CAUSALPERT:GROUNDING LLM HYPOTHESES IN REGULATORY NETWORKS FOR GENE PERTURBATION PREDICTION

Marc Boubnovski Martell[1]*, Josefa Lia Stoisser[1], Lawrence Phillips[1], Aditya Misra[1], Robert Kitchen[1], Jesper Ferkinghoff-Borg[1], Jialin Yu[2], Philip Torr[2], Kaspar Märtens[1]

[1]Novo Nordisk, [2]University of Oxford

ABSTRACT

Predicting transcriptional responses to unseen genetic perturbations is essential for understanding gene regulation and prioritizing large-scale perturbation experiments. Existing approaches either rely on static, potentially incomplete knowledge graphs, or prompt language models for functionally similar genes, retrieving associations shaped by symmetric co-occurrence in scientific text rather than directed regulatory logic. We introduce CausalPert, a lightweight framework that encourages LLM agents to generate directed regulatory hypotheses rather than relying solely on functional similarity. Multiple agents independently propose candidate regulators with associated confidence scores; these are aggregated through a consensus mechanism that filters spurious associations, producing weighted neighborhoods for downstream prediction.

We evaluate CausalPert on Perturb-seq benchmarks across four human cell lines. For perturbation prediction in low-data regimes ($N = 50$ observed perturbations), CausalPert improves Pearson correlation by up to 10.5% over similarity-based baselines. For experimental design, CausalPert-selected anchor genes outperform standard network centrality heuristics by up to 46% in well-characterized cell lines.

1 INTRODUCTION

Predicting transcriptomic responses to genetic perturbations remains a fundamental challenge in computational biology, with applications in understanding gene regulation and identifying therapeutic targets (Replogle et al., 2022; Dixit et al., 2016). While high-throughput screens like Perturb-seq have enabled genome-scale profiling, typical experiments remain small to medium-sized, covering only a fraction of the 20,000 possible single-gene perturbations. The ability to generalize from a small number of observed perturbations to unseen targets (few-shot prediction) has thus emerged as a critical frontier for accelerating biological discovery.

Current methods attempt to bridge this gap by incorporating prior biological knowledge. Graph Neural Network (GNN) approaches, such as GEARS (Roohani et al., 2024) and TxPert (Wenkel et al., 2025), leverage structured Gene Ontology (GO) graphs to propagate perturbation signals. Language Model approaches like LangPert (Märtens et al., 2025) aggregate contextual information from unstructured literature. While recent Foundation Models such as GRNFormer (Qiu et al., 2025) seek to integrate explicit GRN priors into RNA token embeddings, all existing methods share a critical limitation: they rely on static, potentially incomplete knowledge graphs or noisy literature embeddings. When regulatory interactions are missing from curated databases or buried in ambiguous text, these models fail to enforce directional regulatory constraints, leading to poor generalization on understudied genes.

However, whether LLMs are used to generate gene embeddings (Chen & Zou, 2024) or prompted directly for functional similarity, the resulting associations reflect symmetric co-occurrence in scientific text rather than directed regulatory relationships. If genes A and B are frequently co-mentioned

[*]Contact: mbvk@novonordisk.com

in the literature, both approaches will treat them as similar, without distinguishing whether the directed edge is $A \to B$, $B \to A$, or neither Martell et al. (2025). This conflation of co-occurrence with regulatory directionality is a distinction central to causal inference (Pearl, 2009). As a result, predictions risk mirroring co-citation patterns rather than the directional logic of gene regulatory networks.

To overcome this, we introduce CausalPert, a framework that transforms unseen perturbation prediction from similarity-based prompting into a mechanistic reasoning problem. Rather than asking LLMs to identify functionally similar genes as in LangPert (Märtens et al., 2025), CausalPert prompts multiple agents to independently generate directed regulatory hypotheses, which are then aggregated through a consensus mechanism. The framework is algorithmically lightweight, operating entirely at inference time without requiring architectural changes or retraining, and its primary contribution is conceptual: introducing a mechanistic inductive bias that is complementary to model architecture and training scale.

This approach addresses a distinct failure mode from graph-based methods: when curated databases are incomplete but scientific literature is comprehensive, mechanistic reasoning over text provides an alternative to graph-based propagation. While we also explore topological augmentations that benefit smaller language models (e.g., Gemini 3 Flash; Appendix A), we find that for frontier models (e.g., Gemini 3 Pro) the lightweight consensus framework alone is sufficient.

We evaluate CAUSALPERT on two complementary tasks using Perturb-seq data across four human cell lines (Replogle et al., 2022; **?**). First, for few-shot perturbation prediction, we demonstrate that directing LLM reasoning toward regulatory relationships improves generalization by a relative **10.5%** over similarity-based baselines in low-data regimes ($N = 50$). Second, for experimental design, we use CAUSALPERT to autonomously identify an initial set of 50 perturbations to map a regulatory landscape. Here, CAUSALPERT-selected anchors outperform standard network centrality heuristics by up to **46%** in K562, suggesting that LLM-guided target selection can effectively complement structural approaches for experimental prioritization.

2 BACKGROUND

2.1 PRIOR KNOWLEDGE FOR PERTURBATION PREDICTION

The dominant paradigm for perturbation prediction relies on propagating signals through structured knowledge graphs. GNN-based approaches such as GEARS (Roohani et al., 2024) and TxPert (Wenkel et al., 2025) leverage Gene Ontology (GO) (Ashburner et al., 2000) and protein interaction networks (Szklarczyk et al., 2023) to share information between perturbations (Battaglia et al., 2018; Kipf, 2016). While effective, these methods operate under a closed-world assumption: if a regulatory relationship is absent from the graph, the model cannot leverage it. They also treat the interactome as static, failing to capture context-specific rewiring across cell types (Ideker & Krogan, 2012).

An alternative line of work uses LLMs to incorporate biological knowledge without relying on fixed graph structure Phillips et al.; Stoisser et al. (2025). GenePT (Chen & Zou, 2024) and GP+LLM (Märtens et al., 2024) derive gene representations from LLM embeddings, while scGPT (Cui et al., 2024) learns representations from tokenized observational single-cell expression profiles. LangPert (Märtens et al., 2025) prompts LLMs directly to identify functionally similar genes for nearest-neighbor prediction. However, as discussed in Section 1, these approaches reflect symmetric co-occurrence in scientific text rather than directed regulatory relationships: they do not distinguish whether $A \to B$, $B \to A$, or neither.

CausalPert addresses this by treating LLM outputs as noisy hypotheses rather than factual assertions. Rather than relying on a single retrieval, multiple agents independently generate directed regulatory hypotheses, which are filtered through consensus and validated against topological constraints. This decouples co-occurrence from regulatory directionality while retaining the generative flexibility of LLMs.

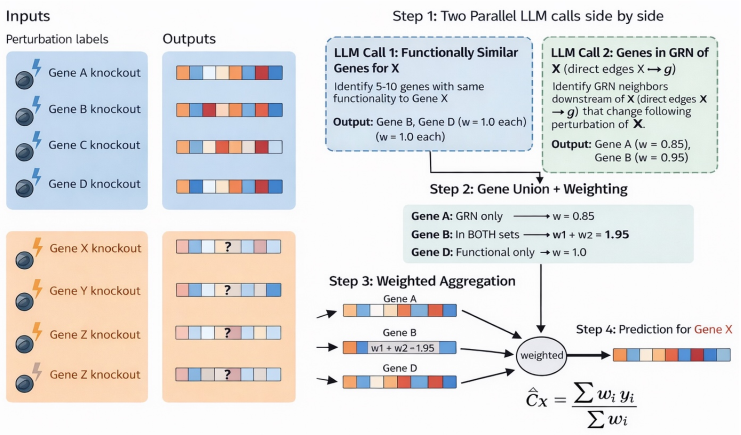
Inputs
Perturbation labels
Outputs
Gene A knockout
Gene B knockout
Gene C knockout
Gene D knockout
Gene X knockout
Gene Y knockout
Gene Z knockout
Gene Z knockout
?
Step 1: Two Parallel LLM calls side by side
LLM Call 1: Functionally Similar Genes for X
Identify 5-10 genes with same functionality to Gene X
Output: Gene B, Gene D (w = 1.0 each) (w = 1.0 each)
LLM Call 2: Genes in GRN of X (direct edges X ⟶ g)
Identify GRN neighbors downstream of X (direct edges X → g) that change following perturbation of X.
Output: Gene A (w = 0.85), Gene B (w = 0.95)
Step 2: Gene Union + Weighting
Gene A: GRN only ⟶ w = 0.85
Gene B: In BOTH sets ⟶ w1 + w2 = 1.95
Gene D: Functional only → w = 1.0
Step 3: Weighted Aggregation
Gene A
Gene B
w1 + w2 = 1.95
Gene D
weighted
Step 4: Prediction for Gene X
$\hat{C}_X = \frac{\sum w_i y_i}{\sum w_i}$

Figure 1: **CausalPert Framework.** Problem setup for predicting gene expression vectors of unseen perturbations. *CausalPert* pipeline: LLMs identify functional similarity and GRN regulators to weight and aggregate training data for few-shot prediction ($\hat{y}_X$).

2.2 Experimental Design for Perturbation Screens

Beyond prediction, selecting which perturbations to perform first is itself a critical design problem. Traditional screens prioritize targets using heuristics such as degree centrality in protein interaction networks (Albert et al., 2000; Barabási & Albert, 1999), but these structural measures do not capture context-specific regulatory logic. Bayesian and active learning approaches (Snoek et al., 2012; Settles, 2009) offer principled alternatives but require an initial dataset to calibrate uncertainty, making them ineffective in cold-start settings (Schein et al., 2002). We show that LLM-generated regulatory hypotheses can also guide experimental prioritization, identifying informative anchor perturbations without requiring any prior screening data.

3 Methods

We propose CausalPert (Figure 1), a framework that transforms the LLM from a passive retrieval engine into a structured hypothesis generator with directional regulatory constraints. We evaluate this framework on two distinct tasks: (1) **Few-Shot Perturbation Prediction**, identifying the effect of a specific perturbation; and (2) **Active Discovery**, identifying which perturbations yield the most information about the system.

3.1 Few-Shot Perturbation Prediction

Given a set of N observed perturbation profiles $\{(\mathbf{y}_j, g_j)\}_{j=1}^{N}$ and a query gene g_t not in the training set, the goal is to predict the transcriptomic response $\hat{\mathbf{y}}_t$. We describe how CausalPert constructs this prediction in two steps.

Step 1: Dual Hypothesis Generation. For each query gene g_t, we prompt the LLM to generate two distinct candidate sets from the training pool:

- **Semantic Set ($\mathcal{N}_{sem}$):** Genes identified by functional similarity to g_t (shared pathways, co-regulation, similar knockout phenotypes). This corresponds to the standard LangPert retrieval strategy (Märtens et al., 2025).
- **Causal Set ($\mathcal{N}_{causal}$):** Genes identified as potential directed regulators of g_t (e.g., upstream transcription factors or direct regulatory targets). For each causal hypothesis r_i, the agent also provides a confidence score $c_i \in [0, 1]$ reflecting its certainty in the directed edge $r_i \to g_t$.

The causal set is the key addition: by explicitly prompting for directed relationships, we encourage the model to reason about regulatory logic rather than symmetric similarity. To reduce sensitivity to

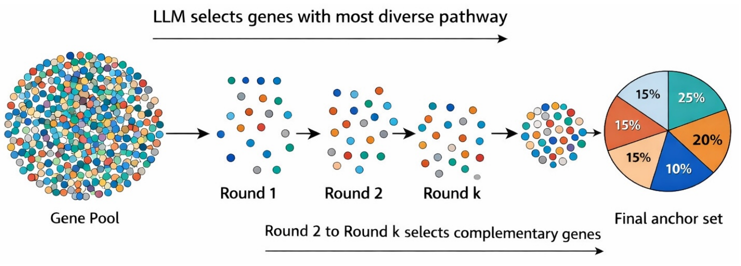
LLM selects genes with most diverse pathway
Gene Pool
Round 1
Round 2
Round k
Final anchor set
15%
25%
15%
20%
15%
10%
Round 2 to Round k selects complementary genes

Figure 2: **Active experiment design.** Iterative hub selection to maximize biological pathway diversity in the final anchor set.

any single LLM call, we execute K independent reasoning chains (we use $K = 3$) and retain genes that appear across multiple chains.

Step 2: Consensus Aggregation. The predicted response $\hat{\mathbf{y}}_t \in \mathbb{R}^d$ is computed as a weighted average of the perturbation profiles of the retrieved genes:

$$\hat{\mathbf{y}}_t = \frac{1}{Z} \sum_{j \in \mathcal{N}_{sem} \cup \mathcal{N}_{causal}} w_j \cdot \mathbf{y}_j \tag{1}$$

where $Z = \sum_j w_j$. We evaluate two weighting schemes:

- **Binary Consensus ($w_j = 1$):** All retrieved genes contribute equally, with consensus arising from the frequency of retrieval across independent chains.
- **Confidence-Weighted:** Semantic neighbors receive unit weight ($w_j = 1$), while causal hypotheses are scaled by the agent's reported confidence c_j ($w_j = c_j$).

This aggregation is intentionally simple: it treats all regulatory relationships additively, without distinguishing activation from repression. Our goal is to test whether encouraging directed regulatory reasoning alone provides a useful inductive bias in sparse regimes, while acknowledging that sign-aware aggregation remains future work.

3.2 Active Discovery of Regulatory Anchors

The second task addresses experimental design: given a budget of k perturbations, which genes should be perturbed first to maximally inform predictions across the full regulatory landscape? Traditional approaches select targets by degree centrality in PPI networks (Barabási & Albert, 1999), but these structural hubs do not necessarily capture context-specific regulatory logic (He & Geng, 2008).

We use CausalPert to identify an informative anchor set $\mathcal{S}$ (Figure 2). Analogous to Task I, we prompt LLM agents to identify candidate master regulators for the specific cellular context, aggregating both semantic and causal signals across independent reasoning chains. To encourage diversity across regulatory pathways, selection proceeds iteratively: in each round, genes are selected to complement the pathways already represented in $\mathcal{S}$. The resulting anchor set is then mapped to the physical PPI interactome to evaluate its information propagation potential.

3.3 Validation Protocol for Active Discovery

To rigorously quantify the information content of the anchors $\mathcal{S}$ selected in Task II, we employ a *Surrogate Evaluation Strategy*. Crucially, we decouple *anchor selection* from *inference* to eliminate the confounding variable of the LLM's generative variance. We do *not* use the CausalPert engine for this step. Instead, we fix the downstream predictor to be a deterministic *Heat Kernel Interpolator* over the physical PPI network (Kondor & Lafferty, 2002).

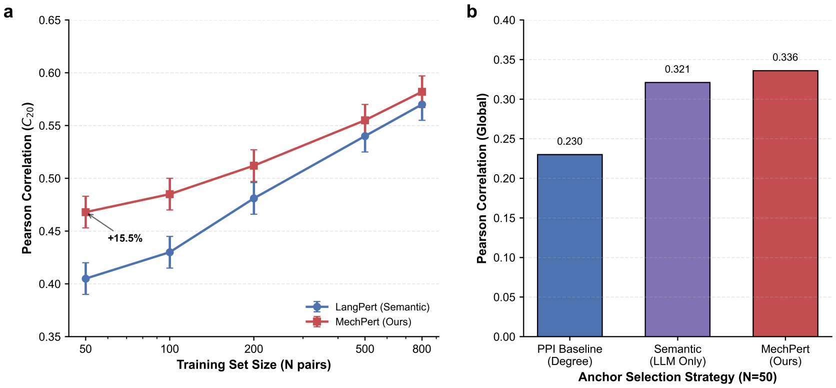

Figure 3: **CAUSALPERT improves low-data generalization and experimental design.** (a) In Jurkat T-cells, our Causal-Consensus model (red) beats the semantic baseline (blue) by +15.5% at N=50, showing causal priors help when data are scarce. (b) In K562, choosing N=50 experimental anchors via geometrically adjudicated consensus gives a +46% gain over PPI-degree heuristics, indicating agentic reasoning finds effective regulatory hubs.

For any unseen gene v, the predicted perturbation effect $\hat{y}_v$ is computed as the weighted average of the anchor effects:

$$\hat{y}_v = \frac{\sum_{s \in \mathcal{S}} K_\beta(s, v) \cdot y_s}{\sum_{s \in \mathcal{S}} K_\beta(s, v)} \tag{2}$$

where $K_\beta(s, v) = [\exp(-\beta \mathbf{L})]_{sv}$ is the heat kernel corresponding to the graph Laplacian $\mathbf{L}$.

By utilizing a fixed, topology-only propagator, we ensure that any performance gain in the CAUSALPERT arm is strictly attributable to the *Topological Superiority* of the selected anchors—proving they are better positioned within the functional manifold to capture global variance than standard structural hubs. This evaluation does not measure true experimental utility, but isolates the effect of anchor selection under a fixed inference mechanism.

4 RESULTS

4.1 STRUCTURAL PRIORS FOR RESOURCE-EFFICIENT MODELS

We first characterized the limits of LLM reasoning within transcriptomic space. We found that the raw predictions of efficient models, such as Gemini 3 Flash and GPT 5 mini, were frequently challenged by biological signal noise. To bridge this gap, we evaluated three adaptations that anchor semantic retrieval in physical manifolds: (1) **3+2 Strategy** (Euclidean PPI expansion), (2) **Topological Harmonizer** (multiscale graph smoothing), and (3) **Spectral Filtering** (Hadamard-gated Poincaré density).

These methods act as structural scaffolding, substituting internal model parameters with Protein-Protein Interaction (PPI) and hyperbolic manifold priors. The impact of this scaffolding is most pronounced in low-data regimes: for the *Jurkat* lineage at $N = 50$, the **Topological Harmonizer** improves the Pearson correlation of the model by **+24.1%** ($0.360 \rightarrow 0.447$). Strikingly, this augmentation allows the efficient model to surpass the few-shot performance floor of the $100\times$ larger *Pro* backbone (0.406). While critical for smaller models, these augmentations provide limited benefit to the largest LLMs, which already possess robust internal representations of these signaling topologies.

4.2 CAUSAL PRIORS IMPROVE SAMPLE EFFICIENCY IN FEW-SHOT REGIMES

To test the hypothesis that causal consensus acts as a robust inductive bias when training data is scarce, we evaluated CAUSALPERT across four divergent cell lines (K562, RPE1, Jurkat, HepG2). We conducted a Scaling Law Analysis by training on subsampled datasets ranging from $N = 50$ to

$N = 800$ perturbation pairs (Figure 3a). We compare three strategies: (1) **LangPert** (standard semantic retrieval), (2) **Binary Consensus** (unweighted voting), and (3) **CAUSALPERT** (confidence-weighted).

As shown in Table 1, CAUSALPERT consistently outperforms the semantic baseline, particularly in the critical cold-start regime ($N = 50$). Aggregated across all cell lines, our confidence-weighted model achieves a +10.4% relative improvement in Pearson correlation ($C_{20} : 0.528 \to \mathbf{0.583}$) compared to the LangPert baseline.

Crucially, the performance gap is context-dependent. In cell lines with sparse prior knowledge (e.g., Jurkat T-cells), the benefit of causal reasoning is magnified: CAUSALPERT improves correlation by +15.5% ($0.405 \to \mathbf{0.468}$), validating our hypothesis that generative priors effectively substitute for missing experimental data. While Binary Consensus performs comparably to Confidence-Weighted aggregation in some regimes, the weighted approach provides strictly lower variance (Standard Error ± 0.031 vs ± 0.035), suggesting that uncertainty calibration stabilizes the consensus mechanism.

Table 1: **Few-Shot Generalization (C_{20} Correlation).** CAUSALPERT (confidence-weighted) outperforms LangPert and binary consensus on pooled C20 correlation across four cell lines (K562, RPE1, Jurkat, HepG2; Mean ± SEM). Best values per (N) are highlighted, and CAUSALPERT shows a notable +10.4% gain in the low-data setting ($N = 50$).

Training Size (N)	LangPert (Sem)	Binary Consensus	CAUSALPERT (Conf)	Rel. Improv.
$N = 50$	0.528 ± 0.032	0.553 ± 0.031	$\mathbf{0.583} \pm 0.031$	**+10.4%**
$N = 100$	0.529 ± 0.030	0.540 ± 0.029	$\mathbf{0.551} \pm 0.029$	**+4.1%**
$N = 200$	0.558 ± 0.029	0.569 ± 0.029	$\mathbf{0.581} \pm 0.029$	**+4.1%**
$N = 500$	0.583 ± 0.028	$\mathbf{0.606} \pm 0.028$	0.589 ± 0.028	+3.9%
$N = 800$	0.605 ± 0.027	$\mathbf{0.634} \pm 0.026$	0.620 ± 0.026	+4.8%

Agentic Selection Outperforms Structural Heuristics. To evaluate the utility of agentic reasoning for experimental design, we compared four anchor selection strategies for $N = 50$ in Table 2: (1) Random Uniform (Maximum Entropy), (2) PPI Degree Centrality (Structural Baseline), (3) Semantic Importance (LLM Only), and (4) CAUSALPERT (Consensus).

Our results reveal a distinct regime of advantage. In well-characterized lineages where the structural prior is dense (K562, RPE1), CAUSALPERT identifies anchors that maximize information flow relative to structural centrality. Specifically, in K562, our consensus strategy yields a +46% improvement over the standard PPI Degree baseline (0.336 vs. 0.230), confirming that semantic hubs are topologically superior to naive structural bottlenecks for signal propagation.

However, we observe that performance is strictly conditioned on the fidelity of the biophysical prior. In contexts with sparser interactomes (HepG2) or highly dynamic signaling (Jurkat), the synergy between text and graph breaks down (0.181 vs. 0.247), and simple Random Uniform sampling—which maximizes global coverage rather than local mechanism—remains the most robust strategy (0.261). This validates the hypothesis that agentic reasoning acts as a structural amplifier: it can enhance the utility of well-characterized biological networks but cannot invent topology where physical ground truth is missing.

Table 2: **Active Experimental Design Performance.** Pearson correlation of global genome prediction using $N = 50$ anchors selected by different strategies. CAUSALPERT provides the best targeted strategy in well-characterized lineages (K562, RPE1), beating standard PPI centrality.

Cell Line	Random (Reference)	PPI Baseline	Semantic (LLM)	CAUSALPERT (Ours)
K562	0.404	0.230	0.321	**0.536**
RPE1	0.580	0.592	0.612	**0.614**
Jurkat	**0.367**	0.315	0.316	0.260
HepG2	**0.261**	0.247	0.199	0.181

Taken together, these results establish a dual advantage for consensus-driven reasoning. First, it enables robust few-shot generalization by imposing severe inductive biases when training data is scarce ($N = 50$). Second, it transforms the experimental design process itself, allowing agents to autonomously identify the most informative causal interventions before a single experiment is run. This moves toward integrating inference and experimental prioritization, suggesting that LLM agents may assist in navigating combinatorial biological design spaces.

4.3 Ablation Study: Contribution of Consensus Components

To rigorously isolate the contribution of consensus aggregation and confidence calibration, we performed an ablation study in the low-data regime ($N = 50$) on the challenging Jurkat cell line (Table 3).

Comparing the breakdown of performance gains reveals a clear hierarchy of inductive biases. The transition from Single-Agent Retrieval (Baseline) to Multi-Agent Binary Consensus yields the most significant lift ($+10.6\%$), suggesting that the simple frequency of hypothesis generation across independent reasoning chains is a powerful filter for stochastic hallucination. However, incorporating the agent's explicit uncertainty estimate via Confidence Weighting provides a critical second-order refinement, boosting performance by an additional $+4.9\%$ (Total $+15.5\%$). This confirms that while ensemble consistency establishes robustness, calibration is necessary to achieve state-of-the-art accuracy in sparse data regimes.

Table 3: **Ablation Study of Consensus Components ($N = 50$).** Jurkat shows most performance gain comes from the consensus aggregation (frequency), with confidence weighting adding further refinement.

Model Variant	Aggregation Logic	Mean C_{20} (Jurkat)	% Gain
LangPert (Baseline)	None (Single Retrieval)	0.405	–
Binary Consensus	Unweighted Voting ($w_i = 1$)	0.448	$+10.6\%$
CausalPert (Full)	**Confidence Weighted** ($w_i = c_i$)	$\mathbf{0.468}$	$\mathbf{+15.5\%}$

4.4 Case Study: Disambiguating Stress Pathways via Mechanistic Reasoning

To demonstrate the interpretability of our approach, we examined the prediction of **NVL** (Nuclear VCP-Like), a critical AAA-ATPase involved in ribosome biogenesis, within the HepG2 liver carcinoma line. This gene presents a challenging test case due to its high lexical similarity to its paralog VCP, which operates in a fundamentally different cellular compartment.

Representation Failure due to Lexical Ambiguity. The semantic baseline, relying on distributional co-occurrence statistics, exhibited **lexical conflation**: it failed to distinguish *NVL* from its paralog *VCP* (Valosin-Containing Protein) due to shared nomenclature and high textual co-occurrence in the training corpus. Consequently, it retrieved functional neighbors associated with the *Ubiquitin-Proteasome System* and *ER-Associated Degradation (ERAD)*. This led to the prediction of an **Unfolded Protein Response (UPR)** signature (characterized by *ATF4, CHOP, XBP1*), which is biologically incorrect for an NVL knockout. This representational failure resulted in a strong negative correlation with the ground truth ($C_{20} = -0.67$), indicating that the model predicted the *opposite* of the true transcriptomic response.

Mechanistic Disambiguation via Causal Consensus. In contrast, our CausalPert agent explicitly adjudicated the regulatory mechanism by leveraging **subcellular context**. The consensus reasoning correctly identified that while NVL shares a protein family with VCP, it specifically functions in the **nucleolus** for *pre-60S ribosomal subunit release*, rather than in the ER for protein quality control.

1. **Mechanism Inferred:** The agent correctly identified that disrupting NVL causes **Nucleolar Stress**, leading to the release of ribosomal proteins (e.g., *RPL5, RPL11*) into the nucleoplasm, where they sequester MDM2.
2. **Outcome Predicted:** Instead of a UPR signature, the agent predicted a stabilization of **p53** and the massive upregulation of its downstream targets (*CDKN1A/p21, MDM2, BTG2*).

By correctly mapping the perturbation to the **p53-dependent Ribosomal Stress Pathway** rather than the **Proteotoxic Stress Pathway**, the Consensus model achieved a strong positive correlation with the ground truth ($C_{20} = 0.74$). This represents a $\Delta C_{20} = 1.41$ swing from failure to success, attributable solely to the difference in reasoning architecture. This case demonstrates that the ability to disentangle distinct stress mechanisms illustrates that disentangling stress mechanisms may improve perturbation prediction.

5 Discussion

Our results reveal a fundamental limitation of treating large language models solely as retrieval engines: semantic proximity in the literature is not isomorphic to functional causality in the cell. While standard LLMs excel at capturing distributional co-occurrence, they inherently struggle to represent the directed, asymmetric nature of gene regulatory networks.

The consistent performance gains observed in the low-data regime ($N \leq 100$) demonstrate that multi-agent consensus provides a robust inductive bias when training data is scarce. This positions CausalPert as a highly effective **'few-shot Cold-Start' engine**. While data-rich regimes ($N \geq 800$) allow supervised models to bridge the performance gap, our approach provides the critical prior knowledge necessary for the initial discovery phase of understudied genes where experimental data is non-existent.

Crucially, hypothesis frequency may act as a heuristic proxy for regulatory plausibility. By aggregating votes from multiple LLM agents, we implicitly filter stochastic hallucinations in favor of reproducible causal signals. This provides a natural explanation for the method's dominance in sparse regimes: the consensus mechanism appears to provide an inductive bias that reduces reliance on purely associative neighbors that lack consistent mechanistic support across diverse reasoning paths.

6 Limitations

CausalPert is evaluated on four human cancer cell lines using single-gene CRISPRi perturbations; generalization to primary cells, gene knockouts, and combinatorial or temporal perturbations remains untested. The method is most effective when partial regulatory structure is available but incomplete. In poorly characterized lineages such as HepG2 and Jurkat, LLM-guided hypothesis generation can underperform random sampling, likely reflecting publication and annotation biases in pretraining corpora. Additionally, since the evaluation benchmark Replogle et al. (2022) is a high-profile publication likely present in the model's training data, future work should evaluate on perturbations from post-training-cutoff publications to rule out memorization effects.

Additionally, the validation of CausalPert has been conducted primarily in controlled cell line models with specific genetic perturbations. Its effectiveness and robustness in complex, heterogeneous, and multi-cellular biological systems remain untested. Extensive validation in primary tissues, disease models, or in vivo settings is necessary to confirm its practical utility in real-world biological and therapeutic contexts.

The framework relies on consensus across multiple LLM agents rather than calibrated confidence estimates. We observe that majority voting performs comparably to or better than confidence-weighted aggregation, indicating that model-reported confidence is not reliably aligned with biological plausibility in this setting.

Finally, causal language in this work refers to directed regulatory hypotheses rather than formal causal identification. CausalPert does not estimate interventional effect sizes or recover a uniquely identified causal graph; its outputs should be interpreted as mechanistically informed hypotheses for experimental prioritization. Future work could incorporate causal effect estimation methods to strengthen these claims.

ACKNOWLEDGMENTS

We acknowledge the use of LLMs to assist with text editing, distinctness, grammatical refinement and diagram editing. All scientific claims and the final text were reviewed and verified by the authors.

REFERENCES

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 2000.

Michael Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 2000.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 1999.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pp. 2023–10, 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.

Yang-Bo He and Zhi Geng. Active learning of causal networks. *JMLR*, 2008.

Trey Ideker and Nevan J Krogan. Differential network biology. *Molecular Systems Biology*, 2012.

TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, volume 2002, pp. 315–322, 2002.

Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Daniel Keitley, Maria Dermit, and Julien Fauqueur. A scalable LLM framework for therapeutic biomarker discovery: Grounding q/a generation in knowledge graphs and literature. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025. URL `https://openreview.net/forum?id=ClewUE4sUK`.

Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturbation models with llm-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Kaspar Märtens, Marc Boubnovski Martell, Cesar A Prada-Medina, and Rory Donovan-Maiye. Langpert: Llm-driven contextual synthesis for unseen perturbation prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Rory Donovan-Maiye, and Kaspar Märtens. Synthpert: Enhancing biological reasoning in llms via synthetic reasoning traces for cellular perturbation prediction.

Mufan Qiu, Xinyu Hu, Fengwei Zhan, Sukwon Yun, Jie Peng, Ruichen Zhang, Bhavya Kailkhura, Jiekun Yang, and Tianlong Chen. Grnformer: A biologically-guided framework for integrating gene regulatory networks into rna foundation models. *arXiv preprint arXiv:2503.01682*, 2025.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Andrew I Schein et al. Methods and metrics for cold-start recommendations. In *SIGIR*, 2002.

Burr Settles. Active learning literature survey. *University of Wisconsin-Madison Technical Report*, 2009.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *NeurIPS*, 2012.

Josefa Lia Stoisser, Lawrence Phillips, Aditya Misra, Tom A Lamb, Philip Torr, Marc Boubnovski Martell, Julien Fauqueur, and Kaspar MÃĪrtens. Towards label-free biological reasoning synthetic dataset creation via uncertainty filtering. *arXiv preprint arXiv:2510.05871*, 2025.

Damian Szklarczyk et al. The string database in 2023. *Nucleic Acids Research*, 2023.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Cian Eastwood, Shawn T Whitfield, Ihab Bendidi, Craig Russell, Liam Hodgson, Yassir El Mesbahi, et al. Txpert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*, 2025.

A Architectural Variants and Adaptations

To evaluate the scaling of different inductive priors, we compare the baseline model against three distinct architectural adaptations. These adaptations were developed primarily for smaller LLM models (Gemini 3 Flash, gpt 5 mini) to augment their ability to retrieve and process sparse biological signals. Preliminary ablations showed that larger LLMs did not benefit significantly from these specific topological expansion strategies due to their superior internal reasoning capabilities; consequently, these methods were not applied to the larger model benchmarks.

A.1 Geometric Data-Augmented Retrieval (The 3+2 Strategy)

While the standard LangPert baseline relies on a Large Language Model (LLM) to retrieve a flexible set of 5, the 3+2 strategy introduces a geometric prior to stabilize retrieval for smaller models. This approach leverages a hybrid reasoning-manifold pipeline consisting of the following steps:

1. **Expert Seed Retrieval:** The model first queries the LLM to identify exactly three "Expert Seeds" from the training pool. These genes are selected based on the highest perceived functional similarity (e.g., shared biological pathways and co-regulation) to the target gene g_t.
2. **Euclidean Manifold Projection:** The identified seeds are projected into a 50-dimensional Euclidean embedding space ($\mathbb{R}^{50}$). These embeddings are pre-trained using Node2Vec on the STRING Protein-Protein Interaction (PPI) interactome, ensuring that proximity in the latent space corresponds to structural and functional relatedness in the proteome.
3. **Neighborhood Completion:** We calculate the Euclidean centroid C of the three expert seeds. To ground the prediction in the local topological density of the manifold, the model identifies the two nearest neighbors ("geometric neighbors") to this centroid from the training pool:

$$C = \frac{1}{3} \sum_{i \in \text{Seeds}} \mathbf{e}_i \tag{A.3}$$

where $\mathbf{e}_i$ represents the embedding vector of the i-th seed.

4. **Robust Aggregation:** The final predicted perturbation profile $\hat{y}_t$ is computed as a simple average of the perturbation vectors from the ensemble of five genes (3 semantic seeds plus 2 geometric neighbors):

$$\hat{y}_t = \frac{1}{5}\left(\sum_{i \in \text{Seeds}} y_i + \sum_{j \in \text{neighbors}} y_j\right) \tag{A.4}$$

By delegating the neighborhood completion to the Euclidean PPI manifold, the 3+2 strategy ensures that the functional context is anchored to the physical interactome, effectively "rescuing" the prediction when the LLM's primary retrieval is sparse or uncertain.

A.1.1 Topological Manifold Harmonizer

Building upon the discrete retrieval of the *LangPert* baseline and the Euclidean-based *Geometric Manifold Retrieval* (the 3+2 strategy), the **Topological Manifold Harmonizer** performs a continuous signal broadcast across the physical interactome. While standard *LangPert* relies exclusively on a small set of agent-selected semantic seeds, and the 3+2 strategy provides a discrete expansion in flat Euclidean space, the Harmonizer variant utilizes graph-theoretic diffusion and hyperbolic geometry to aggregate a broader neighborhood of the top 50 most reachable genes. This process consists of two primary stages:

1. **Topological Signal Propagation:** Using the networkx Python package, the model performs a Personalized PageRank (PPR) traversal on the STRING PPI interactome (G_{PPI}). In this implementation, the three LLM-selected seeds are treated as starting points for a diffusion process. We initialize a personalization vector $\mathbf{p}$ with uniform weight across the seeds and compute the steady-state reachability distribution using a damping factor $\alpha = 0.85$:

$$\mathbf{pr} = \alpha \mathbf{A}\mathbf{pr} + (1 - \alpha)\mathbf{p} \tag{A.5}$$

   where $\mathbf{A}$ represents the transition matrix of the interactome. This traversal identifies the top 50 genes most topologically relevant to the original expert seeds.

2. **Hyperbolic Density Weighting:** To refine the expanded neighborhood, we utilize hierarchical embeddings in a 100-dimensional Poincaré ball, implemented via the gensim package's `PoincareModel`. For each identified neighbor j, we calculate its hyperbolic distance $d_{\mathbb{H}}$ to the Einstein midpoint C of the original semantic seeds:

$$w_j = \exp\left(-\frac{d_{\mathbb{H}}(\mathbf{z}_j, C)^2}{2\sigma^2}\right) \tag{A.6}$$

   where $\mathbf{z}_j$ is the coordinate vector on the Poincaré manifold. To maintain high specificity, the bandwidth σ is set dynamically based on the 20th percentile of local distances.

Relationship to Previous Strategies: Unlike the *LangPert* baseline which assumes independent retrieval, or the *3+2 strategy* which uses flat Euclidean geometry to "rescue" sparse sets with a few extra genes, the **Topological Manifold Harmonizer** utilizes gensim's hyperbolic coordinates to model the latent hierarchy of biological systems. By transitioning from discrete neighbor selection to continuous manifold smoothing via networkx, this approach allows the model to identify relevant functional neighbors that may be semantically distant to an LLM but are structurally essential to the target perturbation.

A.1.2 Spectral Manifold Harmonizer

The **Spectral Manifold Harmonizer** represents our most rigorous adaptation, designed to isolate the highest-fidelity functional signals within the proteome. While the *Topological Manifold Harmonizer* performs a standard smoothing across reachable nodes, the Spectral variant treats the manifold as a conductive surface where signal strength is modulated by local hierarchical density. This method integrates topological heat and manifold conductivity through a multiplicative gating process:

1. **Topological Heat Diffusion:** Similar to the Harmonizer variant, we utilize the networkx library to simulate a diffusion process on the STRING PPI interactome (G_{PPI}). We compute a Personalized PageRank distribution **pr** starting from the semantic seeds. This distribution, which we define as the "Topological Heat," represents the global reachability of all candidate genes from the expert reasoning of the LLM.
2. **Hyperbolic Conductivity Gating:** In parallel, we calculate the hierarchical relevance of each neighbor using coordinates from the gensim *PoincareModel*. For each reachable gene j, we compute a manifold density weight $w_{\text{geo},j}$ based on its hyperbolic distance to the semantic center:
$$w_{\text{geo},j} = \exp\left(-\frac{d_{\mathbb{H}}(\mathbf{z}_j, C)^2}{2\sigma^2_{\text{spec}}}\right) \tag{A.7}$$
where $d_{\mathbb{H}}$ is the Poincaré metric. In the Spectral variant, the distance bandwidth σ_{spec} is constrained to the 15^{th} percentile (rather than 20th) to enforce a tighter structural constraint.
3. **Spectral Weight Product:** The final weighting $w_{\text{final},j}$ for each gene in the ensemble is calculated as the Hadamard (element-wise) product of the topological heat and the manifold conductivity:
$$w_{\text{final},j} = \text{pr}_j \times w_{\text{geo},j} \tag{A.8}$$
By multiplying these scores, the model applies a physical "gate": a gene must be both highly reachable in the interactome hierarchy *and* reside in the correct hierarchical branch of the manifold to influence the prediction.

Relationship to Previous Strategies: The **Spectral Manifold Harmonizer** differs from the standard *Harmonizer* by transitioning from additive signal smoothing to multiplicative signal gating. While the previous variants allow for broader context, the Spectral adaptation is designed for high-fidelity discovery, effectively pruning any topological neighbors that lack hierarchical support in the Poincaré ball. This dual-filter approach ensures that the resulting perturbation vector is maximally consistent with both the agent's semantic logic and the latent tree-like structure of the proteome.

B Results

B.1 Scaling Analysis of Agentic Consensus

The results in Table B.4 demonstrate that the *LangPert* consensus framework consistently enhances the predictive fidelity of the Gemini 3 Pro model across all scaling regimes. In well-characterized cell lines such as **RPE1** and **K562**, we observe a clear hierarchy: **Weighted Consensus** typically provides the most significant gains in low-data environments ($N \leq 200$) by leveraging the model's self-reported confidence to prioritize high-fidelity associations. As the training pool expands to $N = 800$, the **Binary Consensus** variant takes the lead, achieving peak correlations such as 0.8734 in RPE1. This shift indicates that frequency-aware multiset aggregation becomes a superior noise-filter as the candidate pool grows. Even in the more challenging **Jurkat** and **HepG2** cell lines, these consensus mechanisms successfully stabilize predictions, showing that the aggregation of independent reasoning traces allows the model to self-correct and maintain high structural alignment within complex regulatory manifolds.

B.2 Augmenting Small Model Capabilities via Inductive Priors

The scaling results in Table B.5 demonstrate that incorporating structural priors significantly enhances the predictive accuracy of smaller language models. Across all cell lines, we observe a consistent "topological rescue" effect: the **Harmonizer** variant outperforms the baseline in nearly all low-to-mid data regimes, with its most pronounced impact in the *Jurkat* lineage where it improves the correlation by nearly 25% at $N = 50$ (0.4721 vs 0.3826). As data density increases to $N = 800$, the optimal strategy shifts toward the **Spectral Harmonizer** and the **3+2 Strategy**, which provide the highest fidelity for well-characterized networks like *K562* and the challenging *HepG2* line. Collectively, these results show that while performance generally improves with training size N, offloading topological reasoning to a physical manifold allows efficient models to match the performance of significantly larger backbones.

Table B.4: Main Paper Performance Comparison (C_{20} Pearson Correlation ± SEM) using the Gemini 3 Pro backbone. We evaluate the impact of agentic consensus strategies on prediction fidelity across five training pool sizes (N). *LangPert* represents the standard single-agentic retrieval. *Binary Consensus* implements a frequency-aware multiset aggregation across multiple independent reasoning traces. *Weighted Consensus* utilizes the LLM's self-reported confidence scores to weight the contribution of each identified regulator. Maximum values for each row are bolded.

Cell Line	N	LangPert	Binary Consensus	Weighted Consensus
RPE1	50	0.8142 ± 0.0486	0.8170 ± 0.0449	**0.8227 ± 0.0456**
	100	0.8098 ± 0.0471	0.8013 ± 0.0500	**0.8116 ± 0.0483**
	200	0.7756 ± 0.0601	0.7936 ± 0.0604	**0.8003 ± 0.0598**
	500	0.8138 ± 0.0471	**0.8246 ± 0.0494**	0.8172 ± 0.0509
	800	0.7681 ± 0.0767	**0.8734 ± 0.0277**	0.8176 ± 0.0518
K562	50	0.5696 ± 0.0704	0.6063 ± 0.0607	**0.6139 ± 0.0655**
	100	0.6142 ± 0.0692	0.6234 ± 0.0670	**0.6659 ± 0.0638**
	200	0.6410 ± 0.0745	0.6475 ± 0.0769	**0.6725 ± 0.0673**
	500	0.6773 ± 0.0753	**0.7048 ± 0.0689**	0.6725 ± 0.0750
	800	0.7025 ± 0.0678	**0.7203 ± 0.0720**	0.7120 ± 0.0679
Jurkat	50	0.4057 ± 0.0888	**0.5270 ± 0.0740**	0.4685 ± 0.0800
	100	0.3801 ± 0.1054	0.4128 ± 0.1010	**0.4400 ± 0.0966**
	200	0.3777 ± 0.1086	0.3721 ± 0.1129	**0.4077 ± 0.1136**
	500	0.4293 ± 0.1149	0.4266 ± 0.1130	**0.4336 ± 0.1186**
	800	0.4424 ± 0.1191	**0.4473 ± 0.1167**	0.4202 ± 0.1221
HepG2	50	**0.2930 ± 0.0990**	0.2696 ± 0.1013	0.2532 ± 0.1048
	100	0.5498 ± 0.0679	0.5380 ± 0.0634	**0.5529 ± 0.0648**
	200	0.4366 ± 0.0968	**0.4620 ± 0.0950**	0.4460 ± 0.0980
	500	0.4019 ± 0.0961	**0.4691 ± 0.0869**	0.4312 ± 0.0918
	800	0.5083 ± 0.0973	0.4946 ± 0.0857	**0.5294 ± 0.0916**

C Prompt Templates

This section details the exact instruction sets used for both experimental tasks. To ensure high-fidelity biological reasoning, all prompts follow a *Chain-of-Thought* (CoT) protocol requiring the model to articulate mechanistic justifications before returning structured outputs.

C.1 Task 1: Perturbation Prediction

C.1.1 Baseline (LangPert): Functional Similarity Retrieval

Baseline System Prompt

You are a Lead Computational Biologist. Your task is to perform an *analogous function mapping* for gene perturbations. You must ground your similarity assessments in high-confidence mechanistic evidence, prioritizing genes that occupy identical transcriptomic manifolds or pathway positions. Respond strictly in valid JSON.

Table B.5: Scaling Performance Comparison (C_{20} Pearson Correlation ± SEM) using the Gemini 3 Flash backbone. We evaluate the impact of different inductive priors on the predictive fidelity of smaller language models across five training pool sizes (N). *Base LangPert* represents the standard single-agentic retrieval. The *3+2 Strategy* utilizes a discrete Euclidean manifold expansion (3 expert seeds + 2 nearest neighbors). The *Harmonizer* variant implements continuous topological smoothing using Personalized PageRank on the STRING PPI network combined with Poincaré hyperbolic weighting. The *Spectral* variant utilizes a multiplicative Hadamard gate between topological reachability and manifold conductivity to isolate high-fidelity regulatory signals. Maximum values for each configuration (row) are bolded.

Cell Line	N	LangPert	3+2 Strategy	Harmonizer	Spectral
K562	50	0.5535 ± 0.0375	0.5531 ± 0.0377	$\mathbf{0.5898 \pm 0.0291}$	0.5381 ± 0.0374
	100	0.5732 ± 0.0397	0.5737 ± 0.0394	$\mathbf{0.6124 \pm 0.0318}$	0.5320 ± 0.0406
	200	0.6132 ± 0.0359	0.6112 ± 0.0358	$\mathbf{0.6334 \pm 0.0314}$	0.5873 ± 0.0375
	500	0.6525 ± 0.0351	0.6571 ± 0.0351	0.6458 ± 0.0329	$\mathbf{0.6685 \pm 0.0344}$
	800	$\mathbf{0.6673 \pm 0.0355}$	0.6665 ± 0.0353	0.6594 ± 0.0341	0.6653 ± 0.0374
RPE1	50	0.7520 ± 0.0291	0.7519 ± 0.0292	$\mathbf{0.7686 \pm 0.0267}$	0.7416 ± 0.0288
	100	0.7418 ± 0.0300	0.7419 ± 0.0300	$\mathbf{0.7752 \pm 0.0260}$	0.7396 ± 0.0296
	200	0.7484 ± 0.0286	0.7471 ± 0.0287	$\mathbf{0.7789 \pm 0.0263}$	0.7476 ± 0.0280
	500	0.7816 ± 0.0266	0.7816 ± 0.0266	$\mathbf{0.7932 \pm 0.0259}$	0.7845 ± 0.0257
	800	0.8001 ± 0.0246	$\mathbf{0.8020 \pm 0.0244}$	0.7933 ± 0.0261	0.7799 ± 0.0279
Jurkat	50	0.3826 ± 0.0489	0.3832 ± 0.0488	$\mathbf{0.4721 \pm 0.0509}$	0.3926 ± 0.0493
	100	0.4268 ± 0.0513	0.4272 ± 0.0513	$\mathbf{0.4863 \pm 0.0508}$	0.4075 ± 0.0506
	200	0.4381 ± 0.0496	0.4425 ± 0.0497	$\mathbf{0.4723 \pm 0.0522}$	0.4159 ± 0.0486
	500	0.4492 ± 0.0475	0.4503 ± 0.0471	$\mathbf{0.4661 \pm 0.0552}$	0.3835 ± 0.0519
	800	0.4432 ± 0.0494	0.4530 ± 0.0482	$\mathbf{0.4690 \pm 0.0560}$	0.4262 ± 0.0509
HepG2	50	0.1788 ± 0.0465	0.1788 ± 0.0464	$\mathbf{0.1834 \pm 0.0565}$	0.1374 ± 0.0476
	100	0.2616 ± 0.0505	$\mathbf{0.2633 \pm 0.0499}$	0.2091 ± 0.0576	0.2615 ± 0.0482
	200	0.2313 ± 0.0532	0.2189 ± 0.0536	0.2085 ± 0.0577	$\mathbf{0.2397 \pm 0.0551}$
	500	0.2896 ± 0.0502	$\mathbf{0.2997 \pm 0.0496}$	0.2248 ± 0.0551	0.2525 ± 0.0516
	800	$\mathbf{0.3205 \pm 0.0506}$	0.3203 ± 0.0510	0.2570 ± 0.0530	0.2674 ± 0.0543

Baseline User Prompt

Instruction: Identify {`k_range`} functional neighbors for the target gene {`gene`}.
Available Candidates: {`list_of_genes`}
Chain-of-Thought Protocol:

1. **Profile Analysis:** Define the metabolic/signaling role of {`gene`} in the {`context`}.
2. **Pathway Alignment:** For each top-ranked candidate, identify the specific shared interaction (e.g., membership in the same protein complex or signaling cascade).
3. **Adjudication:** Rank neighbors based on the depth of the mechanistic link.

Output Format (JSON): {`"reasoning": {... }, "kNN": ["G1", "G2", ...]`}

C.1.2 CausalPert (Ours): Mechanistic Grounding via Regulatory Anchors

CausalPert System Prompt

You are a Molecular Systems Geneticist. Your objective is not general similarity, but the identification of the *Mechanical Drivers* of a gene's perturbation profile. Priority: Transcription Factors (TFs), downstream targets, or obligate complex partners.

CausalPert User Prompt

Task: Map the regulatory hierarchy of {gene} in {context}.
Candidate Pool: {list_of_genes}
Hierarchical Search Logic:

1. **Upstream:** Identify TFs from the pool that directly regulate {gene}.
2. **Downstream:** Identify primary targets of {gene} (if it is a TF/signaling node).
3. **Complexation:** Identify proteins that form stable, non-redundant complexes with the target.

Output Format (JSON):

```
{
  "regulators": [
    {"gene": "SYMB", "confidence": 0-100, "type": "TF/Target/Partner"}
  ],
  "reasoning": "...",
  "mechanism": "Brief description of the circuit."
}
```

C.2 Task 2: Active Experimental Design

ł

C.2.1 Step 1: Iterative Master Regulator Selection

Step 1: Iterative Candidate Identification

Context: Active discovery of the {cell_line} regulatory network. **Objective:** Select the next 10 most informative ”Master Regulators” to perturb. **Iterative Constraints:**

1. **Novelty:** Prioritize genes not in the already perturbed set: {perturbed_list}.
2. **Regulatory Reach:** Select nodes predicted to have maximum transcriptomic impact ($> 50k$ targets).
3. **Exclusivity:** You MUST select ONLY from the available pool: {gene_list_sample}...

Return strictly a valid JSON array of 10 gene symbols: `["REG1", "REG2", ..., "REG10"]`

C.2.2 Step 2: Evidence-Grounded Target Mapping

Mechanistic Verification (Batch)

Map the primary downstream targets for the candidates: {batch_list}. **Requirement:** For each regulator, provide a confidence interval grounded in literature (ChIP-seq, Perturb-seq, RNA-seq). **Confidence Calibration:** 1.0 (Direct cell-type specific evidence); 0.7 (Lineage-wide evidence); 0.4 (Computational/Motif prediction only). **Output Format (JSON):**

```
{
  "REGULATOR_X": {
    "targets": ["T1", "T2"],
    "confidence": 0.9,
    "logic": "Repression/Activation",
    "evidence_note": "A summary of the supporting literature."
  }
}
```

D Consensus Aggregation Protocol

For the CausalPert model, we execute 3 independent reasoning chains with temperature $T = 0.7$. The final regulator set is computed as:

$$w_r = \sum_{i=1}^{3} \nVdash[r \in \mathcal{R}_i] \cdot c_{i,r} \tag{D.9}$$

where $\mathcal{R}_i$ is the regulator set from chain i and $c_{i,r} \in [0, 1]$ is the self-reported confidence for regulator r in chain i (normalized from the 0–100 scale). Regulators are ranked by w_r and the top-k are selected.

For **Binary Consensus**, we set $c_{i,r} = 1$ for all regulators, effectively counting votes:

$$w_r^{\text{binary}} = \sum_{i=1}^{3} \nVdash[r \in \mathcal{R}_i] \tag{D.10}$$