HYPOTHESIS PORTFOLIO SEARCH: SHARED FEEDBACK FOR GENETIC PERTURBATION DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents can effectively link biological knowledge to genetic perturbation experiments, yet tight budgets leave most plausible interventions untested. Active hit discovery aims to maximize the recovery of assay-specific hits within these limits. However, standard scoring methods select candidates purely by global rank, risking over-concentration on dominant pathways while neglecting under-explored biological mechanisms, whereas prompt-based LLM agents struggle to convert accumulating feedback into quantitative, assay-specific value estimates. To address this, we formulate discovery as two coupled decisions: allocating testing budgets across hypothesis-defined candidate pools, and selecting specific genes within each pool. We propose **Hy**pothesis **P**ortfolio **S**earch, termed **HyPS**, a novel framework that treats hypotheses as revisable search units with dynamic budget allocations. Feedback iteratively updates these units, while a shared lightweight scorer integrates historical responses with biological knowledge to rank untested candidates. We evaluate HyPS retrospectively on six single-gene perturbation screens using nine LLM backbones at matched testing budgets. Averaged equally across screens and backbones, HyPS raises overall hit recall from 8.40% to 12.47%, outperforming the same-backbone retrieval-augmented agent in 53 of 54 comparisons. On a two-gene perturbation benchmark, HyPS increases mean hit-pair recovery by 72.4% over the same comparator across three LLM backbones. Our code will be made publicly available upon publication.

1 INTRODUCTION

Automated scientific agents powered by large language models (LLMs) offer promising avenues for accelerating biological research and therapeutic development (Qi et al., 2026; Huang et al., 2025). By connecting domain knowledge to experimental planning, these agents facilitate systematic investigation into how genetic interventions change cellular behavior. However, experimental capacity remains fundamentally constrained by tight laboratory budgets. This defines the goal of active hit discovery: recovering as many interventions as possible that satisfy an assay-specific criterion under a fixed testing budget (Rubbi et al., 2026). In typical screening settings, experimental capacity is limited to a few hundred tests per run—such as 160 to 640 tests across total search spaces exceeding 19,000 protein-coding genes, with combinations further enlarging the space. Under such stringent constraints, valuable biological hits are easily missed if experimental space exploration is improperly structured.

Current paradigms for active discovery face distinct, complementary limitations. On one hand, conventional Bayesian optimization and acquisition-function-based approaches learn from measured responses to prioritize promising candidates (Shahriari et al., 2016; Lyle et al., 2023; Li et al., 2026). However, standard scoring methods select candidates purely by global rank, risking over-concentration on dominant pathways while neglecting under-explored biological mechanisms. On the other hand, LLM agents leverage massive scientific literature to broaden candidate generation and propose data-based tests (Weng et al., 2026). Yet, prompt-based LLM agents struggle to convert accumulating experimental feedback into quantitative, assay-specific value estimates, often failing to systematically update candidate prioritization as new observations arrive (Gupta et al., 2025). Consequently, existing approaches either lack structural mechanisms to preserve biological diversity or lack explicit quantitative feedback loops to score untested candidates.

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.

Crucially, the ability to propose plausible experiments does not guarantee effective sequential decisions. First, prompt-driven exploration can be unreliable. In bandit experiments, most tested model–prompt configurations failed to explore robustly, whereas a configuration combining reasoning with externally summarized history succeeded (Krishnamurthy et al., 2024). Second, useful feedback needs to be translated into estimates that guide subsequent tests. Bayesian-OPRO provides evidence for this approach: an outcome-updated Gaussian-process surrogate guides LLM proposals and improves search on word and molecular optimization tasks (Agarwal et al., 2025). These advances motivate a more specific question for genetic perturbation discovery: how can LLM agents systematically convert accumulating experimental feedback into joint updates of candidate scope, directional budgets, and testing priorities? ***To address this, we argue for augmenting prompt-based planning with revisable search units, explicit experimental allocations, and quantitative feedback scoring.***

To operationalize this principle, we formulate active hit discovery as ***two coupled decisions: allocating testing budgets across hypothesis-defined candidate pools, and selecting specific genes within each pool.*** We propose **Hy**pothesis **P**ortfolio **S**earch, termed **HyPS**, a novel framework that treats hypotheses as revisable search units with dynamic budget allocations. Under HyPS, LLM-generated hypotheses and biological anchors define distinct candidate pools and thereby determine where tests can be allocated. As feedback arrives, HyPS dynamically updates these search units and reallocates bounded experimental quotas across pools to prevent over-investment in unpromising directions. Simultaneously, a shared lightweight scorer combines accumulated assay responses with fixed biological knowledge graphs to rank untested genes within those quotas and determine what to test. By decoupling direction-level portfolio allocation from gene-level feedback scoring, HyPS enables transparent exploration of new biological mechanisms without discarding prior experimental evidence.

We rigorously evaluate HyPS against retrieval-enabled BioDiscoveryAgent under a benchmark-informed retrospective protocol with matched backbones and assay budgets detailed in Section 5.1. Across six single-gene screens and nine LLM backbones, HyPS raises overall hit recall from 8.40% to 12.47% and non-common-essential recall from 5.57% to 8.15%. On the two-gene benchmark, mean recovered hit pairs increase from 80.6 to 139.0 across three backbones, a 72.4% relative gain. Component studies identify feedback-based scoring as the most consistent driver of performance gains, whereas hypothesis-adaptation controls yield setting-dependent behaviors. Overall, these findings demonstrate that integrating structured portfolio allocation with quantitative feedback scoring significantly resolves the trade-off between broad mechanistic exploration and exploitation under tight budgets, establishing HyPS as a principled, highly effective framework for hypothesis-driven active discovery.

2 Related Work

Active hit discovery and experimental design. GeneDisco benchmarks sequential genetic experiment design (Mehrjou et al., 2022). Within this setting, Probability-of-Hit targets posterior threshold exceedance, while DiscoBAX values effective, diverse intervention sets (Rubbi et al., 2026; Lyle et al., 2023). HyPS addresses a complementary decision: how knowledge defines revisable pools and how response estimates guide selection under directional budgets.

Knowledge-guided agents for perturbation discovery. BioDiscoveryAgent combines history, reflective planning, and retrieval (Roohani et al., 2025). LLMNN updates LLM-proposed centers from feedback, retrieves neighbors, and divides the budget equally (Gupta et al., 2025). HyPS combines persistent hypothesis–anchor units with outcome-dependent quotas and shared within-pool scoring. This distinguishes its decision structure, but without a direct LLMNN comparison does not establish superior performance.

Combining prior knowledge with experimental feedback. LLAMBO uses in-context learning for candidate sampling and surrogate prediction, whereas GOLLuM adapts language-derived representations through a Gaussian-process objective (Liu et al., 2024; Ranković et al., 2026). GEARS uses co-expression and Gene Ontology relations to predict transcriptional responses (Roohani et al., 2024). In contrast, HyPS propagates observed assay responses through fixed relations and fits four scoring parameters. Knowledge supplies structure, not predetermined gene scores. Quotas then govern how those estimates enter the next batch.

3 PROBLEM FORMULATION

We study sequential hit discovery under a fixed assay budget. Inputs are a biological goal, eligible interventions, biological knowledge, a response threshold, and completed assay outcomes. The threshold's provenance in the evaluated implementation is specified in Section 5.1.

Single-gene perturbation. Let $\mathcal{X}$ be the eligible genes. Testing $x \in \mathcal{X}$ yields a phenotypic response $f(x) \in \mathbb{R}$, with batch members tested separately. The target set $\mathcal{H} \subseteq \mathcal{X}$ contains benchmark hits, whose criteria and response directions can differ across assays.

Before round t, the history is $\mathcal{D}_{t-1} = \{(x, f(x)) : x \in \mathcal{O}_{t-1}\}$ for previously tested set $\mathcal{O}_{t-1}$. Starting from an empty history, the agent commits b distinct, untested genes $\mathcal{B}_t \subseteq \mathcal{X} \setminus \mathcal{O}_{t-1}$ before observing their responses. Its objective over T rounds is

$$\max_{\mathcal{B}_1,\dots,\mathcal{B}_T} |\mathcal{O}_T \cap \mathcal{H}|, \qquad \mathcal{O}_T = \bigcup_{t=1}^{T} \mathcal{B}_t. \tag{1}$$

The budget is Tb. Dividing recovered hits by $|\mathcal{H}|$ gives recall, not the hit rate among tested candidates. Benchmark hits are distinct from HyPS's internal response labels.

Two-gene perturbation. Here, $x = (g_i, g_j)$ denotes two genes perturbed jointly in one cell, with measured response $f_{\text{pair}}(x)$. Both $\mathcal{X}_{\text{pair}}$ and $\mathcal{H}_{\text{pair}}$ contain combinations, not individual genes. Equation (1) applies to untested pairs: one budget unit is one joint experiment, and a gene may reappear with a different partner. Joint response does not, by itself, identify a genetic-interaction criterion. In either task, candidate inclusion and test selection are separate decisions: ranking cannot recover excluded candidates, and inclusion does not ensure a test.

4 HYPOTHESIS PORTFOLIO SEARCH

4.1 METHOD OVERVIEW

HyPS separates candidate admission from prioritization in a closed loop (Figure 1). Hypothesis-level search defines pools and quotas, while candidate-level search selects genes within them. Completed assays extend cumulative history for subsequent decisions. Thus, evidence is shared across directions, while selection respects pool-specific quotas. Below we specify the single-gene implementation.

4.2 HYPOTHESIS-LEVEL SEARCH: WHERE TO SEARCH

Hypothesis portfolio and anchor-guided retrieval. HyPS maintains K persistent slots, each with a hypothesis and three anchors. Given the goal, current hypotheses, summaries of assay history, and slot statistics, the LLM keeps, revises, or replaces each unit. Anchors retrieve pools $\mathcal{P}_{k,t}$ through Achilles dependency-profile similarity (Dempster et al., 2019). Tested genes are excluded, though they may remain anchors. Overlaps are resolved in slot order, and unselected candidates remain eligible. Revised anchors can therefore admit previously excluded genes. Slots are revisable search units, not validated mechanisms or fixed categories. Revision or replacement archives the prior version and resets its allocation counts, renewing its exploration bonus while retaining cumulative assay history.

Bounded adaptive quotas. Allocation uses slot-level counts of the strong-effect event $y_\tau(x) = \mathbf{1}[|f(x)| \geq \tau]$, which can differ from benchmark hit membership. A priority heuristic combines each retained version's smoothed event rate with a low-observation bonus. Quotas satisfy

$$\sum_{k=1}^{K} q_{k,t} = b, \qquad q_{\min} \leq q_{k,t} \leq q_{\max}. \tag{2}$$

Each slot receives its floor, then remaining tests follow priority order subject to the cap. These bounds preserve testing opportunities while limiting concentration. Initial quotas are approximately equal. The default $K = 4$, $b = 128$ configuration uses bounds of 16–48.

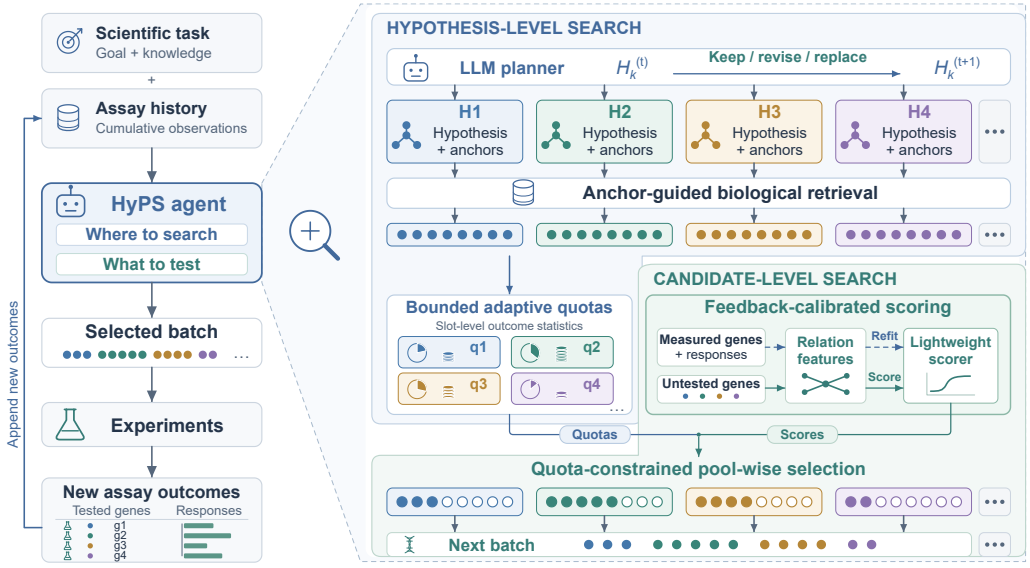


Figure 1: **HyPS's closed-loop discovery framework.** Left: the scientific goal and cumulative assay history inform the agent. Each completed batch appends measured genes and responses to history. Right: the agent expands into two search levels. Hypothesis-level search updates hypothesis–anchor units, retrieves pools, and assigns bounded quotas from slot statistics. H1–H4 identify persistent slots, not preset pathways. From round two, candidate-level search refits one lightweight scorer on all observations via fixed relation features (dashed blue), then scores untested candidates (solid teal). Quotas and scores jointly guide pool-wise selection. Revision retains global history. Round one uses LLM selection. Filled/open circles denote selected/unselected genes. Counts, response bars, and the scorer curve are schematic.

4.3 Candidate-Level Search: What to Test

Feedback-calibrated scoring. From round two, one lightweight scorer is fitted to $\mathcal{D}_{t-1}$, the measured genes and responses from all pools. Robustly normalized response magnitudes propagate through fixed dependency, pathway, and regulatory relations to form standardized features $z_{v,t}(x)$, $v \in \{A, R, S\}$, for measured and untested genes (Dempster et al., 2019; Gillespie et al., 2022; Lo Surdo et al., 2026). Feature construction excludes untested responses and direct self-response. Reactome and SIGNOR affect ranking only, not retrieval or quotas.

Given these features, a ridge-logistic model estimates the strong-effect probability:

$$\hat{p}_t(x) = \sigma\left(\beta_{0,t} + \sum_{v\in\{A,R,S\}} w_{v,t} z_{v,t}(x)\right), \tag{3}$$

where σ is the sigmoid. One intercept and three view weights minimize regularized binary cross-entropy over measured genes. After each batch, features are recomputed and parameters refitted on the expanded history without LLM fine-tuning. Here, feedback-calibrated denotes assay-adapted ranking.

Quota-constrained pool-wise selection. Round one uses pool-wise LLM selection. Thereafter, quotas specify counts and the same fitted scorer determines priorities:

$$\mathcal{S}_{k,t} = \mathrm{Top}_{q_{k,t}}(\mathcal{P}_{k,t}; \hat{p}_t), \qquad \mathcal{B}_t = \bigcup_{k=1}^{K} \mathcal{S}_{k,t}. \tag{4}$$

Ties follow gene-name order. Completed batches add gene–response observations to history for the next round. Equation (4) specifies candidate-only batches. Implementation details for allocation, relation features, response fitting, and the control-reserving variant are collected in Appendices A.1–A.4.

Table 1: **Single-gene hit recall (%, ↑).** All/N/E include all hits or exclude common-essential genes. Panel (a) reports traditional-method scores from BDA. Panel (b) compares same-backbone HyPS and BDA, both with dependency-profile retrieval, over ten five-round runs with batch size 128 (*Scharenberg: 32). Green/red annotations are HyPS/BDA recall ratios, not percentage-point gains. Bold/flames highlight selected configurations, not column-wise maxima. Table 6 in Appendix B.2 gives model sources and price ordering.

Model / Method	**Schmidt1**		**Schmidt2**		**Carnev.**		**Sanchez**		**Scharen.***		**Steinhart**	
(a) Traditional methods	All↑	N/E↑	All↑	N/E↑	All↑	N/E↑	All↑	N/E↑	All↑	N/E↑	All↑	N/E↑
Top Uncertain	5.7	4.2	7.2	5.0	3.7	3.3	3.9	2.8	29.4	23.6	5.8	4.4
🔥**Coreset**	**7.2**	**6.6**	**10.2**	**8.4**	**4.7**	**3.8**	**6.1**	**5.4**	**24.3**	**19.7**	**6.9**	**5.9**
Badge	6.0	5.0	7.7	5.8	4.4	3.6	3.9	3.5	25.8	21.1	4.2	3.8
DiscoBax	8.8	6.9	7.4	5.7	4.5	3.8	5.8	4.9	20.1	20.0	4.7	2.1

(b) BioDiscoveryAgent vs. HyPS, paired by LLM

Model	**Metric**	BDA	**HyPS**	BDA	**HyPS**	BDA	**HyPS**	BDA	**HyPS**	BDA	**HyPS**	BDA	**HyPS**
MiMo-V2.5-Pro	All↑	6.9	8.7↑ 1.26×	10.6	12.8↑ 1.21×	4.4	4.6↑ 1.05×	2.3	4.2↑ 1.83×	4.5	6.3↑ 1.40×	7.7	7.6↓ 0.99×
	N/E↑	7.1	7.7↑ 1.08×	12.7	15.5↑ 1.22×	2.3	2.7↑ 1.17×	2.8	4.9↑ 1.75×	5.7	7.1↑ 1.25×	7.0	7.7↑ 1.10×
DeepSeek-V4-Pro	All↑	5.8	9.0↑ 1.55×	15.9	23.5↑ 1.48×	3.4	5.2↑ 1.53×	3.7	5.8↑ 1.57×	14.0	22.5↑ 1.61×	4.7	6.0↑ 1.28×
	N/E↑	3.0	4.6↑ 1.53×	8.5	13.3↑ 1.56×	2.0	3.1↑ 1.55×	2.0	3.2↑ 1.60×	11.3	17.8↑ 1.58×	3.9	4.7↑ 1.21×
GLM-5.3	All↑	7.5	11.6↑ 1.55×	14.4	16.4↑ 1.14×	4.0	6.3↑ 1.58×	3.8	6.0↑ 1.58×	3.8	4.9↑ 1.29×	4.1	5.3↑ 1.29×
	N/E↑	3.5	5.5↑ 1.57×	7.6	8.9↑ 1.17×	2.0	3.2↑ 1.60×	2.2	3.4↑ 1.55×	2.7	2.6↓ 0.96×	3.6	4.0↑ 1.11×
Claude Haiku 4.5	All↑	7.1	11.2↑ 1.58×	15.0	23.4↑ 1.56×	3.7	5.8↑ 1.57×	3.5	5.6↑ 1.60×	17.8	27.7↑ 1.56×	5.8	7.6↑ 1.31×
	N/E↑	3.8	5.9↑ 1.55×	8.0	12.5↑ 1.56×	1.7	2.6↑ 1.53×	2.4	3.7↑ 1.54×	14.8	23.1↑ 1.56×	5.0	6.4↑ 1.28×
🔥Claude Sonnet 5	All↑	9.1	**14.0**↑ 1.54×	17.8	**27.6**↑ 1.55×	4.0	**6.2**↑ 1.55×	4.6	**7.1**↑ 1.54×	17.8	**27.7**↑ 1.56×	4.9	**6.3**↑ 1.29×
	N/E↑	4.1	**6.4**↑ 1.56×	9.6	**14.9**↑ 1.55×	1.9	**3.0**↑ 1.58×	2.3	**3.6**↑ 1.57×	14.0	**21.8**↑ 1.56×	4.3	**5.3**↑ 1.23×
🔥GPT-5.6 Terra	All↑	8.7	**13.6**↑ 1.56×	17.4	**27.1**↑ 1.56×	4.4	**6.7**↑ 1.52×	3.7	**5.6**↑ 1.51×	14.8	**22.6**↑ 1.53×	5.1	**6.5**↑ 1.27×
	N/E↑	4.3	**6.8**↑ 1.58×	9.4	**14.6**↑ 1.55×	2.1	**3.3**↑ 1.57×	2.4	**3.7**↑ 1.54×	11.4	**17.7**↑ 1.55×	4.0	**4.8**↑ 1.20×
🔥GPT-5.6 Sol	All↑	8.1	**12.5**↑ 1.54×	18.1	**28.1**↑ 1.55×	4.2	**6.5**↑ 1.55×	3.9	**5.4**↑ 1.38×	17.8	**27.7**↑ 1.56×	4.4	**5.4**↑ 1.23×
	N/E↑	4.0	**6.3**↑ 1.58×	9.8	**15.3**↑ 1.56×	2.2	**3.5**↑ 1.59×	1.8	**2.8**↑ 1.56×	14.0	**21.8**↑ 1.56×	3.6	**4.3**↑ 1.19×
Claude Fable 5.1	All↑	7.6	11.8↑ 1.55×	17.5	27.2↑ 1.55×	3.9	6.1↑ 1.56×	4.2	6.5↑ 1.55×	17.4	27.0↑ 1.55×	5.2	6.7↑ 1.29×
	N/E↑	4.0	6.2↑ 1.55×	9.4	14.5↑ 1.54×	1.9	3.0↑ 1.58×	2.1	3.3↑ 1.57×	14.1	21.8↑ 1.55×	4.4	5.5↑ 1.25×
GPT-6 Astra	All↑	8.7	13.5↑ 1.55×	16.6	25.9↑ 1.56×	4.4	6.9↑ 1.57×	3.3	5.2↑ 1.58×	17.0	26.4↑ 1.55×	4.4	5.4↑ 1.23×
	N/E↑	3.8	6.0↑ 1.58×	9.0	14.0↑ 1.56×	2.1	3.3↑ 1.57×	1.5	2.4↑ 1.60×	14.0	21.8↑ 1.56×	3.5	4.1↑ 1.17×

5 Experiments

5.1 Experimental Setup

Datasets and comparators. Six single-gene tasks cover IFNG/IL2 production (Schmidt et al., 2022), adenosine-suppressed T-cell function (Carnevale et al., 2022), tau abundance (Sanchez et al., 2021), lysosomal choline recycling (Scharenberg et al., 2023), and CAR-T proliferation. HyPS and BioDiscoveryAgent (BDA) share nine backbones and use dependency-profile retrieval at matched budgets in all comparisons. Top Uncertain (Gal et al., 2017), Coreset (Sener & Savarese, 2018), BADGE (Ash et al., 2020), and DiscoBAX (Lyle et al., 2023) use scores reported by BDA (Roohani et al., 2025). The two-gene K562 subset of Horlbeck et al. (2018) contains 100,576 pairs and 9,959 target hits, comparing HyPS with BDA across three backbones. Tables 1–2 specify repetitions and budgets. Evaluation reveals stored assay outcomes, with τ derived from the complete benchmark hit set. Dataset mappings, protocol and resource details, and pair-file identifiers are given in Appendices B.1–B.2.

Metrics. All recall measures cumulative recovery of benchmark hits. N/E excludes common-essential genes $\mathcal{C}$, using $\mathcal{H}_{\mathrm{All}} = \mathcal{H}$ and $\mathcal{H}_{\mathrm{N/E}} = \mathcal{H} \setminus \mathcal{C}$:

$$R_c(T) = \frac{|\mathcal{O}_T \cap \mathcal{H}_c|}{|\mathcal{H}_c|}, \qquad c \in \{\mathrm{All}, \mathrm{N/E}\}. \tag{5}$$

Macro averages weight datasets and LLMs equally. Gains use percentage points (pp), except the ratios in Table 1. Pairs use hit counts without N/E filtering. Captions specify uncertainty, with study-specific aggregation in Appendix D.

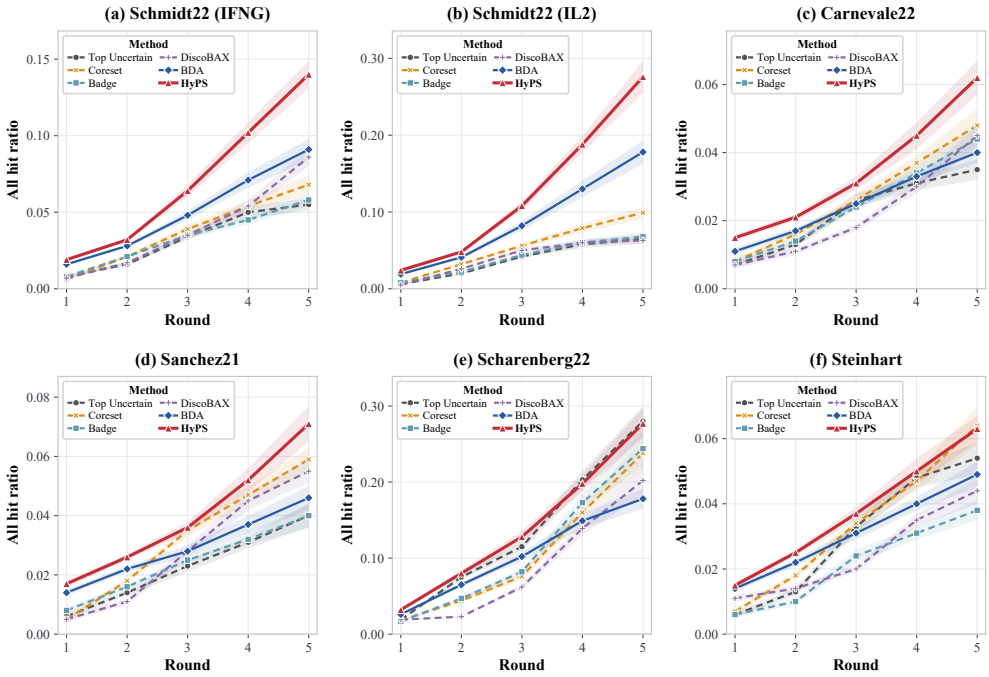


Figure 2: **Five-round discovery trajectories with Claude Sonnet 5.** Both HyPS and BDA use dependency-profile retrieval. Lines show mean cumulative All recall and bands show ± one SD over ten runs. HyPS stays above BDA throughout, with a larger final gap on every dataset. However, Top Uncertain overtakes HyPS on Scharenberg in round four, and Coreset leads on Steinhart at round five. Runs are separate from Table 1.

5.2 Discovery under Matched Budgets

Single-gene discovery. HyPS improves six-dataset mean recall over BDA for every backbone (Table 1). Across all six datasets and nine LLMs, All recall rises from 8.40% to 12.47% and N/E from 5.57% to 8.15%, gaining 4.07/2.58 pp. Each metric improves in 53 of 54 comparisons. The exceptions are MiMo-V2.5-Pro on Steinhart All (7.7% to 7.6%) and GLM-5.3 on Scharenberg N/E (2.7% to 2.6%). Against Coreset, seven of nine HyPS configurations have higher macro All recall and six have higher N/E recall. However, Top Uncertain leads all HyPS configurations on Scharenberg.

Figure 2 also shows that the advantage over BDA is not confined to initial proposals: on IL2, the gap grows from 0.5 pp in round one to 9.8 pp in round five. This supports examining feedback decisions separately, rather than attributing the gains to initialization alone.

Two-gene discovery. Across three equally weighted LLMs, mean recovery increases from 80.6 to 139.0 hit pairs after 640 tests, a 72.4% relative gain (Table 2). Claude Sonnet 5 recovers the most pairs (168.7), whereas GPT-5.6 Sol has the largest relative gain (124.6%). Thus, stronger relative improvement need not imply higher final recovery. Figure 3 shows when these gains emerge.

5.3 Contributions of the Two-Level Design

Decision-level controls. Figure 4 supports adaptive allocation when only the quota rule changes. Feedback scoring improves All for all seven backbones and N/E for six relative to LLM selection. Pool quotas improve All for six and N/E for five, but global selection wins on both metrics for GLM-5.3 and on N/E for GPT-5.6 Terra. In contrast, hypothesis adaptation's opposing All/N/E effects show that greater overall recovery need not improve non-common-essential recovery. Thus, allocation and scoring receive more consistent support.

Table 2: **Two-gene discovery.** Mean cumulative hits (↑) ± SD over ten five-round runs, with 128 pairs per round and dependency-profile retrieval in both agents. Bold marks the higher paired mean. Δ gives R5 additional hits and relative gain over BDA. No N/E filtering applies. Table 7 in Appendix C.1 reports endpoint recall and hit rates.

Model	Method	R1	R2	R3	R4	R5	Δ (R5)
Claude Sonnet 5	BDA	16.3 ± 2.1	34.5 ± 3.3	68.2 ± 4.9	86.7 ± 6.3	116.4 ± 8.1	—
	HyPS	**39.4 ± 2.9**	**83.6 ± 4.8**	**105.3 ± 6.7**	**132.8 ± 8.5**	**168.7 ± 10.8**	**+52.3 (+44.93%)**
GPT-5.6 Sol	BDA	6.2 ± 1.3	18.4 ± 2.1	28.3 ± 2.7	40.5 ± 3.6	45.2 ± 4.2	—
	HyPS	**14.3 ± 1.7**	**34.8 ± 3.0**	**57.4 ± 4.3**	**78.6 ± 5.4**	**101.5 ± 6.9**	**+56.3 (+124.56%)**
GPT-5.6 Terra	BDA	**19.2 ± 2.3**	41.3 ± 3.2	51.6 ± 4.0	64.4 ± 5.1	80.3 ± 6.4	—
	HyPS	16.4 ± 1.8	**44.5 ± 3.7**	**72.3 ± 5.2**	**118.5 ± 7.6**	**146.8 ± 10.1**	**+66.5 (+82.81%)**

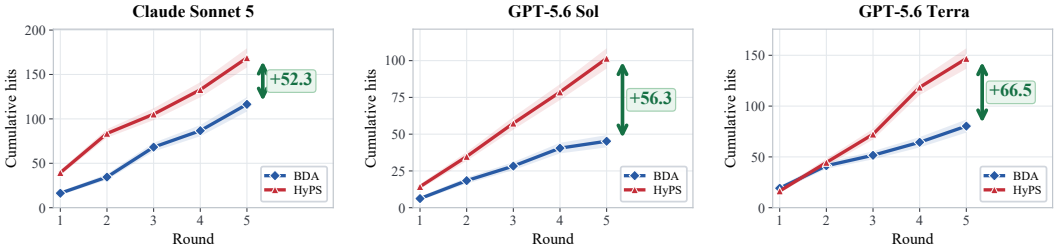


Figure 3: **Two-gene discovery trajectories.** Lines and bands show means and ± one SD from Table 2. HyPS leads from round one with Claude Sonnet 5 and GPT-5.6 Sol, and from round two with GPT-5.6 Terra. Arrows denote additional R5 hits, not recall gains. Table 7 in Appendix C.1 reports endpoint recall and hit rates.

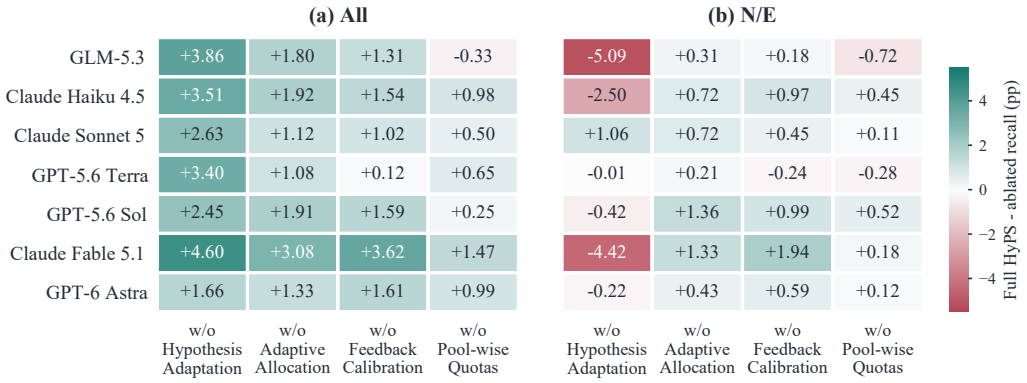


Figure 4: **Decision-level ablations.** Cells show Full HyPS minus control in six-dataset macro recall (pp). Green/red favors HyPS/control. Controls use fixed directions, equal quotas, LLM selection, or global ranking. Adaptation improves All for all backbones but lowers N/E for six, whereas allocation improves both throughout. The hypothesis control freezes the matched initial portfolio. Rows follow price order. Table 8 in Appendix C.2 gives complete scores.

Portfolio size. Figure 5 supports four slots as a practical default at the tested budget. H4 gains 1.54–3.83/0.93–2.29 pp All/N/E over H2 and 0.82–1.10/0.37–0.77 pp over H8. Thus, adding directions alone does not improve discovery. H2/H8 still win individual dataset comparisons, so this aggregate preference does not identify a universal optimum or an exhaustive biological taxonomy.

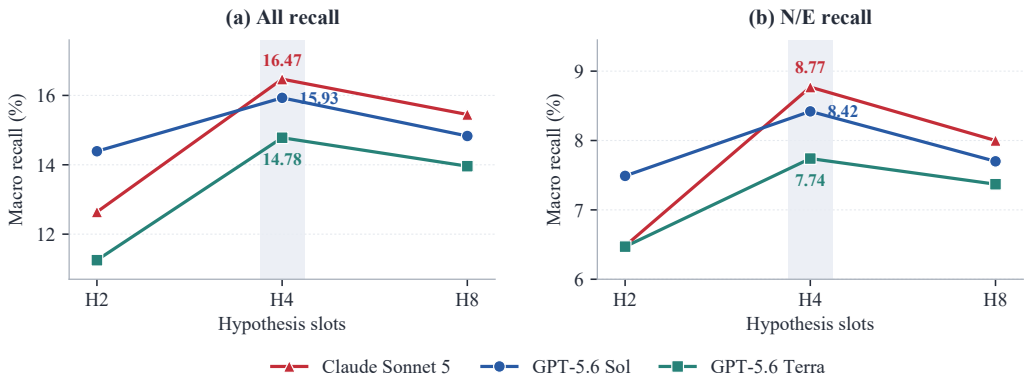


Figure 5: **Sensitivity to portfolio size.** Mean recall over IFNG/IL2/Carnevale from ten five-round runs. H2/H4/H8 use 2/4/8 slots. H4 maximizes both metrics for each backbone. Shading marks H4, not uncertainty. Tables 9–10 in Appendix C.3 give macro and dataset results with SD definitions.

Feedback and relation views. Figure 6 shows larger macro gains from feedback scoring than from jointly adding relation views. Other rules are fixed, although changed selections alter subsequent histories; direct LLM selection still wins Sol's IFNG N/E comparison (51 versus 50 hits).

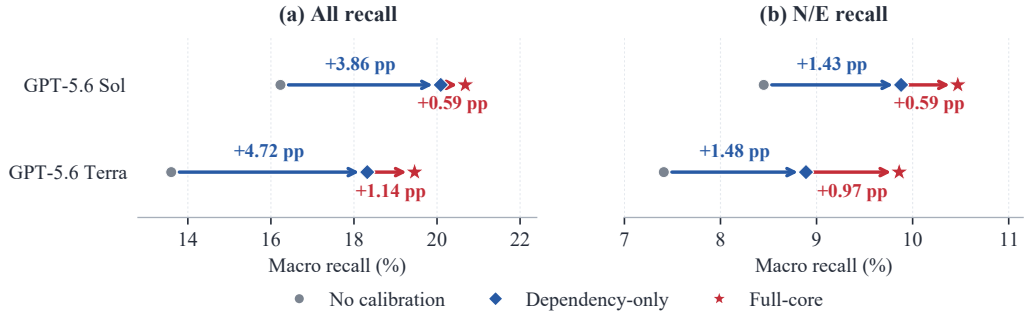


Figure 6: **Feedback scoring and relation views.** Macro recall over IFNG/IL2. Markers denote LLM selection, dependency-only scoring, and scoring with Reactome/SIGNOR added; arrows show successive gains in pp. Tables 11–12 in Appendix C.4 report macro recalls and hit counts.

5.4 Feedback over Longer Searches

Hypothesis construction and revision. Both conditions share four initial hypotheses and anchors; Fixed preserves them, whereas Adaptive may update each slot from assay feedback. At 30 rounds, Adaptive wins all nine dataset–model pairs by 2.71/2.44 pp All/N/E on average (Table 3). A separate five-round study shows smaller mean gains of 0.42/0.14 pp; because its runs are unpaired with the 30-round study, this contrast cannot be attributed to horizon alone (Table 15).

Table 3: **Hypothesis comparison at 30 rounds.** Mean All/N/E recall (%) over IFNG, IL2, and Carnevale under matched initialization. Fixed preserves the initial directions and anchors; Adaptive updates them from feedback. Tables 13–14 in Appendix C.5 give direction definitions and dataset-level results.

Model	Fixed directions	Adaptive hypotheses	Δ (pp)
Claude Sonnet 5	40.32 / 31.39	**43.71 / 34.47**	+3.39 / +3.08
GPT-5.6 Sol	40.69 / 31.62	**43.20 / 33.96**	+2.51 / +2.34
GPT-5.6 Terra	41.17 / 32.28	**43.39 / 34.17**	+2.23 / +1.90

The IL2/Claude Sonnet 5 trace illustrates how revision changes candidate admission: r3 shifts from ubiquitin feedback to RNA processing, admitting 65 of 92 second-round candidates absent from all first-round pools. One, CDC5L, is selected and meets the All-hit criterion. Appendix E provides the complete sequence of hypotheses, allocations, selections, and outcomes.

Complete-framework comparison. Separate thirty-round experiments retain dependency-profile retrieval for both HyPS and BDA. HyPS improves both hit counts in 17 of 18 dataset–model comparisons, tying on Steinhart with Claude Sonnet 5. Mean recall over the three LLMs improves on all six datasets (Table 4). Gains reach 3.98/4.64 pp All/N/E on IL2 but stay below 0.5 pp on Carnevale and Steinhart. The advantage is non-monotonic.

Table 4: **Full-framework comparison at 30 rounds.** Three-LLM mean All/N/E recall (%), with dependency-profile retrieval in both agents. Runs differ from Table 3. Tables 16–17 in Appendices C.7–C.8 give model-wise counts and separately aggregated gain trajectories. Rounding can change displayed differences by 0.01 pp.

Dataset	BDA	HyPS	Reported Δ (pp)
IFNG	37.46 / 29.76	**41.09 / 33.13**	+3.62 / +3.37
IL2	57.34 / 43.59	**61.31 / 48.23**	+3.98 / +4.64
Carnevale	27.08 / 21.70	**27.54 / 22.08**	+0.46 / +0.38
Scharenberg22	87.42 / 87.08	**89.94 / 89.12**	+2.52 / +2.04
Sanchez21	28.32 / 23.86	**31.42 / 26.54**	+3.10 / +2.68
Steinhart GD2	23.94 / 21.84	**24.38 / 22.30**	+0.44 / +0.46

6 Discussion and Limitations

Discussion. The results distinguish two uses of assay feedback: revising which genes enter the search and deciding which receive a test. Adaptive budgets and shared response scoring connect these decisions while preserving prior observations when hypotheses change. Empirically, scoring and allocation receive more consistent support than hypothesis revision, and global selection sometimes outperforms pool quotas. Accordingly, biological structure is useful when it improves access to valuable tests. Future analyses should measure candidate admission and selected-hit recovery separately to explain when this structure helps.

Limitations and future work. The benchmark-derived threshold limits conclusions about deployment with an independently specified criterion. Moreover, the matched fixed-direction control evaluates hypothesis updating as a whole rather than separating keep, revise, and replace decisions. Relation views are also added jointly, and shared scoring lacks a pool-specific control. These effects require separate evaluation. Direct LLMNN/Probability-of-Hit comparisons and total-cost analysis would clarify the scope and efficiency of the gains. Finally, retrospective hit recovery does not validate the proposed mechanisms. Prospective assays are needed to assess biological discovery beyond the benchmark.

7 Conclusion

We introduced HyPS for budget-constrained genetic perturbation discovery by coupling hypothesis planning with feedback-based candidate selection. To this end, revisable hypotheses and anchors define candidate pools with outcome-dependent testing quotas. Within this structure, a shared lightweight scorer combines accumulated assay responses with fixed biological relations to prioritize genes within each pool. Under the benchmark-informed retrospective protocol, HyPS improves average single-gene recall and two-gene hit recovery over the retrieval-augmented comparator at matched budgets. Component studies provide more consistent support for allocation and scoring than for hypothesis adaptation. Overall, these findings support using biological knowledge to structure search and experimental feedback to guide how limited tests are spent.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the writing, translation, and language polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to distinguish the use of LLMs for manuscript preparation from their role as the subject of our research. Regarding the writing process, the writing-assistant LLM was not involved in the ideation, formulation of the research methodology, or theoretical design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the writing-assistant LLM were solely focused on improving the linguistic quality of the paper. The usage of LLMs as the core experimental discovery agents (i.e., the HyPS framework) is extensively documented, evaluated, and disclosed within the technical sections of this paper. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

ETHICS STATEMENT

Our work strictly follows the ethical principles outlined in the ICLR Code of Ethics. As a retrospective evaluation of automated experiment-selection agents, this investigation required no human subjects or live animal experimentation. All datasets used, including the genetic perturbation screens, were sourced from publicly available repositories in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

To facilitate the replication of our findings, we have carefully documented all necessary experimental procedures. Comprehensive details regarding our experimental design—such as agent configurations, scoring modules, and evaluation protocols—are thoroughly outlined in Sections 4 and 5.1 of this manuscript. Furthermore, in-depth documentation of the Hypothesis Portfolio Search (HyPS) framework and its specific implementation steps can be found in the Appendices. The underlying genetic perturbation datasets, including both single-gene and two-gene interaction benchmarks, are derived from open-access sources, which guarantees that our evaluation metrics remain consistent and verifiable. Through these transparent reporting practices, we aim to provide the research community with sufficient information to independently reproduce our study.

REFERENCES

Dhruv Agarwal, Manoj Ghuhan Arivazhagan, Rajarshi Das, Sandesh Swamy, Sopan Khosla, and Rashmi Gangadharaiah. Searching for optimal solutions with LLMs via Bayesian optimization. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 67180–67201, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a79237d6c47feef24c1df723cb41f674-Paper-Conference.pdf.

Anthropic. Introducing Claude Haiku 4.5. Official release page, October 2025. URL https://www.anthropic.com/news/claude-haiku-4-5. Accessed September 24, 2026.

Anthropic. Introducing Claude Fable 5.1 and Claude Mythos 5.1. Official release page, September 2026a. URL https://www.anthropic.com/claude-fable-and-mythos-5-1. Accessed September 24, 2026.

Anthropic. Introducing Claude Sonnet 5. Official release page, June 2026b. URL https://www.anthropic.com/news/claude-sonnet-5. Accessed September 24, 2026.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Julia Carnevale, Eric Shifrut, Nupura Kale, William A. Nyberg, Franziska Blaeschke, Yan Yi Chen, Zhongmei Li, Sagar P. Bapat, Morgan E. Diolaiti, Patrick O'Leary, Shane Vedova, Julia Belk, Bence Daniel, Theodore L. Roth, Stefanie Bachl, Alejandro Allo Anido, Brooke Prinzing, Jorge Ibañez-Vega, Shannon Lange, Dalia Haydar, Marie Luetke-Eversloh, Maelys Born-Bony, Bindu Hegde, Scott Kogan, Tobias Feuchtinger, Hideho Okada, Ansuman T. Satpathy, Kevin Shannon, Stephen Gottschalk, Justin Eyquem, Giedre Krenciute, Alan Ashworth, and Alexander Marson. RASA2 ablation in T cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925): 174–182, 2022. doi: 10.1038/s41586-022-05126-w.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Joshua M. Dempster, Clare Pacini, Sasha Pantel, Fiona M. Behan, Thomas Green, John Krill-Burger, Charlotte M. Beaver, Scott T. Younger, Victor Zhivich, Hanna Najgebauer, Felicity Allen, Emanuel Gonçalves, Rebecca Shepherd, John G. Doench, Kosuke Yusa, Francisca Vazquez, Leopold Parts, Jesse S. Boehm, Todd R. Golub, William C. Hahn, David E. Root, Mathew J. Garnett, Aviad Tsherniak, and Francesco Iorio. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. *Nature Communications*, 10(1):5817, 2019. doi: 10.1038/s41467-019-13805-y.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192. PMLR, 2017.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022. doi: 10.1093/nar/gkab1028.

Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15482–15510, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.838. URL https://aclanthology.org/2025.findings-emnlp.838/.

Max A. Horlbeck, Albert Xu, Min Wang, Neal K. Bennett, Chong Y. Park, Derek Bogdanoff, Britt Adamson, Eric D. Chow, Martin Kampmann, Tim R. Peterson, Ken Nakamura, Michael A. Fischbach, Jonathan S. Weissman, and Luke A. Gilbert. Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967.e22, 2018. doi: 10.1016/j.cell.2018.06.010.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 120124–120158. Curran Associates, Inc., 2024. doi: 10.52202/079017-3818. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d951f73c521d069fefbb73396df01424-Paper-Conference.pdf.

Yanke Li, Tianyu Cui, Tommaso Mansi, Mangal Prakash, and Rui Liao. BioBO: Biology-informed Bayesian optimization for perturbation design. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/feafa00364bf93528a4ea72ebebdb932-Paper-Conference.pdf.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/84b8d9fcb4e262fcd429544697e1e720-Abstract-Conference.html.

Prisca Lo Surdo, Marta Iannuccelli, Klas Karis, Eleonora Meo, Parnian Omidi, Marta Tosoni, Simone Graziosi, Simona Panni, Maria Teresa Fiorillo, Luana Licata, Francesca Sacco, Benjamin M. Gyori, and Livia Perfetto. SIGNOR 4.0: the 2025 update with focus on phosphorylation data. *Nucleic Acids Research*, 54(D1):D682–D690, 2026. doi: 10.1093/nar/gkaf1237.

Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jesson, Stefan Bauer, Yarin Gal, and Patrick Schwab. DiscoBAX: Discovery of optimal intervention sets in genomic experiment design. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23170–23189. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lyle23a.html.

Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. GeneDisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-w2oomO6qgc.

OpenAI. GPT-5.6 Sol model. OpenAI API documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 24, 2026.

OpenAI. GPT-5.6 Terra model. OpenAI API documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Accessed September 24, 2026.

OpenAI. GPT-6 Astra model. OpenAI API documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 24, 2026.

Cong Qi, Wenbo Wang, Siqi Jiang, Qin Liu, Xun Song, Hanzhang Fang, and Zhi Wei. Artificial intelligence agents for biological research: a survey. *Briefings in Bioinformatics*, 27(1):bbag075, 2026. doi: 10.1093/bib/bbag075. URL https://doi.org/10.1093/bib/bbag075.

Bojana Ranković, Ryan-Rhys Griffiths, and Philippe Schwaller. Large language models as uncertainty-calibrated optimizers for experimental discovery. *Nature Machine Intelligence*, 8: 1466–1477, August 2026. doi: 10.1038/s42256-026-01283-z.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HAwZGLcye3.

Andrea Rubbi, Arpit Merchant, Samuel Ogden, Amir Akbarnejad, Pietro Liò, Sattar Vakili, and Mo Lotfollahi. Many needles in a haystack: Active hit discovery for perturbation experiments. arXiv preprint arXiv:2605.10196, 2026. URL https://arxiv.org/abs/2605.10196.

Carlos G. Sanchez, Christopher M. Acker, Audrey Gray, Malini Varadarajan, Cheng Song, Nadire R. Cochran, Steven Paula, Alicia Lindeman, Shaojian An, Gregory McAllister, John Alford, John

Reece-Hoyes, Carsten Russ, Lucas Craig, Ketthsy Capre, Christian Doherty, Gregory R. Hoffman, Sarah J. Luchansky, Manuela Polydoro, Ricardo Dolmetsch, and Fiona Elwood. Genome-wide CRISPR screen identifies protein pathways modulating tau protein levels in neurons. *Communications Biology*, 4(1):736, 2021. doi: 10.1038/s42003-021-02272-1.

Samantha G. Scharenberg, Wentao Dong, Ali Ghoochani, Kwamina Nyame, Roni Levin-Konigsberg, Aswini R. Krishnan, Eshaan S. Rawat, Kaitlyn Spees, Michael C. Bassik, and Monther Abu-Remaileh. An SPNS1-dependent lysosomal lipid transport pathway that enables cell survival under choline limitation. *Science Advances*, 9(16):eadf8966, 2023. doi: 10.1126/sciadv.adf8966.

Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W. Freimer, Raymund Bueno, Vinh Q. Nguyen, Franziska Blaeschke, Chun Jimmie Ye, and Alexander Marson. CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science*, 375(6580): eabj4008, 2022. doi: 10.1126/science.abj4008.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016. doi: 10.1109/JPROC.2015.2494218. URL https://doi.org/10.1109/JPROC.2015.2494218.

Yixuan Weng, Minjun Zhu, Qiujie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. DeepScientist: Advancing frontier-pushing scientific findings progressively. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 47981–48037, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4f64494ecc3442f1c9261baa036378bc-Paper-Conference.pdf.

Xiaomi MiMo Team. MiMo-V2.5-Pro. Official model card, 2026. URL https://huggingface.co/XiaomiMiMo/MiMo-V2.5-Pro. Accessed September 24, 2026.

Z.ai. GLM-5.3. Official model documentation, 2026. URL https://docs.z.ai/guides/llm/glm-5.3. Accessed September 24, 2026.

APPENDIX

CONTENTS

A **Method Details** 15
A.1 Hypothesis Portfolio and Budget Allocation 15
A.2 Biological Relation Operators 15
A.3 Response Features and Parameter Estimation 15
A.4 Prompts and Hypothesis Traces 16
B **Experimental Setup** 17
B.1 Datasets and Evaluation Protocols 17
B.2 Model Sources and Reported Token Prices 18
C **Additional Experimental Results** 19
C.1 Two-Gene Discovery: Endpoint Metrics 19
C.2 Module Controls and Complete Scores 19
C.3 Portfolio-Size Results 20
C.4 Feedback and Relation-View Results 21
C.5 Adaptive Hypotheses at Thirty Rounds 22
C.6 Earlier Five-Round Hypothesis Comparison 23
C.7 Thirty-Round Framework Results by Model 24
C.8 Long-Horizon Gain Trajectories 24
D **Implementation and Statistical Details** 25
E **Complete Agent Trace** 26

A METHOD DETAILS

A.1 HYPOTHESIS PORTFOLIO AND BUDGET ALLOCATION

The standard configuration has four slots and three anchors per slot. Slot identifiers persist, but their hypotheses and anchors can change and are not permanently assigned to four biological categories. Tested genes are excluded from pools and overlap is resolved in slot order. Refinement or replacement resets a slot's allocation counts but retains its observations in the shared response history. The first round uses LLM selection, followed by the response model in Section 4.3 for later rounds.

Let $n_{k,t}$ and $m_{k,t}$ be the retained numbers of tests and strong-effect events for slot k, and $N_t = \sum_j n_{j,t}$. Because revision resets these counts, N_t need not equal the number of all completed assays. The allocation priority is

$$v_{k,t} = \frac{m_{k,t}+1}{n_{k,t}+2} + \alpha\sqrt{\frac{\log(N_t+2)}{n_{k,t}+1}}, \qquad \alpha = 0.35. \tag{6}$$

The first term smooths the observed event rate, while the second favors less-tested slots. The resulting priority is a budget-allocation heuristic, not a probability of hypothesis validity. Each slot receives its minimum quota. Remaining places are assigned one at a time by priority, excluding slots at the cap. Ties favor the smaller current quota. Bounds are 16–48 for $K = 4$, $b = 128$. Other configuration details are discussed in Appendix D.

A.2 BIOLOGICAL RELATION OPERATORS

The dependency-profile view uses nonnegative cosine similarity over Achilles/DepMap profiles (Dempster et al., 2019). The pathway view aggregates shared Reactome membership while downweighting large pathways (Gillespie et al., 2022). The regulatory-neighborhood view uses degree-normalized SIGNOR adjacency after duplicate and self-loop removal and symmetrization (Lo Surdo et al., 2026). Only records annotated as up- or down-regulation with format-valid PubMed identifiers are retained, without an additional filter on the `DIRECT` field. Edge direction and regulatory sign are not retained. These operations define relation evidence rather than causal-effect estimates.

A.3 RESPONSE FEATURES AND PARAMETER ESTIMATION

Signal construction. For each observed gene $g \in \mathcal{O}_{t-1}$, the magnitude of its response is normalized using the historical median M_t and median absolute deviation MAD_t:

$$s_t(g) = \mathrm{clip}\left(\frac{|f(g)| - M_t}{\max(1.4826\,\mathrm{MAD}_t, 0.05)}, -3, 3\right). \tag{7}$$

Positive and negative signals indicate effects above and below the historical median, not molecular activation and inhibition. The bounded signal incorporates relatively weak as well as strong observations. For each fixed relation view $v \in \{A, R, S\}$, the response feature is

$$\begin{aligned} r_{v,t}(x) &= \sum_{\substack{g\in\mathcal{O}_{t-1}\\ g\neq x}} \kappa_v(x,g)s_t(g), \\ z_{v,t}(x) &= \frac{r_{v,t}(x) - \mu_{v,t}}{\max(\sigma_{v,t}, 0.01)}. \end{aligned} \tag{8}$$

Here κ_v is the relation operator in Appendix A.2, and $\mu_{v,t}$ and $\sigma_{v,t}$ are historical normalization statistics. All observed genes can contribute regardless of their original hypothesis slot. Untested outcomes and a tested gene's direct self-response are excluded. Reactome and SIGNOR contribute only to ranking, not retrieval or quotas.

Parameter estimation. The probability model in Eq. (3) uses the sigmoid function. With $\boldsymbol{w}_t = [\beta_{0,t}, w_{A,t}, w_{R,t}, w_{S,t}]^\top$, the parameters minimize

$$\boldsymbol{w}_t = \arg\min_{\boldsymbol{w}} \left[\sum_{g\in\mathcal{O}_{t-1}} \mathrm{BCE}\left(y_\tau(g), p_{\boldsymbol{w},t}(g)\right) + \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t^0\|_\Lambda^2\right], \quad \Lambda = \mathrm{diag}(1,3,3,3). \tag{9}$$

Here BCE is binary cross-entropy, $p_{\boldsymbol{w},t}$ is Eq. (3) with parameters $\boldsymbol{w}$, and $\|u\|_\Lambda^2 = u^\top \Lambda u$. The reference vector $\boldsymbol{w}_t^0$ uses the historical event rate for its intercept and zero view weights. Regularization limits large coefficients when feedback is scarce. Each round recomputes features and estimates four parameters from the full history, keeping the LLM parameters and relation resources fixed.

The scorer supports assay-adapted candidate ranking. Probability calibration has not been evaluated separately.

The reported evaluation derives τ from the complete benchmark hit set by taking its minimum absolute score as the default. Although untested responses are excluded from the features, this threshold summarizes information beyond completed assays and affects both allocation events and scorer labels. The results therefore concern a benchmark-informed retrospective protocol, not label-blind deployment. Threshold values are stored with the run configurations; the retained IL2 trace uses $\tau = 0.32403$. BDA does not use HyPS scorer labels, and both methods are evaluated against the same benchmark hit set.

A.4 Prompts and Hypothesis Traces

Appendix E presents five rounds from an IL2/Claude Sonnet 5 run: hypotheses, candidate lists, selected indices, and measured responses. Prompt–assistant excerpts cover initial planning, first-round r3 selection, and feedback-based replanning. Repeated blocks are marked as omitted. The planner's biological explanations are model outputs and have not been experimentally validated.

In round two, 65 of the 92 genes in the r3 pool were absent from every first-round pool. One newly admitted gene, CDC5L, is selected and has a measured score of $+0.56872$, meeting the All-hit criterion but not the N/E criterion. At the same time, r3's quota increases from 32 to 48 and selection switches from the LLM to the response scorer. The example illustrates how the search changes, but does not isolate the contribution of hypothesis revision.

This run reserves two control tests per slot outside the candidate pools. Thus, r3 uses 30 candidate tests and two controls in round one, then 46 candidate tests and two controls in round two. Its quotas include both, unlike the candidate-only batches in Eq. (4). The trace is the median-All-hit run among three completed seeds, not a ten-run average. It illustrates the same core decision process used by the aggregate experiments, whose reported statistics are computed from separate repeated runs.

B EXPERIMENTAL SETUP

B.1 DATASETS AND EVALUATION PROTOCOLS

Retrospective protocol. Evaluations reveal stored assay outcomes after each selected batch. HyPS's internal threshold τ is derived from the complete benchmark hit set, while scoring features use only responses observed in completed rounds. Appendix A.3 specifies threshold construction, and Appendix D records implementation, resource, and run information.

Retrieval-enabled comparator. Throughout the experiments, BDA denotes BioDiscoveryAgent with biological-knowledge retrieval enabled, including Achilles dependency-profile retrieval. Both HyPS and BDA use this resource in the five-round single-gene, two-gene, and thirty-round comparisons, with matched LLM backbones and assay budgets. Each agent makes its own selections and accumulates its own observed history. The within-HyPS ablations instead change the named decision component or relation views, as specified in their captions.

Dataset mapping. The CAR-T task and 100,576-pair K562 subset follow BioDiscoveryAgent (Roohani et al., 2025). Schmidt1 and Schmidt2 denote the IFNG and IL2 screens from Schmidt et al. (2022), identified by their biological readout names in the ablations. Carnev. denotes the adenosine-condition task from Carnevale et al. (2022), and Sanchez denotes the tau-regulation screen from Sanchez et al. (2021). Scharen. is the 1,061-perturbation benchmark from Scharenberg et al. (2023), internally named `Scharenberg22`. Steinhart, Steinhart GD2, and GD2 D22 identify the same CAR-T proliferation dataset.

The original genetic-interaction maps evaluate 222,784 ordered gene pairs (472×472) per cell line (Horlbeck et al., 2018). The K562 benchmark used here is defined directly by the supplied manifest and contains all $\binom{449}{2} = 100{,}576$ unordered pairs of 449 genes, without self-pairs or duplicates. Evaluation uses ten five-round runs of 128 pairs per batch. Appendix C.1 gives the endpoint metrics after 640 joint interventions.

Pair data and hit definition. The file `ground_truth_Horlbeck.csv` contains two fields: `Gene_pairs`, a tuple of gene symbols in lexicographic order, and `Score`, the stored response used by the retrospective evaluator. The file `topmovers_Horlbeck.npy` lists 9,959 distinct hit pairs, all present in the candidate set. Hit recovery is defined by membership in this list, without N/E filtering. Thus, reproducing the reported metric requires the supplied pair universe and hit list rather than reconstruction of the upstream screen.

The hit set includes 5,026 positive-score and 4,933 negative-score pairs. A single absolute-score cutoff cannot reproduce it, since some non-hits have larger absolute scores than some hits. Pair evaluation therefore treats the supplied hit list as the benchmark definition rather than applying the single-gene event label y_τ.

Table 5 lists the exact files and their hashes. The supplementary implementation provides the pair-construction, retrieval, ranking, and per-round logging pipeline used for both BDA and HyPS, together with the five-round, batch-128 run configurations.

Table 5: **Two-gene benchmark files.** Candidate and hit pairs are unique, canonically ordered tuples. SHA-256 hashes identify the exact files used by the reported runs.

File	Contents
`ground_truth_Horlbeck.csv`	100,576 pairs with `Gene_pairs` and `Score`
`C54185BA5B417353480DC322B51AC786727983298F5FE115D1341D3C6E90F50E`	
`topmovers_Horlbeck.npy`	9,959 hit pairs, all in the candidate file
`042B7D95BA28C9B5DB96069658F09698FDCF63BA09456A4326EA2CCBE1C4129A`	

The single-gene evaluation uses five batches of 128 untested genes, except for Scharenberg's batches of 32. Recall uses the fixed hit sets in Eq. (5). Macro averages weight datasets and then backbones equally, without pooling hit counts across datasets.

All and N/E recall have different denominators, so N/E may exceed All. Neither measures biological novelty. Table 1 annotates the multiplier $R_{\mathrm{HyPS}}/R_{\mathrm{BDA}}$, rounded from displayed recalls. Paired agents share the backbone, retrieval resource, and assay budget, but follow their own experimental trajectories. Component-study scores are analyzed separately.

B.2 Model Sources and Reported Token Prices

Table 6 provides an official technical report, model card, documentation, or release page for each backbone, alongside the reported output-token prices used to order Table 1. Total cost also depends on token usage and tool charges, so the ordering is not a cost-effectiveness comparison.

Table 6: **Model sources and reported output-token prices.** Source links identify the evaluated model names and were accessed on September 24, 2026. Prices are the author-reported USD per million output tokens used to order Table 1, not measured experiment costs. Equal-price models retain their listed order.

Model	Official source	USD / 1M output tokens
MiMo-V2.5-Pro	Model card (Xiaomi MiMo Team, 2026)	0.87
DeepSeek-V4-Pro	Technical report (DeepSeek-AI, 2026)	1.98
GLM-5.3	Model documentation (Z.ai, 2026)	4.40
Claude Haiku 4.5	Release page (Anthropic, 2025)	5.00
Claude Sonnet 5	Release page (Anthropic, 2026b)	10.00
GPT-5.6 Terra	Model documentation (OpenAI, 2026b)	12.00
GPT-5.6 Sol	Model documentation (OpenAI, 2026a)	20.00
Claude Fable 5.1	Release page (Anthropic, 2026a)	50.00
GPT-6 Astra	Model documentation (OpenAI, 2026c)	50.00

C Additional Experimental Results

C.1 Two-Gene Discovery: Endpoint Metrics

Table 7 summarizes round-five recovery. Relative gain divides the additional mean hits by the BDA mean. GT recall divides recovered hits by 9,959 target pairs, whereas the experimental hit rate divides them by 640 tests. Round-wise results use mean ± SD over ten runs. N/E filtering is not applied, and no significance test is reported.

Averaging the three LLMs equally gives 80.6333 hits for BDA and 139.0 for HyPS. The increase in mean hits is $(139.0/80.6333 - 1) \times 100 \approx 72.4\%$, computed before rounding the displayed means. This differs from averaging the three model-specific relative gains.

Table 7: **Endpoint metrics for two-gene discovery.** Mean hits after five rounds and gains over BDA, with dependency-profile retrieval enabled in both HyPS and BDA. GT recall uses 9,959 target pairs, whereas hit rate uses 640 tests. Both are percentages, with arrows indicating BDA → HyPS. N/E does not apply. Data provenance is discussed in Appendix B.1.

Model	BDA hits	HyPS hits	Δ hits	Relative gain	GT recall (%) BDA → HyPS	Hit rate (%) BDA → HyPS
Claude Sonnet 5	116.4	**168.7**	+52.3	+44.93%	1.169 → **1.694**	18.19 → **26.36**
GPT-5.6 Sol	45.2	**101.5**	+56.3	+124.56%	0.454 → **1.019**	7.06 → **15.86**
GPT-5.6 Terra	80.3	**146.8**	+66.5	+82.81%	0.806 → **1.474**	12.55 → **22.94**

C.2 Module Controls and Complete Scores

Table 8 gives six-dataset, equal-weight All / N/E recall percentages for the controls in Section 5.3. Direct selection removes feedback scoring, whereas global selection retains the scorer but removes pool quotas. Within-table differences are percentage points, not relative gains.

The equal-allocation control differs from Full HyPS only in the quota-allocation rule. All other experimental settings, including candidate-pool construction and requested pool sizes, are held constant. The fixed-direction control likewise shares Full HyPS's initial hypotheses and anchors but disables subsequent updates. It therefore evaluates hypothesis adaptation as a compound decision rather than isolating keep, revise, and replace operations.

The module study uses the same assay budget as Table 1 but is a separate repeated evaluation, so its Full HyPS estimates need not equal the main-table estimates (e.g., Claude Sonnet 5: 11.73 / 6.15 versus 14.82 / 9.17). We therefore interpret each control relative to Full HyPS within this study rather than as an arithmetic decomposition of the main-table gains.

Table 8: **Overall module ablation.** Equal-weight All/N/E recall (%) over six datasets. Bold marks the Full HyPS values only, not per-row maxima. Rows follow output-token price. Equal allocation changes only the quota rule. The fixed-direction control shares Full HyPS's initial portfolio and disables subsequent hypothesis updates.

Backbone	w/o Hypothesis Adaptation	w/o Adaptive Allocation	w/o Feedback Calibration	w/o Pool-wise Quotas	Full HyPS
GLM-5.3	9.42 / 11.78	11.48 / 6.38	11.97 / 6.51	13.61 / 7.41	**13.28 / 6.69**
Claude Haiku 4.5	6.73 / 8.22	8.32 / 5.00	8.70 / 4.75	9.26 / 5.27	**10.24 / 5.72**
Claude Sonnet 5	9.10 / 5.09	10.61 / 5.43	10.71 / 5.70	11.23 / 6.04	**11.73 / 6.15**
GPT-5.6 Terra	10.19 / 8.19	12.51 / 7.97	13.47 / 8.42	12.94 / 8.46	**13.59 / 8.18**
GPT-5.6 Sol	11.41 / 9.01	11.95 / 7.23	12.27 / 7.60	13.61 / 8.07	**13.86 / 8.59**
Claude Fable 5.1	10.41 / 12.31	11.93 / 6.56	11.39 / 5.95	13.54 / 7.71	**15.01 / 7.89**
GPT-6 Astra	11.80 / 8.40	12.13 / 7.75	11.85 / 7.59	12.47 / 8.06	**13.46 / 8.18**

C.3 PORTFOLIO-SIZE RESULTS

Table 10 gives H2/H4/H8 results for nine dataset–model combinations, each with ten five-round runs. Table 9 averages these recalls equally across datasets.

Dataset-level ± terms denote SDs over ten runs. In Table 9, they instead average the three dataset SDs and do not measure uncertainty in macro recall. A run-level macro SD requires the joint results across datasets, including their covariance.

On Carnevale, H4 has the highest All recall for Claude Sonnet 5 and GPT-5.6 Sol, whereas GPT-5.6 Terra favors H8. H2 has the highest N/E recall there for GPT-5.6 Sol and GPT-5.6 Terra. GPT-5.6 Terra also favors H8 on both IL2 metrics. Thus, H4 is an aggregate default, not a dataset-wise optimum.

Table 9: **Portfolio-size ablation.** Macro recall (%) over IFNG, IL2, and Carnevale, each evaluated with ten runs of five rounds. H2/H4/H8 use $K = 2/4/8$ slots. Bold marks the highest mean per model and metric. The ± terms are arithmetic averages of dataset-level standard deviations (SDs), not SDs of run-level macro recall. Table 10 reports the dataset-level results.

Model	Metric	H2 ($K = 2$)	H4 ($K = 4$)	H8 ($K = 8$)
Claude Sonnet 5	All	12.64 ± 0.82	**16.47** ± 0.66	15.45 ± 0.55
	N/E	6.48 ± 0.43	**8.77** ± 0.46	8.00 ± 0.35
GPT-5.6 Sol	All	14.39 ± 0.79	**15.93** ± 0.91	14.83 ± 0.82
	N/E	7.49 ± 0.52	**8.42** ± 0.56	7.70 ± 0.51
GPT-5.6 Terra	All	11.25 ± 0.67	**14.78** ± 0.72	13.96 ± 0.73
	N/E	6.47 ± 0.42	**7.74** ± 0.38	7.37 ± 0.40

Table 10: **Dataset-level portfolio-size ablation.** Endpoint recall (%), reported as mean ± standard deviation (SD) over ten runs of five rounds. Bold marks the highest mean per row, including settings where H2 or H8 exceeds H4. Values are recall percentages, not percentage-point gains.

Dataset	Model	Metric	H2 ($K = 2$)	H4 ($K = 4$)	H8 ($K = 8$)
IFNG	Claude Sonnet 5	All	11.23 ± 0.72	**14.82** ± 0.33	13.41 ± 0.88
		N/E	5.36 ± 0.38	**7.56** ± 0.40	6.98 ± 0.54
	GPT-5.6 Sol	All	10.80 ± 0.78	**13.80** ± 1.04	12.36 ± 0.82
		N/E	5.74 ± 0.42	**7.07** ± 0.69	6.19 ± 0.45
	GPT-5.6 Terra	All	2.93 ± 0.22	**15.00** ± 0.75	7.50 ± 0.48
		N/E	3.37 ± 0.25	**7.48** ± 0.38	4.24 ± 0.28
IL2	Claude Sonnet 5	All	20.85 ± 1.35	**27.93** ± 1.23	26.40 ± 0.62
		N/E	11.16 ± 0.72	**15.18** ± 0.75	14.00 ± 0.12
	GPT-5.6 Sol	All	26.96 ± 1.24	**27.57** ± 1.30	25.79 ± 1.22
		N/E	13.65 ± 0.94	**15.25** ± 0.79	14.00 ± 0.85
	GPT-5.6 Terra	All	25.08 ± 1.42	24.77 ± 1.15	**27.06** ± 1.28
		N/E	12.47 ± 0.78	13.10 ± 0.62	**14.76** ± 0.71
Carnevale	Claude Sonnet 5	All	5.83 ± 0.38	**6.65** ± 0.42	6.54 ± 0.16
		N/E	2.92 ± 0.18	**3.57** ± 0.23	3.03 ± 0.40
	GPT-5.6 Sol	All	5.41 ± 0.36	**6.42** ± 0.39	6.33 ± 0.42
		N/E	**3.07** ± 0.21	2.94 ± 0.20	2.92 ± 0.22
	GPT-5.6 Terra	All	5.73 ± 0.37	4.56 ± 0.25	**7.32** ± 0.44
		N/E	**3.57** ± 0.23	2.65 ± 0.15	3.11 ± 0.20

C.4 Feedback and Relation-View Results

Table 12 gives mean All and N/E hit counts with sample SDs over ten runs for GPT-5.6 Sol and GPT-5.6 Terra on IFNG and IL2.

All/N/E hit-set sizes are 920/802 for IFNG and 654/481 for IL2. We convert each mean count to recall using its corresponding denominator, then average the two datasets equally. Table 11 and Figure 6 round these macro recalls to two decimals, with increments computed by subtracting the displayed percentages. No macro SD or significance test is reported.

Full-core exceeds Dependency-only in every row, but No calibration recovers more N/E hits for GPT-5.6 Sol on IFNG. The two added views are evaluated jointly, and different selections can change subsequent histories and pools.

Table 11: **Feedback-based scoring and relation views.** Equal-weight macro recall (%) over IFNG and IL2, computed from Table 12. No calibration uses direct LLM selection, whereas Dependency-only and Full-core use one and three relation views. Increments subtract displayed macro percentages (pp). Bold marks the highest value per model and metric.

	No calibration		Dependency-only		Full-core	
Model	All	N/E	All	N/E	All	N/E
GPT-5.6 Sol	16.23	8.45	20.09	9.88	**20.68**	**10.47**
Increment (pp)	—	—	+3.86	+1.43	+0.59	+0.59
GPT-5.6 Terra	13.60	7.41	18.32	8.89	**19.46**	**9.86**
Increment (pp)	—	—	+4.72	+1.48	+1.14	+0.97

Table 12: **Feedback-based scoring and relation-view counts.** Mean recovered hits ± sample SD over ten runs. No calibration uses direct LLM selection, whereas Dependency-only and Full-core use one and three relation views. Bold marks the highest mean per row.

Dataset	Model	Metric	No calibration	Dependency-only	Full-core
IFNG	GPT-5.6 Sol	All	112.00 ± 5.29	119.30 ± 4.95	**122.70 ± 7.30**
		N/E	**51.00 ± 3.56**	47.30 ± 2.91	50.00 ± 5.16
IFNG	GPT-5.6 Terra	All	86.00 ± 7.85	106.00 ± 6.41	**119.30 ± 7.66**
		N/E	39.30 ± 3.68	44.30 ± 1.89	**50.30 ± 3.68**
IL2	GPT-5.6 Sol	All	132.70 ± 7.65	178.00 ± 5.68	**183.30 ± 2.79**
		N/E	50.70 ± 4.81	66.70 ± 3.59	**70.70 ± 2.58**
IL2	GPT-5.6 Terra	All	116.70 ± 4.92	164.30 ± 8.74	**169.70 ± 7.65**
		N/E	47.70 ± 4.30	59.00 ± 2.94	**64.70 ± 2.87**

C.5 Adaptive Hypotheses at Thirty Rounds

Both conditions receive the same first-round hypotheses and anchors, organized by the four initial directions in Table 13. The Fixed condition preserves these hypothesis–anchor units for all 30 rounds. The Adaptive condition starts identically but may keep, revise, or replace each unit after each batch using accumulated assay feedback. Consequently, this comparison changes only whether the initial portfolio can be updated. The listed scopes define a controlled initialization rather than permanent categories: an adaptive slot may move beyond its initial scope after revision or replacement. Table 14 reports endpoint counts, recalls, and differences in percentage points.

Table 13: **Shared initial directions for the hypothesis-update controls.** Both conditions begin with these directions and the same first-round hypotheses and anchors. The Fixed condition preserves them throughout, whereas the Adaptive condition may keep, revise, or replace each hypothesis–anchor unit after observing assay feedback.

Slot	Role	Initial exploration direction
r1	exploit	receptor-proximal and intracellular signaling
r2	expand	transcriptional and epigenetic regulation
r3	challenge	metabolism, stress, and protein homeostasis
r4	discover	membrane trafficking, cytoskeleton, and organelles

Table 14: **Fixed directions versus adaptive hypotheses after 30 rounds.** Entries are recovered hits, with recall (%) in parentheses. Both conditions share the first-round directions, hypotheses, and anchors. Fixed preserves them; Adaptive permits feedback-driven updates. Differences use the displayed endpoints.

Dataset	Model	Metric	Fixed	Adaptive	Δ hits	Δ pp
IFNG	Claude Sonnet 5	All	351 (38.15)	**383 (41.63)**	+32	+3.48
		N/E	242 (30.17)	**268 (33.42)**	+26	+3.25
IFNG	GPT-5.6 Sol	All	348 (37.83)	**377 (40.98)**	+29	+3.15
		N/E	239 (29.80)	**263 (32.79)**	+24	+2.99
IFNG	GPT-5.6 Terra	All	351 (38.15)	**386 (41.96)**	+35	+3.81
		N/E	244 (30.42)	**272 (33.92)**	+28	+3.50
IL2	Claude Sonnet 5	All	380 (58.10)	**403 (61.62)**	+23	+3.52
		N/E	217 (45.11)	**232 (48.23)**	+15	+3.12
IL2	GPT-5.6 Sol	All	381 (58.26)	**402 (61.47)**	+21	+3.21
		N/E	216 (44.91)	**231 (48.02)**	+15	+3.11
IL2	GPT-5.6 Terra	All	389 (59.48)	**398 (60.86)**	+9	+1.38
		N/E	223 (46.36)	**228 (47.40)**	+5	+1.04
Carnevale	Claude Sonnet 5	All	233 (24.71)	**263 (27.89)**	+30	+3.18
		N/E	164 (18.89)	**189 (21.77)**	+25	+2.88
Carnevale	GPT-5.6 Sol	All	245 (25.98)	**256 (27.15)**	+11	+1.17
		N/E	175 (20.16)	**183 (21.08)**	+8	+0.92
Carnevale	GPT-5.6 Terra	All	244 (25.87)	**258 (27.36)**	+14	+1.49
		N/E	174 (20.05)	**184 (21.20)**	+10	+1.15

C.6 Earlier Five-Round Hypothesis Comparison

In the earlier five-round study on Carnevale, IFNG, and IL2 with three models, both conditions use the same initial portfolio defined in Table 13. After correcting the condition labels, Adaptive has higher across-model macro recall than Fixed by 0.42/0.14 pp All/N/E (Table 15), although the per-model effects vary. The five- and thirty-round studies use separate, unpaired run groups, so their difference cannot be attributed to horizon alone.

Table 15: **Five-round hypothesis comparison.** Macro All/N/E recall (%) over Carnevale, IFNG, and IL2. Both conditions share the initial portfolio; only Adaptive permits feedback-driven updates. Changes are Adaptive minus Fixed directions in pp. Runs are separate from the 30-round comparison.

Model	Fixed directions	Adaptive	Δ (pp)
Claude Sonnet 5	15.37 / 7.79	**16.85 / 8.05**	$+1.48$ / $+0.26$
GPT-5.6 Sol	15.87 / 7.81	**16.17 / 8.20**	$+0.30$ / $+0.39$
GPT-5.6 Terra	**15.09 / 7.48**	14.55 / 7.25	-0.54 / -0.23
Across-model mean	15.44 / 7.69	**15.86 / 7.83**	$+0.42$ / $+0.14$

C.7 THIRTY-ROUND FRAMEWORK RESULTS BY MODEL

Table 16 reports hit counts and within-row count differences for all 18 model–dataset comparisons. Table 4 reports the corresponding recall summary using the dataset-specific hit-set denominators in the evaluation configuration. These HyPS endpoints come from a different experiment group than the fixed-direction comparison, so the two studies are analyzed separately.

Table 16: **Model-wise framework comparison at 30 rounds.** Entries are All / N/E hit counts. HyPS and BDA both enable dependency-profile retrieval. These runs are separate from the five-round comparison. Bold marks the higher count or both counts for ties. HyPS improves both metrics in 17 of 18 comparisons and ties on Steinhart with Claude Sonnet 5.

Dataset	Model	BDA	HyPS	Δ hits
IFNG	Claude Sonnet 5	349 / 239	**379 / 266**	+30 / +27
	GPT-5.6 Sol	327 / 231	**383 / 269**	+56 / +38
	GPT-5.6 Terra	358 / 246	**372 / 262**	+14 / +16
IL2	Claude Sonnet 5	394 / 229	**408 / 235**	+14 / +6
	GPT-5.6 Sol	362 / 199	**401 / 233**	+39 / +34
	GPT-5.6 Terra	369 / 201	**394 / 228**	+25 / +27
Carnevale	Claude Sonnet 5	260 / 190	**262 / 191**	+2 / +1
	GPT-5.6 Sol	247 / 182	**257 / 186**	+10 / +4
	GPT-5.6 Terra	259 / 193	**260 / 198**	+1 / +5
Scharenberg22	Claude Sonnet 5	47 / 43	**48 / 44**	+1 / +1
	GPT-5.6 Sol	45 / 42	**47 / 43**	+2 / +1
	GPT-5.6 Terra	47 / 43	**48 / 44**	+1 / +1
Sanchez21	Claude Sonnet 5	271 / 209	**276 / 212**	+5 / +3
	GPT-5.6 Sol	259 / 204	**302 / 238**	+43 / +34
	GPT-5.6 Terra	255 / 202	**293 / 234**	+38 / +32
Steinhart GD2	Claude Sonnet 5	**36 / 32**	**36 / 32**	0 / 0
	GPT-5.6 Sol	36 / 32	**37 / 33**	+1 / +1
	GPT-5.6 Terra	35 / 31	**36 / 32**	+1 / +1

C.8 LONG-HORIZON GAIN TRAJECTORIES

The All gain peaks at round 20 and N/E at round 25 (Table 17), so neither increases monotonically. The final N/E trajectory gain is 2.19 pp, compared with 2.26 pp from the six-dataset endpoint average. Because the trajectory and endpoint summaries aggregate at different stages, we report them separately rather than treating them as the same estimate.

Table 17: **Long-horizon recall gains.** HyPS minus BDA recall in percentage points (pp), with dependency-profile retrieval enabled in both methods. Maxima are bold. Gains are positive but not monotonic. Appendix C.8 discusses the difference between the final N/E value and the endpoint summary.

Round	Δ All recall (pp)	Δ N/E recall (pp)
1	+0.35	+0.20
5	+0.85	+0.52
10	+1.42	+1.10
15	+1.95	+1.65
20	**+2.58**	+2.15
25	+2.42	**+2.28**
30	+2.35	+2.19

D Implementation and Statistical Details

Implementation record. The standard single-gene configuration uses $K = 4$, $b = 128$, three anchors per slot, and quota bounds of 16–48. Sections 4.2–4.3 and Appendices A.1–A.3 specify portfolio updates, allocation, relation features, and scorer fitting. The supplementary configuration files record undersized-pool handling, retry rules, software versions, random seeds, and round-level selections and responses. They also identify the Achilles/DepMap, Reactome, and SIGNOR snapshots through release and retrieval metadata, hashes, gene mappings, and post-mapping coverage. Appendices A.4 and B.1 document the control-reserving trace and two-gene evaluation interface.

Run and uncertainty records. The supplementary run records underlie every reported aggregate. Table 1 reports ten-run means without SDs, whereas the trajectory and dataset-level studies report SDs where shown. The averaged dataset SDs in Table 9 summarize dataset-level variability and are not macro-recall SDs. Table 12 reports sample SDs across ten runs, while Table 11 reports means only. Accordingly, we do not make significance claims from the macro summaries.

Organization across studies. The main comparison, module controls, portfolio-size study, and thirty-round experiments are separate run groups, each linked to its own configuration and per-run records. Figure 2 uses Claude Sonnet 5 and a dedicated ten-run trajectory evaluation, which explains why its traditional-method endpoints are not entries from Table 1. Appendix C.8 reports the trajectory aggregation separately from the endpoint aggregation. Finally, τ is stored with each HyPS run and follows the benchmark-informed construction in Appendix A.3.

E COMPLETE AGENT TRACE

Run protocol Claude Sonnet 5 / IL2

```
Protocol: HyPS full_core, IL2, Claude Sonnet 5, seed 44.
Five rounds; 128 new single-gene tests per round.
Chosen as the median All-hit seed (173, 184, 185)
among three complete IL2 full-core seeds.
The trace is one run, not a seed average.

Decision loop: four hypothesis regions; independent
Achilles-expanded pools; bounded adaptive allocation;
pool-wise selection; assay feedback and revision.
Round 1 uses the LLM selector. Rounds 2-5 use a
response model refitted on measured feedback.

Artifact chain: the planner prompt contains the task,
recent feedback and prior mechanism identities. Its
answer defines hypotheses/anchors. Achilles expands
each anchor set into a candidate pool. The allocator
commits final quotas. Round 1 selects pool indices by
LLM; later rounds rank each pool with the fitted
response model. The measured scores become the next
planner prompt and response-model training data.

Frozen internal feedback event: |score| >= 0.32403.
GT All/N/E is a separate post-hoc evaluation:
  All denominator=654; N/E denominator=481.
N/E excludes common-essential genes. Do not equate
feedback positives with GT hits.

Privacy: scientific content only. Service-routing metadata,
authentication information, identities, and machine paths
are excluded; experiment-relevant configuration is retained.
This is a structured experimental trace, not a verbatim
service transcript.

Source consistency checks: all 5 proposal snapshots,
the final state, and cumulative sampled-gene arrays
```

Round 1 Claude Sonnet 5 / IL2

```
agree on 640 unique measured genes; no planner fallback.
ROUND 1: plan -> expand -> allocate -> select -> assay

Previous feedback: none (initial round).

Selection mode: llm-selector
Allocation mode: bounded-adaptive
Pipeline: hypothesis_planning, candidate_expansion,
  adaptive_allocation, pool_wise_selection,
  experiment_and_update
Planning quotas / actual quotas / pool-priority:
  r1: 32 / 32 / 0.791
  r2: 32 / 32 / 0.791
  r3: 32 / 32 / 0.791
```

Round 1 (continued) Claude Sonnet 5 / IL2

```
  r4: 32 / 32 / 0.791

Hypotheses, revisions, and observed response:
r1 (exploit; 30 candidates + 2 controls)
  Hypothesis: TCR proximal signaling loss reduces IL-2
    output
  Anchors: LCP2, ZAP70, LAT
  Controls: HPRT1, GAPDH
  Revision: initialize - initial hypothesis
  Feedback: 3/30 candidate; 0/2 control
r2 (expand; 30 candidates + 2 controls)
  Hypothesis: NFAT/calcineurin pathway boundary controls
    IL-2 magnitude
  Anchors: PPP3CA, NFATC2, CALM1
  Controls: ACTB, TUBB
  Revision: initialize - initial hypothesis
  Feedback: 5/30 candidate; 0/2 control
r3 (challenge; 30 candidates + 2 controls)
  Hypothesis: Negative feedback via ubiquitin ligases masks
    true hits
  Anchors: CBLB, ITCH, PDCD1
  Controls: POLR2A, EEF2
  Revision: initialize - initial hypothesis
  Feedback: 1/30 candidate; 1/2 control
r4 (discover; 30 candidates + 2 controls)
  Hypothesis: Metabolic/mTOR sensing independently gates
    IL-2 transcription
  Anchors: MTOR, RPTOR, TSC2
  Controls: B2M, RPL13A
  Revision: initialize - initial hypothesis
  Feedback: 8/30 candidate; 1/2 control

Execution details:
  No control repair recorded.
  Anchor repairs: none
  Control fallbacks: none
  Candidate-pool expansions: 0
  Common-essential cap enabled: False
  Planner fallback: none

Candidate pools before selection (recorded order):
  r1: 92 candidates
  TESC, CMTM7, DBNDD2, EVC2, ZNF880, DAB2IP, BEND2, CLEC4M
  UNC5A, POTEE, HAO2, KIAA1211L, CHCHD5, RABL2A, SULT1C2
  FAHD2A, C1QTNF4, HOXD10, EPS8L3, NELFCD, C16orf92, YPEL3
  SBK1, NFATC2IP, CA5A, RNF40, YPEL1, KIF22, SRCAP, HSD3B7
  PYDC1, GPR61, PLA2G10, MPV17L, ZNF629, CDC42SE1, POLR3E
  RBBP6, GDE1, RNPS1, THOC6, SRRM2, GTF3C1, HAUS4, GFER
  MRPS21, CDT1, NPIPB6, PKMYT1, ANXA9, SETD1A, ZNF785
  POLR3K, PGS1, PTPA, GRIN2A, SMG1, BRCA1, ABCB7, MINDY1
  PPP4C, NUBP2, CD8B, HINFP, WDHD1, ZNF688, MRPL38, SUPT5H
  BRF2, SMG5, EARS2, ZNF207, COPS4, TIMM29, UBQLN4, BICDL2
  ELOB, NUBP1, MMS19, CIAPIN1, OGT, EXOSC5, POLR2F, SMG6
  LCE1E, YEATS4, TNPO3, HAUS3, RSL1D1, PSMC3, CCP110
  C16orf96
  r2: 92 candidates
  ZGRF1, SNAPC3, POLR3B, MED4, LSM6, CCT8, RPL14, LARS
  KARS, TSHZ2, CTCFL, SGK2, FAM209A, ZNF582, TMEM246
```

Round 1 (continued) Claude Sonnet 5 / IL2

```
  KIAA1755, SLC12A5, PRKD1, C1D, ZCCHC17, UPF3A, TCEA3
  LMAN2, IQCF1, PDZK1IP1, RHOQ, ABCF1, AURKA, CDK1, NAA15
  DDX55, CLP1, CDC27, CUL1, PSME1, LSM2, ALG14, URB1, INTS8
  MARS, IARS, UBA3, GTF2B, SNW1, COPS5, RFC2, SRSF11
  SLC9B1, RAD51, BUD31, THOC7, INTS3, MEPCE, TMA16, PFDN6
  SEM1, SASS6, DTL, IMP4, POLR2K, CCNA2, SRP68, NAA50
  EFTUD2, POLR2B, UBTF, PRPF3, GTF2E1, WDR3, RPL5, RAD17
  NOB1, CCT3, MAK16, TRAPPC3, RUVBL1, XAB2, PSMD13, SBNO1
  NMD3, RPS20, RBM22, MIS18BP1, UQCRH, ANKLE2, EIF2B1
  ACTL6A, PLRG1, RPN1, PSMB3, LUC7L3, EIF3A
  r3: 92 candidates
  CEACAM5, ABCA10, CLEC7A, SIDT1, HEG1, CMSS1, TAAR1
  CEACAM1, CLLU1OS, SPRY2, ARHGAP35, ZNF341, IFT52, GHRH
  TTLL9, DEFB124, ACSS2, AFAP1, GAL3ST2, OR2T1, LPCAT1
  CRHR1, ZNF623, MROH6, OR2W1, CCL4, IQSEC3, CES5A, EBLN2
  RRH, LEKR1, CCDC54, ZNF431, TXLNB, RNASE6, GAP43, APLN
  HDX, TNS4, OR51L1, TVP23B, CFHR4, IL33, CEP128, LAMP2
  THEMIS, ZNF157, FAM3C, SLCO1B7, SPINK2, PDLIM5, SULT1C3
  GRIK2, ZNF729, PTBP2, EFCAB1, NLK, AKAIN1, C8orf48
  TMEM155, FOXD4L4, GPX3, NLRP14, FLNA, MAGED2, SMCP, ZNF35
  ABI3BP, RFPL4AL1, CYLC1, COL4A6, GRIA2, PER1, POPDC3
  RAB9A, STARD13, MGST1, KRTAP10-2, SPRED1, CCL20, SASH1
  LRRC70, AP1S2, RGPD3, TRIM73, ZNF780A, TNFSF18, VAMP5
  GPR141, CEP83, OR10Z1, NPY2R
  r4: 92 candidates
  TBC1D3B, RPS29, YARS, DDX56, NOL10, UTP15, PSMA1, RPS17
  RSL24D1, PRMT1, TSC1, DDIT4, NPRL2, KEAP1, PPCDC, EIF2AK4
  KCTD5, NAMPT, ADSS, LCMT1, PRELID1, ETF1, RPS16, RPS11
  CCT4, POLR1C, FARSB, RPL24, TARS, DDX49, WDR43, DHX37
  RPL8, PSMA6, KPNB1, COPA, NIP7, TSR1, PAFAH1B1, TRMT112
  SF1, LYRM4, DDB1, DBR1, RPL27, VCP, NOL9, HSPA9, POLR3A
  USP36, PSMB4, RCC1, EIF1AD, RPL23, CDK11A, RPL18, INTS11
  XPO1, PRPF19, PELP1, EIF2S1, SNRNP200, DYNC1H1, TAF1C
  TOMM22, SNRPD1, RPS27A, SPOUT1, COPB1, RRP12, DDX54
  RPL32, SARS, UBA1, RPS9, NIFK, RUVBL2, AARS, RPS8, TCP1
  NUP93, SAE1, OSGEP, ERH, ARL2, LRR1, SDE2, PSMA3, DDX10
  RPL12, FAU, POLR3H

All selected genes and measured outcomes:
  Columns: gene score flags; rounds 2-5 also p/u.
  F=feedback hit, A=GT All, N=GT N/E, E=CEG,
  C=control; p=predicted response, u=uncertainty.
r1: 32 tested
  01 HPRT1      -0.06477 ----C
  02 GAPDH      -0.10180 ---EC
  03 NFATC2IP   +0.24520 -----
  04 CD8B       -0.01407 -----
  05 SBK1       +0.04529 -----
  06 DAB2IP     -0.22077 -----
  07 GRIN2A     -0.21998 -----
  08 UNC5A      +0.03087 -----
  09 TESC       +0.01688 -----
  10 CDC42SE1   -0.03474 -----
  11 SRCAP      -0.11666 -----
  12 RNF40      +1.03200 FAN--
  13 SETD1A     +0.27820 -----
  14 OGT        +0.01658 ---E-
  15 PPP4C      +0.22190 ---E-
```

Round 1 (continued) Claude Sonnet 5 / IL2

```
  16 SMG1       +0.15249 -----
  17 BRCA1      +0.24001 -----
  18 RBBP6      -0.01497 ---E-
  19 POLR3E     -0.05879 -----
  20 POLR3K     +0.00803 ---E-
  21 POLR2F     +0.13854 -----
  22 SUPT5H     -0.58234 FA-E-
  23 BRF2       +0.02273 -----
  24 GTF3C1     -0.10130 ---E-
  25 RNPS1      +0.02366 -----
  26 PTPA       +0.10477 ---E-
  27 UBQLN4     +0.28637 -----
  28 ELOB       +0.29736 -----
  29 PSMC3      +0.42005 FA-E-
  30 MINDY1     -0.17286 -----
  31 CIAPIN1    +0.00329 -----
  32 COPS4      -0.00421 -----
r2: 32 tested
  01 ACTB       +0.04800 ---EC
  02 TUBB       -0.03774 ---EC
  03 RHOQ       +0.02132 -----
  04 PRKD1      +0.26977 -----
  05 SGK2       -0.15007 -----
  06 AURKA      +0.16864 -----
  07 CDK1       +0.43441 FA-E-
  08 UBA3       -0.34319 FAN--
  09 COPS5      +0.21354 -----
  10 TCEA3      -0.08550 -----
  11 TSHZ2      -0.01111 -----
  12 CTCFL      -0.04795 -----
  13 FAM209A    +0.11620 -----
  14 ZNF582     -0.12187 -----
  15 TMEM246    -0.00442 -----
  16 KIAA1755   -0.04409 -----
  17 SLC12A5    +0.27074 -----
  18 IQCF1      +0.04072 -----
  19 SLC9B1     +0.08985 -----
  20 PDZK1IP1   -0.00878 -----
  21 LMAN2      +0.13047 -----
  22 ALG14      -0.07910 ---E-
  23 UQCRH      -0.08603 -----
  24 DTL        +0.06212 ---E-
  25 CCNA2      +0.16211 ---E-
  26 RAD51      +0.27451 -----
  27 CUL1       -0.30038 -----
  28 EFTUD2     +0.71480 FA-E-
  29 RUVBL1     +0.55881 FAN--
  30 ACTL6A     +0.83375 FA-E-
  31 POLR2B     +0.26680 ---E-
  32 GTF2B      -0.09263 ---E-
r3: 32 tested
  01 POLR2A     +0.11154 ---EC
  02 EEF2       +0.43049 FA-EC
  03 SPRY2      -0.09658 -----
  04 SPRED1     -0.10071 -----
  05 THEMIS     -0.42409 FAN--
  06 NLK        +0.13098 -----
  07 AP1S2      +0.06464 -----
```

Round 1 (continued) Claude Sonnet 5 / IL2

```
  08 LAMP2      +0.00237 -----
  09 RAB9A      -0.14619 -----
  10 FLNA       -0.12035 -----
  11 CLEC7A     -0.16276 -----
  12 CEACAM1    -0.03371 -----
  13 FAM3C      -0.03111 -----
  14 TNFSF18    -0.14001 -----
  15 CCL4       -0.05266 -----
  16 CCL20      +0.01243 -----
  17 IL33       +0.00115 -----
  18 ARHGAP35   -0.15535 -----
  19 AFAP1      -0.00346 -----
  20 LPCAT1     +0.19914 -----
  21 IQSEC3     -0.04321 -----
  22 GAP43      -0.14264 -----
  23 PDLIM5     +0.00681 -----
  24 GPX3       +0.11252 -----
  25 STARD13    +0.23423 -----
  26 SASH1      -0.03834 -----
  27 ZNF341     +0.00117 -----
  28 ACSS2      +0.01348 -----
  29 ZNF623     -0.00683 -----
  30 CEP128     +0.06693 -----
  31 PTBP2      -0.07034 -----
  32 MAGED2     -0.03367 -----
r4: 32 tested
  01 B2M        +0.08004 ----C
  02 RPL13A     +0.68174 FAN-C
  03 TSC1       +0.80855 FAN--
  04 DDIT4      +0.09463 -----
  05 NPRL2      +0.02286 -----
  06 KEAP1      -0.09058 -----
  07 EIF2AK4    -0.05095 -----
  08 NAMPT      -0.11147 ---E-
  09 KCTD5      +0.00311 -----
  10 LYRM4      -0.02334 -----
  11 PRMT1      +0.89754 FA-E-
  12 KPNB1      +0.14065 ---E-
  13 DDB1       +0.23294 ---E-
  14 VCP        +0.66110 FA-E-
  15 USP36      -0.03436 -----
  16 RCC1       -0.05505 ---E-
  17 XPO1       +0.34959 FA-E-
  18 RUVBL2     +0.58122 FA-E-
  19 PSMA1      +0.38757 FA-E-
  20 PPCDC      -0.13556 -----
  21 HSPA9      -0.03406 ---E-
  22 TOMM22     -0.05038 ---E-
  23 UBA1       +0.21104 ---E-
  24 SAE1       -0.03155 ---E-
  25 COPA       +0.05048 ---E-
  26 ARL2       +0.02018 ---E-
  27 ADSS       +0.73901 FA-E-
  28 PSMA6      +0.34373 FA-E-
  29 PSMB4      +0.13663 ---E-
  30 PSMA3      +0.04775 ---E-
  31 SARS       +0.20553 ---E-
  32 TARS       -0.12576 ---E-
```

Round 1 (continued) Claude Sonnet 5 / IL2

```
Round outcome:
  128 new tests; internal feedback hits: 19
  New GT: All 19; N/E 6
  Cumulative GT: All 19/654; N/E 6/481
  Cumulative selected CEG: 42/128
```

Round 1: Hypothesis planning Claude Sonnet 5 / IL2

```
HOW THE FIRST DECISIONS WERE PRODUCED

SOURCE LOG EXCERPT: round_1_planner.log
Line wrapping added. [omitted] marks duplicate bulk data.
Model-generated biological claims are unverified.
==================prompt====================


Human: You are an experimental scientist learning from
  sequential perturbation experiments.

Research problem:
I'm planning to run a genome-wide CRISPR screen to identify
  genes that regulate the production of Interleukin-2
  (IL-2). There are 18,939 possible  genes to perturb and I
  can only perturb 128 genes at a time. For each
  perturbation, I'm able to measure out the log fold change
  in Interleukin-2 (IL-2) normalized read counts which will
  be referred to as the score. I can only do a few rounds
  of experimentation.

Your objective is NOT to maximize one extreme score.
  Discover as many genes as possible whose
absolute experimental effect crosses the predeclared hit
  threshold, while revealing distinct causal
mechanisms.  A region is useful only if its predicted
  positives later beat its own controls.

Threshold-aligned observations from the most recent genes
  actually tested (at most 64):
[]

Plan hypotheses for these mechanism slots. Counts are
  provisional capacity hints; the bounded adaptive
allocator commits the final 128-gene quotas only after
  hypotheses and candidate pools are frozen:
[{"id": "r1", "action": "exploit", "count": 32}, {"id":
  "r2", "action": "expand", "count": 32}, {"id": "r3",
  "action": "challenge", "count": 32}, {"id": "r4",
  "action": "discover", "count": 32}]

Empirical mechanism portfolio (candidate and control
  outcomes from prior rounds):
{"r1": {"candidate_trials": 0, "candidate_hits": 0,
  "control_trials": 0, "control_hits": 0}, "r2":
```

Round 1: Hypothesis planning (continued) Claude Sonnet 5 / IL2

```
  {"candidate_trials": 0, "candidate_hits": 0,
  "control_trials": 0, "control_hits": 0}, "r3":
  {"candidate_trials": 0, "candidate_hits": 0,
  "control_trials": 0, "control_hits": 0}, "r4":
  {"candidate_trials": 0, "candidate_hits": 0,
  "control_trials": 0, "control_hits": 0}}

Persistent mechanism-slot identities from the immediately
  preceding round:
{}

The exploit region may extend an established mechanism.
  Expand tests its boundary.  Challenge tests
a plausible rival explanation of established hits.
  Discover regions must be mechanistically distinct
from established regions and from one another: do not
  merely rename a pathway or list direct neighbors.
Each region must include a compact falsifiable hypothesis
  of at most 8 words, exactly 3 anchor genes,
and its assigned number of negative-control genes expected
  NOT to hit:
{"r1": 2, "r2": 2, "r3": 2, "r4": 2}. Region IDs are
  persistent
mechanism slots. Set revision=preserve only when the
  hypothesis text is unchanged. Use refine or replace
when the hypothesis changes, and give an evidence-based
  revision_reason; changed hypotheses start a new
mechanism version with fresh allocation statistics. On the
  first round use replace with reason
"initial hypothesis".
For every region, return 3-6 valid HGNC anchor genes. A
  separate Achilles tool will expand anchors only within
  that region. Do not use anchors from one region to
  populate another region.

Return only one JSON object:
{
  "map_update": {"established": [], "uncertain": [],
  "falsified": [], "unexplained_hits": []},
  "regions": [
    {"id": "r1", "action": "exploit", "hypothesis":
  "testable causal statement", "revision": "replace",
  "revision_reason": "evidence-based reason", "anchors":
  ["HGNC"], "controls": ["HGNC"]}
  ]
}


Assistant:
```json
{
  "map_update": {"established": [], "uncertain": [],
  "falsified": [], "unexplained_hits": []},
  "regions": [
    {
      "id": "r1",
      "action": "exploit",
      "hypothesis": "TCR proximal signaling loss reduces
```
```

Round 1: Hypothesis planning (continued) — Claude Sonnet 5 / IL2

```
  IL-2 output",
      "revision": "replace",
      "revision_reason": "initial hypothesis",
      "anchors": ["LCP2", "ZAP70", "LAT", "PLCG1"],
      "controls": ["HPRT1", "GAPDH"]
    },
    {
      "id": "r2",
      "action": "expand",
      "hypothesis": "NFAT/calcineurin pathway boundary
  controls IL-2 magnitude",
      "revision": "replace",
      "revision_reason": "initial hypothesis",
      "anchors": ["PPP3CA", "NFATC2", "CALM1", "PPP3R1"],
      "controls": ["ACTB", "TUBB"]
    },
    {
      "id": "r3",
      "action": "challenge",
      "hypothesis": "Negative feedback via ubiquitin
  ligases masks true hits",
      "revision": "replace",
      "revision_reason": "initial hypothesis",
      "anchors": ["CBLB", "ITCH", "PDCD1"],
      "controls": ["POLR2A", "EEF2"]
    },
    {
      "id": "r4",
      "action": "discover",
      "hypothesis": "Metabolic/mTOR sensing independently
  gates IL-2 transcription",
      "revision": "replace",
      "revision_reason": "initial hypothesis",
      "anchors": ["MTOR", "RPTOR", "TSC2", "AKT1"],
      "controls": ["B2M", "RPL13A"]
    }
  ]
}
```
END SOURCE LOG EXCERPT

Round 1: r3 candidate selection — Claude Sonnet 5 / IL2

```
SOURCE LOG EXCERPT: round_1_r3_selector_1.log
Line wrapping added. [omitted] marks duplicate bulk data.
Model-generated biological claims are unverified.
==================prompt====================


Human: You are selecting an experiment batch from
  independently generated candidate pools.

Research problem:
I'm planning to run a genome-wide CRISPR screen to identify
```

Round 1: r3 candidate selection (continued) Claude Sonnet 5 / IL2

```
  genes that regulate the production of Interleukin-2
  (IL-2). There are 18,939 possible  genes to perturb and I
  can only perturb 128 genes at a time. For each
  perturbation, I'm able to measure out the log fold change
  in Interleukin-2 (IL-2) normalized read counts which will
  be referred to as the score. I can only do a few rounds
  of experimentation.

Use the discovery map and observed hit/non-hit evidence to
  choose genes. Do not maximize one extreme
effect or collapse all choices into the established
  mechanism. Preserve the exact region quotas.

Discovery map:
{"established": [], "uncertain": [], "falsified": [],
  "unexplained_hits": []}

Required region quotas:
[{"id": "r3", "action": "challenge", "count": 8}]

Mechanism hypotheses:
{"r3": {"action": "challenge", "hypothesis": "Negative
  feedback via ubiquitin ligases masks true hits",
  "anchors": ["CBLB", "ITCH", "PDCD1"]}}

Candidate pools, one per mechanism. Select only their
  integer indices:
[omitted: indexed pool; complete r3 pool above]

Return only JSON: {"selection_indices_by_region": {"r1":
  [0, 1]}}. Each integer must be a distinct valid index
  from that region's pool.
Return exactly every required gene once, with each region
  containing its exact quota. A gene may only be
selected from its region's candidate pool when pools are
  provided.

[omitted: batch exclusion list; all selections above]
Return at least 8 novel candidates; you may return up to 16
  candidates so invalid or repeated symbols can be removed.

Assistant:
Looking at the r3 challenge region, the hypothesis is about
  negative feedback via ubiquitin ligases (CBLB, ITCH,
  PDCD1 as anchors). I need to select genes that could
  plausibly test/challenge this mechanism - related to
  immune signaling, ubiquitination, feedback regulation, or
  T-cell function, without collapsing into obvious
  established anchors themselves.

Looking at the pool, genes with plausible immune/signaling
  relevance for challenging the negative feedback
  hypothesis: SPRY2 (RTK feedback inhibitor), SPRED1 (RAS
  feedback inhibitor), THEMIS (T-cell signaling), NLK
  (negative regulator of signaling), AP1S2 (adaptor
  complex/trafficking), LAMP2 (lysosomal degradation),
  RAB9A (vesicular trafficking, could relate to receptor
  degradation), FLNA (scaffold for signaling), CCL20/CCL4
```

Round 1: r3 candidate selection (continued) — Claude Sonnet 5 / IL2

```
  (immune chemokines, potential unexplained hits), IL33
  (immune cytokine).

```json
{"selection_indices_by_region": {"r3": [9, 78, 45, 56, 82,
  44, 74, 63, 25, 79, 42, 35]}}
```
END SOURCE LOG EXCERPT
```

Round 1: Index-to-gene mapping — Claude Sonnet 5 / IL2

```
Index-to-gene check (r3 selector call 1):
  9: SPRY2 (accepted)
  78: SPRED1 (accepted)
  45: THEMIS (accepted)
  56: NLK (accepted)
  82: AP1S2 (accepted)
  44: LAMP2 (accepted)
  74: RAB9A (accepted)
  63: FLNA (accepted)
The remaining proposed indices were backups; the
quota was eight candidates for this selector call.
Later calls completed the r3 batch shown above.
```

Round 2 — Claude Sonnet 5 / IL2

```
ROUND 2: plan -> expand -> allocate -> select -> assay

Previous round candidate feedback hits:
  r1: 3/30
  r2: 5/30
  r3: 1/30
  r4: 8/30

Selection mode: experiment-trained-pool-wise
Allocation mode: bounded-adaptive
Pipeline: hypothesis_planning, candidate_expansion,
  adaptive_allocation, pool_wise_selection,
  experiment_and_update
Planning quotas / actual quotas / pool-priority:
  r1: 32 / 16 / 0.253
  r2: 32 / 16 / 0.315
  r3: 32 / 48 / 1.211
  r4: 32 / 48 / 1.211

Hypotheses, revisions, and observed response:
r1 (expand; 14 candidates + 2 controls)
  Hypothesis: TCR proximal signaling loss reduces IL-2
    output
  Anchors: LCP2, ZAP70, LAT
```

Round 2 (continued) Claude Sonnet 5 / IL2

```
  Controls: ALDOA, EEF1A1
  Revision: preserve - unchanged hypothesis, low but
    nonzero hit rate (3/30) justifies boundary expansion
  Feedback: 6/14 candidate; 0/2 control
r2 (exploit; 14 candidates + 2 controls)
  Hypothesis: NFAT/calcineurin pathway boundary controls
    IL-2 magnitude
  Anchors: PPP3CA, NFATC2, CALM1
  Controls: RPLP0, TUBB4B
  Revision: preserve - unchanged hypothesis, 5/30 hit rate
    supports continued exploitation
  Feedback: 5/14 candidate; 0/2 control
r3 (discover; 46 candidates + 2 controls)
  Hypothesis: RNA splicing/mRNA processing controls IL-2
    transcript levels
  Anchors: SF3B1, U2AF1, SRSF1
  Controls: GUSB, PPIA
  Revision: replace - ubiquitin ligase feedback hypothesis
    falsified (1/30 hits, sole hit THEMIS is TCR-adaptor
    not ligase); pivoting to distinct RNA-processing
    mechanism, mechanistically unrelated to TCR signaling,
    calcineurin, or mTOR/proteostasis
  Feedback: 26/46 candidate; 0/2 control
r4 (challenge; 46 candidates + 2 controls)
  Hypothesis: Proteasome/nuclear transport, not mTOR,
    drives effect
  Anchors: PSMB5, XPO1, VCP
  Controls: TBP, RPL19
  Revision: refine - strongest hits (VCP, XPO1, RUVBL2,
    PSMA1/A6, ADSS) are proteostasis/transport/nucleotide
    synthesis genes, not canonical mTOR components
    (MTOR/RPTOR/TSC2 themselves untested here); testing
    rival explanation that proteasome/export machinery, not
    mTOR sensing per se, is the causal driver
  Feedback: 22/46 candidate; 1/2 control

Execution details:
  r1 control repair: ALDOA, EEF1A1
  r2 control repair: RPLP0, TUBB4B
  r3 control repair: PPIA
  r4 control repair: TBP, RPL19
  Anchor repairs: none
  Control fallbacks: none
  Candidate-pool expansions: 0
  Common-essential cap enabled: False
  Planner fallback: none

Feedback-trained response selector:
  Fit: all-score multiview ridge logistic; 128 measured
  genes; 19 internal positives; threshold 0.32403.
  Views: Achilles, Reactome, SIGNOR; frozen features,
  coefficients learned from measured assay scores.
  On-policy candidate audit: n=120; observed hit
  rate=0.492; Brier=0.345.

Candidate pools before selection (recorded order):
  r1: 92 candidates
  CMTM7, DBNDD2, EVC2, ZNF880, BEND2, CLEC4M, POTEE, HAO2
```

Round 2 (continued) Claude Sonnet 5 / IL2

```
  KIAA1211L, CHCHD5, RABL2A, SULT1C2, FAHD2A, C1QTNF4
  HOXD10, EPS8L3, NELFCD, C16orf92, YPEL3, CA5A, YPEL1
  KIF22, HSD3B7, PYDC1, GPR61, PLA2G10, MPV17L, ZNF629
  GDE1, THOC6, SRRM2, HAUS4, GFER, MRPS21, CDT1, NPIPB6
  PKMYT1, ANXA9, ZNF785, PGS1, ABCB7, NUBP2, HINFP, WDHD1
  ZNF688, MRPL38, SMG5, EARS2, ZNF207, TIMM29, BICDL2
  NUBP1, MMS19, EXOSC5, SMG6, LCE1E, YEATS4, TNPO3, HAUS3
  RSL1D1, CCP110, C16orf96, DNAJC8, WDR88, DHX8, TWF1, VMP1
  HNRNPM, POLD3, SUPT6H, SNRNP25, PPP1R8, NELFA, BARD1
  ALG1L, ELP3, EXOSC6, CDC6, MRPL14, COPS8, TBL3, CORO6
  MRPL49, POLR3C, ILF3, SUGP1, SNAPC1, TARDBP, MED29, SRSF7
  PTPN23, POLE2
  r2: 92 candidates
  ZGRF1, SNAPC3, POLR3B, MED4, LSM6, CCT8, RPL14, LARS
  KARS, C1D, ZCCHC17, UPF3A, ABCF1, NAA15, DDX55, CLP1
  CDC27, PSME1, LSM2, URB1, INTS8, MARS, IARS, SNW1, RFC2
  SRSF11, BUD31, THOC7, INTS3, MEPCE, TMA16, PFDN6, SEM1
  SASS6, IMP4, POLR2K, SRP68, NAA50, UBTF, PRPF3, GTF2E1
  WDR3, RPL5, RAD17, NOB1, CCT3, MAK16, TRAPPC3, XAB2
  PSMD13, SBNO1, NMD3, RPS20, RBM22, MIS18BP1, ANKLE2
  EIF2B1, PLRG1, RPN1, PSMB3, LUC7L3, EIF3A, CACNB3, CDC73
  SACM1L, HNRNPC, SMC3, ERH, CCNH, RBMX2, NARS, PSMA5, BMS1
  NIFK, POLE, PSMA4, RPL3, SLU7, MED8, POLR1A, FARSA, TCP1
  DDX23, RFC3, LSM8, IK, SART3, HAUS6, SEC61G, XRN2, TOPBP1
  COPB1
  r3: 92 candidates
  SNRNP200, RAN, PSMD7, SF3A1, HNRNPK, RNGTT, SRSF2
  TBC1D3B, SNRPD1, PRELID1, YARS, POLR2L, SNRPA1, RPL12
  SF3B5, CDC45, EIF2S1, SF3B3, PSMD14, SF1, ETF1, SMU1
  RPS16, RPS19, PRPF8, RPS11, RPL23, WEE1, SNRPB, POLR2I
  PCNA, SNRPF, RPS29, RPS28, CCT4, NEDD1, SNRPE, PHF5A
  CDC5L, SF3B4, RPS27A, RPL4, FARSB, RPL9, HCFC1, POLR2D
  RRM1, SFPQ, PRPF31, WDR43, KIF11, RPL17, POLR2C, PAFAH1B1
  VPS25, FAU, RPL37, HSPE1, U2AF2, RPS13, RRM2, RPL18
  RPS3A, POLR3A, RPLP2, TXNL4A, ATP6V0C, SNRPD2, NUDT21
  RPL10A, RPL11, EIF5, PRPF38A, CDC20, RPL27, SPC24, PSMA2
  PSMA7, RPL7A, CDC123, RPS9, SRSF3, DONSON, PRPF19
  RSL24D1, PUF60, NIP7, POLR2E, INTS11, COPB2, RNPC3, DDX41
  r4: 92 candidates
  THAP1, RPS17, RPL24, AARS, SNRPG, ANAPC11, NEDD8, GPN1
  UTP15, EIF2S3, DYNC1H1, MTBP, GINS2, DHX37, DDX56, SDE2
  CDC23, PSMD4, PRPF6, RPL8, DDX49, NSF, RPA1, NUP93, ISY1
  TSR1, SRP54, CSE1L, CDK9, DNAJC17, CDK11A, PDRG1, DHDDS
  ISCU, CTCF, CHAF1B, SPOUT1, RPL7, GINS1, NUP85, CHMP6
  POLR1C, SF3A2, SEC13, AURKB, ATP6V0B, RPL32, HARS
  RABGGTA, RPS2, MCM2, ZMAT2, SMC2, PRIM1, RPS6, RRP12
  ANAPC2, SEC61A1, RPL18A, LRR1, TIMELESS, DBR1, CDC16
  PPWD1, INTS9, DDX54, DDX10, ESPL1, RPL31, RAE1, RPS24
  EIF3I, RPL23A, SAP30BP, CARS, MED11, TSR2, PPAN, QARS
  UBL5, NUTF2, SSU72, NUP133, NCBP1, RPS15A, TRMT112, CDC7
  CHEK1, NFS1, BUD23, RPS12, TUBGCP2

All selected genes and measured outcomes:
  Columns: gene score flags; rounds 2-5 also p/u.
  F=feedback hit, A=GT All, N=GT N/E, E=CEG,
  C=control; p=predicted response, u=uncertainty.
r1: 16 tested
  01 ALDOA      -0.02442 ----C p=-- u=--
```

Round 2 (continued) Claude Sonnet 5 / IL2

```
  02 EEF1A1     +0.20723 ----C p=-- u=--
  03 BARD1      +0.03653 ----- p=0.332 u=0.180
  04 PKMYT1     +0.45712 FA-E- p=0.261 u=0.139
  05 KIF22      -0.12488 ----- p=0.253 u=0.167
  06 NELFA      +0.00454 ----- p=0.209 u=0.049
  07 NELFCD     +0.03483 ----- p=0.205 u=0.047
  08 POLD3      +0.52072 FAN-- p=0.191 u=0.047
  09 CDC6       +0.46684 FAN-- p=0.189 u=0.046
  10 POLE2      +0.09347 ----- p=0.188 u=0.047
  11 CDT1       +0.35598 FA-E- p=0.183 u=0.046
  12 SUPT6H     -0.35785 FA-E- p=0.175 u=0.046
  13 COPS8      -0.05493 ----- p=0.175 u=0.045
  14 HAUS4      +0.31646 ----- p=0.174 u=0.045
  15 HAUS3      +0.37935 FAN-- p=0.174 u=0.045
  16 EXOSC5     +0.10532 ----- p=0.173 u=0.046
r2: 16 tested
  01 RPLP0      +0.25585 ---EC p=-- u=--
  02 TUBB4B     +0.24619 ----C p=-- u=--
  03 SEM1       +0.35038 FAN-- p=0.587 u=0.250
  04 PSMB3      -0.00537 ---E- p=0.567 u=0.240
  05 PSMA5      +0.33729 F--E- p=0.567 u=0.240
  06 PSMA4      +0.22661 ---E- p=0.566 u=0.240
  07 PSMD13     +0.15295 ---E- p=0.560 u=0.238
  08 CDC27      +0.16980 ---E- p=0.334 u=0.197
  09 POLR2K     +0.09758 ----- p=0.253 u=0.064
  10 NIFK       +0.26479 ----- p=0.249 u=0.136
  11 CCNH       +0.14323 ---E- p=0.233 u=0.056
  12 RFC2       +0.30456 ---E- p=0.208 u=0.049
  13 RFC3       +0.49916 FAN-- p=0.207 u=0.049
  14 RPS20      +0.43328 FA-E- p=0.188 u=0.047
  15 POLE       +0.56658 FAN-- p=0.188 u=0.047
  16 RPL5       +0.09846 ----- p=0.188 u=0.047
r3: 48 tested
  01 GUSB       -0.04105 ----C p=-- u=--
  02 PPIA       +0.01792 ----C p=-- u=--
  03 RPS27A     +0.47663 FAN-- p=0.655 u=0.267
  04 RAN        +0.12271 ---E- p=0.648 u=0.500
  05 PSMA2      +0.29018 ---E- p=0.567 u=0.240
  06 PSMA7      -0.00316 ---E- p=0.566 u=0.240
  07 PSMD14     +0.21094 ---E- p=0.564 u=0.238
  08 PSMD7      +0.57379 FAN-- p=0.562 u=0.237
  09 RRM2       +0.49593 FAN-- p=0.280 u=0.178
  10 POLR2I     +0.38523 FA-E- p=0.259 u=0.066
  11 POLR2C     +0.35633 FA-E- p=0.259 u=0.066
  12 POLR2D     +0.56226 FA-E- p=0.259 u=0.066
  13 NEDD1      +0.31500 ----- p=0.257 u=0.136
  14 POLR2E     +0.15590 ---E- p=0.254 u=0.064
  15 POLR2L     -0.02675 ---E- p=0.254 u=0.064
  16 CDC20      +0.29234 ---E- p=0.252 u=0.071
  17 CDC5L      +0.56872 FA-E- p=0.232 u=0.104
  18 KIF11      +0.36433 FA-E- p=0.230 u=0.103
  19 WEE1       -0.17798 ---E- p=0.230 u=0.065
  20 PCNA       +0.19575 ---E- p=0.213 u=0.050
  21 RPL18      +0.77455 FA-E- p=0.189 u=0.047
  22 RPS29      +0.22790 ----- p=0.188 u=0.047
  23 RPS11      +0.63731 FA-E- p=0.188 u=0.047
  24 RPL27      +0.87409 FA-E- p=0.188 u=0.047
  25 RPS16      +0.37759 FA-E- p=0.188 u=0.047
```

Round 2 (continued) Claude Sonnet 5 / IL2

```
  26 FAU         +0.49086 FA-E- p=0.188 u=0.047
  27 RPL23       +0.08725 ---E- p=0.188 u=0.047
  28 RPL4        +0.34618 FA-E- p=0.188 u=0.047
  29 RPL17       +0.48554 FAN-- p=0.188 u=0.047
  30 RPS9        +0.12483 ----- p=0.188 u=0.047
  31 RPL12       +0.33573 F--E- p=0.188 u=0.047
  32 RPS19       +0.22907 ---E- p=0.188 u=0.047
  33 RPL7A       +0.60798 FAN-- p=0.188 u=0.047
  34 RPL9        +0.44310 FAN-- p=0.188 u=0.047
  35 RPS3A       +0.38784 FAN-- p=0.188 u=0.047
  36 RPS28       -0.16369 ----- p=0.188 u=0.047
  37 RPL11       +0.04622 ---E- p=0.187 u=0.047
  38 RPL37       +0.68446 FAN-- p=0.187 u=0.046
  39 RPL10A      +0.86359 FA-E- p=0.187 u=0.046
  40 RPLP2       +0.51805 FA-E- p=0.187 u=0.046
  41 RPS13       +0.92491 FA-E- p=0.187 u=0.046
  42 PAFAH1B1    -0.25810 ---E- p=0.183 u=0.046
  43 RNGTT       +0.02730 ---E- p=0.182 u=0.046
  44 CDC45       +0.07549 ----- p=0.181 u=0.046
  45 PRPF19      +0.38417 FA-E- p=0.176 u=0.046
  46 SPC24       +0.57943 FA-E- p=0.176 u=0.046
  47 HCFC1       +0.05903 ---E- p=0.176 u=0.046
  48 ETF1        -0.44534 FAN-- p=0.175 u=0.046
r4: 48 tested
  01 TBP         +0.08113 ----C p=-- u=--
  02 RPL19       +1.20560 FA-EC p=-- u=--
  03 CDK9        +0.05982 ---E- p=0.992 u=0.072
  04 CDC23       +0.75313 FAN-- p=0.392 u=0.298
  05 CDC16       +0.48024 FA-E- p=0.391 u=0.298
  06 CHEK1       -0.06993 ---E- p=0.259 u=0.104
  07 ESPL1       +0.12687 ----- p=0.233 u=0.104
  08 CDC7        +0.61931 FAN-- p=0.219 u=0.073
  09 RPA1        +0.22904 ---E- p=0.217 u=0.051
  10 ANAPC11     -0.06527 ----- p=0.216 u=0.051
  11 ANAPC2      +0.23391 ---E- p=0.215 u=0.051
  12 NCBP1       +0.22746 ---E- p=0.210 u=0.050
  13 RPS2        +0.33243 F--E- p=0.198 u=0.047
  14 SEC13       +0.18842 ---E- p=0.195 u=0.047
  15 NUP85       -0.10742 ---E- p=0.194 u=0.047
  16 NUP133      +0.48955 FA-E- p=0.194 u=0.047
  17 RPS6        +0.47048 FA-E- p=0.191 u=0.047
  18 NUP93       +0.34690 FA-E- p=0.190 u=0.047
  19 RPL8        +0.81099 FA-E- p=0.189 u=0.047
  20 DYNC1H1     -0.16739 ----- p=0.188 u=0.047
  21 MCM2        +1.05330 FAN-- p=0.188 u=0.047
  22 RPS17       +0.82364 FAN-- p=0.188 u=0.047
  23 RPL32       +0.67605 FAN-- p=0.188 u=0.047
  24 RPL24       +0.77733 FA-E- p=0.187 u=0.046
  25 RPS24       +0.82772 FAN-- p=0.187 u=0.046
  26 RPL23A      +0.79110 FAN-- p=0.187 u=0.046
  27 RPL7        +0.31451 ----- p=0.187 u=0.046
  28 RPS12       +0.17717 ---E- p=0.187 u=0.046
  29 RPS15A      +0.20883 ---E- p=0.187 u=0.046
  30 RPL31       +0.39774 FAN-- p=0.187 u=0.046
  31 RAE1        +0.42340 FA-E- p=0.187 u=0.046
  32 RPL18A      +0.31208 ---E- p=0.187 u=0.046
  33 AURKB       +0.17232 ---E- p=0.186 u=0.046
  34 PRIM1       +0.10731 ---E- p=0.183 u=0.046
```

Round 2 (continued) Claude Sonnet 5 / IL2

```
  35 SSU72      +0.13397 ---E- p=0.179 u=0.046
  36 CHMP6      +0.85209 FA-E- p=0.179 u=0.046
  37 SMC2       +0.03376 ---E- p=0.177 u=0.046
  38 ISY1       -0.21369 ----- p=0.176 u=0.046
  39 MTBP       +0.57118 FAN-- p=0.176 u=0.046
  40 NEDD8      +0.08974 ---E- p=0.176 u=0.046
  41 EIF2S3     +0.09786 ---E- p=0.175 u=0.046
  42 SNRPG      +0.30238 ----- p=0.175 u=0.046
  43 TUBGCP2    +0.58513 FA-E- p=0.173 u=0.046
  44 GINS2      +0.38156 FA-E- p=0.173 u=0.046
  45 SEC61A1    -0.17194 ----- p=0.173 u=0.046
  46 PRPF6      +0.45058 FAN-- p=0.173 u=0.046
  47 GINS1      +0.46934 FAN-- p=0.173 u=0.046
  48 TIMELESS   +0.10878 ----- p=0.172 u=0.046

Round outcome:
  128 new tests; internal feedback hits: 60
  New GT: All 57; N/E 26
  Cumulative GT: All 76/654; N/E 32/481
  Cumulative selected CEG: 114/256
```

Round 2: Feedback-driven replanning Claude Sonnet 5 / IL2

```
HOW FEEDBACK CHANGED THE NEXT PLAN

SOURCE LOG EXCERPT: round_2_planner.log
Line wrapping added. [omitted] marks duplicate bulk data.
Model-generated biological claims are unverified.
==================prompt====================


Human: You are an experimental scientist learning from
  sequential perturbation experiments.

Research problem:
I'm planning to run a genome-wide CRISPR screen to identify
  genes that regulate the production of Interleukin-2
  (IL-2). There are 18,939 possible  genes to perturb and I
  can only perturb 128 genes at a time. For each
  perturbation, I'm able to measure out the log fold change
  in Interleukin-2 (IL-2) normalized read counts which will
  be referred to as the score. I can only do a few rounds
  of experimentation.

Your objective is NOT to maximize one extreme score.
  Discover as many genes as possible whose
absolute experimental effect crosses the predeclared hit
  threshold, while revealing distinct causal
mechanisms.  A region is useful only if its predicted
  positives later beat its own controls.

Threshold-aligned observations from the most recent genes
  actually tested (at most 64):
[omitted: recent observation records; see outcomes above]
```

Round 2: Feedback-driven replanning (continued) Claude Sonnet 5 / IL2

```
Plan hypotheses for these mechanism slots. Counts are
  provisional capacity hints; the bounded adaptive
allocator commits the final 128-gene quotas only after
  hypotheses and candidate pools are frozen:
[{"id": "r1", "action": "expand", "count": 32}, {"id":
  "r2", "action": "exploit", "count": 32}, {"id": "r3",
  "action": "discover", "count": 32}, {"id": "r4",
  "action": "challenge", "count": 32}]

Empirical mechanism portfolio (candidate and control
  outcomes from prior rounds):
{"r1": {"candidate_trials": 30, "candidate_hits": 3,
  "control_trials": 2, "control_hits": 0}, "r2":
  {"candidate_trials": 30, "candidate_hits": 5,
  "control_trials": 2, "control_hits": 0}, "r3":
  {"candidate_trials": 30, "candidate_hits": 1,
  "control_trials": 2, "control_hits": 1}, "r4":
  {"candidate_trials": 30, "candidate_hits": 8,
  "control_trials": 2, "control_hits": 1}}

Persistent mechanism-slot identities from the immediately
  preceding round:
{"r1": {"mechanism_version": 1, "hypothesis": "TCR proximal
  signaling loss reduces IL-2 output", "anchors": ["LCP2",
  "ZAP70", "LAT"]}, "r2": {"mechanism_version": 1,
  "hypothesis": "NFAT/calcineurin pathway boundary controls
  IL-2 magnitude", "anchors": ["PPP3CA", "NFATC2",
  "CALM1"]}, "r3": {"mechanism_version": 1, "hypothesis":
  "Negative feedback via ubiquitin ligases masks true
  hits", "anchors": ["CBLB", "ITCH", "PDCD1"]}, "r4":
  {"mechanism_version": 1, "hypothesis": "Metabolic/mTOR
  sensing independently gates IL-2 transcription",
  "anchors": ["MTOR", "RPTOR", "TSC2"]}}

The exploit region may extend an established mechanism.
  Expand tests its boundary.  Challenge tests
a plausible rival explanation of established hits.
  Discover regions must be mechanistically distinct
from established regions and from one another: do not
  merely rename a pathway or list direct neighbors.
Each region must include a compact falsifiable hypothesis
  of at most 8 words, exactly 3 anchor genes,
and its assigned number of negative-control genes expected
  NOT to hit:
{"r1": 2, "r2": 2, "r3": 2, "r4": 2}. Region IDs are
  persistent
mechanism slots. Set revision=preserve only when the
  hypothesis text is unchanged. Use refine or replace
when the hypothesis changes, and give an evidence-based
  revision_reason; changed hypotheses start a new
mechanism version with fresh allocation statistics. On the
  first round use replace with reason
"initial hypothesis".
For every region, return 3-6 valid HGNC anchor genes. A
  separate Achilles tool will expand anchors only within
  that region. Do not use anchors from one region to
  populate another region.
```

Round 2: Feedback-driven replanning (continued) Claude Sonnet 5 / IL2

```
Return only one JSON object:
{
  "map_update": {"established": [], "uncertain": [],
  "falsified": [], "unexplained_hits": []},
  "regions": [
    {"id": "r1", "action": "expand", "hypothesis":
  "testable causal statement", "revision": "replace",
  "revision_reason": "evidence-based reason", "anchors":
  ["HGNC"], "controls": ["HGNC"]}
  ]
}


Assistant:
Looking at the data: r4 (mTOR/metabolic sensing +
  proteostasis) is massively productive (8/30 hits,
  including very strong ones like TSC1, PRMT1, VCP, ADSS,
  XPO1, RUVBL2, PSMA1/A6). r3 (ubiquitin ligase feedback)
  is nearly dead (1/30, and that hit THEMIS is
  TCR-proximal, not feedback-related) -- this hypothesis is
  essentially falsified and needs replacement with a
  distinct mechanism. r1 and r2 show moderate yields but I
  don't have their gene-level detail here, so I'll keep
  their established hypotheses but adjust action framing
  based on stated rates (r1 3/30 modest, r2 5/30 decent).

```json
{
  "map_update": {
    "established": ["r4: proteostasis/mTOR-adjacent
  nucleocytoplasmic transport and proteasome control IL-2
  (TSC1, PRMT1, VCP, XPO1, RUVBL2, PSMA1, PSMA6, ADSS)"],
    "uncertain": ["r1: TCR proximal signaling boundary (low
  hit rate, needs expansion)", "r2: NFAT/calcineurin
  pathway (moderate hit rate, exploit further)"],
    "falsified": ["r3: ubiquitin ligase negative feedback
  (1/30 hits, hit gene THEMIS is TCR-proximal not a ligase,
  mechanism not supported)"],
    "unexplained_hits": ["THEMIS (r3, TCR-adjacent adaptor,
  not a ubiquitin ligase -- mechanism mismatch)"]
  },
  "regions": [
    {
      "id": "r1",
      "action": "expand",
      "hypothesis": "TCR proximal signaling loss reduces
  IL-2 output",
      "revision": "preserve",
      "revision_reason": "unchanged hypothesis, low but
  nonzero hit rate (3/30) justifies boundary expansion",
      "anchors": ["LCP2", "ZAP70", "LAT"],
      "controls": ["ACTB", "GAPDH"]
    },
    {
      "id": "r2",
      "action": "exploit",
      "hypothesis": "NFAT/calcineurin pathway boundary
```
```

Round 2: Feedback-driven replanning (continued) Claude Sonnet 5 / IL2

```
  controls IL-2 magnitude",
      "revision": "preserve",
      "revision_reason": "unchanged hypothesis, 5/30 hit
  rate supports continued exploitation",
      "anchors": ["PPP3CA", "NFATC2", "CALM1"],
      "controls": ["POLR2A", "TUBB"]
    },
    {
      "id": "r3",
      "action": "discover",
      "hypothesis": "RNA splicing/mRNA processing controls
  IL-2 transcript levels",
      "revision": "replace",
      "revision_reason": "ubiquitin ligase feedback
  hypothesis falsified (1/30 hits, sole hit THEMIS is
  TCR-adaptor not ligase); pivoting to distinct
  RNA-processing mechanism, mechanistically unrelated to
  TCR signaling, calcineurin, or mTOR/proteostasis",
      "anchors": ["SF3B1", "U2AF1", "SRSF1"],
      "controls": ["GUSB", "HPRT1"]
    },
    {
      "id": "r4",
      "action": "challenge",
      "hypothesis": "Proteasome/nuclear transport, not
  mTOR, drives effect",
      "revision": "refine",
      "revision_reason": "strongest hits (VCP, XPO1,
  RUVBL2, PSMA1/A6, ADSS) are
  proteostasis/transport/nucleotide synthesis genes, not
  canonical mTOR components (MTOR/RPTOR/TSC2 themselves
  untested here); testing rival explanation that
  proteasome/export machinery, not mTOR sensing per se, is
  the causal driver",
      "anchors": ["PSMB5", "XPO1", "VCP"],
      "controls": ["B2M", "RPL13A"]
    }
  ]
}
```
END SOURCE LOG EXCERPT

Rounds 2-5 do not have LLM selector prompts:
their within-pool choices use the fitted response
model; its fit, audit and per-gene p/u are above.
The same planner-feedback loop continues in rounds
3-5; their decisions and outcomes are recorded above.

Round 3 Claude Sonnet 5 / IL2

```
ROUND 3: plan -> expand -> allocate -> select -> assay

Previous round candidate feedback hits:
  r1: 6/14
  r2: 5/14
  r3: 26/46
  r4: 22/46

Selection mode: experiment-trained-pool-wise
Allocation mode: bounded-adaptive
Pipeline: hypothesis_planning, candidate_expansion,
  adaptive_allocation, pool_wise_selection,
  experiment_and_update
Planning quotas / actual quotas / pool-priority:
  r1: 32 / 32 / 0.791
  r2: 32 / 32 / 0.791
  r3: 32 / 32 / 0.791
  r4: 32 / 32 / 0.791

Hypotheses, revisions, and observed response:
r1 (discover; 30 candidates + 2 controls)
  Hypothesis: Cytokine secretion/vesicle trafficking limits
    IL-2 release
  Anchors: RAB27A, STX4, VAMP7
  Controls: MYH9, RPL37A
  Revision: replace - TCR proximal signaling gave only 20%
    hit rate; testing distinct secretory mechanism
  Feedback: 10/30 candidate; 1/2 control
r2 (challenge; 30 candidates + 2 controls)
  Hypothesis: Ribosomal/translation loss mimics hits via
    general fitness
  Anchors: RPL5, RPS6KB1, EEF2
  Controls: RPL13, RPS18
  Revision: replace - r3/r4 hits dominated by RPL/RPS
    genes; testing if effect is IL-2-specific or general
    essentiality confound
  Feedback: 11/30 candidate; 2/2 control
r3 (exploit; 30 candidates + 2 controls)
  Hypothesis: Ribosomal protein loss reduces IL-2 via
    translation capacity
  Anchors: RPL10A, RPS13, RPL37
  Controls: UBC, YWHAZ
  Revision: refine - 57% hit rate concentrated in RPL/RPS
    genes, not general splicing factors
  Feedback: 17/30 candidate; 0/2 control
r4 (expand; 30 candidates + 2 controls)
  Hypothesis: Cell-cycle/APC and nuclear pore genes boost
    IL-2 loss
  Anchors: CDC23, NUP133, MCM2
  Controls: PGK1, LDHA
  Revision: refine - hits cluster on APC/C, nucleoporins,
    replication genes, not proteasome/XPO1 as predicted
  Feedback: 15/30 candidate; 0/2 control

Execution details:
  r1 control repair: RPL37A
  r2 control repair: RPS18
```

Round 3 (continued) Claude Sonnet 5 / IL2

```
  r3 control repair: UBC, YWHAZ
  r4 control repair: PGK1, LDHA
  Anchor repairs: none
  Control fallbacks: none
  Candidate-pool expansions: 0
  Common-essential cap enabled: False
  Planner fallback: none

Feedback-trained response selector:
  Fit: all-score multiview ridge logistic; 256 measured
  genes; 79 internal positives; threshold 0.32403.
  Views: Achilles, Reactome, SIGNOR; frozen features,
  coefficients learned from measured assay scores.
  On-policy candidate audit: n=120; observed hit
  rate=0.442; Brier=0.251.

Candidate pools before selection (recorded order):
  r1: 92 candidates
  KRTAP22-2, ZNF880, ZNF669, DTWD1, ZNF30, ODAM, RGS1
  C12orf40, FAM83B, PROM1, STXBP3, SNAP23, GRHL2, RAB10
  NMT1, RAB6A, MYBL2, RABIF, PIK3CA, ERBB2, SPRY3, COL4A6
  TBC1D19, COL4A5, GUCY2F, STX17, CACNG7, OCM, HMHB1
  ZNF611, UPF1, CAP1, DNM2, PTPN23, ATP6AP2, UXT, BCAS2
  WNK1, ATP6AP1, ATP1A1, TFAP4, NELL1, CAPZB, PPIL1, CLTC
  ACSM2B, WDR26, ATP6V1G1, TSG101, IFNE, GRIA2, CCDC130
  PPP1R15B, ZRSR2, PPEF1, RPA2, TMEM175, ARIH1, GRK2
  TBC1D26, HNRNPM, FAM107B, INS, OLR1, NUDC, SCAP, RPAIN
  VAMP5, UBAP1, MCM3AP, BNIP3L, ZNF853, CENPM, MAPK8
  COL28A1, NAPA, MRPL34, NXF1, CASC3, RAB41, ATXN3L, IGBP1
  SCLT1, EPC1, DEFB124, ZNF35, CDC37, DMRT2, C2orf72, IK
  TEAD3, WDR60
  r2: 92 candidates
  LSM2, CCT3, RBM22, YARS, CUEDC1, NME1, SCPEP1, PRR29
  FBF1, PRR15L, UBE2O, C17orf58, OR4D1, MPO, SNRNP200
  EIF2S1, HNRNPK, SF3A1, SNRPD1, RSL24D1, TBC1D3B, HSPE1
  DHX37, PHF5A, ANKLE2, PRPF8, FARSB, AARS, PRELID1, DDX56
  CCT4, SF3A2, SNRPB, POLR1C, INTS11, POLR3A, CDC123, MAK16
  TSR1, SF1, XAB2, GPN1, ERH, SRSF2, DDX54, SNRPF, COPB1
  RRM1, CHAF1B, NUDT21, KARS, WDR43, RABGGTB, DDX49, SNRPA1
  SDE2, BUD31, BIRC5, PUF60, POLR1A, UTP15, SNRPE, NPIPB6
  SNRPD2, SMU1, PRPF31, RPL3, EIF3A, NOL10, TXNL4A, VPS25
  SF3B4, MED11, CTCF, DBR1, DDX10, RRP12, ATP6V0C, CLNS1A
  LUC7L3, DDX41, RPL14, UBTF, SRSF3, PFDN2, NARS, SF3B5
  HARS, COPB2, PRPF38A, ISCU, SNW1
  r3: 92 candidates
  NIP7, CDK7, RNPC3, TRMT112, DUX4, RPL15, DONSON, RPS8
  CDCA8, EIF5, MED30, ZMAT2, FBL, LRR1, BUB3, DUT, PLK1
  SFPQ, NCBP2, NCAPG, TSR2, U2AF2, WDR74, SRSF7, SS18L2
  DNAJC17, UBL5, BYSL, PRPF38B, RPLP1, PDRG1, CPSF1, GPN3
  SAP30BP, ANAPC4, RPL10, NFS1, USP5, QARS, CPSF4, RABGGTA
  SF3B3, INTS3, SPOUT1, RPL27A, FOXD4, IMP3, NUTF2, EIF1AD
  XRCC6, SNRPC, RPL35, CARS, HSPA5, ZPR1, WDR5, ATP6V0B
  PMF1, CDK11A, PSMC5, DHX15, WDR75, ZNHIT2, RACGAP1, PPWD1
  DHX8, RPS3, MMS22L, HIST1H2BJ, SAP18, EIF3G, EIF4A3
  SNRNP25, PSMB2, IMP4, INTS4, RPS14, ZNF830, NSF, COPS6
  MIS18A, WDR82, EIF3I, RAD21, TONSL, GEMIN5, RSL1D1, HAUS5
  SF3B1, MED8, CNIH4, RBM25
  r4: 92 candidates
```

Round 3 (continued) Claude Sonnet 5 / IL2

```
  MARS, PHB, NSA2, MCM5, POLR3H, DYNC1I2, CSE1L, TRAPPC3
  LSM7, SRP54, GPN2, SNRNP70, ATP6V1F, TTC27, IARS, RPA3
  SMC3, AHCTF1, MCM7, CPSF2, GTF2A2, BUD23, YKT6, CENPK
  FCF1, DHDDS, TRNT1, HSPD1, CLP1, NDC80, INTS9, SNRNP27
  CENPN, UTP20, BANF1, CCT2, RPS4X, GINS4, CHAF1A, TAF6
  POLD1, SNAPC1, RPS15, PSMD4, TANGO6, TOP2A, EPRS, BMS1
  SYMPK, THOC7, DAD1, SNRPD3, PPAN, RPAP1, SNRNP35, UBA2
  POLR2G, DDX55, TOMM40, SNAPC2, DARS, SMC1A, POLR1B, KIF23
  ABCE1, RFC5, ATP2A2, LSM3, TNPO3, INTS1, PRC1, WDR12
  NOC3L, NVL, SSRP1, TRRAP, LSM8, PSMD3, PSMD6, CHERP, SLU7
  NAA50, URB1, SBNO1, ATP6V1A, WBP11, CENPW, TUT1, PELP1
  GUK1, DKC1, EIF2B2

All selected genes and measured outcomes:
  Columns: gene score flags; rounds 2-5 also p/u.
  F=feedback hit, A=GT All, N=GT N/E, E=CEG,
  C=control; p=predicted response, u=uncertainty.
r1: 32 tested
  01 MYH9       -0.06413 ----C p=-- u=--
  02 RPL37A     +0.73140 FA-EC p=-- u=--
  03 HNRNPM     +0.23661 ----- p=0.530 u=0.184
  04 RPA2       +0.20762 ---E- p=0.519 u=0.068
  05 ARIH1      +0.14951 ---E- p=0.300 u=0.045
  06 WNK1       -0.11798 ----- p=0.282 u=0.043
  07 CLTC       +0.25547 ---E- p=0.272 u=0.038
  08 TSG101     +0.88005 FAN-- p=0.271 u=0.038
  09 UPF1       +0.70601 FA-E- p=0.270 u=0.038
  10 UBAP1      +0.26599 ----- p=0.268 u=0.037
  11 NUDC       +0.75578 FA-E- p=0.267 u=0.038
  12 ATP6V1G1   +0.08852 ----- p=0.262 u=0.038
  13 NXF1       +0.53730 FA-E- p=0.262 u=0.039
  14 CDC37      -0.08625 ---E- p=0.261 u=0.039
  15 CASC3      +0.01550 ----- p=0.259 u=0.032
  16 IK         +0.05566 ----- p=0.258 u=0.039
  17 DNM2       +0.57393 FA-E- p=0.257 u=0.037
  18 ZRSR2      +0.06166 ----- p=0.257 u=0.039
  19 CAPZB      -0.08522 ----- p=0.257 u=0.039
  20 MCM3AP     +1.03070 FA-E- p=0.255 u=0.039
  21 RPAIN      -0.03809 ----- p=0.253 u=0.039
  22 ATP6AP2    -0.04607 ----- p=0.252 u=0.039
  23 NAPA       -0.03054 ---E- p=0.252 u=0.039
  24 GRK2       +0.07935 ----- p=0.251 u=0.070
  25 PPIL1      +0.51173 FAN-- p=0.251 u=0.037
  26 PTPN23     +0.86963 FA-E- p=0.251 u=0.039
  27 IGBP1      +0.06971 ---E- p=0.250 u=0.039
  28 ATP6AP1    +0.27823 ----- p=0.250 u=0.039
  29 BCAS2      +0.56641 FAN-- p=0.250 u=0.037
  30 CAP1       +0.43024 FAN-- p=0.249 u=0.036
  31 MRPL34     +0.26492 ----- p=0.249 u=0.038
  32 UXT        +0.02720 ---E- p=0.249 u=0.038
r2: 32 tested
  01 RPL13      +0.64954 FA-EC p=-- u=--
  02 RPS18      +0.45801 FA-EC p=-- u=--
  03 RPL3       +0.42418 FA-E- p=0.533 u=0.056
  04 RPL14      +0.60828 FA-E- p=0.532 u=0.056
  05 EIF3A      +0.18986 ---E- p=0.332 u=0.036
  06 EIF2S1     +0.01586 ---E- p=0.303 u=0.044
  07 BIRC5      +0.29587 ---E- p=0.289 u=0.038
```

Round 3 (continued) Claude Sonnet 5 / IL2

```
  08 SNRPF      +0.57494 FA-E- p=0.287 u=0.038
  09 SNRPB      +0.35521 FAN-- p=0.287 u=0.038
  10 SNRPE      +0.23781 ----- p=0.287 u=0.038
  11 SNRPD1     +0.33868 F--E- p=0.283 u=0.038
  12 SNRPD2     +0.39007 FA-E- p=0.282 u=0.038
  13 SNW1       +0.27949 ---E- p=0.278 u=0.038
  14 XAB2       +0.41351 FA-E- p=0.276 u=0.039
  15 WDR43      +0.24109 ---E- p=0.271 u=0.039
  16 DHX37      +0.28540 ---E- p=0.271 u=0.039
  17 UTP15      +0.58644 FA-E- p=0.271 u=0.039
  18 DDX49      +0.21527 ---E- p=0.271 u=0.039
  19 HNRNPK     -0.04235 ---E- p=0.271 u=0.039
  20 TSR1       +0.22836 ---E- p=0.270 u=0.039
  21 SRSF2      -0.24000 ---E- p=0.269 u=0.039
  22 SF3B4      +0.18031 ----- p=0.268 u=0.039
  23 SF3B5      +0.26283 ---E- p=0.268 u=0.039
  24 SNRNP200   +0.77563 FA-E- p=0.267 u=0.039
  25 ATP6V0C    +0.06285 ---E- p=0.267 u=0.039
  26 PRPF8      +0.44952 FA-E- p=0.266 u=0.039
  27 PHF5A      +0.33530 F--E- p=0.266 u=0.039
  28 SF3A1      +0.16884 ----- p=0.266 u=0.039
  29 TXNL4A     -0.25048 ---E- p=0.266 u=0.039
  30 PUF60      -0.07018 ---E- p=0.266 u=0.039
  31 SNRPA1     +0.05698 ---E- p=0.266 u=0.039
  32 SRSF3      +0.00737 ---E- p=0.266 u=0.039
r3: 32 tested
  01 UBC        +0.05580 ----C p=-- u=--
  02 YWHAZ      -0.05288 ----C p=-- u=--
  03 PSMB2      +0.17036 ---E- p=0.642 u=0.074
  04 PSMC5      +0.34163 FA-E- p=0.641 u=0.073
  05 RPS3       +0.54166 FA-E- p=0.607 u=0.066
  06 RPS8       +0.68948 FA-E- p=0.594 u=0.066
  07 RPS14      +0.59932 FAN-- p=0.594 u=0.067
  08 RPL27A     +0.27340 ---E- p=0.540 u=0.058
  09 RPL15      +0.92308 FAN-- p=0.532 u=0.056
  10 RPLP1      +0.66193 FA-E- p=0.532 u=0.056
  11 RPL10      +0.71376 FAN-- p=0.532 u=0.056
  12 RPL35      +0.98263 FA-E- p=0.531 u=0.056
  13 CDK7       +0.20938 ---E- p=0.435 u=0.078
  14 PLK1       +0.33065 F--E- p=0.430 u=0.075
  15 ANAPC4     +0.48695 FA-E- p=0.369 u=0.036
  16 NCBP2      +0.00488 ---E- p=0.361 u=0.036
  17 EIF3I      +0.44982 FA-E- p=0.332 u=0.036
  18 EIF3G      +0.08066 ---E- p=0.331 u=0.036
  19 BUB3       +0.48196 FA-E- p=0.304 u=0.037
  20 EIF4A3     +0.33471 F--E- p=0.304 u=0.037
  21 DUT        +0.44096 FAN-- p=0.292 u=0.043
  22 EIF5       -0.12001 ----- p=0.286 u=0.038
  23 TRMT112    +0.07340 ---E- p=0.281 u=0.038
  24 WDR5       +0.81915 FAN-- p=0.278 u=0.038
  25 CPSF4      +1.12870 FA-E- p=0.277 u=0.038
  26 CDCA8      +0.16746 ----- p=0.274 u=0.039
  27 U2AF2      +0.64932 FA-E- p=0.273 u=0.039
  28 GEMIN5     +0.00313 ---E- p=0.272 u=0.039
  29 NIP7       +0.21907 ---E- p=0.271 u=0.039
  30 ZNF830     +0.12813 ----- p=0.270 u=0.039
  31 IMP4       +0.15595 ---E- p=0.270 u=0.039
  32 WDR75      +0.31231 ---E- p=0.270 u=0.039
```

Round 3 (continued) Claude Sonnet 5 / IL2

```
r4: 32 tested
  01 PGK1        -0.15864 ---EC p=-- u=--
  02 LDHA        -0.05393 ----C p=-- u=--
  03 PSMD3       +0.26707 ---E- p=0.641 u=0.074
  04 PSMD6       +0.42751 FAN-- p=0.640 u=0.074
  05 RPS4X       +0.70020 FA-E- p=0.593 u=0.066
  06 RPS15       +0.27580 ----- p=0.592 u=0.066
  07 MCM7        +0.00874 ---E- p=0.452 u=0.100
  08 RPA3        -0.01871 ----- p=0.429 u=0.041
  09 RFC5        +0.44367 FA-E- p=0.412 u=0.039
  10 POLR2G      +0.40972 FA-E- p=0.373 u=0.036
  11 POLD1       +0.30433 ----- p=0.369 u=0.036
  12 MCM5        +0.43350 FA-E- p=0.319 u=0.036
  13 DYNC1I2     +0.62441 FA-E- p=0.298 u=0.037
  14 BANF1       +0.26762 ---E- p=0.285 u=0.039
  15 SNRPD3      +0.03772 ---E- p=0.285 u=0.038
  16 SSRP1       -0.04000 ----- p=0.283 u=0.038
  17 ABCE1       +0.27711 ---E- p=0.277 u=0.038
  18 DAD1        -0.41773 FA-E- p=0.276 u=0.038
  19 GINS4       +0.58941 FA-E- p=0.276 u=0.038
  20 DKC1        +0.39585 FA-E- p=0.275 u=0.038
  21 CENPK       -0.01856 ----- p=0.274 u=0.038
  22 CENPN       +0.39802 FAN-- p=0.274 u=0.038
  23 AHCTF1      -0.16032 ----- p=0.271 u=0.038
  24 TAF6        -0.13790 ---E- p=0.271 u=0.039
  25 NDC80       +0.04924 ---E- p=0.271 u=0.039
  26 SMC3        +0.95558 FAN-- p=0.271 u=0.038
  27 SMC1A       +0.68087 FA-E- p=0.271 u=0.038
  28 PELP1       +0.19325 ----- p=0.270 u=0.039
  29 UTP20       +0.32990 F--E- p=0.270 u=0.039
  30 BUD23       +0.24384 ---E- p=0.270 u=0.039
  31 BMS1        +0.55990 FAN-- p=0.270 u=0.039
  32 FCF1        +0.39702 FAN-- p=0.270 u=0.039

Round outcome:
  128 new tests; internal feedback hits: 56
  New GT: All 51; N/E 15
  Cumulative GT: All 127/654; N/E 47/481
  Cumulative selected CEG: 197/384
```

Round 4 Claude Sonnet 5 / IL2

```
ROUND 4: plan -> expand -> allocate -> select -> assay

Previous round candidate feedback hits:
  r1: 10/30
  r2: 11/30
  r3: 17/30
  r4: 15/30

Selection mode: experiment-trained-pool-wise
Allocation mode: bounded-adaptive
Pipeline: hypothesis_planning, candidate_expansion,
  adaptive_allocation, pool_wise_selection,
```

Round 4 (continued) Claude Sonnet 5 / IL2

```
  experiment_and_update
Planning quotas / actual quotas / pool-priority:
  r1: 32 / 38 / 1.152
  r2: 32 / 37 / 1.152
  r3: 32 / 16 / 0.680
  r4: 32 / 37 / 1.152

Hypotheses, revisions, and observed response:
r1 (challenge; 36 candidates + 2 controls)
  Hypothesis: Secretory/vesicle trafficking loss lowers
    IL-2 secretion
  Anchors: STX11, UNC13D, SYT7
  Controls: SDHA, IPO8
  Revision: replace - Prior trafficking anchors
    underperformed (10/30 hits, low specificity); testing
    rival mechanism: exocytic SNARE machinery vs general
    ribosomal confound
  Feedback: 10/36 candidate; 0/2 control
r2 (discover; 35 candidates + 2 controls)
  Hypothesis: Mitochondrial OXPHOS dysfunction impairs IL-2
    induction
  Anchors: NDUFA9, SDHB, UQCRC1
  Controls: HMBS, C1orf43
  Revision: replace - Ribosomal-mimicry hypothesis
    confounded with r3 findings; pivoting to distinct
    metabolic mechanism not yet tested
  Feedback: 1/35 candidate; 0/2 control
r3 (exploit; 14 candidates + 2 controls)
  Hypothesis: Ribosomal protein loss reduces IL-2 via
    translation capacity
  Anchors: RPL10A, RPS13, RPL37
  Controls: CHMP2A, VPS29
  Revision: preserve - Strong hit rate (17/30) confirms
    mechanism; extend to remaining ribosomal
    paralogs/subunits
  Feedback: 5/14 candidate; 1/2 control
r4 (expand; 35 candidates + 2 controls)
  Hypothesis: Cell-cycle/APC and nuclear pore genes boost
    IL-2 loss
  Anchors: ESPL1, STAG2, PDS5B
  Controls: EMC7, REXO4
  Revision: refine - High hit rate (15/30) but mixed with
    cohesin/replication genes; narrow boundary test to
    chromosome segregation/cohesin complex specifically
  Feedback: 9/35 candidate; 0/2 control

Execution details:
  r1 control repair: SDHA, IPO8
  r2 control repair: HMBS, C1orf43
  r3 control repair: CHMP2A, VPS29
  r4 control repair: EMC7, REXO4
  Anchor repairs: none
  Control fallbacks: none
  Candidate-pool expansions: 0
  Common-essential cap enabled: False
  Planner fallback: none

Feedback-trained response selector:
```

Round 4 (continued) Claude Sonnet 5 / IL2

```
  Fit: all-score multiview ridge logistic; 384 measured
  genes; 135 internal positives; threshold 0.32403.
  Views: Achilles, Reactome, SIGNOR; frozen features,
  coefficients learned from measured assay scores.
  On-policy candidate audit: n=120; observed hit
  rate=0.208; Brier=0.168.

Candidate pools before selection (recorded order):
  r1: 92 candidates
  CARD14, CLVS2, WISP3, LRFN2, CTGF, PRKCG, HDAC9, SLC35D3
  ANOS1, OLR1, CCDC137, TEPSIN, RAC3, SLC25A10, CEP131
  SF3B6, PDK2, SLC35G3, LRRC37A3, NANOG, SDHAF2, VPS51
  NELFB, CCDC84, KLC2, CD5, ACTR2, INCENP, TAF1C, EIF4E
  GTF2A2, SCGB1D1, TIMM10, DDX41, TMEM258, NVL, SF3B2
  NOL10, ANKRD36B, ALG1, EIF1AD, RASGRP2, SNRNP70, SLC22A25
  CBWD2, LTBP3, NOP16, NFS1, COPS6, NCL, NPIPB6, ZPR1, EPRS
  ACTR8, RPL36, ALG2, SRP72, NOL9, HYOU1, LSM3, POLR1B
  ALG13, MAP3K11, HAUS5, SCGB2A1, TSR2, SBDS, NSA2, VPS37C
  TAF7, FBL, KAT8, DDOST, CCDC86, ISCU, TBC1D3B, ARCN1
  CCDC115, AFG3L2, ZNHIT2, BUD31, CIAO1, CHERP, PPA1
  DNAJC17, RRN3, NOC3L, MPHOSPH10, LSM2, ELP6, ELP5, INTS4
  r2: 92 candidates
  NDUFS5, NDUFB4, NDUFB3, NDUFA11, NDUFA2, COX7C, NDUFB8
  MRPL38, CYCS, SDHC, ADSL, GMPS, SDHD, OGDH, TFRC, UQCRC2
  CYC1, UQCRB, UQCRFS1, COX5B, COX6C, MRPL21, MRPS11
  MRPL33, BCS1L, MRPL10, MRPS23, HARS2, MRPL15, MRPL43
  COQ4, COA7, MRPL41, YARS2, MRPS24, MRPS16, PTCD3, MRPL16
  ISCA2, UQCRQ, MRPS14, MRPL57, MARS2, LRPPRC, MRPL55
  MRPS12, RARS2, COX17, NDUFB6, MRPS5, MRPL4, MRPL53
  MRPS22, FARS2, MRPL13, SUPV3L1, MRPL39, TARS2, KIAA0391
  MRPL18, MRPL14, MRPL12, MRPS28, MRPL32, MRPL28, MRPL2
  MRPS21, AARS2, GFM1, MRPS9, NDUFA6, EARS2, SARS2, NDUFAF3
  NDUFS2, MTG2, MRPL27, MRPS18B, MTPAP, MRPL49, METTL17
  MRPS26, VARS2, MRPL35, MRPL3, MRPS10, MRPS27, LARS2
  COX15, QRSL1, HSPE1, TFAM
  r3: 92 candidates
  CDC123, RSL24D1, FARSB, PRELID1, RNPC3, DUX4, DBR1
  DONSON, SF1, RBM22, ANKLE2, CCT3, MED30, RRM1, NUDT21
  COPB1, POLR1C, CCT4, ZMAT2, PRPF38A, LRR1, AARS, COPB2
  SFPQ, KARS, DDX56, CHAF1B, POLR3A, CTCF, NCAPG, MAK16
  INTS11, WDR74, SRSF7, MED11, SS18L2, PRPF31, SDE2, YARS
  UBTF, SMU1, UBL5, BYSL, PRPF38B, PDRG1, CPSF1, GPN3
  SAP30BP, SF3A2, PFDN2, USP5, QARS, RABGGTA, SF3B3, LUC7L3
  INTS3, POLR1A, SPOUT1, FOXD4, IMP3, NUTF2, XRCC6, SNRPC
  CARS, HSPA5, ATP6V0B, PMF1, CDK11A, VPS25, DHX15, RACGAP1
  PPWD1, DHX8, MMS22L, HIST1H2BJ, SAP18, SNRNP25, DDX54
  HARS, NSF, MIS18A, WDR82, ERH, RAD21, TONSL, RSL1D1
  SF3B1, MED8, CNIH4, RBM25, SRP54, GPN1
  r4: 92 candidates
  PAXIP1, PAGR1, MAU2, PPP2R3C, FGFR1OP, ACTR3, RAC1, RSU1
  MOB4, ARPC3, ARPC4, VEZT, TLN1, TUBA1C, DNTTIP2, MARS
  POLR3H, KIF23, TRRAP, PRC1, TOP2A, GPN2, RRP12, GUK1
  SNRNP27, SNAPC1, SNAPC2, URB1, PAM16, PHB, SNAPC5, DDX55
  CLP1, RPAP1, PMPCB, YKT6, TRNT1, THOC7, TANGO6, BUB1B
  CENPW, SRBD1, LSM7, NARS, RBBP5, UBA2, TRAPPC3, INTS9
  SYMPK, RPS25, HSPD1, PPAN, NAA50, TTC27, MDN1, NMD3
  ATP6V1F, NUS1, TCP1, CNOT3, DHDDS, PDCD11, OSGEP, CRCP
  WBP11, DDX10, MASTL, PSMG4, DMAP1, ATP6V1A, DYNLRB1
```

Round 4 (continued) Claude Sonnet 5 / IL2

```
  HAUS1, CPSF2, ZMAT5, TNPO3, IARS, MCM6, EXOSC6, PFDN6
  PWP2, NACA, EXOSC4, SPC25, BOP1, GTF2E1, TOP3A, POP5
  POLR3B, PSMD4, CHAF1A, CLNS1A, SBNO1

All selected genes and measured outcomes:
  Columns: gene score flags; rounds 2-5 also p/u.
  F=feedback hit, A=GT All, N=GT N/E, E=CEG,
  C=control; p=predicted response, u=uncertainty.
r1: 38 tested
  01 SDHA       -0.01729 ----C p=-- u=--
  02 IPO8       -0.03072 ----C p=-- u=--
  03 RPL36      +0.71602 FA-E- p=0.642 u=0.054
  04 EIF4E      +0.03554 ----- p=0.396 u=0.029
  05 NCL        +0.51324 FAN-- p=0.327 u=0.029
  06 DDOST      -0.19635 ---E- p=0.325 u=0.030
  07 INCENP     +0.17468 ----- p=0.324 u=0.030
  08 TMEM258    -0.14237 ---E- p=0.324 u=0.030
  09 NOL9       +0.11239 ---E- p=0.324 u=0.030
  10 SF3B2      +0.28804 ---E- p=0.322 u=0.030
  11 HAUS5      +0.25107 ---E- p=0.319 u=0.031
  12 CHERP      +0.48800 FA-E- p=0.319 u=0.031
  13 BUD31      +0.08475 ---E- p=0.317 u=0.031
  14 SF3B6      +0.08784 ----- p=0.317 u=0.030
  15 CEP131     +0.05877 ----- p=0.317 u=0.030
  16 MPHOSPH10  +0.46522 FA-E- p=0.316 u=0.046
  17 SRP72      -0.54843 FAN-- p=0.316 u=0.030
  18 VPS37C     -0.13273 ----- p=0.315 u=0.028
  19 FBL        +0.40312 FAN-- p=0.313 u=0.060
  20 ISCU       -0.01270 ---E- p=0.312 u=0.031
  21 LSM2       -0.17720 ---E- p=0.311 u=0.032
  22 HYOU1      -0.17571 ----- p=0.310 u=0.031
  23 COPS6      -0.08211 ---E- p=0.310 u=0.031
  24 NELFB      -0.19657 ---E- p=0.310 u=0.031
  25 SNRNP70    +0.33315 F--E- p=0.310 u=0.031
  26 DDX41      +0.46584 FA-E- p=0.310 u=0.031
  27 TAF7       -0.36679 FAN-- p=0.309 u=0.031
  28 LSM3       +0.04563 ----- p=0.308 u=0.031
  29 ACTR8      -0.11467 ----- p=0.307 u=0.031
  30 TBC1D3B    +0.12835 ----- p=0.305 u=0.032
  31 KAT8       +0.46678 FA-E- p=0.305 u=0.031
  32 NPIPB6     -0.14642 ----- p=0.304 u=0.032
  33 DNAJC17    +0.04130 ----- p=0.304 u=0.032
  34 ACTR2      -0.06208 ---E- p=0.304 u=0.031
  35 TSR2       +0.18577 ----- p=0.304 u=0.032
  36 EIF1AD     +0.19368 ----- p=0.304 u=0.032
  37 ZPR1       +0.15403 ---E- p=0.303 u=0.032
  38 NOL10      +0.05410 ---E- p=0.303 u=0.032
r2: 37 tested
  01 HMBS       -0.05968 ----C p=-- u=--
  02 C1orf43    -0.12733 ----C p=-- u=--
  03 HSPE1      -0.02808 ----- p=0.307 u=0.032
  04 CYCS       +0.01360 ---E- p=0.306 u=0.031
  05 MRPS21     -0.00918 ----- p=0.304 u=0.031
  06 YARS2      +0.05522 ---E- p=0.304 u=0.031
  07 MRPL27     +0.08298 ----- p=0.303 u=0.031
  08 MRPL21     -0.04163 ----- p=0.303 u=0.031
  09 MRPL4      +0.02675 ---E- p=0.301 u=0.030
  10 MRPL10     -0.09543 ----- p=0.301 u=0.030
```

Round 4 (continued) Claude Sonnet 5 / IL2

```
  11 COX7C      +0.11069 ----- p=0.301 u=0.030
  12 MRPL53     -0.07357 ---E- p=0.301 u=0.030
  13 MRPS16     +0.19963 ----- p=0.300 u=0.030
  14 SUPV3L1    -0.16140 ---E- p=0.300 u=0.031
  15 MRPL49     -0.01425 ----- p=0.300 u=0.030
  16 MRPL12     +0.06556 ----- p=0.299 u=0.030
  17 TARS2      -0.03604 ----- p=0.299 u=0.030
  18 MRPS27     +0.11619 ----- p=0.299 u=0.030
  19 MRPL55     +0.12334 ----- p=0.299 u=0.030
  20 MRPL33     +0.24698 ----- p=0.299 u=0.030
  21 MRPL35     +0.09726 ----- p=0.298 u=0.030
  22 MRPS23     +0.30109 ----- p=0.298 u=0.030
  23 SDHD       +0.19894 ----- p=0.298 u=0.031
  24 PTCD3      +0.06218 ----- p=0.298 u=0.030
  25 MRPL43     -0.17870 ---E- p=0.298 u=0.030
  26 COX5B      +0.07463 ---E- p=0.298 u=0.030
  27 MRPL38     +0.20809 ---E- p=0.298 u=0.030
  28 SARS2      -0.07542 ---E- p=0.297 u=0.030
  29 MRPL57     +0.19446 ---E- p=0.297 u=0.030
  30 MRPL16     -0.01350 ----- p=0.297 u=0.030
  31 MARS2      +0.04652 ---E- p=0.297 u=0.030
  32 COX17      +0.18560 ----- p=0.297 u=0.031
  33 MRPL39     +0.33297 F---- p=0.296 u=0.030
  34 MRPS5      +0.05122 ----- p=0.296 u=0.030
  35 MRPS12     +0.05508 ----- p=0.296 u=0.030
  36 MRPL18     +0.12930 ---E- p=0.296 u=0.029
  37 MRPS22     -0.01965 ----- p=0.296 u=0.030
r3: 16 tested
  01 CHMP2A     +0.05745 ---EC p=-- u=--
  02 VPS29      -0.75477 FAN-C p=-- u=--
  03 SRSF7      -0.23586 ---E- p=0.384 u=0.202
  04 RAD21      +0.08024 ---E- p=0.352 u=0.033
  05 CPSF1      +0.08828 ---E- p=0.330 u=0.030
  06 IMP3       +0.42437 FA-E- p=0.327 u=0.030
  07 BYSL       +0.04345 ---E- p=0.325 u=0.030
  08 PMF1       +0.43751 FAN-- p=0.324 u=0.030
  09 SF3B3      +0.36213 FA-E- p=0.324 u=0.031
  10 NUDT21     +0.16965 ---E- p=0.323 u=0.031
  11 SF3A2      +0.64724 FA-E- p=0.320 u=0.031
  12 SRP54      -0.19208 ----- p=0.320 u=0.030
  13 DHX15      +0.25874 ---E- p=0.320 u=0.031
  14 SF3B1      +0.30481 ---E- p=0.319 u=0.031
  15 RBM22      +0.61349 FAN-- p=0.317 u=0.031
  16 VPS25      -0.07726 ---E- p=0.315 u=0.031
r4: 37 tested
  01 EMC7       +0.08015 ----C p=-- u=--
  02 REXO4      -0.29464 ----C p=-- u=--
  03 RPS25      +0.49069 FAN-- p=0.703 u=0.059
  04 BUB1B      -0.02411 ---E- p=0.357 u=0.029
  05 RBBP5      -0.05938 ----- p=0.355 u=0.029
  06 MCM6       +0.32802 F---- p=0.348 u=0.029
  07 TUBA1C     -0.10361 ----- p=0.338 u=0.027
  08 SYMPK      +0.67820 FA-E- p=0.327 u=0.030
  09 CPSF2      +0.57743 FA-E- p=0.327 u=0.030
  10 PDCD11     +0.43068 FAN-- p=0.327 u=0.030
  11 PWP2       +0.01529 ---E- p=0.327 u=0.030
  12 EXOSC4     +0.25170 ---E- p=0.326 u=0.030
  13 EXOSC6     +0.02358 ---E- p=0.326 u=0.030
```

Round 4 (continued) — Claude Sonnet 5 / IL2

```
  14 BOP1       +0.50964 FAN-- p=0.324 u=0.030
  15 CLP1       +0.05566 ---E- p=0.324 u=0.030
  16 TOP3A      -0.07420 ----- p=0.321 u=0.030
  17 HAUS1      -0.01636 ---E- p=0.318 u=0.030
  18 WBP11      +0.13133 ----- p=0.315 u=0.031
  19 KIF23      +0.25644 ---E- p=0.315 u=0.031
  20 TOP2A      +0.13454 ---E- p=0.313 u=0.031
  21 SPC25      +0.30707 ---E- p=0.313 u=0.049
  22 ATP6V1F    +0.03149 ----- p=0.311 u=0.031
  23 ATP6V1A    -0.06805 ---E- p=0.311 u=0.031
  24 TRRAP      +0.06885 ---E- p=0.311 u=0.031
  25 CNOT3      +0.43689 FA-E- p=0.310 u=0.031
  26 MASTL      +0.17139 ---E- p=0.309 u=0.031
  27 ZMAT5      +0.19000 ---E- p=0.309 u=0.031
  28 SNRNP27    +0.62990 FA-E- p=0.309 u=0.031
  29 NACA       +0.04307 ----- p=0.309 u=0.031
  30 LSM7       +0.09202 ---E- p=0.309 u=0.031
  31 THOC7      +0.17317 ----- p=0.308 u=0.031
  32 CLNS1A     +0.46589 FA-E- p=0.307 u=0.031
  33 TRNT1      +0.08753 ----- p=0.307 u=0.031
  34 CENPW      +0.02623 ----- p=0.306 u=0.031
  35 HSPD1      -0.01716 ---E- p=0.306 u=0.031
  36 GTF2E1     -0.04720 ----- p=0.306 u=0.031
  37 POP5       +0.20184 ---E- p=0.306 u=0.031

Round outcome:
  128 new tests; internal feedback hits: 26
  New GT: All 23; N/E 10
  Cumulative GT: All 150/654; N/E 57/481
  Cumulative selected CEG: 261/512
```

Round 5 — Claude Sonnet 5 / IL2

```
ROUND 5: plan -> expand -> allocate -> select -> assay

Previous round candidate feedback hits:
  r1: 10/36
  r2: 1/35
  r3: 5/14
  r4: 9/35

Selection mode: experiment-trained-pool-wise
Allocation mode: bounded-adaptive
Pipeline: hypothesis_planning, candidate_expansion,
  adaptive_allocation, pool_wise_selection,
  experiment_and_update
Planning quotas / actual quotas / pool-priority:
  r1: 32 / 32 / 0.791
  r2: 32 / 32 / 0.791
  r3: 32 / 32 / 0.791
  r4: 32 / 32 / 0.791

Hypotheses, revisions, and observed response:
r1 (expand; 30 candidates + 2 controls)
```

Round 5 (continued) Claude Sonnet 5 / IL2

```
  Hypothesis: Vesicle tethering/exocytosis machinery loss
    blocks IL-2 release
  Anchors: STX11, UNC13D, STXBP2
  Controls: ACTG1, KRT8
  Revision: refine - 10/36 hit rate strong; testing
    boundary genes beyond core SNARE/Ca2+ sensors into
    tethering complexes
  Feedback: 12/30 candidate; 0/2 control
r2 (discover; 30 candidates + 2 controls)
  Hypothesis: Chromatin remodeling loss impairs IL-2
    transcription
  Anchors: SMARCA4, ARID1A, SMARCB1
  Controls: KRT18, KRT19
  Revision: replace - mito-ribosome/OXPHOS hypothesis
    nearly falsified (1/35 hits); pivoting to distinct
    mechanism (chromatin/transcription) not yet tested
  Feedback: 8/30 candidate; 0/2 control
r3 (exploit; 30 candidates + 2 controls)
  Hypothesis: Spliceosome and 3'-end RNA processing loss
    reduces IL-2
  Anchors: SF3B1, CPSF1, RBM22
  Controls: UBB, RPL6
  Revision: refine - 22/44 hit rate high, hits span
    splicing (SF3B1/SF3A2/RBM22) and
    cleavage/polyadenylation (CPSF1/NUDT21); broadening
    exploit within RNA processing core
  Feedback: 11/30 candidate; 1/2 control
r4 (challenge; 30 candidates + 2 controls)
  Hypothesis: Ribosome biogenesis/rRNA processing loss
    lowers IL-2, not mitosis
  Anchors: BOP1, PDCD11, SYMPK
  Controls: RPL21, RPL22
  Revision: refine - hits concentrate in ribosome
    biogenesis/exosome/CPSF (SYMPK, PDCD11, BOP1, CPSF2,
    RPS25) while pure mitotic genes (TOP2A, BUB1B, MASTL,
    HAUS1) missed; rival mechanism to prior cell-cycle
    framing
  Feedback: 7/30 candidate; 1/2 control

Execution details:
  r1 control repair: ACTG1, KRT8
  r2 control repair: KRT18, KRT19
  r3 control repair: UBB, RPL6
  r4 control repair: RPL21, RPL22
  Anchor repairs: none
  Control fallbacks: none
  Candidate-pool expansions: 0
  Common-essential cap enabled: False
  Planner fallback: none

Feedback-trained response selector:
  Fit: all-score multiview ridge logistic; 512 measured
  genes; 161 internal positives; threshold 0.32403.
  Views: Achilles, Reactome, SIGNOR; frozen features,
  coefficients learned from measured assay scores.
  On-policy candidate audit: n=120; observed hit
  rate=0.317; Brier=0.211.
```

Round 5 (continued) Claude Sonnet 5 / IL2

```
Candidate pools before selection (recorded order):
  r1: 92 candidates
  CARD14, CLVS2, WISP3, LRFN2, CTGF, PRKCG, HDAC9, SLC35D3
  ANOS1, OLR1, CCDC137, TEPSIN, RAC3, SLC25A10, PDK2
  SLC35G3, LRRC37A3, NANOG, CAMSAP3, FCER2, PLVAP, P2RY11
  CLEC17A, KDM4B, ZNF700, RAB11B, CDKN2D, GMIP, NOL11
  RPAP1, G6PC, COPB2, NUP214, PPY, SPDYE1, UBTF, ADAM11
  POLA1, SDC1, OTOP1, SLC4A5, NUTF2, EIF3B, UBE2O, UNK
  ZNF492, EIF4B, H3F3C, CTCF, CCT5, TUBGCP4, EIF4G1
  SLC26A8, PSMD2, GALP, ZNF674, SPOUT1, PRPF4, MEPCE, INTS1
  SCPEP1, KNL1, NUF2, YARS, PPIL4, VPS13D, HSPB9, RRP12
  PSMC1, AARS, JAZF1, DDX11, SNRNP35, IMPDH1, NFS1, RBM25
  GSDMA, BANP, ANKRD36B, SMG5, DDX10, LSM6, SBNO1, DDX54
  ANKLE2, NOP14, POTEG, EIF2B1, SLC35G6, DMRTA2, IFITM3
  KARS
  r2: 92 candidates
  SMARCE1, SMARCC1, DPF2, SMNDC1, PRELID1, TSEN54, LRR1
  PRPF31, RSL24D1, RRM1, ZMAT2, DDX56, CCT4, PRPF38B, PPWD1
  ATP6V0B, FARSB, CDC123, WDR82, LUC7L3, CARS, INTS3, PAM16
  MED11, MIS18A, SAP18, MED30, XRCC6, RNPC3, COPB1, HARS
  SAP30BP, NCAPG, POLR3H, PRPF38A, GPN2, ZBTB11, CHAF1B
  POLR1C, INTS11, SS18L2, EIF3D, ERH, MMS22L, PMPCB, POLR1A
  VARS, SNAPC5, INTS9, CCT3, PFDN2, SNRPC, SNAPC1, DMAP1
  URB1, NSF, TAF1C, DBR1, NUS1, PDRG1, RPP21, MAK16, SF1
  THOC3, DUX4, HIST1H2BJ, DHX8, SFPQ, ZNHIT2, RABGGTA
  PSMG4, SDE2, ALG2, SLC7A6OS, PHB, DONSON, POLR3A, TTC27
  MED6, HAUS6, SF3A3, SNAPC2, GPN1, NEPRO, TANGO6, RPS27
  CSNK2B, NOC3L, GPN3, SMU1, DHDDS, NOB1
  r3: 92 candidates
  TRAPPC3, CDK11A, MARS, RACGAP1, CSE1L, UBL5, MED8, IARS
  WDR74, YKT6, NSA2, NARS, QARS, NMD3, SNRNP25, PRC1, DDX55
  SLU7, FOXD4, CWC22, UBA2, RPL26, PPAN, RABGGTB, TONSL
  TCP1, NOP58, PSMD4, GTF2A2, RSL1D1, CHAF1A, TNPO3, POLR3B
  WDR55, NAA50, POLR1B, INTS4, CCT2, WDR12, CRNKL1, MFAP1
  EIF2B2, SUPT16H, TUT1, USP5, RPN1, ATP2A2, NOP2, POLR3C
  DARS, EBNA1BP2, RBM19, MDN1, TRAPPC4, EPRS, NVL, NRF1
  TRAPPC1, NOP10, RBM8A, GUK1, NPLOC4, NLE1, TICRR, MCM4
  CKAP5, LSM8, SRSF1, SBDS, POLR3F, COPS3, TBCB, SYF2, LARS
  OSGEP, GRPEL1, WARS, TOMM40, GNL2, TBCD, DYNLRB1, ANAPC1
  DCTN5, SRBD1, PMPCA, EIF3E, RPS23, EXOSC2, KRI1, PLRG1
  ECD, WDR18
  r4: 92 candidates
  PAK1IP1, RPP40, TIMM10, EXOSC8, HEATR1, NOC4L, DHX33
  KRR1, CIAO1, WDR3, NOP16, NAT10, PFDN6, XRN2, NOP56
  EXOSC10, CRCP, GTPBP4, RPP14, NOL6, TMA16, TRIAP1, EEF1G
  SOD1, C1orf109, ZNF720, LONP1, MIS18BP1, RRS1, SHQ1
  RANGAP1, RPP30, CHCHD4, CSTF1, PAF1, NAA15, BRIX1, CEBPZ
  TAF5, UTP3, SPATA5, GRWD1, DDX42, RRN3, ERCC2, MTOR
  RBM17, ELP6, DCTN6, PLA2G10, PES1, EXOSC7, RPL34, DPAGT1
  URI1, TINF2, NHP2, DDX23, ELP5, ILF2, GGTLC2, IPO7, RNMT
  WDR77, COPZ1, PI4KA, SMC4, PPP1R10, TRAPPC11, SRRM1
  CPSF3, RBM28, UPF2, CMTR1, SART3, SNAPC3, RPS7, TBL3
  DDX18, RPSA, MED4, DPH3, TARDBP, PCBP2, RFT1, UTP18
  SRP14, GPS1, CCDC84, DDX47, TBCC, ILF3

All selected genes and measured outcomes:
  Columns: gene score flags; rounds 2-5 also p/u.
  F=feedback hit, A=GT All, N=GT N/E, E=CEG,
```

Round 5 (continued) Claude Sonnet 5 / IL2

```
  C=control; p=predicted response, u=uncertainty.
r1: 32 tested
  01 ACTG1      +0.08889 ----C p=-- u=--
  02 KRT8       -0.10457 ----C p=-- u=--
  03 PSMD2      +0.76766 FAN-- p=0.648 u=0.053
  04 PSMC1      +0.13890 ----- p=0.647 u=0.053
  05 EIF4G1     +0.45271 FAN-- p=0.405 u=0.027
  06 NUP214     +0.16475 ---E- p=0.376 u=0.025
  07 EIF3B      +0.47816 FA-E- p=0.356 u=0.024
  08 EIF4B      +0.32525 F---- p=0.326 u=0.023
  09 POLA1      +0.24326 ----- p=0.320 u=0.023
  10 NOP14      +0.33250 F---- p=0.281 u=0.024
  11 NUF2       +0.18963 ---E- p=0.280 u=0.024
  12 SMG5       +0.40496 FAN-- p=0.276 u=0.024
  13 NOL11      +0.11364 ---E- p=0.274 u=0.023
  14 PPIL4      +0.17797 ----- p=0.270 u=0.024
  15 RBM25      +0.46968 FAN-- p=0.269 u=0.024
  16 ANKLE2     +0.20013 ----- p=0.268 u=0.025
  17 SDC1       +0.00509 ----- p=0.267 u=0.023
  18 CTCF       -0.05940 ----- p=0.266 u=0.025
  19 SNRNP35    -0.16279 ---E- p=0.266 u=0.024
  20 BANP       -0.14328 ----- p=0.266 u=0.024
  21 TUBGCP4    +0.27092 ----- p=0.265 u=0.024
  22 PRPF4      +0.64776 FA-E- p=0.263 u=0.024
  23 LSM6       +0.37079 FAN-- p=0.261 u=0.024
  24 COPB2      +0.23605 ---E- p=0.258 u=0.025
  25 YARS       -0.06168 ---E- p=0.257 u=0.025
  26 KNL1       +0.08194 ----- p=0.257 u=0.035
  27 KARS       +0.11366 ---E- p=0.257 u=0.025
  28 AARS       +0.36999 FA-E- p=0.257 u=0.025
  29 IFITM3     +0.00286 ----- p=0.257 u=0.025
  30 INTS1      +0.01179 ---E- p=0.256 u=0.025
  31 DDX54      +0.38541 FAN-- p=0.256 u=0.025
  32 SPOUT1     +0.37628 FA-E- p=0.256 u=0.025
r2: 32 tested
  01 KRT18      -0.00351 ----C p=-- u=--
  02 KRT19      -0.10000 ----C p=-- u=--
  03 RPS27      +0.12923 ----- p=0.720 u=0.057
  04 EIF3D      +1.57460 FA-E- p=0.356 u=0.024
  05 HAUS6      +0.01822 ----- p=0.303 u=0.046
  06 CSNK2B     -0.12573 ----- p=0.303 u=0.024
  07 SF3A3      +0.33665 F--E- p=0.284 u=0.024
  08 SMNDC1     -0.08610 ----- p=0.283 u=0.024
  09 RPP21      -0.16410 ---E- p=0.279 u=0.024
  10 NOB1       +0.37803 FAN-- p=0.278 u=0.024
  11 SAP18      +0.14005 ---E- p=0.271 u=0.024
  12 SNRPC      +1.24800 FAN-- p=0.269 u=0.024
  13 RNPC3      -0.01778 ----- p=0.267 u=0.024
  14 XRCC6      -0.22206 ---E- p=0.267 u=0.025
  15 SF1        +0.06864 ---E- p=0.266 u=0.025
  16 PRPF31     +0.49100 FA-E- p=0.266 u=0.025
  17 SMU1       +0.45338 FA-E- p=0.266 u=0.025
  18 ZMAT2      +0.45278 FAN-- p=0.266 u=0.025
  19 LUC7L3     +0.12211 ---E- p=0.266 u=0.025
  20 PRPF38A    +0.14771 ---E- p=0.265 u=0.025
  21 DHX8       +0.06290 ---E- p=0.265 u=0.025
  22 PPWD1      +0.29332 ---E- p=0.265 u=0.025
  23 MED30      +0.25780 ---E- p=0.265 u=0.024
```

Round 5 (continued) Claude Sonnet 5 / IL2

```
  24 MED6       -0.16400 ----- p=0.264 u=0.024
  25 ATP6V0B    +0.06105 ---E- p=0.263 u=0.025
  26 THOC3      +0.53639 FA-E- p=0.263 u=0.024
  27 MED11      +0.04522 ---E- p=0.263 u=0.025
  28 NCAPG      +0.17098 ----- p=0.262 u=0.025
  29 FARSB      +0.13152 ---E- p=0.260 u=0.025
  30 LRR1       -0.03053 ---E- p=0.260 u=0.025
  31 MIS18A     -0.00752 ----- p=0.260 u=0.025
  32 RABGGTA    +0.24749 ----- p=0.259 u=0.025
r3: 32 tested
  01 UBB        +0.11361 ----C p=-- u=--
  02 RPL6       +0.61968 FA-EC p=-- u=--
  03 RPS23      +0.43396 FA-E- p=0.702 u=0.056
  04 RPL26      +0.99306 FAN-- p=0.622 u=0.050
  05 ANAPC1     +0.23423 ----- p=0.391 u=0.029
  06 NPLOC4     +0.17937 ---E- p=0.360 u=0.077
  07 MCM4       +0.01386 ---E- p=0.358 u=0.064
  08 EIF3E      +0.21122 ----- p=0.329 u=0.046
  09 RBM8A      +0.32923 F--E- p=0.311 u=0.024
  10 CKAP5      +0.37029 FA-E- p=0.292 u=0.024
  11 PLRG1      +0.60280 FA-E- p=0.288 u=0.029
  12 NOP10      -0.04207 ----- p=0.286 u=0.024
  13 NOP58      +0.22033 ----- p=0.286 u=0.024
  14 SUPT16H    -0.11126 ----- p=0.286 u=0.024
  15 SRSF1      +0.63768 FA-E- p=0.285 u=0.024
  16 RPN1       -0.01813 ----- p=0.283 u=0.024
  17 EXOSC2     +0.04772 ---E- p=0.282 u=0.024
  18 EBNA1BP2   +0.38750 FAN-- p=0.279 u=0.024
  19 WDR12      +0.42737 FA-E- p=0.279 u=0.024
  20 WDR18      +0.06885 ----- p=0.278 u=0.024
  21 NOP2       +0.54848 FA-E- p=0.273 u=0.024
  22 CWC22      +0.22022 ----- p=0.273 u=0.024
  23 CRNKL1     -0.03229 ---E- p=0.273 u=0.024
  24 SLU7       +0.13025 ---E- p=0.269 u=0.024
  25 PRC1       +0.37505 FAN-- p=0.268 u=0.027
  26 SNRNP25    +0.01035 ---E- p=0.267 u=0.024
  27 UBL5       -0.00166 ---E- p=0.265 u=0.025
  28 CDK11A     -0.06064 ----- p=0.264 u=0.025
  29 COPS3      +0.15958 ---E- p=0.264 u=0.024
  30 MFAP1      +0.42193 FAN-- p=0.264 u=0.024
  31 LSM8       +0.16211 ---E- p=0.264 u=0.024
  32 TUT1       +0.20017 ---E- p=0.264 u=0.024
r4: 32 tested
  01 RPL21      +0.33179 F---C p=-- u=--
  02 RPL22      -0.23380 ----C p=-- u=--
  03 RPS7       +1.02160 FA-E- p=0.697 u=0.056
  04 RPSA       +0.28251 ----- p=0.695 u=0.056
  05 RPL34      +0.67099 FAN-- p=0.621 u=0.050
  06 UPF2       +0.04038 ---E- p=0.419 u=0.166
  07 UTP3       +0.28674 ----- p=0.370 u=0.122
  08 RANGAP1    -0.01845 ---E- p=0.301 u=0.023
  09 ERCC2      -0.09847 ---E- p=0.296 u=0.023
  10 CPSF3      +0.04118 ---E- p=0.293 u=0.024
  11 DDX42      +0.14429 ----- p=0.288 u=0.024
  12 PCBP2      +0.66526 FAN-- p=0.285 u=0.024
  13 TINF2      +0.07063 ----- p=0.285 u=0.024
  14 CSTF1      +0.17042 ----- p=0.284 u=0.024
  15 MTOR       -0.00960 ----- p=0.284 u=0.024
```

Round 5 (continued) Claude Sonnet 5 / IL2

```
  16 RBM17      +0.00400 ---E- p=0.283 u=0.024
  17 NHP2       +0.01256 ---E- p=0.283 u=0.023
  18 NOP56      -0.08073 ---E- p=0.282 u=0.024
  19 HEATR1     +0.06112 ---E- p=0.282 u=0.024
  20 WDR3       +0.33659 F--E- p=0.282 u=0.024
  21 DDX47      +0.56027 FA-E- p=0.282 u=0.024
  22 EXOSC8     +0.07498 ---E- p=0.281 u=0.024
  23 KRR1       +0.11311 ---E- p=0.281 u=0.024
  24 NOC4L      +0.10582 ---E- p=0.281 u=0.024
  25 EXOSC7     +0.21759 ---E- p=0.281 u=0.024
  26 NOL6       +0.52455 FA-E- p=0.280 u=0.024
  27 PAF1       +0.30253 ----- p=0.280 u=0.024
  28 TBL3       +0.12564 ---E- p=0.279 u=0.024
  29 EXOSC10    +0.07530 ----- p=0.279 u=0.024
  30 UTP18      +0.01673 ----- p=0.279 u=0.023
  31 TARDBP     -0.45886 FAN-- p=0.279 u=0.039
  32 RPP40      +0.17914 ----- p=0.279 u=0.024

Round outcome:
  128 new tests; internal feedback hits: 40
  New GT: All 34; N/E 16
  Cumulative GT: All 184/654; N/E 73/481
  Cumulative selected CEG: 326/640
```

Final state and interpretation Claude Sonnet 5 / IL2

```
FINAL STATE AND INTERPRETATION

Final external result: All 184/654 (28.1% recall);
N/E 73/481 (15.2% recall); selected CEG 326/640.
The high selected-CEG count limits claims of
mechanism-specific discovery from All hits alone.

Final portfolio (candidate hits/trials; controls):
  r1: 12/30; controls 0/2
  r2: 8/30; controls 0/2
  r3: 11/30; controls 1/2
  r4: 7/30; controls 1/2

Final discovery map:
  established: r3: pre-mRNA splicing/3'-end processing loss
  reduces IL-2, r1: secretory/vesicle trafficking loss
  lowers IL-2
  falsified: r2: mitochondrial OXPHOS dysfunction impairs
  IL-2 (mito-ribosome subset barely hits)
  uncertain: r4: ribosome biogenesis / RNA processing
  complex loss alters IL-2 (not cell-cycle per se)
  unexplained_hits: MRPL39 (r2, isolated hit amid
  mito-ribosome non-hits)

Evidence boundary: one trace shows agent behavior and
observations, not a causal effect of hypothesis update
or of any one evidence view. Those claims require the
```