EVOCIRCUIT: ACTIVE LEARNING FOR PROTEIN OPTIMIZATION VIA MECHANISTIC INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (pLMs) provide rich representations for protein engineering, but active optimization must adapt these representations to target-specific objectives from only a small number of experimental measurements. Existing protein optimization approaches improve candidate selection through activity prediction or search over learned fitness landscapes, but do not explicitly exploit task-relevant sparse representations within the underlying pLM. We introduce EvoCircuit, an active-learning framework informed by mechanistic interpretability that uses sparse features extracted from cross-layer transcoders of ESM-2. At each round, EvoCircuit identifies task-relevant features through associations between feature activations and experimental activity in measured variants. These features induce a task-adaptive similarity over candidate sequences, guiding batch acquisition alongside predictor-based ranking. Across retrospective active-learning simulations on single-mutant protein fitness landscapes, EvoCircuit improves the discovery of high-activity variants over predictor-based selection under matched assay budgets, with effects varying across landscapes and predictors. Ablations examine the contributions of task-adaptive feature selection and acquisition design, including comparisons with embedding-based and random-feature controls. Experiments on a combinatorial multi-mutant landscape extend the evaluation to predictor-based and reward-guided search baselines, revealing differences between hit discovery and peak-fitness optimization. Our results suggest that sparse internal representations of pretrained pLMs can complement activity prediction as inductive signals for data-efficient protein optimization.

1 INTRODUCTION

Protein engineering often requires optimization over large sequence spaces while experimental measurements remain costly and limited, making the selection of which variants to evaluate next a central challenge. Large pLMs learn representations that capture information about protein structure, function, and the effects of sequence variation (Alley et al., 2019; Rives et al., 2021; Meier et al., 2021). These representations have been used to train target-specific predictors and to guide iterative search over protein fitness landscapes (Biswas et al., 2021; Jiang et al., 2024; Yang et al., 2025). This raises the question of whether task-relevant information within these broad representations can be identified and used more directly for candidate selection.

Mechanistic interpretability provides tools for decomposing dense model representations into sparse, more localized features. In language models, sparse autoencoders have been used to extract interpretable features from internal activations, and feature-level interventions have shown in specific settings that manipulating selected features can produce corresponding changes in model outputs (Huben et al., 2024; Templeton et al., 2026). Transcoders extend this approach by learning sparse approximations of intermediate model computations, enabling feature-level analysis of how model outputs are produced (Dunefsky et al., 2024). More recently, cross-layer transcoders have been used to connect sparse features across layers and represent model computations as circuits associated with particular outputs (Ameisen et al., 2025).

These approaches have recently been extended to protein language models. Sparse features extracted from ESM-2 have been associated with structural motifs, functional domains, binding sites, and other biological properties (Simon & Zou, 2025), and have also been used for functional annotation

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.

of protein sequences (Silberg et al., 2025). Cross-layer circuit analysis has further identified compact sets of ESM-2 features associated with protein properties and fitness. Importantly, interventions on these circuits have been shown to guide protein design toward higher-fitness variants relative to the evaluated baselines (Tsui et al., 2026).

We introduce **EvoCircuit**, an active-learning framework that uses task-relevant sparse features within ESM-2 to guide candidate selection. At each round, EvoCircuit identifies features using two sources of task-specific evidence: their contribution to the current predictive surrogate and their association with activities observed among measured variants. We obtain these features from predictor-conditioned circuits extracted with cross-layer transcoders. The selected features define a task-adaptive similarity between candidate sequences, which is used to guide exploration alongside predictor-based ranking during batch selection. As new measurements are collected, both the activity predictor and the feature representation used for acquisition are updated.

Our contributions are as follows.

- We introduce **EvoCircuit**, an active-learning framework that selects sparse CLT features of ESM-2 using their associations with measured activity.
- We evaluate EvoCircuit on four single-mutant and three combinatorial fitness landscapes under matched assay budgets. The results identify predictor- and landscape-dependent gains in high-activity variant discovery, while distinguishing discovery yield from peak-fitness optimization.
- We use matched representation controls and fixed-state ablations to assess the contribution of activity-based feature selection. A feature-level case study further connects feature interpretation and removal to changes in specific candidate-acquisition decisions.

2 Related Work

2.1 Protein Language Models for Protein Engineering

Protein language models (pLMs) provide pretrained representations of protein sequences that can be transferred to downstream prediction and design tasks. Early sequence models such as UniRep learned protein representations through autoregressive sequence modeling (Alley et al., 2019), while ESM and ESM-2 use masked language modeling to learn contextual residue representations from large protein sequence collections (Rives et al., 2021; Lin et al., 2023b). Other large-scale models, including ProtTrans and ProGen, have explored related masked or autoregressive objectives for protein representation and generation (Elnaggar et al., 2022; Madani et al., 2023; Nijkamp et al., 2023). Depending on the model and task, downstream methods use residue-level hidden states or pooled sequence embeddings as representations, and model likelihoods as sequence-level scores. These representations have been used in several forms of protein engineering. pLM likelihoods can score mutation effects without task-specific measurements (Meier et al., 2021), while pretrained embeddings can be combined with experimental labels to predict quantitative properties such as activity or fitness. Low-N protein engineering, for example, combines pretrained representations with small labeled datasets and family-specific adaptation to support supervised protein optimization (Biswas et al., 2021). More broadly, pLMs can be kept fixed, adapted to a target family, or fine-tuned for downstream prediction. These studies establish pretrained sequence representations as a common basis for learning protein properties when experimental data are limited.

2.2 Active Protein Optimization

Protein optimization requires selecting new variants from measurements collected in previous rounds. EVOLVEpro represents a predictor-driven approach: it trains a regression model on pLM embeddings and measured activities, then selects variants with the highest predicted activity in subsequent rounds (Jiang et al., 2024). ALDE instead formulates selection as batch Bayesian optimization, using predictive uncertainty together with predicted fitness to balance exploration and exploitation (Yang et al., 2025). FolDE further considers the composition of the selected batch, combining a pLM-based prior with a constant-liar strategy that reduces redundancy among candidates (Roberts et al., 2025). Together, these methods represent complementary approaches based on predicted activity, predictive uncertainty, and diversity-aware batch selection. Other work has

explored the structure of the candidate space more explicitly. CLADE uses hierarchical clustering to sample from distinct regions of sequence space before supervised fitness prediction (Qiu et al., 2021). Similarity-aware approaches have also used kernel-based diversity criteria, including determinantal point processes, to select high-scoring but distinct candidates within a batch (Vornholt et al., 2024). Related model-guided approaches search local sequence neighborhoods or optimize learned fitness functions to propose new variants (Emami et al., 2023; Kirjner et al., 2024). These methods illustrate that exploration can be defined through several sources of information, including uncertainty, sequence-space structure, candidate similarity, and within-batch diversity.

2.3 Mechanistic Interpretability of Protein Language Models

Sparse feature methods have progressed from identifying features within individual model activations to analyzing computations within and across layers (Huben et al., 2024; Dunefsky et al., 2024; Ameisen et al., 2025). Sparse autoencoders primarily expose features represented at a particular activation, while transcoders and cross-layer transcoders support analysis of how sparse features contribute to intermediate and cross-layer computations. Recent work has applied these methods to protein language models. Sparse autoencoders trained on ESM-2 identify features associated with protein families and biological properties (Adams et al., 2025). InterPLM similarly finds sparse ESM-2 features corresponding to structural motifs, functional domains, binding sites, and other biological concepts, and uses these features for annotation and sequence steering (Simon & Zou, 2025). Sparse autoencoders have been applied to ESM-2 representations at both the protein and residue levels, while transcoders have been used to model sparse transformations between protein-level representations across layers (Gujral et al., 2025). Sparse pLM features have also been used directly for functional annotation of protein sequences (Silberg et al., 2025). Circuit-level work extends this analysis across layers. ProtoMech applies cross-layer transcoders to ESM-2 to extract compact circuits associated with protein prediction tasks and biological properties, and uses identified circuits to steer protein design toward higher-fitness variants (Tsui et al., 2026). Prior work therefore evaluates sparse representations mainly through biological interpretation, predictive relevance, and intervention. EvoCircuit instead evaluates whether task-selected sparse features are useful for sequential acquisition under a fixed experimental budget.

We propose **EvoCircuit**, a mechanistic-interpretability-guided active learning pipeline for protein engineering. Unlike conventional score-based acquisition methods that rely solely on predictor outputs, EvoCircuit incorporates sparse internal representations from a pretrained protein language model (pLM) to guide candidate selection. By combining predictor scores with task-relevant pLM features, EvoCircuit introduces information from the model's internal representations into the active learning loop. Appendix A provides an overview of the full pipeline.

3 EvoCircuit

3.1 Sparse Internal Representations of the pLM

Mechanistic interpretability provides access to internal model representations beyond the scalar outputs used for candidate ranking. In particular, cross-layer transcoders (CLTs) decompose intermediate pLM computations into sparse latent features and use them to reconstruct downstream MLP activations, providing an interpretable representation of computation across layers. EvoCircuit uses these sparse model-internal features as coordinates for active learning, allowing candidate selection to incorporate information from intermediate pLM computations; details of the CLT architecture and reconstruction objective are provided in Appendix B. From this representation, EvoCircuit identifies internal features whose activations are associated with experimentally measured activity and uses them to construct a task-adaptive representation for candidate selection. The activity predictor remains unchanged and is retained for candidate ranking. A CLT feature is indexed by the pair $g = (\ell, j)$ of a layer ℓ and latent index j; let $z_{g,p}(x)$ denote its activation at residue position p of sequence x. We define a sequence-level representation by averaging over residue positions,

$$\phi_g(x) = \frac{1}{L(x)} \sum_{p=1}^{L(x)} z_{g,p}(x),$$

excluding special tokens. These representations are computed from sequence using a frozen ESM-2 backbone and CLT, without candidate activity labels.

3.2 Task-Adaptive Feature Selection

Although CLT activations are sparse, the full latent dictionary contains hundreds of thousands of features, most of which are inactive or constant across the sequences measured in a single campaign. EvoCircuit therefore selects a compact feature set at each active-learning round using only measurements available to the campaign.

At round t, let $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{n_t}$ denote the measured sequences and their activities. Before evaluating feature–activity associations, we construct a candidate pool from CLT features whose sequence-level activations vary across the observed sequences. Candidate-pool construction uses the observed sequences x_i, but not their activities y_i, separating initial screening of available features from subsequent selection based on the downstream objective. From these varying features, we sample a fixed number uniformly within each layer using layer-balanced quotas to form $\mathcal{C}_t$. This prevents differences in the number of varying features across depth from determining the layer composition of the pool, while allowing task relevance to be assessed from observed activities. We sample uniformly rather than ranking features by a label-free criterion such as observed activation variance, which would be computed on sequences already chosen by the campaign and make the pool depend on the acquisition policy's history. The candidate pool is drawn independently at each round. Exact pool sizes, layer quotas, and corresponding ablations are reported in the Appendix.

We next identify features whose activations vary systematically with measured activity, computing $\rho_{t,g} = \mathrm{Spearman}(\{\phi_g(x_i)\}_{i=1}^{n_t}, \{y_i\}_{i=1}^{n_t})$ for each $g \in \mathcal{C}_t$.

We use rank correlation to capture monotonic feature–activity associations without fitting an additional parametric model over the sparse features. Because these associations can be unstable when only a small number of measurements are available, we additionally evaluate their consistency under bootstrap resampling. Let $\rho_{t,g}^{(r)}$ denote the correlation obtained from bootstrap replicate r, for $r = 1, \ldots, R_{\mathrm{boot}}$ resamples of $\mathcal{D}_t$ drawn with replacement. We define the sign-stability score

$$q_{t,g} = \frac{1}{R_{\mathrm{boot}}} \sum_{r=1}^{R_{\mathrm{boot}}} \mathbf{1}\left[\rho_{t,g}^{(r)} \neq 0 \ \wedge \ \mathrm{sign}\left(\rho_{t,g}^{(r)}\right) = \mathrm{sign}(\rho_{t,g})\right].$$

Features are retained when both association magnitude and bootstrap stability meet predefined thresholds, $\mathcal{E}_t = \{g \in \mathcal{C}_t : |\rho_{t,g}| \geq \tau_\rho, \ q_{t,g} \geq \tau_q\}$, and the panel keeps at most K features with the largest absolute correlations, $\mathcal{F}_t = \mathrm{Top}_K(\mathcal{E}_t; |\rho_{t,g}|)$. Both positively and negatively associated features are eligible. The resulting $\mathcal{F}_t$ is a compact, round-specific set of internal pLM coordinates used to guide candidate acquisition.

3.3 Feature-Guided Candidate Acquisition

The selected features specify which internal pLM coordinates are associated with the current objective, but do not by themselves determine which unmeasured sequences should be assayed next. EvoCircuit uses these coordinates to define a task-adaptive distance between candidate sequences and experimentally observed high-activity variants, and combines the resulting candidate neighborhoods with the host activity predictor.

The A highest-activity variants observed in $\mathcal{D}_t$ serve as anchors for candidate retrieval. For each selected feature g, its observed activation scale and the panel's median scale are $\sigma_{t,g} = \mathrm{SD}_{x_i \in \mathcal{D}_t}[\phi_g(x_i)]$ and $m_t = \mathrm{median}_{g \in \widetilde{\mathcal{F}}_t} \sigma_{t,g}$, where $\widetilde{\mathcal{F}}_t \subseteq \mathcal{F}_t$ excludes coordinates whose scale is numerically degenerate. We define the distance between a candidate x and an anchor a as

$$d_t(x, a) = \frac{1}{|\widetilde{\mathcal{F}}_t|} \sum_{g \in \widetilde{\mathcal{F}}_t} \frac{(\phi_g(x) - \phi_g(a))^2}{\sigma_{t,g}^2 + (\gamma m_t)^2}.$$

Under this distance, two sequences are close when they induce similar activation patterns along the internal coordinates currently associated with measured activity. Normalization by the observed

feature scales prevents coordinates with larger activation magnitudes from dominating retrieval, while the ridge term controlled by γ stabilizes distances for low-variance features.

Let $\mathcal{U}_t$ denote the set of candidates not requested in any previous round and $s_t(x)$ the score assigned by the host activity predictor. EvoCircuit preserves a fraction of each batch for direct predictor-based selection. For a batch size B and allocation parameter $\lambda \in [0, 1]$, define $b_{\text{global}} = \lfloor \lambda B \rfloor$ and $\mathcal{G}_t = \text{Top}_{b_{\text{global}}}(\mathcal{U}_t; s_t)$.

The remaining candidates are retrieved using the feature-space distance. For each anchor, candidates in $\mathcal{U}_t \setminus \mathcal{G}_t$ are ordered by increasing $d_t(x, a)$, and retrieval is cycled across anchors until a shortlist $\mathcal{S}_t$ of size M is formed. The host predictor then ranks candidates within this feature-guided shortlist, giving the final batch $\mathcal{B}_t = \mathcal{G}_t \cup \text{Top}_{B-b_{\text{global}}}(\mathcal{S}_t; s_t)$.

This construction uses feature-guided retrieval to complement rather than replace the activity predictor: one portion of the batch preserves globally high-scoring candidates, while the remainder is selected from regions identified through activity-associated internal pLM features. After the batch is measured, the new observations are added to give $\mathcal{D}_{t+1}$, and the predictor, candidate feature pool, selected feature set, and anchors are updated for the next round. All concrete acquisition budgets and hyperparameter settings are reported in the Appendix.

4 Experiments

4.1 Datasets

We evaluate on seven measured protein fitness landscapes. Because every variant is measured in advance, we replay directed-evolution campaigns *in silico* against recorded ground truth: a method requests a variant and the evaluator returns its stored measurement. Appendix D gives library sizes, thresholds and assay provenance.

Single-mutant deep mutational scans. Four assays, all taken from ProteinGym v1, cover a varied set of proteins and readouts, each with between $6{,}000$ and $8{,}600$ measured substitutions. **AmiE** is the aliphatic amidase of *P. aeruginosa*, scored by growth with isobutyramide as the sole nitrogen source (Wrenbeck et al., 2017). **CYP2C9** is human cytochrome P450 2C9, scored by an activity-based probe in humanized yeast (Amorosi et al., 2021); we use its activity map rather than the companion abundance map. **GCK** is human glucokinase, scored by complementation growth of a hexokinase-deficient yeast strain (Gersing et al., 2023). **RgDAO** is D-amino acid oxidase from *R. gracilis*, scored by enzyme proximity sequencing (Vanella et al., 2024). For all four the candidate pool is exactly the measured pool, so no request is ever unanswerable.

Combinatorial active-site libraries. Three libraries mutate a handful of active-site positions to all 20 amino acids, which makes epistasis rather than breadth the difficulty. **GB1** varies four positions of the B1 domain of protein G, read out as IgG-Fc binding enrichment normalized to the wild type (Wu et al., 2016). **TrpB3I** and **TrpB4** vary three and four positions of the β-subunit of tryptophan synthase from *T. maritima*, starting from the engineered parent Tm9D8* and read out as growth rate of a tryptophan auxotroph (Johnston et al., 2024); we follow the three-site library naming of Li et al. (2025). GB1 covers 93% of its $160{,}000$-variant space, TrpB3I 97% of $8{,}000$, and TrpB4 over 99% of $160{,}000$. The three differ sharply in difficulty: among measured variants of mutation depth ≥ 2, 2.43% of GB1 and 1.41% of TrpB3I exceed their parent, against 0.68% of TrpB4.

4.2 Metrics

Each campaign acquires 112 measurements in seven batches of 16 on top of a shared initial set (§4.3). We score a campaign by how many desirable variants it finds within that budget, and how early.

Discovery counts. A single-mutant discovery has activity at or above the dataset's 90th percentile. A combinatorial discovery has finite fitness strictly above the parent at mutation depth ≥ 2, excluding variants differing from the parent at only one of the mutated positions. We also count discoveries

against measured percentiles—top-5% and top-1% for the single-mutant datasets, and top-10%, top-5%, and top-1% for the combinatorial landscapes, again restricted to depth-≥ 2 variants. Thresholds are computed once per landscape from the measured pool and held fixed. They are not nested: depending on the landscape, a variant can exceed the parent yet fall outside the top decile, or sit in the top 1% yet fall below the parent, so the two counts need not move together.

Area under the discovery curve. Final counts alone do not distinguish a method that finds its hits in the first batch from one that finds them in the last. *Discovery AUC* averages cumulative discoveries over the budget,

$$A_H = \frac{1}{112} \sum_{r=1}^{7} \frac{H_{r-1} + H_r}{2} 16,$$

where H_r is the number of new discoveries after batch r and $H_0 = 0$; it is a count, not a proportion. *Best-fitness AUC* substitutes best-so-far fitness, includes the initial measurements, and is range-normalized per landscape with evaluator-only bounds, since fitness scales are not comparable across landscapes. It was the primary endpoint of the original single-mutant and TrpB4 protocols, and we report it alongside the discovery endpoints—six per single-mutant contrast and ten per combinatorial contrast (Appendix F).

4.3 Experimental setup

Budget and initialization. Every campaign shares 64 initial measurements with its controls, then makes seven batches of 16, for 112 new and 176 total. Predictors are refit each round on that configuration's own observations. All configurations sharing a seed start from an identical initial set, so every contrast is a within-seed difference. Single-mutant experiments use five paired seeds per dataset in each of two initialization cohorts, one unscreened and one conditioned on initial quality, reported separately; each combinatorial landscape uses twenty paired seeds starting from the parent plus 63 sampled variants. Requests falling outside the measured pool consume budget without replacement. Additional detail is in Appendix E.

Baselines. We factor each method into a predictor that scores variants and a batch rule that selects the next 16, and evaluate EvoCircuit by swapping only the rule. Greedy ensemble-mean top-16 is the common control, and each published method's native rule is retained as a separate baseline. The single-mutant datasets use three predictors: a random forest on ESM-2 650M embeddings, which is the EVOLVEpro predictor (Jiang et al., 2024); TabPFN (Hollmann et al., 2025) on the same embeddings; and FolDE (Roberts et al., 2025). The two ESM-2 predictors each run under four rules—greedy, EvoCircuit, an ESM-distance retrieval control, and a count- and layer-matched random CLT panel—and FolDE under its native rule, greedy and EvoCircuit, giving thirteen configurations per dataset and cohort together with random selection and ESM-2 zero-shot scoring (Lin et al., 2023a). The combinatorial landscapes reuse these and add the ALDE DNN and gradient-boosted ensembles on one-hot inputs (Yang et al., 2025), each under Thompson sampling, greedy and EvoCircuit, and a one-hot random forest.

5 Results

5.1 Single-mutant discovery

Table 1 reports the four single-mutant datasets under unscreened initialization. Adding EvoCircuit improves both predictors on every discovery endpoint except one. New top-10% discoveries rise from 26.15 to 29.15 for the random forest, which is the EVOLVEpro predictor, and from 31.65 to 33.15 for TabPFN. Top-5% discoveries rise from 13.60 to 15.15 and from 16.95 to 19.30, and discovery AUC from 13.28 to 14.71 and from 15.87 to 17.20, so the discoveries also arrive earlier in the budget. The exception is the random forest's top-1% count, which does not move. Neither do the best-fitness endpoints, on a different scale, for either predictor, so the effect is specific to discovery yield rather than to the single best variant found. Dataset means also disagree, with the random forest declining on RgDAO and TabPFN on CYP2C9. With five seeds per dataset we read these as a shared direction rather than as individually resolved effects.

Improving a fixed predictor does not make the resulting configuration the strongest. FolDE's ESM-2 ranker reaches 41.20 discoveries under greedy selection and is unchanged by EvoCircuit, while ESM-2 zero-shot scoring, which uses no assay data, reaches 74.6 on CYP2C9 — the one dataset whose discovery threshold sits at the neutral level: wild-type activity is at the 88.4th percentile, and the top-10% cutoff is 1.019 on a scale topping out at 1.305, so a discovery there is a substitution that leaves the enzyme intact rather than improves it. The same ranking reaches 24.6, 26.8, and 18.8 on the other three, where the top decile lies well above wild type. These comparisons show a tendency rather than a ranking: in most settings tested, guiding acquisition with sparse features from inside ESM-2 finds more discoveries than ranking by predicted activity alone.

Table 1: Single-mutant discovery under unscreened initialization (mean ± SD, five paired seeds). New top-10% discoveries and AUC cover 112 additional requests after 64 shared initial measurements. AUC is the mean cumulative discovery count (higher is better). Summary columns average the four dataset means. **Blue bold** marks a higher EvoCircuit mean than greedy selection with the same predictor, not statistical significance.

		New top-10% discoveries				Mean over datasets	
Predictor	Acquisition	AmiE	CYP2C9	GCK	RgDAO	Count	AUC
No assay fitting	Random	13.8 ± 1.3	9.0 ± 3.5	9.8 ± 2.3	10.2 ± 3.6	10.70	5.16
	ESM-2 zero-shot	24.6 ± 0.9	74.6 ± 2.2	26.8 ± 1.5	18.8 ± 1.3	36.20	19.46
RF	EVOLVEpro	30.0 ± 10.7	28.4 ± 12.3	21.2 ± 4.7	25.0 ± 4.7	26.15	13.28
	+ EvoCircuit	**35.4 ± 5.0**	**33.6 ± 14.6**	**24.0 ± 1.6**	23.6 ± 3.3	**29.15**	**14.71**
TabPFN	EVOLVEpro	36.6 ± 10.5	43.6 ± 12.8	24.2 ± 8.4	22.2 ± 5.1	31.65	15.87
	+ EvoCircuit	**37.4 ± 9.1**	42.4 ± 17.4	**28.4 ± 10.2**	**24.4 ± 6.3**	**33.15**	**17.20**
FolDE	Constant liar	44.0 ± 15.9	61.6 ± 5.2	26.0 ± 3.5	28.2 ± 3.3	39.95	20.03
	Greedy	43.2 ± 12.5	65.6 ± 0.5	28.6 ± 4.6	27.4 ± 5.7	41.20	21.30
	+ EvoCircuit	**45.8 ± 8.4**	62.4 ± 3.6	**32.2 ± 9.4**	24.0 ± 3.2	41.10	21.18

5.2 Combinatorial landscapes and predictor transfer

We extend the comparison to three combinatorial landscapes, reporting discovery AUC alongside the final counts so that the rate of discovery is visible as well as its total; the full table is in the appendix (Table H.1). The clearest gains are on GB1 with the two one-hot ensembles: ALDE's DNN raises above-parent discoveries from 70.95 to 80.90 and discovery AUC from 30.17 to 34.57, and its XGBoost counterpart from 63.85 to 71.80 and from 25.08 to 28.29. For ALDE the upper tail moves with the yield, with top-1% discoveries rising from 54.60 to 65.95. The XGBoost ensemble carries the effect to TrpB4, where above-parent discoveries rise from 43.90 to 51.45 and discovery AUC from 16.45 to 20.98. The picture is landscape-dependent rather than uniformly positive. GB1 is where the method works broadly: every predictor improves its top-1% count there, four of the five improve above-parent discoveries, and the two one-hot ensembles account for the largest gains. TrpB4 is narrower but carries the strongest rate improvement, the XGBoost ensemble's discovery AUC. TrpB3I gains nothing for any predictor, and FolDE's discovery AUC there falls from 27.79 to 26.40, the one comparison that moves clearly in the wrong direction. Across the fifteen comparisons the two predictors that regress on mean-pooled ESM-2 embeddings respond least, and TabPFN is the only one whose GB1 yield does not rise even as its top-1% count does.

Two things separate the predictors that gain from those that do not. The gain is larger where the matched greedy control is weaker: on GB1 the weakest baselines give the largest gains while FolDE and TabPFN, the two strongest, change by less than two discoveries, and the same ordering holds on TrpB4. It also splits by predictor input, since the predictors that gain most read one-hot encodings while the two that gain least regress on the mean-pooled ESM-2 embeddings from which EvoCircuit's own features are computed. Section 5.3 tests that second reading directly. TrpB3I fits neither account, and we leave it as an exception. Finding more variants above a threshold is not the same as reaching a higher maximum. On GB1, ALDE's mean final best fitness falls from 6.732 to 6.076 as its discovery count rises, and the XGBoost ensemble falls from 7.008 to 6.245. Best-fitness AUC, the primary endpoint of the original protocols, does not improve on any landscape or predictor. The two objectives separate here, and the discovery endpoints are the ones EvoCircuit addresses.

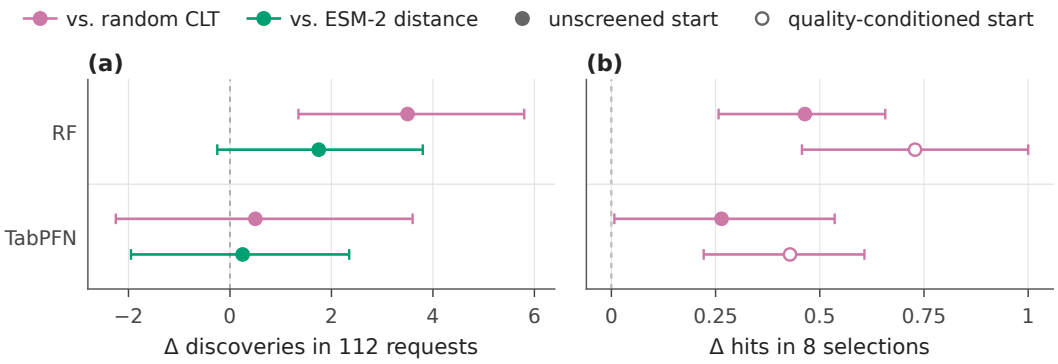


Figure 1: Feature-selection and representation comparisons on four single-mutant datasets. (a) Full campaigns with unscreened starts, comparing activity-selected CLT features with matched random CLT features or ESM-2 embedding distance (five seeds per dataset). Each method updates using its own observations. (b) Single-batch comparisons of activity-selected and matched random CLT features at saved EvoCircuit states. The observed history, predictor, anchors, and budget are fixed. Each point averages seven rounds within one dataset, initialization condition, and predictor. Points show mean paired differences (activity-selected CLT minus control), with pointwise 95% bootstrap intervals. Panel (a) counts new top-10% discoveries over 112 requests. Panel (b) counts hits among eight feature-guided selections, not cumulative gains.

5.3 Contribution of sparse feature selection

Sections 5.1 and 5.2 show that adding EvoCircuit finds more discoveries without saying what produces the difference. We remove one part of the method at a time, mainly on the single-mutant scans where every control is available (Figure 1). Choosing features by their association with measured activity is what matters. Holding the basis at the CLT features and swapping only the selection rule for a matched random set, the random forest reaches 29.15 new top-10% discoveries against 25.65 (Figure 1a); TabPFN moves in the same direction, but the complete campaigns do not resolve it. Those campaigns diverge after the first round, so the difference could also come from the paths they took. We therefore rewind to saved rounds and change nothing but the features, leaving the measurements, the predictor, the anchors and the budget as they were. The activity-selected set still finds more hits in that batch, for both predictors and both initializations (Figure 1b), and whichever anchor, shortlist and batch rule we use. Because this is a single batch, it locates the source of the difference without estimating a cumulative gain.

The features are not redundant with the predictor's own representation. The same random forest on one-hot encodings, with no language model in it at all, rises from 61.75 to 73.85 above-parent discoveries on GB1 and from 48.70 to 55.90 on TrpB4, where its ESM-2 counterpart barely moves; on the single-mutant scans the ESM-2 predictors are the ones that gain, so this holds of these libraries rather than in general. Whether the coordinates must come from the CLT we cannot settle: distance over the 1,280 ESM-2 embedding coordinates reaches 27.40 against 29.15 with the random forest, an interval spanning zero (Figure 1a). The sparse basis matches dense retrieval using at most eight coordinates, but matching is not exceeding.

What the selected features add is that each is a single CLT feature with a layer and an index, so it can be described, checked against saved activation profiles, and removed to see whether the selection changes. Section 5.4 carries out all three on one stored trajectory. A dense embedding supports no comparable operation.

5.4 Feature interpretation and candidate selection

Section 5.3 showed that the sparse basis changes which candidates are acquired. Because each coordinate is a single CLT feature rather than a direction in a dense embedding, the same choice leaves the acquisition open to inspection. We describe each feature from the reference proteins it activates on, following InterPLM (Simon & Zou, 2025), after the campaign and without showing the

model any GCK or assay data. This section reads two consecutive rounds of one saved trajectory, chosen after their outcomes were known, so it describes three requests rather than an effect; the procedure, a third acquisition and the full record are in Appendix K.

What the coordinate responds to. `L14_F09977`, selected at rounds 4 and 5, is active at three of GCK's 465 positions — F74, C220 and E221 — and its values over the 8,570 library variants fall into two groups with no variant in between, the smaller holding 614. In every saved profile the lower value is the one that has lost the activation at C220, so the coordinate tracks that context rather than a residue: substitutions within ten residues of C220–E221 flip it six times more often than distant ones, yet replacing C220 itself leaves it on in 16 of 18 cases. Its description calls it a feature of cyclic-nucleotide-binding-domain contexts, a family GCK does not belong to; what transfers is the local claim, a peak at E, K or A, which lands on E221. All three sites sit in strands that carry nucleotide contacts: C220 and E221 open the strand 220–237, which holds the ATP contact T228 and the substrate contact N231, and F74 lies in the strand holding the ATP contacts at 78–83.

What it did in the campaign. At round 4 the switched state is shared by four of the sixteen measured high-activity anchors and by D217A, which entered the shortlist through Q219M and was requested although the random forest ranked it 103rd of the 8,506 eligible candidates. It returned an activity of 2.151, a top-10% discovery. That measurement makes D217A an anchor at round 5, where it admits D217T, ranked 95th, returning 2.675 and again a discovery, and S411A, ranked 42nd and closer still, returning 0.407 and not. Dropping the feature while holding the measurements and the predictor's ranking fixed removes D217A at round 4, where the eight requests taken from the shortlist fall from three discoveries to one, and both round-5 candidates, where they fall from five to three. The window the switch marks is where this library's activating substitutions are: switched variants inside it are discoveries in 41 of 123 cases, against 10.0% across the library and 2 of 167 at the annotated substrate contacts.

What it does not show. The state is not activity. S411A carries it from 190 residues away, sits closer to the anchor than D217T does, and is not a discovery; it substitutes inside the annotated ATP-binding segment 411–415, so the annotation accounts for that miss and the feature does not. Neither is it a mechanism: we could not place the coordinate in an attribution graph over the CLT features at all, and those graphs track the random forest's ranking too weakly to pass a check we fixed in advance. And none of it is new evidence about GCK, since every variant named here was already measured in the library the trajectory ran on. What the case establishes is the other half of what the sparse basis buys: the request itself can be read, while the predictor's ranking and the model computation behind the features stay closed.

6 Conclusion

We introduced **EvoCircuit**, an active-learning framework in which sparse cross-layer-transcoder features of ESM-2 determine which experiment is performed next. At each round the method selects features whose activations track measured activity and uses them to retrieve candidates near the highest-activity variants observed so far, alongside predictor-based ranking. Because it modifies only the batch rule, it applies to existing methods without replacing their predictors; we evaluated it by swapping the rule across five predictors on four single-mutant scans and three combinatorial libraries. EvoCircuit improves discovery endpoints in most settings evaluated, most clearly where the matched greedy control is weaker and where the predictor does not regress on the embeddings our features are computed from, although the effect is landscape-dependent and absent on TrpB3I. Ablations attribute this to selecting features by their association with measured activity rather than to the sparse basis alone. Because each coordinate is a named feature, the resulting acquisition can also be described from reference proteins, checked against the target's own activations and annotation, and removed to observe the change in the request. These findings indicate that the sparse internal representations of a pretrained pLM can serve as an inductive signal complementary to activity prediction, and that mechanistic interpretability can contribute to a protein optimization campaign while it is running rather than only after it has concluded.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and editing, including drafting portions of the paper and improving English grammar and readability. We also used generative AI tools for literature retrieval and discovery; all references and relevant claims identified with AI assistance were manually verified against the original sources.

For research execution, code agents were used to implement author-specified methods, automate parts of the experimental pipeline, and assist with the implementation of demonstrations. The research questions, methodological specifications, experimental design, and interpretation of experimental results were developed by the authors rather than generated by AI tools. All AI-assisted code, text, and other generated artifacts were reviewed and, where applicable, tested and validated by the authors before inclusion or use.

We did not use generative AI tools to generate synthetic datasets or to formulate or prove mathematical claims. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Clara J. Amorosi, Melissa A. Chiasson, Matthew G. McDonald, Lai Hong Wong, Katherine A. Sitko, Gabriel Boyle, John P. Kowalski, Allan E. Rettie, Douglas M. Fowler, and Maitreya J. Dunham. Massively parallel characterization of cyp2c9 variant enzyme activity and abundance. *The American Journal of Human Genetics*, 108(9):1735–1751, 2021. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2021.07.001. URL https://www.sciencedirect.com/science/article/pii/S000292972100269X.

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.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, $44(10_{Part_2})$: 7112–7127, $October 2022. ISSN 0162-8828. doi: . URL https://doi.org/10.1109/TPAMI.2021.3095381.

Patrick Emami, Aidan Perreault, Jeffrey Law, David Biagioni, and Peter St. John. Plug amp; play directed evolution of proteins with gradient-based discrete mcmc. *Machine Learning: Science and Technology*, 4(2):025014, April 2023. ISSN 2632-2153. 10.1088/2632-2153/accacd. URL http://dx.doi.org/10.1088/2632-2153/accacd.

Sarah Gersing, Matteo Cagiada, Marinella Gebbia, Anette P. Gjesing, Atina G. Coté, Gireesh Seesankar, Roujia Li, Daniel Tabet, Jochen Weile, Amelie Stein, Anna L. Gloyn, Torben Hansen, Frederick P. Roth, Kresten Lindorff-Larsen, and Rasmus Hartmann-Petersen. A comprehensive map of human glucokinase variant activity. *Genome Biology*, 24(1):97, 2023. 10.1186/s13059-023-02935-8. URL https://doi.org/10.1186/s13059-023-02935-8.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025. 10.1073/pnas.2506316122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2506316122.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R Sgrizzi, Lukas Villiger, Alisan Kayabolen, BJ Kim, Josephine K Carscadden, Masahiro Hiraizumi, Hiroshi Nishimasu, et al. Rapid in silico directed evolution by a protein language model with evolvepro. *Science*, 387(6732): eadr6006, 2024.

Kadina E. Johnston, Patrick J. Almhjell, Ella J. Watkins-Dulaney, Grace Liu, Nicholas J. Porter, Jason Yang, and Frances H. Arnold. A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences*, 121(32):e2400439121, 2024. 10.1073/pnas.2400439121. URL https://doi.org/10.1073/pnas.2400439121. doi: 10.1073/pnas.2400439121.

Andrew Kirjner, Jason Yim, Raman Samusevich, Shahar Bracha, Tommi Jaakkola, Regina Barzilay, and Ila Fiete. Improving protein optimization with smoothed fitness landscapes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 46568–46586, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cbb7a23649b25001e797a726cf75498e-Paper-Conference.pdf.

Francesca-Zhoufan Li, Jason Yang, Kadina E. Johnston, Emre Gürsoy, Yisong Yue, and Frances H. Arnold. Evaluation of machine learning-assisted directed evolution across diverse combinatorial landscapes. *Cell Systems*, 16(9), 2025. 10.1016/j.cels.2025.101387. URL https://doi.org/10.1016/j.cels.2025.101387. doi: 10.1016/j.cels.2025.101387.

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

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

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Vijay Naik. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41:1099 – 1106, 2023. URL https://api.semanticscholar.org/CorpusID:256304602.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. 10.1016/j.cels.2023.10.002. URL https://doi.org/10.1016/j.cels.2023.10.002. doi: 10.1016/j.cels.2023.10.002.

Yuchi Qiu, Jian Hu, and Guo-Wei Wei. Cluster learning-assisted directed evolution. *Nature Computational Science*, 1:809–818, 12 2021. 10.1038/s43588-021-00168-y.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Jacob B. Roberts, Catherine R. Ji, Isaac Donnell, Thomas D. Young, Allison N. Pearson, Graham A. Hudson, Leah S. Keiser, Mia Wesselkamper, Peter H. Winegar, Janik Ludwig, Sarah H. Klass, Isha V. Sheth, Ezechinyere C. Ukabiala, Maria C. T. Astolfi, Benjamin Eysenbach, and Jay D. Keasling. Low-n protein activity optimization with folde, 2025. URL https://arxiv.org/abs/2510.24053.

Jacob Silberg, Elana Simon, and James Zou. Towards functional annotation with latent protein language model features. In David A Knowles and Peter K Koo (eds.), *Proceedings of the 20th Machine Learning in Computational Biology meeting*, volume 311 of *Proceedings of Machine Learning Research*, pp. 280–293. PMLR, 10–11 Sep 2025. URL https://proceedings.mlr.press/v311/silberg25a.html.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Darin Tsui, Kunal Talreja, Daniel Saeedi, and Amirali Aghazadeh. Protein circuit tracing via cross-layer transcoders. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Ki0eWT1sFS.

Rosario Vanella, Christoph Küng, Alexandre A. Schoepfer, Vanni Doffini, Jin Ren, and Michael A. Nash. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. *Nature Communications*, 15(1):1807, 2024. 10.1038/s41467-024-45630-3. URL https://doi.org/10.1038/s41467-024-45630-3.

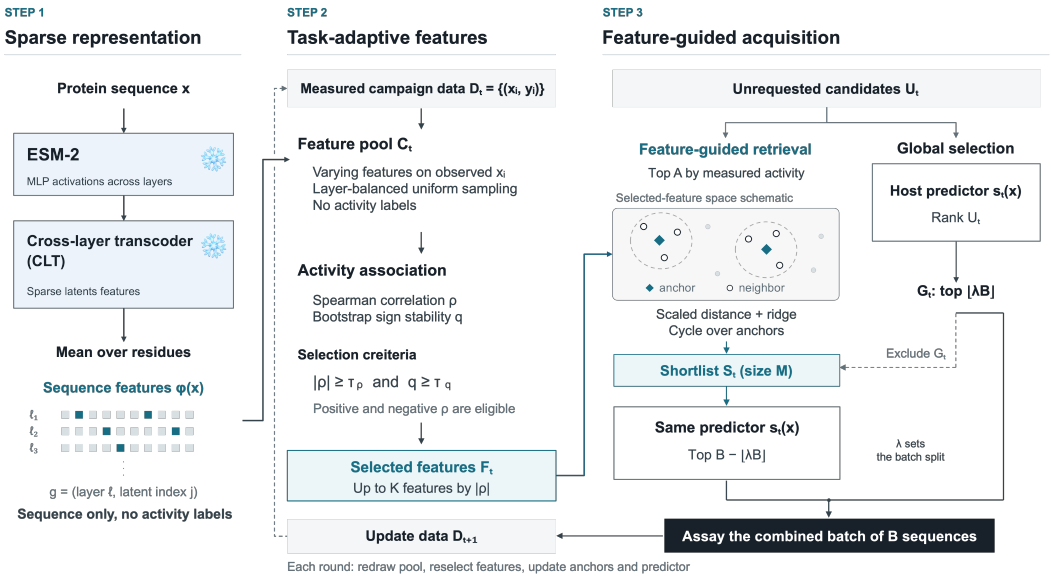
Tobias Vornholt, Mojmir Mutny, Gregor W. Schmidt, Christian Schellhaas, Ryo Tachibana, Sven Panke, Thomas R. Ward, Andreas Krause, and Markus Jeschek. Enhanced sequence-activity mapping and evolution of artificial metalloenzymes by active learning. *ACS Central Science*, 10(7):1357–1370, 2024. 10.1021/acscentsci.4c00258. URL https://doi.org/10.1021/acscentsci.4c00258.

Emily E. Wrenbeck, Laura R. Azouz, and Timothy A. Whitehead. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications*, 8(1):15695, 2017. 10.1038/ncomms15695. URL https://doi.org/10.1038/ncomms15695.

Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, Ren Sun, and Richard A Neher. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016. 10.7554/eLife.16965. URL https://doi.org/10.7554/eLife.16965.

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.

A Overview of the EvoCircuit



B Cross-layer transcoder and sparse feature extraction

B.1 Architecture and reconstruction objective

The backbone is ESM-2 t33 650M UR50D, frozen throughout. The cross-layer transcoder (CLT) has one encoder per transformer layer, 33 layers of $12{,}800$ latent features each, giving a dictionary of $422{,}400$ features over a $1{,}280$-dimensional residue stream. Each encoder receives the normalized MLP input after the attention residual addition; its own input normalization is the identity. The encoder carries learned b_ℓ^{pre} and b_ℓ^{enc} vectors even though the linear module has no additional bias parameter.

Let $m_{\ell,p}$ denote the native MLP output at layer ℓ and position p, before the residual addition. Using one-based layer indices, the decoder reconstructs it from the current layer and the three preceding ones,

$$\widehat{m}_{\ell,p} = b_\ell^{\mathrm{pre}} + \sum_{r=\max(1,\ell-3)}^{\ell} D_{r\to\ell}\, z_{r,p}, \tag{1}$$

so a source feature contributes to reconstructing its own layer and the following three MLP outputs. Decoder feature directions are unit-normalized. Training optimizes

$$\mathcal{L}_{\mathrm{CLT}} = \frac{1}{33}\sum_{\ell=1}^{33}\left[\frac{\mathrm{MSE}(\widehat{m}_\ell, m_\ell)}{\mathrm{Var}_{\mathrm{all}}(m_\ell) + 10^{-8}} + 0.03125\,\mathcal{L}_\ell^{\mathrm{AuxK}}\right], \tag{2}$$

where the mean squared error and the population variance are taken over all valid token and hidden-coordinate entries in the layer. The AuxK term is an unnormalized squared error against the detached reconstruction residual, using up to 640 dead latents with a dead-feature threshold of 500 forward passes; it is zero for layers without an applicable auxiliary reconstruction. The configured loss uses hard TopK sparsity and no ℓ_1 penalty. Pretraining reconstructs model activations and has no fitness-label objective.

C EvoCircuit implementation details

C.1 Candidate pool, layer quotas, and numerical conventions

The 2,048-feature pool allocates 63 features to each of the first two layers and 62 to each remaining layer. Only features with positive observed population standard deviation are eligible. A layer with insufficient eligible features contributes all of them, and the remaining budget is sampled from the unused eligible inventory; the implementation requires at least 2,048 variable features overall. The pool is refreshed from the current trajectory, seed and round, so it is drawn independently at each round.

C.2 Representation controls and position-resolved selection

The random-CLT control samples from its own observed-variable 2,048-feature pool, matching the selected panel's feature count and per-layer counts; overlap with the selected panel is allowed. The ESM-distance control instead uses all 1,280 final-layer residue-mean embedding coordinates with the same anchor and acquisition construction. It is distinct from the ESM-2 zero-shot sequence-scoring baseline, which ranks variants by likelihood and performs no retrieval.

Table C.1: Acquisition settings used for all reported EvoCircuit arms.

Quantity	Value	Role
CLT layers / width	33/12,800	Frozen feature inventory
Token activation budget	24	Sparse encoder
Candidate feature budget	2,048	Per-round sampling
Selected feature budget	At most 8	Acquisition coordinates
Association / stability thresholds	0.2/0.8	Feature screening
Bootstrap replicates	200	Sign stability
Observed anchors	16	Neighborhood construction
Distance ridge factor	0.25	Scale stabilization
Shortlist size	128	Candidate retrieval
Batch allocation	$8 + 8$	Global / shortlist score ranking

D Datasets

Table D.1 and Table D.2 give the library sizes, thresholds and assay provenance for the seven landscapes. Thresholds are computed once per landscape from its measured pool and held fixed across every method, seed and round; they are empirical percentiles of the measurements, not of the legal sequence space. For the four single-mutant scans the candidate pool is exactly the measured pool, so no request is unanswerable. For the three combinatorial libraries a request for an unmeasured combination consumes budget without replacement and is not a success.

CYP2C9 is the one dataset whose discovery threshold sits at the neutral level: its activity is expressed relative to wild type, the top-10% cut falls at 1.019, and the scale tops out at 1.305, so a discovery there is a substitution that leaves the enzyme intact rather than one that improves it. On the other three the top decile lies well above wild type.

E Campaign protocol

E.1 Budget and initialization

Every campaign shares 64 initial measurements with its controls and then makes seven batches of 16, for 112 new and 176 total requests. All configurations sharing a seed start from an identical initial set, so every contrast is a within-seed difference. Predictors and feature selection use each arm's own observations, refit each round.

Table D.1: Single-mutant deep mutational scans. All four assays are taken from ProteinGym v1 and the candidate pool is exactly the measured pool, so no request is unanswerable. Activity is the normalized ProteinGym `DMS_score`; higher is better. Thresholds are empirical percentiles of the measured pool, computed once per landscape and held fixed across all methods and seeds. The top-10% cut defines a discovery.

				Library		Threshold			
Dataset	Protein	Readout	Source	Variants	Length	Top-10%	Top-5%	Top-1%	Max
AmiE	*P. aeruginosa* aliphatic amidase	Growth on isobutyramide	Wrenbeck et al. (2017)	6,227	346	0.1090	0.2031	0.4215	0.8438
CYP2C9	Human cytochrome P450 2C9	Activity-based probe	Amorosi et al. (2021)	6,142	490	1.019	1.081	1.188	1.305
GCK	Human glucokinase	Yeast complementation growth	Gersing et al. (2023)	8,570	465	1.589	2.019	3.049	7.068
RgDAO	*R. gracilis* D-amino acid oxidase	Enzyme proximity sequencing	Vanella et al. (2024)	6,396	364	0.02206	0.06205	0.1477	0.4449

Table D.2: Combinatorial active-site libraries: GB1 (Wu et al., 2016), TrpB3I and TrpB4 (Johnston et al., 2024). Each library mutates the listed positions to all 20 amino acids, giving a legal space of GB1 160,000; TrpB3I 8,000; TrpB4 160,000. Measured reports the covered fraction of that space; requests for unmeasured combinations consume budget without replacement. Depth-≥ 2 counts exclude the parent and every variant differing from it at only one mutated position, and both the above-parent count and the percentile thresholds are computed on that restricted set. The two criteria are not nested. The TrpB parent Tm9D8* is an engineered starting point, not a natural wild type.

		Parent		Library			Threshold		
Landscape	Sites	Combo	Fitness	Measured	Depth ≥ 2	Above parent	Top-10%	Top-1%	Max
GB1	39, 40, 41, 54	`VDGV`	1.000	149,361 (93.4%)	149,284	3,623 (2.43%)	0.04606	2.150	8.762
TrpB3I	182, 183, 184	`YVF`	0.5252	7,784 (97.3%)	7,726	109 (1.41%)	0.2081	0.5649	0.9997
TrpB4	183, 184, 227, 228	`VFVS`	0.4081	159,129 (99.5%)	159,052	1,083 (0.68%)	0.04250	0.3491	1.000

Unscreened single-mutant starts comprise 16 uniformly sampled variants followed by 48 requests from a shared random-forest warmup; they are *not* a uniform random draw of 64. Quality-conditioned starts additionally meet historical initial-quality thresholds. Conditioning uses initial measurements, and its screening cost lies outside the 112-request continuation budget, so the two cohorts are reported separately and never pooled. Combinatorial starts comprise the parent plus 63 uniformly sampled distinct variants.

E.2 Seeds and cohort policy

Single-mutant experiments use five paired seeds per dataset in each of the two initialization cohorts: four datasets $\times$ two cohorts $\times$ five seeds gives 40 unique initial states and 520 complete trajectories across the 13 configurations. Each combinatorial landscape uses 20 paired seeds, giving 60 unique initial states and 1,200 reported trajectories.

Two cohort decisions are worth stating plainly because they limit confirmatory interpretation. TrpB4's 20 seeds are a first block of 10 plus a 10-seed extension decided after the first block's results were seen. TrpB3I's first run (seed 43001) is excluded from every method and endpoint, and the reported seeds are 43002–43021; the full 1,217-trajectory record including that run is retained for historical analysis only. Added controls reuse these same initializations, so all comparisons here are retrospective, and the stage sensitivities in Appendix J report what the staging costs.

E.3 Predictors and baselines

We factor each method into a predictor that scores variants and a batch rule that selects the next 16, and evaluate EvoCircuit by swapping only the rule. Greedy ensemble-mean top-16 is the common control; each published method's native rule is retained as a separate baseline. Table E.1 lists the configurations as they were actually run.

ALDE uses one-hot predictor inputs: a 20-dimensional indicator per mutable site, concatenated across sites. Its Thompson-sampling rule draws one of the five ensemble members independently for each of the 16 batch slots; its greedy rule ranks the ensemble mean. FolDE's ESM-2 adaptation replaces the native ESMC backbone and continues from the shared initial 64 requests in place of its native zero-shot initial batch; the ESMC-300M comparison is in Appendix H. Single-mutant FolDE arms share only seed-specific label-free pretrained weights, and all subsequent fits use each arm's own observed labels. When a one-hot predictor is paired with EvoCircuit, the acquisition still reads ESM-2/CLT features; only the scoring model is one-hot.

E.4 Missing measurements

A request for a variant outside the measured pool consumes budget, is not replaced, is never imputed as zero, and is not counted as a success. This applies only to the combinatorial landscapes; the single-mutant candidate pools are exactly their measured pools.

F Endpoints and statistical procedure

F.1 Endpoint definitions

Each contrast is evaluated on a fixed family of endpoints, and the Holm correction below is applied within that family.

The six single-mutant endpoints are: new top-10% discoveries; top-5% and top-1% discoveries, at the dataset's measured 95th and 99th percentiles; discovery AUC, defined in Eq. equation 4.2; final best fitness; and best-fitness AUC.

The ten combinatorial endpoints are: above-parent discoveries; top-10%, top-5% and top-1% discoveries; the discovery AUC of each of those four counts; final best fitness; and best-fitness AUC. All four count endpoints and their AUCs are restricted to measured variants of mutation depth at least two.

Table E.1: Predictor and baseline configurations. ESM-2 embeddings are final-layer residue means excluding BOS/EOS, computed in bf16 with float32 pooling. The zero-shot baseline is a wild-type-marginal score, not a masked-marginal one: one unmasked forward pass of the parent gives the logits, and a variant scores as the sum over its mutated sites of $\log p(\text{mutant} \mid \text{parent}) - \log p(\text{parent} \mid \text{parent})$. It is additive, uses no labels, and receives no feedback between rounds.

Predictor	Input	Configuration
Random forest (EVOLVEpro)	ESM-2 650M embedding	100 trees, `friedman_mse`, `max_features` $= 1.0$, bootstrap, fixed `random_state`
TabPFN	ESM-2 650M embedding	TabPFN v8.5.0 regressor, 8 estimators, `fit_with_cache`, float32 inputs
ALDE DNN	One-hot sites	5 members, $[80, 30, 30, 1]$, LeakyReLU(0.01), float64, lr 10^{-3}, 300 steps, 0.9 train fraction, per-member split
ALDE XGB	One-hot sites	5 members, `reg:squarederror`, 100 trees, `hist`, early stopping after 10 rounds
FolDE	ESM-2 650M embedding	`JBEI/foldy` @ `b0add68`; 5 members, hidden $[100, 50]$, dropout 0.2, lr 3×10^{-4}, weight decay 10^{-5}, 50 pretrain and 200 train epochs, patience 40; constant liar over the top $\min(5000, \lvert\mathcal{U}_t\rvert)$
ESM-2 zero-shot	Sequence only	ESM-2 650M wild-type marginals; no assay labels, no refitting

Discovery AUC is a trapezoidal integral of the cumulative new-discovery curve from 64 to 176 requests divided by 112, so it has units of cumulative discoveries and is not a proportion. Best-fitness AUC integrates range-normalized best-so-far fitness over the same interval, includes the initial measurements at step 0, and uses evaluator-only normalization bounds because fitness scales are not comparable across landscapes. Discovery AUC and best-fitness AUC are distinct endpoints and are never combined.

F.2 Uncertainty and hypothesis tests

Single-mutant summaries weight the four datasets equally, and the standard deviations in the tables describe five seeds. Paired 95% intervals use 12,000 percentile-bootstrap draws, stratified within the four fixed single-mutant datasets or within each combinatorial landscape. Because the datasets are fixed rather than sampled, these intervals describe variation over seeds and do not estimate generalization to new proteins.

Tests are exact two-sided paired sign flips on the within-seed differences, with Holm correction applied within each contrast across its six or ten endpoints. The correction is not paper-wide, and no endpoint was promoted to primary after seeing results. All comparisons are exploratory unless explicitly identified otherwise. Intervals are pointwise, not simultaneous.

F.3 Corrections and validation scope

Two corrections are recorded against earlier exports. A previous external-baseline export duplicated the step-1 best-so-far value at step 0 on 21 trajectories, affecting best-fitness AUC only; the other discovery metrics and final best are unchanged. A historical TrpB4 runner reported all-depth columns rather than the canonical depth-≥ 2 evaluator. The package uses the corrected ALDE host RNG results and the corrected own-history random-CLT data; superseded versions are retained for reference only.

Validation recomputes the core combinatorial and distance-control endpoints from acquired variant IDs and rederives the single-mutant endpoints from stored curves, with external-baseline curves rebuilt from acquired IDs and initial ledgers. It is not a fresh replay of predictor training or acquisition.

G COMPLETE SINGLE-MUTANT RESULTS

The two initialization cohorts are kept separate throughout. All four datasets, both ESM-2 predictors, the representation controls and the external baselines are retained.

Main-text Table 1 gives the per-dataset top-10% counts for the unscreened cohort. The tables below add the quality-conditioned cohort, the remaining five endpoints averaged over the four datasets, and the per-dataset discovery AUC. The per-dataset breakdown of the top-5%, top-1% and best-fitness endpoints is in the released per-seed tables rather than here.

Table G.1: Single-mutant metric means across the four fixed datasets, random cohort; 20 paired campaigns per method. Best metrics include the initial observations; discovery counts exclude them.

Acquisition	Hits	Hit AUC	Top-5%	Top-1%	Best (norm.)	Best AUC
No assay fitting						
Random	10.700	5.157	4.900	0.900	0.795	0.766
ESM-2 zero-shot	36.200	19.464	23.000	5.150	0.876	0.856
RF (ESM-2)						
Greedy / EVOLVEpro	26.150	13.282	13.600	3.400	0.876	0.835
+ EvoCircuit	29.150	14.711	15.150	3.350	0.854	0.821
Random CLT	25.650	13.361	12.950	3.250	0.862	0.830
ESM-distance	27.400	14.236	14.250	3.750	0.868	0.835
TabPFN (ESM-2)						
Greedy / EVOLVEpro	31.650	15.868	16.950	4.050	0.859	0.819
+ EvoCircuit	33.150	17.204	19.300	5.500	0.852	0.822
Random CLT	32.650	16.361	17.850	4.400	0.862	0.820
ESM-distance	32.900	16.700	19.250	4.900	0.887	0.836
FolDE (ESM-2)						
Constant liar	39.950	20.032	21.700	4.400	0.905	0.852
Greedy	41.200	21.300	24.200	5.700	0.941	0.874
+ EvoCircuit	41.100	21.179	24.200	6.700	0.917	0.868

H COMPLETE COMBINATORIAL RESULTS AND PREDICTOR TRANSFER

Results remain separated by landscape and predictor; nothing is pooled across landscapes. Missing requests consume the assay budget; one-hot predictors still use ESM-2/CLT features inside EvoCircuit acquisition. For each landscape the first table gives the discovery counts and request accounting and the second the remaining AUC and best-fitness endpoints, so the two together cover all ten endpoints of Appendix F.1. The upper-tail quality profile is shown for GB1, where the effect is largest; the TrpB3I and TrpB4 profiles are in the released figure-source tables.

I FEATURE AND REPRESENTATION CONTROLS

These controls separate three things that the main comparison confounds: which coordinates are selected, which representation defines retrieval distance, and how activations are pooled over positions. The TrpB4 distance-control cohort is a separate 80-trajectory cohort and is not added to the main seed count. The figure panels corresponding to the first two tables are the appendix figures of the released package; only the frozen-state analysis, which is not a cumulative policy comparison, is reproduced here.

J INITIALIZATION AND EXECUTION HISTORY

Initial-quality interventions and outcome-informed execution extensions are reported separately from the main comparison. Execution stages are not new independent holdouts, and the stage-sensitivity tables exist so that the effect of the staging described in Appendix E.2 is visible rather than absorbed.

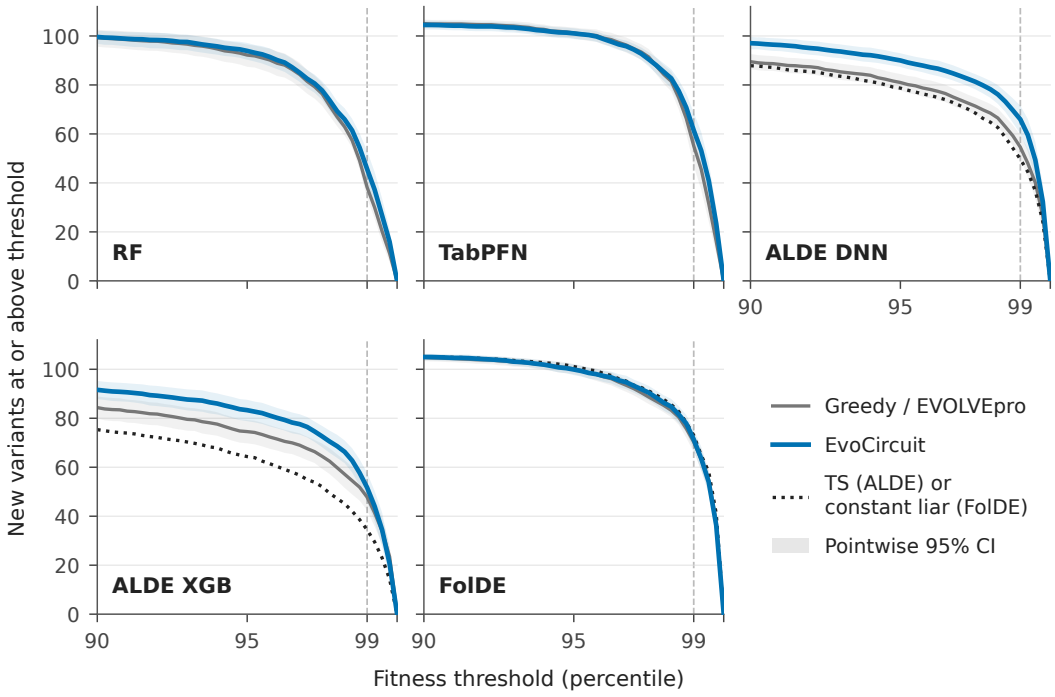


Figure H.1: GB1: upper-tail quality of discovered variants for every main predictor. For each threshold, curves count distinct new variants with fitness at least that landscape percentile, over the 90th through 100th percentiles, so the 90, 95 and 99 values exactly reproduce the top-10%, top-5% and top-1% discovery counts. Only measured variants with at least two substitutions define thresholds and count as discoveries. All 112 requests remain in the budget, including missing measurements; initial requests are excluded. Lines are means over 20 paired seeds and shaded bands are pointwise 95% bootstrap intervals over campaigns, not over independent variants. Dotted lines show TS for ALDE and constant liar for FolDE. This exploratory distribution view does not establish significance across thresholds.

Table H.1: Adding EvoCircuit to a fixed predictor: mean new discoveries within 112 requests after the shared initial 64, over 20 paired seeds per landscape (TrpB3I excludes its first run). *Above parent*: fitness strictly exceeds the starting reference protein; *Top-1%*: fitness reaches the 99th percentile of the landscape; *Disc. AUC*: the above-parent count averaged over the 112-request budget, in units of discoveries rather than a proportion, so that a method finding the same variants earlier scores higher. Only measured variants with at least two substitutions are counted and define percentile thresholds. Higher is better throughout, and the thresholds need not be nested. **Blue bold** indicates a higher EvoCircuit mean than its matched greedy control (EVOLVEpro for RF/TabPFN), not significance. ALDE supports TS and Greedy; constant liar is FolDE's original selection rule. TS: Thompson sampling; XGB: five-model XGBoost ensemble. One-hot describes predictor inputs; EvoCircuit additionally uses ESM-2/CLT. Random/zero-shot require no assay fitting. Paired uncertainty and best-fitness endpoints appear in the appendix.

		GB1			TrpB3I			TrpB4		
Predictor / input	Acquisition	Above parent	Top-1%	Disc. AUC	Above parent	Top-1%	Disc. AUC	Above parent	Top-1%	Disc. AUC
No assay fitting	Random	2.50	0.75	1.22	1.40	1.10	0.66	0.55	1.05	0.22
	ESM-2 zero-shot	16.00	3.00	6.29	27.80	23.80	16.18	54.00	65.00	28.43
RF / ESM-2	EVOLVEpro	75.40	38.20	33.32	33.00	26.90	16.80	64.60	73.40	29.90
	+ EvoCircuit	**76.95**	**45.65**	33.27	32.55	26.40	**17.04**	64.40	73.30	29.59
TabPFN / ESM-2	EVOLVEpro	90.35	55.15	41.34	51.10	41.70	26.07	77.00	85.60	34.69
	+ EvoCircuit	90.00	**61.50**	40.46	49.90	40.65	25.19	74.65	84.40	32.41
ALDE DNN / one-hot	TS	67.80	49.70	29.70	50.70	44.05	25.14	54.45	61.85	22.70
	Greedy	70.95	54.60	30.17	52.40	43.95	26.19	62.70	70.00	25.66
	+ EvoCircuit	**80.90**	**65.95**	**34.57**	**54.05**	**45.75**	**27.30**	**65.30**	**74.10**	**27.59**
ALDE XGB / one-hot	TS	51.35	34.45	18.88	37.70	33.20	15.48	36.25	41.55	13.27
	Greedy	63.85	47.80	25.08	38.00	32.50	17.79	43.90	52.05	16.45
	+ EvoCircuit	**71.80**	**51.70**	**28.29**	**38.95**	**33.15**	**18.93**	**51.45**	**59.00**	**20.98**
FolDE / ESM-2	Constant liar	90.65	73.00	42.50	52.05	42.90	26.90	72.60	80.20	30.89
	Greedy	88.35	70.30	41.77	52.40	42.95	27.79	71.75	79.45	30.95
	+ EvoCircuit	**90.00**	**71.35**	41.30	51.65	42.65	26.40	**74.40**	**82.00**	**31.19**

K Feature interpretation case study

The record behind Section 5.4: one saved GCK trajectory — random forest predictor, quality-conditioned cohort, seed 919410 — read at rounds 4 and 5, and at round 8, which the main text does not discuss.

K.1 Replay and verification

At each round the acquisition is recomputed from that round's saved inputs — the feature representation, the predictor scores and the observed ledger — using the reference procedure of Algorithm **??** and the distance of Eq. equation **??**. The replay reproduces the saved 16-variant batch and the saved 128-candidate shortlist exactly before any feature is removed, and the removal results agree with an independent 32-replay record made earlier on the same trajectory. Candidate labels are never read during selection.

Each round screens 2,048 observed-variable CLT features with balanced layer quotas; 361 pass the association screen at round 4, 278 at round 5 and 141 at round 8, and the eight with the strongest absolute association form the representation. The features discussed here are among them, at bootstrap sign frequency 1.0: `L14_F09977` at -0.487 and -0.441 for rounds 4 and 5, `L03_F00835` at -0.461 and -0.427, and `L01_F05597` at 0.355 for round 8.

K.2 How feature descriptions are produced

Descriptions follow the reference-protein annotation of InterPLM (Simon & Zou, 2025), whose objective we keep: a description that predicts where the feature fires in proteins it was not shown.

Table H.2: Full discovery and request-accounting means on GB1. Missing measurements consume request budget. One-hot hosts retain ESM-2/CLT access in their EvoCircuit acquisition module.

Acquisition	Parent hits	Top-10%	Top-5%	Top-1%	Hit AUC	Missing
No assay fitting						
Random	2.50	10.20	4.85	0.75	1.22	7.40
ESM-2 zero-shot	16.00	43.05	29.00	3.00	6.29	0.00
ALDE XGB (one-hot)						
Greedy	63.85	84.35	74.55	47.80	25.08	1.90
TS	51.35	75.35	64.35	34.45	18.88	2.70
+ EvoCircuit	71.80	91.60	83.30	51.70	28.29	1.60
ALDE DNN (one-hot)						
TS	67.80	87.85	78.70	49.70	29.70	2.30
Greedy	70.95	89.50	81.00	54.60	30.17	1.80
+ EvoCircuit	80.90	97.05	90.00	65.95	34.57	1.55
Random CLT	70.85	90.55	82.10	56.30	29.94	1.75
ESM-distance	81.10	96.30	90.25	63.20	35.51	0.90
FolDE (ESM-2)						
Constant liar	90.65	105.15	101.10	73.00	42.50	0.35
Greedy	88.35	104.90	100.15	70.30	41.77	0.35
+ EvoCircuit	90.00	105.00	99.90	71.35	41.30	0.10
RF (ESM-2)						
Greedy / EVOLVEpro	75.40	99.30	92.20	38.20	33.32	0.30
+ EvoCircuit	76.95	99.60	93.90	45.65	33.27	0.05
TabPFN (ESM-2)						
Greedy / EVOLVEpro	90.35	105.05	101.25	55.15	41.34	0.25
+ EvoCircuit	90.00	104.50	101.05	61.50	40.46	0.00
Random CLT	83.60	101.70	96.15	55.50	37.19	0.20
RF (one-hot)						
Greedy	61.75	84.60	74.45	44.40	25.65	1.00
+ EvoCircuit	73.85	95.45	87.80	51.20	31.34	0.85

Table I.1: Single feature controls. Dataset-stratified paired intervals over four datasets, five seeds each. Holm correction covers six endpoints within each contrast, not all comparisons in the paper.

Comparison	Metric	Delta	95% CI	p (Holm)
Unscreened / RF - random CLT	new_hits	+3.50	[+1.35, +5.80]	0.05363
Unscreened / RF - ESM-distance	new_hits	+1.75	[-0.25, +3.80]	0.853
Unscreened / TabPFN - random CLT	new_hits	+0.50	[-2.25, +3.60]	1
Unscreened / TabPFN - ESM-distance	new_hits	+0.25	[-1.95, +2.35]	1
Unscreened / RF - random CLT	hit_auc	+1.35	[+0.23, +2.52]	0.1868
Unscreened / RF - ESM-distance	hit_auc	+0.47	[-0.46, +1.48]	1
Unscreened / TabPFN - random CLT	hit_auc	+0.84	[-0.50, +2.26]	1
Unscreened / TabPFN - ESM-distance	hit_auc	+0.50	[-0.22, +1.22]	1
Quality-conditioned / RF - random CLT	new_hits	+4.50	[+2.40, +6.80]	0.008286
Quality-conditioned / RF - ESM-distance	new_hits	+3.20	[+1.05, +5.50]	0.08457
Quality-conditioned / TabPFN - random CLT	new_hits	+0.60	[-2.20, +3.55]	1
Quality-conditioned / TabPFN - ESM-distance	new_hits	+1.50	[+0.10, +2.95]	0.4604
Quality-conditioned / RF - random CLT	hit_auc	+2.01	[+0.89, +3.22]	0.05091
Quality-conditioned / RF - ESM-distance	hit_auc	+0.82	[-0.32, +2.12]	0.734
Quality-conditioned / TabPFN - random CLT	hit_auc	+0.67	[-0.71, +2.11]	1
Quality-conditioned / TabPFN - ESM-distance	hit_auc	+1.05	[+0.20, +1.94]	0.3512

Table I.2: GB1 acquisition controls, mean ± SD over all 20 paired seeds. Selected and random CLT match feature count and layer allocation; ESM-distance uses all 1,280 mean embedding coordinates. ESM-distance was not run for GB1 TabPFN. The controls augment previously used initializations.

Selection rule	Parent hits	Hit AUC	Top-1%
ALDE DNN			
Top-16	71.0 ± 8.2	30.2 ± 5.2	54.6 ± 11.2
Random CLT	70.8 ± 10.5	29.9 ± 7.0	56.3 ± 10.6
ESM-distance	81.1 ± 8.5	35.5 ± 5.9	63.2 ± 16.0
+ EvoCircuit	80.9 ± 7.5	34.6 ± 4.8	66.0 ± 13.6
TabPFN			
Top-16	90.3 ± 7.5	41.3 ± 5.6	55.1 ± 19.1
Random CLT	83.6 ± 6.8	37.2 ± 4.9	55.5 ± 11.4
+ EvoCircuit	90.0 ± 7.3	40.5 ± 4.9	61.5 ± 17.9

Table I.3: TrpB4 RF distance controls, mean ± SD over their own 20 seeds; only ten overlap with the main RF cohort. Endpoints were recomputed from unique acquired IDs under the same finite, depth-≥ 2 definition. The historical hamming arm uses standardized one-hot distance; it is distinct from the excluded position-locality baseline. The source column distinct counts distinct requested sequences including initialization, not site-amino-acid diversity.

Acquisition	Parent hits	Hit AUC	Top-1%	Best
Selected CLT	63.5 ± 14.4	29.3 ± 7.2	72.0 ± 11.3	0.812 ± 0.075
Random CLT	61.5 ± 13.0	27.3 ± 6.7	69.8 ± 10.7	0.793 ± 0.054
ESM-distance	61.4 ± 13.2	28.4 ± 7.9	69.0 ± 11.4	0.811 ± 0.080
Standardized one-hot distance	65.3 ± 11.4	29.3 ± 5.8	73.2 ± 9.3	0.816 ± 0.087

Table I.4: Mutation-site minus residue-mean CLT, paired by seed. The residue-mean representation is the common main method. Mutation-site selection changes feature-site coordinates; this is not an isolated pooling-only ablation.

Landscape / predictor	Metric	n	Delta	95% CI
TrpB3I				
RF	above_parent	20	+2.000	[-0.650, +4.350]
RF	parent_auc	20	-0.186	[-1.421, +0.971]
RF	top1	20	+1.950	[-0.300, +4.050]
RF	final_best	20	+0.014	[-0.002, +0.042]
TabPFN	above_parent	20	+0.950	[-0.950, +3.300]
TabPFN	parent_auc	20	+0.932	[-0.418, +2.493]
TabPFN	top1	20	+0.900	[-0.750, +3.100]
TabPFN	final_best	20	+0.008	[+0.000, +0.023]
TrpB4				
RF	above_parent	20	-2.700	[-7.400, +1.750]
RF	parent_auc	20	-1.229	[-3.154, +0.614]
RF	top1	20	-1.700	[-5.850, +2.250]
RF	final_best	20	-0.014	[-0.042, +0.014]

Table I.5: History-correct single-state feature control. Each value averages new top-10% hits among eight shortlist selections across seven stored EvoCircuit rounds per initialization. Only the CLT panel changes; the random panel uses the focal candidate pool and layer quotas. Intervals resample five initializations within each of four fixed datasets (12,000 draws). Predictors and initialization conditions stay separate. These counterfactual one-batch gains are not closed-loop cumulative effects; no biological causal conclusion follows.

Condition / predictor	n	Selected	Random	Delta	95% CI
Quality-conditioned / RF	20	2.429	1.700	+0.729	[+0.457, +1.000]
Quality-conditioned / TabPFN	20	2.207	1.779	+0.429	[+0.221, +0.607]
Unscreened / RF	20	2.157	1.693	+0.464	[+0.257, +0.657]
Unscreened / TabPFN	20	2.164	1.900	+0.264	[+0.007, +0.536]

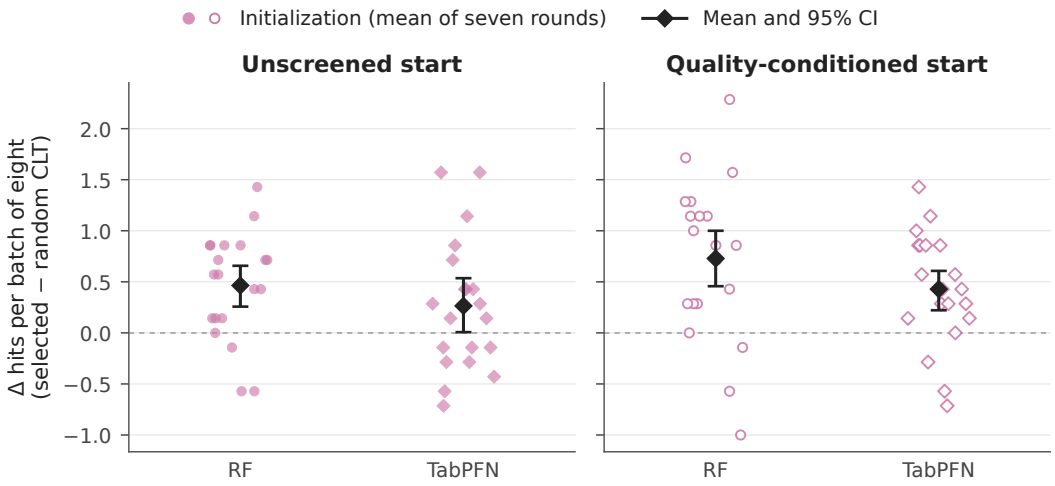


Figure I.1: Feature contribution at fixed EvoCircuit observation histories. Only the CLT panel changes; random panels match the focal candidate pool, feature count and layer quotas. Each point averages seven rounds within one dataset/initialization/predictor, giving 20 points per group (four datasets, five seeds each); circles are RF and diamonds TabPFN. Conditions and predictors remain separate. Black diamonds and bars show means and pointwise 95% dataset-stratified bootstrap intervals. These one-batch counterfactual comparisons do not estimate cumulative policy gains or biological causality.

Table J.1: TrpB4 fixed-30-seed initialization intervention. High and random conditions are separate; high starts use fitness-informed initialization. These 180 paths are not appended to the main unconditional cohort. Values are from the existing completed export; this packaging step checked keys and contrast arithmetic, not a new acquisition replay.

Condition / acquisition	Parent hits	Top-5%	Top-1%	Best
high / Selected CLT	64.0 ± 10.6	99.9 ± 5.7	74.7 ± 8.3	0.810 ± 0.076
high / Predictor only	62.4 ± 11.9	98.4 ± 5.8	72.2 ± 8.4	0.805 ± 0.070
high / Random CLT	59.2 ± 11.1	97.1 ± 6.9	69.1 ± 9.0	0.814 ± 0.082
random / Selected CLT	55.9 ± 12.5	93.3 ± 7.0	66.0 ± 9.8	0.775 ± 0.065
random / Predictor only	58.8 ± 12.6	94.8 ± 6.7	68.2 ± 10.6	0.801 ± 0.068
random / Random CLT	54.4 ± 13.1	90.7 ± 8.3	63.8 ± 10.4	0.783 ± 0.060

Table J.2: High-minus-random interaction from the registered 30-seed study. The primary parent-hit selected-minus-random-CLT interaction was not significant (p=0.2435). Positive within-high contrasts are secondary. Intervals and p-values are retained from the source export; exact integer sign-flip or Monte Carlo for continuous metrics, as recorded in the accompanying CSV.

Metric / contrast	High	Random	Interaction	95% CI	p
above_parent / clt-randomclt	+4.767	+1.533	+3.233	[-1.933, +8.300]	0.2435
above_parent / clt-natural	+1.600	-2.933	+4.533	[-1.467, +10.267]	0.1472
top1 / clt-randomclt	+5.633	+2.200	+3.433	[-0.700, +7.501]	0.1274
top1 / clt-natural	+2.500	-2.133	+4.633	[-0.701, +9.667]	0.09838
top5 / clt-randomclt	+2.833	+2.633	+0.200	[-3.100, +3.500]	0.924
top5 / clt-natural	+1.500	-1.467	+2.967	[+0.233, +5.867]	0.05607
final_best / clt-randomclt	-0.004	-0.009	+0.004	[-0.030, +0.038]	0.8144
final_best / clt-natural	+0.005	-0.026	+0.031	[-0.005, +0.066]	0.1074
auc_best_normalised / clt-randomclt	+0.005	-0.010	+0.014	[-0.005, +0.035]	0.1883
auc_best_normalised / clt-natural	+0.002	-0.021	+0.023	[+0.001, +0.046]	0.05603

Table J.3: Descriptive source-stage sensitivity for TrpB3I. Differences are EvoCircuit minus the same greedy host. Stages document execution history and are not newly independent holdouts. This historical table retains every stored seed, including the first TrpB3I run; the main cohort and its uncertainty summaries use only the subsequent 20 TrpB3I seeds.

Host / source stage	n	Delta hits	Delta top-1%	Delta best
RF / ESM-2				
first1	1	+1.000	+2.000	+0.000
remaining20	20	-0.450	-0.500	-0.014
all	21	-0.381	-0.381	-0.013
TabPFN / ESM-2				
first1	1	-1.000	-2.000	+0.000
remaining20	20	-1.200	-1.050	-0.008
all	21	-1.190	-1.095	-0.007
RF / one-hot				
first1	1	+9.000	+9.000	+0.000
remaining20	20	+0.900	+0.700	+0.029
all	21	+1.286	+1.095	+0.028
ALDE / one-hot				
first1	1	+4.000	+4.000	+0.000
remaining20	20	+1.650	+1.800	+0.008
all	21	+1.762	+1.905	+0.007
FolDE / ESM-2				
first1	1	+2.000	+0.000	+0.000
remaining20	20	-0.750	-0.300	+0.000
all	21	-0.619	-0.286	+0.000

Table J.4: Descriptive source-stage sensitivity for TrpB4. Differences are EvoCircuit minus the same greedy host. Stages document execution history and are not newly independent holdouts. This historical table retains every stored seed, including the first TrpB3I run; the main cohort and its uncertainty summaries use only the subsequent 20 TrpB3I seeds.

Host / source stage	n	Delta hits	Delta top-1%	Delta best
RF / ESM-2				
shortlist_extension10	10	+1.500	+2.000	-0.033
shortlist_initial10	10	-1.900	-2.200	+0.012
all	20	-0.200	-0.100	-0.010
TabPFN / ESM-2				
shortlist_extension10	10	-2.600	-2.100	+0.030
shortlist_initial10	10	-2.100	-0.300	+0.004
all	20	-2.350	-1.200	+0.017
RF / one-hot				
shortlist_extension10	10	+7.500	+6.700	+0.049
shortlist_initial10	10	+6.900	+8.600	+0.001
all	20	+7.200	+7.650	+0.025
ALDE / one-hot				
shortlist_extension10	10	-2.300	+0.400	-0.061
shortlist_initial10	10	+7.500	+7.800	+0.005
all	20	+2.600	+4.100	-0.028
FolDE / ESM-2				
shortlist_extension10	10	+2.300	+0.900	-0.016
shortlist_initial10	10	+3.000	+4.200	+0.020
all	20	+2.650	+2.550	+0.002

Each is issued as an annotation-supported candidate requiring review, not as a function assignment. The payload for one feature holds the 50 Swiss-Prot proteins with the strongest activation, each with its 21-residue peak window, normalized activation tier and annotations; up to 25 annotation terms enriched against matched controls and 25 associated with controls only, with Benjamini–Hochberg values; a recurrence consensus over the peak windows; and the distribution of relative peak positions. A language model returns a schema-checked object naming the activation pattern, the high- and low-activation predictors, the peak signature and anchor residues, a confidence level, caveats, and the evidence identifiers it used; the validator rejects outputs that fail the schema or cite evidence that was not supplied.

Two boundaries follow from the payload. Control contrasts are annotation-level rather than individual inactive proteins, so control-associated terms indicate contexts without activation among the supplied activators rather than measured inactivity; and a stated confidence refers to the reference evidence, not to transfer into the target. The payload contains no GCK sequence, no assay measurement and no attribution from the graphs below, the target protein being used only to choose which features to describe, and all 844 features here were described after the campaign had finished, 46 of them among those the acquisition selected.

K.3 The acquisitions

D is the distance of Eq. equation **??** to the anchor whose neighbour list admitted the candidate; a discovery is a new top-10% variant, as in Section 5.1.

The round-8 acquisition rests on two other features, and both carry statements that can be checked against GCK's saved profiles. `L06_F04406` activates where a cysteine sits three residues earlier: the offset holds at 12 of its 13 active native positions in E322K and accounts for 97.5% of the feature's activation, the largest at C233. `L01_F05597` responds to asparagine- and glutamate-centred hydrophobic contexts and peaks at N204, which UniProt annotates as a substrate-binding residue with structural evidence (**??**) — evidence about the residue, not about the feature.

Table K.1: The acquisitions examined. Predictor rank is the candidate's position in the random forest's ranking of eligible unmeasured variants at that round, so a large rank is one the predictor alone would not have requested.

Round	Candidate	Predictor rank	Anchor	D	Activity	Outcome
4	D217A	103 / 8,506	Q219M	0.714	2.151	discovery
5	D217T	95 / 8,490	D217A	0.182	2.675	discovery
5	S411A	42 / 8,490	D217A	0.109	0.407	not a discovery
8	E322K	500 / 8,442	D267A	0.063	1.887	discovery

K.4 What the four descriptions claim, and how they fare in GCK

Table K.2 sets each description against GCK. The pattern is the same in all four: the family label is drawn from reference proteins unrelated to a hexokinase and does not transfer, while the local statement — which residue the peak sits on, and in what immediate context — can be checked and holds. For `L14_F09977` the label is a cyclic-nucleotide-binding domain, which GCK does not have; what survives is a peak at E, K or A, landing on E221. For `L03_F00835` the label is EGF and metallothionein, extracellular and metal-binding families GCK is not a member of; what survives is activation on cysteines, and GCK's two strongest, C230 and C233, form a C-x-x-C pair of the kind that fills the reference peak windows, although the reference's more specific C at offset -2 becomes -3 here. The two round-8 features behave the same way. Support for the family label does not predict whether it transfers: `L14_F09977` carries the most uniform one in the set, shared by all 50 of its reference proteins, and it is the one furthest from GCK. The comparison runs in one direction only, since the payload carries sequence and annotation alone: whether the reference peaks sit in strands comparable to GCK's is not something this evidence can answer.

Table K.2: What each description claims, split into the family-level label drawn from the reference proteins and the local statement about the peak. Counts in the family column are the reference proteins carrying that term out of 50. Confidence is the level the description states about its *reference* evidence, not about transfer into GCK.

Feature	Family-level claim	Local claim, and GCK	Conf.
`L14_F09977`	cNMP-binding domain (IPR018490, 50/50); 43/50 ion channels	peak at E, K or A; GCK's largest of three at E221	moderate
`L03_F00835`	EGF (IPR000742, 20/50), metallothionein (IPR017854, 16/50)	peak on C, C frequent at -2; GCK active at C230/C233/C252/C457, C230–C233 at -3	moderate
`L06_F04406`	extracellular space (GO:0005615, 18/50), cystine-knot cytokine (10/50)	C three residues before the peak; 12 of 13 active positions in E322K	moderate
`L01_F05597`	plasma membrane (GO:0005886, 24/50)	N/E-centred hydrophobic window; GCK peaks at N204	high

K.5 The activation state and the window it marks

Beyond the numbers in Section 5.4: substitutions within ten residues of C220–E221 fall below the empty interval in 32.1% of cases ($n = 383$) against 4.9% more than twenty residues away ($n = 7{,}827$), a 6.6-fold enrichment, with D217 substitutions below in 17 of 19 cases; 62% of all variants below the interval are nonetheless more than twenty residues away, S411A among them. At round 4 the lower group holds four of the sixteen anchors, all substitutions at Q219, and 31 of the 128 shortlisted candidates; at round 5 it holds four anchors and 32 shortlisted candidates.

None of the feature's three positions is itself an annotated functional site, but all three sit in strands that carry such sites (**?**). F74 lies inside the strand 72–92 of the small subdomain, which contains the ATP contacts at 78–83; C220 and E221 are the first two residues of the strand 220–237 of the large subdomain, which contains the ATP contact T228 and the substrate contact N231. The helix 205–214 precedes that strand, so D217 and Q219, the substitutions that flip the state most often, fall in the short loop between the two.

Table K.3: The measured library against the activation state, read after the fact. *Switched* is the lower of the two groups the feature's values fall into, and *the window* is within ten residues of C220–E221. Discovery rate is the top-10% rate within each group. The groups are not a partition of the library and these counts estimate no effect; they describe where this library's activating substitutions lie.

Group	Variants	Discoveries	Rate
Whole library	8,570	857	10.0%
Switched	614	74	12.1%
inside the window	123	41	33.3%
inside the window, excluding D217 and Q219	95	28	29.5%
outside the window	491	33	6.7%
Not switched, inside the window	260	40	15.4%
Annotated substrate contacts	167	2	1.2%
Annotated ATP contacts	355	30	8.5%
segment 411–415, which contains S411A	93	8	8.6%

K.6 Removal at a fixed observed state

A removal deletes one representation coordinate and recomputes the observed-set standard deviations, the ridge, the distances, the shortlist and the batch under the original procedure. It does not refit the predictor or simulate later rounds, and because it rescales the surviving coordinates and changes which candidates compete, it shows dependence on the panel rather than an additive contribution. Table K.4 also covers `L03_F00835`, the other feature selected at rounds 4 and 5: its removal changes the batch without excluding any of the three candidates.

Table K.4: Feature removal at the same observed state. *Guided hits* counts discoveries among the eight requests taken from the feature-guided shortlist, before and after the removal; the remaining eight requests of each batch are the predictor's own top-ranked candidates and do not change. The last column lists which of the examined candidates leave the batch.

Round	Feature removed	Guided hits	Candidates that leave the batch
4	`L14_F09977`	$3 \to 1$	D217A
4	`L03_F00835`	$3 \to 4$	none
5	`L14_F09977`	$5 \to 3$	D217T, S411A
5	`L03_F00835`	$5 \to 3$	none
8	`L06_F04406`	$1 \to 1$	E322K
8	`L01_F05597`	$1 \to 0$	E322K

K.7 Why the attribution graphs are not offered as an explanation

For each variant we hold an attribution graph over CLT feature–residue coordinates, built to reproduce a surrogate of the predictor. None of the features the acquisition selected appears in those graphs, so we recomputed each one with the eight selected features admitted. Both round-8 features and `L03_F00835` then occupy residue-level paths, the round-8 pair keeping its native peak at profile cosines of 0.829 and 0.627, while `L14_F09977` is inactive at both rounds 4 and 5, so no path can be drawn for the coordinate that decided those requests. The recomputed graphs also track the saved predictor's ranking too weakly to meet the criterion fixed in advance, at every round examined (Table K.5), so a path locates a feature inside a conditional computation and does not explain the ranking or the measured activity. Across the twenty-five case–round combinations examined, 136 of 200 selected-feature coordinates are active in the recomputed graph and 60 keep the peak position of the native profile, which is further reason to treat presence in a graph and native activation as separate facts.

Table K.5: Rank agreement with the saved random forest, as Spearman ρ over the n variants held at each round. *Native features* and *Full CLT* score the representations directly; *Original graph* and *Recomputed* score the attribution surrogate before and after the selected features are admitted. *RF top-16 kept* counts how many of the predictor's own top 16 the surrogate also places in its top 16. No round meets the criterion fixed in advance, so the graphs are reported as a negative result.

Round	n	Native features	Full CLT	Original graph	Recomputed	RF top-16 kept
4	96	0.737	0.674	-0.050	-0.008	3 / 16
5	96	0.693	0.530	0.364	0.364	5 / 16
8	95	0.739	0.509	0.323	0.325	4 / 16
10	93	0.734	0.640	0.056	0.050	3 / 16

K.8 Scope

The trajectory and the rounds were chosen after their outcomes were known, so this describes four selections rather than estimating an effect, and claims no significance. No new assay, predictor fit, GPU inference or model call was made for this analysis: D217A, D217T, S411A and E322K were all measured in the library the trajectory ran on, and their activities are used only to score selections that were already complete. Where an external annotation appears it describes a residue of GCK, never a feature, and the reference descriptions name protein families that we do not transfer to GCK.

L Interactive feature explorer

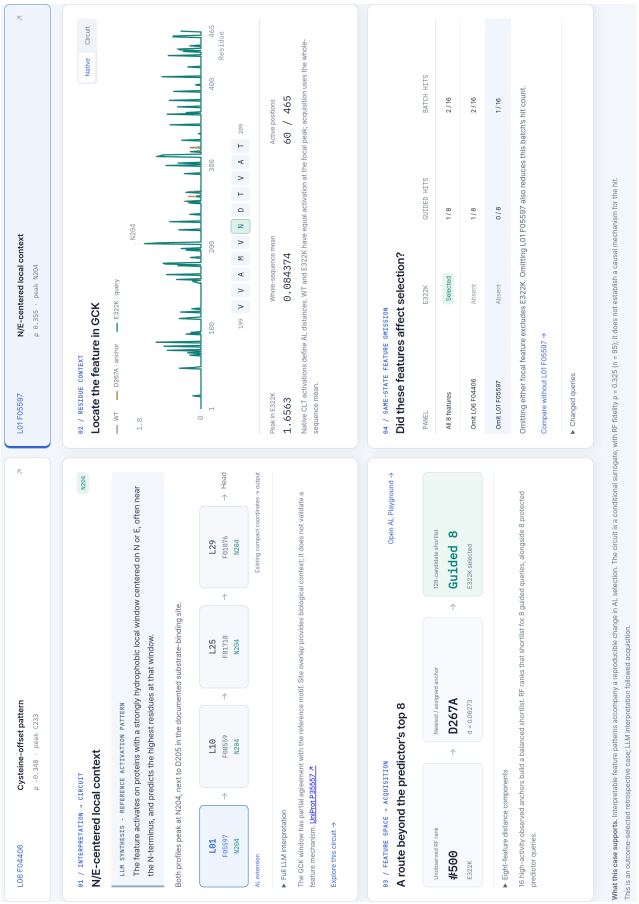


Figure L.1: The interactive feature explorer, shown on the round-8 acquisition of Appendix K. Clockwise from top left: the reference-protein description of `L01_F05597` and the circuit path it occupies; its activation profile over GCK's 465 residues for the wild type, the anchor D267A and the query E322K; the route by which E322K entered the batch from rank 500 under the random forest; and the effect of omitting each focal feature at the same observed state. The explorer reads stored activations and saved campaign records only: it runs no model, requests no assay, and produces no evidence beyond what the appendix already reports.