LLM-guided acquisition improves pathway-specific Perturb-seq design under experimental budgets

Anonymous Authors[1]

Abstract

We study LLM-guided acquisition for budgeted Perturb-seq design, where each perturbation is a costly wet-lab experiment and acquisition strategy is central to sample efficiency. We find that with modern perturbation-response predictors, common acquisition methods based on curated priors or uncertainty often fail to beat random selection. We therefore propose a two-stage strategy for experimental planning: candidates are shortlisted by ensemble-based epistemic uncertainty, then re-ranked by an LLM using pathway context, candidate annotations, and diversity criteria, with a concise rationale per selection. On public Perturb-seq benchmarks, our method achieves the best predictive performance on pathway genes for 5 of 5 evaluated pathways in K562 and 2 of 4 in RPE1; in blinded LLM-as-judge evaluation with three independent judges, our selections are preferred over random and BALD and tie with IterPert, indicating the edge over IterPert is predictive rather than in judged coherence. Ablations show the LLM relies on supplied biological context rather than gene-symbol memorisation, and performance is comparable across two reasoning-capable backbones. Together, these results suggest base LLMs already encode useful structure for Perturb-seq experimental planning, with clear headroom from retrieval, fine-tuning, and richer agentic loops.

1. Introduction

Genetic perturbation studies are a central tool for uncovering gene function, regulatory networks, and disease mechanisms (Datlinger et al., 2017; Gasperini et al., 2019). Modern single-cell CRISPR screening platforms, most notably Perturb-seq (Dixit et al., 2016), couple targeted genetic perturbations with single-cell RNA sequencing to measure transcriptomic responses at scale (Replogle et al., 2022; Adamson et al., 2016; Datlinger et al., 2017). In target discovery, the goal is to determine perturbations that move cells toward therapeutically relevant states. As these screens scale to thousands of perturbations and millions of cells, the challenge becomes not only predicting responses, but deciding which perturbations are worth testing next under tight experimental budgets.

We view this decision-making problem as budgeted active learning for perturbation selection. Each perturbation is a costly experimental intervention on a target gene, producing a post-intervention transcriptomic response relative to control cells. Because the number of perturbations that can be assayed is small relative to the size of the candidate space, acquisition decisions must be informative early in the loop, biologically relevant, and interpretable enough to support experimental scrutiny and iteration (Huang et al., 2023). This is especially important in lab-in-the-loop settings, where selection decisions are not merely optimisation steps, but part of an ongoing scientific process.

A natural expectation is that acquisition rules based on curated priors or uncertainty estimates should outperform naïve baselines in sequential Perturb-seq design (Huang et al., 2023). However, as we show later in Section 5.1, this advantage is not reliable with modern perturbation-response predictors. Performance is highly model dependent, random selection is often a strong baseline, and uncertainty-based acquisition can substantially underperform. This creates a practical tension: the acquisition strategy must work well under severe budget constraints, but it must also remain intelligible enough for biologists to interpret, contextualise, and when necessary, override (Mosqueira-Rey et al., 2023).

This motivates the use of large language models (LLMs) for experimental planning. LLMs are increasingly used as biological priors in genomics: gene-level text embeddings support functional prediction tasks (Chen & Zou, 2024b; Hedley et al., 2026), while perturbation-response models have begun to use LLM-derived representations to improve generalisation to unseen perturbations (Märtens et al., 2024; Chen & Zou, 2024a; Ramakrishnan et al., 2026; Wang et al.,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

2024). More recently, natural-language context has been used more directly for perturbation reasoning and design (Märtens et al., 2025; Phillips et al., 2026; Martell et al., 2026), a direction also reflected in recent community efforts and challenges (Genentech, 2026). These developments suggest that LLMs may offer not only useful priors over candidate perturbations, but also a more interpretable basis for acquisition.

We address this experimental planning problem with a two-stage LLM-guided acquisition strategy. We first estimate epistemic uncertainty with a deep ensemble to shortlist candidate perturbations, then pass these to a reasoning LLM that re-ranks by gene function, pathway context, and functional diversity, while producing a concise rationale for each selection. Across two cell lines (K562 and RPE1) from the Replogle et al. (2022) Perturb-seq dataset, our method is the best-performing acquisition strategy in the largest number of evaluated settings. Ablations show that both the uncertainty shortlist and the biological annotations supplied to the LLM contribute to performance, and a multi-judge evaluation using three independent LLMs supports the biological coherence of the selected gene sets.

Contributions. We make three main contributions.

1. We show that standard acquisition rules (uncertainty-based, kernel-prior) are unreliable under modern perturbation-response predictors: random selection is a surprisingly strong baseline, and these methods frequently underperform it.
2. We propose a two-stage acquisition strategy that shortlists candidates by ensemble epistemic uncertainty, then uses a reasoning LLM to select a batch by gene function, pathway context, and diversity, returning a mechanistic rationale for each choice.
3. Across two cell lines and five pathways, our method achieves the best performance in seven of the nine pathway configurations. Ablations show that both the uncertainty shortlist and biological annotations are load-bearing, and a multi-judge evaluation confirms the biological coherence of selected gene sets.

Broader impacts and dual-use considerations of our work are discussed in Appendix: Impact.

2. Related Work

LLMs for Biological Analysis and Discovery. LLM-based systems in biology have largely focused on automating analysis workflows through tool use, retrieval, and multi-agent orchestration. AutoBA (Zhou et al., 2024), CellAgent (Xiao et al., 2024), BioMaster (Su et al., 2025) and Biomni (Huang et al., 2025) exemplify this tool-augmented paradigm but are not specialised for budgeted perturbation selection. Close to our setting, PerTurboAgent (Hao et al., 2025) and BioDiscoveryAgent (Roohani et al., 2024b) use LLMs to propose perturbations in iterative discovery loops, motivating agentic experiment planning; however, they provide limited integration with uncertainty-aware acquisition and do not explicitly connect selections to expected improvements in a downstream predictor.

Perturbation Prediction Models. Perturbation response modelling has progressed from latent-variable generative approaches such as CPA (Lotfollahi et al., 2023) and sVAE+ (Lopez et al., 2023) to methods using structured priors, pretraining or richer conditioning. GEARS (Roohani et al., 2024a) uses gene-relationship graphs, while scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024) adapt foundation models to single-cell and perturbation settings. Recent benchmarking suggests that generalisation to unseen perturbations remains difficult and gains over simple baselines are not guaranteed (Ahlmann-Eltze et al., 2025). Recent models therefore incorporate additional biological context, including LLM-derived gene embeddings in LLMPert (Märtens et al., 2024), LLMHistPert (Ramakrishnan et al., 2026), and scLAMBDA (Wang et al., 2024), or population-level distribution matching in STATE (Adduri et al., 2025).

Active Learning for Genetic Perturbation Screens. Perturb-seq design is naturally a budgeted, batched active learning problem (Hacohen et al., 2022; Huang et al., 2023), where each round selects a small perturbation panel under experimental constraints. Standard uncertainty-based acquisition, often using deep ensembles, can be limited by noisy biological measurements and batch redundancy; related work addresses these issues through epistemic/aleatoric uncertainty separation Scherer et al. (2025), BatchBALD-style redundancy control (Kirsch et al., 2019), biological priors such as ITERPERT (Huang et al., 2023), and biology-informed Bayesian optimisation (Li et al., 2025). Empirically, acquisition gains can be predictor-dependent and modest relative to simple baselines (Mittal et al., 2019; Panagopoulos et al., 2025), motivating methods that combine sample efficiency with interpretable rationales for expert oversight.

3. Method

3.1. Problem Overview

We study budgeted batch active learning for Perturb-seq screen design. Given an offline dataset $\mathcal{D} = \{(p_i, \mathbf{y}_i)\}_{i=1}^{N}$ of perturbation–response pairs, we fix a held-out test set $\mathcal{D}_{\text{test}}$ and treat the remainder as an acquisition pool $\mathcal{D}_{\text{pool}}$.

Starting from a small initial training set $\mathcal{L}_0 \subset \mathcal{D}_{\text{pool}}$, the learner iteratively selects batches $\mathcal{S}_t$ of size b from the unacquired pool $\mathcal{U}_t = \mathcal{D}_{\text{pool}} \setminus \mathcal{L}_t$, retrains a predictor $f_{\theta_t} : \mathcal{P} \to \mathbb{R}^G$, and evaluates on $\mathcal{D}_{\text{test}}$. Our objective is to maximise cumulative performance over the acquisition trajectory. Depending on the metric over which we evaluate, this is equivalent to either minimising or maximising the area under the learning curve. This then rewards strategies that improve prediction early and sustain it throughout acquisition.

3.2. Perturbation Prediction Model

We use LLMPERT (Märtens et al., 2024; Ramakrishnan et al., 2026) as the perturbation response model. Each perturbation p is represented by a frozen embedding $\mathbf{e}_p \in \mathbb{R}^d$ formed by concatenating a text-based gene function embedding derived from NCBI gene descriptions (following GenePT (Chen & Zou, 2024b)) with a protein sequence embedding from ProtT5 (Elnaggar et al., 2022). LLMPERT predicts a perturbation-induced mean expression shift $\Delta_\theta(\mathbf{e}_p) \in \mathbb{R}^G$ relative to a fixed control baseline $\mathbf{y}^{\text{ctrl}} \in \mathbb{R}^G$:

$$\hat{\mathbf{y}}_\theta(p) = \mathbf{y}^{\text{ctrl}} + \Delta_\theta(\mathbf{e}_p), \tag{1}$$

where Δ_θ is an MLP trained on acquired perturbations by minimising mean-squared error to observed mean responses. Conditioning on LLM-derived gene representations improves generalisation to unseen perturbations relative to graph-based priors (Roohani et al., 2024a; Märtens et al., 2024; Ramakrishnan et al., 2026). We focus on mean prediction to isolate the effect of the acquisition policy; the framework extends naturally to distributional predictors (Ramakrishnan et al., 2026).

3.3. Epistemic Uncertainty Estimation from Deep Ensemble Prediction Models

To estimate epistemic uncertainty for shortlisting candidate perturbations, we use deep ensembles of the LLMPERT predictor. At each acquisition round t, we independently train an ensemble $\{f_{\theta_m}\}_{m=1}^M$ on the acquired set $\mathcal{L}_t$, using the same architecture and training protocol but different random initialisations $\theta_1, \ldots, \theta_M$. For each unacquired candidate perturbation $p \in \mathcal{U}_t$, ensemble member m predicts a mean expression response $\hat{\mathbf{y}}_{\theta_m}(p) \in \mathbb{R}^G$. We compute the ensemble predictive mean and the per-gene ensemble variance as

$$\bar{\mathbf{y}}(p) = \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{y}}_{\theta_m}(p); \tag{2}$$

$$\sigma^2(p) = \frac{1}{M} \sum_{m=1}^{M} (\hat{\mathbf{y}}_{\theta_m}(p) - \bar{\mathbf{y}}(p))^2 \in \mathbb{R}^G, \tag{3}$$

respectively. This variance captures disagreement between models trained on the same acquired data and therefore acts as a practical proxy for epistemic uncertainty, in contrast to aleatoric variability, which reflects irreducible response heterogeneity. We aggregate this vector into a scalar perturbation-level score by averaging across genes, $u(p) = \frac{1}{G} \sum_{g=1}^{G} \sigma_g^2(p)$, and use $u(p)$ to rank or shortlist candidates for acquisition.

3.4. LLM-guided Active Learning

We formulate perturbation selection as a closed-loop, LLM-guided active learning process that iteratively updates experimental design decisions using predictive uncertainty and accumulated experimental context. At acquisition round t, a perturbation-response predictor is trained on the acquired set $\mathcal{L}_t$, and a deep ensemble estimates epistemic uncertainty over the unacquired pool $\mathcal{U}_t$ (section 3.3). The top-k most uncertain candidates are then passed to an LLM together with biological annotations, pathway context, and the current experimental state. Based on these inputs, the LLM selects the next batch of perturbations, $\mathcal{S}_t \subset \mathcal{U}_t$, with $|\mathcal{S}_t| = b$, and provides mechanism-focused rationales linking the selection to the pathway objective.

The selected perturbations are added to the training set via $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{S}_t$, updating the predictor and uncertainty estimates in subsequent rounds.

LLM inputs. At round t, the LLM receives three types of inputs: candidate perturbation profiles, target pathway context, and the current experimental state.

Candidate perturbation profiles describe each shortlisted perturbation p using gene name, epistemic uncertainty information, Gene Ontology annotations, an LLM-generated biological summary, and inferred pathway relationships from protein interaction databases. Full annotation details are given in Section 4.4.

The pathway context contains the target pathway name, description, and known functional modules. The experimental state contains the acquired set $\mathcal{L}_t$ of perturbations profiled up to round t, allowing the LLM to avoid redundant selections and account for previously covered modules.

LLM-guided selection objective. The LLM selects a batch $\mathcal{S}_t$ of b perturbations to improve prediction of genes in the target pathway. The LLM balances three criteria: *module-level coverage*, *pathway relationships*, and *uncertainty–relevance trade-off* (Appendix B.1 for prompting details). Specifically, it instructs the LLM to cover a minimum number of functional modules; prioritise candidates with pathway connections such as direct members, upstream regulators, or downstream effectors; and prefer perturbations that are both uncertain and functionally rele-

vant.

Interpretation generation of selected perturbations. In addition to batch selection, the LLM generates a concise, gene-level rationale for each selected perturbation $p \in \mathcal{S}_t$ (Appendix C). The rationale links the gene to the target pathway using the available pathway context, candidate annotations, and acquired set $\mathcal{L}_t$, highlighting expected information gain through pathway membership, regulatory influence, downstream effects, or related cellular processes. This records the biological reasoning behind each selected gene and supports inspection of how the acquisition process explores pathway roles and functional modules across rounds.

We make explicit our use of LLMs throughout this work in Appendix: LLM Usage.

4. Experimental Details

4.1. Data

We evaluate on two Perturb-seq CRISPRi datasets from Replogle et al. (2022): K562 (chronic myeloid leukaemia; 1,075 perturbations) and RPE1 (retinal pigment epithelium; 1,516 perturbations). These cell lines differ in lineage, transcriptional programmes, and perturbation coverage, providing a test of cross-context robustness. Following standard GEARS-style preprocessing (Roohani et al., 2024a), we retain the top 5,000 highly variable genes (HVGs) in each dataset and use log-normalised expression counts.

4.2. Pathway Evaluation Datasets

We evaluate pathway-specific acquisition on five gene sets from the MSigDB Hallmark collection (Liberzon et al., 2015): Heme Metabolism, G2M Checkpoint, Ribosome, Apoptosis, and IFN-γ Response. These pathways were selected on two criteria: biological diversity and sufficient overlap with the dataset's 5,000 output genes (≥ 30 evaluable genes) to support reliable per-pathway evaluation. The selected pathways span distinct biological processes, from tightly co-regulated machinery (Ribosome) to broad regulatory programmes (G2M Checkpoint). For each pathway, we measure prediction accuracy only on the subset of pathway genes retained in the output gene set. In K562, this yields 115, 137, 50, 67, and 66 evaluable genes respectively. In RPE1, we evaluate four of the five pathways (74, 157, 84, and 75 genes); Ribosome is excluded for this cell line due to insufficient representation (29 genes).

4.3. Evaluation

We evaluate held-out perturbations using mean squared error (MSE) between predicted and observed expression deltas relative to the control mean, computed on pathway output genes only. We use pathway-gene MSE as the primary metric because the acquisition objective is to improve quantitative prediction of pathway-relevant responses, including both direction and magnitude of the perturbational effect. We plot learning curves of MSE as a function of the number of acquired perturbations and summarise sample efficiency by the area under the learning curve (AUC).

While other evaluation metrics have been proposed for perturbation prediction, MSE is the most natural choice here because it directly measures quantitative error in the predicted response and remains straightforward to interpret across acquisition settings. We report Pearson correlation in Appendix D as a complementary metric, since it captures agreement in the overall pattern of gene-level responses. By contrast, we do not adopt recent discrimination-based scores as a metric, since their interpretation has been found to depend more strongly on the choice of similarity measure and on effect scaling (Liu et al., 2026).

4.4. Candidate Annotations

Each candidate perturbation presented to the LLM includes:

Uncertainty signal. Rank and scalar epistemic uncertainty score $u(p)$ from the deep ensemble (Section 3.3). The top-$k = 200$ candidates by uncertainty are shortlisted for LLM re-ranking.

GO annotations. Gene Ontology terms associated with each gene, providing standardised functional descriptors.

Biological summary. A one-sentence functional description ($\leq$20 words) generated by Qwen3.5-35B-A3B for each of the 1,075 perturbation genes, cached and reused across all experiments.

Pathway relationship. In pathway-targeted mode, each candidate is labelled with its relationship to the target pathway, inferred from OmniPath (Türei et al., 2016) protein interaction data: *direct_member* (gene is in the pathway gene set), *upstream_regulator* (directed edge from candidate to a pathway gene), *downstream_target* (directed edge from a pathway gene to candidate), *interacting_partner* (undirected or weak interaction), or *unknown*. Relationships are queried in batches of 40 candidates across the OmniPath, PathwayExtra, KinaseExtra, and DoRothEA datasets.

4.5. Acquisition Strategies

We compare four strategies:

Random. Uniform random selection from $\mathcal{U}_t$.

BALD (Bayesian Active Learning by Disagreement). Select perturbations that maximise mutual information between predictions and model parameters, estimated via dis-

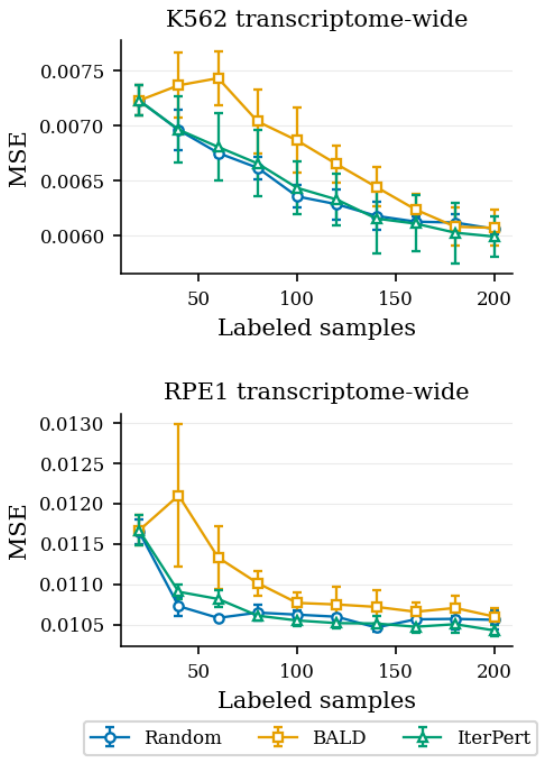

Figure 1. Transcriptome-wide MSE learning curves on K562 and RPE1, comparing random sampling, BALD, and ITERPERT. Points show mean ± s.e. over 3 random seeds.

agreement among ensemble members (Houlsby et al., 2011).

ITERPERT-style (kernel prior). Following Huang et al. (2023), candidates are scored by combining uncertainty with a diversity penalty based on biological similarity kernels. Specifically, each perturbation is scored by its epistemic uncertainty minus a weighted maximum similarity to the currently acquired set $\mathcal{L}_t$, where similarity is computed from gene embeddings.

LLM-guided (pathway-stratified). Our method. The LLM receives the top-200 uncertain candidates with annotations described above, the target pathway description and functional modules, and the current acquired set. It selects b perturbations balancing module coverage, biological relevance, and diversity, returning a structured rationale per selection. We use Qwen3.5 35B A3B with reasoning enabled, served locally via vLLM.

In ablation experiments, we additionally test: (i) LLM-guided with stripped annotations (gene symbols and uncertainty only, no GO terms, summaries, or pathway relationships), and (ii) a second LLM backbone (Claude Opus 4.7) to assess robustness to model choice.

4.6. Experimental Protocol

Each experiment follows a batched active learning protocol with batch size $b = 20$ and $T = 10$ acquisition rounds, for a total budget of 200 perturbations. We initialise with an empty training set ($|\mathcal{L}_0| = 0$): the first batch is selected by the acquisition function without any trained predictor. For BALD and ITERPERT, which require ensemble uncertainty estimates, the first round defaults to uniform random selection. For LLM-guided acquisition, the LLM selects the initial batch using only biological annotations, without uncertainty scores. This cold-start protocol tests whether the LLM can make useful selections before any experimental data is available. A fixed held-out test set containing 10% of perturbations is reserved for evaluation, consistent across all runs. Results are averaged over 3 random seeds.

For model architecture, training, ensemble, and embedding details, see Appendix A. LLM prompts are provided in Appendix B.1.

5. Results

5.1. Baseline acquisition methods do not beat random

On transcriptome-wide prediction, random sampling is competitive with or better than uncertainty-based (BALD) and prior-based (ITERPERT) acquisition (Figure 1, Table 3). On K562, BALD achieves an MSE AUC of 1.215 versus 1.161 for random, a 4.7% increase in accumulated error, while ITERPERT is statistically indistinguishable from random (1.162 vs 1.161). On RPE1, BALD underperforms random by 3.4%, and ITERPERT is within SE of random on both MSE (1.919 vs 1.918) and Pearson-Δ (112.40 vs 112.63). These results are consistent with prior observations that acquisition gains are strongly predictor-dependent (Mittal et al., 2019; Panagopoulos et al., 2025): when the base predictor generalises well via LLM-derived embeddings, uncertainty-based scoring provides little additional benefit, and sophisticated kernel-based diversity penalties produce, at best, marginal gains over random.

5.2. LLM-guided acquisition improves pathway prediction

LLM-guided pathway-stratified acquisition achieves the lowest MSE AUC on all 5 K562 pathways and 2 of 4 RPE1 pathways (Table 1, Figure 2). On K562, the agent improves on all other baselines across Heme Metabolism (2.399 vs 2.458), G2M Checkpoint (1.750 vs 1.802), IFN-γ Response (1.262 vs 1.270), Ribosome (4.386 vs 4.464), marginally on Apoptosis (1.5858 vs 1.5863). On RPE1, the agent wins on G2M Checkpoint (5.086 vs 5.184, $\Delta = 0.098$) and IFN-γ Response (2.401 vs 2.451, $\Delta = 0.050$), while ITERPERT is stronger on Heme Metabolism (1.561 vs 1.614, $\Delta = 0.053$) and Apoptosis (4.509 vs 4.806).

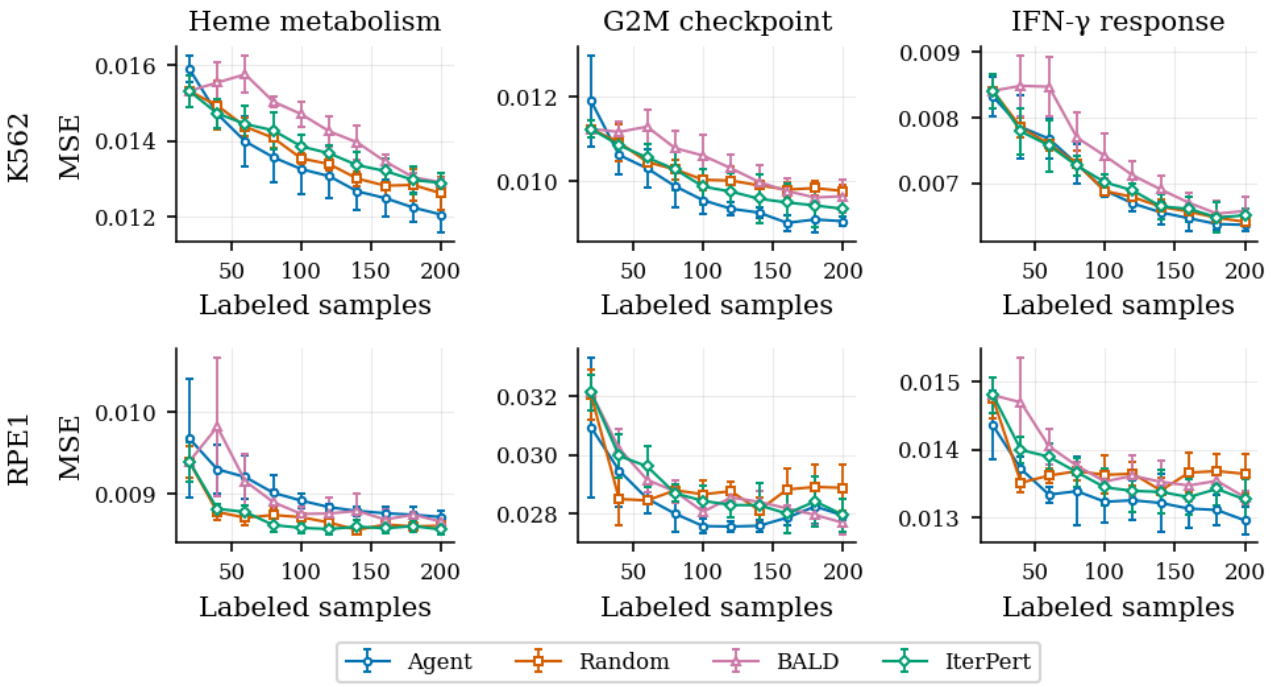

Figure 2. Pathway-specific MSE learning curves comparing LLM-guided acquisition to baselines. Top row: K562. Bottom row: RPE1. Points show mean ± s.e. over 3 random seeds.

5.3. Ablations

We ablate two components of the method (Table 5).

Biological annotations help where the agent outperforms random. Stripping GO terms, gene summaries, and pathway relationships from the prompt increases MSE AUC on several pathways where the full agent strongly beats random (K562 Heme $+0.054$, G2M $+0.049$, Ribosome $+0.166$; RPE1 G2M $+0.078$, IFN-γ $+0.030$) and decreases MSE where it does not (K562 Apoptosis -0.020; RPE1 Heme -0.028, Apoptosis -0.256). The annotations help precisely where the agent's selections are already improving over random, and hurt where they are not. This suggests that the LLM's biological context is not a uniform prior; its value depends on the configuration.

LLM backbone is interchangeable. Replacing Qwen3.5-35B-A3B with Claude Opus 4.7 yields comparable MSE AUC across four tested configurations (Table 6): K562 Heme ($2.399 \rightarrow 2.467$, $+2.8\%$), K562 G2M ($1.750 \rightarrow 1.818$, $+3.9\%$), RPE1 Heme ($1.614 \rightarrow 1.595$, -1.2%), and RPE1 G2M ($5.086 \rightarrow 5.137$, $+1.0\%$). Differences are within 4% in all cases, with Qwen winning $6/8$ metric-cells and Opus winning $2/8$. Both are reasoning-capable models; the result suggests the method is not dependent on a specific LLM backbone.

5.4. LLM-as-judge evaluation

To assess the biological coherence of acquired gene sets independent of predictive performance, we conduct blinded pairwise comparisons between pathway-stratified selections from the agent and size-matched sets from each baseline (Random, BALD, IterPert) across K562 and RPE1 pathways. Each comparison is judged independently by three LLMs (GPT-4o, Claude Sonnet 4.6, Llama 3.3) to avoid circularity from using the selector model as judge. Full prompt in Appendix B.2. By majority vote, agent selections are preferred over Random in 76.7% of 258 comparisons ($p < 0.001$, binomial test) and over BALD in 62.0% of 242 comparisons ($p < 0.001$). Against IterPert, agent and baseline selections are statistically indistinguishable (45.8% of 251 comparisons, $p = 0.20$), consistent with IterPert also incorporating biological priors into its acquisition. The agent's advantage over IterPert is therefore in predictive performance ($5/5$ K562 pathways, Section 5.2) rather than judged coherence. Per-pathway and per-judge breakdowns are reported in Appendix D.5.

6. Discussion

Random is a strong baseline because modern predictors are strong. Across both K562 and RPE1 transcriptome-wide, neither uncertainty-based (BALD) nor prior-based (ITERPERT) acquisition consistently improves over random

Pathway	Agent	Random	BALD	IterPert
K562				
Heme	**2.399 ± 0.055**	2.458 ± 0.024	2.596 ± 0.014	2.491 ± 0.045
G2M	**1.750 ± 0.040**	1.835 ± 0.024	1.879 ± 0.039	1.802 ± 0.051
IFN-γ	**1.262 ± 0.016**	1.270 ± 0.008	1.336 ± 0.025	1.274 ± 0.025
Apoptosis	**1.586 ± 0.018**	1.586 ± 0.011	1.650 ± 0.022	1.595 ± 0.033
Ribosome	**4.386 ± 0.190**	4.480 ± 0.049	4.980 ± 0.080	4.464 ± 0.137
RPE1				
Heme	1.614 ± 0.014	1.565 ± 0.004	1.611 ± 0.028	**1.561 ± 0.004**
G2M	**5.086 ± 0.057**	5.190 ± 0.038	5.184 ± 0.010	5.195 ± 0.073
IFN-γ	**2.401 ± 0.036**	2.461 ± 0.023	2.485 ± 0.021	2.451 ± 0.027
Apoptosis	4.806 ± 0.077	4.511 ± 0.008	4.782 ± 0.151	**4.509 ± 0.014**

Table 1. Pathway-specific MSE AUC (↓). **Bold** indicates best method per row; values are mean ± SE over 3 seeds. The agent achieves the lowest MSE AUC on 5/5 K562 pathways and 2/4 RPE1 pathways.

sampling (Section 5.1). This is consistent with a broader pattern in active learning: when the base predictor generalises well, uncertainty signal is washed out and acquisition gains collapse to noise (Mittal et al., 2019; Panagopoulos et al., 2025). In the foundation-model era, this reframes what acquisition is for. The question is no longer "which points has the predictor not seen?" — under a strong predictor, the answer is "it doesn't matter much" — but "which biologically meaningful subspace should we sample within?" The agent's gains on pathway-stratified prediction (Section 5.2) follow from constraining the hypothesis space, not from better-calibrated uncertainty.

Vanilla LLMs already encode useful biological structure. The agent uses a generic instruction-tuned LLM with no perturbation-specific adaptation: no fine-tuning on Perturb-seq data, no domain-adapted tokenizer, no curated retrieval index beyond the GO terms and pathway annotations supplied in-context. Despite this, it achieves the lowest MSE AUC on 5/5 K562 pathways and 2/4 RPE1 pathways, and is competitive with ITERPERT, a method designed specifically for perturbation acquisition with multi-modal biological priors, across the remainder. The result suggests base LLMs to some extent already encode enough pathway-level biological structure to guide perturbation selection, and that this structure transfers across reasoning-capable backbones (Qwen and Opus give comparable results, see Section 5.3).

Priors matter; flexibility distinguishes the agent from ITERPERT. Judge evaluation prefers agent selections over random (76.7%) and BALD (62.0%), but is indistinguishable from ITERPERT (45.8%, $p = 0.20$). Explicit biological reasoning produces selections judges recognise as more coherent than uncertainty- or distance-based heuristics; between two prior-aware methods, judged coherence saturates. The agent's edge over ITERPERT therefore lives in predictive performance (5/5 K562 pathways), not in looking more biological. We read this as flexibility: ITERPERT's priors are fixed by its kernel over multi-modal embeddings, while the agent reweights priors in-context per pathway and per round. Both encode biology; only one adapts to the task.

Scaling and adaptation. The current method represents a lower bound on what LLMs can contribute: the asymmetric ablation result (annotations help on every pathway where the agent already beats random and hurt on every pathway where it does not) suggests the LLM's biological knowledge is unevenly distributed across pathways rather than missing in aggregate. Two non-exclusive interpretations are consistent with the data: the LLM has stronger priors for well-characterised pathways (e.g., Heme Metabolism, G2M Checkpoint on K562) and weaker priors elsewhere; or some pathways are inherently noisier given current predictors, and biological context pushes the model toward false confidence. Either way, the result points to clear levers. Speculatively, we expect the agent–random gap to widen with stronger reasoning models: the capabilities the agent leans on — multi-step reasoning over annotations, holding several pathway hypotheses in working memory, calibrated comparison across candidates — all track general reasoning, which has improved sharply across recent generations. Retrieval-augmented prompting and biology-specialised or fine-tuned LLMs offer a complementary axis, targeting annotation-density failures directly rather than waiting for base-model progress.

Cell-type effects. The agent wins 5/5 on K562 but only 2/4 on RPE1. K562, a chronic myeloid leukemia line, is among the most extensively studied cell models in biology; RPE1 has a smaller literature footprint. A natural hunch is that the LLM's training distribution overweights K562 biology, and that the method's edge will narrow as it is applied to less-studied cell types. This is the regime where LLM-guided acquisition would be most useful — novel cell models have the least prior data to learn from — and is precisely where the current approach is weakest. We flag

this as the central limitation. Mitigating it almost certainly requires injecting cell-type-specific biological context that the base model lacks, which again points back to retrieval or adaptation.

Evaluating agents in biological discovery. Predictive metrics measure whether an acquisition improves a downstream model; they are silent on whether selections cohere as biology. The two axes are partially decoupled: our agent and IterPert are indistinguishable under blinded judging ($p = 20$) but differ on predictive performance across all 5 K562 pathways. Reporting only one risks endorsing methods that are biologically vacuous or experimentally inert. LLM-as-judge is an imperfect instrument — judges share training corpora and over-represent well-studied pathways — which we partially mitigate by ensembling three model families and requiring majority vote. Wet-lab validation remains ground truth; that said, LLM judges are a triage layer flagging whether selections a biologist would plausibly entertain, before committing experimental budget.

Limitations. Our results are reported on only two Perturb-seq datasets; acquisition behaviour and predictor performance may differ across more cell types, perturbation modalities, and protocols. Pathway-focused evaluation is limited to five Hallmark pathways, and generalisation to other pathway classes (e.g., signalling with feedback/cross-talk or highly redundant programmes) remains to be tested. We report only one acquisition budget and initial set size; sweeping both would reveal more about agent behaviour in different settings. While LLM-as-judge suggests coherent gene sets, the accompanying rationales have not been validated by experimental biologists and may contain fluent but incorrect claims. Finally, because the screen and related summaries are public, there is a residual risk of dataset-specific leakage influencing the LLM (or the judge).

Future work. Pairing acquisition with distributional predictors (e.g., histogram-based outputs (Ramakrishnan et al., 2026), generative models (Wang et al., 2024; Klein et al., 2025; Chi et al., 2026)) would separate aleatoric from epistemic uncertainty, giving the agent access to a signal BALD currently fails to exploit under generalising predictors. Second, biology-specialised or fine-tuned LLMs for re-ranking and rationale generation directly target the annotation-density failures we observe on RPE1. Finally, extending pathway conditioning to multi-objective settings, such as multiple pathways jointly, or pathway–global trade-offs, would test whether the agent's flexibility advantage over ITERPERT scales with task complexity.

7. Conclusions

We introduce an LLM-guided active learning method for budgeted Perturb-seq experimental planning, shortlisting perturbations by ensemble-based epistemic uncertainty and re-ranking with an LLM that balances module coverage with pathway relevance. Random sampling is a surprisingly strong baseline under modern predictors, and standard uncertainty- and prior-driven rules underperform; our method matches or beats top baselines while producing pathway-grounded rationales for each selection. In blinded multi-judge evaluation, our selections are preferred over random and BALD but tie with ITERPERT: the edge over ITERPERT is predictive, not aesthetic. We read this as flexibility: ITERPERT's priors are fixed by its kernel, while the LLM reweights priors in-context per pathway and per round, conditioning on the candidate shortlist, biological annotations, and prior selections. This points to a broader role for LLMs in biological discovery: as adaptive priors that can be coupled to strong predictors, expose their reasoning for scientific scrutiny, and shift focus as experimental goals evolve across rounds. The framing also bears on how such systems should be evaluated. Predictive metrics and judged biological coherence are partially decoupled, and reporting only one risks endorsing methods that look plausible but fail to inform downstream models, or vice versa. We see LLM judging as a cheap, reproducible triage signal that complements predictive evaluation, while wet-lab validation remains ground truth. The natural next step is prospective wet-lab verification, alongside scaling to additional cell lines, retrieval-augmented context, and fine-tuned biological reasoners.

References

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., et al. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv 2025.06.26.661135*, 2025. doi: 10.1101/2025.06.26.661135.

Ahlmann-Eltze, C., Huber, W., and Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02772-6.

URL https://www.nature.com/articles/s41592-025-02772-6.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Chen, Y. and Zou, J. Genepert: Leveraging genept embeddings for gene perturbation prediction. *bioRxiv*, 2024a. doi: 10.1101/2024.10.27.620513. URL https://www.biorxiv.org/content/early/2024/10/29/2024.10.27.620513.

Chen, Y. and Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv 2023.10.16.562533*, 2024b. ISSN 2692-8205. doi: 10.1101/2023.10.16.562533. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10614824/.

Chi, C., Xia, J., Huang, Y., Ouyang, Z., Tan, C., Liu, Y., Zhou, J., Yu, C., Yuan, L., Li, S., Zang, Z., and Li, S. Z. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rvpDHfoTd2.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. Pooled crispr screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1):377–390, 2019.

Genentech. BioReasoning Challenge 2026, 2026. URL https://genentech.github.io/BioReasoningChallenge/. MLGenX Workshop at ICLR 2026. Accessed: 2026-05-04.

Hacohen, G., Dekel, A., and Weinshall, D. Active learning on a budget: Opposite strategies suit high and low budgets, 2022. URL https://arxiv.org/abs/2202.02794.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://www.nature.com/articles/s41592-024-02305-7.

Hao, M., Lee, Y., Wang, H., Scalia, G., and Regev, A. Perturboagent: A self-planning agent for boosting sequential perturb-seq experiments. *bioRxiv*, 2025. doi: 10.1101/2025.05.25.656020. URL https://www.biorxiv.org/content/early/2025/05/27/2025.05.25.656020.

Hedley, J. G., Torr, P., and Märtens, K. GenePT revisited: Do better text embeddings make better gene embeddings? In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026. URL https://openreview.net/forum?id=fDLe47paJ1.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Huang, K., Lopez, R., Hütter, J.-C., Kudo, T., Rios, A., and Regev, A. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. *bioRxiv*, 2023. doi: 10.1101/2023.12.12.571389. URL https://www.biorxiv.org/content/early/2023/12/13/2023.12.12.571389.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., Yin, D., Marwaha, S., Carter, J. N., Zhou, X., Wheeler, M., Bernstein, J. A., Wang, M., He, P., Zhou, J., Snyder, M., Cong, L., Regev, A., and Leskovec, J. Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv 2025.05.30.656746*, 2025. doi: 10.1101/2025.05.30.656746. URL https://www.biorxiv.org/content/10.1101/2025.05.30.656746v1.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019. URL https://arxiv.org/abs/1906.08158.

Klein, D., Fleck, J. S., Bobrovskiy, D., Zimmermann, L., Becker, S., Palma, A., Dony, L., Tejada-Lapuerta, A., Huguet, G., Lin, H.-C., Azbukina, N., Sanchís-Calleja, F., Uscidda, T., Szalata, A., Gander, M., Regev, A., Treutlein, B., Camp, J. G., and Theis, F. J. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://www.biorxiv.org/content/early/2025/04/17/2025.04.11.648220.

Li, Y., Cui, T., Mansi, T., Prakash, M., and Liao, R. Biobo: Biology-informed bayesian optimization for perturbation design, 2025. URL https://arxiv.org/abs/2509.19988.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015.

Liu, Q., Zhang, Q., Du, J.-H., Zhao, S., and Wang, J. Effects of distance metrics and scaling on the perturbation discrimination score. In *ICLR 2026 Workshop on Generative AI in Genomics (Gen²): Barriers and Frontiers*. OpenReview, 2026. URL https://openreview.net/forum?id=p3Y9B90R5x.

Lopez, R., Tagasovska, N., Ra, S., Cho, K., Pritchard, J. K., and Regev, A. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. *arXiv 2211.03553*, 2023. doi: 10.48550/arXiv.2211.03553. URL http://arxiv.org/abs/2211.03553.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., Shendure, J., McFaline-Figueroa, J. L., Boyeau, P., Wolf, F. A., Yakubova, N., Günnemann, S., Trapnell, C., Lopez-Paz, D., and Theis, F. J. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. ISSN 1744-4292. doi: 10.15252/msb.202211517. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10258562/.

Martell, M. B., Stoisser, J., Phillips, L., Misra, A., Kitchen, R., Ferkinghoff-Borg, J., Yu, J., Torr, P., and Märtens, K. Causalpert: Grounding llm hypotheses in regulatory networks for gene perturbation prediction. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026.

Märtens, K., Donovan-Maiye, R., and Ferkinghoff-Borg, J. Enhancing generative perturbation models with LLM-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Märtens, K., Martell, M. B., Prada-Medina, C. A., and Donovan-Maiye, R. LangPert: LLM-Driven Contextual Synthesis for Unseen Perturbation Prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Mittal, S., Tatarchenko, M., Özgün Çiçek, and Brox, T. Parting with illusions about deep active learning, 2019. URL https://arxiv.org/abs/1912.05361.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

Panagopoulos, G., Lutzeyer, J. F., Ennadir, S., Vazirgiannis, M., and Pang, J. Efficient data selection for training genomic perturbation models, 2025. URL https://arxiv.org/abs/2503.14571.

Phillips, L., Martell, M. B., Misra, A., Stoisser, J., Prada-Medina, C., Donovan-Maiye, R., and Märtens, K. Synthpert: Enhancing llm biological reasoning via synthetic reasoning traces for cellular perturbation prediction. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026.

Ramakrishnan, K., Hedley, J. G., Qu, S., Dokania, P. K., Torr, P., Prada-Medina, C. A., Fauqueur, J., and Märtens, K. Beyond mean shifts: Predicting distributional responses to unseen genetic perturbations. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026. URL https://openreview.net/forum?id=TSrxbWcF7Q.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6): 927–935, 2024a. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6.

Roohani, Y., Lee, A., Huang, Q., Vora, J., Steinhart, Z., Huang, K., Marson, A., Liang, P., and Leskovec, J. BioDiscoveryAgent: An AI Agent for Designing

Genetic Perturbation Experiments. *arXiv 2405.17631*, 2024b. doi: 10.48550/arXiv.2405.17631. URL http://arxiv.org/abs/2405.17631.

Scherer, P., Kirsch, A., and Taylor-King, J. P. When three experiments are better than two: Avoiding intractable correlated aleatoric uncertainty by leveraging a novel bias–variance tradeoff, 2025. URL https://arxiv.org/abs/2509.04363.

Su, H., Long, W., and Zhang, Y. Biomaster: Multi-agent system for automated bioinformatics analysis workflow. *bioRxiv 2025.01.23.634608*, 2025. doi: 10.1101/2025.01.23.634608.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, 2016. doi: 10.1038/nmeth.4077.

Wang, G., Liu, T., Zhao, J., Cheng, Y., and Zhao, H. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024. doi: 10.1101/2024.12.04.626878. URL https://www.biorxiv.org/content/early/2024/12/08/2024.12.04.626878.

Xiao, Y., Liu, J., Zheng, Y., Xie, X., Hao, J., Li, M., Wang, R., Ni, F., Li, Y., Luo, J., et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv 2407.09811*, 2024.

Zhou, J., Zhang, B., Li, G., Chen, X., Li, H., Xu, X., Chen, S., He, W., Xu, C., Liu, L., et al. An ai agent for fully automated multi-omic analyses. *Advanced Science*, 11 (44):2407094, 2024.

Appendix

Impact

This work may have positive societal impact by improving the efficiency and interpretability of experimental design in functional genomics and target discovery, which could help accelerate biological discovery and, in the long term, therapeutic development. Potential negative impacts are limited but include over-reliance on imperfect acquisition systems, which could bias experimental resources toward misleading or incomplete biological hypotheses, especially if deployed without expert oversight. More broadly, as with other AI methods for biology, improved decision-making tools for perturbation selection could contribute to capabilities that may be misused in harmful biological contexts. In this work, however, the method is a prioritisation tool for controlled experimental settings, not a system for autonomous biological design or deployment, which limits direct risk.

LLM Usage

We acknowledge the use of large language models (LLMs) in two distinct roles in this work. First, LLMs are part of the proposed method itself. An LLM is used within the acquisition strategy to re-rank perturbation candidates after uncertainty-based shortlisting, using pathway context, candidate annotations, and diversity criteria, and to produce a concise rationale for each selected perturbation. In addition, LLMs are used in a separate blinded evaluation analysis, where three independent models act as judges of the biological coherence of selected gene sets. No LLM is used to generate the underlying perturbation-response training labels or the primary quantitative evaluation metrics; the main predictive results are based on downstream model performance on held-out perturbations. Second, LLMs were used in the preparation of the manuscript only for limited editorial assistance, including rewording and minor polishing of author-written text, and for light coding support such as formatting figures or helper scripts. All scientific ideas, method design, experiments, analysis, interpretation, and conclusions are entirely the work of the authors.

A. Implementation

Model & Training. We use LLMPERT as an MLP-based mean predictor with hidden dimensions $(1024, 512, 256)$. Each hidden layer applies Layer Normalization (Ba et al., 2016) followed by LeakyReLU (negative slope 0.01); the output layer is linear. The model takes concatenated gene embeddings as input and predicts expression deltas relative to the control mean. Models are trained with AdamW (Loshchilov & Hutter, 2017) using a learning rate of 10^{-3}, cosine annealing (minimum learning rate 10^{-5}), weight decay 10^{-3}, dropout 0.3, batch size 1024, and 200 epochs per round.

Ensemble. We use $M = 5$ ensemble members with independent random initialisations, and estimate epistemic uncertainty via ensemble disagreement.

Gene embeddings. Each perturbed gene is represented by a 2,560-dimensional embedding formed by concatenating 1,024-dimensional ProtT5 protein sequence embeddings (Elnaggar et al., 2022) and 1536-dimensional GenePT text-based embeddings of functional descriptions (Chen & Zou, 2024b). Embeddings are precomputed and kept fixed during training.

LLM configuration & prompts. The LLM-guided acquisition and LLM-as-judge panel use Qwen3.5 35B A3B and Llama 3.3 70B run locally with VLLM, Claude Opus 4.7 and Claude Sonnet 4.6 via Anthropic's API, and ChatGPT 4o via OpenAI's API, with default temperature for selection, and reasoning on for LLM-guided acquisition. Prompt templates for selection are provided in Appendix B.1.

Computational resources. All experiments were run on an internal compute cluster. The main perturbation-acquisition experiments used NVIDIA Quadro RTX 6000 GPUs (24GB VRAM), with a full 10-round acquisition experiment taking approximately 80 minutes, depending on the acquisition function. The total compute for the reported main experiments is approximately 64 GPU-hours, comprising 3 seeds × 2 datasets × 3 baseline acquisition functions, together with 3 seeds × 2 datasets × 5 pathway-specific runs. LLM inference was performed on the same internal cluster using a single NVIDIA H200 GPU (∼150 GB VRAM).

Software. Models are implemented in PyTorch; data handling uses Scanpy. Code will be released upon publication.

B. Prompts

B.1. Pathway Acquisition LLM Prompt

Prompt template used by the agent to select the next batch of perturbations. Placeholders in braces are instantiated per (pathway, round, seed).

LLM-based next round perturbation selection

```
You are designing the next round of a CRISPR screen to characterize
the {target_pathway} pathway.

OBJECTIVE
---------
Select exactly {n_samples} genes that will most improve prediction of
genes in {target_pathway}.

Some {target_pathway} genes are held out for evaluation (you don't know
which ones). Your selections should provide information that generalizes
across the pathway.

CURRENT STATE
-------------
Training round: {round_num} of {total_rounds}
Already profiled: {acquired_str if acquired_str else "[none yet]"}

TARGET PATHWAY BIOLOGY
----------------------
{pathway_description}

Key functional modules within {target_pathway}:
{pathway_modules}

CANDIDATES
----------
Format: gene | uncertainty_rank | uncertainty_score | GO_terms |
        relation_to_target_pathway | bio_summary

{candidates_block_with_relation}

SELECTION STRATEGY
------------------
1. MODULE COVERAGE
   - {target_pathway} has distinct functional modules (listed above)
   - Sampling one gene per module gives broader coverage than
     clustering in one module
   - Unknown which modules contain held-out genes, so cover all

2. PATHWAY RELATIONSHIPS
   - Direct pathway members: inform other members via shared function
   - Upstream regulators: reveal how pathway responds to input signals
   - Downstream effectors: reveal pathway output / consequences

3. UNCERTAINTY x RELEVANCE
   - High uncertainty alone may be pathway-irrelevant noise
   - Prefer: high uncertainty AND functionally connected to
     {target_pathway}

AVOID
-----
- Clustering selections in single module (held-out genes may be elsewhere)
- Ignoring pathway relevance for raw uncertainty
- Redundant selections (genes in same complex behave similarly)
```

```
CONSTRAINTS
-----------
- Exactly {n_samples} genes from candidate list
- Cover at least {min_modules} distinct functional modules

OUTPUT FORMAT
-------------
Return ONLY this JSON:
{
  "selected": ["GENE1", "GENE2", ...],
  "module_coverage": {
    "module_name": ["GENE1", ...],
    ...
  },
  "rationale": "2-3 sentences on module coverage strategy and
               expected information gain."
}

No markdown fences. No text outside JSON.
```

B.2. LLM-as-judge Prompt

System message

You are an expert computational biologist specializing in gene expression analysis, perturbation biology, and machine learning for biological systems.
Your task is to evaluate gene selection choices for training a machine learning model that predicts gene expression changes after gene perturbations (knockouts/knockdowns).
The model is being trained specifically to predict changes in genes belonging to a particular biological pathway. Your evaluation should consider:

1. **Biological relevance**: Are the selected genes likely to provide informative training signal for the target pathway?
2. **Mechanistic insight**: Do the genes have known regulatory relationships with the pathway?
3. **Diversity**: Do the genes cover different aspects/mechanisms relevant to the pathway?
4. **Training utility**: Will perturbations of these genes help the model learn generalizable patterns?

You will be shown two sets of genes and asked to judge which set would be more useful for training.

User message (RPE1 / heme_metabolism / iter 1 / seed 42)

Target Pathway: heme_metabolism
Sample of pathway genes (20 of 200): AHSP, CA1, NUDT4, RNF123, HBB, BMP2K, MFHAS1, PGLS, MOCOS, ACP5, TOP1, CCDC28A, LMO2, SLC6A8, CDR2, ATG4A, MYL4, FOXJ2, HBD, SLC22A4
Context: A machine learning model is being trained to predict gene expression changes specifically for genes in the heme_metabolism pathway. At training iteration 1, the acquisition function selected genes to add to the training set.
Set A (Acquisition Function Selection), 20 genes: HSPA9, ZBTB17, PSMD12, VPS35, PRPF39, ATP6V0B, SMC1A, MCM6, RFC5, NARS1, RPL13A, RPS3A, PSMC3, CCT7, CLTC, ZNF787, SF3A2, AFG3L2, RPLP0, COPS4
Set B (random Selection), 20 genes: EIF1AX, WDR43, CAMLG, DAXX, TBP, SRSF3, ARPC2, MRPL38, MED20, GMIP, LAMTOR4, UBA52, PPWD1, NCAPG, GEMIN8, KIF18B, RTCB, HAUS7, NEDD8, BUD31
Please evaluate which set of genes would be more useful for training a model to predict expression changes in the heme_metabolism pathway, then respond with a JSON object containing `preferred_set` ("A" or "B"), `confidence` ("high"/"medium"/"low"), `reasoning`, `set_a_strengths`, `set_b_strengths`, and `key_differentiators`.

Figure 3. LLM-as-judge prompt template. The system message is fixed; the user message is instantiated per (dataset, pathway, iteration, seed) tuple. **Set A** is the agent's acquired batch and **Set B** is the baseline's acquired batch (here, random; analogous prompts are used for BALD and IterPert). Example shown for RPE1 / heme metabolism / iteration 1 / seed 42; both gene lists are size 20 (matched batch size).

C. Sample LLM Acquisition Reasoning

For heme metabolism pathway, seed 42, budget 20, iteration 5:

Gene	Reasoning
PCNA	DNA replication/repair patterns; critical for iron homeostasis in heme biosynthesis; key regulator not previously selected
RPS3	Ribosomal processing insight; essential for heme-related protein synthesis; different ribosomal protein from those profiled
NUP62	Nuclear pore complex patterns; key for gene regulation linked to heme pathways; first nuclear transport member
RPL7	Ribosomal gene regulation; critical for translation affecting heme metabolism; differs by ribosomal position
PSMA7	Proteasome activity patterns; influences degradation of heme synthesis regulators; unique component not profiled
WDR82	Transcriptional regulation; impacts heme metabolism gene expression; presents transcriptional complex diversity
RPL19	Ribosomal assembly insights; crucial for heme enzyme synthesis efficiency; non-redundant variance
SRRT	RNA processing pathways; impacts RNA stability for heme gene expression; adds distinct stability data
INTS5	Mediator complex and transcription patterns; vital for upstream transcription influences; unique mediator gene
COPB2	Vesicular transport insights; affects iron transport and indirectly heme biosynthesis; functional diversity
PSMD11	Proteasome assembly insights; influences protein regulation affecting heme genes; distinct assembly role
EXOSC5	Exosome component insight; role in RNA processing for heme gene regulation; RNA processing diversity
XPO1	Nuclear export processes; essential for transport of heme pathway transcription factors; distinct export pattern
RPS18	Ribosomal RNA processing patterns; impacts translation of heme metabolism proteins; novel ribosomal coverage
RPL36	Translation machinery patterns; affects synthesis rates of heme proteins; unique ribosome protein profiling
DNAJC17	Chaperone-mediated folding; crucial for heme biosynthesis protein stability; only selected chaperone
ERH	RNA splicing patterns; relevant for properly spliced heme transcripts; distinct splice-related gene
PPIE	Peptidyl-prolyl isomerase activity; impacts heme protein folding; unique folding insights
CNOT3	mRNA decay patterns; essential for heme gene expression stability; first mRNA decay member selected
PSMC2	ATPase activity in proteasome; regulation of heme protein degradation; distinct ATPase mechanism

Table 2. LLM gene selections with pathway-stratified reasoning (Heme Metabolism, Init=20, Step 1).

D. Additional results

D.1. Transcriptome-wide baselines

	K562		RPE1	
Method	MSE↓	PearsonΔ↑	MSE↓	PearsonΔ↑
Random	**1.161 ± 0.006**	**75.59 ± 0.43**	**1.918 ± 0.005**	**112.63 ± 0.22**
BALD	1.215 ± 0.016	69.18 ± 0.23	1.984 ± 0.026	110.75 ± 0.57
IterPert	1.162 ± 0.027	72.66 ± 0.83	1.919 ± 0.006	112.40 ± 0.28

Table 3. Transcriptome-wide AUC for baseline acquisition strategies. **Bold** indicates best per column. Random is the strongest baseline on both cell lines, though within SE of ITERPERT on RPE1. All runs use 3 seeds; values are mean ± SE.

D.2. Pathway-specific full results

Dataset	**Pathway**	**Method**	MSE↓	PearsonΔ↑
K562	Heme	Agent	**2.399 ± 0.055**	**68.69 ± 2.33**
		Random	2.458 ± 0.024	68.67 ± 1.11
		BALD	2.596 ± 0.014	60.53 ± 0.51
		IterPert	2.491 ± 0.045	63.05 ± 1.37
K562	G2M	Agent	**1.750 ± 0.040**	**84.50 ± 0.82**
		Random	1.835 ± 0.024	84.37 ± 0.26
		BALD	1.879 ± 0.039	79.95 ± 0.98
		IterPert	1.802 ± 0.051	82.65 ± 0.86
K562	IFN-γ	Agent	**1.262 ± 0.016**	82.30 ± 0.36
		Random	1.270 ± 0.008	**82.74 ± 1.33**
		BALD	1.336 ± 0.025	76.91 ± 0.77
		IterPert	1.274 ± 0.025	78.45 ± 0.38
K562	Apoptosis	Agent	**1.586 ± 0.018**	86.97 ± 1.65
		Random	1.586 ± 0.011	**88.17 ± 0.43**
		BALD	1.650 ± 0.022	82.70 ± 0.32
		IterPert	1.595 ± 0.033	84.68 ± 0.85
K562	Ribosome	Agent	**4.386 ± 0.190**	**70.37 ± 3.21**
		Random	4.480 ± 0.049	69.29 ± 2.68
		BALD	4.980 ± 0.080	57.09 ± 2.21
		IterPert	4.464 ± 0.137	66.32 ± 2.20
RPE1	Heme	Agent	1.614 ± 0.014	96.22 ± 0.38
		Random	1.565 ± 0.004	97.76 ± 0.10
		BALD	1.611 ± 0.028	96.02 ± 0.90
		IterPert	**1.561 ± 0.004**	**98.05 ± 0.21**
RPE1	G2M	Agent	**5.086 ± 0.057**	135.91 ± 0.20
		Random	5.190 ± 0.038	**136.60 ± 0.06**
		BALD	5.184 ± 0.010	135.87 ± 0.22
		IterPert	5.195 ± 0.073	136.41 ± 0.27
RPE1	IFN-γ	Agent	**2.401 ± 0.036**	102.83 ± 0.71
		Random	2.461 ± 0.023	**104.03 ± 0.29**
		BALD	2.485 ± 0.021	102.29 ± 0.42
		IterPert	2.451 ± 0.027	103.76 ± 0.69
RPE1	Apoptosis	Agent	4.806 ± 0.077	108.51 ± 0.48
		Random	4.511 ± 0.008	**112.21 ± 0.36**
		BALD	4.782 ± 0.151	109.87 ± 1.31
		IterPert	**4.509 ± 0.014**	111.85 ± 0.36

Table 4. Pathway-specific MSE AUC (↓) and Pearson delta AUC (↑) across acquisition strategies. **Bold** indicates best method per (dataset, pathway, metric); values are mean ± SE over 3 seeds.

D.3. Biological annotation ablation

		Full		Stripped	
Dataset	**Pathway**	MSE↓	PearsonΔ↑	MSE↓	PearsonΔ↑
K562	Heme	**2.399 ± 0.055**	**68.69 ± 2.33**	2.453 ± 0.023	66.73 ± 1.63
K562	G2M	**1.750 ± 0.040**	**84.50 ± 0.82**	1.799 ± 0.022	81.85 ± 0.90
K562	IFN-γ	1.262 ± 0.016	**82.30 ± 0.36**	**1.253 ± 0.004**	80.98 ± 0.56
K562	Apoptosis	1.586 ± 0.018	**86.97 ± 1.65**	**1.566 ± 0.000**	86.73 ± 0.83
K562	Ribosome	**4.386 ± 0.190**	**70.37 ± 3.21**	4.552 ± 0.055	68.21 ± 1.76
RPE1	Heme	1.614 ± 0.014	96.22 ± 0.38	**1.585 ± 0.032**	**96.84 ± 0.56**
RPE1	G2M	**5.086 ± 0.057**	135.91 ± 0.20	5.164 ± 0.066	**135.94 ± 0.15**
RPE1	IFN-γ	**2.401 ± 0.036**	102.83 ± 0.71	2.431 ± 0.059	**103.64 ± 0.73**
RPE1	Apoptosis	4.806 ± 0.077	108.51 ± 0.48	**4.550 ± 0.070**	**111.27 ± 1.01**

Table 5. Biological annotation ablation. **Full**: agent with GO terms, gene summaries, and pathway relationships. **Stripped**: gene symbols and uncertainty scores only. **Bold** indicates better per (row, metric); values are mean ± SE over 3 seeds.

D.4. LLM backbone comparison

		Qwen-3.5		Claude Opus 4.7	
Dataset	**Pathway**	MSE↓	PearsonΔ↑	MSE↓	PearsonΔ↑
K562	Heme	**2.399 ± 0.055**	**68.69 ± 2.33**	2.467 ± 0.048	65.71 ± 1.96
K562	G2M	**1.750 ± 0.040**	**84.50 ± 0.82**	1.818 ± 0.040	80.69 ± 2.52
RPE1	Heme	1.614 ± 0.014	**96.22 ± 0.38**	**1.595 ± 0.017**	95.37 ± 1.17
RPE1	G2M	**5.086 ± 0.057**	135.91 ± 0.20	5.137 ± 0.122	**136.15 ± 0.09**

Table 6. LLM backbone comparison. Replacing Qwen3.5 35B A3B with Claude Opus 4.7 yields comparable performance across four configurations. **Bold** indicates better per (row, metric); values are mean ± SE over 3 seeds.

D.5. LLM-as-judge per-pathway breakdown

All comparisons are agent (pathway-stratified) vs baseline, size-matched, on K562 and RPE1 pathways. Sample sizes vary because RPE1 Ribosome was filtered (intersection with predicted output too small to report) and per-pathway prompt counts depend on overlap between acquired sets. Majority is the per-prompt majority across the three judges. Errors are binomial standard errors of the win rate, $\mathrm{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$, with n the number of decided votes per cell.

			% preferring agent selections		
Pathway	n	**Majority**	**GPT-4o**	**Sonnet**	**Llama**
Apoptosis	57	66.7 ± 6.2	75.4 ± 5.7	68.4 ± 6.2	47.4 ± 6.6
G2M	57	80.7 ± 5.2	93.0 ± 3.4	57.9 ± 6.5	80.7 ± 5.2
Heme	57	77.2 ± 5.6	94.7 ± 3.0	63.2 ± 6.4	71.9 ± 6.0
IFN-γ	57	84.2 ± 4.8	87.7 ± 4.3	63.2 ± 6.4	78.9 ± 5.4
Ribosome	30	73.3 ± 8.1	80.0 ± 7.3	53.3 ± 9.1	70.0 ± 8.4
All	258	**76.7 ± 2.6**	86.8 ± 2.1	62.0 ± 3.0	69.8 ± 2.9

Table 7. Agent vs Random. Per-pathway and per-judge preference rates (± SE).

			% preferring agent selections		
Pathway	n	**Majority**	**GPT-4o**	**Sonnet**	**Llama**
Apoptosis	57	38.6 ± 6.4	45.6 ± 6.6	43.9 ± 6.6	28.1 ± 6.0
G2M	57	59.6 ± 6.5	71.9 ± 6.0	49.1 ± 6.6	43.9 ± 6.6
Heme	50	36.0 ± 6.8	40.0 ± 6.9	46.0 ± 7.0	14.0 ± 4.9
IFN-γ	57	38.6 ± 6.4	40.4 ± 6.5	43.9 ± 6.6	31.6 ± 6.2
Ribosome	30	63.3 ± 8.8	70.0 ± 8.4	53.3 ± 9.1	56.7 ± 9.0
All	251	$\mathbf{45.8 \pm 3.1}$	52.2 ± 3.2	46.6 ± 3.1	33.1 ± 3.0

Table 8. Agent vs IterPert. Per-pathway and per-judge preference rates ($\pm$ SE). Majority preference is statistically indistinguishable from chance ($p = 0.20$, binomial test).

			% preferring agent selections		
Pathway	n	**Majority**	**GPT-4o**	**Sonnet**	**Llama**
Apoptosis	57	61.4 ± 6.4	63.2 ± 6.4	73.7 ± 5.8	33.3 ± 6.2
G2M	57	73.7 ± 5.8	77.2 ± 5.6	66.7 ± 6.2	61.4 ± 6.4
Heme	41	56.1 ± 7.8	70.7 ± 7.1	63.4 ± 7.5	29.3 ± 7.1
IFN-γ	57	56.1 ± 6.6	54.4 ± 6.6	71.9 ± 6.0	31.6 ± 6.2
Ribosome	30	60.0 ± 8.9	60.0 ± 8.9	60.0 ± 8.9	63.3 ± 8.8
All	242	$\mathbf{62.0 \pm 3.1}$	65.3 ± 3.1	68.2 ± 3.0	42.6 ± 3.2

Table 9. Agent vs BALD. Per-pathway and per-judge preference rates ($\pm$ SE). Majority preference for the agent is significant ($p < 0.001$, binomial test).