When to Trust Your AI Co-Scientist? Learning Prior Reliability in Closed-Loop Bayesian Optimization

Anonymous authors
Paper under double-blind review

Abstract

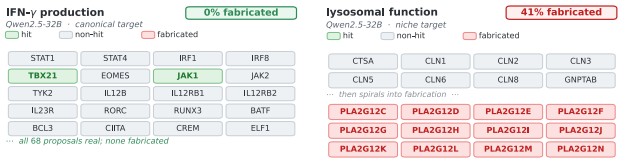
AI "co-scientists" promise to accelerate discovery by proposing what to test next from knowledge distilled across millions of publications. Their reliability, however, depends on literature coverage: on under-studied targets they fabricate plausible candidates and fail silently, so a practitioner cannot tell a useful prior from a harmful one before committing costly experiments. Larger models mitigate but cannot eliminate failures where the literature is thin. We propose HARIBO, which learns a prior's reliability from the experiments themselves. HARIBO adds the prior as a single feature of the Bayesian optimization (BO) surrogate. Its coefficient, fit to the sparse observations already being collected, provides an interpretable measure of trust that exploits, ignores, or inverts the prior as the data dictate. Across gene perturbation, protein variant effects, and antibody developability, and across language-model agents from 7B to frontier scale and protein foundation models, HARIBO improves on BO wherever the prior carries signal and matches it where the prior is noise. Under a systematically wrong prior, every scheme that fixes the prior's weight in advance, by a constant, schedule, or model confidence, falls below BO, while HARIBO stays above it. The model is consulted before the first experiment and never again, which is enough: re-querying it every round instead, at 20–32$\times$ the requests and 124–234$\times$ the tokens, helps the resulting closed-loop agent on one of three domains and leaves it below knowledge-free BO on the other two. By learning trust from data rather than from model confidence, HARIBO lets AI co-scientists enter discovery loops safely.

1 Introduction

The new wave of AI "co-scientists" promises to accelerate scientific discovery by proposing biological hypotheses, experimental designs, and candidate interventions (Wang et al., 2023; Gottweis et al., 2026; Swanson et al., 2025; Ghareeb et al., 2026; Bran et al., 2023; Boiko et al., 2023). Building on a long line of work on automating scientific hypothesis generation and experimentation (King et al., 2004), this promise is especially compelling in *closed-loop discovery*, such as pooled CRISPR perturbation screens (Dixit et al., 2016; Wang & Doudna, 2023; Przybyla & Gilbert, 2022), protein engineering, and antibody optimization, where researchers iteratively propose candidates, perform costly experiments, and refine the next round of designs. Because misidentifying a target is a leading cause of downstream failure (Nelson et al., 2015), prioritizing the right candidates early is critical. By drawing on knowledge distilled from millions of publications, AI co-scientists can potentially guide experimental search far beyond what the limited measurements of a single campaign allow (Hu et al., 2025; Hou & Ji, 2024).

However, this knowledge is *coverage-conditioned*: it is only as good as the literature on the target. As shown in Fig. 1, the same AI co-scientist that proposes only real regulators for a well-studied target fabricates 26 of 64 candidates (41%) for an under-studied phenotype, and nothing in the generated batch identifies which ones. Nor does the model catch such errors when asked to evaluate candidates. We mix real genes with plausible but non-existent symbols and ask it to rate each gene's knockout effect; on the niche screen, it scores 28% of the fabricated genes at $5/10$ or higher (Table 1). Asking the model again therefore does not tell a practitioner which proposals to trust before costly laboratory validation. Blindly following the prior can waste a limited experimental budget, while discarding it forfeits potentially useful knowledge.

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.



(a) IFNG (well-studied) (b) Scharenberg22 (niche)

Figure 1: **AI Co-Scientist hallucinates more as coverage thins.** (a) On the well-studied IFN-γ screen (Schmidt et al., 2022), every proposed gene is real; (b) on the niche lysosomal choline recycling screen (Scharenberg et al., 2023), 26 of 64 are fabricated (41%). Nothing in the output marks which ones, so the failure is silent.

Table 1: **AI Co-Scientist assigns substantial scores to fabricated genes on a niche screen.** We mix real genes with plausible but non-existent symbols (e.g. `SLC25A80`) and ask Qwen2.5-32B to rate each gene's knockout effect (0–10). Entries show mean scores for real hits and fabricated genes, and the fraction of fabricated genes scored at least 5.

phenotype (coverage)	real hit	**fabricated**	fakes scored ≥ 5
IFNG (well-studied)	3.23	1.70	5%
Carnevale22 (moderate)	2.55	1.30	0%
Scharenberg22 (niche)	4.35	**3.00**	**28%**

Two observations sharpen this failure mode. First, fabrication is driven by how many candidates the model is asked for. A closed-loop screen sets its batch size from the experimental budget, and a co-scientist asked to fill that batch fabricates 41% of it on the screen of Fig. 1 and up to 66% across screens at a 128-gene demand (Appendix C.1). Asked instead only for candidates it is confident about, it fabricates almost none, but returns only a dozen genes out of $\sim$15k, too few to guide subsequent rounds. Second, model scale alone does not resolve the problem: a frontier co-scientist stops fabricating on these screens, yet its proposals for a thinly studied phenotype remain barely enriched over chance, because the limit there is the literature rather than the reader (Appendix C.1).

This raises a fundamental question for AI-assisted discovery: *how much should an AI co-scientist be trusted?* Existing closed-loop optimization methods provide no satisfactory answer. Bayesian optimization (BO) and active learning rely solely on experimentally observed data, which makes them robust but unable to exploit external scientific knowledge (Shahriari et al., 2015; Frazier, 2018; Mehrjou et al., 2021; Notin et al., 2023). Conversely, AI co-scientists inject prior knowledge but offer no reliable way to tell when it should be trusted. Confidence estimates are often poorly calibrated (Ji et al., 2023), and Table 1 shows that a model's own scores do not reliably separate real candidates from fabricated ones. This rules out the obvious remedy of weighting the prior by the model's self-reported confidence (Yuan et al., 2026), which we find shifts confidently the wrong way under a bad prior (Sec. 4.2). Deciding how far to defer to external evidence over observed data is an open problem (Wu et al., 2024). The question is therefore not whether to use AI-generated priors, but how to incorporate them while determining their reliability from the experimental evidence.

To address this, we propose **HARIBO** (**H**arnessing **A**I Co-scientist priors **R**obustly **I**nto **B**ayesian **O**ptimization). HARIBO takes a minimalistic approach: rather than redesigning the optimizer, it adds the AI prior as a single explicit feature of the BO surrogate. The coefficient on that feature, refit in closed form from the observations the loop is already collecting, acts as an interpretable trust. It grows when the prior is informative, shrinks toward zero when it is not, and turns negative under a systematically misleading prior. Because the prior is obtained zero-shot before any measurement and never refreshed, this adds almost no cost to the loop. Across gene perturbation, protein variant effects, and antibody developability, with both LLM-agent and foundation-model priors, HARIBO improves on standard BO wherever the prior carries signal and matches it where the prior is noise.

To summarize, we make the following contributions:

- We identify a failure mode of AI co-scientists: their reliability depends on literature coverage, and neither their proposals nor their own scores flag fabricated candidates (Fig. 1, Table 1). Large fixed batch demands amplify this failure when coverage is thin, and a frontier model, while no longer fabricating, remains only weakly enriched over chance on poorly studied targets.
- We propose HARIBO, a simple, interpretable method that learns the reliability of an AI-generated prior from experimental observations. Prop. 1 characterizes its risk under informative and adversarial priors in the population linear setting.
- Across LLM and foundation-model priors, biological domains, and model scales, we show that HARIBO improves on standard BO when the prior carries signal, tracks it when the prior is noise, and withstands a systematically inverted prior, where every scheme that fixes the prior's weight in advance falls below BO, while consulting the model only before the first experiment – 20–32$\times$ fewer requests and 124–234$\times$ fewer tokens than re-querying it every round.

2 RELATED WORK

AI co-scientists, LLM agents, and foundation-model priors. Automating scientific hypothesis generation has a long history (King et al., 2004), and recent "AI co-scientists" and tool-augmented LLM agents (Wang et al., 2023; Gottweis et al., 2026; Swanson et al., 2025; Ghareeb et al., 2026; Bran et al., 2023; Boiko et al., 2023) now propose nanobodies, gene targets, molecules, and full experimental plans, and design gene-editing protocols (Qu et al., 2026) or query biomedical databases (Jin et al., 2024). The knowledge that makes them attractive as priors has been probed directly: LLMs recall gene-set function (Hu et al., 2025), annotate cell types (Hou & Ji, 2024), and are benchmarked for biology research (Laurent et al., 2024), while self-supervised foundation models provide sequence-level priors from protein language models (Meier et al., 2021; Lin et al., 2023) to single-cell models such as scGPT (Cui et al., 2024). Yet all these priors degrade on under-represented inputs and hallucinate confidently (Ji et al., 2023), and how far to trust such external evidence over observed data is contested (Wu et al., 2024). Such systems are typically evaluated on bespoke case studies rather than against BO baselines on shared closed-loop benchmarks, and re-prompt the model every round; BIODISCOVERYAGENT (Roohani et al., 2025) formalizes the gene-perturbation loop this way. We instead treat any such agent's output as one *untrusted prior score* per candidate and benchmark against plain BO, consulting the model in a single round before the first experiment; on two of the three domains that is already better than re-querying it every round, and on the third we report what the extra rounds buy (Sec. 4.4).

Closed-loop experiment design. Closed-loop screen design has traditionally relied on active learning and Bayesian optimization over learned surrogates: GeneDisco benchmarks acquisition functions for perturbation experiments (Mehrjou et al., 2021), DiscoBAX targets diverse intervention sets (Lyle et al., 2023), related work folds multi-modal priors into the acquisition (Huang et al., 2024) or predicts perturbation outcomes (Roohani et al., 2023), and ProteinGym is the analogous protein-fitness benchmark (Notin et al., 2023). These explorers are robust but knowledge-free. Ours is a label-fitted ridge surrogate with greedy batch acquisition, standard non-GP pool-based surrogate BO for large discrete design spaces (SMAC (Hutter et al., 2011), DNGO (Snoek et al., 2015), GeneDisco (Mehrjou et al., 2021)); a GP-UCB explorer (Srinivas et al., 2009) gives the same conclusions.

Priors and LLMs in Bayesian optimization. To inject a prior, πBO and BOPrO (Hvarfner et al., 2022; Souza et al., 2021) assume it is *benevolent*; we target an untrusted, possibly adversarial prior and let the data set its weight. The closest work, LLM-guided BO (LGBO) (Yuan et al., 2026), shifts the surrogate mean by the LLM's *self-reported* confidence and re-queries every round; because self-report is uncalibrated (Table 1), it shifts confidently the wrong way on a bad prior, whereas we learn trust from labels and thus *invert* rather than amplify an adversarial prior (Sec. 4.2). GOLLuM (Ranković et al., 2026) instead uses the LLM as a *representation*, fine-tuning its embeddings through the GP marginal likelihood, which needs embedding- or weight-level access and per-campaign gradient updates. HARIBO takes a fixed scalar score per candidate from any chat-only model and learns its signed contribution from labels, so the two are complementary: a GOLLuM-adapted embedding could serve as our ϕ.

3 METHODOLOGY

Notation. A finite pool $\mathcal{X}$ is screened over R rounds; each round selects a batch of B candidates, and an assay returns their *continuous* measured property y (a fitness or log-fold-change). Each candidate x carries an AI co-scientist *prior* $\pi(x)$, computed once and zero-shot, and a domain-specific candidate feature $\phi(x) \in \mathbb{R}^d$ (STRING→PCA for genes; one-hot→PCA for proteins and antibodies). Hits are the benchmark's *topmover* set (0.8–5.0% of the panel by screen) and we report their cumulative recall, dropping CEGv2 core-essential genes from the hit set and its denominator: they are up to $7\times$ over-represented there and register a viability defect rather than a phenotype-specific discovery. The benchmark keeps them; Sec. 4.3 reports both scorings and what the filter is worth. Hits are an *evaluation* target and never enter scoring.

Computing the prior $\pi(x)$. Each prior is one frozen scalar per candidate. For genes a single *round* of zero-shot elicitation returns a proposal batch, and a proposed gene gets $\pi{=}1$ against $\pi{=}0$

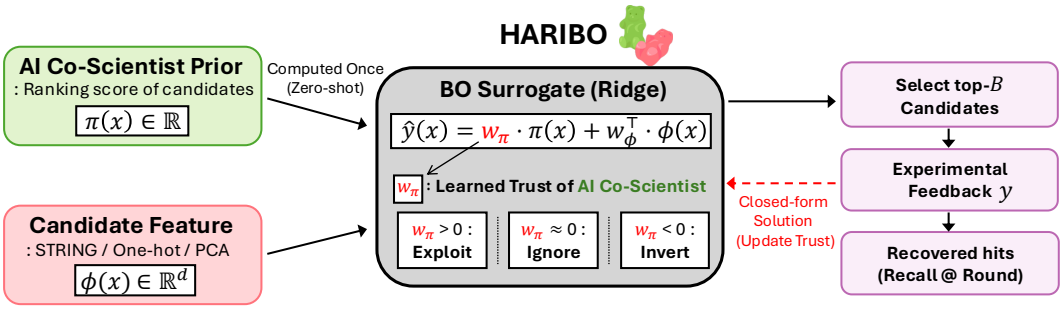


Figure 2: Overview of HARIBO. An AI co-scientist produces a prior $\pi(x)$ in a single zero-shot round before any measurement, while each candidate is represented by domain-specific features $\phi(x)$. HARIBO adds the prior as one extra feature of the BO surrogate and learns its coefficient w_π, the trust placed in the co-scientist. As feedback y arrives each round, the surrogate is refit and w_π is re-estimated from the observations, so the learned trust exploits informative priors ($w_\pi > 0$), ignores uninformative ones ($w_\pi \approx 0$), and inverts adversarial ones ($w_\pi < 0$).

elsewhere: a *nomination indicator* over the ∼15k pool.[1] We use an indicator rather than a graded rank. Position is informative in itself (the correlation between list position and hit status is -0.16 on Sonnet 5), but encoding it does not pay: a ∼120-name list is exhausted by the first B=128 batch, so what survives into later rounds is which genes were named, not their order, and ranking by the model's genuine emission order scores below the indicator on both models (Appendix A.1). For proteins π is the ESM2-650M zero-shot fitness of each single mutant, and for antibodies its mean per-residue log-likelihood over both chains; in Sec. 4.1 the agent instead rates each candidate 0–10 in one prompt listing the pool. Every prior is z-scored per assay so its scale cannot dominate the d features. The three prompts are reproduced verbatim in Appendix B.

HARIBO: the prior as a feature. The natural way to use a prior is a *fixed blend* $s_\lambda = (1-\lambda)\, z(\widehat{f}) + \lambda\, z(\pi)$ with $\lambda = \frac{1}{2}$, where $\widehat{f}$ is the no-prior prediction and $z(\cdot)$ standardizes; it helps under a good prior and degrades under a wrong one, because the weight on π is fixed before any measurement. HARIBO instead makes the prior an additional *feature*, so its weight is learned from the observations (Fig. 2). Writing $u(x) = [1; \pi(x); \phi(x)] \in \mathbb{R}^{d+2}$ and stacking the observed candidates into (U, y) at round t, we fit a *partially* penalized ridge surrogate: the d feature coordinates carry a penalty α chosen each round by leave-one-out cross-validation on the observed labels (over a log-spaced grid on $[10^0, 10^5]$), so the surrogate regularizes itself as labels accumulate, while the prior coordinate carries *none* (Appendix A.2).[2]

$$\theta_t = \begin{bmatrix} b \\ w_\pi \\ w_\phi \end{bmatrix} = \left(U^\top U + P_\alpha\right)^{-1} U^\top y, \qquad P_\alpha = \mathrm{diag}(0, 0, \alpha, \ldots, \alpha). \tag{1}$$

We predict $\widehat{y}(x) = b + w_\pi\, \pi(x) + w_\phi^\top \phi(x)$ and acquire the top-B unmeasured candidates by $\widehat{y}$. *Pure BO*, our no-prior baseline throughout, is the same surrogate and acquisition on ϕ alone ($w_\pi \equiv 0$); HARIBO adds only the prior coordinate. Trust is thus one scalar among d+2 parameters, re-estimated in closed form each round, while the prior model stays frozen and receives no gradients.

Why one coefficient suffices. Let $L[\cdot \mid \phi]$ denote the population best linear predictor (linear projection onto $\{1, \phi\}$), and define the residual prior, the residual outcome, and the population trust coefficient

$$\tilde{\pi} = \pi - L[\pi \mid \phi], \qquad \tilde{y} = y - L[y \mid \phi], \qquad \gamma^\star = \frac{\mathrm{Cov}(\tilde{\pi}, \tilde{y})}{\mathrm{Var}(\tilde{\pi})}. \tag{2}$$

The following population least-squares result characterizes the *unique* contribution of the prior beyond the candidate features and the cost of a fixed weight.

[1] A round is not a request: the harness re-prompts until the batch is full, so round one is 1.7 requests on Claude Sonnet 5 and 5.0 on Qwen2.5-32B (Appendix C.1).

[2] Neither zero is an artefact of the solver; w_π is obtained in closed form rather than from `Ridge`, which applies one α to every column (Appendix A.2).

Proposition 1 (HARIBO absorbs an untrusted prior)**.** *Assume finite second moments and* $\mathrm{Var}(\tilde{\pi}) > 0$ *(the prior is not a linear function of the features), and consider the population least-squares limit* ($\alpha \to 0$)*.* ***(a)*** *The* HARIBO *coefficient on* π *equals* $\gamma^\star$*: it is* 0 *when the prior is uninformative beyond the features (*ignore*), positive when it is informative (*exploit*), and negative when it is adversarial (*invert*).* ***(b)*** *The* HARIBO *score* $s^\star = L[y \mid \phi, \pi]$ *is the best linear predictor of* y *from* $[\phi, \pi]$*, and no blend* $\bar{s}_\lambda = h(\phi) + \lambda\pi$ *with linear* h *attains it unless* $\lambda = \gamma^\star$*. Under joint Gaussianity,* $s^\star = \mathbb{E}[y \mid \phi, \pi]$*, so ranking by* $s^\star$ *selects the batch with the largest expected total outcome.* ***(c)*** *Even with its feature part chosen optimally, a blend with weight* λ *has excess risk* $(\lambda - \gamma^\star)^2 \mathrm{Var}(\tilde{\pi})$ *over* $s^\star$*; the unadjusted blend, whose feature part is the no-prior projection* $h = L[y \mid \phi]$*, incurs an additional* $\lambda^2\, \mathbb{E}\big[L[\pi \mid \phi]^2\big]$*. For an adversarial prior (*$\gamma^\star < 0$*) and any positive* λ*, the first term grows with how wrong the prior is: the blend is most damaged exactly where* HARIBO *most helps.*

Proof. (a) By Frisch–Waugh–Lovell, the coefficient of π in the projection of y onto $\mathrm{span}\{1, \phi, \pi\}$ equals the simple-regression coefficient of $\tilde{y}$ on $\tilde{\pi}$, namely $\gamma^\star$; the sign statements follow from $\mathrm{sign}\,\gamma^\star = \mathrm{sign}\,\mathrm{Cov}(\tilde{\pi}, \tilde{y})$. (b) Least squares returns the L^2 projection of y onto $\{c_0 + a^\top\phi + c\,\pi\}$, every $\bar{s}_\lambda$ with linear h lies in that class, and the projection's π-coordinate is $\gamma^\star$ by (a). Under joint Gaussianity the projection equals $\mathbb{E}[y \mid \phi, \pi]$, so by linearity its top-B candidates maximize the expected batch total. (c) By (a), $s^\star = L[y \mid \phi] + \gamma^\star\tilde{\pi}$, so for $s = g(\phi) + \lambda\pi$ the error splits as $\big(g - L[y \mid \phi] + \lambda L[\pi \mid \phi]\big) + (\lambda - \gamma^\star)\tilde{\pi}$ with the terms orthogonal ($\tilde{\pi}$ is uncorrelated with every linear function of ϕ). Minimizing over g leaves $(\lambda - \gamma^\star)^2\,\mathrm{Var}(\tilde{\pi})$; fixing $g = L[y \mid \phi]$ instead adds $\lambda^2\,\mathbb{E}\big[L[\pi \mid \phi]^2\big]$. □

Two caveats: λ here multiplies the raw prior under squared error, whereas the experiments standardize and score by recall, and under adaptive acquisition the fitted coefficient need not equal $\gamma^\star$ when the model is misspecified.

So w_π asks not whether π correlates with y, but whether the part of π the features leave unexplained predicts the part of y they leave unexplained – one scalar, with no threshold or blend weight to tune. Letting it vary across candidate space buys nothing measurable (Appendix A.3), and k prior columns give k coefficients, so sources can be attached without deciding in advance which is good.

Remark (finite α and the unpenalized prior). In practice the features carry a cross-validated penalty and the prior coordinate carries none, so w_π is the exact finite-α coefficient rather than the least-squares one; penalizing the prior as well would shrink a good prior exactly as hard as a wrong one, and the features need a penalty orders of magnitude above the prior's own scale. Proposition 1 is a least-squares limit, but its *sign* reading survives at any finite α (Appendix A.2).

Round one as a design problem. Round one has no labels, so w_π is not merely unestimated but *undefined*, and the batch acquired then is what identifies it – a design problem, not a hedging one. Under homoscedastic uncorrelated errors of variance σ^2, Frisch–Waugh–Lovell applied to ordinary least squares within the batch S gives

$$\mathrm{Var}(\widehat{w}_\pi) = \sigma^2 \Big/ \sum_{x \in S} \tilde{\pi}_S(x)^2, \qquad \tilde{\pi}_S = \pi - \widehat{L}_S[\pi \mid \phi], \tag{3}$$

with $\widehat{L}_S$ the least-squares projection onto $\{1, \phi\}$ within S.[3] Because it uses only π and ϕ, it can be evaluated before any label is seen. For a binary prior orthogonal to the features within the batch, $\tilde{\pi}_S$ is π centred within S, so $\sum_{x \in S} \tilde{\pi}_S(x)^2 = |S|\rho(1-\rho)$ for a nominated fraction ρ. This quantity is maximized at $\rho = 1/2$, minimizing $\mathrm{Var}(\hat{w}_\pi)$: HARIBO splits round one evenly. Otherwise the even split is a heuristic motivated by Eq. 3, and both extremes fail in mirror image: an exploratory batch of 128 from a 15k pool holds about one nominated gene, an all-prior batch no un-nominated contrast (Sec. 4.3 measures both). This concerns *identification* alone, not whether the prior is any good, which round one cannot know. For continuous priors round one takes the ρB candidates with largest π and fills the rest uniformly (Appendix A.4).

[3] Two qualifications. The criterion is the OLS variance, not the exact variance of the finite-α ridge estimate, and its derivation needs $|S| > d + 2$ – true on the gene screens (128 against 52) but not on the small antibody assays ($|S| \approx 6$, $d = 50$), where an OLS projection inside the batch would saturate. Identification survives there anyway, because the code projects π on ϕ with the feature fit's penalty rather than by OLS.

Beyond the ridge surrogate. The construction is not specific to the ridge: placing π as an input dimension of a GP and acquiring by GP-UCB gives the same robustness conclusions, with trust read off that dimension's lengthscale rather than a signed scalar (Appendix C.2).

4 EXPERIMENTS

Setup. We evaluate on three closed-loop discovery domains (propose a batch, measure it, refit; Table 2). The six gene screens span well-studied to niche phenotypes: IFNG and IL2 (T-cell cytokine production; (Schmidt et al., 2022)), Carnevale22 (resistance to tumor-microenvironment suppression; (Carnevale et al., 2022)), Sanchez21 and Sanchez21_down (neuronal tau levels; (Sanchez et al., 2021)), and Steinhart (CRISPRa in GD2 CAR-T cells); Scharenberg22 (Scharenberg et al., 2023) serves only the example of Sec. 1. Protein and antibody assays are from ProteinGym (Notin et al., 2023) and FLAb (Chungyoun et al., 2024), where a hit is a candidate in the top 5% of measured values. Features ϕ are PCA-50 embeddings, and we report the cumulative hit recall of §3 over 20 seeds, treating datasets as the replication unit and comparing methods with paired Wilcoxon signed-rank tests. Besides pure BO, we compare against four ways of folding a prior into BO, all run in the same loop on the same features, prior and seeds and differing only in how the prior's weight is set: a fixed blend ($\lambda = \frac{1}{2}$); πBO (Hvarfner et al., 2022) and BOPrO (Souza et al., 2021), which decay it with the round index and with the number of observations; and LGBO (Yuan et al., 2026), which scales it by the model's self-reported confidence.

Table 2: **Three closed-loop discovery domains.** Each uses R=10 rounds.

Domain	Genes	Proteins	Antibodies
assays	6 CRISPR screens	184 ProteinGym DMS	43 FLAb developability
size (each)	∼15k genes	up to 4k single mutants	∼100–4k antibodies
candidate	a gene	point mutant (e.g. A123P)	antibody (heavy+light)
outcome y	perturbation LFC	DMS fitness	developability property
features ϕ	STRING	one-hot sequence	one-hot heavy+light
prior π	agent nomination indicator	ESM2-650M zero-shot	ESM2 naturalness
batch / round	128	n÷20	n÷20

4.1 MAIN RESULT: LEARNED TRUST (w_π) PAYS OFF ACROSS DOMAINS

HARIBO should turn an untrusted prior into a safe one: exploit it where it helps and ignore it where it does not. We first feed the *same* LLM co-scientist prior (Qwen2.5-32B) to every method in all three domains (Table 3, Fig. 3). The agent is competent at gene *selection* but weak at fine granularity (protein residues, antibody CDRs), where its prior barely correlates with the outcome (corr $+0.08/\approx 0$).[4] HARIBO beats plain BO wherever the prior carries signal (gene 0.133 vs 0.111, protein 0.602 vs 0.560) and beats the prior arm in all three domains. On antibodies, where the prior is uninformative, the prior arm and the fixed blend fall *below* plain BO (0.471 and 0.607 against 0.625) while HARIBO matches it (0.626, p=0.48): the data shrink its weight to zero rather than hold it at one half. Recovering the no-prior baseline when the prior is noise is the safety property the method is built for, and it is what no method with a pre-fixed weight can do. The same holds with a frontier agent: with Claude Sonnet 5 as the co-scientist, whose nominations carry more signal, HARIBO rises to 0.140 on gene, above both the prior arm (0.116) and pure BO.

Which prior to trust. Can a domain-specific foundation model fill the agent's gap at fine granularity? Replacing the agent with ESM2 lifts protein recall substantially (Appendix C.3; HARIBO 0.725 vs. BO 0.560, Table 4), whereas scGPT carries little zero-shot signal on genes and ESM2 little on antibody CDRs, where HARIBO defaults to BO (Fig. 6). Agents and foundation models are thus reliable at complementary granularities, and HARIBO does not need to know in advance which one it has been given. The learned coefficient also reads as a diagnosis of the prior. Fitting both priors as separate columns on a matched protein set (n=212), the per-assay weights track each prior's actual

[4] Giving the agent its published retrieval and database tools (literature search, pathway enrichment, similar-gene search) does not change this: in a separate tool ablation on the same six screens, the one-round prior reaches 0.130 with tools against 0.125 without (p=0.56).

Table 3: **The AI co-scientist prior across three domains** (recall@R10, 20 seeds). The same agent prior is fed to every method; **bold** marks the row best. HARIBO is above the prior arm everywhere, above pure BO wherever the prior is informative, and within 0.007 of the row best; the fixed blend that edges it here collapses by 0.150 under an adversarial prior (Table 4). We omit corr(π, y) for the sparse gene nomination indicator ($\sim$120 of $\sim$15k genes).

domain (prior)	pure BO	agent arm	fixed blend	HARIBO	corr(π, y)
gene (Qwen2.5-32B agent, 6 screens)	0.111	0.099	0.133	**0.133**	sparse
gene (Claude Sonnet 5 agent, 6 screens)	0.111	0.116	**0.147**	0.140	sparse
protein (Qwen2.5-32B, n=184)	0.560	0.518	**0.605**	0.602	+0.08
antibody (Qwen2.5-32B, n=43)	0.625	0.471	0.607	**0.626**	≈ 0

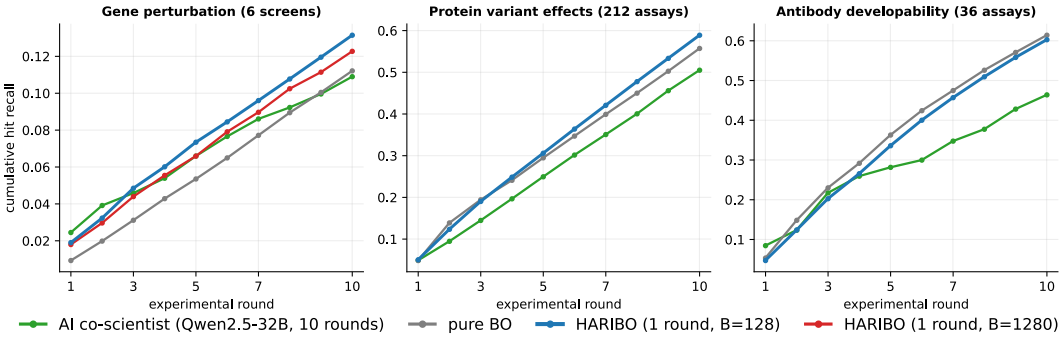


Figure 3: **One round against ten**, per experimental round. The agent is re-queried with the measured outcomes every round, against pure BO and HARIBO, whose prior is elicited in round one only: the legend counts *query* rounds, all arms run ten measurement rounds. B is the elicited list length, defined only on gene. Batches follow the pool (Table 2), so recall compares within a panel, not across. Panels use the assays the agent can be prompted on (n=212 protein; 36 of 43 antibody).

correlation with fitness (Spearman +0.86 for ESM2, +0.72 for the agent), and the ESM2 weight exceeds the agent's on 93% of assays: w_π is an interpretable estimate of how much a given prior deserves to be believed on a given assay, not only a device for ranking.

4.2 ROBUSTNESS TO WRONG AND ADVERSARIAL PRIORS

Fig. 4 shows the learned coefficient round by round. It is never told which prior it has been handed: on a screen where the nomination is informative the weight rises and stays, on a prior elicited for a different screen it settles between, and under −ESM it goes negative and stays there. The ordering it discovers matches the ordering of the priors' actual quality on 11 of the 12 screen×model cells, including the one cell where a misapplied list is genuinely the more useful of the two and the weight prefers it (Appendix C.4, Fig. 7).

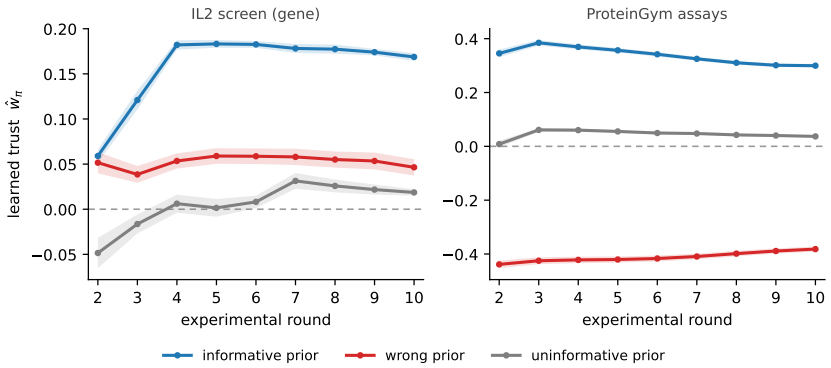


Figure 4: **The learned trust, round by round.** Standardised $\hat{w}_\pi$=corr$(\tilde{\pi}, \tilde{y})$, mean ± 1 s.e. over 10 seeds, for an informative, a wrong and an uninformative prior in the same loop. The wrong prior is another screen's nomination on gene and the sign-flipped score on protein, so only the protein panel goes negative.

The safety property above matters most under a *systematically wrong* prior, the regime that poisons methods with a pre-fixed trust (Table 4). Under an adversarial −ESM prior the fixed blend collapses 0.150 below no-prior BO ($p < 10^{-24}$), while HARIBO learns $w_\pi < 0$, *inverts* the prior, and ends 0.145 above BO, as Prop. 1 predicts. Every method that sets the weight in advance is poisoned, losing 0.07–0.16; the only baseline that survives is the one that also reads the weight off the labels, which is an ablation of HARIBO rather than a competing design; the decay schedules of πBO and BOPrO help only in that they eventually stop listening. The trade-off is sharpest for πBO, the only method that edges HARIBO on the good prior (0.730 vs. 0.725) and the one that collapses hardest on the wrong one: 0.005 gained in the benign case against 0.162 lost in the adversarial one is a bet no practitioner can size before measuring. LGBO shows why self-report is no substitute for labels: scaling its shift by the model's confidence, it moves confidently the wrong way and ends 0.132 below BO. The same pattern appears on antibodies, where ESM2 is naturally adversarial on some assays (correlation with fitness from +0.52 to −0.73): the fixed blend drops to 0.608 against BO's 0.625 while HARIBO holds 0.630. Across both prior families and signs, its worst case never falls below BO (0.626 vs. 0.625 on antibodies), against 0.150 and 0.162 for the fixed blend and πBO.

Table 4: **Adversarial robustness** (protein, n=184, recall@R10, 20 paired seeds; no-prior BO 0.560). Fixing the prior's weight by a schedule or by self-report loses far more on a wrong prior than it gains on a good one. The penultimate row is the obvious repair, a blend whose λ is read off the labels and may go negative; it works, which is the point, and HARIBO still beats it by 0.017 and 0.022 ($p \leq 0.002$) because it weights only the part of the prior the features do not already carry.

method	how the prior's weight is set	good (+ESM)	adversarial (−ESM)	Δ vs. BO, adv.
fixed blend	constant, chosen in advance	0.711	0.410	−0.150
πBO (Hvarfner et al., 2022)	decays with the round index	**0.730**	0.398	−0.162
BOPrO (Souza et al., 2021)	decays with the number of labels	0.636	0.490	−0.070
LGBO (Yuan et al., 2026)	the model's self-reported confidence	0.706	0.428	−0.132
blend with a *learned* λ	refit from labels each round, sign-free	0.703	0.688	+0.128
HARIBO (ours)	**learned from labels**	**0.725**	**0.705**	**+0.145**

Nor can the blend's weight be chosen in advance: the best fixed λ is 0.4, 0.6 and 0 under the weak, good and adversarial priors (Appendix C.5), and HARIBO stays within 0.004 of that oracle in both benign rows while beating every fixed λ by 0.146 where the weight must go negative.

4.3 ABLATION AND SENSITIVITY STUDY

The essential-gene filter. Keeping the CEGv2 genes the benchmark counts as hits raises every arm, pure BO most (0.111 → 0.121), and shrinks HARIBO's gene margin from +0.022 to +0.013 on Qwen2.5-32B and from +0.029 to +0.015 on Claude Sonnet 5 (p=0.031 → 0.31). Every sign is unchanged; the significance of the gene margin rests on the filter.

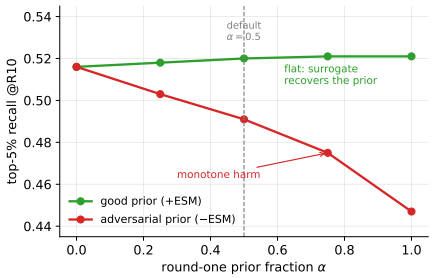


Figure 5: **The round-one split is a robustness choice.** Protein recall@R10 as the prior fraction of the first batch varies: nearly flat under a good prior (+ESM), monotonically worse under an adversarial one (−ESM), so an even split is safe under both.

Round one: identification, not recall. Sec. 3 motivates an even first batch as the design that identifies w_π fastest. Against $\gamma^\star$ computed from all labels, a balanced round one recovers the sign of w_π on 98% of gene screens, against 32% for an exploratory one (about one nominated gene per 128) and 75% for an all-prior one (no un-nominated contrast; Table 8). Identification and recall are different axes: the balanced design also beats the exploratory one on recall (0.139 vs. 0.125, 11 of 12 screen-model cells, $p = 0.016$), but all-prior reaches the same 0.139. What the even split buys is a correctly signed trust coefficient, which matters when the prior is wrong: on protein (Fig. 5), recall is nearly flat in the prior fraction under a good prior and degrades monotonically under an adversarial one, so an even split is worst-case safe. Sharpening the estimate further, by selecting the prior half with largest $\tilde{\pi}^2$ or reserving part of every round for identification, cuts the error on $\widehat{w}_\pi$ by up to five-fold but never raises recall, and lowers it on protein, since budget spent on identification is budget not spent on acquisition (Appendix A.4). One balanced batch is enough.

Scale, prompting and implementation. None of these changes the conclusion. Running the same loop with agents from 7B to frontier scale, HARIBO reaches 0.123, 0.133 and 0.140 on gene with Qwen2.5-7B, Qwen2.5-32B and Claude Sonnet 5, all above pure BO's 0.111 (Table 9): the gain grows with the prior's quality, and on the two open models the prior is worth *less* than no prior at all (0.094 and 0.099 against 0.111) yet HARIBO still ends above BO. Better prompting is no substitute: none of five round-one designs beats the agent's own prompt on Qwen2.5-32B, and only their five-call union does (Appendix C.1). One implementation choice matters: leaving the prior coordinate unpenalized is worth $+0.023$ and $+0.027$ on protein ($p \leq 10^{-4}$; Appendix A.2). How the nomination is encoded does not, as long as it is not ordered by the model's emission rank, which is worse than a plain indicator on both models (Appendix A.1).

4.4 EFFICIENCY OF HARIBO

Because HARIBO consumes the prior as a feature, the model is queried in *one* round and never again. A measurement round is not a single request: the harness re-prompts until the batch is full (Algorithm 1 of Roohani et al., 2025), so ten rounds are 162 requests on Qwen2.5-32B and 34 on Claude Sonnet 5. Per gene screen that is 20–32$\times$ our requests and 124–234$\times$ our tokens (Table 5), a ratio that shrinks severalfold on a metered API, since our single round is almost all output while the agent's stream re-sends a growing history as input; HARIBO then runs a CPU-only ridge, 39 ms for ten rounds on the $14{,}582$-gene pool, against a 65 GB model held loaded throughout. What the extra rounds buy depends on the domain rather than on the method.[5] On Qwen2.5-32B they buy nothing: one round plus HARIBO leads on gene (0.133 vs. 0.109, all six screens, p=0.031), protein (0.589 vs. 0.505, n=212) and antibody (0.603 vs. 0.464, n=36), and on the last two the agent finishes *below* no-prior BO. With Claude Sonnet 5 the agent gains on gene (0.169 vs. 0.140) but still loses on protein (0.650 vs. 0.703, p=1.8×10^{-4}, n=194) and antibody (0.449 vs. 0.602, p=5.8×10^{-4}, n=36), where it is also slightly worse than the 32B agent. Scaling the proposer helps a closed-loop agent on one domain of three; where its gene advantage comes from, what the extra rounds do and do not buy, and how much of it a wider single call recovers are in Appendices C.6, C.7 and C.8.

Table 5: **What ten rounds cost**, per gene screen. Requests and tokens are server-logged, not estimated; exact token counts are $1{,}389{,}467$ and $5{,}925$ on Qwen2.5-32B and $283{,}553$ and $2{,}278$ on Sonnet 5.

	Qwen2.5-32B		Sonnet 5	
	req.	tok.	req.	tok.
AI co-scientist	162	1.39M	34	284k
HARIBO	**5.0**	**5.9k**	**1.7**	**2.3k**
reduction	32$\times$	234$\times$	20$\times$	124$\times$

5 CONCLUSION

AI co-scientists hold enormous promise for accelerating closed-loop scientific discovery, but their reliability is fundamentally coverage-conditioned: they may confidently hallucinate and fail silently exactly where experimental guidance is needed most. This paper argues that the key challenge is not whether to use AI-generated priors, but how much to trust them. To this end, we introduced HARIBO, a minimalistic yet highly interpretable framework that incorporates an AI prior as evidence within Bayesian optimization and learns its reliability directly from experimental observations. Consequently, HARIBO automatically exploits informative priors, ignores uninformative ones, and even inverts systematically misleading priors, so that the benefit of AI guidance is retained where it exists and an uninformative prior returns the loop to the no-prior baseline rather than degrading it.

Limitations. While we evaluate HARIBO across diverse biological discovery datasets, prospective validation in real laboratory discovery campaigns remains an important next step. Our theoretical analysis assumes a linear surrogate; extending the guarantees and trust learning to nonlinear surrogates remains future work. Learning the trust also does not manufacture knowledge the prior lacks: where a frontier co-scientist's coverage of a phenotype is already deep, re-querying it every round still finds more hits than one round plus a knowledge-free surrogate, and which regime a screen falls in cannot be read off before measuring it.

[5] These assay counts differ from Table 3's n=184 because every arm here is scored on the assays the agent can be prompted on: the 250-variant matched protein cache (n=212, or the 194 a frontier agent completed without declining) and the 36 of 43 antibody assays whose pool fits in one prompt. Mixing the two pools would flatter whichever arm was scored on the smaller one.

AI USE STATEMENT

We used generative AI tools to check grammar and improve the clarity of text drafted by the authors. We also used language-model agents as the AI co-scientists studied in this work: they generated or scored candidate priors through the prompts described in the paper. We evaluated their outputs against the assay data and checked the reported results. The models and prompting procedures used in the experiments are specified in the main text and appendix. The authors reviewed all AI-assisted content and take responsibility for the final manuscript, analyses, and conclusions.

ETHICS STATEMENT

This work evaluates closed-loop selection methods retrospectively using existing biological assay datasets. It does not report new laboratory experiments or recruit human participants. Our analysis highlights the risk of relying on fabricated or uninformative AI-generated candidates in biological discovery. We do not claim that the retrospective results establish safety or effectiveness in a prospective laboratory setting.

REPRODUCIBILITY STATEMENT

Section 4 specifies the prior construction, surrogate, and acquisition procedure. The experimental setup and comparisons are described in the main text, with additional implementation details, ablations, and analyses in the appendix. We report results over multiple random seeds and identify the source datasets and models used in each domain. Source code of HARIBO is provided here: `https://anonymous.4open.science/r/ddk3ajdkl`.

REFERENCES

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools, 2023. URL `https://arxiv.org/abs/2304.05376`.

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, et al. RASA2 ablation in T cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925):174–182, 2022.

Michael Chungyoun, Jeffrey Ruffolo, and Jeffrey Gray. FLAb: Benchmarking deep learning methods for antibody fitness prediction. *BioRxiv*, pp. 2024–01, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods*, 21(8):1470–1480, 2024.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Peter I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Wenpin Hou and Zhicheng Ji. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods*, pp. 1–4, 2024.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature methods*, 22(1):82–91, 2025.

Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In *International Conference on Research in Computational Molecular Biology*, pp. 17–37. Springer, 2024.

Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pp. 507–523. Springer, 2011.

Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. πBO: Augmenting acquisition functions with user beliefs for bayesian optimization. *arXiv preprint arXiv:2204.11051*, 2022.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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 *International Conference on Machine Learning*, pp. 23170–23189. PMLR, 2023.

Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. Genedisco: A benchmark for experimental design in drug discovery. *arXiv preprint arXiv:2110.11875*, 2021.

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

Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36: 64331–64379, 2023.

Laralynne Przybyla and Luke A Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022.

Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C Cousins, William A Johnson, Xiaotong Wang, Mihir Shah, et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, 10(2):245–258, 2026.

Bojana Ranković, Ryan-Rhys Griffiths, and Philippe Schwaller. Large language models as uncertainty-calibrated optimizers for experimental discovery. *Nature Machine Intelligence*, pp. 1–12, 2026.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, 2023.

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*, volume 2025, pp. 26417–26466, 2025.

Carlos G Sanchez, Christopher M Acker, Audrey Gray, Malini Varadarajan, Cheng Song, Nadire R Cochran, Steven Paula, Alicia Lindeman, Shaojian An, Gregory McAllister, et al. Genome-wide CRISPR screen identifies protein pathways modulating tau protein levels in neurons. *Communications biology*, 4(1):736, 2021.

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.

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.

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, 2015.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pp. 2171–2180. PMLR, 2015.

Artur Souza, Luigi Nardi, Leonardo B Oliveira, Kunle Olukotun, Marius Lindauer, and Frank Hutter. Bayesian optimization with a prior for the optimum. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 265–296. Springer, 2021.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Joy Y Wang and Jennifer A Doudna. CRISPR technology: A decade of genome editing is only the beginning. *Science*, 379(6629):eadd8643, 2023.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37: 33402–33422, 2024.

Xinzhe Yuan, Zhuo Chen, Jianshu Zhang, Huan Xiong, Nanyang Ye, Yuqiang Li, and Qinying Gu. Unleashing llms in bayesian optimization: Preference-guided framework for scientific discovery. In *The Fourteenth International Conference on Learning Representations*, 2026.

APPENDIX

This appendix collects the material that supports the main text. Subsections appear in the order they are first cited.

- **§A, Implementation details.** §A.1 how the gene nomination is encoded; §A.2 where the ridge penalty is applied, and the exact finite-α coefficient; §A.3 why one trust coefficient rather than several; §A.4 the round-one design, in full.
- **§B, Prompts.** §B.1 gene nomination; §B.2 protein scoring; §B.3 antibody scoring.
- **§C, Additional experiments.** §C.1 scale, architecture and prompting; §C.2 a GP explorer in place of the ridge; §C.3 per-round curves under a foundation-model prior; §C.4 the learned trust on every gene screen; §C.5 sweeping the blend's fixed weight; §C.6 what ten rounds cost, and where the frontier agent's gene advantage comes from; §C.7 what the extra rounds buy, and what did not close the gap; §C.8 scaling the elicitation.

A IMPLEMENTATION DETAILS

A.1 ENCODING THE GENE PRIOR

The agent returns a set of gene symbols. Two things could be encoded: the fact of nomination, and a gene's position within the returned list. We encode the first. Position is not uninformative: pooled over the six screens the correlation between list position and hit status is -0.16 for Claude Sonnet 5 and -0.03 for Qwen2.5-32B (p=0.44), so a frontier agent does put likelier genes earlier. But encoding it does not pay at this width. A list of $\sim$120 names is exhausted by the first 128-gene batch, so what survives into rounds two to ten is which genes were named, not the order in which they were, and Table 6 bears that out: the indicator is ahead on $4/6$ screens for both models, by $+0.008$ on Qwen2.5-32B and -0.003 on Sonnet 5 in the mean; at $4/6$ screens neither difference is significant. What the table does settle is that ordering by the model's *genuine* emission rank, recovered from the run logs, is worse than the indicator on both models. One caveat applies to any measurement on saved agent output: the batch is assembled with a set difference, so the stored within-round order is a hash order rather than the model's emission order, which is why the two orderings are reported separately.

Table 6: **Round-one prior encodings** (gene, HARIBO recall@R10, 20 paired seeds). "stored rank" is the published encoding, whose ordering is a hash ordering; "true-rank (floored)" uses the emission order recovered from the run logs, mapped to $[0.5, 1]$. No difference between the indicator and the stored rank is significant over six screens. Ordering by the genuine emission rank loses to the indicator on both models.

encoding of π for a nominated gene	Qwen2.5-32B	Claude Sonnet 5	note
stored rank $(N-\text{rank})/N$ (published)	0.125	0.143	ordering is a hash order
indicator π=1 (ours)	**0.133**	**0.140**	ahead on $4/6$ screens
true-rank (floored), emission order	0.126	0.129	position is real but does not pay
pure BO (no prior)	0.111	0.111	

A.2 WHERE THE RIDGE PENALTY IS APPLIED

Neither zero is a solver artefact. The intercept is unpenalized because `scikit-learn` centres the design before applying the penalty, checked against the closed form to 4×10^{-15}. The prior coordinate cannot be freed the same way, since `Ridge` applies one α to every column, so w_π is solved in the closed form below. On the small antibody assays, where an OLS projection inside the batch would saturate, the penalized projection keeps $\sum \tilde{\pi}_S^2$ between 2 and 5 on all 43 assays: it is the design argument of Eq. 3 that narrows there, not the estimate.

The exact finite-α coefficient. In practice the features carry a cross-validated penalty $\alpha > 0$. On the observed rows, let $X = [\mathbf{1}, \Phi]$ and $A_\alpha = \text{diag}(0, \alpha, \ldots, \alpha)$, and define $M_\alpha = I - X(X^\top X + A_\alpha)^{-1}X^\top$. Eliminating the intercept and feature coefficients from Eq. 1 gives the exact finite-α

coefficient $w_\pi = \pi^\top M_\alpha y / (\pi^\top M_\alpha \pi)$. Unlike the ordinary least-squares residual maker, M_α is not idempotent, so regressing $M_\alpha y$ on $M_\alpha \pi$ does not generally recover w_π. A penalty α' on the prior coordinate would multiply its unpenalized coefficient by $D_\alpha/(D_\alpha + \alpha')$, where $D_\alpha = \pi^\top M_\alpha \pi$. That pulls a wrong prior toward being ignored but shrinks a good one just as hard, and the features need a penalty orders of magnitude above the prior's own scale. Proposition 1 is a least-squares limit, but its *sign* reading survives: M_α is positive semidefinite, so $D_\alpha > 0$ unless π is fitted exactly, and w_π takes the sign of $\pi^\top M_\alpha y$.

The penalty is not cosmetic at these sample sizes. At round two the gene surrogate has 128 observations and 51 standardized columns; with a fixed α=1 it interpolates the observed rows and then extrapolates over the $\sim$14.6k unmeasured genes. Measured on IFNG with Sonnet 5, the feature part of the prediction has standard deviation 2.99 across candidates against a learned prior contribution of 1.20, so the nominees still on the table sit at median rank 3692 and are never acquired. Cross-validating α each round fixes this without introducing a tuned constant, and it helps pure BO as much as HARIBO, so it is not a change that flatters the method.

Applying that same α to the prior coordinate is a separate error. The value chosen to tame $\sim$50 noisy feature columns is orders of magnitude larger than the prior's own scale, and it shrinks w_π below the $\gamma^\star$ of Prop. 1, which is by construction the *unpenalized* coefficient of $\tilde{\pi}$. Eq. 1 therefore uses $P_\alpha = \mathrm{diag}(0, 0, \alpha, \ldots, \alpha)$, zeroing the intercept and the prior coordinate alike. The gain from freeing that coordinate is confined to *dense* priors: on protein it is worth $+0.023$ under the agent prior (p=10^{-4}) and $+0.027$ under an adversarial one (p=10^{-8}), while on the sparse gene indicator it changes nothing ($+0.000$ and -0.001; Table 7). A sparse binary prior is already close to orthogonal to the features, so the shared penalty has little of w_π to remove; a dense prior competes with them directly.

Table 7: **Where the ridge penalty is applied** (recall@R10, 20 paired seeds). *one fixed* α=1: the repo's original default, on every coordinate. *CV α, all*: cross-validated each round, on every coordinate including the prior. *CV α, features*: cross-validated on the features only, prior coordinate free – the setting used throughout the paper. p-values are paired Wilcoxon between the last two columns.

domain	method	α=1, all	CV α, all	CV α, features	p
gene (Sonnet 5, 6 screens)	pure BO	0.108	0.111	0.111	–
	HARIBO	0.131	**0.140**	**0.140**	0.84
gene (Qwen2.5-32B)	pure BO	0.108	0.111	0.111	–
	HARIBO	0.125	0.132	**0.133**	0.56
protein (n=184, agent prior)	pure BO	0.554	0.560	0.560	–
	HARIBO	0.578	0.579	**0.602**	10^{-4}
protein (n=184, −ESM)	pure BO	0.554	0.560	0.560	–
	HARIBO	0.644	0.678	**0.705**	10^{-8}

A.3 WHY A SINGLE TRUST COEFFICIENT

The coverage diagnosis says prior reliability varies over candidate space, so a constant w_π looks like the wrong model. We tested that by letting trust vary, adding one to fifty interaction columns to the design: $\mathrm{zs}(\pi^2)$, so trust depends on where in the prior's own ranking a candidate sits; $\mathrm{zs}(\pi \cdot \mathrm{dens})$, with dens the local feature-space density as a coverage proxy; $\pi\phi_{1:5}$; and the full $\pi\phi_{1:50}$, which estimates $\gamma^\star(x) = \gamma + v^\top \phi(x)$ outright. Recall on 201 protein assays under a good prior was 0.523 for the scalar against 0.517, 0.516, 0.518 and 0.499, and under an adversarial prior 0.492 against 0.497, 0.487, 0.494 and 0.493. A separate variant that makes the prior branch itself vector-valued (per-position mean and spread of the log-likelihood ratio alongside the scalar) is also null, within 0.007 of the scalar on both signs. Only the full interaction reaches significance, and it *hurts* the good-prior case (-0.024, p=5×10^{-5}) while buying nothing adversarial. These runs predate the protocol of the main tables (fixed α=1, batch 64), so the levels are not comparable to Table 4; the comparison within the table is paired and is what the conclusion rests on. What governs acquisition is the ranking w_π induces, and a scalar already reorders the pool correctly whenever its sign is right.

A.4 ROUND-ONE DESIGN: THE FULL TABLE

Sec. 4.3 reports the three headline designs. Table 8 adds two refinements: selecting the prior half by largest residual $\tilde{\pi}^2$, and reserving a share of *every* round for identification rather than only round one. Both sharpen the estimate and neither changes recall. That is the expected pattern – acquisition depends on the ranking w_π induces, not on how precisely it is known – and it also means the recall column should not be read as following the identification column: the all-prior design identifies worse than the balanced one and reaches the same recall.

Table 8: **Round-one design: identification and recall** (gene, 6 screens $\times$ 2 proposers, 20 seeds). $\gamma^\star$ is the population trust coefficient of Eq. 2, computed from all labels; $\widehat{w}_\pi$ is what the loop estimates after round one. "$\tilde{\pi}^2$-balanced" keeps the even split but picks the prior half by largest residual instead of by prior rank.

round-one design	sign of $\widehat{w}_\pi$ correct	$\lvert\widehat{w}_\pi - \gamma^\star\rvert$	recall@R10
exploratory (ρ=0)	32%	0.039	0.125
all-prior (ρ=1)	75%	0.041	0.139
balanced (ρ=1/2, ours)	**98%**	**0.033**	**0.139**
$\tilde{\pi}^2$-balanced	95%	0.016	0.140
balanced + 10% of every round	100%	0.011	0.140
balanced + 25% of every round	100%	0.007	0.139

B PROMPTS

Every prior in the paper comes from one of the three prompts below, issued once per screen or assay before any measurement. The gene prompt is BIODISCOVERYAGENT's own, reproduced unchanged so that our prior and the re-querying baseline are elicited identically; the protein and antibody prompts are ours, and score each candidate rather than returning a list, because those pools are enumerable. Braces mark the fields substituted per screen or assay.

B.1 GENE: NOMINATION

```
You are a scientist working on problems in drug discovery.
Research Problem: {task} There are 18,939 possible genes to perturb and I can only
perturb {B} genes at a time. For each perturbation, I'm able to measure out {measurement}
which will be referred to as the score. I can only do a few rounds of experimentation.
Always respond in this format exactly:
1. Reflection: Thoughts on previous results and next steps.
2. Research Plan: The full high level research plan, with current status and reasoning
behind each proposed approach. It should be at most 5 sentences.
3. Solution: Propose a list of predicted genes to test separated by commas in this
format: 1. <Gene name 1>, 2. <Gene name 2> ...
Do not include any genes from this prompt (since they're already tested).
```

The prior is the nomination indicator over the returned symbols. The re-querying baseline uses the same prompt with the measured score and hit status of every previously proposed gene appended from round two onward, which is what makes it a closed-loop competitor rather than a prior (§C.7). For the wide-elicitation configuration of §C.8 only {B} changes, from 128 to $B \times R$=1280.

B.2 PROTEIN: PER-VARIANT SCORING

```
You are an expert protein biochemist analyzing a deep mutational scanning experiment.
Protein: {name} (UniProt {uniprot}, organism: {organism}).
Measured property: {assay} ({selection type}). A HIGHER measured score means the variant
is MORE functional/stable/fit for this assay.
For each single mutation (format WT-position-MUT, e.g. A123P), predict its effect on the
measured property on a 0-10 scale: 0 = strongly decreases/abolishes, 5 = neutral, 10 =
strongly increases. Use your knowledge of THIS specific protein's structure and function.
If you are unfamiliar with this protein, output 5.
Output EXACTLY one line per mutation as 'MUTATION: score'. No other text.
Mutations: {40 variants per request}
```

B.3 ANTIBODY: PER-CANDIDATE SCORING

```
You are an antibody engineer assessing developability. Property: {property}. A HIGHER
score means the antibody is MORE favorable for this property.
For each antibody, predict a 0-10 score (0=poor, 5=average, 10=excellent) from its
heavy (VH) and light (VL) sequences. Use your knowledge of antibody biophysics (CDR
composition, hydrophobicity, charge, liabilities). If unsure, output 5.
Output EXACTLY one line per antibody as 'Ab<j>: score'.
Ab0: HEAVY={VH} | LIGHT={VL}
Ab1: ...
```

The protein and antibody prompts are issued in fixed-size chunks so that each request stays well inside the context window; the union of the chunks covers the pool exactly once, and the returned scores are z-scored per assay before entering the surrogate. A candidate the model declines to score, or scores unparseably, is left at the pool mean.

C ADDITIONAL EXPERIMENTS

C.1 SCALE, ARCHITECTURE AND PROMPTING

A round is not a request. The harness re-prompts within a round until the batch is full, so round one costs 1.7 requests on Claude Sonnet 5, which fills it in its first response, and 5.0 on Qwen2.5-32B, whose first response names only 26 in-panel genes. Keeping only that first response costs Sonnet 5 nothing (0.142 against 0.140) but drops Qwen2.5-32B to the no-prior baseline, which is why the round rather than the request is the unit we report.

Table 9: **The prior gets better with scale; the loop is what makes it safe** (gene, 6 screens, recall@R10, 20 seeds, pure BO $= 0.111$ for every row). "pure Prior" measures the nominated genes first and then draws at random, which is all a sparse 128-name list can do once it is exhausted. On both open models that is *worse* than having no prior, and HARIBO still ends above BO. Qwen2.5-7B returns no in-panel gene at all on Steinhart, where the prior is empty and HARIBO reduces exactly to BO.

proposer	in-panel names	hit enrichment	pure Prior	HARIBO	Δ vs. BO
Qwen2.5-7B	98	2.4×	0.094	0.123	+0.012
Qwen2.5-32B	119	3.1×	0.099	0.133	+0.022
Claude Sonnet 5	121	5.3×	0.116	**0.140**	**+0.029**

Table 10: **Fabrication is produced by the demand for a full batch.** Fraction of proposed symbols that no HGNC lookup resolves, Qwen2.5-32B, one call, mean of two replicates. *Free-length*: the model is asked for what it is confident about and stops on its own, returning 7–15 symbols. 128-*gene demand*: it is asked for the batch size a closed-loop screen fixes, and returns 94–190. Scharenberg22, the screen of Fig. 1, is left out because its free-length rate is the one that is not stable across replicates; Fig. 1 is a separate run with a different phenotype wording and a 64-gene demand.

screen (phenotype)	free-length prompt	128-**gene demand**
IFNG (IFN-γ production, well studied)	0%	**0**%
Steinhart (CRISPRa, GD2 CAR-T)	6%	8%
Sanchez21_down (tau levels in neurons)	0%	56%
Sanchez21 (tau levels in neurons)	0%	**66**%
IL2 (IL-2 production)	0%	61%
Carnevale22 (adenosine suppression)	0%	64%

A batch demand, not model scale, produces fabrication. Asked for what it is confident about, Qwen2.5-32B returns 7–15 symbols and invents almost none (Table 10, left). Asked for 128, it returns 94–190 and invents 56–66% of them on four of the six screens. The mechanism is padding rather than ignorance: on the lysosomal screen of Fig. 1 it exhausts the real phospholipases at `PLA2G12B` and then continues the letter, and on the tau screens it emits `PPP2R5P`, `PPP2R1C`, `DCTNIP1` and similar variants of real families. Two screens resist: IFNG, where the literature holds enough real candidates, and Steinhart, where the model returns far fewer than asked.

Scale narrows the problem without dissolving it. Claude Sonnet 5 fabricates 0% on every screen at the same 128-gene demand, so the failure of Fig. 1 is not inherent to AI co-scientists. What does not change with scale is coverage: on a thinly-studied phenotype Sonnet 5's proposals are enriched only $1.8\times$ over the panel base rate, against $\sim 10\times$ on the two best-covered screens, because the limit there is the literature and not the reader.

Prompt engineering does not replace learned trust. Five round-one elicitation designs were scored against the agent's own prompt on the same six screens, 3 prompt samples $\times$ 20 paired seeds each: the agent's prompt plus an explicit confidence ordering, plus HIGH/MED/LOW tiers, an open-form list with no target count, and open-form with each of those two additions. On Qwen2.5-32B none of the five beats it, the agent's prompt reaching 0.131 against 0.109–0.123; only the union of all five, which is a five-call ensemble rather than a prompt, edges ahead at 0.133. (Those figures come from that sweep's own reference arm, hence 0.131 rather than Table 3's 0.133.) The open-form designs are the informative failure: dropping the target count raises per-gene hit precision on $4/6$ screens while roughly halving the list, 57 in-panel names on Steinhart against 126, and it is depth that round ten pays for. Self-assigned confidence tiers are also not monotone in hit precision, the per-candidate analogue of the self-report failure of §4.2.

C.2 A GP EXPLORER INSTEAD OF THE RIDGE

The construction of Sec. 3 is not specific to the ridge. Place π as an input dimension of a GP over $u = [\pi; \phi]$, with posterior mean

$$\widehat{y}_{\mathrm{GP}}(x) = k\big(u(x), U\big)^{\top}\big(K + \sigma_n^2 I\big)^{-1} y, \tag{4}$$

and acquire by GP-UCB ($\widehat{y}_{\mathrm{GP}} + \beta\sigma_{\mathrm{GP}}$, $\beta = 1$) (Srinivas et al., 2009). Trust becomes the inverse lengthscale $1/\ell_\pi^2$ of that dimension, learned by marginal likelihood, and an adversarial prior needs no sign parameter because the GP can fit a decreasing effect of π. Influence is still read off the data, but it is no longer one signed scalar, so that interpretability is specific to the linear surrogate; conclusions match the ridge (Table 11).

Table 11: **HARIBO is explorer-agnostic** (protein, n=104, recall@R10). Both surrogates choose their own regularization – the ridge by cross-validation, the GP by marginal likelihood – and the two land within 0.025 of each other on every row.

	ridge	GP
pure BO	0.651	0.678
HARIBO (+ESM)	**0.791**	0.768
HARIBO (−ESM)	**0.769**	0.754

Swapping the ridge for a GP-UCB explorer (Srinivas et al., 2009) leaves the conclusions intact. The GP is the better no-prior explorer (0.678 against 0.651), the ridge is slightly better once the prior is attached, and the two land within 0.025 of each other on every row. Adding π as an input dimension of a GP over $[\phi; \pi]$ also recovers above GP-BO under an adversarial prior, so absorbing the prior as evidence is a property of the construction rather than of the ridge.

C.3 PER-ROUND CURVES UNDER A FOUNDATION-MODEL PRIOR

Sec. 4.1 reports the endpoint of each curve; the trajectories are below. The pattern is the same one the learned weight produces elsewhere: where the foundation model carries signal HARIBO separates from pure BO within two or three rounds, and where it does not the two curves stay on top of each other for all ten.

C.4 THE LEARNED TRUST ON EVERY GENE SCREEN

Fig. 7 shows the learned trust on all six gene screens.

C.5 SWEEPING THE BLEND'S FIXED WEIGHT

A fixed blend is the strongest simple baseline whenever its weight happens to sit near $\gamma^\star$, and unlike HARIBO it pays no estimation error. Table 12 asks whether that weight can be chosen before any

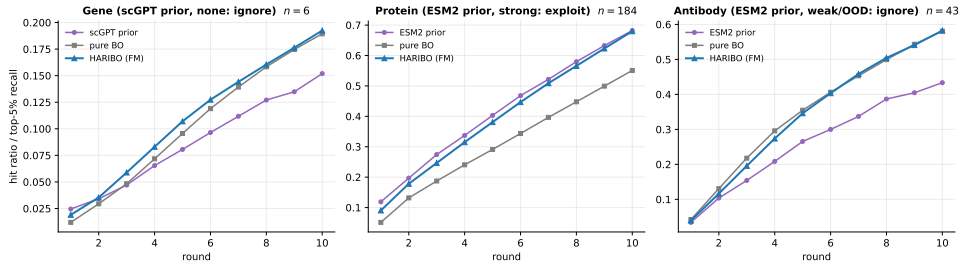


Figure 6: **Foundation-model prior across domains**, per round. HARIBO (FM) exploits the strong ESM2 prior on protein residues and tracks pure BO where the foundation model carries no usable signal (scGPT on genes; ESM2 on antibody CDRs, where the prior alone falls below BO).

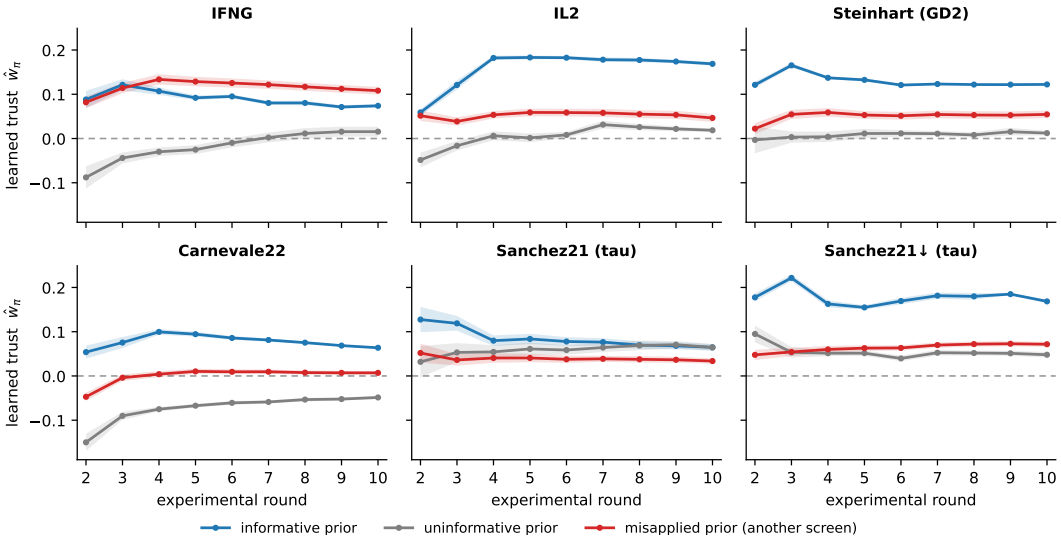


Figure 7: **Standardised $\hat{w}_\pi$ per gene screen** (Qwen2.5-32B nominations, mean ± 1 s.e. over 10 seeds). The weight tracks usefulness, not provenance: it is never told which list was elicited for which screen. On IFNG the model's own nomination is weaker than the lists elicited for the other screens ($1.90\times$ hit enrichment against $2.94\times$) and the weight correctly prefers the misapplied one. Across the twelve screen$\times$model cells the ordering of $\hat{w}_\pi$ matches the ordering of prior enrichment in eleven. On Sanchez21 all three priors converge to the same value, which is the right answer: no monotone function of that screen's score ranks its hits (AUC 0.500), so there is nothing for any prior to be right about.

measurement. It cannot: the optimum is a different number in each regime, and committing to 0.5 gains 0.002 over HARIBO on the weak prior, gives up 0.015 on the good one, and loses 0.295 on the adversarial one, landing 0.149 below no-prior BO.

Table 12: **The blend's weight cannot be chosen in advance** (protein, n=184 assays, 10 seeds, recall@R10). $s_\lambda = (1-\lambda)\,\mathrm{z}(\hat{\mu}) + \lambda\,\mathrm{z}(\pi)$ for a *fixed* λ; HARIBO reads the weight off the labels and may make it negative. The best λ is a different number in each row, and which row an assay belongs to is unknown until it has been measured.

prior	λ=0	0.25	**0.4**	**0.5**	**0.6**	1.0	best λ	HARIBO
agent, weak (corr +0.08)	0.559	0.588	**0.605**	0.604	0.577	0.518	0.40	0.602
ESM, good (corr +0.44)	0.559	0.661	0.701	0.711	**0.730**	0.682	0.60	0.726
ESM, adversarial	**0.559**	0.457	0.430	0.410	0.384	0.317	0.00	**0.705**

C.6 What ten rounds cost, and where the gene advantage comes from

With Claude Sonnet 5 the ten-round agent leads on gene, 0.169 against 0.140 for one round plus HARIBO (Fig. 8). That advantage is not uniform across screens. On the three whose round-one nominations are most enriched it is worth $+0.013$ to $+0.084$ over HARIBO; on the three least enriched it is worth between -0.007 and $+0.003$, i.e. nothing. Which of the two a screen is cannot be read off before it has been measured, so preferring the agent is a wager on coverage that cannot be settled in advance, the condition this paper is about. A wider single call recovers much of it, at a cost described in §C.8.

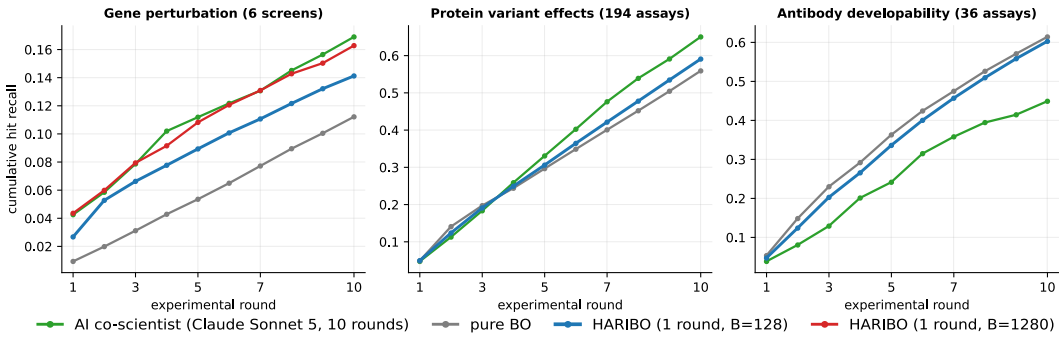


Figure 8: **A frontier proposer does not change the conclusion outside gene.** Fig. 3 repeated with Claude Sonnet 5 as both the agent and the source of our prior. The agent gains on gene, where a wider single call recovers most of the gap, and loses on protein and antibody, where knowledge-free BO also beats it.

C.7 Depth: what extra LLM calls buy

Sec. 4.4 compares one zero-shot call against the published re-querying loop. Two facts qualify that comparison. From round two the agent's prompt carries the measured score and hit status of every gene it previously proposed, so it is not a prior but a closed-loop competitor consuming the same labels our surrogate consumes, with a language model standing where we put a ridge. And the two arms are matched on measurement budget: the agent proposes $\sim$1,125 candidates over ten rounds on Qwen2.5-32B and $\sim$1,180 on Sonnet 5, against our 1,280.

Where the agent does lead, on gene with a frontier proposer, the binding constraint is our feature space rather than the trust: pure BO on STRING-PCA50 reaches 0.111, close to the base rate for ten batches of 128 against $\sim$558 hits, so once the $\sim$120 nominated genes are exhausted in round one the surrogate must generalise from STRING features that carry little signal on these phenotypes while the agent keeps reading the measurement table (§C.6).

What did not close the gap. Recorded so they need not be repeated: propagating the nomination through DepMap co-dependency (-0.003), distilling the agent's nomination rule onto the panel with a logistic or forest classifier ($+0.003$), a column recomputed each round from the hits confirmed so far ($+0.004$), and a random-forest surrogate (-0.012, worse than the ridge). The distillation result is the informative one: applied to genes the agent did *not* name, the learned rule's top-1000 precision is 0.037 against a base rate of 0.035. The agent's knowledge is not recoverable from this feature space, which is why no feature-side expansion of a short prior substitutes for asking the model for more.

C.8 Scaling the elicitation

Asking for $B \times R = 1280$ names in the same single call, as many as the loop will measure, is the only change; the encoding becomes a floored rank because a flat indicator degrades as the list grows, and the cold start takes the top 128 strictly in the model's stated order. The effect is conditional on the proposer's capability, in both directions (Fig. 9).

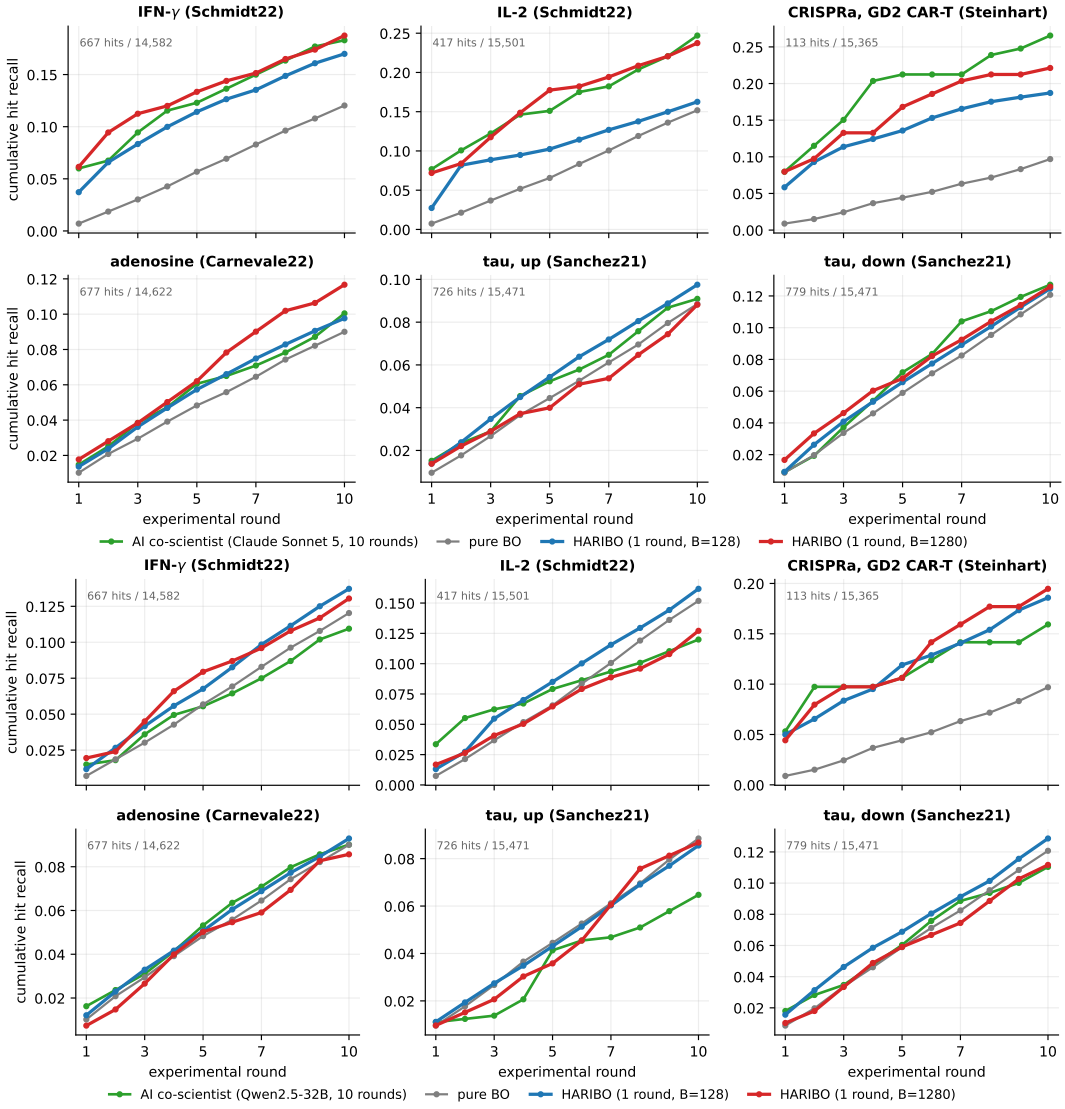


Figure 9: **A wider single call, per gene screen.** Claude Sonnet 5 (top) and Qwen2.5-32B (bottom). Sonnet fills the request with 1,122 in-panel genes and rises from 0.140 to 0.163 on 5/6 screens, overtaking the ten-round agent on IFNG and adenosine. Qwen2.5-32B *falls*, 0.133 to 0.123: asked for 1,280 names it emits 1,280 numbered entries but 90% are symbols mygene cannot resolve (`PTPN314...PTPN373`, `ATF7L1000...`), leaving 70 in-panel genes against 119 at the published width. Its `finish_reason` is `stop` with 9k tokens of headroom, so this is the limit of what the model knows, not of its context.

The wider ask also removes what the loop contributes. At B=1280 the prior on its own reaches 0.160 on Sonnet 5 against HARIBO's 0.163: a margin of +0.003, where at B=128 it is +0.024. The reason is arithmetic: a list as long as the measurement budget is exhausted by round ten whatever the acquisition does, and the prior arm scores the same whether its ranking is as elicited, reversed or shuffled. On Qwen2.5-32B the margin survives at either width (+0.035 and +0.031), because a fabricated list still has to be discounted. We therefore report the published width in the main text.