{S}}\, z_i = \log\sum_{i\in\mathcal{S}}\exp(z_i)$ denotes the LogSumExp operator, and $\mathcal{L}_{\mathrm{logit}} = \frac{1}{2|\widetilde{C}|}\sum_{b\in\widetilde{C}} z_b^2$ regularizes finite logits to prevent probability saturation. This multi-positive objective equips downstream policy learning with a directional biochemical prior toward functionally viable reactions while strictly preserving alternative valid repair pathways.

Engineering implementation. The reaction-edit actor initializes from our pretraining checkpoint using a 12-layer, 768-dimensional GINE encoder over bipartite reaction–metabolite graphs, containing 59.3M trainable parameters out of 92.7M total. Organism context is supplied by frozen ESM2 representations (Fig.11): up to 512 uniformly sampled proteins per proteome are represented by fixed 256-dimensional truncations of the normalized 480-dimensional ESM2 embeddings, aggregated by trainable attention, and mapped into the 768-dimensional graph embedding via a zero-initialized gated MLP residual, preserving pretrained reaction logits at initialization without learning explicit protein-to-reaction alignments. For scalable ADD training, each local batch constructs a shared candidate panel of nominal size 2,048 that retains all supervised batch positives and draws up to 75% of remaining slots from cached topology-based hard candidates (256 per graph) alongside vocabulary negatives; validation scores the complete ADD vocabulary without candidate sampling, while REMOVE scoring evaluates only present reaction nodes. Training uses 10,837 partial states weighted inversely by each organism's state count ($w_s = 1/n_x$) to balance supervisory influence. Optimization spans 8 GPUs with local batch size 4 (effective global batch size 32) for up to five epochs using AdamW (learning rate 3×10^{-5}, weight decay 10^{-4}, gradient clipping at 1.0) under BF16 autocast with FP32 loss reductions. We evaluate checkpoints on 410 development states and select checkpoints lexicographically by operation-macro, model-equal hit rate (Hit@1). SFT-A strictly outputs reaction candidate rankings, leaving decision thresholds, edit cardinality, and complete repair decoding to the downstream SFT-Soft stage.

Key results. Table 6 summarizes reaction-retrieval performance. The full Pretrained + ESM2 configuration achieves the strongest accuracy and likelihood fit, lifting operation-macro Hit@1 from 0.016 on pretrained initialization to 0.207 while attaining the lowest development loss (3.248). Removing pretraining degrades Hit@1 to 0.114 and inflates loss to 4.524, while removing ESM2 drops Hit@1 to 0.143 with loss 3.922. This advantage is concentrated in ADD retrieval, reaching 0.264 Hit@1 against 0.068 without pretraining and 0.125 without ESM2. Both metabolic pretraining and proteome context thus prove essential for prioritizing viable reactions and minimizing partial-label ranking loss.

E.3.2 SFT-B: Complete Set-level Reaction-Edit Supervision

Algorithm design. SFT-B converts the single action ranking actor learned in SFT-A into a one shot predictor of the complete remaining repair set. Suppressing state and operation indices, let $\mathcal{C}$ be the legal candidate set, $Y \subseteq \mathcal{C}$ the target edits from one certified completion, $N = \mathcal{C} \setminus Y$ the non target candidates. Unlike SFT-A, SFT-B treats each teacher completion as a complete binary target. Alternative valid completions remain separate training examples rather than being merged. Both variants share a class balanced binary cross entropy loss and a hard negative ranking loss. Let v denote the candidate score vector, and let $H_h \subseteq N$ be the top h negatives. With $\mathrm{softplus}(u) = \log(1+e^u)$,

$$\mathcal{L}_{\mathrm{BCE}}(v) = \frac{1}{2}\left[\frac{1}{|Y|}\sum_{a\in Y}\mathrm{softplus}(-v_a) + \frac{1}{|N|}\sum_{b\in N}\mathrm{softplus}(v_b)\right],$$

$$\mathcal{L}_{\mathrm{rank}}(v) = \frac{1}{|Y||H_h|}\sum_{a\in Y}\sum_{b\in H_h}\mathrm{softplus}(v_b - v_a).$$

The balanced BCE prevents the large negative vocabulary from dominating the objective, while the ranking loss separates target edits from competitive negatives. When one class is empty, BCE uses only the available class. Likewise, boundary losses are evaluated only when both their positive and negative aggregates are nonempty; otherwise the corresponding boundary term is set to zero. Ranking is applied only when both classes are present. Hard-SFT evaluates these losses on raw logits z, while Soft-SFT evaluates them on threshold adjusted scores u.

Hard-SFT factorizes set prediction into reaction ranking and discrete cardinality prediction. The actor outputs reaction logits z and a categorical count distribution $q(k)$. Decoding predicts $\widehat{k} = \arg\max_k q(k)$ and returns $\widehat{Y} = \mathrm{TopK}(z, \widehat{k})$. The count head supports up to $K_{\max}$ edits (13 for ADD, 1 for REMOVE), trained with cross-entropy $\mathcal{L}_{\mathrm{count}} = -\log q(m)$ and normalized expected error $\mathcal{L}_{\mathrm{dist}} = |\sum_k kq(k) - m|/K_{\max}$. Hard-SFT also enforces separation around the decision boundary at temperature $t = 0.25$ via $\mathcal{L}_{\mathrm{bdry}} = t\,\mathrm{softplus}((b^- - b^+ + 1)/t)$, where $b^+ = -t\log(|Y|^{-1}\sum_{a\in Y} e^{-z_a/t})$ and $b^- = t\log(|H_{32}|^{-1}\sum_{b\in H_{32}} e^{z_b/t})$ smoothly aggregate target and top-32 negative logits. The complete objective is

$$\mathcal{L}_{\mathrm{Hard}} = \mathcal{L}_{\mathrm{BCE}}(z) + 0.40\mathcal{L}_{\mathrm{rank}}(z) + 0.20\mathcal{L}_{\mathrm{bdry}} + 0.25\mathcal{L}_{\mathrm{count}} + 0.50\mathcal{L}_{\mathrm{dist}}.$$

Hard-SFT therefore learns reaction identity and repair size through separate outputs. This gives explicit cardinality control, but count errors directly cause under or over editing and the finite support imposes a fixed upper bound on repair size.

Soft-SFT replaces the discrete count head with a state conditioned inclusion threshold. The actor outputs reaction logits z_a, while an auxiliary head predicts an operation specific threshold τ. The adjusted score is $u_a = z_a - \tau$, giving inclusion probability $p_a = \sigma(u_a)$ and deterministic prediction $\widehat{Y} = \{a \in \mathcal{C} : u_a \geq 0\}$. Repair size therefore emerges from the candidate scores and the learned threshold rather than from a separate count prediction. Soft-SFT applies the shared BCE and ranking losses to u, with H_8 used for ranking. Its expected cardinality is $\bar{k} = \sum_{a\in\mathcal{C}} p_a$. Using smooth L_1 loss ρ, cardinality is supervised by $\mathcal{L}_{\mathrm{logcount}} = \rho(\log(1+\bar{k}) - \log(1+m))$ and $\mathcal{L}_{\mathrm{relative}} = \rho((\bar{k}-m)/(m+1))$. The former stabilizes supervision across different set sizes, while the latter remains sensitive to small targets. Soft-SFT also directly trains the zero inclusion boundary using $\mathcal{L}_{\mathrm{bdry}} = \frac{1}{2}[|Y|^{-1}\sum_{a\in Y}\mathrm{softplus}(0.5 - u_a) + |H_{64}|^{-1}\sum_{b\in H_{64}}\mathrm{softplus}(u_b + 0.5)]$. The complete objective is

$$\mathcal{L}_{\mathrm{Soft}} = \mathcal{L}_{\mathrm{BCE}}(u) + 0.20\mathcal{L}_{\mathrm{rank}}(u) + 0.15\mathcal{L}_{\mathrm{bdry}} + 0.50\mathcal{L}_{\mathrm{logcount}} + 0.02\mathcal{L}_{\mathrm{relative}}.$$

By coupling reaction scoring and repair sizing within continuous scores u, Soft-SFT removes the discrete cardinality limit and directly produces the Bernoulli inclusion probabilities for downstream one-shot RL.

Engineering implementation. Both variants initialize their shared 12-layer, 768-dimensional GNN actor, ADD/REMOVE scoring heads, and frozen ESM2 conditioning from the identical selected SFT-A checkpoint across seeds 42–44, randomly initializing only the newly added count or threshold module. Training views are matched to their decoding assumptions across 1,078 shared GEMs: Hard-SFT trains on 5,874 early-subset states with ADD-loss reweighting, whereas Soft-SFT trains on 40,151 all-subset states. Hard-SFT restricts training to early-subset states ($|S| \leq 2$) to prevent its categorical count head from developing a small-cardinality prior collapse induced by late-stage combinatorial subsets, aligning the training cardinality distribution with initial-state deployment. Conversely, Soft-SFT leverages all-subset states to train its continuous threshold across diverse residual repair scales. Operation-specific Soft thresholds are initialized by maximizing edit F1 on initial training examples and recalibrated per-epoch on a fixed training subset without consulting evaluation splits. Both methods score the full ADD vocabulary under per-GEM legality masks, restrict REMOVE scoring to present reactions, and evaluate or deploy directly from original unedited GEMs. Optimization runs on eight GPUs with effective batch size 128 using AdamW (weight decay $1e{-}4$, gradient clipping at 1.0) with layer-wise learning rates of $5e{-}6$ for the backbone, $2e{-}5$ for actor heads, and $1e{-}4$ for the count or threshold module.

Key results. For downstream evaluation, checkpoints across seeds are selected on the development split by joint edit F1, with final held-out test evaluations presented in Table 1. Crucially, a matched-batch diagnostic before any RL update confirms that Soft-SFT provides a far more informative initialization for policy optimization, yielding nonzero within-group reward variance for $51.6\% \pm 4.9\%$ of sampled organism groups compared with $32.0\% \pm 5.1\%$ for Hard-SFT. This verifies that the adaptive-threshold policy exposes a richer verifier signal to kick-start policy exploration.

Table 7: **Shared RL configuration.**

Parameter	Setting
Actor initialization	Soft-SFT
Total / trainable parameters	92.85M / 59.46M
Training / validation GEMs	9,196 / 1,087
RL seeds	42, 43, 44
Training devices	8 GPU ranks
GEMs per update	32
Rollouts per GEM	32
Training budget	3 nominal epochs, 864 updates
Actor updates per rollout batch	1
Optimizer	AdamW
Learning rate	$2e{-}6$
Weight decay	$1e{-}4$
Gradient norm clipping	1.0
GRPO clipping	$\epsilon = 0.2$
Advantage stabilizer	$1e{-}6$
Bernoulli sampling temperature	1.0
FBA growth threshold	$1e{-}3$
Validation cadence	Every 5 updates and at update 864

E.4 Reinforcement Learning Details

Shared RL training setup. Alongside the GFlowNet (Appendix D.2), Vanilla GRPO, MaxEnt, and QuotientPO (Appendix A) share a unified one-shot policy optimization framework with detailed configurations summarized in Table 7. Initialized from the selected Soft-SFT checkpoint, the policy samples a complete repair set in a single forward pass via independent Bernoulli decisions over legal edits using fine-tuned inclusion thresholds, dispensing with value critics, reference KL penalties, or discrete cardinality heads. Vanilla GRPO optimizes this verifier-guided objective directly, serving as the common backbone for exploration ablations.

Action-entropy baseline. MaxEnt augments GRPO with the exact Shannon entropy of the factorized Bernoulli policy, summed analytically in nats over all legal ADD and REMOVE decisions without vocabulary normalization, Monte Carlo estimation, or additional verifier overhead. Because summing unnormalized entropy across a large reaction vocabulary risks simultaneously elevating many individually unlikely actions and causing severe repair cardinality explosion, its coefficient requires explicit calibration. In development experiments, coefficients of $1e{-}3$ and $5e^{-4}$ triggered severe ADD cardinality inflation, whereas $1e{-}4$ remained stable but yielded insufficient exploration. We therefore fixed the intermediate coefficient of $2e{-}4$ across all seeds based strictly on validation behavior, ensuring that the hyperparameter choice was locked before test evaluation without consulting held-out splits.

Runtime overhead. We benchmark one complete optimizer update on a single node with $8\times$ NVIDIA RTX PRO 6000 Blackwell GPUs, using the same RL training settings for all methods. All runs use the same HiGHS backend, 64 single-threaded FBA workers, FP32 execution, and no validation or external logging. QuotientPO requires 23.22 s per update versus 21.32 s for Vanilla GRPO, corresponding to an 8.9% overhead. Most additional cost comes from verifier-side processing, including canonicalization, QPI and kernel scoring.

Table 8: **Per-update runtime** under identical hardware and rollout settings.

Method	Update (s)	FBA drain (s)	Overhead
Vanilla GRPO	21.323	2.229	–
QuotientPO	23.218	4.186	+8.9%

Table 9: **Performance across GEM reconstruction templates.** Results are mean ± standard deviation across three seeds.

Template	Method	Success@1	Success@8	Success@32	Greedy	Improved	Success Cores@32
Default ($N = 697$)							
	Vanilla GRPO	7.37±0.58	12.72±0.79	16.79±1.08	**7.84±0.36**	28.83±1.08	0.516±0.003
	MaxEnt	5.69±0.96	11.81±1.58	15.93±1.28	7.32±1.52	28.04±2.33	0.462±0.088
	GFlowNet	1.77±0.22	6.07±0.54	10.38±0.33	2.54±0.30	11.48±0.70	0.265±0.017
	QuotientPO	**7.89±0.50**	**14.25±0.22**	**18.75±0.82**	7.80±0.33	**30.01±1.19**	**0.582±0.006**
Gram-positive ($N = 491$)							
	Vanilla GRPO	**7.13±0.20**	13.04±0.35	16.70±2.12	8.15±0.41	30.14±0.72	0.573±0.019
	MaxEnt	5.16±1.73	10.93±1.01	16.02±0.85	7.13±1.43	30.17±1.06	0.483±0.087
	GFlowNet	2.58±0.31	7.06±0.31	10.93±0.12	2.72±0.82	14.33±0.64	0.297±0.017
	QuotientPO	6.93±0.20	**14.39±0.62**	**19.48±0.77**	**9.03±0.77**	**31.44±1.08**	**0.604±0.003**
Gram-negative ($N = 507$)							
	Vanilla GRPO	7.43±0.89	14.79±0.99	19.26±0.99	8.02±0.41	30.21±0.49	0.507±0.027
	MaxEnt	6.11±0.59	13.48±1.40	17.82±2.43	6.77±0.97	29.50±1.88	0.550±0.027
	GFlowNet	3.81±0.30	8.22±0.11	12.10±0.30	2.89±0.11	13.89±0.46	0.377±0.008
	QuotientPO	**8.15±0.30**	**15.98±0.20**	**20.97±1.01**	**8.55±0.23**	**32.08±0.56**	**0.617±0.032**
Archaeal ($N = 517$)							
	Vanilla GRPO	7.80±0.30	14.96±2.34	19.34±2.42	7.93±0.70	32.11±0.23	0.409±0.057
	MaxEnt	6.96±0.89	14.77±1.24	19.08±1.45	7.93±1.01	30.17±2.76	0.427±0.022
	GFlowNet	5.22±0.19	12.77±0.51	19.79±0.40	4.84±0.39	21.08±0.15	0.433±0.017
	QuotientPO	**8.00±0.30**	**16.89±0.81**	**21.66±0.39**	**8.77±0.40**	**33.38±0.36**	**0.476±0.030**

E.5 Evaluation Details

All methods are evaluated on 2,212 held-out initial GEMs. For each GEM, we generate 32 stochastic repair sets in FP32 with dropout disabled under fixed per-organism seeds, alongside a separate greedy prediction. Canonical repair cores are uniformly recomputed for all baselines using the deterministic reduction described in Appendix A. Success@K is computed over nested prefixes of the 32 rollouts, greedy success is evaluated separately, and diversity metrics are calculated over these frozen outputs. All metrics are computed per GEM and macro-averaged, reporting the mean and sample standard deviation across three training seeds.

F Additional Experimental Results

F.1 Robustness across reconstruction templates.

QuotientPO consistently improves Success@32 over Vanilla GRPO across all four CarveMe templates, with gains of +1.96, +2.78, +1.71, and +2.32 percentage points on default, Gram-positive, Gram-negative, and archaeal models, respectively (Table 9). Successful-core coverage also improves in every template, indicating that the overall gain is not driven by any single reconstruction setting, including the archaeal subset.

F.2 Solution Space Geometry: SFT versus Policy Optimization

To analyze exploration geometry, each raw repair is represented by its signed reaction edits under pairwise Jaccard distance $d(A, B) = 1 - |A \cap B|/|A \cup B|$ ($d(\varnothing, \varnothing) = 0$), and 32 stochastic rollouts from SFT-Soft and QuotientPO are jointly embedded with the unedited GEM and a certified MILP teacher reference via 2D metric multidimensional scaling (MDS)[4]. Identical edits share averaged coordinates on aligned axes, with all reported teacher distances evaluated in the ambient Jaccard

[4] https://github.com/scikit-learn/scikit-learn

Table 10: **Success-controlled mechanism coverage.** Expected distinct canonical cores (D_r) evaluated across r successful draws on paired cohorts.

r	Paired GEMs (42/43/44)	Vanilla GRPO	QuotientPO
2	295/269/323	1.425 ± 0.019	$\mathbf{1.461 \pm 0.020}$
4	247/228/258	1.960 ± 0.037	$\mathbf{2.050 \pm 0.056}$
8	209/186/206	2.659 ± 0.076	$\mathbf{2.782 \pm 0.114}$
16	147/123/130	3.256 ± 0.033	$\mathbf{3.502 \pm 0.203}$

space. Across six representative organisms (Figure 13), SFT-Soft rollouts yield few successful completions and cluster tightly around the certified teacher reference. In contrast, QuotientPO substantially increases repair success and disperses into multiple disconnected clusters, moving beyond the supervised demonstration to recover valid repairs across separated regions of the edit space.

F.3 Extended Solution-Space Visualizations Across RL Baselines

Figure 14 compares exploration behavior across Vanilla GRPO, MaxEnt, GFlowNet, and QuotientPO projected into both raw edit and quotient mechanism spaces over six representative GEMs. Standard RL baselines and action-entropy exploration suffer from severe mode collapse, frequently rediscovering identical redundant mechanisms despite surface-level action variations. In contrast, QuotientPO consistently disperses across the quotient landscape, discovering substantially more distinct, successful canonical repair cores across diverse biological targets.

F.4 Success-Controlled Mechanism Coverage

Successful-core yield conflates solve rates with conditional mechanism diversity. To isolate the latter, we measure the expected distinct canonical cores under an equalized budget of r successful repairs, restricted to paired cohorts where both methods achieve at least r successes among 32 rollouts. For M successful repairs with core a occurring s_a times, the expected distinct core count across r draws without replacement is

$$D_r = \sum_a \left[1 - \frac{\binom{M-s_a}{r}}{\binom{M}{r}} \right].$$

Table 10 shows that QuotientPO consistently outperforms Vanilla GRPO across all budgets, with the margin widening from $+0.035$ at $r = 2$ to $+0.246$ at $r = 16$. This confirms that the diversity gain does not merely reflect a higher solve rate. We truncate at $r = 16$ because requiring all 32 rollouts to succeed leaves only 10/8/4 paired GEMs across seeds 42/43/44.

F.5 Finite-Sample Credit Resolution Analysis

Metric definitions. We compute the training diagnostics from each rollout group of $N = 32$ repairs. Let n_a denote the multiplicity of canonical core a in the group. Under exact Rényi credit, two distinct cores receive identical credit if and only if they have the same multiplicity, so we identify ties directly by $n_a = n_b$ rather than by floating-point comparison. Weighting each distinct-core pair by the number of corresponding sampled pairs, $w_{ab} = n_a n_b$, the exact-credit tie rate is

$$\mathrm{TieRate} = \frac{\sum_{a<b} w_{ab}\, \mathbb{I}[n_a = n_b]}{\sum_{a<b} w_{ab}}.$$

Importantly, $\mathrm{TieRate}$ is an *exact-count* diagnostic for every method: it measures how often distinct cores receive identical multiplicity-based credit, $n_a = n_b$, within rollout groups generated by that policy. Thus, the Kernel Rényi-2 row does not measure ties under the kernel credit itself. Instead, it quantifies how often exact counting would still fail to distinguish cores sampled from the kernel-trained policy. The complementary $\mathrm{ResolvedRate}$ measures the fraction of these exact-count ties that are separated by the graded kernel credit.

To measure how much structure is hidden within these ties, we recompute a kernel credit from the same full rollout group. With signed-Jaccard similarity $K_{ab} = \exp[-(1 - J(a,b))/0.25]$, the local kernel density of core a is $\rho_a = \frac{(n_a-1)+\sum_{b\neq a} n_b K_{ab}}{N-1}$, $P_K = \frac{1}{N}\sum_a n_a \rho_a$, giving the normalized kernel cost $c_a^K = 2(\rho_a/P_K - 1)$. Following the training scale, we compare $\widetilde{c}_a^K = 0.02\ \mathrm{clip}(c_a^K, -8, 8)$ and define the kernel-resolved rate as the fraction of exact-credit ties satisfying $|\widetilde{c}_a^K - \widetilde{c}_b^K| > 0.001$:

$$\mathrm{ResolvedRate} = \frac{\sum_{a<b} w_{ab}\, \mathbb{I}[n_a = n_b]\, \mathbb{I}[|\widetilde{c}_a^K - \widetilde{c}_b^K| > 0.001]}{\sum_{a<b} w_{ab}\, \mathbb{I}[n_a = n_b]}.$$

Thus, TieRate measures the coarseness of exact finite-sample credit, whereas ResolvedRate is a conditional diagnostic of graded structure hidden inside exact ties and is not a higher-is-better performance metric. The successful-core statistics in Table 2 are computed separately from successful stochastic repairs only. For a GEM with $M \geq 2$ successful draws, let s_a be the number assigned to successful core a. Exact collision and mean pairwise Jaccard distance are

$$\mathrm{Collision} = \frac{\sum_a s_a(s_a - 1)}{M(M-1)}, \qquad \bar{d}_J = \frac{\sum_{a<b} 2 s_a s_b\, [1 - J(a,b)]}{M(M-1)}.$$

Repeated draws retain their empirical multiplicity, so duplicate cores increase collision and contribute zero Jaccard distance rather than being deduplicated. All quantities are first computed per GEM and then macro-averaged over GEMs for which the corresponding statistic is defined. Undefined cases are excluded rather than imputed as zero.

Full training trajectory. Table 11 traces credit resolution across updates 0 to 864. As training proceeds, Exact Rényi-2 becomes increasingly coarse under finite rollout groups, with its exact-count tie rate rising from 5.67% to 10.23%. Kernel Rényi-2 instead reshapes the rollout distribution toward finer exact-count structure, ending at 7.99%. More importantly, exact multiplicity substantially understates the available geometric signal: at step 864, kernel credit resolves 78.20% of exact-count ties in Exact Rényi-2 and 84.86% in Vanilla GRPO. These results demonstrate that kernel relaxation promotes a less degenerate rollout geometry and extracts graded credit among otherwise indistinguishable cores, translating directly into superior downstream performance (Table 12).

Table 11: **Full training trajectory of finite-sample credit resolution.** Exact-count tie rates measure distinct-core pairs indistinguishable under multiplicity credit within rollout groups; kernel-resolved rates report the fraction of these ties separated by graded kernel credit.

	Exact-credit ties (%)				**Kernel-resolved ties (%)**			
Method	**0**	**200**	**400**	**864**	**0**	**200**	**400**	**864**
Vanilla GRPO	5.67	8.19	8.31	8.59	30.61	60.55	57.50	84.86
Discrete Shannon	5.67	8.60	9.99	9.72	30.61	59.23	59.77	80.20
Exact Rényi-2	5.67	7.34	9.06	10.23	30.61	59.71	63.71	78.20
Kernel Rényi-2	5.67	8.57	8.60	**7.99**	30.61	63.81	57.23	72.78

Table 12: **Downstream performance of quotient exploration objectives.**

Method	Success@32 ↑	Success Cores@32 ↑
Vanilla GRPO	17.93 ± 1.54	0.50 ± 0.01
Discrete Shannon	18.61 ± 1.09	0.52 ± 0.01
Exact Rényi-2	18.04 ± 1.08	0.54 ± 0.03
Kernel Rényi-2	$\mathbf{20.10 \pm 0.52}$	$\mathbf{0.57 \pm 0.02}$

F.6 Multi-Seed Robustness and Ablation Results

Table 13 provides the complete evaluations corresponding to Table 3, confirming consistent improvements for QuotientPO.

Table 13: **Multi-seed robustness and ablation results.** Success@32 (%) are reported as mean ± standard deviation over three seeds.

(a) RLOO Base

Method	Success@32
Vanilla	15.75 $\pm$1.20
MaxEnt	14.29 $\pm$0.95
GFlowNet	11.57 $\pm$0.14
QuotientPO	**17.21** $\pm$**0.81**

(b) Policy Architecture

Architecture	RL	Success@32
Soft	Vanilla	17.93 $\pm$1.54
	QuotientPO	20.10 $\pm$0.52
Hard	Vanilla	5.92 $\pm$0.09
	QuotientPO	6.65 $\pm$0.23

(c) Design Ablation

Variant	Success@32
Default	20.10 $\pm$0.52
Raw-space kernel	18.16 $\pm$0.61
w/o Quotient term	17.54 $\pm$1.81
w/o QPI term	19.62 $\pm$0.72
w/o Redundancy penalty	19.94 $\pm$1.79

(d) Hyperparameter Sensitivity

Param.	Setting	Success@32
Default	$h = 0.25,\ \beta = 0.02$	20.10 $\pm$0.52
h	0.125	19.41 $\pm$0.70
	0.5	18.57 $\pm$0.74
β	0.01	19.05 $\pm$1.12
	0.04	20.00 $\pm$0.74

G Biological Analysis and Case Studies

Six additional cases complement the main-text sucrose and nitrate examples, spanning two cases across three progressive coverage tiers. Each comparison evaluates the identical GEM, checkpoint, and seed over $N = 32$ stochastic rollouts per method (V: Vanilla GRPO; Q: QuotientPO). For each focal alternative (the second pathway in each case), we track its sample-level representation along the three-stage evaluation pipeline: raw proposals containing the focal module (R), verified successful raw repairs containing the focal module (S), and successful canonicalized repairs in which the focal module remains present (C), satisfying $R \geq S \geq C$. Importantly, R, S, and C count sample occurrences rather than distinct canonical cores; the total number of distinct successful canonical cores is reported separately as K_{core}. The three coverage tiers characterize the stage at which Vanilla fails to retain the focal mechanism while QuotientPO does ($C_Q > 0$): Tier 1 (retention gap: $S_V > 0, C_V = 0$) captures focal alternatives that already occur in successful Vanilla proposals but are eliminated during canonicalization; Tier 2 (viability gap: $R_V > 0, S_V = 0$) captures focal modules proposed by Vanilla only within unsuccessful repairs; and Tier 3 (discovery gap: $R_V = 0$) captures focal modules absent from Vanilla raw proposals.

H Limitations and Future Work

Limitations. Our study adopts a controlled setting to isolate phenotype-supervised GEM repair. All GEMs are reconstructed with CarveMe in the BiGG reaction system, which ensures a consistent reaction vocabulary and modeling convention across organisms but does not yet test transfer across reconstruction pipelines. We also use standardized category-specific media as reproducible proxies for BacDive growth conditions. This abstraction is necessary because many BacDive assays provide heterogeneous or incomplete culture descriptions, but it does not capture all strain-specific environmental details. Finally, our verifier focuses on steady-state metabolic feasibility and does not model regulatory or kinetic effects.

Future work. Future work can extend the framework to GEMs reconstructed by other pipelines, richer or uncertainty-aware culture-condition models, and additional biological constraints beyond FBA. More broadly, combining quotient-space repair with targeted experimental validation could connect phenotype-consistent computational repairs to testable metabolic hypotheses. Finally, quotient-space exploration may extend well beyond metabolic modeling to broader scientific domains such as molecular design, reaction mechanism discovery, and symbolic physical modeling, where macroscopic observations underdetermine underlying structures and search will benefit from operating directly over functional equivalence classes.

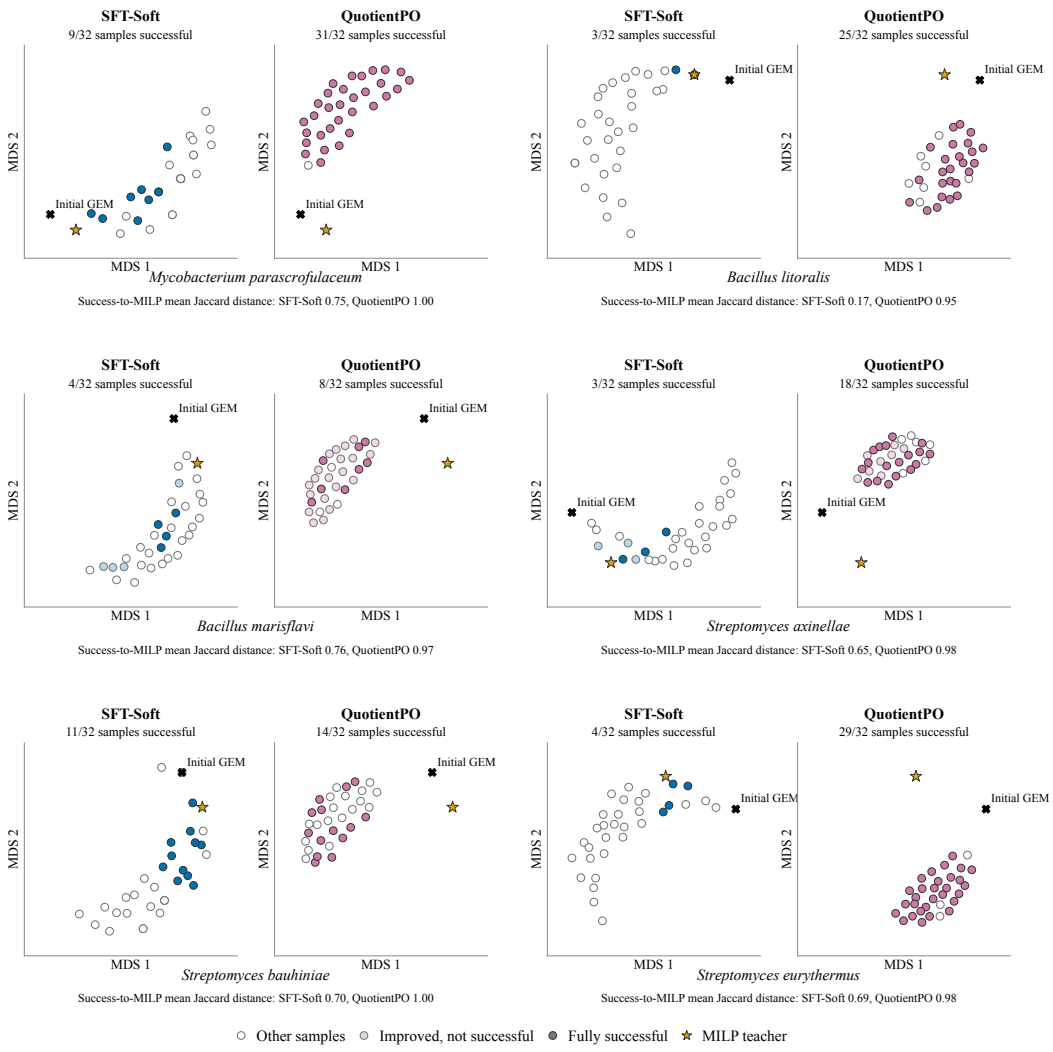


Figure 13: **Cases of SFT versus RL Optimization.**

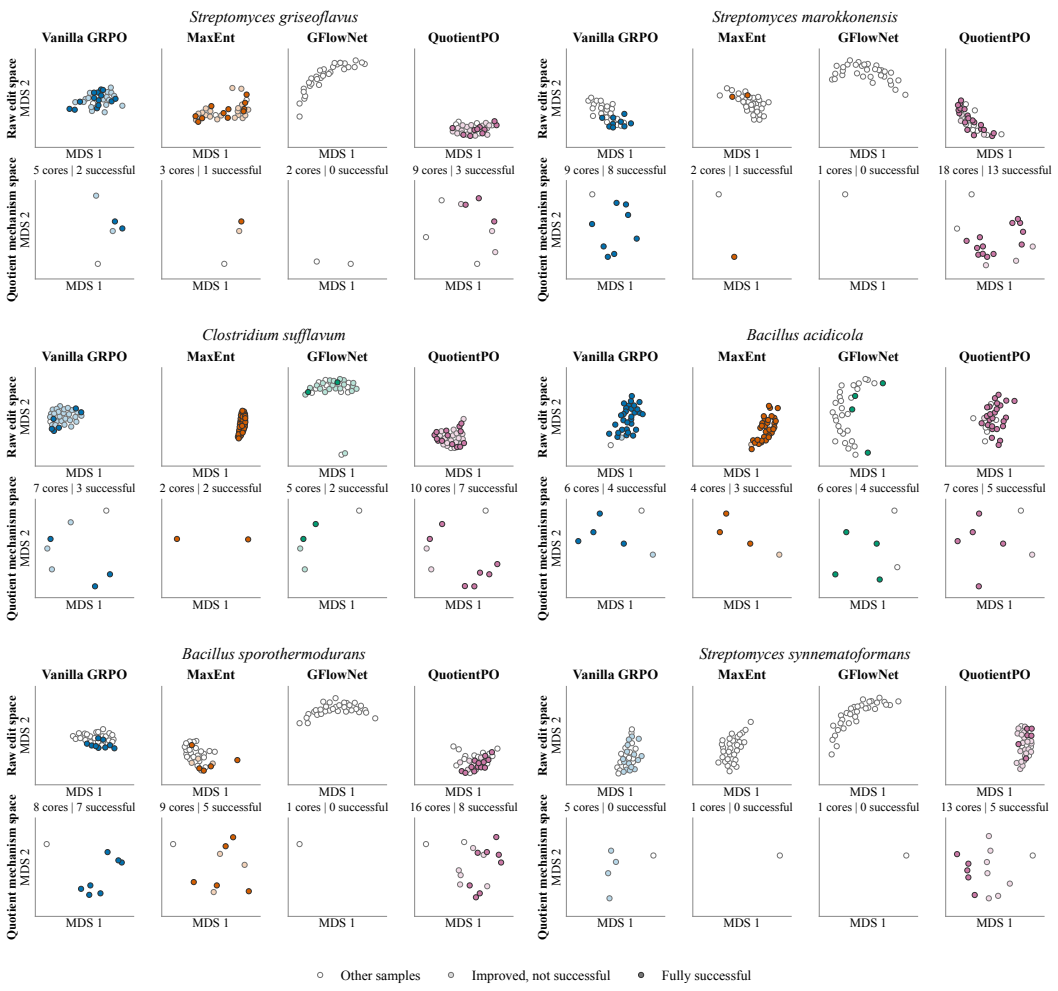


Figure 14: **Cases of solution-space visualization across RL baselines.**

Table 14: **Tier 1: retained-solution coverage.** The focal alternative appears in successful Vanilla raw proposals ($S_V > 0$) but is retained only by QuotientPO after canonical reduction ($C_V = 0$, $C_Q > 0$).

Repair alternatives & encoded chemistry	**Functional verification & policy diagnosis**
Case 1 (C1): *Mycobacterium terrae* **Target:** Urea utilization **Focal pipeline** ($R \to S \to C$)**:** V: $13 \to 2 \to 0$ vs Q: $13 \to 2 \to 1$ **Yield** ($N_{\text{succ}} \mid K_{\text{core}}$)**:** V: $4 \mid 1$ Q: $4 \mid 2$	

Shared deletions (3):
`-HEX1`, `-MALTabc`, `-NARK`
Pathway 1: Uptake blockade
`-UREAt`
$\text{urea}_e \longrightarrow \text{urea}_c$
Pathway 2: Carboxylation blockade (focal)
`-UREASE`
$\text{urea} + \text{ATP} + \text{HCO}_3^- \longrightarrow \text{allophanate} + \text{ADP} + \text{P}_\text{i} + \text{H}^+$
Note: `UREASE` encodes urea carboxylase in the GEM. Both canonical cores contain exactly four deletions.
	Phenotypic verification
	Both cores reduce urea-supported growth from 0.202 to zero while preserving nitrate-supported growth of 0.545 and all remaining targets. Reintroducing the focal deletion restores the failed urea-growth phenotype, confirming its functional necessity.
	Search behavior diagnosis
	Vanilla produces successful raw proposals containing this intracellular branch ($S_V = 2$), but completely eliminates it during canonical reduction ($C_V = 0$). In contrast, QuotientPO successfully retains both functional intervention sites ($C_Q = 1$).

Case 2 (C2): *Streptomyces inusitatus* **Target:** Alternative proline supply **Focal pipeline** ($R \to S \to C$)**:** V: $32 \to 32 \to 0$ vs Q: $32 \to 30 \to 2$ **Yield** ($N_{\text{succ}} \mid K_{\text{core}}$)**:** V: $32 \mid 2$ Q: $30 \mid 2$	

Shared deletion (1): `-UREAt`
Pathway 1: Direct conversion
`+ORNCD`
$\text{ornithine} \longrightarrow \text{proline} + \text{NH}_4^+$
Pathway 2: Semialdehyde/P5C route (focal)
`+G5SADs`
$\text{GSA} \longrightarrow \text{P5C} + \text{H}_2\text{O} + \text{H}^+$
$\text{P5C} \xrightarrow{\text{existing P5CR/P5CRx}} \text{proline}$
Note: GSA denotes glutamate-5-semialdehyde; P5C denotes Δ^1-pyrroline-5-carboxylate. Both canonical cores contain two edits.
	Phenotypic verification
	Both two-edit cores restore proline supply and recover citrate-supported growth from zero to 0.831 while preserving the inactive urea target. Flux witnesses independently confirm active throughput along both pathways.
	Search behavior diagnosis
	Vanilla already contains the focal G5SADs route in successful proposals but loses it after canonicalization ($C_V = 0$); QuotientPO retains it ($C_Q = 2$). Thus, this is a retention rather than discovery gap. ORNCD and ORNCD_2 are treated as one functional pathway family.

Table 15: **Tier 2: successful-combination coverage.** Vanilla proposes the focal module ($R_V > 0$) but does not place it in a successful complete repair ($S_V = 0$), whereas QuotientPO does.

Repair alternatives & encoded chemistry	**Functional verification & policy diagnosis**
Case 3 (C3): *Lactobacillus farraginis* **Target:** Alternative phosphate donors for UTP synthesis **Focal pipeline** ($R \to S \to C$)**:** V: $4 \to 0 \to 0$ vs Q: $10 \to 1 \to 1$ **Yield** ($N_{\mathrm{succ}} \mid K_{\mathrm{core}}$)**:** V: $0 \mid 0$ Q: $2 \mid 2$	

Shared edits (3):
`+MOHMT`, `+NTRIR2y`, `-NTRIR3pp`
Pathway 1: ATP donor
`+NDPK2`
$\mathrm{ATP} + \mathrm{UDP} \longrightarrow \mathrm{ADP} + \mathrm{UTP}$
Pathway 2: PEP donor (focal)
`+PYK2`
$\mathrm{PEP} + \mathrm{UDP} + \mathrm{H}^{+} \longrightarrow \mathrm{pyruvate} + \mathrm{UTP}$
Both complete cores contain four edits.
	Phenotypic verification
	The alternatives provide UDP-to-UTP phosphorylation using different phosphate donors.
	Both QuotientPO cores satisfy all six targets, with glucose growth 0.540171 and 0.539515, respectively, and nitrate growth 0.888279.
	Flux witnesses carry approximately 0.16655 through `NDPK2` or 0.16635 through `PYK2`.
	Search behavior diagnosis
	Vanilla proposes `PYK2` four times, but these proposals lack the coordinated repair required for success. Adding `NTRIR2y` restores the nitrate-positive target.

Case 4 (C4): *Streptomyces ferralitis* **Target:** Alternative nitrogen donors for asparagine **Focal pipeline** ($R \to S \to C$)**:** V: $24 \to 0 \to 0$ vs Q: $29 \to 3 \to 1$ **Yield** ($N_{\mathrm{succ}} \mid K_{\mathrm{core}}$)**:** V: $0 \mid 0$ Q: $3 \mid 3$	

Shared edits (3):
`+ARGDC`, `-FFSD`, `-SUCRtex`
Pathway 1: Glutamine donor
`+ASNS1`
$\mathrm{Asp} + \mathrm{Gln} + \mathrm{ATP} + \mathrm{H_2O} \longrightarrow \mathrm{Asn} + \mathrm{Glu} + \mathrm{AMP} + \mathrm{PP_i} + \mathrm{H}^{+}$
Pathway 2: Ammonium donor (focal)
`+ASNS2`
$\mathrm{Asp} + \mathrm{NH_4^{+}} + \mathrm{ATP} \longrightarrow \mathrm{Asn} + \mathrm{AMP} + \mathrm{PP_i} + \mathrm{H}^{+}$
	Phenotypic verification
	The two reactions complete asparagine synthesis using different nitrogen donors.
	Both QuotientPO cores satisfy the glucose-positive and sucrose-negative targets, with glucose growth 1.342551 and 1.345126.
	Search behavior diagnosis
	Vanilla frequently proposes the focal `ASNS2` branch, but the corresponding repair sets remain unsuccessful. One such proposal also introduces extracellular sucrose hydrolysis through `SUCRe`; removing this additional bypass restores the sucrose-negative target.
	Thus, successful use of the focal biosynthetic branch depends on its surrounding repair combination.

Table 16: **Tier 3: raw-module coverage.** The complete focal module is absent from Vanilla raw proposals ($R_V = 0$) but appears in successful QuotientPO repairs.

Repair alternatives & encoded chemistry	**Functional verification & policy diagnosis**
Case 5 (C5): *Streptomyces olivochromogenes* **Target:** Alternative control of citrate-supported growth **Focal pipeline** ($R \to S \to C$)**:** V: $0 \to 0 \to 0$ vs Q: $2 \to 2 \to 2$ **Yield** ($N_{\text{succ}} \mid K_{\text{core}}$)**:** V: $2 \mid 1$ Q: $8 \mid 3$	

Shared edit (1):
`-UREAt`
Alternative 1: Entry restriction
`-CITt14`, `-FE3t`
CITt14 directly supports citrate uptake, while FE3t provides a reversible iron route coupled to the remaining FEDCabc citrate-entry pathway.
Alternative 2: Intracellular node (focal)
`-SUCOAS`
$\text{succinyl-CoA}+\text{ADP}+\text{P}_\text{i} \longrightarrow \text{succinate}+\text{CoA}+\text{ATP}$
	Phenotypic verification
	Both cores eliminate the initial urea and citrate false positives while preserving glucose- and sucrose-supported growth. The entry-restriction core retains glucose/sucrose growth at $0.630548/1.261097$, whereas the SUCOAS core retains $0.436023/0.872046$.
	Restoring either `CITt14` or `FE3t` breaks only the citrate-negative target in the first core. Restoring `SUCOAS` does the same in the focal core. Restoring `UREAt` affects only the urea-negative target.
	Search behavior diagnosis
	Vanilla never proposes `-SUCOAS` in this rollout panel. QuotientPO discovers it twice. Both samples are successful and canonicalize to the same focal core. Thus, this comparison exposes an intracellular repair alternative absent from Vanilla raw proposals, rather than two distinct mechanisms.

Case 6 (C6): *Streptomyces durbertensis* **Target:** Alternative quinone-precursor supply **Focal pipeline** ($R \to S \to C$)**:** V: $0 \to 0 \to 0$ vs Q: $5 \to 1 \to 1$ **Yield** ($N_{\text{succ}} \mid K_{\text{core}}$)**:** V: $1 \mid 1$ Q: $7 \mid 5$	

Shared edits (5):
`+HSERTA`, `+HSST`, `+THZPSN2_SC`, `-NO3abc`, `-NO3tex`
Pathway 1: Aromatic branch
`+PHPYROX`
$\text{phenylpyruvate} \longrightarrow \text{benzaldehyde}$
$\longrightarrow \text{benzoate} \longrightarrow \text{benzoyl-CoA}$
$\xrightarrow{\text{reverse BSCT}} \text{OSB-CoA}$
Pathway 2: Three-ADD branch (focal)
`+SHCHCS`, `+SUCBZS`, `+SUCBZL_1`
$\text{isochorismate} + \text{SSA}$
$\longrightarrow \text{SHCHC} \longrightarrow \text{OSB}$
$\longrightarrow \text{OSB-CoA}$
Note: SSA denotes succinic semialdehyde and OSB denotes *o*-succinylbenzoate.
	Phenotypic verification
	Both branches converge on OSB-CoA and downstream menaquinone-precursor supply.
	The resulting cores satisfy all four targets and yield glucose growth of approximately 0.454946.
	Minimum-total-flux witnesses use the corresponding precursor branch at approximately 4.55×10^{-5}, matching the quinone demand induced by biomass production.
	Search behavior diagnosis
	The complete three-ADD module does not occur in Vanilla raw proposals, whereas QuotientPO proposes it five times and retains one successful canonical core. Its five total distinct cores therefore correspond to multiple edit sets, rather than five distinct quinone-supply mechanisms.