EGRET: EXPLORING GENERALIST AND SPECIALIST REPERTOIRES VIA POSTERIOR LINEAGE SAMPLING OVER EPISTATIC TREES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learning-guided protein engineering usually optimizes one sequence at a time, although related conditions may be better served by descendants that share early mutations and specialize after branching. We introduce *assay-limited generalist-to-specialist diversification*: given a common scaffold, a finite sequence library, and limited sequence–condition measurements, a method jointly selects a condition niche and a rooted mutation tree under a shared construction budget. The objective measures how much accessible specialist headroom the tree captures beyond the best reachable generalist. This decision couples specialists by shared lineage cost and exposes condition-wise surrogate maximization to the optimizer's curse. EGRET uses a hierarchical epistatic posterior and posterior lineage sampling for assay allocation. It commits a final tree using measured leaf–condition utilities, providing a verifiable lower bound on its value. Diverse experiments evaluate repertoire quality, explanation faithfulness, and decision traceability across protein-engineering scenarios with varied objectives and assay budgets.

1 INTRODUCTION

Learning-guided protein engineering searches large sequence spaces under limited experimental budgets, often alternating model updates with targeted assays (Yang et al., 2025). Recent approaches combine pretrained representations with offline sequence optimization (Chen et al., 2023), guide generative models toward high-fitness variants (Bogensperger et al., 2025), and use interpretable concepts or motif importance to steer sequence edits (Ismail et al., 2025; Pandey et al., 2026). Multi-objective methods also generate diverse candidates spanning competing properties (Jain et al., 2023; Tang et al., 2025). These advances motivate a further design question: which related variants should be constructed together when their value comes from serving different conditions?

Consider an enzyme required to act on several substrates. Promiscuous activity can preserve an existing function while mutations improve a secondary one, providing a route from a generalist scaffold to specialized descendants (Aharoni et al., 2005). Which descendants are reachable, however, can depend on permissive substitutions and residue-level epistasis (Starr et al., 2017). When these descendants are constructed from a common scaffold, early mutations may be shared before their lineages branch. Reusing a shared construction step can make a family affordable even when building each endpoint separately would exceed the budget. Conversely, a diverse set of individually strong variants need not be jointly affordable. The relevant design object is therefore a rooted mutation tree, evaluated by leaf condition coverage and shared-edge construction cost.

We formalize this problem as *assay-limited generalist-to-specialist diversification*. Given a scaffold, a finite aligned sequence library, and a budget of sequence–condition measurements, a method selects a condition niche and a specialist tree. We compare its utility with that of the best generalist reachable under the same construction budget, and normalize the gain by the accessible specialist headroom. This formulation couples niche selection, terminal variants, and branching structure: a useful specialist must improve coverage while leaving enough budget to construct complementary descendants that jointly cover the remaining conditions.

Learning for this decision requires both a structured landscape model and an acquisition policy that accounts for the entire repertoire. Interpretable sequence models can expose which concepts or

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.

motifs guide design (Ismail et al., 2025; Pandey et al., 2026); here, mutation effects shared across conditions must also be distinguished from effects associated with specialization. Meanwhile, reliable uncertainty at the level of individual protein variants remains challenging (Groth et al., 2024). Taking maxima across many predicted variants can amplify optimistic errors through the optimizer's curse (Smith & Winkler, 2006), and sharing construction costs makes the value of an assay depend on which other variants join the tree. A strong predicted endpoint alone is consequently insufficient evidence for a strong repertoire.

EGRET addresses these challenges through posterior lineage sampling. An interpretable epistatic model decomposes sequence–condition predictions into site, residue, and pair effects, with shared effects and condition-specific deviations. Posterior landscape draws are converted into feasible trees, and sequences are prioritized by their support across promising repertoires. Adaptive screening allocates measurements across conditions, while final commitment uses measured utilities. Together, these components connect interpretable landscape learning to construction-aware exploration and experimentally supported repertoire selection.

Contributions. ① **A new sequence-design problem.** We formulate assay-limited generalist-to-specialist diversification, measuring the advantage of a jointly constructible specialist tree over the best generalist reachable under the same construction budget. ② **A new decision-level method.** We introduce posterior lineage sampling, which directs assays toward sequences that repeatedly support high-value feasible trees across posterior landscape draws. This aligns exploration with the final repertoire decision while reserving commitment for experimentally supported utility. ③ **A broad empirical evaluation.** We evaluate performance across proteins, condition panels, objectives, and assay budgets against optimization and active-learning baselines, with controlled ablations and audits of repertoire quality, explanation faithfulness, and decision traceability.

2 Related Work

Appendix A provides an expanded discussion of the biological motivation, practical applications, and methodological distinctions, including a comparison matrix in Table 9.

Learning-guided biological sequence design. Recent methods use pretrained representations and generative priors to search for functional sequences. Chen et al. (2023) combine a pretrained language-model proxy with bidirectional learning for offline biological sequence optimization. VL-GPO couples a flow-matching prior over mutations with a fitness predictor to optimize proteins in a continuous latent space (Bogensperger et al., 2025). EGRET considers a finite library with sequential assay access and evaluates terminal variants jointly through condition coverage and shared construction cost. Its central decision is therefore which specialists should be constructed together.

Multi-objective and diverse sequence design. Multi-Objective GFlowNets generate diverse Pareto candidates, including through an active-learning variant (Jain et al., 2023). PepTune combines discrete diffusion with Monte Carlo Tree Guidance to optimize multiple therapeutic peptide properties (Tang et al., 2025). These approaches already go beyond returning a single optimum: they explore trade-offs and maintain candidate diversity. EGRET targets a different set objective, in which each condition is served by the best leaf and complementary specialists share a construction budget. In particular, PepTune's tree organizes the generative search, whereas EGRET's returned tree specifies shared mutation steps from a common scaffold. Diversity or Pareto quality alone does not determine this tree's utility or cost.

Interpretable and uncertainty-aware decision making. Interpretable genotype–phenotype models expose functional effects of sequence changes (Tonner et al., 2022), and recent design methods use biological concepts or motif importance to guide edits (Ismail et al., 2025; Pandey et al., 2026). In parallel, Bayesian protein optimization uses predictive uncertainty to select experiments (Groth et al., 2024; Bal et al., 2025), with structured-action precedents in combinatorial posterior sampling (Wang & Chen, 2018). EGRET connects these directions at the repertoire decision: a conditional epistatic landscape distinguishes broad from condition-specific effects, while posterior draws reveal which sequences support high-value feasible trees. Consequently, both explanations and assays are tied to the same construction-coupled decision.

3 PROBLEM FORMULATION

Budgeted setting. Let $\mathcal{X}$ be a finite sequence library and $\mathcal{C}$ a family of related conditions. The hidden utility $f(x, c) \in [0, 1]$ is revealed one entry at a time: one query of sequence x under condition c consumes one assay. Starting from a designated scaffold x_0, a separate construction budget R limits which descendants can be built and which mutation trees are feasible. After at most B assays, a method returns a niche $C \subseteq \mathcal{C}$ with $|C| = M$ and a diversification tree rooted at x_0.

Generalist-to-specialist diversification. Let $\mathcal{X}_R \subseteq \mathcal{X}$ contain the individual sequences constructible from x_0 within budget R. Let $\mathcal{T}_R$ contain the feasible trees rooted at x_0 with between two and K non-root leaves, where feasibility accounts for allowed edits and charges every shared edge only once. For $T \in \mathcal{T}_R$, write ∂T for its leaves and define

$$V(C, T) := \frac{1}{M} \sum_{c \in C} \max_{x \in \partial T} f(x, c). \tag{1}$$

Thus, T captures mutations shared before branching into condition-specific descendants.

The relevant generalist is the best single sequence reachable with the same budget, while the condition-wise accessible ceiling chooses a separate reachable sequence for each condition:

$$G_R(C) := \max_{x \in \mathcal{X}_R} \frac{1}{M} \sum_{c \in C} f(x, c), \qquad U_R(C) := \frac{1}{M} \sum_{c \in C} \max_{x \in \mathcal{X}_R} f(x, c). \tag{2}$$

We restrict attention to niches with $U_R(C) - G_R(C) \geq \tau > 0$, ensuring sufficient headroom for meaningful normalization. The score and selected repertoire are

$$\Phi_R(C, T) := \frac{V(C, T) - G_R(C)}{U_R(C) - G_R(C)}, \qquad (C^\star, T^\star) \in \arg\max_{C,T} \Phi_R(C, T). \tag{3}$$

The tree budget couples the repertoire members through their shared mutational history: choosing one descendant changes which other descendants remain reachable. Optimizing the niche, branching structure, and terminal sequences is therefore a joint graph and coverage problem.

Biological interpretation. A promiscuous enzyme can serve as x_0, substrates as the conditions, and the leaves of T as variants specialized through a shared directed-evolution campaign. Promiscuous activities often seed new enzymatic functions (Aharoni et al., 2005), while epistasis and permissive substitutions constrain which functional descendants are accessible (Starr et al., 2017). The objective therefore asks whether a fixed mutation budget is better spent on one broad enzyme or on a small family of specialists with complementary activity profiles across the selected substrate niche.

4 METHODOLOGY

EGRET separates model-guided exploration from measured-value commitment. From complete seed profiles, it repeatedly updates a site-coherent posterior, samples feasible lineages, adapts screening width to the condition panel, and constructs a verified tree. These are not independent modules: the hierarchical posterior generates complete landscapes, constrained optimization pushes those draws into leaf probabilities, and the leaf probabilities are converted into cell-level assays using the remaining lineage capacity and the condition spectrum. This separation matters because model error is amplified both by condition-wise maximization and by coupling the resulting champions under a shared-edge budget.

4.1 SITE-COHERENT EPISTATIC POSTERIOR

An interpretable conditional landscape. Let $z(x) = (1, z_{\text{site}}(x), z_{\text{res}}(x), z_{\text{pair}}(x))$ indicate edited sites, observed residue substitutions, and data-supported substitution pairs. EGRET decomposes the coefficient in condition c into a consensus effect and a condition-specific deviation:

$$\boldsymbol{\theta}_c = \bar{\boldsymbol{\theta}} + \boldsymbol{\delta}_c, \qquad y_{xc} := f(x, c) = z(x)^\top \boldsymbol{\theta}_c + \epsilon_{xc}, \quad \epsilon_{xc} \sim \mathcal{N}(0, \sigma_c^2). \tag{4}$$

Equation 4 is explainable by construction: its dot product attributes a prediction to sites, residues, and residue pairs, while $\bar{\boldsymbol{\theta}}$ identifies effects shared across conditions and $\boldsymbol{\delta}_c$ isolates condition-specific departures. This retains the goal of interpretable genotype–phenotype landscapes (Tonner et al., 2022), but makes the attribution directly relevant to shared trunks and specialized branches.

From hierarchy to posterior. Let $\mathcal{D} = \{(x, c, y_{xc})\}$ denote all assays observed so far. The Gaussian likelihood in Eq. 4 and a structured Gaussian prior jointly give the negative log-posterior, up to an irrelevant additive normalization constant,

$$-\log p(\bar{\boldsymbol{\theta}}, \{\boldsymbol{\delta}_c\} \mid \mathcal{D}) \doteq \frac{1}{2} \sum_{(x,c)\in\mathcal{D}} \frac{(y_{xc} - z(x)^\top \boldsymbol{\theta}_c)^2}{\sigma_c^2} + \frac{\lambda}{2} \left(\|\bar{\boldsymbol{\theta}}\|_{P_\rho}^2 + s \sum_c \|\boldsymbol{\delta}_c\|_{P_\rho}^2 \right). \tag{5}$$

Here $\|u\|_P^2 = u^\top P u$ and $P_\rho = \mathrm{diag}(0, I_{\mathrm{site}}, \rho I_{\mathrm{res}}, \rho I_{\mathrm{pair}})$. Thus ρ shrinks residue and pair corrections toward their site effects, whereas the fixed ratio $s > 1$ shrinks condition deviations toward the consensus. The ridge $\lambda > 0$ and site-coherence ratio $\rho \geq 1$ are selected by generalized cross-validation over grids induced by the same ridge candidates; the identity matrices act on their named feature blocks. The condition residual scale σ_c is estimated from the fit. A residue coordinate is created only after that substitution is observed. Before then, the site coordinate supplies the mean, while residue-level uncertainty must still be restored.

Calibrated posterior prediction. The quadratic form in Eq. 5 yields a Gaussian posterior with mean $\widehat{\boldsymbol{\theta}}_{c,\mathcal{D}}$ and covariance $\Sigma_{c,\mathcal{D}}$. Linear propagation through Eq. 4 therefore gives

$$\begin{aligned} \mathbb{E}[\widetilde{f}(x,c) \mid \mathcal{D}] &= \underbrace{z(x)^\top \widehat{\boldsymbol{\theta}}_{c,\mathcal{D}}}_{\text{hierarchical mean}}, \\ \mathrm{Var}[\widetilde{f}(x,c) \mid \mathcal{D}] &= \underbrace{z(x)^\top \Sigma_{c,\mathcal{D}} z(x)}_{\text{posterior uncertainty}} + \underbrace{r(x)\tau_c^2}_{\text{unseen-residue uncertainty}}. \end{aligned} \tag{6}$$

Here $r(x)$ counts substitutions without an observed residue coordinate and τ_c^2 is their empirical residue-level variance, floored by the variance implied by the fitted residue prior. The residual scale σ_c^2 already scales the coefficient posterior through Eq. 5; it is not added again because $\widetilde{f}$ denotes a draw of the fitted landscape, not a new residual realization. Thus shrinkage may transfer a site's mean to an unseen residue, but not the site's confidence. A joint draw now produces the complete random landscape optimized in the next subsection. The Gaussian update itself is standard; EGRET's contribution is the combination of observed-only residue coordinates, site-level fallback, and restored residue-level variance, which makes an unseen substitution informative without making it spuriously certain.

4.2 Posterior Lineage Sampling

Decision-aware acquisition. The posterior in Eq. 6 is over landscapes, while the downstream decision is a feasible tree. Write $T_C^\star(g)$ for the highest-value feasible tree under landscape g. For posterior draw s, let $\widetilde{T}_C^{(s,j)}$ be the tree of zero-based rank j, with $j = 0$ denoting the optimum, and let $\mathcal{G}_C^{(s)}$ contain the strongest candidates by mean utility across C under that draw. EGRET first pushes the posterior through tree optimization, then smooths the resulting leaf support:

$$\begin{aligned} \pi_C(x) &:= \Pr\Big\{ x \in \partial T_C^\star(\widetilde{f}) \mid \mathcal{D} \Big\} \approx \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\Big\{ x \in \partial \widetilde{T}_C^{(s,0)} \Big\}, \\ a_C(x) &:= \frac{1}{S} \sum_{s=1}^{S} \left[\sum_{j=0}^{J-1} \frac{\mathbf{1}\{x \in \partial \widetilde{T}_C^{(s,j)}\}}{1+j} + \beta\, \mathbf{1}\{x \in \mathcal{G}_C^{(s)}\} \right]. \end{aligned} \tag{7}$$

The trees $\widetilde{T}_C^{(s,j)}$ are ordered by decreasing value in Eq. 1 over $\mathcal{T}_R$ under $\widetilde{f}^{(s)}$. The principal term $\pi_C(x)$ is therefore a posterior probability over a *decision*, not an uncertainty score for $f(x, c)$. The actual acquisition $a_C(x)$ retains J trees with harmonic decay and adds the generalist guard with fixed weight $\beta = 0.35$ in all experiments. Thus a candidate is favored only when it repeatedly participates in an affordable repertoire, while near ties and broadly useful variants are not discarded. This differs from selecting one super-arm per posterior draw (Wang & Chen, 2018): EGRET aggregates support across sampled, near-optimal lineages before buying any assay. Proposition C.1 shows that this distinction can be strict even for $M = K = 2$: mean-landscape planning may select a lineage whose posterior optimality decays exponentially in the number of competing modes, whereas repeated lineage draws recover those modes at a quantified rate.

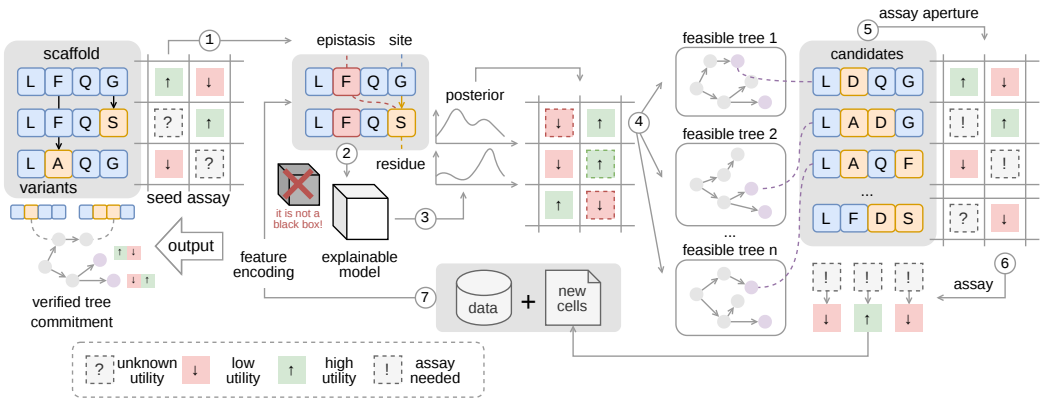


Figure 1: EGRET overview. Complete seed profiles are used to fit a site-coherent epistatic posterior over sequence utilities. Posterior samples are passed to a cost-aware tree search that identifies constructible repertoires and ranks candidate sequences. Adaptive screening allocates a limited assay budget across sequences and conditions. Measured utilities then determine the final repertoire commitment, closing the loop from landscape inference to experimental selection.

Reachability-aware candidate generation. The candidate pool is itself drawn from the posterior: mean ranking can exclude an uncertain site before tree optimization ever sees it. Feasibility adds a second correction. If $m(x)$ is the mutation set of x, reaching it consumes at least $\lceil |m(x)|/w \rceil$ width-w rounds, and each additional leaf uses at least one of the remaining rounds. Therefore any feasible tree containing x obeys

$$|\partial T| \le \kappa(x) := \min\left\{K, R - \left\lceil \frac{|m(x)|}{w} \right\rceil + 1\right\}. \tag{8}$$

$\kappa(x)$ is not another utility score: it is the maximum repertoire size that remains possible after admitting x. EGRET preserves the ranking from the acquisition $a_C(x)$, but reserves one of every K slots for $\kappa(x) = K$. This is the smallest floor that keeps a full K-leaf solution in the search without replacing posterior evidence by a depth heuristic.

Cost-aware tree construction. The optimizer in Eq. 7 groups one-endpoint expansions by cheapest feasible construction cost and reserves beam capacity for each represented cost. A high-valued but exhausted partial tree therefore cannot eliminate a cheaper state that can still add a specialist. For fixed endpoints, EGRET moves shared mutations toward the trunk and aligns branch points with edit-round boundaries, lowering cost without changing leaf utilities; the same search supplies both posterior leaf support and the final constructible tree.

4.3 Spectral Assay Aperture and Adaptive Allocation

Data-adaptive screening width. Equation 7 ranks sequence rows, whereas the oracle charges individual sequence–condition cells. Opening more conditions per row therefore trades screening breadth for profiling depth. Let R_C be the condition correlation matrix from complete seed profiles, and let B_{screen} be the cells available to the screening lane in the current round. Its participation ratio directly determines the aperture and hence the number of screened rows:

$$h(C) := \text{clip}_{[1,M]}\left(\text{round}\,\frac{(\text{tr}\,R_C)^2}{\|R_C\|_F^2}\right), \qquad n_{\text{screen}}(C) := \left\lfloor \frac{B_{\text{screen}}}{h(C)} \right\rfloor. \tag{9}$$

Here $\text{clip}_{[1,M]}$ truncates its rounded argument to the integer interval $[1, M]$; when a correlation matrix cannot yet be estimated, EGRET uses the conservative full aperture $h(C) = M$. The participation ratio equals the number of active modes when they are equally strong, so $h(C)$ moves from one assay for a redundant panel to M assays for independent conditions. The second expression follows from the cell constraint $h(C)n_{\text{screen}}(C) \le B_{\text{screen}}$. EGRET takes the $n_{\text{screen}}(C)$ rows with largest $a_C(x)$. It remains to decide which $h(C)$ cells to open within each selected row. Let $A(x, c) \in \{0, 1\}$ indicate whether a cell has already been assayed. Among the unopened cells,

EGRET chooses the $h(C)$ conditions with largest predicted improvement $\widehat{f}(x, c)$ over the best value already measured in condition c; if fewer remain, it opens all of them. Consequently, Eq. 7 chooses *which sequences*, Eq. 9 chooses *how many conditions*, and measured improvement chooses *which conditions*. This chain converts posterior support for a feasible tree into a concrete cell-level assay batch without introducing another acquisition score.

Promotion and local refinement. Global correlation need not persist in the specialist upper tail, so the narrow aperture in Eq. 9 can leave a promising sequence measured only where it was first selected. Promotion closes this gap by profiling top endpoint candidates among partially observed rows, allowing them to compete across conditions using evidence rather than posterior completion. Its connection to the tree objective is the measured marginal coverage

$$\Delta_{\mathcal{D}}(x \mid C, T) := \frac{1}{M} \sum_{c \in C} \left[A(x, c) y_{xc} - \max_{z \in \partial T} A(z, c) y_{zc} \right]_+, \tag{10}$$

where $[u]_+ = \max\{u, 0\}$. A partial leader contributes whenever it exceeds the current measured leaf coverage; completing its row tests whether that gain extends across the niche.

Once a verified tree exists, EGRET spends remaining follow-up capacity on one-substitution leaf replacements that preserve the budget. This evaluates the same marginal gain near the measured tree without discarding its shared trunk. Both lanes run over finite candidate lists to saturation; unused cells return to posterior lineage sampling rather than defining a separate acquisition score.

4.4 Verified Commitment

Measurement-backed selection. For an assayed cell, $y_{xc} = f(x, c)$ is known. Since utilities lie in $[0, 1]$, assigning zero to an unmeasured cell gives a lower bound preserved by maximization. Let $\widehat{G}_R(C)$ and $\widehat{U}_R(C)$ be the accessible anchors estimated during search. EGRET commits by replacing the unknown tree value in Eq. 3 with its measured certificate:

$$\begin{aligned} \underline{V}_{\mathcal{D}}(C, T) &:= \frac{1}{M} \sum_{c \in C} \max_{x \in \partial T} A(x, c) y_{xc} \le V(C, T), \\ (\widehat{C}, \widehat{T}) &\in \arg \max_{C, T \in \mathcal{T}_R} \frac{\underline{V}_{\mathcal{D}}(C, T) - \widehat{G}_R(C)}{\widehat{U}_R(C) - \widehat{G}_R(C)}. \end{aligned} \tag{11}$$

The cross-niche comparison inherits uncertainty from the estimated anchors, but the reported value of every committed tree contains no imputed leaf utility and needs at most KM measurements.

Connection to acquisition. The commitment rule is the endpoint of the same chain, rather than an independent verification module. Posterior lineage sampling prioritizes sequences likely to support the optimal feasible tree, the aperture buys early evidence across more sequences, and promotion completes only the profiles that can change the tree. Proposition C.2 formalizes this connection: the posterior regret of restricting commitment to fully measured sequences is controlled by the optimal-tree leaf probability left outside that measured pool. Hence the model decides which evidence to acquire, while Eq. 11 ensures that no prediction is counted as terminal evidence.

5 Experiments

Evaluation setup. We compare EGRET with nine literature baselines spanning submodular tree orienteering (STO) (Ghuge & Nagarajan, 2020), combinatorial Bayesian optimization (BOCS) (Baptista & Poloczek, 2018), graph-product BO (COMBO) (Oh et al., 2019), categorical trust-region search (Casmopolitan) (Wan et al., 2021), protein-specific proximal exploration (PEX) (Ren et al., 2022), probabilistic reparameterization (PR) (Daulton et al., 2022), tensor-output bandit BO (CBBO) (Huang et al., 2026), set-level Bayesian algorithm execution (DiscoBAX) (Lyle et al., 2023), and game-theoretic combinatorial BO (GameOpt) (Bal et al., 2025); a budget-matched random control completes the comparison. We preserve each source method's model and acquisition, adapting only the interface needed to return a niche, a reachable mutation tree, and a cell batch; the full evaluation protocol is given in Appendix D, with method-specific adaptations in Appendix D.2.

Table 1: Normalized headroom capture Φ_R on the benchmark (8% assay budget); five-seed means (standard deviations), with bold denoting the best per instance and summary row.

Dataset *Venue*	Random –	STO SODA 2020	BOCS ICML 2018	COMBO NeurIPS 2019	Casmopolitan ICML 2021	PEX ICML 2022	PR NeurIPS 2022	CBBO ICLR 2026	DiscoBAX ICML 2023	GameOpt ICLR 2025	**EGRET** (ours)
AncSR1	−0.763 (0.338)	0.062 (0.197)	−0.015 (0.194)	0.006 (0.311)	0.066 (0.489)	−0.084 (0.342)	0.123 (0.255)	0.210 (0.039)	0.436 (0.233)	0.106 (0.272)	**0.839** (0.176)
AncSR2	−1.431 (0.156)	0.493 (0.285)	0.664 (0.110)	0.413 (0.198)	0.447 (0.206)	0.306 (0.423)	0.357 (0.233)	−0.041 (0.245)	0.819 (0.083)	0.789 (0.144)	**0.859** (0.139)
Azoarcus	−13.285 (2.063)	−0.260 (0.526)	0.405 (0.366)	0.060 (0.074)	0.222 (0.152)	0.693 (0.236)	−0.285 (0.510)	−0.015 (0.029)	0.098 (0.196)	0.587 (0.180)	**0.959** (0.025)
CH65+Expr	−2.894 (1.333)	0.175 (0.296)	−0.176 (0.243)	0.308 (0.086)	0.308 (0.086)	−0.043 (0.286)	0.143 (0.046)	0.285 (0.075)	−0.112 (0.487)	0.198 (0.133)	**0.864** (0.000)
CH65	−6.841 (1.849)	−0.094 (0.333)	−0.419 (0.311)	−0.094 (0.333)	−0.094 (0.333)	−0.029 (0.601)	−0.425 (0.395)	−0.094 (0.333)	−0.280 (0.502)	0.412 (0.494)	**0.754** (0.000)
CR9114	−4.986 (0.634)	0.440 (0.156)	0.677 (0.252)	0.440 (0.156)	0.440 (0.156)	0.457 (0.032)	0.289 (0.201)	0.440 (0.156)	0.305 (0.326)	0.685 (0.000)	**0.887** (0.041)
DAOx	−3.137 (0.889)	0.216 (0.139)	−1.318 (0.901)	0.216 (0.139)	0.224 (0.137)	0.201 (0.161)	−0.022 (0.280)	0.223 (0.138)	0.274 (0.274)	−0.394 (0.399)	**0.363** (0.121)
FP switch	−0.501 (0.368)	0.467 (0.286)	0.635 (0.380)	0.488 (0.301)	0.570 (0.297)	0.807 (0.042)	0.407 (0.366)	0.613 (0.202)	0.605 (0.315)	0.308 (0.418)	**0.887** (0.000)
Hsp90	−1.261 (0.389)	0.387 (0.453)	−0.581 (0.556)	0.285 (0.444)	0.347 (0.439)	0.020 (0.429)	−0.087 (0.373)	0.543 (0.101)	0.022 (0.253)	−0.445 (0.669)	**0.571** (0.048)
KCNE1	−1.469 (0.762)	0.650 (0.125)	−0.051 (0.158)	**0.698** (0.071)	**0.698** (0.071)	0.398 (0.432)	0.389 (0.406)	**0.698** (0.071)	0.450 (0.527)	−0.018 (0.165)	0.660 (0.178)
KCNJ2	−1.495 (0.501)	0.339 (0.122)	**0.504** (0.189)	0.252 (0.131)	0.252 (0.131)	0.119 (0.313)	0.119 (0.313)	0.252 (0.131)	0.282 (0.252)	0.325 (0.222)	0.499 (0.130)
KRAS	−2.907 (0.672)	−0.332 (0.118)	0.108 (0.154)	−0.332 (0.118)	−0.285 (0.112)	−0.402 (0.210)	−0.447 (0.353)	−0.332 (0.118)	−0.347 (0.205)	−1.271 (0.426)	**0.228** (0.112)
Omicron escape	−3.643 (1.960)	0.575 (0.230)	−0.452 (0.201)	0.575 (0.230)	0.575 (0.230)	0.575 (0.230)	0.575 (0.230)	0.575 (0.230)	0.575 (0.230)	0.575 (0.230)	**0.751** (0.306)
Omicron+ACE2	−1.571 (1.449)	0.599 (0.198)	−0.043 (0.152)	0.599 (0.198)	0.609 (0.192)	0.484 (0.177)	0.465 (0.185)	0.599 (0.198)	0.455 (0.137)	0.350 (0.099)	**0.839** (0.197)
Potts	−2.407 (0.447)	0.106 (0.211)	0.140 (0.308)	0.185 (0.145)	0.135 (0.202)	0.068 (0.254)	0.165 (0.150)	−0.001 (0.160)	0.492 (0.085)	−0.043 (0.142)	**0.677** (0.021)
BA.2 backgrounds	−1.278 (0.890)	0.129 (0.124)	−0.165 (0.305)	−0.083 (0.184)	0.437 (0.154)	0.476 (0.104)	−0.193 (0.369)	0.216 (0.052)	0.517 (0.105)	−0.125 (0.269)	**0.598** (0.057)
JN.1/XBB	−0.754 (0.239)	0.049 (0.134)	−0.267 (0.137)	0.059 (0.059)	0.022 (0.108)	0.022 (0.108)	0.022 (0.108)	−0.069 (0.109)	0.114 (0.156)	−0.114 (0.123)	**0.134** (0.166)
Lineage+ACE2	−1.371 (0.628)	0.331 (0.089)	0.587 (0.024)	0.426 (0.046)	0.426 (0.046)	**0.617** (0.000)	0.436 (0.055)	0.426 (0.046)	0.563 (0.061)	0.336 (0.214)	0.617 (0.008)
Lineage escape	−3.028 (0.770)	−0.027 (0.309)	0.291 (0.104)	0.129 (0.314)	0.223 (0.197)	0.401 (0.000)	0.251 (0.094)	0.223 (0.197)	0.342 (0.092)	0.274 (0.102)	**0.401** (0.027)
S309 backgrounds	−0.740 (0.293)	0.223 (0.150)	0.278 (0.230)	0.239 (0.055)	0.167 (0.161)	0.290 (0.148)	0.184 (0.108)	0.050 (0.210)	0.382 (0.118)	0.176 (0.254)	**0.428** (0.060)
Mean / instances led	−2.788 / 0	0.226 / 0	0.040 / 1	0.243 / 1	0.290 / 1	0.269 / 1	0.123 / 0	0.240 / 1	0.300 / 0	0.136 / 0	**0.641 / 17**

Table 2: **Decision-quality diagnostics.** Macro-averages over twenty instances and five seeds at an 8% assay budget. Columns separate realized quality from observational support, then report niche ambition, specialist diversity, and runtime; bold marks the best non-random value reported in each diagnostic column.

	Headline			Evidence behind the commitment			Niche ambition		Repertoire		Cost
Method	$\Phi_R \uparrow$	$A \uparrow$	Rank↓	$\Phi_{R,\mathrm{cert}} \uparrow$	Measured↑	$\lvert\hat{\Phi}_R - \Phi_R\rvert \downarrow$	$\Phi_{R,\mathrm{orc}} \uparrow$	Price↓	$q^+ \uparrow$	$N_{\mathrm{eff}} \uparrow$	s/run↓
Random	−2.788	−0.308	11.00	−7.446	0.074	–	0.846	0.459	0.117	1.73	0.0
STO	0.226	0.044	6.05	0.226	**1.000**	0.495	0.842	0.110	0.565	2.15	4.8
BOCS	0.040	0.032	6.50	−2.851	0.433	0.638	0.842	0.121	0.462	2.05	20.5
COMBO	0.243	0.046	5.58	0.227	0.985	0.505	0.839	0.109	0.553	2.15	11.6
Casmopolitan	0.290	0.053	5.17	0.277	0.984	0.474	0.856	0.105	0.571	**2.16**	14.0
PEX	0.269	0.050	5.55	0.112	0.860	0.563	0.856	0.103	0.536	2.08	24.0
PR	0.123	0.033	7.55	−0.407	0.798	0.575	0.835	0.121	0.487	1.95	13.2
CBBO	0.240	0.040	5.85	0.220	0.988	0.414	0.787	0.100	0.549	2.11	**4.1**
DiscoBAX	0.300	0.065	4.78	0.239	0.905	0.385	0.824	0.079	0.501	1.94	6.9
GameOpt	0.136	0.033	6.72	−1.625	0.615	0.615	0.850	0.116	0.514	1.92	102.6
EGRET	**0.641**	**0.112**	**1.25**	**0.640**	0.948	**0.193**	**0.880**	**0.048**	**0.603**	2.16	10.9

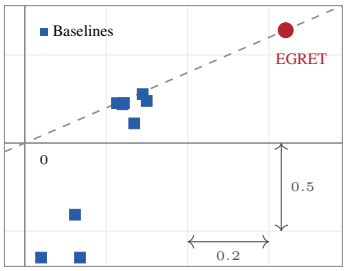


Figure 2: **Measured agreement.** Realized score versus the score supported by measured cells. EGRET is red; the dashed line marks equality.

5.1 Benchmark Comparison

Headline comparison. Table 1 reports Φ_R on twenty instances. On average across instances, EGRET exceeds the strongest per-instance baseline by 0.141 and attains mean rank 1.25, compared with 4.78 for the next-ranked method. It leads on $17/20$ instances; the three losses occur against different methods, and two are smaller than 0.01. Baseline rankings correlate across instances (mean Spearman 0.53), showing shared patterns of difficulty.

Decision quality diagnostics. Table 2 decomposes the headline result into evidence, niche ambition, repertoire structure, and cost. EGRET leads $17/20$ instances, ranks first on average, and its measured-only score nearly equals its realized score with 94.8% of the committed block observed. It also retains the highest oracle score at the lowest lineage price, showing that its gain survives both measurement and construction rather than existing only in surrogate predictions. Its runtime remains moderate; GameOpt costs nearly an order of magnitude more without matching the decision quality. Figure 2 makes the evidence gap explicit: EGRET lies almost exactly on the identity line.

5.2 Robustness Interventions

Matched intervention study. Table 3 pairs seven interventions with their untouched controls. EGRET retains a positive margin under reduced libraries, weaker epistasis, condition-wise calibration, and high condition conflict, although halving the library narrows it. Destroying mutational locality removes the advantage, while held-out conditions yield a near-zero, inconclusive margin, identifying the main transfer limits. At $\sigma = 0.10$, all methods collapse toward zero and EGRET no longer separates from the envelope. Here, rival denotes the non-random per-instance envelope, n the number of eligible instances, and m the number of compared methods; bootstrap intervals quantify uncertainty, while "led" counts instances won by EGRET against that envelope.

Table 3: **Robustness interventions.** Matched control–intervention comparisons across seven benchmark perturbations; bold differences have 95% intervals excluding zero.

			Untouched control					Under intervention				
Intervention	n	m	setting	EGRET	rival	EGRET − rival [95% CI]	led	setting	EGRET	rival	EGRET − rival [95% CI]	led
Assay-observation noise	20	11	$\sigma = 0$	0.631	0.450	**+0.181** [+0.106, +0.262]	18/20	$\sigma = 0.10$	−0.006	0.026	−0.032 [−0.116, +0.055]	9/20
Candidate-library coverage	19	11	full library	0.635	0.504	**+0.131** [+0.069, +0.202]	16/19	half retained	0.678	0.604	**+0.073** [+0.015, +0.135]	12/19
Epistasis strength	11	11	$\lambda = 1$	0.648	0.530	**+0.118** [+0.036, +0.228]	8/11	$\lambda = 0$	0.714	0.625	**+0.090** [+0.024, +0.166]	7/11
Condition-wise calibration	19	4	$s = 1.0$	0.624	0.412	**+0.212** [+0.121, +0.318]	17/19	$s = 2.0$	0.556	0.384	**+0.172** [+0.085, +0.268]	14/19
Mutational locality	16	11	canonical	0.670	0.498	**+0.172** [+0.102, +0.251]	15/16	destroyed	0.430	0.450	−0.020 [−0.134, +0.113]	5/16
Condition visibility	9	6	observed	0.553	0.380	**+0.173** [+0.024, +0.333]	8/9	held out	0.256	0.250	+0.005 [−0.054, +0.066]	4/9
Condition conflict	9	4	low	0.566	0.278	**+0.289** [+0.149, +0.445]	9/9	high	0.475	0.306	**+0.169** [+0.064, +0.329]	9/9

Table 4: **Locality response.** Mean Φ_R before and after locality destruction; D_m is the paired change.

Method	Canonical	Destroyed	D_m
Random	−3.025	−2.943	0.082
STO	0.184	0.198	0.014
BOCS	0.141	0.090	−0.051
COMBO	0.214	0.254	0.040
Casmopolitan	0.267	0.252	−0.015
PEX	0.290	0.218	−0.072
PR	0.129	0.146	0.017
CBBO	0.193	0.230	0.037
DiscoBAX	0.310	0.234	−0.076
GameOpt	0.203	0.155	−0.048
EGRET (ours)	**0.670**	**0.430**	−0.241

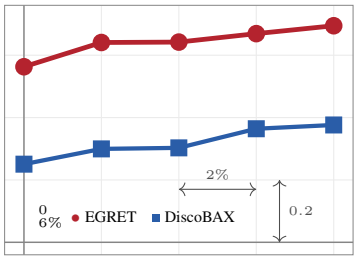


Figure 3: **Budget sensitivity.** Mean Φ_R across five assay budgets from 6% to 14%.

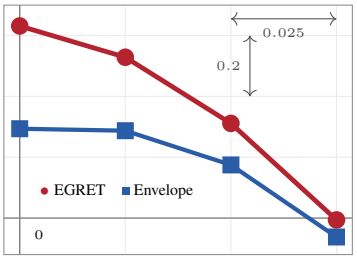


Figure 4: **Noise sensitivity.** Mean Φ_R across paired assay corruption levels $\sigma = 0$–0.10.

Failure mechanisms. Table 4 isolates the structural assumption behind EGRET's advantage across sixteen eligible instances and five seeds. Its score falls from 0.670 to 0.430 when the same variants are reassigned to destroy mutational locality, eliminating the 0.172 canonical margin. Figure 3 shows that this behavior is not an artifact of the chosen assay budget: EGRET leads DiscoBAX from 6% through 14% and improves as more cells become observable. Figure 4 exposes a different limit. Both methods approach zero as corruption grows, and their separation becomes inconclusive at $\sigma = 0.10$. The three interventions therefore distinguish missing lineage structure, scarce observations, and unreliable observations.

5.3 Generalization Stress Tests

Three observation interventions. The parenthesized numbers in the column-group headers are eligible-instance counts, not additional metrics. *(i) Support.* The matched-support columns give every method the mutation-token coordinates induced by EGRET's structured representation, while "features held out" removes that namespace; the shrinking margin isolates the value of the representation itself. *(ii) Precision.* The readout is quantized to 64, 32, 16 or 8 levels only on the method-visible side, so the hidden evaluator remains canonical; EGRET stays separated through 16 levels but not at 8. *(iii) Conditions.* A tree is constructed without one or more conditions and then rescored on them, testing transfer beyond the observed panel. The hidden matrix, candidate space and evaluator stay fixed throughout; only acquisition-time information changes. Figure 5 complements this test by adding conditions; its intervals expose seed variability against the strongest baseline.

Table 5: **Observation stress tests.** Mean Φ_R under restricted support, precision and condition visibility, with the hidden evaluator and candidate space held fixed throughout.

	Support (18)		Precision, levels (20)				Conditions (5 / 4)	
Method	matched	held out	64	32	16	8	1	2–3
Random	−2.853	−2.853	−2.788	−2.788	−2.788	−2.788	–	–
STO	0.117	−0.126	0.240	0.205	0.178	0.067	–	–
BOCS	−0.025	−0.210	−0.011	0.008	0.066	−0.033	–	–
COMBO	0.210	−0.080	0.243	0.210	0.184	0.110	0.132	0.068
Casmopolitan	0.188	−0.075	0.252	0.250	0.239	0.091	0.144	0.101
PEX	0.157	−0.095	0.257	0.255	0.203	0.165	0.159	0.091
PR	0.054	−0.237	0.122	0.065	0.107	0.039	–	–
CBBO	0.176	−0.018	0.226	0.232	0.183	0.108	0.083	0.034
DiscoBAX	0.229	0.025	0.301	0.287	0.263	0.248	0.225	0.189
GameOpt	0.108	−0.188	0.083	0.079	0.106	0.002	–	–
EGRET	**0.495**	**0.245**	**0.628**	**0.609**	**0.565**	**0.423**	**0.250**	**0.262**

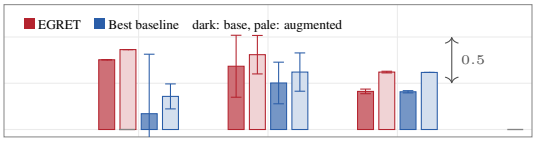


Figure 5: **Panel augmentation.** Mean $\Phi_R \pm$SD over ten seeds against the strongest baseline.

Table 6: **Interpretable model comparison.** Six accounts fitted to the same frozen assay ledger. ρ_{out} and RMSE score cells the campaign never acquired and $\rho_{d\geq 3}$ only distant variants; Δ_{add} and ρ_{pair} probe the interaction block; ρ and stability are agreement with hidden marginal effects and across seeds; nomination and $K^\star$ are the best repertoire from the top 24 variants and the substitutions reaching 90% of it. Term counts are scale descriptors only.

		Prediction			Interaction		Faithfulness		Decision		
Account	Form	ρ_{out}	RMSE	$\rho_{d\geq 3}$	Δ_{add}	ρ_{pair}	ρ	stability	nomination	$K^\star$	terms
EGRET	site+residue+pairs	**+0.744**	**0.043**	**+0.675**	+0.056	+0.140	+0.709	+0.824	**0.801**	10	301
RFA	ref-free 1st+2nd	+0.716	0.051	+0.650	+0.034	**+0.211**	+0.704	+0.708	0.746	13	152
CoVES	site×residue	+0.690	0.058	+0.636	+0.000	—	**+0.789**	+0.778	0.714	14	150
SQUID	local additive+pairs	+0.737	0.045	+0.662	+0.119	+0.126	+0.700	+0.804	0.777	8	306
AutoSciLab	latent equation	+0.371	0.096	+0.290	*n/a*	—	+0.277	**+0.885**	0.589	16	4
IDEAS	position saliency	+0.599	0.069	+0.545	*n/a*	—	+0.168	+0.450	0.720	11	28 283

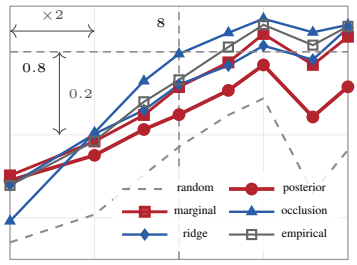


Figure 6: **Explanation-guided construction.** Coverage using only the top-K substitutions named by each account.

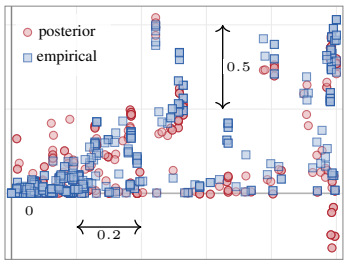


Figure 7: **Committed-edge calibration.** Predicted versus measured coverage loss after cutting a tree edge.

Table 7: **Prospective reliability.** Faithfulness of each account refit from a prefix of the oracle's assay log; Δ is the gain from 25% to completion.

	ρ with hidden effects, at log fraction				
Account	25%	50%	75%	100%	Δ
EGRET marginal	**+0.781**	**+0.839**	**+0.867**	**+0.892**	+0.111
Ridge on log	+0.726	+0.775	+0.794	+0.811	+0.086
LIME	+0.723	+0.773	+0.794	+0.811	+0.087
Permutation	+0.691	+0.755	+0.783	+0.794	+0.103
Empirical marginal	+0.666	+0.737	+0.771	+0.803	+0.136
EGRET posterior	+0.640	+0.685	+0.693	+0.709	+0.069
Hashed surrogate	+0.633	+0.659	+0.677	+0.687	+0.055
EGRET site half	+0.551	+0.597	+0.604	+0.617	+0.066
Occlusion	+0.433	+0.462	+0.465	+0.473	+0.040
Random	+0.005	−0.007	−0.038	+0.041	+0.036

5.4 Actionable Explanation Analysis

Controlled account evaluation. We freeze the assay ledger and committed tree from each EGRET campaign, and then fit every account to the same measurements under the same second-order budget. The comparison includes RFA (Park et al., 2024), CoVES (Ding et al., 2024a), SQUID (Seitz et al., 2024), AutoSciLab (Desai et al., 2025), and IDEAS (Pandey et al., 2026). Differences in performance therefore arise from the accounts, rather than from differences in the underlying evidence.

In Table 6, EGRET leads all three predictive diagnostics and nomination, giving the strongest combination of extrapolation and decision utility. RFA, CoVES, and AutoSciLab each lead one specialized diagnostic, but none carries it through the construction decision.

Decision-facing explanation tests. Figure 6 builds only from the substitutions each account names, Figure 7 compares predicted edge costs with the measured loss of cutting those edges, and Table 7 refits each account from prefixes of the assay log. EGRET keeps its construction value, its posterior costs track measured edge effects, and its marginal account is the most faithful from the first quarter on. Appendices E.1, E.4, and E.7 add lineage, mutation, and per-campaign traces.

5.5 Ablation Studies

Table 8 isolates the landscape model while holding the rest of EGRET fixed. Condition pooling and the loss of pairwise interactions cause the clearest failures, and generic surrogates transfer poorly into the lineage decision. Predictive diagnostics do not track repertoire quality consistently, supporting an acquisition rule that evaluates models through feasible downstream trees. Several component-wise intervals include zero, so the evidence favors the structured posterior as a whole more than any component alone.

Table 8: **Posterior ablations.** Twenty datasets, five seeds; ρ_s is rank correlation, P@50 top-50 precision, and Cov_{80} interval coverage.

Variant	$\Phi_R\uparrow$		$\Delta\Phi_R$	$\rho_s\uparrow$	P@50↑	Cov_{80}
EGRET (full model)	**0.641**		**0.000**	**0.588**	0.283	0.366
Structural ablations						
Unpooled residue effects	0.617	−0.024 [−0.056, +0.005]	0.558	0.277	0.404
Additive mutation effects	0.566	−0.075 [−0.129, −0.026]	**0.588**	0.252	0.294
Independent condition models	0.614	−0.027 [−0.060, +0.003]	**0.588**	0.287	0.360
Fully pooled condition model	0.553	−0.088 [−0.179, −0.021]	0.438	0.168	0.369
Fixed shrinkage parameters	0.604	−0.037 [−0.088, +0.007]	0.579	0.225	0.353
Alternative surrogates						
Hashed-feature Bayesian ridge	0.540	−0.101 [−0.145, −0.059]	0.560	0.202	**0.926**
Hamming diffusion GP	0.523	−0.118 [−0.191, −0.053]	0.500	**0.289**	0.979
Hamming kernel regression	0.580	−0.060 [−0.118, −0.011]	0.489	0.152	0.665
Neural network ensemble	0.570	−0.071 [−0.164, −0.003]	0.519	0.256	0.374

AI USE STATEMENT

During the preparation of this paper, we drafted the manuscript and used large language models (LLMs), specifically ChatGPT and Claude, for language refinement and minor improvements to clarity and organization, guided by our drafts, ideas, and instructions. We also used these tools to support our literature review and to help identify potentially related prior work when assessing the novelty of our contribution. In addition, we designed and prepared the figures and tables, with these tools making only minor refinements to their presentation; all data shown were derived exclusively from actual experimental results, and no fabricated or synthetically generated experimental data were included in this paper.

All tool-assisted outputs were carefully reviewed and independently verified by us for factual accuracy, technical correctness, and consistency with the underlying experiments. The core ideas, methodology, experimental design, and scientific contributions of this work were conceived and developed by us. We have used these tools responsibly and transparently, in accordance with the principles of research integrity and with the goal of supporting progress in the research community.

REPRODUCIBILITY STATEMENT

We release supplementary materials needed to reproduce the main tables and figures. Upon acceptance, we will release the complete source code, experimental scripts, configurations, and execution logs for all reported results. We commit to providing the data processing and evaluation details necessary for independent reruns to reproduce the reported results within the stated variation.

REFERENCES

Amir Aharoni, Leonid Gaidukov, Olga Khersonsky, Stephen McQ. Gould, Cintia Roodveldt, and Dan S. Tawfik. The "evolvability" of promiscuous protein functions. *Nature Genetics*, 37(1): 73–76, 2005. doi: 10.1038/ng1482. URL https://doi.org/10.1038/ng1482.

Melis Ilayda Bal, Pier Giuseppe Sessa, Mojmir Mutny, and Andreas Krause. Optimistic games for combinatorial bayesian optimization with application to protein design. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/434743b59a2ef42d657e61e06f0dd97b-Abstract-Conference.html.

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 462–471. PMLR, 2018. URL https://proceedings.mlr.press/v80/baptista18a.html.

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-N protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, 2021. doi: 10.1038/s41592-021-01100-y.

Lea Bogensperger, Dominik Narnhofer, Ahmed Allam, Konrad Schindler, and Michael Krauthammer. A variational perspective on generative protein fitness optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 4700–4712. PMLR, 2025. URL https://proceedings.mlr.press/v267/bogensperger25a.html.

James C. Bowden, Sergey Levine, and Jennifer Listgarten. Leveraging discrete function decomposability for scientific design. In *Proceedings of the Fourteenth International Conference on Learning Representations*, volume 2026, pp. 74809–74845, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ab4881aee9d4f2ab458236364c937d-Abstract-Conference.html. Code: https://github.com/james-bowden/DADO.

Can Chen, Yingxue Zhang, Xue Liu, and Mark Coates. Bidirectional learning for offline model-based biological sequence design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 5351–5366. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ao.html.

Willow Coyote-Maestas, David Nedrud, Yungui He, and Daniel Schmidt. Determinants of trafficking, conduction, and disease within a K+ channel revealed through multiparametric deep mutational scanning. *eLife*, 11:e76903, 2022. doi: 10.7554/eLife.76903.

Samuel Daulton, Xingchen Wan, David Eriksson, Maximilian Balandat, Michael A. Osborne, and Eytan Bakshy. Bayesian optimization over discrete and mixed spaces via probabilistic reparameterization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12760–12774, 2022. doi: 10.52202/068431-0927. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/531230cfac80c65017ad0f85d3031edc-Abstract-Conference.html.

Saaketh Desai, Sadhvikas Addamane, Jeffrey Y. Tsao, Igal Brener, Laura P. Swiler, Remi Dingreville, and Prasad P. Iyer. AutoSciLab: A self-driving laboratory for interpretable scientific discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 146–154, 2025. doi: 10.1609/aaai.v39i1.31990. URL https://ojs.aaai.org/index.php/AAAI/article/view/31990.

David Ding, Ada Y. Shaw, Sam Sinai, Nathan Rollins, Noam Prywes, David F. Savage, Michael T. Laub, and Debora S. Marks. Protein design using structure-based residue preferences. *Nature Communications*, 15(1):1639, 2024a. doi: 10.1038/s41467-024-45621-4. URL https://www.nature.com/articles/s41467-024-45621-4.

Kerr Ding, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, Peng Liu, Yang Yang, and Yunan Luo. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nature Communications*, 15 (1):6392, 2024b. doi: 10.1038/s41467-024-50698-y. URL https://www.nature.com/articles/s41467-024-50698-y.

Andre J. Faure, Aina Martí-Aranda, Cristina Hidalgo-Carcedo, Antoni Beltran, Jörn M. Schmiedel, and Ben Lehner. The genetic architecture of protein stability. *Nature*, 634(8035):995–1003, 2024. doi: 10.1038/s41586-024-07966-0. URL https://www.nature.com/articles/s41586-024-07966-0.

Julia M. Flynn et al. Comprehensive fitness maps of Hsp90 show widespread environmental dependence. *eLife*, 9:e53810, 2020. doi: 10.7554/eLife.53810.

Chase R. Freschlin, Sarah A. Fahlberg, Pete Heinzelman, and Philip A. Romero. Neural network extrapolation to distant regions of the protein fitness landscape. *Nature Communications*, 15 (1):6405, 2024. doi: 10.1038/s41467-024-50712-3. URL https://www.nature.com/articles/s41467-024-50712-3.

Rohan Ghuge and Viswanath Nagarajan. Quasi-polynomial algorithms for submodular tree orienteering and other directed network design problems. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1039–1048, 2020. doi: 10.1137/1.9781611975994.63.

Peter Mørch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29514–29565, 2024. doi: 10.52202/079017-0929. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/34547650b2ca69d91f3b3c3ae8b21962-Abstract-Conference.html.

Mohammad Haddadnia, Yuvan Chali, Abhilash Jayaraj, Constance Kraay, Joana Reis, Felix Strieth-Kalthoff, and Haribabu Arthanari. Target-aware bandit allocation for scalable surrogate optimization in chemical space. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. doi: 10.48550/arXiv.2606.26657. URL https://doi.org/10.48550/arXiv.2606.26657. Code: https://github.com/Mohaddadnia/boba.

Michael J. Harms and Joseph W. Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203–207, 2014. doi: 10.1038/nature13410.

Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283, 2024. doi: 10.1038/s41587-023-01763-2.

Jingru Huang, Haijie Xu, Jie Guo, Manrui Jiang, and Chen Zhang. Combinatorial bandit bayesian optimization for tensor outputs. In *Proceedings of the Fourteenth International Conference on Learning Representations*, volume 2026, pp. 44648–44676, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a06bfb11c00bd8a2a1cd7a522138f86-Abstract-Conference.html.

Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, Julius Adebayo, Samuel Stanton, Héctor Corrada Bravo, Kyunghyun Cho, and Nathan C. Frey. Concept bottleneck language models For protein design. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8cf2d1a2fef23281af0f50d1a04dfd7c-Abstract-Conference.html.

Moksh Jain, Sharath Chandra Raparthy, Alex Hernández-García, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective GFlowNets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 14631–14653. PMLR, 2023. URL https://proceedings.mlr.press/v202/jain23a.html.

Jiaqi Luo, Kerr Ding, and Yunan Luo. Pareto-optimal sampling for multi-objective protein sequence design. *iScience*, 28(3):112119, 2025. doi: 10.1016/j.isci.2025.112119.

Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jesson, Stefan Bauer, Yarin Gal, and Patrick Schwab. DiscoBAX: Discovery of optimal intervention sets in genomic experiment design. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 23170–23189. PMLR, 2023. URL https://proceedings.mlr.press/v202/lyle23a.html.

Alief Moulana, Thomas Dupic, and Michael M. Desai. Epistasis and background dependence in the evolution of Omicron variants of the SARS-CoV-2 spike protein. *Molecular Biology and Evolution*, pp. msag100, 2026. doi: 10.1093/molbev/msag100.

Alief Moulana et al. The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution. *eLife*, 12:e83442, 2023. doi: 10.7554/eLife.83442.

Ayesha Muhammad et al. High-throughput functional mapping of variants in an arrhythmia gene, KCNE1, reveals novel biology. *Genome Medicine*, 16:73, 2024. doi: 10.1186/s13073-024-01340-5.

Changyong Oh, Jakub M. Tomczak, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using the graph cartesian product. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2cb6b10338a7fc4117a80da24b582060-Abstract.html.

Akash Pandey, Wei Chen, and Sinan Keten. Interpretability driven evolutionary approach for the design of biological sequences. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. Poster. Code: https://github.com/pandeyakash23/IDEAS.

Yeonwoo Park, Brian P. H. Metzger, and Joseph W. Thornton. The simplicity of protein sequence–function relationships. *Nature Communications*, 15(1):7953, 2024. doi: 10.1038/s41467-024-51895-5. URL https://www.nature.com/articles/s41467-024-51895-5.

Gianluca Peri, Clémentine Gibard, Nicholas H. Shults, Kent Crossin, and Eric J. Hayden. Dynamic RNA fitness landscapes of a group I ribozyme during changes to the experimental environment. *Molecular Biology and Evolution*, 39(3):msab373, 2022. doi: 10.1093/molbev/msab373.

Angela M. Phillips et al. Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife*, 10:e71393, 2021. doi: 10.7554/eLife.71393.

Angela M. Phillips et al. Hierarchical sequence-affinity landscapes shape the evolution of breadth in an anti-influenza receptor binding site antibody. *eLife*, 12:e83628, 2023. doi: 10.7554/eLife.83628.

Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10:4213, 2019. doi: 10.1038/s41467-019-12130-8.

Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, and Jian Peng. Proximal exploration for model-guided protein sequence design. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 18520–18536. PMLR, 2022. URL https://proceedings.mlr.press/v162/ren22a.html.

Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013. doi: 10.1073/pnas.1215251110.

Evan E. Seitz, David M. McCandlish, Justin B. Kinney, and Peter K. Koo. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nature Machine Intelligence*, 6(6):701–713, 2024. doi: 10.1038/s42256-024-00851-5. URL https://www.nature.com/articles/s42256-024-00851-5.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Tyler N. Starr, Lora K. Picton, and Joseph W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409–413, 2017. doi: 10.1038/nature23902.

Matt Sternke, Katherine W. Tripp, Soumya Prakash Behera, Justin P. Nguyen, and Doug Barrick. Protein stability is determined by single-site bias rather than pairwise covariance. *Nature Chemical Biology*, 22(8):1252–1261, 2026. doi: 10.1038/s41589-026-02270-6. URL https://www.nature.com/articles/s41589-026-02270-6.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. PepTune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 59017–59065. PMLR, 2025. URL https://proceedings.mlr.press/v267/tang25n.html.

Peter D. Tonner, Abe Pressman, and David Ross. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. *Proceedings of the National Academy of Sciences*, 119(26):e2114021119, 2022. doi: 10.1073/pnas.2114021119.

Rosario Vanella, Sean Boult, Christoph Küng, and Michael A. Nash. Decoding the substrate specificity landscape of a promiscuous enzyme through multi-substrate mutational scanning. *Nature Communications*, 17:3245, 2026. doi: 10.1038/s41467-026-69913-z.

Xingchen Wan, Vu Nguyen, Huong Ha, Binxin Ru, Cong Lu, and Michael A. Osborne. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 10663–10674. PMLR, 2021. URL https://proceedings.mlr.press/v139/wan21b.html.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5114–5122. PMLR, 2018.

Daniel M. Weinreich, Nigel F. Delaney, Mark A. DePristo, and Daniel L. Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, 2006. doi: 10.1126/science.1123539.

Chenchun Weng, André J. Faure, Albert Escobedo, and Ben Lehner. The energetic and allosteric landscape for KRAS inhibition. *Nature*, 626(7999):643–652, 2024. doi: 10.1038/s41586-023-06954-0.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658. URL https://ojs.aaai.org/index.php/AAAI/article/view/3982.

Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 116(18):8852–8858, 2019. doi: 10.1073/pnas.1901979116.

Jason Yang, Ravi G. Lal, James C. Bowden, Raul Astudillo, Mikhail A. Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H. Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, 2025. doi: 10.1038/s41467-025-55987-8. URL https://www.nature.com/articles/s41467-025-55987-8.

Hui Yuan, Chengzhuo Ni, Huazheng Wang, Xuezhou Zhang, Le Cong, Csaba Szepesvari, and Mengdi Wang. Bandit theory and thompson sampling-guided directed evolution for sequence optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38291–38304, 2022. doi: 10.52202/068431-2774. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa3c139cf8084de7bfd944f1c90c8695-Abstract-Conference.html.

APPENDIX CONTENTS

A Extended Related Work **16**

A.1 Positioning EGRET 16

A.2 Learning-Guided Protein Engineering 16

A.3 Functional Diversification 17

A.4 Interpretable Fitness Landscapes 18

A.5 Repertoire Optimization 18

A.6 Decision-Aware Experimental Design 19

B Algorithmic Specification **20**

B.1 Campaign Overview 20

B.2 Posterior Update 20

B.3 Posterior Lineage Search 21

B.4 Assay Allocation Pipeline 22

C Theoretical Foundation **24**

C.1 Analytical Derivations 24

C.2 Further Theoretical Guarantees 25

D Experimental Details **29**

D.1 Benchmark Inventory 29

D.2 Compared Baseline Adaptations 30

D.3 Common Evaluation Protocol 30

D.4 Scope of Our Experimental Studies 31

E Additional Results **32**

E.1 Qualitative Lineage Analysis 32

E.2 Budgeted Decision Dynamics 34

E.3 Geometry of Specialist Repertoires 36

E.4 Explainable Mutation Analysis 37

E.5 Robust Search under Structural Commitments 38

E.6 Posterior Calibration 39

E.7 Decision Traceability 40

E.8 Hyperparameter Sensitivity 41

E.9 Failure Cases 43

F Further Discussion **44**

F.1 Code Availability 44

F.2 Social Impact 44

F.3 Limitations 44

F.4 Future Work 44

F.5 Ethical Statement 45

A Extended Related Work

We begin with EGRET's position in the literature, then examine the biological motivation and methodological precedents underlying its design.

A.1 Positioning EGRET

From endpoints to repertoires. The central question in EGRET is which related specialists should be constructed together. Learning-guided optimization improves individual variants (Yang et al., 2025; Bogensperger et al., 2025), while multi-objective search and library design produce useful collections (Jain et al., 2023; Tang et al., 2025; Ding et al., 2024b). These advances motivate a further step: jointly choosing the conditions to serve, the leaves that cover them, and the mutation steps they can share. Table 9 locates this decision relative to the surrounding literature. EGRET couples condition-wise coverage to a shared construction budget and evaluates the resulting repertoire against a generalist reachable with the same budget.

Table 9: EGRET within the related literature. ✓ supported; △ partial; – outside scope.

	Landscape learning				Repertoire design			
Research direction	Multiple conditions	Assay selection	Explainable model	Uncertainty estimates	Set utility	Shared edits	Generalist baseline	Verified utility
Fitness optimization	△	△	△	✓	–	–	–	–
Landscape modeling	△	–	✓	△	–	–	–	–
Multi-objective design	✓	△	△	△	△	–	–	–
Library diversification	–	–	△	–	✓	–	–	–
Combinatorial optimization	–	✓	△	✓	△	–	–	–
EGRET	✓	✓	✓	✓	✓	✓	✓	✓

Reading the comparison. The rows summarize representative published formulations, including interpretable landscape models (Tonner et al., 2022; Faure et al., 2024) and combinatorial Bayesian optimization (Wang & Chen, 2018; Bal et al., 2025). Partial support denotes a narrower treatment or a capability present in only some of the cited studies. The first three columns cover multiple conditions or target properties, sequential assay selection, and whether the learned landscape model exposes an explainable structure (e.g. additive or pairwise mutation effects) rather than an opaque predictor. The next column covers set-level diversity or coverage. The last three ask whether designs reuse mutation steps, use a budget-matched reachable generalist as a reference, and provide a measured-value coverage lower bound. Generating several candidates therefore does not by itself establish a set objective, nor does experimental validation establish that coverage bound. The entries describe the original formulations, before our benchmark's adaptations to produce feasible trees.

Implications for EGRET. This positioning motivates acquisition at the level of a constructible repertoire. EGRET samples landscapes, identifies promising feasible trees, and allocates assays according to the support for their leaves; final commitment uses measured utilities. Its contribution is the coupling of this decision rule with the diversification problem, rather than posterior sampling, epistatic modeling, or set optimization in isolation. The following subsections connect this design to biological motivation, prior methods, and practical applications.

A.2 Learning-Guided Protein Engineering

Data-efficient fitness optimization. Gaussian-process modeling established a connection between protein fitness prediction and uncertainty-guided experimental design (Romero et al., 2013). Subsequent work used combinatorial libraries to learn beneficial combinations of mutations (Wu et al., 2019), while transfer learning reduced the amount of labeled data needed to engineer a target protein (Biswas et al., 2021). More recently, ALDE demonstrated an iterative batch Bayesian-optimization workflow on an epistatic enzyme landscape, experimentally improving a non-natural cyclopropanation reaction (Yang et al., 2025). This progression makes assay-efficient search a practical component of protein engineering.

These methods motivate EGRET's sequential use of experimental feedback, but fitness improvement for a target application does not specify how several related variants should divide responsibility across conditions. Even a batch of individually promising sequences may contain redundant specialists or require too many separate construction steps. EGRET therefore assigns value to a jointly feasible repertoire, with each assay informing both the condition coverage of a candidate and its compatibility with other descendants.

Generative search priors. Pretrained sequence models supply useful information before task-specific assays are available, especially when experimental evidence is still sparse. General protein language models have guided antibody evolution (Hie et al., 2024); bidirectional learning combines a pretrained proxy with offline biological sequence optimization (Chen et al., 2023); and VLGPO couples a generative prior with a fitness model through a variational formulation (Bogensperger et al., 2025). These approaches expand the ability to propose plausible, high-scoring endpoints for downstream experimental evaluation. Their proposal mechanisms are complementary to EGRET's decision problem: EGRET assumes a finite candidate library and asks which descendants merit measurement and joint construction under limited experimental resources. A stronger generator or representation could enrich that library with additional plausible candidates, but does not by itself determine a niche, a branching structure, or the allocation of a shared construction budget in advance.

Simplicity under assay constraints. EGRET deliberately favors a compact, interpretable Bayesian model over a complex black-box deep predictor as the core of its assay loop. This choice is motivated by the limited observations available within each campaign and the need to trace predictions to shared mutation effects or condition-specific deviations. Interpretable genotype–phenotype models already provide evidence that predictive power need not require opaque representations (Tonner et al., 2022; Faure et al., 2024). EGRET adopts this principle through explicit site, residue, and pair effects with structured shrinkage. Pretrained models remain useful sources of candidate sequences or prior information, but the repertoire decision does not depend on training a large generative model for every campaign.

The resulting design also supports a tractable connection between uncertainty, acquisition, and commitment. The linear-Gaussian structure admits explicit posterior calculations, while Propositions C.1 and C.2 characterize lineage-mode recovery and bound posterior regret through unmeasured leaf mass. These guarantees concern the decision procedure under the stated assumptions; they do not establish that a simple surrogate always predicts better than a deep network. EGRET's aim is to obtain useful repertoires with limited assays using a transparent model whose role in the final decision can be analyzed and checked against measured evidence.

A.3 Functional Diversification

Evolutionary accessibility. Protein promiscuity provides a biological starting point for diversification: weak secondary activities can provide material for the evolution of new functions (Aharoni et al., 2005). However, the value of a mutation depends on its background. Experimental studies have shown that only a restricted subset of mutational paths may be selectively accessible (Weinreich et al., 2006), and that permissive substitutions and historical contingency shape functional transitions (Harms & Thornton, 2014; Starr et al., 2017). These findings motivate representing related specialists through explicit shared ancestry rather than treating their mutation sets as independent design choices in realistic evolutionary settings.

EGRET turns this motivation into an engineering constraint: mutations installed before a branch are charged once, so endpoint selection changes the remaining capacity for diversification. This is a construction model, not a model of natural selection. Admissible edit paths need not be monotonically improving, and terminal utility measurements do not certify intermediate functionality. The distinction keeps the comparison with evolutionary accessibility precise while identifying a useful extension: incorporating measured viability or stability constraints on internal nodes.

Practical specialist repertoires. Enzyme engineering offers a direct use case. A laboratory developing catalysts for a panel of related substrates may retain a broadly active scaffold, then reuse early mutations before producing substrate-specific descendants. ALDE's experimentally validated biocatalysis workflow illustrates the value of learning from limited measurements (Yang et al., 2025),

while MODIFY designs starting enzyme libraries by jointly considering predicted fitness and sequence diversity (Ding et al., 2024b). EGRET addresses a subsequent allocation question: whether the available construction effort should support one broad catalyst or a small family whose members cover different substrates. Here the repertoire utility reflects choosing the best available catalyst for each substrate, rather than assuming that all catalysts operate together in one reaction.

A related prospective use is engineering a family of antibodies evaluated against an antigen panel, building on language-model-guided antibody evolution (Hie et al., 2024). Each antigen can be assigned to its strongest measured leaf, while shared sequence edits reduce the modeled construction cost. This interpretation matches a selectable reagent repertoire; it does not establish cocktail efficacy, which would require interaction, dosage, and safety data outside the present objective. In both applications, the cost advantage is most relevant when laboratories can reuse intermediate constructs; independent synthesis of every endpoint would require a different cost model.

A.4 Interpretable Fitness Landscapes

Structured epistasis. Interpretable genotype–phenotype models aim to recover useful sequence rules without sacrificing predictive power (Tonner et al., 2022). Recent experimental work shows that, for the proteins studied, additive free-energy effects with sparse pairwise energetic couplings can explain substantial variation in large sequence spaces (Faure et al., 2024). This supports structured models as a serious alternative to unconstrained predictors, while distinguishing energetic additivity from additivity of the measured phenotype. EGRET does not infer free energies: its site, residue, and pair coefficients describe the utility landscape directly. Pooling these coefficients across conditions expresses shared activity, whereas condition-specific deviations represent potential specialization. These are predictive attributions within the model, not independent evidence of a causal biochemical mechanism alone.

A complementary 2026 study separates single-site biases from pairwise covariance in `Potts`-based designs, finding different contributions to stability and catalytic activity in the tested proteins (Sternke et al., 2026). Together, these results argue against assuming that the same interaction structure is uniformly useful for every phenotype. They motivate EGRET's shrinkage of residue and pair effects and its condition-specific deviations, while leaving their adequacy to be assessed on each underlying target landscape.

Interpretability as a design signal. Concept bottleneck language models expose biological concepts that can be used to control protein generation (Ismail et al., 2025). IDEAS uses an explainable model to identify important sequence motifs and guide evolutionary edits (Pandey et al., 2026). Such methods show that explanations can influence which sequences are proposed, rather than merely annotate a completed search. EGRET uses a different explanatory unit: a substitution or interaction can contribute broadly across a condition panel or favor a specific branch. Posterior lineage sampling translates these uncertain effects into support for constructible repertoires. The resulting explanation connects a mutation's predicted role to a joint design decision, without requiring that motif importance alone determine the final leaves.

Uncertainty beyond observed variants. Accurate interpolation is insufficient when optimization selects unusual combinations of mutations. Experimental evaluation of neural-network extrapolation found declining reliability away from the training regime and mismatches between held-out prediction metrics and design outcomes (Freschlin et al., 2024). Kermut combines protein-informed Gaussian-process kernels and examines the difficulty of obtaining useful uncertainty at the individual-variant level (Groth et al., 2024). These results motivate EGRET's explicit uncertainty for unobserved residues and its separation of exploration from commitment. The structured posterior is still an approximation; its role is to expose plausible alternative lineages, while the final coverage certificate uses observed utilities rather than a maximum over imputed values.

A.5 Repertoire Optimization

Pareto search. Multi-Objective GFlowNets generate diverse Pareto candidates, with an active-learning variant that incorporates expensive evaluations (Jain et al., 2023). PepTune uses multi-objective guidance in discrete diffusion to design therapeutic peptides (Tang et al., 2025), while

Pareto-optimal sampling has also been studied specifically for protein sequences (Luo et al., 2025). These methods already address competing properties and multiple candidate solutions. EGRET's distinction is the utility assigned to the returned set: conditions can be served by different leaves, and each leaf is useful only in relation to the coverage and construction cost of the others.

In particular, a strong compromise on a Pareto front can be a useful generalist, while complementary specialists can improve the maximum attained separately in each condition. Pareto quality alone determines neither this coverage nor joint affordability. A search tree also differs from a construction tree: PepTune's tree guides generation, whereas EGRET's edges represent reusable mutation rounds. The latter makes shared ancestry part of the returned solution and its budget.

Diverse library design. MODIFY is a close conceptual neighbor because it optimizes a library rather than only an individual endpoint. Its fitness–diversity objective supports broad exploration of useful enzyme variants (Ding et al., 2024b). Nevertheless, sequence diversity is not equivalent to functional complementarity: distant sequences may favor the same conditions, while nearby sequences may specialize differently. For a fixed set of conditions, EGRET's condition-wise maximum rewards precisely the coverage supplied by the leaf set. Furthermore, the marginal construction cost depends on which shared trunks already exist. A fixed per-sequence cost or a diversity penalty cannot in general encode this dependence, motivating joint optimization of the repertoire and its mutation tree under shared lineage constraints.

A.6 Decision-Aware Experimental Design

Structured posterior sampling. Combinatorial Thompson sampling provides a direct precedent for sampling uncertain outcomes and optimizing a structured action (Wang & Chen, 2018). Decision-focused learning similarly emphasizes the quality of downstream combinatorial decisions rather than prediction error alone (Wilder et al., 2019). In protein design, GameOpt uses a game-theoretic formulation to optimize an upper-confidence-bound acquisition function in combinatorial sequence spaces (Bal et al., 2025). EGRET builds on these ideas; neither posterior sampling nor combinatorial acquisition is itself a novelty claim. Its action is a construction-coupled specialist tree, and its acquisition statistic aggregates leaf membership across sampled feasible repertoires before allocating sequence–condition measurements.

Assay allocation. Batch active learning can choose several informative variants at a time (Yang et al., 2025; Jain et al., 2023). EGRET additionally distinguishes selecting a sequence from paying to measure its complete condition profile. Correlated conditions permit broader initial screening, but incomplete profiles can hide a specialist's contribution to another condition. Spectral screening followed by profile completion addresses this tension under a cell-level budget. Because the value of a measurement depends on alternative leaves and shared construction steps, independent uncertainty ranking does not capture the whole downstream decision.

Measured-value commitment. Optimization can select candidates whose estimated values are unusually optimistic, an instance of the optimizer's curse (Smith & Winkler, 2006). EGRET is exposed to this effect when it maximizes across leaves and searches across feasible trees. Its response is to use posterior predictions for exploration but observed, nonnegative utilities for the final lower bound on repertoire coverage. This certifies repertoire utility under the stated assay model, but does not provide a confidence interval for the normalized diversification score: the accessible generalist and specialist anchors used to compare niches remain estimated. The distinction identifies exactly which part of the final decision is supported by measured evidence.

B ALGORITHMIC SPECIFICATION

This section makes the dependencies between the four components in Section 4 explicit. At the beginning of round t, the only learned state is the assay record $\mathcal{D}_t$; the posterior, candidate lineages, and acquisition order are recomputed from it. One round is therefore the following composition:

$$\mathcal{D}_t \xrightarrow{\text{fit}} p(\widetilde{f} \mid \mathcal{D}_t) \xrightarrow{\text{tree search}} \{a_C(x)\} \xrightarrow{\text{allocate}} \mathcal{B}_t \xrightarrow{\text{assay}} \mathcal{D}_{t+1}. \tag{12}$$

Here $\mathcal{B}_t$ is a set of sequence–condition cells, not a set of complete profiles. This distinction allows EGRET to screen broadly while reserving complete profiles for contrasts and terminal evidence.

B.1 CAMPAIGN OVERVIEW

Before any assay, the scaffold and construction limits determine the reachable set $\mathcal{X}_R$, the mutation set $m(x)$ of every candidate, and its leaf capacity $\kappa(x)$ from Eq. 8. These objects never change during a campaign. Algorithm 1 then alternates model fitting, lineage-level acquisition, and measurement until the cell budget is exhausted.

Algorithm 1 EGRET assay campaign

Require: library $\mathcal{X}$, conditions $\mathcal{C}$, scaffold x_0, assay budget B, tree limits (R, K, w)
Ensure: niche $\widehat{C}$, feasible tree $\widehat{T}$
1: Compute $\mathcal{X}_R$, $m(x)$, and $\kappa(x)$ for every reachable x
2: $\mathcal{D}_0 \leftarrow$ measure complete seed profiles from $\mathcal{X}_R$
3: **for** each round t while assays remain **do**
4: **Update:** fit to $\mathcal{D}_t$; obtain $\widehat{f}$ and a sampler
5: **Focus:** shortlist niches; retain their active conditions
6: **Sample:** draw a working pool; optimize sampled feasible trees
7: **Rank:** aggregate leaf support as $a_C(x)$; stratify using $\kappa(x)$
8: **Allocate:** refine, promote, spectrally screen, then explore
9: **Measure:** assay $\mathcal{B}_t$; append outcomes to obtain $\mathcal{D}_{t+1}$
10: **end for**
11: **Commit:** assemble measured trees; add feasible hedge leaves
12: Normalize with niche anchors; select $(\widehat{C}, \widehat{T})$ by Eq. 11
13: **return** $(\widehat{C}, \widehat{T})$

Complete-profile seed. The seed rows are measured across every condition. They have two roles that a collection of unrelated cells cannot replace: they provide matched sequence backgrounds for estimating the deviations $\boldsymbol{\delta}_c$, and they make the condition correlation matrix R_C identifiable. Subsequent rows may be partial because the shared consensus can learn from a single observed cell; complete profiles remain the evidence used for condition contrasts and for the spectral aperture.

B.2 POSTERIOR UPDATE

Observed-only feature refinement. At round start, EGRET rebuilds $z(x)$ from the current record. It includes a site coordinate once that site has been touched, a residue coordinate only for an observed substitution, and the most frequent substitution pairs supported by complete rows. For example, suppose $m(x) = \{i{:}a, j{:}b\}$, where substitution $i{:}a$ has been observed, $j{:}b$ is unobserved but site j has been touched through another residue, and their pair is not retained. If e denotes a one-hot coordinate in its named block, then

$$\begin{aligned} z(x) &= \big(1, \underbrace{e_i + e_j}_{\text{site block}}, \underbrace{e_{i:a}}_{\text{residue block}}, \underbrace{0}_{\text{pair block}}\big), \\ \mathbb{E}[\widetilde{f}(x,c) \mid \mathcal{D}] &= \widehat{\theta}_{0,c} + \widehat{\theta}^{\text{site}}_{i,c} + \widehat{\theta}^{\text{site}}_{j,c} + \widehat{\theta}^{\text{res}}_{i:a,c}, \qquad r(x) = 1. \end{aligned} \tag{13}$$

The missing residue $j{:}b$ contributes no fitted coordinate, but $r(x) = 1$ adds one residue-uncertainty term in Eq. 6. Thus site evidence transfers without transferring residue confidence. If site j were also unseen, its mean contribution would fall back to the intercept.

Approximate posterior fit. Let C_{act} denote the current active condition set, $C_x := \{c \in C_{\mathrm{act}} : A(x,c) = 1\}$, and $\bar{y}_c$ the mean of the observed cells in condition c. EGRET approximates the joint model in Eq. 5 through two consecutive fits:

$$\widehat{\widetilde{\boldsymbol{\theta}}} = \arg\min_{v} \sum_{x:C_x \neq \emptyset} \underbrace{\frac{|C_x|}{|C_{\mathrm{act}}|}}_{\text{observed fraction}} \left[\underbrace{\frac{1}{|C_x|} \sum_{c \in C_x} (y_{xc} - \bar{y}_c)}_{\text{observed contrast}} - z(x)^\top v \right]^2 + \lambda \|v\|^2_{P_\rho},$$
$$\widehat{\boldsymbol{\delta}}_c = \arg\min_{d} \sum_{x:C_x = C_{\mathrm{act}}} \left[y_{xc} - z(x)^\top (\widehat{\widetilde{\boldsymbol{\theta}}} + d) \right]^2 + s\lambda \|d\|^2_{P_\rho}, \qquad c \in C_{\mathrm{act}}. \tag{14}$$

The first fit weights each observed contrast by its measured fraction; complete rows have unit weight. The second holds the consensus fixed and fits deviations on complete profiles. These sequential estimates generally differ from the joint posterior mean. Condition means are restored through the intercept. Generalized cross-validation selects λ and ρ on complete profiles. Gaussian sampling is centered at the resulting coefficients, with covariance from the consensus factorization scaled by each condition's residual variance. This defines a working posterior approximation; residual residue dispersion supplies τ_c in Eq. 6.

Decision-facing shortlist. For each niche C, let L_C be the at-most-K touched rows obtained by greedy measured coverage and score it with the same normalized measured certificate used for commitment in Eq. 11. EGRET retains the three highest-scoring niches and takes their union as C_{act}. It then uses one posterior draw over $\mathcal{X}_R$ to form the working pool $\mathcal{P}$ from verified rows, condition leaders, and broad generalists. This early stochastic truncation matters: a plausible sequence excluded from $\mathcal{P}$ cannot be recovered by the subsequent tree optimizer.

B.3 Posterior Lineage Search

Posterior pushforward. For each leading niche, the posterior is passed through the constrained tree optimizer before it becomes an acquisition score:

$$\widetilde{f}^{(s)} \sim p(\widetilde{f} \mid \mathcal{D}) \quad \longmapsto \quad \left\{ \widetilde{T}_C^{(s,j)} \right\}_{j=0}^{J-1} := \operatorname*{Top}_{J,\, T \in \mathcal{T}_R} V_{\widetilde{f}^{(s)}}(C,T) \quad \longmapsto \quad a_C(x). \tag{15}$$

The last map is the rank-weighted leaf count in Eq. 7, including its generalist guard; EGRET sums it over the leading niches. Consequently, a sequence receives support only when it is useful jointly with affordable partners. High marginal variance without membership in a strong feasible tree is insufficient for final selection.

Coverage-preserving extension. The optimizer begins with feasible pairs because a valid repertoire contains at least two distinct leaves, then extends each partial tree by one endpoint. For a sampled landscape g, write $V_g(C,L)$ for Eq. 1 evaluated on leaf set L. After proposing x, the new score follows from

$$V_g(C, L \cup \{x\}) = \frac{1}{M} \sum_{c \in C} \max\left\{ \max_{z \in L} g(z,c), g(x,c) \right\},$$
$$V_g(C, L \cup \{x\}) - V_g(C,L) = \frac{1}{M} \sum_{c \in C} \Big[g(x,c) - \max_{z \in L} g(z,c) \Big]_+ . \tag{16}$$

The second line uses $\max\{u,v\} - u = [v-u]_+$ and shows that a new leaf matters only where it beats the current champion. Algorithm 2 applies this gain only after verifying that the merged mutation tree respects (R,K,w). For fixed leaves, shared substitutions are placed toward the trunk and branch points are aligned with edit-round boundaries, reducing cost without changing Eq. 16.

Cost stratification is essential during extension: a high-valued tree with no remaining construction capacity must not eliminate a cheaper state that can still add another specialist. The returned trees are then rank-weighted in Eq. 7; Algorithm 2 itself adds no surrogate acquisition score.

Algorithm 2 Cost-stratified lineage beam

Require: sampled landscape g, working pool $\mathcal{P}$, niche C, tree limits (R, K, w)
Ensure: ranked feasible trees for this draw
1: Form the feasible two-leaf trees from promising members of $\mathcal{P}$
2: Group by construction cost; retain the best beam states per group
3: **for** leaf count $k = 3, \ldots, K$ **do**
4: Extend every retained leaf set by one member of $\mathcal{P}$
5: Discard repeated leaves and trees violating (R, w)
6: Score each survivor by Eq. 16
7: Deduplicate; retain a cost-stratified beam
8: **end for**
9: **return** the best distinct trees, breaking value ties by lower cost

Capacity protection. Summing support over niches yields one sequence order for the round. Sorting this support alone can fill the order with deep endpoints. EGRET keeps the order within each group but reserves one of every K positions for a candidate with $\kappa(x) = K$. By Eq. 8, that candidate can still belong to a full-size repertoire. This is a feasibility floor, not a second utility score: posterior evidence determines the order inside both groups.

B.4 Assay Allocation Pipeline

Row-to-cell conversion. The ranked list specifies sequence rows, whereas the budget charges cells. From complete active profiles, Eq. 9 computes $h(C)$, the effective number of distinct condition modes. A small $h(C)$ permits a shallow screen across many rows; a large $h(C)$ demands nearly complete profiles. For a ranked row x, the remaining width and its chosen conditions are obtained together for the current screen:

$$\begin{aligned} h_x &:= \left[h(C) - \sum_{d \in C} A(x,d) \right]_+ , \\ (c_1, \ldots, c_{h_x}) &:= \mathrm{Top}_{h_x} \left(\widehat{f}(x,c) - \max_{z : A(z,c)=1} y_{zc} : c \in C,\ A(x,c) = 0 \right). \end{aligned} \tag{17}$$

EGRET adds $(x, c_1), \ldots, (x, c_{h_x})$ to the batch. Thus $a_C(x)$ selects the row, $h(C)$ fixes its total assay depth, and improvement over the measured champion selects the columns. Existing cells reduce h_x, so none is bought twice. This separation keeps the row-level acquisition score distinct from the cell-level allocation rule: the former determines which candidate enters the batch, whereas the latter determines which observations are still needed to evaluate its contribution. A partially profiled row can therefore remain eligible for later completion without violating the same assay budget.

Promotion lane. Both verification lanes change the same condition-wise measured maximum. For compactness, write $v_c(L) := \max_{u \in L} A(u,c) y_{uc}$, let C_x contain the conditions currently observed for a partial row x, and let $\mathcal{D}^+$ be the record after profiling it. Expanding the definition of Eq. 10 over observed and missing cells gives

$$\begin{aligned} \Delta_{\mathcal{D}^+}(x \mid C, T) &= \frac{1}{M} \left(\sum_{c \in C_x} [y_{xc} - v_c(\partial T)]_+ + \sum_{c \in C \setminus C_x} [y_{xc} - v_c(\partial T)]_+ \right) \\ &= \Delta_{\mathcal{D}}(x \mid C, T) + \frac{1}{M} \sum_{c \in C \setminus C_x} [y_{xc} - v_c(\partial T)]_+ \geq \Delta_{\mathcal{D}}(x \mid C, T). \end{aligned} \tag{18}$$

The first summation is precisely the gain already visible in the screen; the second is the additional, nonnegative gain unlocked by promotion. Hence profiling cannot reduce the candidate's measured marginal coverage in the current campaign. The decomposition separates evidence already purchased from evidence obtained only by completing the partial row, while preserving the same condition-wise certificate throughout the promotion decision. Since the newly exposed contribution is nonnegative, the screening evidence remains valid after the row is profiled.

Rather than restating the masking identity, characterize the finite candidate set passed to promotion. For a fixed quota q, let $\mathrm{Top}_q^{(c)}(\mathcal{P}(C))$ denote the rows in $\mathcal{P}(C)$ with the largest values of $s_c(x)$ for condition c on the active panel:

$$\begin{aligned} \mathcal{P}(C) &:= \left\{ x \in \mathcal{X}_R : 0 < \sum_{c \in C} A(x,c) < M \right\}, \\ s_c(x) &:= A(x,c) y_{xc}, \qquad \mathcal{P}_q(C) := \bigcup_{c \in C} \mathrm{Top}_q^{(c)}\big(\mathcal{P}(C)\big). \end{aligned} \tag{19}$$

This matches EGRET's implementation: $\mathcal{P}(C)$ contains reachable partial rows, and $s_c(x)$ ranks them with unopened cells assigned zero. The union keeps condition-wise leaders, removes duplicates, and exhausts a finite candidate list before unused budget returns to screening or exploration.

Refinement lane. For incumbent leaf z, set $L_{-z} := \partial T \setminus \{z\}$ and let $T_{z \to x}$ replace it by a fully profiled, feasible one-substitution neighbor x. Subtracting the two certificates and applying $\max\{u, v\} = u + [v - u]_+$ to both maxima yields

$$\begin{aligned} M[\underline{V}_{\mathcal{D}^+}(C, T_{z\to x}) - \underline{V}_{\mathcal{D}}(C,T)] &= \sum_{c \in C} \left(\max\{v_c(L_{-z}), y_{xc}\} - \max\{v_c(L_{-z}), A(z,c) y_{zc}\} \right) \\ &= \sum_{c \in C} \left(v_c(L_{-z}) + [y_{xc} - v_c(L_{-z})]_+ \right) \\ &\quad - \sum_{c \in C} \left(v_c(L_{-z}) + [A(z,c) y_{zc} - v_c(L_{-z})]_+ \right) \\ &= \sum_{c \in C} \left([y_{xc} - v_c(L_{-z})]_+ - [A(z,c) y_{zc} - v_c(L_{-z})]_+ \right). \end{aligned} \tag{20}$$

The common coverage $v_c(L_{-z})$ cancels in the last step. What remains is the candidate's unique measured coverage beyond the unchanged leaves, minus the incumbent's unique measured coverage. Promotion and refinement thus alter the same certificate rather than introduce separate acquisition objectives within one unified framework.

Algorithm 3 Cell-batch allocation

Require: record $\mathcal{D}_t$, ranked rows, active conditions, round allowance
Ensure: new-cell batch $\mathcal{B}_t$
1: $\mathcal{B}_t \leftarrow \emptyset$; compute aperture $h(C)$ from complete profiles
2: Add full profiles of feasible one-step refinements
3: Add missing cells of the top partial rows in each active condition
4: **for** each ranked row while screening allowance remains **do**
5: Add up to $h(C)$ unopened cells by measured-champion improvement
6: **end for**
7: Spend the reserved exploration cells in the public row order
8: **return** $\mathcal{B}_t$

Algorithm 3 implements this order. Refinement and promotion range over finite lists, so unused allowance passes to the next lane without another allocation ratio. The lineage-ranked screen follows, and the fixed public row order receives the exploration reserve last.

Measured terminal assembly. After the last round, EGRET shortlists niches again and assembles a tree for each one from every touched row. Because utilities lie in $[0, 1]$, replacing an unmeasured cell by zero yields exactly the lower-bound objective $\underline{V}_{\mathcal{D}}$ in Eq. 11; posterior values never enter the committed tree value. If a leaf can be added without violating (R, K, w), then

$$\partial T \subseteq \partial T' \quad \Longrightarrow \quad \underline{V}_{\mathcal{D}}(C, T') \geq \underline{V}_{\mathcal{D}}(C, T), \tag{21}$$

because each condition-wise maximum is taken over a superset. Unused feasible leaf capacity is therefore a cost-free hedge: measured gain ranks additions, posterior gain breaks zero-gain ties, and only measured utility changes the certificate. EGRET normalizes candidates by their shortlist anchors and returns the best $(\widehat{C}, \widehat{T})$. The result remains traceable: assays certify value, mutation paths expose cost, and $z(x)^\top \boldsymbol{\theta}_c$ attributes the exploration signal.

C Theoretical Foundation

C.1 Analytical Derivations

Identifiable reference posterior. The joint model in Eq. 5 provides a reference for the two-stage approximation in Section B.2. Stack the consensus and condition deviations into $\boldsymbol{\vartheta} = (\bar{\boldsymbol{\theta}}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{|\mathcal{C}|})$. For an observed cell (x, c), let a_{xc} be the sparse row satisfying $a_{xc}^\top \boldsymbol{\vartheta} = z(x)^\top(\bar{\boldsymbol{\theta}} + \boldsymbol{\delta}_c)$, and collect the prior weights in $Q = \mathrm{blockdiag}(P_\rho, sP_\rho, \ldots, sP_\rho)$. All inverses and densities below are restricted to the orthogonal complement of the redundant direction that shifts the consensus intercept and every deviation intercept equally and oppositely. Define

$$H := \sum_{(x,c)\in\mathcal{D}} \frac{a_{xc}a_{xc}^\top}{\sigma_c^2} + \lambda Q, \qquad b := \sum_{(x,c)\in\mathcal{D}} \frac{a_{xc}y_{xc}}{\sigma_c^2}. \tag{22}$$

Completing the square in Eq. 5 in $\boldsymbol{\vartheta}$, up to additive constants, gives

$$\begin{aligned} -\log p(\boldsymbol{\vartheta} \mid \mathcal{D}) &\doteq \frac{1}{2}\boldsymbol{\vartheta}^\top H\boldsymbol{\vartheta} - b^\top\boldsymbol{\vartheta} \\ &= \frac{1}{2}(\boldsymbol{\vartheta} - H^{-1}b)^\top H(\boldsymbol{\vartheta} - H^{-1}b). \end{aligned} \tag{23}$$

With positive noise variances and observations in every condition, H is positive definite on this subspace. The reference posterior therefore has

$$\boldsymbol{\vartheta} \mid \mathcal{D} \sim \mathcal{N}(H^{-1}b, H^{-1}), \qquad \widehat{\boldsymbol{\theta}}_{c,\mathcal{D}} = L_c H^{-1}b, \qquad \Sigma_{c,\mathcal{D}} = L_c H^{-1} L_c^\top, \tag{24}$$

where the selector L_c maps $\boldsymbol{\vartheta}$ to $\bar{\boldsymbol{\theta}} + \boldsymbol{\delta}_c$. Linear propagation, also valid for the working Gaussian approximation in Eq. 6, gives $\mathrm{Var}[z(x)^\top\boldsymbol{\theta}_c \mid \mathcal{D}] = z(x)^\top\Sigma_{c,\mathcal{D}}z(x)$.

Unseen-residue variance. The design omits a residue coordinate until that substitution has been assayed. To state the resulting correction explicitly, let $\eta_{1c}, \ldots, \eta_{r(x)c}$ be the missing residue effects in condition c. They are modeled as conditionally independent, zero-mean corrections with variance τ_c^2, independent of the fitted coefficients. A sampled utility therefore has the decomposition

$$\widetilde{f}(x, c) = z(x)^\top\boldsymbol{\theta}_c + \sum_{j=1}^{r(x)} \eta_{jc}. \tag{25}$$

Using bilinearity of covariance and the stated independence assumptions,

$$\begin{aligned} \mathrm{Var}[\widetilde{f}(x,c) \mid \mathcal{D}] &= \mathrm{Var}[z(x)^\top\boldsymbol{\theta}_c \mid \mathcal{D}] + \sum_{j=1}^{r(x)} \mathrm{Var}[\eta_{jc}] \\ &\quad + 2\sum_{j=1}^{r(x)} \mathrm{Cov}[z(x)^\top\boldsymbol{\theta}_c, \eta_{jc}] + 2\sum_{i<j} \mathrm{Cov}[\eta_{ic}, \eta_{jc}] \\ &= z(x)^\top\Sigma_{c,\mathcal{D}}z(x) + r(x)\tau_c^2. \end{aligned} \tag{26}$$

The two covariance sums vanish rather than being silently ignored. In EGRET, τ_c is the larger of the empirical residue dispersion and the scale $\sigma_c/\sqrt{\lambda\rho}$ implied by the fitted residue prior. Consequently, strong site-level shrinkage cannot make an unobserved residue spuriously certain.

Leaf-capacity bound. Let P_x be the root-to-x path in a feasible tree. Because one construction edge installs at most w substitutions, $|P_x| \geq \lceil |m(x)|/w \rceil$. No other leaf can lie on P_x, and every one of the remaining $|\partial T| - 1$ leaves requires a distinct edge outside that path. Therefore

$$\begin{aligned} |E(T)| &\geq |P_x| + |\partial T| - 1 \\ &\geq \left\lceil \frac{|m(x)|}{w} \right\rceil + |\partial T| - 1, \qquad |E(T)| \leq R \\ &\implies \quad |\partial T| \leq \min\left\{ K, R - \left\lceil \frac{|m(x)|}{w} \right\rceil + 1 \right\}. \end{aligned} \tag{27}$$

This is precisely Eq. 8. It is a necessary capacity bound, not a claim that every sequence attaining it can be combined with arbitrary other leaves.

Range of the spectral aperture. Let $r_C = \operatorname{rank}(R_C)$ and let $\xi_1, \dots, \xi_{r_C} > 0$ be the nonzero eigenvalues of the positive semidefinite correlation matrix. Spectral identities give $\operatorname{tr} R_C = \sum_j \xi_j$ and $\|R_C\|_F^2 = \sum_j \xi_j^2$. The lower and upper bounds follow separately from

$$\begin{aligned} \left(\sum_j \xi_j\right)^2 &= \sum_j \xi_j^2 + 2\sum_{i<j} \xi_i \xi_j \geq \sum_j \xi_j^2, \\ \left(\sum_j \xi_j\right)^2 &\leq r_C \sum_j \xi_j^2 \qquad \text{(Cauchy–Schwarz).} \end{aligned} \tag{28}$$

Dividing by $\sum_j \xi_j^2 > 0$ yields

$$1 \leq \frac{(\operatorname{tr} R_C)^2}{\|R_C\|_F^2} \leq \operatorname{rank}(R_C) \leq M. \tag{29}$$

The lower equality occurs for a rank-one panel. The upper equality in the middle inequality occurs exactly when all nonzero eigenvalues are equal; in particular, $R_C = I_M$ gives aperture M. These endpoint cases justify the clipping range in Eq. 9.

Computational complexity. Let P denote the candidate endpoint count, M the selected niche size, U the mutation-token dimension, K the leaf cap, and W the global beam width. The global beam cap W is shared across cost strata, and the construction budget R constrains feasibility without introducing an additional iteration factor. Processing S posterior draws sequentially scales runtime linearly with S while preserving peak auxiliary space. After working-pool formation, the bounds for Algorithm 2 in Table 10 are expressed in terms of P; posterior fitting, pool construction, niche selection, and cell allocation incur additional overhead over the library of size $N = |\mathcal{X}_R|$.

Table 10: Leading-order complexity of lineage search. Costs are per Algorithm 2 call with fixed verification caps; logarithmic sorting factors and input storage are excluded.

Operation	Workload	Time complexity	Auxiliary space
Pair initialization	Endpoint pairs	$\mathcal{O}(P^2(M+U))$	$\mathcal{O}(P^2M)$
Beam extension	Beam candidates	$\mathcal{O}(KWP(M+U))$	$\mathcal{O}(WPM)$
Tree verification	Candidate trees	$\mathcal{O}(K^3U^2)$	$\mathcal{O}(K^2U^2)$
Lineage search	Full search call	$\mathcal{O}((P^2+KWP)(M+U)+K^3U^2)$	$\mathcal{O}(PM(P+W)+K^2U^2)$

C.2 Further Theoretical Guarantees

Why sample lineages rather than optimize a mean landscape. Probability matching by itself does not explain why action-level posterior sampling is needed: one must also show what the posterior mean can erase. The following separation uses the smallest genuinely branching repertoire, $M = K = 2$, and respects the shared-edge budget. For a complete landscape F, write $T_C^\star(F)$ for the tie-broken optimizer in Eq. 7.

Proposition C.1 (Lineage-mode recovery and mean-planning separation)**.** *Let $q \geq 2$, and let H be a continuous distribution on $[0,1]$ with mean μ. For any $a \in (\mu, 1)$ satisfying $H(a) < 1$, there exists a two-condition lineage instance and a posterior with uncertain specialist trees $T_1, \dots, T_q$ and a decoy tree T_0 such that:*

$$\Pr\{T_0 = T_C^\star(F) \mid \mathcal{D}\} = p_0 := H(a)^q, \qquad \Pr\{T_i = T_C^\star(F) \mid \mathcal{D}\} = p := \frac{1-p_0}{q}, \quad 1 \leq i \leq q, \tag{30}$$

whereas optimizing the posterior-mean landscape selects T_0 uniquely. Nevertheless, the expected posterior opportunity hidden by that mean decision is

$$\mathbb{E}[V_F(C, T_C^\star(F)) - V_F(C, T_0) \mid \mathcal{D}] = \int_a^1 [1 - H(t)^q]\, dt > 0. \tag{31}$$

After S independent posterior-lineage draws under the stated scheme, if D_S is the number of distinct uncertain trees observed, then

$$\mathbb{E}[D_S] = q\big[1-(1-p)^S\big], \qquad \Pr\{D_S = q\} = \sum_{j=0}^{q}(-1)^j\binom{q}{j}(1-jp)^S \geq 1 - qe^{-pS}. \quad (32)$$

Thus posterior-mean planning can concentrate on a lineage whose posterior optimality vanishes exponentially in q, while posterior lineage sampling recovers all material lineage modes after $S \geq p^{-1}\log(q/\delta)$ draws with probability at least $1-\delta$, for any $\delta \in (0,1)$.

Proof. Take $w = 1$ and $R = 4$. For each $i \in \{0, \dots, q\}$, construct two terminal sequences whose paths share a private two-edit trunk and then use distinct one-edit branches. Mutation sets belonging to different values of i are disjoint. The tree T_i joining the two terminals of mode i therefore costs $2+1+1 = 4$ edges and is feasible. Any tree containing terminals from two different modes costs at least $3+3 = 6 > R$, so the lineage budget makes the modes mutually exclusive.

For $i \geq 1$, draw U_i independently from H. In tree T_i, give its first terminal utility U_i in the first condition and zero in the second; reverse these utilities for its second terminal. Define the decoy T_0 analogously but replace U_i by the constant a. Consequently,

$$V_F(C, T_i) = U_i \quad (1 \leq i \leq q), \qquad V_F(C, T_0) = a. \quad (33)$$

Every uncertain tree has posterior-mean value $\mu < a$, so the mean landscape selects T_0. On a posterior draw, however, the decoy is optimal exactly when $\max_i U_i \leq a$, which has probability $H(a)^q$. Conditional on an uncertain tree winning, continuity removes ties and exchangeability gives every T_i equal probability $(1-H(a)^q)/q$. This proves Eq. 30.

Writing $U_{(q)} = \max_i U_i$, the loss of the mean-selected decoy against the draw-wise optimal tree is $(U_{(q)} - a)_+$. The tail-integral identity and independence yield

$$\mathbb{E}[(U_{(q)} - a)_+] = \int_a^1 \Pr\{U_{(q)} > t\}\, dt = \int_a^1 [1 - H(t)^q]\, dt, \quad (34)$$

which is strictly positive because $H(a) < 1$. Notice that this is an *acquisition* separation: EGRET uses the sampled trees to expose competing lineages and still commits only after measurement.

It remains to count the modes exposed by repeated posterior lineage sampling. Each uncertain tree is returned on any draw with probability p. If I_i indicates that T_i has appeared at least once, then $\mathbb{E}[I_i] = 1-(1-p)^S$; summing over i proves the expectation in Eq. 32. For the exact coverage probability, apply inclusion–exclusion to the events that individual uncertain trees are absent. Any fixed set of j such trees is missed on all S draws with probability $(1-jp)^S$, giving the stated alternating sum. Finally, a union bound and $1-p \leq e^{-p}$ give

$$\Pr\{D_S < q\} \leq q(1-p)^S \leq qe^{-pS}. \quad (35)$$

Solving the last expression for a failure probability δ completes the proof. □

From posterior leaf mass to verified regret. We now connect the acquisition in Eq. 7 to the certificate in Eq. 11. Fix a niche C, let $\mathcal{X}_R$ be the N reachable sequences, and regard the unknown complete landscape F as a draw from the posterior given $\mathcal{D}$. We use fixed tie-breaking throughout. Thus $T_C^\star(F)$ is a well-defined optimal tree, and $x_c^\star(F)$ is a well-defined leaf attaining the maximum for condition c in that tree. Its marginal leaf probability is precisely $\pi_C(x) = \Pr\{x \in \partial T_C^\star(F) \mid \mathcal{D}\}$.

Let $\mathfrak{Q}_b$ be any nonempty family of admissible sets containing b sequences. This family may, for example, impose the capacity reservation in Eq. 8. From S independent posterior landscapes, let $\widehat{\pi}_C$ be the empirical frequency in Eq. 7, before the optional runner-up smoothing, and define the population and empirical pools by

$$Q_C^\star \in \arg\max_{Q \in \mathfrak{Q}_b} \sum_{x \in Q} \pi_C(x), \qquad \widehat{Q}_C \in \arg\max_{Q \in \mathfrak{Q}_b} \sum_{x \in Q} \widehat{\pi}_C(x). \quad (36)$$

After every cell in $Q \times C$ is assayed, denote by $\widehat{T}_Q(F)$ the tree maximizing the measured-value certificate in Eq. 11. The posterior simple regret of this commitment is

$$\mathcal{R}_C(Q) := \mathbb{E}\Big[V_F(C, T_C^\star(F)) - V_F(C, \widehat{T}_Q(F)) \;\Big|\; \mathcal{D}\Big]. \tag{37}$$

Proposition C.2 (Verified leaf-mass guarantee)**.** *For every admissible pool $Q \in \mathfrak{Q}_b$,*

$$\mathcal{R}_C(Q) \le \frac{1}{M}\sum_{c\in C} \Pr\{x_c^\star(F) \notin Q \mid \mathcal{D}\} \le \min\left\{1, \sum_{x\notin Q} \pi_C(x)\right\}. \tag{38}$$

Moreover, for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the same S independent posterior landscapes used to form $\widehat{\pi}_C$ in the preceding construction,

$$\mathcal{R}_C(\widehat{Q}_C) \le \min\left\{1, \sum_{x\notin Q_C^\star} \pi_C(x) + 2K\sqrt{\frac{\log(2|\mathfrak{Q}_b|/\delta)}{2S}}\right\}. \tag{39}$$

Hence the empirical EGRET pool competes with the best pool satisfying the same combinatorial restrictions in posterior leaf mass, without requiring independence among leaf events or among the leaves of a sampled tree.

Proof. For a realized landscape F, abbreviate $T_C^\star(F)$ by $T^\star$ and let $\underline{V}_Q$ denote the certificate after measuring $Q \times C$. Since $\widehat{T}_Q$ maximizes that certificate and every certificate is a lower bound on its true tree value,

$$\begin{aligned} 0 &\le V_F(C, T^\star) - V_F(C, \widehat{T}_Q) \\ &\le V_F(C, T^\star) - \underline{V}_Q(C, \widehat{T}_Q) \le V_F(C, T^\star) - \underline{V}_Q(C, T^\star). \end{aligned} \tag{40}$$

The final difference can be controlled condition by condition. If the tie-broken champion $x_c^\star(F)$ belongs to Q, its true utility occurs in the measured maximum and the difference for condition c is zero. Otherwise that difference is at most one because utilities are in $[0,1]$. Therefore, pointwise in F, for every fixed condition c,

$$V_F(C, T^\star) - \underline{V}_Q(C, T^\star) \le \frac{1}{M}\sum_{c\in C} \mathbf{1}\{x_c^\star(F) \notin Q\}. \tag{41}$$

Taking the conditional expectation proves the first inequality in Eq. 38. For every condition, an unmeasured champion is necessarily an optimal-tree leaf outside Q. A union bound, which does not assume independent leaf membership, consequently gives

$$\begin{aligned} \Pr\{x_c^\star(F) \notin Q \mid \mathcal{D}\} &\le \Pr\{\partial T_C^\star(F) \not\subseteq Q \mid \mathcal{D}\} \\ &\le \sum_{x\notin Q} \Pr\{x \in \partial T_C^\star(F) \mid \mathcal{D}\} = \sum_{x\notin Q} \pi_C(x). \end{aligned} \tag{42}$$

Together with the trivial range $\mathcal{R}_C(Q) \le 1$, this proves the second inequality in Eq. 38.

It remains to account explicitly for estimating the empirical leaf probabilities. A coordinatewise argument would accumulate error over all b selected sequences. Instead, for each admissible pool, define its sampled leaf overlap by

$$Z_s(Q) := |\partial \widetilde{T}_C^{(s,0)} \cap Q|, \qquad \frac{1}{S}\sum_{s=1}^{S} Z_s(Q) = \sum_{x\in Q} \widehat{\pi}_C(x), \qquad \mathbb{E}[Z_s(Q) \mid \mathcal{D}] = \sum_{x\in Q} \pi_C(x). \tag{43}$$

The overlaps are independent across posterior draws and lie in $[0, K]$, even though the membership indicators within one tree can be arbitrarily dependent. Hoeffding's inequality followed by a union bound over the admissible family therefore implies that, with probability at least $1-\delta$,

$$\max_{Q\in\mathfrak{Q}_b} \left| \sum_{x\in Q} \widehat{\pi}_C(x) - \sum_{x\in Q} \pi_C(x) \right| \le \Delta_S, \qquad \Delta_S := K\sqrt{\frac{\log(2|\mathfrak{Q}_b|/\delta)}{2S}}. \tag{44}$$

On this event, the empirical optimality of $\widehat{Q}_C$ gives

$$\sum_{x \in Q_C^\star} \pi_C(x) \le \sum_{x \in Q_C^\star} \widehat{\pi}_C(x) + \Delta_S \le \sum_{x \in \widehat{Q}_C} \pi_C(x) + 2\Delta_S. \tag{45}$$

Subtracting both selected masses from the common total $\sum_{x \in \mathcal{X}_R} \pi_C(x)$ shows that the mass outside $\widehat{Q}_C$ is at most the mass outside $Q_C^\star$ plus $2\Delta_S$. Applying Eq. 38 completes the proof. □

Approximate tree optimization. The exact statement above isolates the acquisition and assay effects, but the same reasoning is stable under the cost-stratified beam in Section 4.2. Suppose its returned tree is within ϵ_{opt} of the best value for every posterior landscape and, when used at commitment, within ϵ_{opt} of the best certificate. Use the beam-returned trees to form the membership frequencies, and let $x_{c,\mathrm{beam}}(F)$ be their tie-broken condition champion. The two solver gaps enter only through two additive terms:

$$V_F(C, T_C^\star(F)) - V_F(C, \widehat{T}_Q^{\mathrm{beam}}(F)) \le 2\epsilon_{\mathrm{opt}} + \frac{1}{M} \sum_{c \in C} \mathbf{1}\{x_{c,\mathrm{beam}}(F) \notin Q\}\,. \tag{46}$$

Indeed, the sampled-tree gap contributes the first ϵ_{opt}; replacing exact certificate maximization by the beam contributes the second; and the remaining measured-versus-true difference is bounded exactly as in Eq. 41. The rest of the proof is unchanged, so Eq. 39 acquires only the additive term $2\epsilon_{\mathrm{opt}}$ (and is clipped at one). This explicitly separates finite posterior-sampling error from combinatorial solver error.

Consequences for EGRET's design. The first inequality in Eq. 38 is stronger than an all-leaves-covered argument: EGRET need only measure the leaf that realizes each condition's optimum, not every leaf of $T_C^\star(F)$. Leaf membership is nevertheless a valid acquisition statistic because it upper-bounds these unknown, condition-specific champion events while retaining their shared-tree feasibility. For a desired Monte Carlo excess of at most γ, the final term in Eq. 39 is at most γ once $S \ge 2K^2\gamma^{-2}\log(2|\mathfrak{Q}_b|/\delta)$. In particular, $|\mathfrak{Q}_b| \le \binom{N}{b} \le (eN/b)^b$ turns this into an explicit bound in N and b. The resulting leading quadratic dependence is on the maximum number of leaves K, rather than the number of assayed rows b produced by a coordinatewise argument.

The abstract constraint family makes the result compatible with the capacity rule. Requiring exactly $\lceil b/K \rceil$ members of Q to satisfy $\kappa(x) = K$ makes the interleaving described after Eq. 8 the empirical optimizer in Eq. 36: it takes the highest-frequency members within each of the two capacity strata, whenever both strata contain enough candidates. Proposition C.2 then compares EGRET with the best posterior-mass pool under the same insurance against exhausting the leaf budget. Finally, Q consists of completed profiles; this is why spectral screening must be followed by promotion before its rows can tighten the verified-regret bound. Runner-up smoothing can change near-tied rankings, but the guarantee itself deliberately concerns the principal posterior membership term.

D EXPERIMENTAL DETAILS

D.1 BENCHMARK INVENTORY

Benchmark collection. Table 11 summarizes twenty sequence–condition matrices spanning synthetic epistasis, antibody binding and escape, enzyme activity, ribozyme function, fluorescence, DNA response elements, and ion-channel phenotypes. The finite library, condition panel, and scaffold follow the problem formulation in Section 3. Each instance contains a finite aligned sequence library, a fixed condition panel, and a designated scaffold x_0. Deterministic preparation orients utilities so that larger values are preferred and produces the complete matrix used for evaluation. The size labels report the audited scale of each loaded library after preparation.

Table 11: Benchmark inventory. Dataset names hyperlink to their download sources. The entries span varied biological targets and condition structures.

Dataset	Conditions	Description	Scale	License	References
`Potts`	8 niches	Synthetic epistasis	Large	Generated	This work
`CR9114`	H1, H3, influenza B	Influenza affinity	Large	CC BY 4.0	Phillips et al. (2021)
`CH65`	3 antigens	Influenza affinity	Large	CC BY 4.0	Phillips et al. (2023)
`CH65+Expr`	3 antigens + expression	Affinity–expression trade-off	Large	CC BY 4.0	Phillips et al. (2023)
`Omicron escape`	4 antibodies	Multi-antibody escape	Large	Source terms	Moulana et al. (2023)
`Omicron+ACE2`	4 antibodies + ACE2	Escape–ACE2 trade-off	Large	Source terms	Moulana et al. (2023)
`Lineage escape`	3 antibodies	Background-dependent escape	Medium	Source terms	Moulana et al. (2026)
`Lineage+ACE2`	3 antibodies + ACE2	Lineage-escape trade-off	Medium	Source terms	Moulana et al. (2026)
`BA.2 backgrounds`	7 backgrounds	Background-dependent ACE2	Small	Source terms	Moulana et al. (2026)
`S309 backgrounds`	7 backgrounds	Background-dependent S309 escape	Small	Source terms	Moulana et al. (2026)
`JN.1/XBB`	4 antibodies + ACE2	Deep lineage escape trade-off	Large	Source terms	Moulana et al. (2026)
`FP switch`	2 colours	Fluorescence switching	Medium	CC BY 4.0	Poelwijk et al. (2019)
`AncSR1`	16 DNA elements	Ancestral response activation	Large	Source terms	Starr et al. (2017)
`AncSR2`	16 DNA elements	Alternate ancestral background	Large	Source terms	Starr et al. (2017)
`Azoarcus`	8 $MgCl_2$ levels	Magnesium-dependent activity	Large	Source terms	Peri et al. (2022)
`DAOx`	5 D-amino acids	Substrate-specific activity	Medium	Source terms	Vanella et al. (2026)
`Hsp90`	6 environments	Environment-specific selection	Large	CC BY 4.0	Flynn et al. (2020)
`KCNJ2`	Surface + function	Trafficking–function trade-off	Medium	MIT	Coyote-Maestas et al. (2022)
`KCNE1`	Expression + function	Expression–function trade-off	Medium	MIT	Muhammad et al. (2024)
`KRAS`	6 binding partners	Effector-specific binding	Large	MIT	Weng et al. (2024)

Structural configuration. Each benchmark instance specifies its protocol before the campaign. The condition panel defines the evaluation axes; the niche size M, tree budget R, leaf cap K, and edit width w determine the feasible repertoire, and the assay budget B controls the opened sequence–condition cells. Table 12 lists these quantities and their operational roles, following the construction objective defined in Section 3. For a selected condition set C, the common validator checks that the proposed endpoints form a rooted tree satisfying (R, K, w), with all parameters fixed throughout the campaign under the declared protocol.

Procedural `Potts` landscape. The `Potts` control uses seed $31{,}415$ to generate a 24-residue scaffold and 20,000 unique sequences, with up to ten substitutions sampled from a geometric distribution ($p = 0.35$) at distinct positions. Its eight-condition utility matrix combines shared and condition-specific residue effects with pairwise couplings, drawn with standard deviations 0.65, 0.45, and 0.22, respectively, and normalized by $\sqrt{24}$.

Table 12: Operational configuration of a prepared EGRET instance. The entries specify the sequence library, condition panel, repertoire constraints, construction budget, and assay ceiling used by the declared benchmark protocol.

Quantity	Symbol	Operational definition	Protocol role
Sequence library	N_{seq}	Unique aligned sequences retained after deterministic preparation	Search space
Condition panel	N_{cond}	Number of fixed assay conditions in the prepared matrix	Evaluation axes
Matrix size	N_{cell}	$N_{\mathrm{seq}} \times N_{\mathrm{cond}}$ sequence–condition cells exposed to the evaluator	Available cells
Niche size	M	Number of conditions selected for one returned niche	Niche constraint
Tree budget	R	Maximum total rooted mutation-tree cost used to construct a repertoire	Construction constraint
Leaf cap	K	Maximum number of terminal specialist leaves in the rooted tree	Repertoire constraint
Edge edit width	w	Maximum number of substitutions introduced by one tree edge	Locality constraint
Assay ceiling	B	$\min\{N_{\mathrm{cell}}, \max(8N_{\mathrm{cond}}, \mathrm{round}(0.08N_{\mathrm{cell}}))\}$ cells/campaign	Information budget

Eligibility and specialist headroom. A benchmark instance must permit meaningful diversification. For a candidate niche C, let $\mathcal{X}_R$ denote the sequences constructible from x_0 within the construction budget R. The best reachable generalist and condition-wise specialist ceiling are

$$G_R(C) := \max_{x \in \mathcal{X}_R} \frac{1}{M} \sum_{c \in C} f(x, c), \qquad U_R(C) := \frac{1}{M} \sum_{c \in C} \max_{x \in \mathcal{X}_R} f(x, c). \tag{47}$$

These quantities correspond to the accessible anchors defined in Eq. 2 and support the normalized objective in Eq. 3. Eligible niches satisfy $U_R(C) - G_R(C) \geq 0.05$, preventing the normalized objective from being dominated by an arbitrarily small generalist–specialist gap. Because $U_R(C)$ optimizes each condition independently, its maximizing variants need not be jointly realizable in one rooted mutation tree. A returned repertoire must still satisfy the construction budget, leaf cap, and admissible-edit constraints as a jointly feasible rooted mutation tree.

Preparation is deterministic: duplicate aligned keys are removed, one condition order is retained, and every retained cell must be finite. The record stores the source identifier, utility orientation, library and panel sizes, and the prepared instance key. The resulting record is used by the algorithmic pipeline described in Appendix B. Missing cells and ambiguous keys are exclusions, not imputed values; changing the library, orientation, or condition panel creates a new key.

D.2 Compared Baseline Adaptations

Adaptation principle. The comparison treats each baseline as a decision procedure operating on a prepared EGRET instance. Its original acquisition logic determines the evidence requested during a campaign, and the adaptation expresses that logic through the niche and rooted tree representation used by the benchmark. This keeps the comparison focused on how different proposal principles allocate information when useful endpoints must be assembled under a shared construction budget. The resulting ports and their declared settings are summarized in Table 13, with additional variants documented in Appendix E.

Table 13: Compared baseline adaptations. Source principles and venues identify the cited method; adaptations identify the finite-library, assay-metered interface.

Method	Venue	Source principle	Benchmark adaptation
STO (Ghuge & Nagarajan, 2020)	SODA	Tree orienteering	Complete-profile rows; greedy marginal coverage per round.
BOCS (Baptista & Poloczek, 2018)	ICML	Quadratic BO / Thompson	Mutation-feature surrogate; finite search over niche trees.
COMBO (Oh et al., 2019)	NeurIPS	Diffusion-kernel GP / UCB	Hamming graph GP; UCB tree search.
Casmopolitan (Wan et al., 2021)	ICML	Categorical trust region	Hamming trust region; restarts from feasible trees.
PEX (Ren et al., 2022)	ICML	Proximal mutation exploration	Mutation ridge; frontier profiling with one-edit candidates.
PR (Daulton et al., 2022)	NeurIPS	Probabilistic reparameterization	Categorical leaf slots; score-function search over leaves.
CBBO (Huang et al., 2026)	ICLR	Tensor-output CMAB-UCB2	Sequence–condition GP; super-arm updates across niches.
DiscoBAX (Lyle et al., 2023)	ICML	Greedy subset + BAX	Feasible subsets; posterior disagreement over candidates.
GameOpt (Bal et al., 2025)	ICLR	Cooperative-game UCB	Mutable-position players; Hedge proposals over candidates.
Random	Control	Random control	Unopened cells; random draws from feasible niche/trees.

Interpretation of the adaptations. The table records the translation layer between each source algorithm and the decision problem studied here. Internal choices enter the comparison through the candidates they make available, the assay requests they generate, and the repertoire decisions those candidates can support. These choices are interpreted through the repertoire objective in Section 3 and the acquisition procedure in Section 4. This perspective keeps the adapted methods tied to the paper's construction aware objective and gives the listed interfaces a direct operational interpretation.

D.3 Common Evaluation Protocol

Prepared instance. Every comparison begins from a prepared benchmark instance with a public sequence library, a designated scaffold, a fixed condition panel, and a fixed utility orientation. The structural parameters (R, K, w) are part of this instance definition and determine which repertoires can be constructed. The assay ceiling B controls how much sequence and condition information can be opened during a campaign. Table 12 records these operational quantities, whose role follows the problem formulation in Section 3. These settings are established before a method starts proposing measurements for the campaign.

Assay accounting. Each method may choose its own proposal sequence, but all requests enter the same assay record. Repeated requests are removed, opened cells consume the remaining allowance, and completed rows are retained whenever the budget permits. Residual cells follow the declared seeded order, which makes the information available to each campaign reproducible. The row to cell transition and adaptive screening logic are described in Section 4.3 and Appendix B.4. The resulting measured mask grows monotonically as the campaign proceeds.

Hyperparameter record. Campaign settings are fixed before each run, with hidden cells held out during tuning. Table 14 lists the values needed for reproducibility, and the method YAML files preserve the complete configuration used across the benchmark. These settings govern the model, acquisition, and search stages specified in Appendix B. The record covers model, acquisition, search, round, and profile settings for each evaluated method.

Table 14: Fixed method hyperparameters used by the benchmark ports. A dash denotes that no value is reported for the corresponding field.

Method	Profiles	Rounds	Starts	Model	Acquisition	Search/other
STO	–	–	12	–	–	tree-search fixed
BOCS	8	8	8	$d = 128$; $\lambda = 1$	Thompson $= .8$	–
COMBO	10	6	8	$\beta = .8$; $\lambda = .05$	UCB $= 1$	bases $= 64$
Casmopolitan	8	8	8	$\beta = .8$; $\lambda = .05$	UCB $= 1$	bases $= 64$; radius $= 5$ $[2, 8]$
PEX	8	8	8	$d = 128$; $\lambda = 2$	–	neighbours $= 48$
PR	8	6	–	$\lambda = 2$	UCB $= 1$; rate $= .5$	gradient $= 24$; MC $= 16$; draws $= 12$
CBBO	8	8	–	$\beta = .8$; $\lambda = .05$	$\beta_{\mathrm{UCB}} = 1$; $\delta = .1$	bases $= 64$; niche $= 512$
DiscoBAX	8	6	–	$d = 128$; $\lambda = 2$	–	posterior draws $= 8$
GameOpt	8	6	–	$d = 128$; $\lambda = 2$	UCB $= 1$	restarts $= 4$; iterations $= 12$; rate $= 4$
EGRET (ours)	–	6	–	$\lambda \in \{.5, 2, 8, 32\}$	seed $= .18$; explore $= .15$	draws $= 8$; runner-up $= 3$; beam $= 48$; shortlist $= 10$
Random	–	–	–	–	–	–

D.4 Scope of Our Experimental Studies

Each prepared instance enters an experimental study after satisfying checks for unique aligned sequences, a complete finite matrix, specialist headroom, and a feasible rooted tree. These checks determine the set of instances contributing to each reported denominator. Table 15 summarizes the resulting study scope. The main paper experiments use the shared benchmark inventory and assay ceiling. Appendix E lists the declared subsets, arms, and exclusions for the additional analyses conducted under this protocol.

Denominators and interpretation. The headline comparison aggregates twenty prepared rows across five seeds. Missing runs are recorded as failures in the corresponding denominator. Interventions and ablations retain their transformed keys and declared study configurations. The eligibility checks and study specific denominators are listed in Table 15, with additional analyses described in Appendix E. Reported scores summarize computational evaluations under the specified finite library protocols; wet lab validation and transfer require separate evidence.

Table 15: Scope of the Appendix E studies. Counts and arms are study-specific.

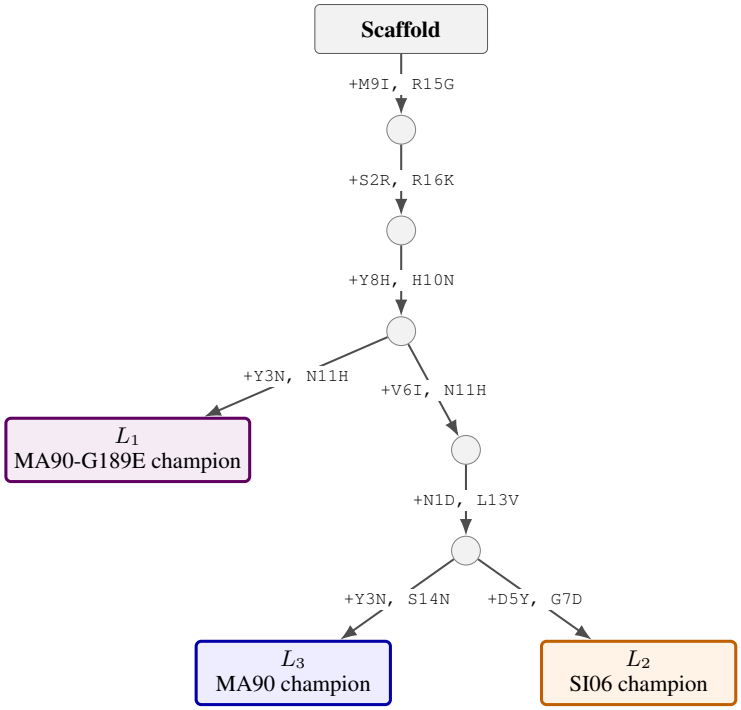
Reference	Study	Design	Purpose and compared arms
Section E.1	Qualitative lineage analysis	`CH65`/`Azoarcus` replays.	Posterior lineage, measured commitment, and tree sharing; illustrative cases.
Section E.2	Budgeted decision dynamics	Four landscapes; 5 seeds; fixed-niche probes.	Niche switches, anytime scores, ledgers, and specialist contributions.
Section E.3	Specialist-repertoire geometry	Four three-condition panels.	Repertoire trade-offs and interaction terms.
Section E.4	Explainable mutation analysis	10 instances across four assay types; 3 seeds.	EGRET, EGRET-mean, IDEAS, IDEAS-rand; commitment audit.
Section E.5	Robust structural search	16 instances; 5 seeds; 8% budget; 2340 campaigns.	EGRET, DADO variants, and BoBa variants under matched commitment.
Section E.6	Posterior calibration	Six instances; 3 seeds; standard 8% budget.	Landscape draws and variance arms on unbought cells; post-hoc audit.
Section E.7	Decision traceability	Five instances; 3 seeds; standard 8% budget.	EGRET, TS-DE variants, AutoSciLab; provenance and replay.
Section E.8	Hyperparameter sensitivity	Twenty instances; structural and configuration sweeps.	Locality, resource, capacity, refitting, and exploration controls.
Section E.9	Failure cases	Twenty instances; controlled protocol interventions.	Initialization and upper-tail sensitivity across benchmark instances.

E ADDITIONAL RESULTS

E.1 QUALITATIVE LINEAGE ANALYSIS

CH65 lineage and specialization. Figure 8 visualizes a deterministic EGRET run on the CH65 influenza landscape. The panel comprises three H1 strains – MA90, its G189E escape variant, and SI06 – and the repertoire assigns a different terminal variant to each. All three descendants retain the first six mutations, after which the MA90-G189E specialist branches early while the MA90 and SI06 specialists reuse a deeper four-mutation trunk. The result therefore exposes the construction decision that a flat list of final sequences hides: mutations shared before a branch are installed only once under shared lineage constraints.

The run spends all eight available construction steps. Its measured leaf profiles show complementary specialization: L_1 is strongest on MA90-G189E, L_2 on SI06, and L_3 on MA90. The resulting repertoire captures substantial specialist headroom ($\Phi_R = 0.754$), while the posterior estimate $\widehat{\Phi}_R = 0.803$ is mildly optimistic. This gap illustrates why EGRET uses its posterior to explore lineages but bases the final condition-to-leaf assignment on measured utilities.



Leaf utility	MA90	MA90-G189E	SI06	Assigned condition
L_1	0.723	**0.998**	0.667	MA90-G189E
L_2	0.758	0.847	**0.967**	SI06
L_3	**0.989**	0.955	0.867	MA90

Figure 8: **A committed EGRET lineage on CH65.** Shared edges encode reusable mutations; bold utilities mark the measured condition assignments.

Posterior explanation of the CH65 lineage. The left panel of Figure 9 exposes the coefficients behind the same run. Generalized cross-validation selects $\lambda = 0.5$ and site coherence $\rho = 1$, so this combinatorial library receives no additional pooling across residues at a position; the displayed values can therefore be read directly as mutation effects. N11H and Y3N have the largest positive MA90-G189E main effects, consistent with their placement on the branch leading to the MA90-G189E champion. Their joint MA90-G189E interaction is nevertheless strongly antagonistic (-0.229), correcting the additive model's otherwise excessive confidence. Conversely, Y4C×D5Y is positive in every condition. The model thus represents both reusable mutations and condition-dependent epistasis instead of explaining a leaf as a ranked list of single edits.

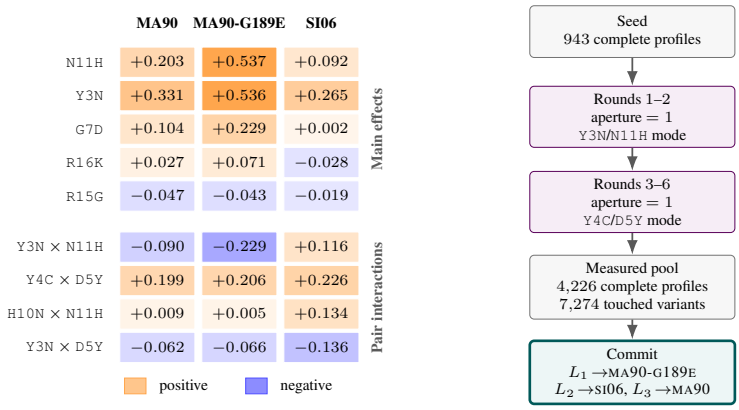
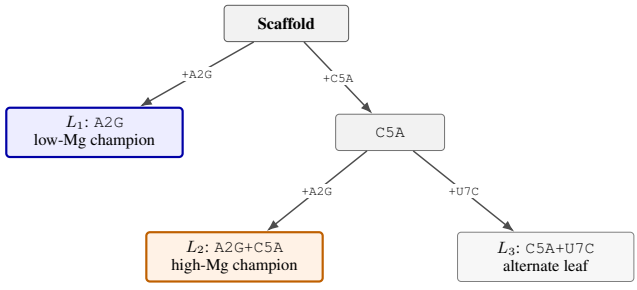


Figure 9: **How EGRET explains the `CH65` seed-0 decision.** The left panel visualizes selected signed main and pairwise effects from the final condition-specific posterior. The right panel traces the dominant sampled lineage motifs through all six acquisition rounds and the subsequent measured-value commitment.

The right panel of Figure 9 follows several consequential decisions. After 943 complete seed profiles, the spectral aperture is one in every round, reflecting strong panel-level redundancy and permitting broad screening. The most frequently sampled lineage changes from a `Y3N/N11H` family in rounds 1–2 to a `Y4C/D5Y` family in rounds 3–6 as new assays update the posterior. These sampled modes determine where EGRET measures; they are not treated as the final answer. Promotion produces 4,226 complete profiles among 7,274 touched variants, and verified commitment then reassembles the measured candidates into L_1, L_2, L_3. None of these three leaves is a round-level posterior-vote leader, which makes the separation between explainable exploration and evidence-based commitment explicit.

Compact `Azoarcus` ribozyme lineage. Figure 10 provides a smaller example on the `Azoarcus` ribozyme landscape across five $MgCl_2$ concentrations. The single mutant `A2G` covers the two low-Mg conditions, while the double mutant `A2G+C5A` covers 6–16 mM. EGRET also returns `C5A+U7C`; it shares the `C5A` construction step with the high-Mg champion but does not win a condition under the hidden evaluation. This extra leaf makes residual exploration error visible without preventing the measured assignment from recovering nearly all accessible specialist headroom under the same budget.



Leaf utility	$MgCl_2$ concentration (mM)					Assigned range
	1	2	6	8	16	
L_3	0.163	0.393	0.909	0.875	0.824	–
L_2	0.167	0.457	**1.000**	**0.997**	**1.000**	Mg 6–16 mM
L_1	**0.994**	**1.000**	0.849	0.957	0.888	Mg 1–2 mM

Figure 10: **A compact EGRET lineage on `Azoarcus`.** Shared `C5A` intermediate supports the high-Mg specialists; bold utilities mark the measured condition assignments.

E.2 Budgeted Decision Dynamics

When the answer was already decided. Figure 11 replays verified commitment at each frozen assay state. Selection uses measured values only; the hidden matrix is consulted afterwards, and the final replay matches the returned tree. `CH65` stabilizes early, whereas most instances continue changing late in the budget. The posterior is optimistic throughout, and `Azoarcus` also shows a separate failure: an infeasible niche can score zero even when its assembled tree is good.

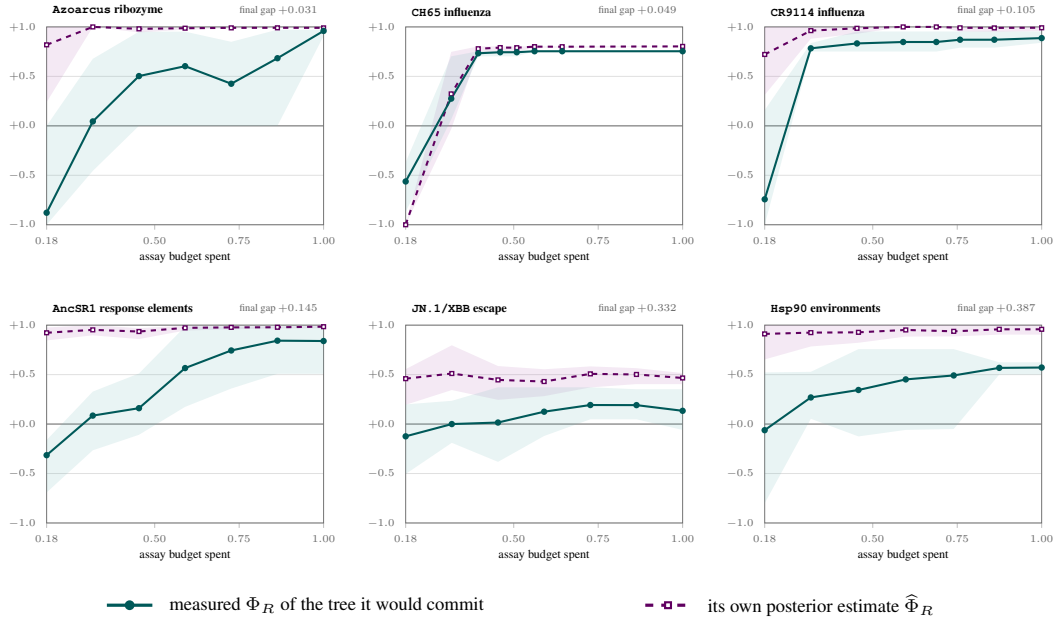
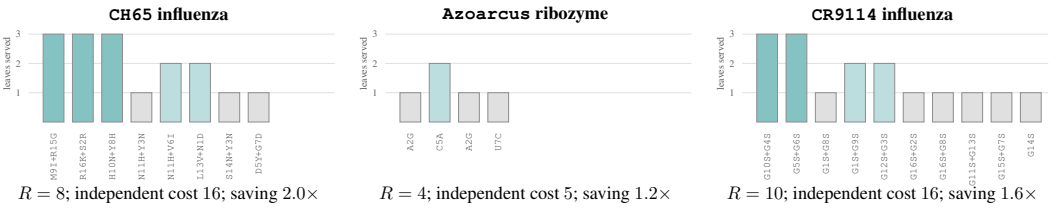


Figure 11: **Anytime verified commitment.** Curves replay the measured-only commitment as the assay budget grows; lines show five-seed means and bands show seed ranges across the six benchmark landscapes. The displayed domain starts at the first recorded commitment (about 18% budget). Panels are ordered by the final calibration gap $\widehat{\Phi}_R - \Phi_R$.

What the branching structure buys. Figure 12 makes the sharing mechanism explicit across three landscapes. Bar height counts descendant leaves, so a tall trunk represents a construction step reused by several specialists; shallow trees approach a star. On `CH65` this topology halves the cost of building the three endpoints separately.



$R = 8$; independent cost 16; saving 2.0× | $R = 4$; independent cost 5; saving 1.2× | $R = 10$; independent cost 16; saving 1.6×

Figure 12: **Construction ledgers across landscapes.** Bars show committed steps and their descendant leaves; shared trunks reduce the cost of building the same endpoints independently.

The normalized ledgers separate topology from plotting scale: a tall bar means that one mutation step supports several descendants, whereas a unit-height bar is private to one endpoint. `CH65` therefore spends its budget on a reusable trunk before paying for the terminal edits; `Azoarcus` is shallower and gains only a modest reuse factor, while `CR9114` has enough depth for an intermediate branch. This is the construction-level reason a tree can fit specialist endpoints that would exceed the same mutation budget when built independently.

Where sharing pays. The saving is a property of the available construction budget, not a universal constant of EGRET (Figure 13). Across the twenty instances, the ten cases with $R = 3$ average only $1.10\times$ and never exceed $1.33\times$; five are exactly $1.0\times$, because three steps leave little depth for a reusable trunk. Deeper cases average $1.46\times$, reaching $1.76\times$ on `CR9114` at $R = 14$ and $2.00\times$ on `CH65`. The tree formulation therefore earns its keep when the budget is deep enough to branch, whereas shallow panels reduce nearly to endpoint selection.

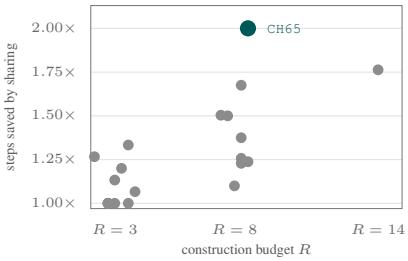


Figure 13: **Where sharing pays.** Steps the committed leaves would cost built independently, divided by the steps the tree spends, averaged over five seeds on each of the 20 instances.

Niche revisions during acquisition. Figure 14 audits the selected condition sets in the existing replay trajectories, using five seeds on each of four landscapes with a nontrivial niche choice. A cell counts changes between consecutive snapshots; it does not measure changes to the tree within a fixed niche. `AncSR1` revises its niche in every seed over the first three transitions, then largely settles. `Azoarcus` keeps revising later, consistent with the niche-feasibility issue in Figure 11.

Revisions do not guarantee improved hidden-matrix scores: five of thirteen switching transitions decrease Φ_R on each of `Lineage escape` and `Hsp90`, compared with three of twenty-two on `Azoarcus`. These are descriptive paired transitions, not a causal effect of switching: both the available assays and the assembled tree can change between snapshots. The diagnostic identifies niche stability as a distinct aspect of convergence that an apparently stable tree score can hide.

	1	2	3	4	5	6
Azoarcus	5	5	4	2	4	2
AncSR1	5	5	5	4	1	1
Lineage escape	3	2	3	1	2	2
Hsp90	5	4	1	1	2	0

Figure 14: **Niche revisions during acquisition.** Each cell counts seeds changing their selected condition set (0–5). Columns follow six successive replay transitions from the seed assays to the final commitment.

Contribution of individual specialists. Figure 15 removes each terminal leaf in turn from six seed-0 commitments with three terminal leaves each. For every intervention, we retain only the original root-to-leaf paths needed by the two surviving endpoints, validate the pruned tree, and rescore it on the same selected conditions. No refitting or reallocation of the freed construction steps is allowed. The reported loss is the decrease in mean best-leaf utility, so it measures each endpoint's marginal coverage of the fixed niche without changing the normalization anchors or selected conditions.

	1	2	3
CH65	0.077	0.033	0.014
CR9114	0.146	0.108	0.021
Azoarcus	0.274	0.048	0.000
Hsp90	0.123	0.034	0.024
AncSR1	0.261	0.182	0.020
DAOx	0.050	0.027	0.026

Figure 15: **Leave-one-leaf-out contribution.** The loss in mean best-leaf utility after each deletion, with the niche fixed (seed 0). Columns rank losses within each dataset.

Endpoints contribute on `CH65`, `CR9114`, `Hsp90`, `AncSR1`, and `DAOx`. `Azoarcus` has a zero-loss deletion: its alternate leaf adds no coverage beyond the two champions; deleting its most consequential specialist loses 0.274. This separates indispensable specialists from redundant endpoints in the repertoire.

The added landscapes also distinguish concentrated from distributed coverage. `AncSR1` relies primarily on two specialists: its smallest deletion loss is over nine times smaller than the second-largest. `DAOx` spreads its marginal coverage more evenly, with the largest loss less than twice the smallest. Thus, three terminal leaves can encode different degrees of dependence on individual specialists. The columns rank contributions within each run, so they do not identify corresponding mutations across datasets. These fixed-niche interventions describe the returned repertoires; they do not predict how a refitted model would spend the construction steps freed by pruning.

E.3 Geometry of Specialist Repertoires

The views below read a committed repertoire in per-condition coordinates, so they cover the four benchmark instances whose niche is three conditions wide: two axes and a colour. They illustrate these cases rather than summarize the full benchmark.

Repertoires in response space. Figure 16 plots each panel in its own measured coordinates rather than a projection, so a position is a pair of utilities and needs no explained-variance caveat. The committed leaves occupy different corners of the reachable cloud in every case, but the shape of that separation differs. On `CR9114` it is stark: L_3 reaches 0.971 on H1 while scoring 0.000 on H3 and influenza B, so no single sequence could stand in for the set. On `CH65` the three H1 strains are correlated and the leaves stay in the same upper-right region, yet remain complementary, with L_3 leading on MA90 at 0.989, L_1 on the G189E escape variant at 0.998 and L_2 on SI06 at 0.967; the generalist anchor coincides with L_3 there, which is why the headroom is narrow. The other two panels are trade-offs, and the repertoire splits along the trade-off: `CH65+Expr` sets full expression with weak binding $(0.448, 0.392, 1.000)$ against its inverse $(0.989, 0.955, 0.379)$, and `Lineage escape` separates a REGN10987 specialist from an AZD8895 one.

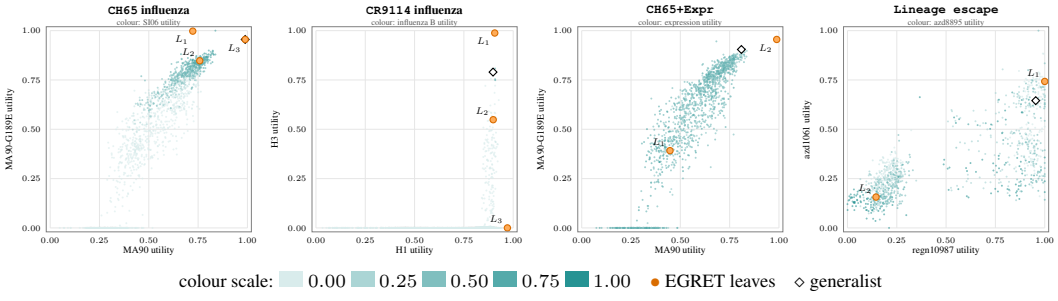


Figure 16: **Specialists in directly interpretable response coordinates.** Each point is a reachable sequence. Position encodes the first two conditions of the niche and colour the third, named above each panel, so the axes are measured utilities rather than a projection. Background points are uniformly sampled; all seed-0 leaves and the best reachable generalist are shown. On `CH65` the generalist coincides with L_3.

Fitted coefficients behind the repertoires. Figure 17 pools the fitted coefficient means for the same four runs. All four concentrate main effects away from zero while keeping pair interactions tight around it, the shape the site-coherent prior is meant to produce. Sign balance separates them: `CH65` fits 71% of its main effects positive and `CH65+Expr` 69%, so their mutations mostly add, whereas `CR9114` fits only 35% positive, consistent with a broadly neutralizing antibody whose mutations trade one subtype against another. The interactions matter unevenly. Summed over the committed leaves they move the predicted profile by 0.787 on average on `CR9114`, with one correction reaching +2.216, against 0.332 on `CH65` and under 0.09 on the two trade-off panels, so an additive read would misplace the endpoints on the influenza panels while being nearly harmless elsewhere. These histograms describe variation across effects and not posterior uncertainty.

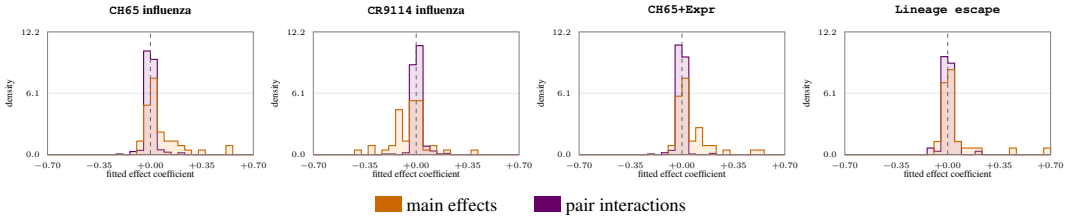


Figure 17: **Distribution of interpretable posterior coefficients.** Histograms pool fitted coefficient means across the three conditions, normalized separately for each component with identical bins. They describe variation across effects, not posterior uncertainty.

E.4 Explainable Mutation Analysis

IDEAS (Pandey et al., 2026) uses explanations as a search signal: a COLOR importance model is refit each round, and its position and replacement scores guide AdaLead-style mutations. We evaluate this operator under the same finite-library assay protocol and construction budget as EGRET, translating its scores into feasible candidate acquisitions. The comparison spans ten antibody, enzyme, expression, and partner-binding instances, with three seeds each. Table 16 evaluates the returned repertoire and the faithfulness of fitted mutation scores. Measured ceiling is the best condition-wise utility in the fully profiled pool; committed shortfall is its gap to the returned tree. Distinct winners, depth, and sites summarize the pool.

Table 16: **EGRET against IDEAS.** Mean over ten small and mid-sized instances, three seeds each, at an 8% assay budget. Best values are bold; runner-up values are underlined.

	EGRET	EGRET-mean	IDEAS	IDEAS-rand
Mutation				
$\Phi_R\uparrow$	<u>0.518</u>	**0.534**	0.385	0.405
Measured ceiling↑	0.931	0.932	<u>0.936</u>	**0.938**
Distinct winners↑	**3.500**	3.367	**3.500**	<u>3.433</u>
Committed shortfall↓	**0.013**	<u>0.014</u>	0.050	0.046
Profiled depth	2.498	2.500	3.031	3.051
Sites touched	0.744	0.728	0.865	0.863
Explanation				
ρ_s vs. truth↑	**0.594**	<u>0.575</u>	0.080	0.049
$\lvert\rho_s\rvert$ vs. truth↑	**0.283**	<u>0.260</u>	0.149	0.154
P@10 (wide)↑	<u>0.213</u>	**0.240**	0.007	0.000
Seed stability↑	**0.646**	<u>0.627</u>	0.551	0.565

Controls and budget-aware commitment. Each method also runs with its explainer fitted every round but not steering. EGRET-mean explores on the posterior mean instead of sampling lineages, whereas IDEAS-rand mutates uniformly instead of following the importance scores. Table 16 shows no measurable loss from either control: EGRET-mean differs by -0.016 and IDEAS-rand by -0.020. Their families remain separated with both explainers off ($+0.129$), locating the difference in acquisition rather than explanation quality.

The mutation rows of Table 16 separate raw variant quality from feasible commitment. IDEAS reaches a slightly higher measured ceiling (0.936 versus 0.931), but EGRET converts its pool into a stronger tree (0.518 versus 0.385) with lower shortfall (0.013 versus 0.050). The largest gaps occur on `FP switch` and `Omicron+ACE2`, where scattered variants are difficult to assemble within one shared construction budget under matched constraints.

Faithful mutation effects. The explanation rows of Table 16 compare mutation scores with hidden marginal effects: ρ_s tests signed rank agreement, $\lvert\rho_s\rvert$ removes sign, P@10 tests top-ten overlap, and seed stability measures agreement across runs. EGRET's signed posterior effects are more faithful and more stable, especially when pair terms capture background-dependent effects.

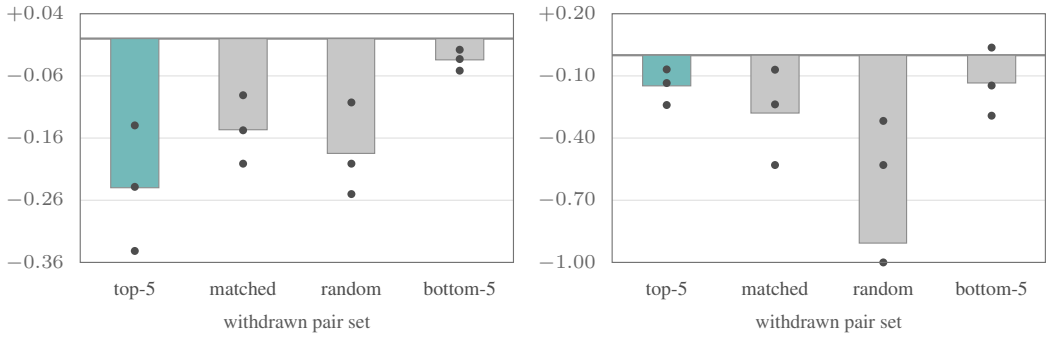


Figure 18: **Pair withdrawal.** Changes in measured ceiling (left) and Φ_R (right). Bars average three instances and three seeds; dots show instance means.

Pair withdrawal. Figure 18 supports EGRET's ranking of pairs relevant to reachable utility. Top-5 withdrawal lowers measured ceiling by 0.240, versus 0.147 for row-count-matched random pairs and 0.034 for bottom-5 pairs. The top-5 versus matched difference (-0.093; 95% interval $[-0.174, -0.013]$) identifies consequential pairs beyond library-size effects, supporting the representation's relevance to search. The Φ_R difference is inconclusive ($+0.131$; $[-0.098, +0.361]$), since withdrawal also changes normalization anchors.

Interpretable pair effects. Figure 19 shows that EGRET's readable pair terms recover measured biological interactions. Against double-mutant epistasis, $u(ab)-u(a)-u(b)+u(\emptyset)$, the 90 audited pairs achieve 76.7% sign agreement and $\rho_s = 0.699$ (Pearson 0.862). The posterior thus captures interaction direction and relative strength. On `FP switch`, agreement reaches 0.87 and $\rho_s = 0.846$; versus 0.70 and 0.369 on `Lineage escape`, showing variation across landscapes. Estimated magnitudes are smaller on average (0.056 versus 0.074 measured), consistent with ridge shrinkage while retaining directional information. These results support EGRET's interpretable representation of background-dependent mutation effects. The audit covers three instances with complete mutant quartets, excluding the seven single-mutant panels that lack the required combinations.

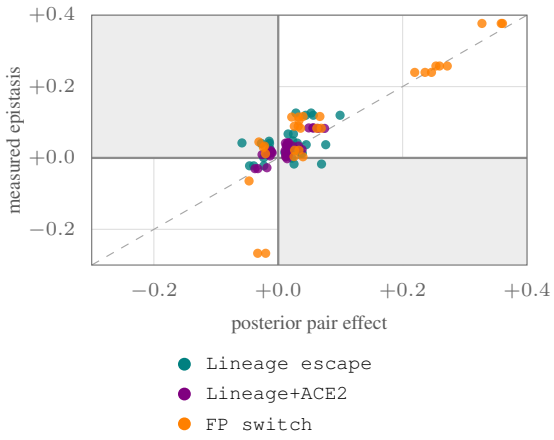


Figure 19: **Measured epistasis audit.** Posterior pair effects versus double-mutant cycles. Dashed: equality; shaded: sign disagreement.

E.5 Robust Search under Structural Commitments

Recent methods gain efficiency through fixed structure. DADO (Bowden et al., 2026) factorizes its search distribution over a prescribed interaction graph, whereas BoBa (Haddadnia et al., 2026) allocates inference and assays across a fixed library partition. EGRET instead learns pair terms each round and adapts screening to the condition spectrum. We test how these choices affect the repertoire returned from a finite, assay-metered library, separating the value of adaptive structure from predictive fit alone under matched assay and construction budgets.

Protocol. We run 2340 campaigns over sixteen combinatorial instances, five seeds and a common 8% assay budget. Four single-mutant panels are excluded because they contain no observed mutation pairs from which to build DADO's graph. Each method retains its source model and acquisition but uses the same measured-value commitment rule. DADO-joint removes clique credit, DADO-refit relearns the decomposition each round, and BoBa-uni replaces bandit allocation with uniform allocation. Every method selects its entire assay allowance itself.

Conversion, not discovery. Table 17 separates variant discovery from feasible commitment. EGRET leads Φ_R at 0.670 against 0.519 for the strongest baseline and ranks 1.75 on average against 3.06. Yet DADO-joint attains a similar measured ceiling (0.937 versus 0.935), so EGRET's gain does not come from finding better individual variants. It converts the profiled pool more effectively: committed shortfall is 0.012, compared with 0.041–0.045, and its worst-condition margin is −0.045, compared with −0.078 to −0.112. These gaps show the value of pricing shared construction edges during acquisition. Strong variants alone do not form a strong repertoire when the construction budget cannot assemble them together.

Table 17: **Robust search under structural commitments.** Mean over sixteen combinatorial instances, 5 seeds each, at an 8% assay budget under a common oracle and commitment rule. Best values are bold; runner-up values are underlined.

	EGRET	DADO	DADO-joint	DADO-refit	BoBa	BoBa-uni
Decision quality						
$\Phi_R\uparrow$	**0.670**	0.476	0.494	0.519	0.475	0.443
Mean rank↓	**1.750**	4.188	3.875	3.062	3.625	4.500
Tree value $V\uparrow$	**0.923**	0.876	0.895	0.882	0.878	0.877
Measured ceiling↑	0.935	0.918	**0.937**	0.926	0.923	0.921
Committed shortfall↓	**0.012**	0.042	0.041	0.045	0.045	0.043
Worst-condition margin↑	**−0.045**	−0.102	−0.085	−0.078	−0.112	−0.096
Repertoire						
Distinct winners↑	2.325	2.413	2.337	**2.425**	2.362	2.350
Leaf spread↑	**5.496**	5.192	4.958	5.021	5.058	5.142
Trunk share	0.506	0.531	0.532	0.533	0.537	0.528
Model						
ρ_s vs. truth↑	**0.759**	0.734	0.753	0.727	0.654	0.662
Top-50 optimism → 0	−0.033	−0.002	**+0.002**	−0.006	−0.007	−0.009
Top-50 true value↑	0.603	0.596	0.594	**0.603**	0.593	0.589

Model fidelity does not explain the gap. EGRET and DADO-joint have similar rank agreement with hidden utility (0.759 versus 0.753), despite a 0.176 gap in Φ_R. EGRET also understates its top rows by 0.033, so optimism does not explain the advantage. Model accuracy and feasible decision quality are therefore distinct.

Perturbing the campaign. Table 18 perturbs assay noise and the two budgets. Under clean evaluation, EGRET leads Φ_R in all five conditions. At half the construction budget it remains positive at 0.134, while DADO and BoBa fall to -0.381 and -0.352: without pricing a shared trunk, they profile pools the tighter budget cannot assemble. EGRET has best worst-seed result (0.244), smallest seed spread (0.171), and lowest degenerate-repertoire rate (0.110). Degeneracy means no captured headroom or a single-leaf solution. The tightened-budget row omits `S309 backgrounds` and `JN.1/XBB` because their halved budgets fail the joint-constructibility check in this protocol.

Table 18: **Robust search under stress.** Assay noise, assay budget and construction budget are perturbed separately; evaluation uses clean measurements. Best values are bold; runner-up values are underlined.

	EGRET	DADO	DADO-joint	DADO-refit	BoBa	BoBa-uni
Φ_R under perturbation						
Nominal	**0.670**	0.476	0.494	<u>0.519</u>	0.475	0.443
Assay noise $\sigma = 0.01$	**0.614**	0.466	0.497	<u>0.510</u>	0.461	0.458
Assay noise $\sigma = 0.05$	**0.421**	0.245	<u>0.288</u>	0.278	0.180	0.220
Half assay budget (4%)	**0.488**	0.290	0.249	<u>0.423</u>	0.321	0.278
Half construction budget ($R/2$)	**0.134**	−0.381	−0.409	−0.178	−0.352	<u>−0.170</u>
Retention: fraction of the method's own nominal Φ_R						
Assay noise $\sigma = 0.01$	0.917	0.979	<u>1.005</u>	0.984	0.970	**1.034**
Assay noise $\sigma = 0.05$	**0.628**	0.516	<u>0.583</u>	0.537	0.379	0.496
Half assay budget (4%)	<u>0.728</u>	0.609	0.503	**0.817**	0.675	0.627
Half construction budget ($R/2$)	**0.200**	−0.801	−0.828	<u>−0.344</u>	−0.742	−0.385
Dispersion, pooled over all five conditions						
Worst-seed $\Phi_R\uparrow$	**0.244**	−0.133	−0.203	<u>0.107</u>	−0.042	−0.125
Seed s.d.↓	**0.171**	0.275	0.322	<u>0.184</u>	0.238	0.270
Degenerate repertoires↓	**0.110**	0.226	0.215	<u>0.174</u>	0.215	0.228

Table 19: **Robust search, per instance.** Nominal Φ_R on each of the sixteen instances, mean over 5 seeds with the seed standard deviation beside it. Per-instance differences at this assay budget are within seed noise on the easier landscapes; the aggregate is what the comparison rests on. The final row reports mean / instances led.

	EGRET	DADO	DADO-joint	DADO-refit	BoBa	BoBa-uni
`S309 backgrounds`	0.428 ±0.07	0.337 ±0.14	<u>0.460</u> ±0.15	**0.471** ±0.10	0.338 ±0.21	0.141 ±0.26
`BA.2 backgrounds`	<u>0.598</u> ±0.06	0.552 ±0.13	0.304 ±0.20	**0.634** ±0.00	0.531 ±0.10	0.443 ±0.22
`Lineage escape`	**0.401** ±0.03	<u>0.351</u> ±0.02	0.049 ±0.61	0.320 ±0.02	0.285 ±0.08	0.163 ±0.28
`Lineage+ACE2`	**0.617** ±0.01	0.429 ±0.16	0.360 ±0.21	0.501 ±0.18	<u>0.575</u> ±0.02	0.408 ±0.24
`FP switch`	**0.887** ±0.00	0.580 ±0.37	<u>0.747</u> ±0.27	0.251 ±0.30	0.229 ±0.33	0.669 ±0.35
`Azoarcus`	0.959 ±0.03	0.954 ±0.03	0.867 ±0.20	**0.981** ±0.02	<u>0.963</u> ±0.02	0.668 ±0.32
`Potts`	<u>0.677</u> ±0.02	0.660 ±0.05	0.677 ±0.02	0.663 ±0.03	**0.688** ±0.00	0.639 ±0.11
`AncSR2`	0.859 ±0.16	0.768 ±0.24	0.874 ±0.09	<u>0.927</u> ±0.03	**0.943** ±0.02	0.916 ±0.05
`Omicron escape`	**0.751** ±0.34	0.331 ±0.38	<u>0.544</u> ±0.42	0.274 ±0.41	0.373 ±0.36	0.389 ±0.35
`Omicron+ACE2`	**0.839** ±0.22	0.486 ±0.07	<u>0.599</u> ±0.20	0.437 ±0.07	0.489 ±0.08	0.518 ±0.06
`JN.1/XBB`	0.134 ±0.19	<u>0.138</u> ±0.25	−0.032 ±0.19	**0.139** ±0.08	0.031 ±0.15	0.039 ±0.14
`AncSR1`	**0.839** ±0.20	−0.045 ±0.42	<u>0.354</u> ±0.58	0.171 ±0.53	0.238 ±0.52	0.206 ±0.31
`KRAS`	**0.228** ±0.13	−0.160 ±0.29	−0.214 ±0.40	<u>0.182</u> ±0.13	−0.076 ±0.27	−0.219 ±0.12
`CR9114`	**0.887** ±0.05	0.860 ±0.05	0.843 ±0.06	0.862 ±0.05	0.845 ±0.06	<u>0.883</u> ±0.07
`CH65+Expr`	**0.864** ±0.00	0.748 ±0.00	<u>0.777</u> ±0.06	0.748 ±0.00	0.510 ±0.21	0.600 ±0.18
`CH65`	**0.754** ±0.00	0.628 ±0.26	0.703 ±0.00	<u>0.737</u> ±0.02	0.638 ±0.18	0.621 ±0.18
Mean / instances led	**0.670 / 10**	0.476 / 0	0.494 / 0	<u>0.519</u> / <u>4</u>	0.475 / 2	0.443 / 0

Landscape-level evidence and trade-offs. Table 19 shows that the aggregate is not driven by one landscape: EGRET leads ten of sixteen instances, against four for DADO-refit and two for BoBa. Its margin is largest on genuine trade-off niches — `AncSR1`, `Omicron escape`, `FP switch` and `KRAS` — and smallest on `Azoarcus` and `Potts`, where all methods recover most accessible headroom. `JN.1/XBB` remains hard because its deep, thirty-three-site library contains few jointly constructible pairs at this assay budget. This pattern matches EGRET's construction-aware specialization objective.

BoBa retains its computational advantage, evaluating 1.78 reachable libraries per campaign versus 5.33 for DADO and 59.59 for EGRET. The baselines retain more nominal score under mild assay noise, but EGRET achieves the highest absolute Φ_R. Within DADO, refitting reaches 0.519, versus 0.494 for joint credit and 0.476 for the fixed decomposition. This ordering implicates stale estimated structure, with the conclusion limited to that setting.

E.6 Posterior Calibration

Equation 6 separates coefficient uncertainty from uncertainty in unobserved substitutions. Site shrinkage transfers site-level means without transferring assayed-residue confidence, so lineage sampling retains residue-level uncertainty when comparing candidates.

We run EGRET on six instances at the standard 8% budget, using three seeds, and score empirical draw quantiles on unbought cells of reachable rows. Four arms alter only the sampling variance: the shipped landscape draw, that draw with residual noise restored, removal of the unseen-residue term without refitting, and a homoscedastic reference. The audit remains invisible to search.

Landscape versus predictive draws. Figure 20 also clarifies what EGRET samples. Its shipped draw represents a fitted landscape, so it deliberately omits residual noise and falls below predictive coverage. Restoring that scale raises nominal-0.90 coverage from 0.56 to 0.77 and improves CRPS from 0.0919 to 0.0870. Thus the search uses comparatively narrow ranking draws, while verified commitment protects the returned tree by scoring terminal choices only from measured cells.

Unseen-residue uncertainty. Figure 21 isolates this term. Removing it lowers nominal-0.90 coverage from 0.56 to 0.34 and halves mean interval width. The loss concentrates on deep panels whose held-out sequences contain unseen substitutions. It is negligible on `Azoarcus` and `CH65`, whose designs expose nearly every residue token. The mechanism thus activates only in its intended regime.

Benefits of site coherence. Disabling the site/residue hierarchy increases held-out error from 2.66 to 3.36 residual scales. Its higher coverage comes from intervals 60% wider with worse CRPS; site coherence therefore improves the mean and sharpens uncertainty.

Adaptive acquisition makes in-sample residuals understate error on declined cells. The audit therefore explicitly characterizes EGRET's campaign-conditioned search distribution, not an i.i.d. test population. This view matters for posterior-guided acquisition because it targets unresolved choices. The six datasets cover the deep and combinatorial regimes.

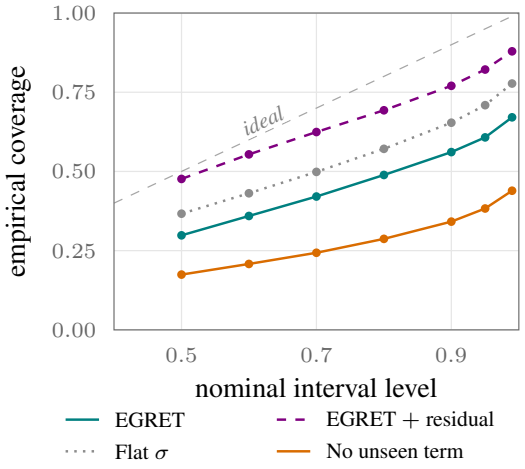


Figure 20: **Reliability on unbought cells.** 6 instances, 3 seeds each, on held-out cells of reachable rows. A draw of the fitted landscape sits below the diagonal at every level (0.56 at a nominal 0.90); restoring the residual scale Eq. 6 deliberately omits gives the posterior predictive, at 0.77.

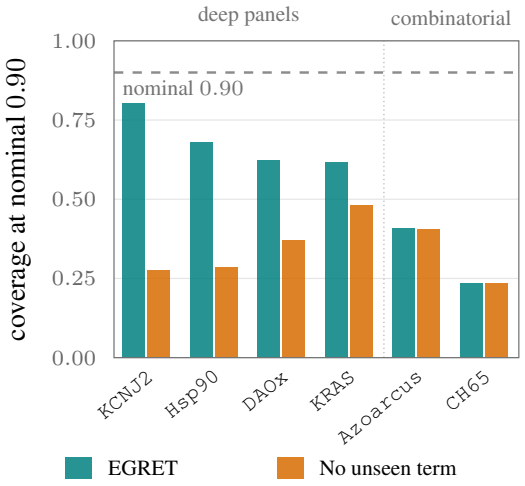


Figure 21: **What the unseen-residue term is worth.** Coverage at a nominal 0.90 with the term and with it removed, nothing refitted in between, so the whole difference is that one term: 0.56 against 0.34, at mean widths 0.198 and 0.091. The loss falls on the deep panels; on the combinatorial libraries the term is inert.

E.7 Decision Traceability

Decision quality alone does not show why a campaign bought particular assays. We compare EGRET with two methods that retain explicit accounts: TS-DE (Yuan et al., 2022) records the posterior coefficients that steer mutation, while AutoSciLab (Desai et al., 2025) reports a response-aligned equation. Each method keeps its own model and acquisition, proposes only reachable rows, and commits through the same measured-value rule. Differences therefore trace to acquired evidence rather than evaluation. AutoSciLab's equation is audited only as an explanation; commitment still uses measurements, avoiding circular agreement. We then test EGRET's substitution effects, TS-DE's sampled coefficients, and AutoSciLab's equation against their final decisions.

Is the stated account the operative one? Table 20 answers it with a pair of probes that only mean something together. Restricting the search to the half of the substitution vocabulary a method's own account ranks highest, EGRET still builds a repertoire worth $+0.145$; every baseline falls below the reachable generalist (-0.239 to -0.333). Withdrawing that half instead costs EGRET 0.478 more than withdrawing a matched random half, against $+0.117$ and $+0.129$ for the two TS-DE arms and -0.063 for AutoSciLab. EGRET's named effects therefore identify substitutions that are both sufficient for useful search and necessary for its strongest decisions. Its commitment is also almost entirely measured: support is 0.924 and the certification gap only 0.002. The ported methods reach 1.000 support because they profile and commit on whole rows; their near-zero reconstruction gaps follow from applying essentially the same rule twice. EGRET instead improves on a generic optimizer reading its log, reaching 0.553 rather than 0.518. Its 0.091 reconstruction gap thus reflects useful structure in verified commitment rather than unsupported model credit under the same protocol.

Table 20: **Decision traceability.** Mean over 5 instances and 3 seeds at an 8% assay budget. Provenance measures verified support; reconstruction tests whether the log reproduces the decision; sufficiency and necessity audit the substitutions named by each account. Best values are bold and runner-up values are underlined.

	EGRET	TS-DE	TS-DE-greedy	AutoSciLab
Decision				
$\Phi_R\uparrow$	**0.553**	0.142	0.141	<u>0.225</u>
Provenance				
Measured support of the tree↑	<u>0.924</u>	**1.000**	1.000	1.000
Certification gap↓	<u>0.002</u>	**0.000**	0.000	0.000
Certified $\Phi_R\uparrow$	**0.550**	0.142	0.141	<u>0.225</u>
Reconstruction				
Φ_R re-derived from the log↑	**0.518**	0.182	0.198	<u>0.251</u>
Reconstruction gap↓	0.091	**0.039**	0.057	<u>0.044</u>
Account				
Sufficiency: Φ_R on its own half↑	**0.145**	−0.308	−0.333	<u>−0.239</u>
Necessity: drop on withdrawal↑	**1.013**	0.680	0.639	<u>0.793</u>
Necessity over a matched random set↑	**0.478**	0.117	<u>0.129</u>	−0.063
Substitutions named	321	321	321	321

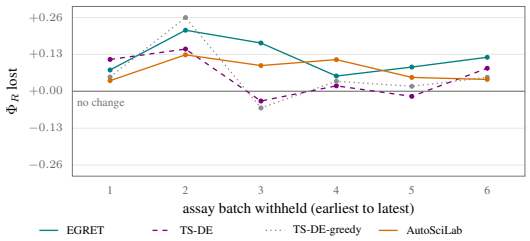


Figure 22: **Which assays the decision rests on.** Withholding each of 6 chronological assay batches and re-deriving the repertoire. Curves show mean Φ_R lost over 5 instances and 3 seeds; negative values mean removal improves the commitment across the evaluated methods.

Which assays the answer rests on. Figure 22 withholds each sixth of the assay log in turn and re-derives the repertoire from the rest. EGRET loses value after every withdrawal, from $+0.053$ to $+0.213$, so each campaign segment contributes to the final answer. Its evidence is distributed across the record while still showing which batches were most consequential. Both TS-DE arms become negative at the third and fifth batches (-0.034 and -0.058), meaning their commitment improves after some measurements are removed. Their sampled trajectories can therefore collect evidence that the final decision is better off ignoring. AutoSciLab remains positive but nearly flat around $+0.09$, consistent with repeated acquisition from a similar latent region. EGRET provides the stronger trace: every batch supports the result, yet their unequal effects reveal how the decision accumulated rather than merely confirming that assays occurred.

E.8 Hyperparameter Sensitivity

This section examines how EGRET's decisions respond to changes in the protocol settings that govern search and repertoire construction. The experiments assess the consistency of measured outcomes across configurations and identify which features of the reported behavior remain stable as resource limits, structural constraints, and experimental conditions vary. This analysis provides context for interpreting these results across the broader evaluation.

Lineage construction controls. Figure 23 shows that EGRET retains a positive measured margin as both locality and construction budget vary. Increasing the edit width from 1 to 5 lowers the paired difference from 0.445 to approximately 0.30, with all intermediate settings remaining positive. The budget sweep reaches its largest margin at $R = 5$ and then settles near 0.30 to 0.33 through $R = 14$. The result gives a stable picture across the tested tree configurations. The sweep uses clean observations and an 8% assay budget, with DiscoBAX fixed under its default constraints. Points summarize paired five seed differences across twenty instances, and every 95% bootstrap interval excludes zero throughout the reported comparison.

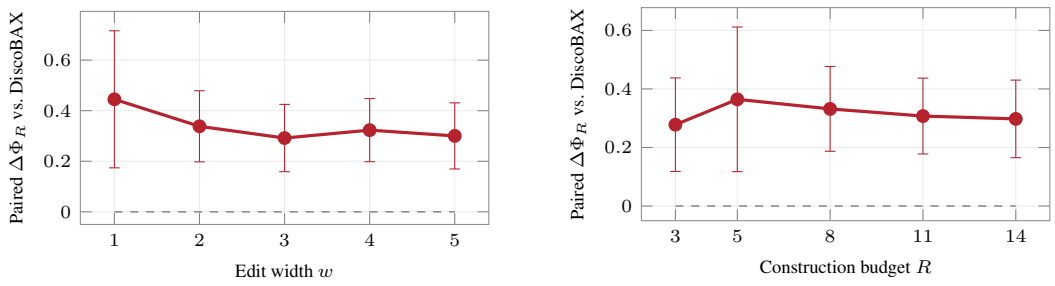


Figure 23: **Tree constraint sensitivity.** Paired performance differences under changes to edit width and construction budget. The assay allowance and evaluation procedure remain fixed, showing how lineage locality and construction capacity shape repertoire advantage across the benchmark.

Specialist capacity. Figure 24 examines the number of specialist leaves available to the returned tree. Across twenty instances and eleven methods, EGRET varies only from 0.583 to 0.636 over the tested leaf capacities. Method ordering remains unchanged, and the Random control lies below the plotted range. Most of the change occurs when capacity first increases, after which EGRET remains comparatively stable. Its persistent separation from DiscoBAX, whose means stay near 0.25 to 0.32, indicates that the comparative result is robust to this structural choice under the benchmark protocol.

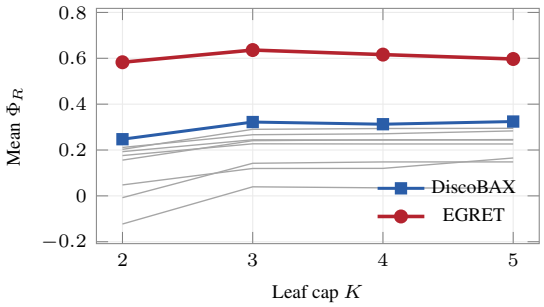


Figure 24: **Leaf cap sensitivity.** Mean repertoire value across the tested leaf cap settings, with EGRET's uncertainty interval shown by shading.

Response profiles across settings. Figure 25 gives the response profiles on a common vertical scale. The capacity and exploration curves remain fairly level around their shipped values, with lower scores toward the edges of their grids. The open square marks the unpooled hierarchy at the default pair capacity; its score remains close to the pooled configuration. The refit curve declines steadily as updates become less frequent and approaches the baseline envelope at its lowest setting. This shape makes the sensitivity pattern easy to distinguish: the model tolerates moderate changes in capacity and exploration, whereas the timing of posterior updates directly affects the quality of the returned repertoire.

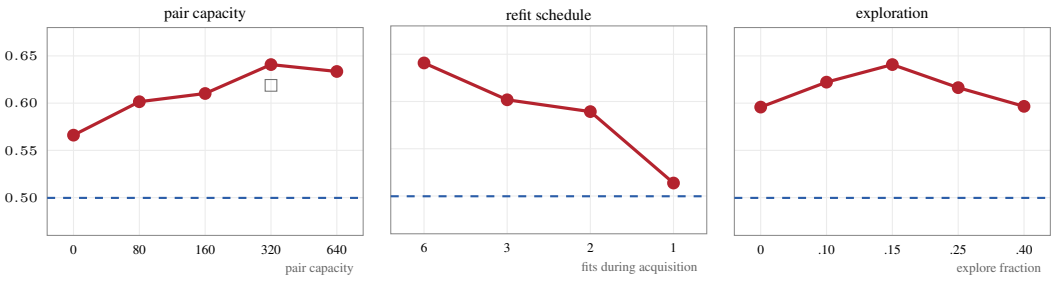


Figure 25: **Configuration response profiles.** Mean repertoire quality across the three hyperparameter sweeps, shown on a common scale. Each panel varies one setting with the remaining configuration fixed throughout the corresponding sweep.

Configuration sweep summary. Table 21 complements the response curves with paired comparisons at each tested setting. It organizes the evidence around surrogate capacity, posterior refitting, and exploration, using the shipped configuration as the common reference. The reported intervals and instance-level wins show which changes remain within local variation and which materially alter the measured advantage.

Surrogate capacity remains stable near the shipped value: doubling the pair budget changes Φ_R by -0.007, while unpooling changes it by -0.022. Exploration follows a peak around 0.15; the adjacent settings remain within uncertainty of the default, and both endpoints yield lower scores. Refit frequency has a larger effect. Updating every two rounds reduces Φ_R by 0.039, and fitting only at the seed and final rounds reduces it by 0.127. The latter schedule leaves an advantage of 0.014 with an interval that includes zero. The clearest sensitivity therefore lies in the timing of posterior updates, while nearby capacity and exploration settings produce smaller changes across the full configuration grid.

Table 21: **Configuration sweeps.** Repertoire quality across surrogate capacity, refit frequency, and exploration settings.

	Φ_R	Δ vs. shipped	vs. envelope		time
Setting	mean	default	Δ	wins	(s)
Surrogate capacity					
0 pairs	0.566	**−0.075** [−0.129, −0.026]	**+0.066** [+0.027, +0.111]	13/20	10.1
80	0.601	**−0.039** [−0.081, −0.007]	**+0.102** [+0.041, +0.175]	14/20	17.9
160	0.610	−0.031 [−0.087, +0.002]	**+0.110** [+0.042, +0.184]	16/20	18.2
shipped	0.641	–	**+0.141** [+0.079, +0.213]	17/20	18.6
640	0.634	−0.007 [−0.022, +0.000]	**+0.134** [+0.075, +0.200]	17/20	18.4
320, unpooled	0.619	−0.022 [−0.055, +0.008]	**+0.119** [+0.060, +0.187]	17/20	16.6
Refit schedule					
shipped	0.641	–	**+0.141** [+0.079, +0.212]	17/20	20.4
every 2 rounds	0.602	**−0.039** [−0.072, −0.008]	**+0.102** [+0.024, +0.187]	12/20	16.8
seed + midpoint	0.589	**−0.052** [−0.085, −0.019]	**+0.090** [+0.030, +0.159]	15/20	15.3
seed + final	0.514	**−0.127** [−0.215, −0.054]	+0.014 [−0.082, +0.106]	10/20	15.8
Exploration					
0.00	0.596	**−0.045** [−0.088, −0.009]	**+0.096** [+0.011, +0.177]	15/20	21.7
0.10	0.622	−0.019 [−0.057, +0.019]	**+0.122** [+0.051, +0.199]	17/20	23.2
shipped	0.641	–	**+0.141** [+0.079, +0.211]	17/20	22.1
0.25	0.616	−0.024 [−0.056, +0.005]	**+0.117** [+0.056, +0.186]	17/20	21.4
0.40	0.597	**−0.044** [−0.093, −0.001]	**+0.097** [+0.015, +0.185]	13/20	22.4

E.9 Failure Cases

This subsection examines boundary conditions revealed by controlled changes to initialization and upper-tail decisions. The analysis identifies where outcomes are most sensitive to these choices.

Concentrated initialization. Figure 26 resolves the clustered-seed intervention by benchmark instance. `AncSR2` decreases consistently by 0.342, and `Potts` increases by 0.054. Grey bars mark mixed seed directions; `AncSR1` averages -0.352 across outcomes from -0.833 to $+0.378$. The untouched control has an average five-seed spread of $0.08\ \Phi_R$, and subsequent acquisition restores full group coverage on nineteen of twenty instances. The pattern concentrates initialization sensitivity in a small part of the benchmark amid substantial run-level variation observed elsewhere.

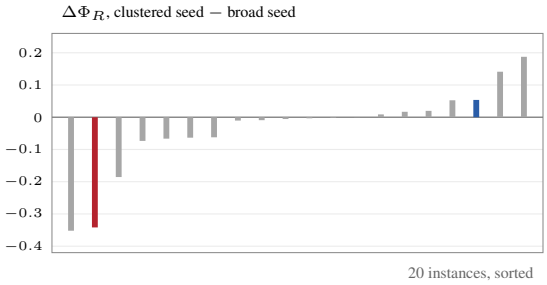


Figure 26: **Initialization sensitivity by instance.** Sorted paired effects of clustered and broad seeding, with consistent seed directions highlighted.

Aggregate intervention profile. Table 22 summarizes controlled changes to initialization and upper-tail processing. Concentrating every seed profile in one cluster changes Φ_R from 0.614 to 0.579, with a paired interval of $[-0.094, +0.019]$ and one consistent decline among twenty instances. Removing one cluster produces 0.618 and no consistent declines. An exhaustive condition screen preserves the aggregate score of 0.636. Suppressing partial-profile promotion lowers it to 0.550, corresponding to a paired change of $-0.086\,[-0.164, -0.017]$ and two consistent declines. These results locate the clearest sensitivity in the promotion step and show that the tested initialization changes are usually absorbed during acquisition.

Table 22: **Initialization and upper-tail controls.** Paired effects of seed composition, condition screening, and partial-profile promotion across the benchmark instances.

	Φ_R		Advantage	Per-instance		
Regime	mean	Δ vs. control	Δ vs. control	hurt	worst of those	
Initialization bias						
native seeding	0.614	–	–	–	–	–
one group seeded	0.579	−0.035 [−0.094, +0.019]	−0.015 [−0.034, +0.000]	1/20	`AncSR2`	−0.342
one group absent	0.618	+0.004 [−0.036, +0.045]	+0.001 [−0.010, +0.014]	0/20	–	–
Upper tail						
unmodified	0.636	–	–	–	–	–
screen every condition	0.636	0.000 [−0.039, +0.036]	+0.002 [−0.006, +0.012]	0/20	–	–
never promote a partial profile	0.550	**−0.086** [−0.164, −0.017]	**−0.011** [−0.023, −0.001]	2/20	`KRAS`	−0.582

F Further Discussion

F.1 Code Availability

Implementation and reproducibility. After acceptance, we will release the EGRET implementation, configuration files, benchmark-preparation scripts, and evaluation utilities in a public GitHub repository. The release will cover posterior inference, lineage search, adaptive assay allocation, and measured-value commitment. Appendix B gives the algorithmic details, while Appendix D describes the benchmark and evaluation protocols. Fixed dataset versions and random seeds will be included where redistribution is permitted. For restricted data, the repository will provide the expected input format and preparation steps needed to reconstruct the benchmark interface.

F.2 Social Impact

Potential benefits and safeguards. In benign protein-engineering settings, EGRET is intended to make limited assay budgets more informative while keeping repertoire decisions traceable. Its construction-aware representation links measured evidence to shared mutation paths and condition-specific branches, while biological review remains necessary. The main implications are:

- *Resource efficiency.* Shared construction steps may reduce material use and assay costs when several useful endpoints can be reached within the same experimental campaign.
- *Decision traceability.* Lineage records expose which measurements supported the returned repertoire and how shared mutations contributed to different condition-specific branches.
- *Responsible use.* Because mutation prioritization may also have dual-use implications, EGRET should remain a decision-support method subject to biosafety review, access controls, and human approval before synthesis or release.

F.3 Limitations

Scope and approximation conditions. EGRET is evaluated in a finite-library setting in which sequence measurements, assay conditions, and construction rules are specified before each campaign. The resulting certificates are therefore most directly interpreted within the tested decision setting and its associated sequence–condition distributions:

- *Model assumptions.* The Gaussian, hierarchical, site-coherent posterior favors interpretable effects but may be less expressive for stronger interactions, non-Gaussian responses, or assay-specific structure.
- *Data coverage.* Despite diverse libraries, biological contexts, and condition panels, finite datasets cover only part of the broader sequence–condition space.
- *Distributional generalization.* Learning and decision results reflect tested distributions; transfer may vary with shifts in library composition, measurement reliability, or resources.

F.4 Future Work

Cross-task and cross-environment studies. A broader evaluation of EGRET could span protein engineering, enzyme optimization, antibody and reagent design, and related sequence–function tasks. Community benchmarks with varied sequence spaces, condition panels, assay modalities, library structures, construction budgets, and laboratory settings would help assess how calibration and decision quality transfer across biological objectives.

Prospective laboratory deployment. Prospective campaigns should reflect the practical conditions under which assays and constructions are carried out. Relevant settings include noisy or unreliable measurements, limited assay budgets, batch effects, incomplete infrastructure, and constraints on throughput, personnel, and construction. Community-contributed datasets and prospective studies would help ground these evaluations in routine laboratory practice.

F.5 Ethical Statement

Intended scientific benefit. We developed EGRET in the hope that more efficient and traceable use of biological measurements will support research that benefits the broader life sciences community. Its intended applications are peaceful scientific inquiry, responsible biological engineering, and the development of knowledge and tools that improve human and environmental well-being. Any experimental deployment should remain subject to appropriate biosafety review, institutional oversight, and applicable regulation.

Responsibility for downstream use. The authors cannot monitor or control independent uses of the paper, software, or associated materials after their release. Users and their institutions are responsible for assessing risks, obtaining required approvals, and ensuring that any application complies with relevant ethical, legal, biosafety, and biosecurity requirements. The authors disclaim responsibility for applications that depart from the scientific purposes and safeguards stated in this work, including harmful, unlawful, or otherwise inappropriate uses.