

BIODISCOVERYAGENT: AN AI AGENT FOR DESIGNING GENETIC PERTURBATION EXPERIMENTS

Anonymous authors

Paper under double-blind review

ABSTRACT

Agents based on large language models have shown great potential in accelerating scientific discovery by leveraging their rich background knowledge and reasoning capabilities. In this paper, we introduce *BioDiscoveryAgent*, an agent that designs new experiments, reasons about their outcomes, and efficiently navigates the hypothesis space to reach desired solutions. We demonstrate our agent on the problem of designing genetic perturbation experiments, where the aim is to find a small subset out of many possible genes that, when perturbed, result in a specific phenotype (e.g., cell growth). Utilizing its biological knowledge, *BioDiscoveryAgent* can uniquely design new experiments without the need to train a machine learning model or explicitly design an acquisition function as in Bayesian optimization. Moreover, *BioDiscoveryAgent* using Claude 3.5 Sonnet achieves an average of 21% improvement in predicting relevant genetic perturbations across six datasets, and a 46% improvement in the harder task of non-essential gene perturbation, compared to existing Bayesian optimization baselines specifically trained for this task. Our evaluation includes one dataset that is unpublished, ensuring it is not part of the language model’s training data. Additionally, *BioDiscoveryAgent* predicts gene combinations to perturb more than twice as accurately as a random baseline, a task so far not explored in the context of closed-loop experiment design. The agent also has access to tools for searching the biomedical literature, executing code to analyze biological datasets, and prompting another agent to critically evaluate its predictions. Overall, *BioDiscoveryAgent* is interpretable at every stage, representing an accessible new paradigm in the computational design of biological experiments with the potential to augment scientists’ efficacy.

1 INTRODUCTION

Scientific discovery often resembles a search problem, where multiple hypotheses are tested to find one that leads to informative outcomes (Simon et al., 1981; Langley et al., 1987). AI agents, particularly those built using large language models (LLMs), show promising capabilities for this task (Wang et al., 2023). Previous studies have demonstrated that LLMs can effectively learn from the scientific literature (Wang et al., 2023; Boiko et al., 2023; Moor et al., 2023), compare various experimental plans (Liu et al., 2024; 2023; Huang et al., 2023b) and use tools such as executing code (Schick et al., 2024) or accessing databases (Jin et al., 2024).

Biomedical research, which heavily relies on iterative experimentation, stands to benefit significantly from such agents. One key challenge in this field is identifying drug targets—typically proteins that, when engaged by a drug, lead to a desired change in disease phenotype. Misidentification of these drug targets is a major cause of failure in clinical trials (Nelson et al., 2015). CRISPR-based genetic perturbation experiments are instrumental in addressing this challenge. These experiments involve the repression or activation of genes that code for proteins, followed by the measurement of the resulting biological effects, helping to identify drug targets that could reverse disease effects (Przybyla & Gilbert, 2022). These *perturbation screens* or *forward genetics screens* (Schneeberger, 2014; Moresco et al., 2013), have been transformative in areas such as drug target discovery (Wang & Doudna, 2023), elucidating disease mechanisms (Mamedov et al., 2023), cell engineering (Lim, 2022), gene therapy (Kalos et al., 2011), and immunotherapy (Goodman et al., 2022).

However, experimentally perturbing every single gene is costly. A perturbation screen typically perturbs around 19,000 protein-coding genes, yet anywhere between a handful of genes to a few thousand may exhibit the desired phenotype. When perturbing combinations of genes, this search space is even larger. By strategically designing these experiments in smaller batches that prioritize genes likely to result in meaningful phenotypic effects, it is possible to enhance the efficiency of the search process (King et al., 2004; Cleary et al., 2017; Huang et al., 2023a; Roohani et al., 2023).

Recent work has benchmarked Bayesian optimization algorithms for this task, but this requires training bespoke machine learning models that are often difficult to interpret, on small datasets (Mehriou et al., 2021; Lyle et al., 2023) (Figure 1a). Moreover, these models are unable to leverage the vast biological prior knowledge contained within the scientific literature that is valuable during early experimentation. In contrast, LLMs have been exposed to the scientific literature and can recall biomedical knowledge (Gao et al., 2024). Such models have shown state of the art performance in information retrieval for clinical question answering (Moor et al., 2023), patient matching for clinical trials (Wornow et al., 2024), gene set identification (Hu et al., 2023), gene function prediction (Chen & Zou, 2023) and cell type annotation using gene expression data (Hou & Ji, 2024). However, LLM-based agents have not yet been used for closed-loop biological experiment design.

For effective experiment design, agents require both domain-specific knowledge and the ability to interpret and reason over experimental results. While LLMs demonstrate strong capabilities in these areas, their full applicability across diverse biological contexts requires access to not only the literature but also external sources like tabular datasets. A careful balance must be struck between granting the LLM freedom to explore the action space of genes to perturb and ensuring the selected genes are biologically valid and well-motivated. The LLM must carry information across successive prompts to maintain a consistent experimental strategy. Additionally, the agent’s decision-making should be interpretable, ideally including literature citations and mechanisms for human feedback.

Here, we overcome some of these challenges and introduce *BioDiscoveryAgent*, an agent that designs genetic perturbation experiments using only an LLM paired with a suite of tools (Figure 1b). In each round, the agent constructs a prompt that includes both the task description and experimental results from previous rounds. This prompt is fed into the LLM, and the response identifies genes to perturb for the next round of experiments. *BioDiscoveryAgent* can also leverage different tools: it can search the scientific literature for relevant articles, execute code to analyze datasets that are inaccessible through textual sources and prompt another agent to critique the predictions of the initial agent.

BioDiscoveryAgent uniquely designs genetic perturbation experiments without relying on a specifically trained machine learning model or an explicitly defined acquisition function. It accomplishes this by leveraging an LLM that can effectively integrate information from both its prior knowledge and experimental results. Moreover, *BioDiscoveryAgent* using Claude 3.5 Sonnet outperforms baseline methods for experiment design, identifying 21% more experimental *hits* after five experimental rounds of 128 genes each, which is approximately 17 additional phenotypically relevant genes per dataset. On the harder task of only predicting non-essential genes, it identifies 46% more hits than baselines. *BioDiscoveryAgent* (Claude 3.5 Sonnet) also shows more than twice as high performance in predicting responses to combinatorial gene perturbations compared to a random baseline, exploring a new setting not previously considered. Unlike conventional approaches that are entirely black-box, the agent’s decision-making is fully transparent at every stage and can be enhanced through using tools such as LLM-based model critique. Overall, *BioDiscoveryAgent* utilizes its vast biological knowledge along with the ability to reason over insights from previous experimental results to offer an accessible and interpretable method for designing genetic perturbation experiments.

2 PROBLEM FORMULATION

We study the following closed-loop experimental design task: an agent determines which genes to perturb in each experimental round, using the results to plan subsequent rounds. The objective is to maximize the number of *hit* genes or gene combinations that are detected over the full course of rounds. In particular, we test our model on two real-world experimental settings:

Single-gene perturbation: Let \mathcal{G} represent the set of all protein-coding genes in humans, where $|\mathcal{G}| \approx 19,000$. The effect of perturbing a gene $g \in \mathcal{G}$, is denoted by a phenotypic response function $f(g)$, where $f : \mathcal{G} \rightarrow \mathbb{R}$ maps each gene to a real-valued phenotypic response. In this setting, we

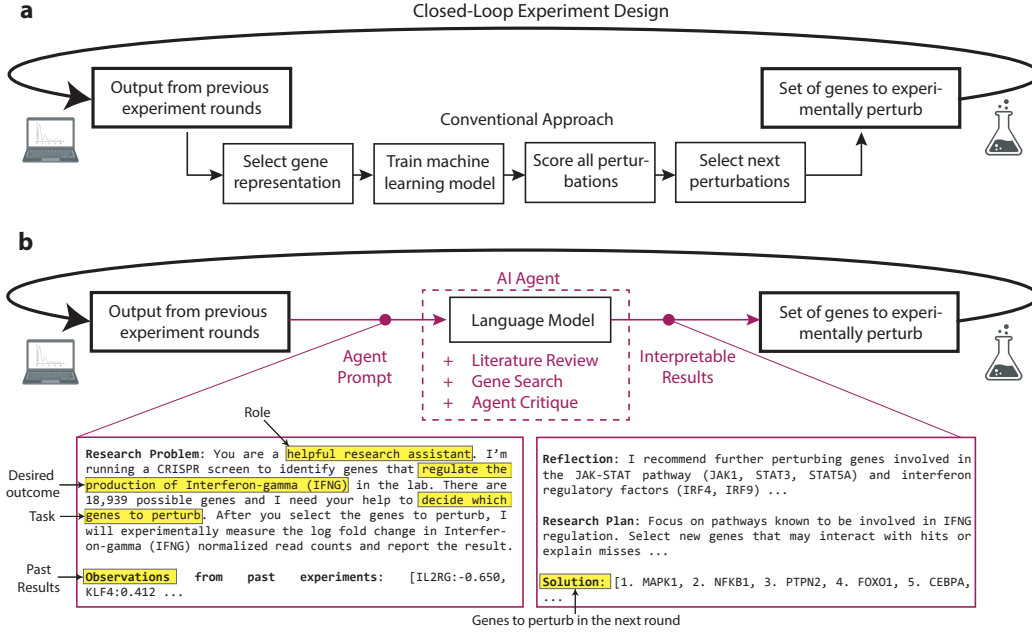


Figure 1: **An AI agent for closed-loop experiment design.** (a) Conventional Bayesian optimization approach for experiment design involves training a machine learning model in every experimental round, scoring all perturbations and defining an acquisition function for selecting genes to perturb in the next round. (b) Overview of BioDiscoveryAgent. In every round, the agent is given a prompt that describes the research problem, its role and task as well as experimental results from previous rounds. It generates a set of genes to perturb in the next round and provides reasoning for its prediction. The agent primarily makes use of a large language model for its predictions but also has access to additional tools such as the ability to search for relevant scientific papers on the internet.

define an experimental round as the perturbation of a set of genes $\{g_1, g_2, \dots, g_B\} \subseteq \mathcal{G}$. Each gene in this set is perturbed individually within separate cells. The goal of the experiment is to identify those genes that upon perturbation produce a desired phenotype, $f(g) > \tau$ for some threshold τ . These genes are referred to as *hits* for that phenotype.

The goal for the agent is to guide the design of these experiments, such that over $t = 1, 2, \dots, T$ experimental rounds, the total number of hits identified can be maximized (Mehrjou et al., 2021). The experiment design procedure involves selecting b genes to perturb at each round t . Let B_t refer to the set of genes selected at round t . Following this selection, the phenotypic response of perturbing each of these genes $f(g)$ for each $g \in B_t$ is revealed. In the next round ($t + 1$), the agent has access to all phenotypic responses for genes tested in previous rounds: $1, 2, \dots, t$.

At the end of T rounds, the total hit ratio is computed as the fraction of true hits that were discovered through the series of experiments (similar to recall). Let $\mathcal{G}_a = \cup_{t=1}^T B_t^+$, where $B_t^+ = \{g \in B_t : f(g) > \tau\}$ represents the hits identified in round t . The hit ratio can then be formally expressed as hit ratio = $\frac{|\mathcal{G}_a|}{|\mathcal{G}_p|}$, where \mathcal{G}_p is the set of all true hits for the phenotype, $\mathcal{G}_p = \{g \in \mathcal{G} : f(g) > \tau\}$.

Two-gene perturbation: We also consider a new problem setting of predicting gene pairs (2-gene combinations) to perturb. Here, each query consists of two genes perturbed in a single cell simultaneously. This formulation is both more computationally challenging given the larger search space as well as biologically impactful. Let $\mathcal{G}^2 = \mathcal{G} \times \mathcal{G}$ denote the set of all possible gene pairs that can be perturbed. Each element in \mathcal{G}^2 is a pair of genes (g_a, g_b) where $g_a, g_b \in \mathcal{G}$. The perturbation experiment in this context is represented by a function $f_c : \mathcal{G}^2 \rightarrow \mathbb{R}$. This function f_c maps each gene pair to a real-valued phenotypic response, which measures the combined effect of perturbing both genes simultaneously. The goal for the agent in this setting is to identify *pairs* of genes whose

combined perturbation results in the desired phenotypic results, i.e. $f_c(g_a, g_b) > \tau$. Thus, *hits* also correspond to gene pairs instead of single genes.

3 BIODISCOVERYAGENT

We now present BioDiscoveryAgent which uses an LLM to automate the scientific discovery process for this biological setting. This includes accessing scientific knowledge, generating hypotheses, planning experiments and interpreting results (Algorithm 1). In the single gene setting, at each step t , the agent’s objective is to select a batch of B genes for testing in the next step. The agent receives a prompt that describes general information about the experimental setup and the biological hypothesis being tested (Figure 1b, Appendix A, B). The results from each experiment are incorporated into the next prompt, along with the same information about the experimental setup. This prompt creation draws from established methods in developing other LLM-based agents, such as pre-action reasoning (Yao et al., 2022), reflective thinking (Shinn et al., 2023), and stepwise planning (AutoGPT, 2023).

We tested 9 different LLMs across varying levels of complexity for use in BioDiscoveryAgent (Claude v1 (Anthropic, 2023), Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus (Anthropic, 2024b), Claude 3.5 Sonnet (Anthropic, 2024a), GPT-3.5-Turbo (OpenAI, 2023), GPT-4o (OpenAI, 2024a), o1-mini (OpenAI, 2024b), o1-preview (OpenAI, 2024c)). Due to the lack of transparent parameter counts from most organizations, we used the price per token as a proxy for model size. Unless otherwise noted, results are presented for the best performing high and low cost models which are Claude 3.5 Sonnet (Price per 1M output tokens: \$15) and Claude 3 Haiku (Price per 1M output tokens: \$1.25) respectively.

Processing gene list and experimental observations: When choosing genes for perturbation, it is not always feasible to include all possible genes to choose from. At the same time, we want to avoid arbitrarily narrowing down the gene list based on our existing knowledge. Therefore, we adopt a two-step approach: Initially, we allow BioDiscoveryAgent to suggest genes without restrictions (Appendix Figure 4a). This enables it to draw on its comprehensive understanding of biology freely. If the agent is unable to produce the required gene list after many trials due to invalid or repeated sampling of genes (Appendix Figure 4b,c), BioDiscoveryAgent summarizes the list of all remaining genes and adds it to the prompt to aid in gene selection (Appendix Figure 4d). The summarized gene list is designed to cover a broad range of biological pathways and functions. A similar summarization technique is used for the experimental observations when they exceed the LLM’s context.

Agent Response Format: To ensure interpretability and to guide the agent’s thought process, a consistent response format is defined across all prompts. We direct the LLM to structure its responses into several parts: *Reflection*, *Research Plan*, *Solution* (Appendix A, Figure 1b), similar to (Huang et al., 2023b). *Research Plan* helps in effective planning and monitoring progress. Through the *Reflection* and *Research Plan* entries, the model is able to provide additional reasoning behind a particular prediction. This also helps to rule out predictions that may be hallucinations or not well-motivated. *Solution* contains a formatted list of genes to perturb next.

3.1 AGENT TOOLS

The primary mechanism by which BioDiscoveryAgent interacts with the user is through natural language (Appendix Figure 5a). The agent also has access to additional resources to aid in this decision-making beyond what is stored within its weights. These include access to the biomedical literature, a second agent to critique its predictions, and the ability to analyze tabular datasets:

Literature search: As part of the input, the user can choose to have the agent query the literature at each experimental cycle to inform its predictions (Appendix Figure 5b,c). In this case, the agent uses the PubMed API (Wobben, 2020) to search for papers containing the most pertinent literature for the experiments that the agent was asked to design. The search terms are chosen by the agent. Once the top 5 papers have been identified, the agent summarizes the information within these papers including parsed title, abstract, methods, results, and conclusion sections. It then attaches the summarization to the prompt and use it to identify additional genes to perturb for the given experimental round. The citations to these papers are retained and returned along with the model predictions.

Gene search based on biological databases: LLMs are trained on text-based data and do not have access to many biological databases that are stored in the form of tabular data. We provide the agent with the ability to query databases to search for other genes with similar biological properties as hit genes from previous experimental rounds (Appendix Figure 5d). First, the API is called to perform enrichment analysis for biological processes on the Reactome 2022 database (Gillespie et al., 2022) to identify the most relevant biological pathways. For each identified pathway, the agent queries other genes that participate in the same pathway. The top K genes that appear most frequently in these pathways are then selected and concatenated to the prompt for the main agent. Gene search based on other criteria (co-essentiality profiles, expression patterns, KEGG pathways) were also tested (Appendix Table 5). Among these, Reactome enrichment analysis gave the best result.

AI critic: An LLM prompted to behave as a research assistant may respond very differently from one that is prompted to mainly critique a prediction. To benefit from these contrastive prompting strategies, we make use of an AI critic similar to the idea by (Weng et al., 2023). The goal is to identify mistakes and enhance the quality of the final prediction made by the agent (Appendix Figure 5e). At every round, once BioDiscoveryAgent selects a batch of genes to be tested, a critic agent (which is also an LLM) is prompted to critique the choice of the main agent i.e. it can change some or all the genes in the batch and come up with a new set of genes.

4 EXPERIMENTS

We assess model performance using data from past genetic perturbation experiments. We simulate the perturbation of a gene g by retrieving the relevant observation of the perturbation-induced phenotype $f(g)$ from this dataset. In every experimental round we perturb 128 genes, representing a reasonably sized small-scale biological screen. Since each round of experimentation can incur additional costs and introduce unwanted experimental variation, we focus our evaluations on fewer experimental rounds (5) to more accurately reflect a real biological setting. For each dataset, after each round, we calculate the hit ratio as the proportion of discovered hits out of the total true hits for that dataset.

4.1 DATASETS AND BASELINES

For the single-gene perturbation setting, we make use of five different datasets spread across different cell types, publication dates and data generation sites. Each of the datasets contains the phenotypic response of knocking-down over 18,000 individual genes in distinct cells, with the exception of (Scharenberg et al., 2023) which contains data for 1061 perturbations. All datasets were released after 2021, apart from one dataset (CAR-T¹) which is so far unpublished.

Each of the datasets measure a distinct biological process. The (Schmidt et al., 2022) dataset measures the changes in the production of two key cytokines involved in immune signaling: Interferon- γ (IFNG) and Interleukin-2 (IL-2) under different genetic perturbations performed in primary human T-cells. The (Carnevale et al., 2022) dataset includes perturbation screens for identifying genes that render T cells resistant to inhibitory signals encountered in the tumor microenvironment. Unpublished data (CAR-T dataset) studies the impact of genome-wide perturbations on CAR-T cell proliferation. The (Scharenberg et al., 2023) dataset measures the effect of perturbation on mediating lysosomal choline recycling in pancreatic cells, and the (Sanchez et al., 2021) dataset studies the change in expression of endogenous tau protein levels in neurons.

For the two-gene perturbation task, we use a dataset from a screen that knocked down 100,576 gene pairs in K562 cells (Horlbeck et al., 2018). For each gene pair, we are interested in the synergistic effects on cell fitness upon combinatorial knockdown. Synergy is determined by the deviation between the observed cell fitness and the expected fitness, which is calculated by summing the average impact of knocking down each gene in the combination individually (Horlbeck et al., 2018).

For baseline models, we use the GeneDisco benchmark (Mehrjou et al., 2021) and the DiscoBax method (Lyle et al., 2023). Every baseline includes a multi-layer perceptron M for predicting experimental outcomes using gene features. This is then combined with one of seven different acquisition functions for designing each round of experiments using Bayesian optimization (Appendix E).

¹generated by the authors of this paper

5 RESULTS

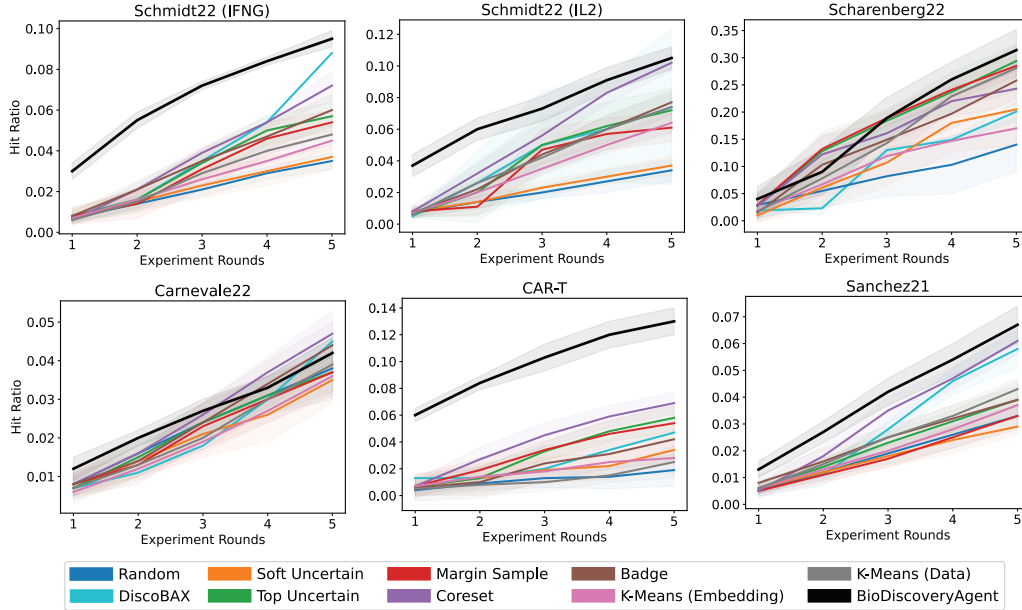


Figure 2: **Per-round performance comparison to machine learning baselines for 1-gene perturbation experiments:** Each line averages the hit ratio from 10 runs, with error bars indicating 1 standard deviation. 128 genes were predicted per round over 5 rounds. For (Scharenberg et al., 2023), a batch size of 32 was used due to its smaller size of 1061 perturbations.

BioDiscoveryAgent outperforms baselines based on hit ratio for 1-gene perturbation experiments: We evaluate the performance of BioDiscoveryAgent without any tools (*No-Tools*) against a range of machine learning benchmarks, including random sampling. When measuring hit ratios across all genes at experimental round 5, BioDiscoveryAgent consistently demonstrates superior performance compared to the best baseline approach, across 7 out of 9 LLMs tested (Table 1, Appendix Table 7). The best performing LLM, Claude 3.5 Sonnet, outperforms the best baseline for each dataset by 21% on average. Performance improvement is observed across 5 of the 6 datasets. This gap is especially large at earlier rounds, where the LLM can leverage its biological knowledge to select genes, in contrast to baseline methods that suffer from the cold start problem (Figure 2).

To verify that the model isn’t relying on uninformative hits, we measured model performance in predicting hits that are non-essential genes. Essential genes are likely to be detected as hits under any perturbation screen given the strong phenotypic effect of perturbing these genes. On the other hand, the response of non-essential genes to perturbation, is harder to predict and often more useful biologically. When filtering for non-essential genes, BioDiscoveryAgent (Claude 3.5 Sonnet) shows an improvement across all 6 datasets with an even higher average performance improvement of 46% over baselines (Table 1, Appendix Table 7).

Additionally, we tested the model’s tendency to predict distinct genes when prompted differently and upon observing different experimental results. We observe a low Jaccard similarity between all predicted genes after five rounds of experiments for any pair of datasets (Appendix Figure 6), suggesting that gene selection is not invariant to the task prompt and experimental observations.

BioDiscoveryAgent can guide 2-gene combinatorial perturbation experiments: In addition to 1-gene perturbation experiments, we also demonstrate that BioDiscoveryAgent can guide 2-gene combinatorial perturbation experiments, which is significantly more difficult due to the much larger combinatorial search space (100,576 gene pairs considered in (Horlbeck et al., 2018)). As shown in Table 2, BioDiscoveryAgent (Claude 3.5 Sonnet) significantly outperforms the random sampling baseline by 170% on average.

Model	Schmidt1		Schmidt2		CAR-T [†]		Scharen.*		Carnev.		Sanchez	
	All	N/E	All	N/E	All	N/E	All	N/E	All	N/E	All	N/E
Random	0.037	0.035	0.031	0.031	0.033	0.020	0.160	0.136	0.036	0.039	0.034	0.034
Human	0.032	0.036	0.029	0.034	0.033	0.034	0.113	0.122	0.042	0.049	0.025	0.026
Baseline Models												
Soft Uncertain	0.037	0.037	0.037	0.035	0.034	0.032	0.205	0.161	0.031	0.035	0.029	0.029
Top Uncertain	0.057	0.042	0.072	0.050	0.058	0.044	0.294	0.236	0.037	0.033	0.039	0.028
Margin Sample	0.054	0.040	0.061	0.047	0.054	0.045	0.285	0.227	0.036	0.032	0.033	0.028
Coreset	0.072	0.066	<u>0.102</u>	0.084	0.069	0.059	0.243	0.197	0.047	0.038	0.061	0.054
Badge	0.060	0.050	0.077	0.058	0.042	0.038	0.258	0.211	0.044	0.036	0.039	0.035
K-Means (E)	0.045	0.035	0.064	0.053	0.028	0.024	0.170	0.154	0.036	0.029	0.037	0.032
K-Means (D)	0.048	0.035	0.074	0.060	0.025	0.021	0.281	0.240	0.039	0.030	0.043	0.037
DiscoBax	0.088	0.069	0.074	0.057	0.047	0.021	0.201	0.200	0.045	0.038	0.058	0.049
BioDiscoveryAgent (No-Tools)												
Claude 3 Haiku	0.064	0.072	0.072	0.083	0.116	0.119	0.209	0.200	0.032	0.034	0.056	0.056
GPT-3.5-Turbo	0.044	0.048	0.061	0.073	0.064	0.066	0.230	0.188	0.032	0.034	0.039	0.038
Claude v1	0.067	0.086	0.089	0.110	0.110	0.126	0.292	0.222	0.038	<u>0.045</u>	0.053	0.055
o1-mini	0.081	0.092	0.096	0.120	0.127	<u>0.139</u>	0.279	0.286	0.040	<u>0.041</u>	0.074	0.073
Claude 3 Sonnet	0.076	0.082	0.088	0.111	0.115	0.118	0.302	0.265	0.041	0.042	0.064	<u>0.063</u>
Claude 3.5 Sonnet	0.095	0.107	0.104	<u>0.122</u>	<u>0.130</u>	0.133	0.326	0.292	0.042	0.044	0.066	<u>0.063</u>
GPT-4o	0.059	0.068	0.074	0.090	0.116	0.120	<u>0.311</u>	<u>0.286</u>	0.037	0.040	0.030	0.043
o1-preview	0.081	0.091	0.091	0.114	0.141	0.145	0.283	0.259	0.041	0.043	<u>0.068</u>	0.063
Claude 3 Opus	<u>0.094</u>	<u>0.106</u>	0.104	0.132	0.126	0.129	0.247	0.206	<u>0.043</u>	0.043	0.059	0.058

Table 1: **Performance comparison to machine learning baselines for 1-gene perturbation experiments.** Hit ratio for experimental round 5 averaged over 10 runs, with 128 genes predicted in each round. 'All' refers to results across all genes, 'N/E' only considers non-essential genes. The best performing machine learning model and the best performing LLM for BioDiscoveryAgent across all datasets are both highlighted in gray. *For (Scharenberg et al., 2023), a batch size of 32 was used due to its smaller size of 1061 perturbations. Schmidt1 refers to the screen measuring Interferon- γ (IFNG) and Schmidt2 measures Interleukin-2 (IL-2) following perturbation (Schmidt et al., 2022). [†]CAR-T refers to an unpublished dataset. LLMs are sorted by price per 1M output tokens (Table 4). See Appendix Table 7 for error intervals.

Model	Round 1	Round 2	Round 3	Round 4	Round 5
Random	2.6 \pm 1.43	5.7 \pm 2.83	8.9 \pm 3.67	12.8 \pm 3.74	16.4 \pm 3.8
BioDiscoveryAgent (No-Tools)					
Claude 3 Haiku	5.1 \pm 2.2	14.8 \pm 2.9	23.6 \pm 4.1	34.1 \pm 5.8	45.1 \pm 5.3
Claude 3.5 Sonnet	8.4 \pm 1.36	15.7 \pm 1.1	23.1 \pm 2.59	31.7 \pm 2.9	40.0 \pm 3.77

Table 2: **Performance evaluation for 2-gene perturbation experiments.** 32 gene pairs out of 100,576 pairs predicted per round averaged over 10 runs (Horlbeck et al., 2018). Evaluation using cumulative number of hits across 5 experimental rounds. Error intervals correspond to 1SD.

The performance of BioDiscoveryAgent can be augmented with tool-use, but the improvement varies by choice of LLM: We evaluated the impact of integrating three different tools—literature search, gene search, and AI critic into BioDiscoveryAgent (Table 3, Appendix Table 8). For the agent using Claude 3.5 Sonnet, tools applied individually or in combination did not lead to significant changes in performance and in some cases hurt performance. In contrast, for Claude 3 Haiku, incorporating all tools consistently enhanced performance.

The effects of different tools varied. In the case of Claude 3 Haiku, using only the literature search tool often resulted in fixation on a few simple keywords and irrelevant papers, leading to less effective searches. However, the literature search still provided verifiable citations for the gene prediction process, which is beneficial for scientists. Performance significantly improved with the use of a gene search based on shared biological pathways. This improvement likely stems from the gene search relying on tabular datasets not fully represented in text-based sources used for language model

Tools Used	Schmidt1	Schmidt2	CAR-T	Scharen.	Carnev.	Sanchez
Random	0.037	0.031	0.033	0.160	0.036	0.034
BioDiscoveryAgent (Claude 3.5 Sonnet) (Does not benefit from these tools)						
No-Tools	0.095	0.104	0.130	0.326	0.042	0.066
Literature	0.096 (+1%)	0.098 (-6%)	0.138 (+6%)	0.309 (-5%)	0.042 (+0%)	0.069 (+5%)
AI Critic	0.088 (-7%)	0.092 (-12%)	0.126 (-3%)	0.309 (-5%)	0.042 (+0%)	0.059 (-11%)
Gene Search	0.096 (+1%)	0.100 (-4%)	0.123 (-5%)	0.348 (+7%)	0.043 (+2%)	0.062 (-6%)
All-Tools	0.096 (+1%)	0.090 (-13%)	0.121 (-6%)	0.234 (-28%)	0.043 (+2%)	0.054 (-18%)
BioDiscoveryAgent (Claude 3 Haiku) (Benefits from these tools)						
No-Tools	0.064	0.072	0.116	0.209	0.032	0.056
Literature	0.053 (-17%)	0.069 (-4%)	0.091 (-22%)	0.164 (-22%)	0.035 (+9%)	0.057 (+2%)
AI Critic	0.061 (-5%)	0.070 (-3%)	0.113 (-3%)	0.219 (+5%)	0.043 (+34%)	0.054 (-4%)
Gene Search	0.080 (+25%)	0.098 (+36%)	0.114 (-2%)	0.249 (+14%)	0.046 (+44%)	0.065 (+16%)
All-Tools	0.084 (+31%)	0.099 (+38%)	0.128 (+10%)	0.259 (+24%)	0.043 (+34%)	0.058 (+4%)

Table 3: **Agent performance improvements when using different tools for 1-gene perturbation experiments.** Results show hit ratio for experimental round 5 averaged over 10 runs. See Table 1 caption for notes on specific datasets. See Appendix Table 8 for error intervals.

Tools Used	Schmidt1	Schmidt2	CAR-T	Scharen.	Carnev.	Sanchez	Price per 1M Token (\$)
Claude 3 Haiku	+31%	+38%	+10%	+24%	+34%	+4%	1.25
GPT-3.5-Turbo	+41%	+64%	-2%	-5%	+16%	+15%	3.00
Claude v1	+42%	+37%	+4%	+14%	+42%	+9%	11.00
o1-mini	+2%	-22%	-10%	-5%	0%	-3%	12.00*
Claude 3 Sonnet	-3%	+3%	-9%	0%	+15%	+14%	15.00
Claude 3.5 Sonnet	+1%	-13%	-6%	-28%	+2%	-18%	15.00
GPT-4o	-17%	-14%	-11%	-8%	-5%	-13%	15.00
o1-preview	-36%	-13%	-30%	+2%	-41%	-54%	60.00*
Claude 3 Opus	+2%	-3%	-8%	+18%	-5%	-7%	75.00

Table 4: **Effect of tools on different LLMs.** Results show increase in hit ratio when using all-tools as compared to results for the same agent when using no-tools. Models are sorted by increasing price per 1M output tokens. *For o1-mini and o1-preview, output tokens include internal reasoning tokens generated by the models that are not visible in API responses (OpenAI, 2024b). Results are shown for experimental round 5 averaged over 10 runs. See Appendix Table 9 for the actual hit rates.

training, which the *No-Tools* agent cannot access. Using only the AI critic tool slightly improved performance over the No-Tools model by diversifying predictions or focusing on specific gene sets.

We further evaluated the impact of incorporating tools across a broader range of LLMs (Table 4, Appendix Table 9). We observe that while some models such as Claude v1, Claude 3 Haiku and GPT3.5-Turbo showed a significant improvement in performance with the incorporation of all tools (24.7%, 23.5% and 21.5% respectively), others such as Claude 3.5 Sonnet, Claude 3.5 Opus and GPT4o showed a decrease in performance or no significant change (-10%, -1% and -11% respectively). This contrast suggests a potential relationship between model size and the benefit of tool use—smaller models appear to benefit more from tool-assisted retrieval, whereas larger models may derive this information directly from their trained weights.

We further investigated this relationship by measuring the percentage of new genes predicted by Claude 3 Haiku only when using tools, that are also predicted by Claude 3.5 Sonnet with no-tools (Appendix Figure 7). For some datasets, we see a large proportion of such tool-derived genes being predicted by Claude 3.5 Sonnet without any tool use (14-28%). This suggests that larger models are capable of intrinsically retrieving additional biological information that is otherwise provided by the tools. However, this wasn’t the case in all datasets indicating that some orthogonal information from the tools remains uncaptured by training on scientific text alone. Therefore, a more strategic design and selection of tools could lead to further improvements in agent performance for larger models.

BioDiscoveryAgent accounts for prior knowledge and observations in decision-making : Next, we investigate the use of prior knowledge versus observations from previous experiments in the agent’s decision-making. We examine three scenarios using BioDiscoveryAgent (Claude 3.5 Sonnet): 1) *Prompt + Observation*, where the agent utilizes both previous experiment results and detailed information about the experiment’s goal; 2) *Prompt Only*, where the agent ignores all experiment results; 3) *Observation Only*, where the agent is unaware of the current experiment’s goal and only conditioned on observations. To accurately capture these trends at fine resolution, we use a larger number of experimental rounds with smaller number of perturbations in each round: 30 rounds of experimentation with 32 genes in each round on the IFNG dataset from (Schmidt et al., 2022).

Results show that *Prompt + Observation* outperforms the other two scenarios, highlighting the significance of integrating prior knowledge and observations (Figure 3a). Interestingly, *Prompt + Observation* and *Prompt Only* benefit from prior knowledge early on, unlike *Observation Only*, which lacks the experiment’s goal, underscoring the vital role of prior knowledge in the initial experiment phases. However, as experiments progress, *Observation Only* surpasses *Prompt Only*, showcasing the agent’s capacity to adapt based on observations. For further validation, we also perform the same experiment using a different LLM and observe a similar trend (Appendix Figure 9, Table 11)

Additionally, we find that access to observations results in more similar gene predictions across different trials compared to experiments without access to observations (Figure 3b). This consistency was quantified using the Jaccard similarity index between all predicted genes after 30 rounds of experimentation. This suggests that observations significantly influence BioDiscoveryAgent’s decision-making, leading to more uniform choices across separate trials.

BioDiscoveryAgent provides interpretable predictions with references to the literature. BioDiscoveryAgent provides interpretable predictions at various stages. In one such example (Appendix G, Figure 5), the agent is tasked with identifying genes regulating the production of Interferon-gamma (IFNG). The agent explicitly reasons that it will focus on genes involved in mitochondrial respiration and the electron transport chain (Appendix Figure 5c). Utilizing the literature search tool, it accesses relevant literature to support its predictions, citing specific papers and line numbers (Appendix Figure 5b). For example, the agent highlights *STUB1* as a gene for potential perturbation, supported by references to specific lines in a relevant scientific paper. Alternatively, the agent is also able to indicate when the paper being summarized does not contain actionable information: "In this segment, I cannot find specific gene targets or pathways to focus on for the initial experiments" (Appendix G).

The LLM critic tool provides valuable insights into the predicted set of genes, identifying potential issues such as randomness or too narrow of a focus. In this example, the critic LLM suggests diversifying the selected genes by also considering pathways involved in interferon regulation, such as NF- κ B signaling and MAPK signaling (Appendix Figure 5e). Thus, not only does the critic LLM provide clear reasons for selecting specific genes but also further opens avenues for human-in-the-loop feedback by a subject-matter expert. In another example, the critic LLM guides the agent to limit the randomness in gene selection: "The selection of genes seems somewhat random and not focused enough on likely candidates based on known lysosomal and endolysosomal genes..." (Appendix G).

6 RELATED WORK

Artificial intelligence has demonstrated significant potential across various scientific domains, from realistic simulations of human behavior (Park et al., 2023b) to exploring mathematical function spaces (Romera-Paredes et al., 2023). Research has shown the utility of these models in mining and querying the scientific literature (Lála et al., 2023; Schick et al., 2023), as well as in general research tasks such as analyzing large datasets, reasoning about data, and generating reports (Shakked & Zhang, 2023; Lehr et al., 2024). Additionally, closed-loop AI-driven lab experimentation has made notable advances, particularly in chemical synthesis (Boiko et al., 2023; M. Bran et al., 2024) and materials discovery (Tom et al., 2024).

In the biological domain, research has shown that LLMs can capture significant information about biological pathways and processes (Park et al., 2023a), and are valuable in simulating biological processes at various scales (Schaefer et al., 2023). There have been benchmarking efforts for answering questions in genomics (Hou & Ji, 2023), performing bioinformatic analyses (Sarwal et al., 2023) as well as assessing broader biological research capabilities (Laurent et al., 2024). Despite long-

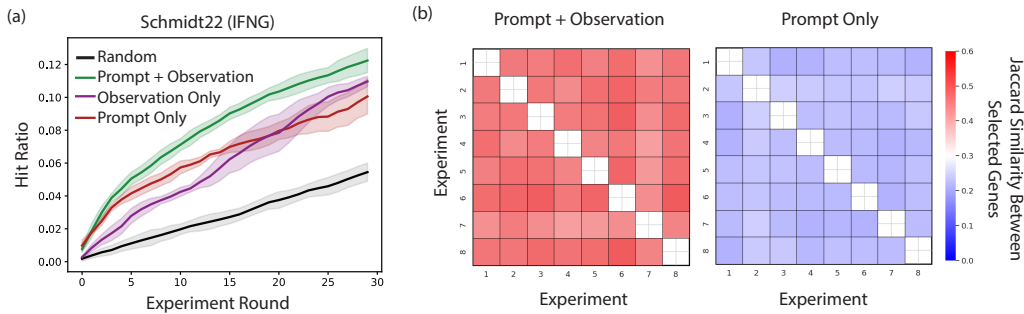


Figure 3: **Agent accounts for both prior knowledge and observations in decision-making** Three scenarios are considered: BioDiscoveryAgent (Claude 3.5 Sonnet) has access to task description and experimental observations (Prompt + Observation); only has access to experimental observations (Observation Only); only has access to the task description (Prompt Only). (a) Hit ratio at each experimental round across 30 rounds with 32 genes predicted in each. Each line corresponds to the average over 8 runs with error bars representing 1SD. (b) Jaccard similarity index between all predicted genes at 30 rounds of experimentation. Each cell corresponds to a different model run.

standing interest in developing autonomous AI systems for genetic perturbation experiment design (King et al., 2004), the use of LLM-based agents has yet to be explored for closed-loop biological experiment design. Some work has focused on designing individual gene editing experiments (Huang et al., 2024), primarily by optimizing experimental parameters and protocols. In contrast, our work, BioDiscoveryAgent, aims to generate and refine biological hypotheses through a series of perturbation experiments, leveraging the agent’s biological understanding and reasoning capabilities. Although agents have not been used in this setting, past research has investigated machine learning and Bayesian optimization techniques for similar purposes (Mehrjou et al., 2021; Lyle et al., 2023).

7 DISCUSSION

BioDiscoveryAgent represents a new paradigm in the design of biological experiments, aimed at augmenting scientists’ capabilities. Traditionally, this process employs a multi-stage pipeline using an acquisition function combined with a machine learning model. This model, often challenging to interpret, requires manual design and periodic retraining with handpicked gene features. In contrast, we demonstrate that an agent powered by an LLM can simplify the entire experimental design process into a single prompt from the researcher. Through its enhanced interpretability at every stage, the researcher can meaningfully engage with the model and augment its predictions.

Moreover, the agent is equipped with prior biological knowledge, solving the cold start problem at the beginning of a new round of experiments while efficiently utilizing observational data as the rounds progress. By using various tools, it can explicitly search for and integrate information from the scientific literature and existing tabular datasets. This capability results in a research assistant capable of speeding up biological research through utilizing information from diverse sources.

While BioDiscoveryAgent shows strong capabilities in enhancing experiment design, it still has room for improvement. It tends to perform better with certain cell types over others, likely due to variations in scientific literature coverage. Most performance benefits are observed in the early stages of experimentation rather than in prolonged rounds. Additionally, there is a need to develop better tools or fine-tuning methods to improve information extraction from non-text-based sources.

Overall, BioDiscoveryAgent offers a complementary approach to existing experimental design methods, significantly improving model performance in the low data regime while enhancing overall model interpretability. Amid increasing interest in active experimental design for biological research, BioDiscoveryAgent provides compelling evidence that language models could become essential components of such systems. By facilitating improved reasoning and interpretability and leveraging a broad understanding of the scientific literature, AI tools like BioDiscoveryAgent are poised to become integral to experimental design strategies in the near future.

REFERENCES

- Anthropic. Introducing claude, 2023. URL <https://www.anthropic.com/index/introducing-claude>.
- Anthropic. Introducing claude 3.5 sonnet, 2024a. URL <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Anthropic. Introducing the next generation of claude, 2024b. URL <https://www.anthropic.com/news/claude-3-family>.
- AutoGPT. AutoGPT: An experimental open-source attempt to make gpt-4 fully autonomous. <https://github.com/Significant-Gravitas/Auto-GPT>, 2023.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.
- Julia Carnevale, Eric Shifrut, Nupura Kale, William A Nyberg, Franziska Blaesche, Yan Yi Chen, Zhongmei Li, Sagar P Bapat, Morgan E Diolaiti, Patrick O’Leary, et al. Rasa2 ablation in t cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925):174–182, 2022.
- Yiqun T Chen and James Zou. Genetp: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, 2023.
- Brian Cleary, Le Cong, Anthea Cheung, Eric S Lander, and Aviv Regev. Efficient generation of transcriptomic profiles by random composite measurements. *Cell*, 171(6):1424–1436, 2017.
- Shanghai Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.
- Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D’Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acid Research*, 50,D1 (2022): D687–D692, 2022.
- Daniel B Goodman, Camillia S Azimi, Kendall Kearns, Alexis Talbot, Kiavash Garakani, Julie Garcia, Nisarg Patel, Byungjin Hwang, David Lee, Emily Park, et al. Pooled screening of car t cells identifies diverse immune signaling domains for next-generation immunotherapies. *Science translational medicine*, 14(670):eabm1463, 2022.
- Max A Horlbeck, Albert Xu, Min Wang, Neal K Bennett, Chong Y Park, Derek Bogdanoff, Britt Adamson, Eric D Chow, Martin Kampmann, Tim R Peterson, et al. Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967, 2018.
- Wenpin Hou and Zhicheng Ji. Geneturing tests gpt models in genomics. *BioRxiv*, 2023.
- Wenpin Hou and Zhicheng Ji. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pp. 1–4, 2024.
- Mengzhou Hu, Sahar Alkhairy, Ingo Lee, Rudolf T Pillich, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Research Square*, 2023.
- Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.
- Kexin Huang, Romain Lopez, Jan-Christian Hutter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. *bioRxiv*, pp. 2023–12, 2023a.

- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as ai research agents. *ArXiv*, abs/2310.03302, 2023b. URL <https://api.semanticscholar.org/CorpusID:263671541>.
- Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btac075, 2024.
- Michael Kalos, Bruce L Levine, David L Porter, Sharyn Katz, Stephan A Grupp, Adam Bagg, and Carl H June. T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Science translational medicine*, 3(95):95ra73–95ra73, 2011.
- Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acid Research*, 45,D1 (2017): D353-D361, 2017.
- Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.
- Alexander Lachmann, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hoyjin J. Lee, Lily Wang, Moshe C. Silverstein, and Avi Ma’ayan. Massive mining of publicly available rna-seq data from human and mouse. *Nature Communications*, 9(1), 1366., 2018.
- Jakub Lála, Odhran O’Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.
- Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. Scientific discovery: Computational explorations of the creative processes. 1987. URL <https://api.semanticscholar.org/CorpusID:61752618>.
- Jon M Laurent, Joseph D Janizek, Michael Ruza, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnampati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.
- Steven A Lehr, Aylin Caliskan, Suneragiri Liyanage, and Mahzarin R Banaji. Chatgpt as research scientist: Probing gpt’s capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35):e2404328121, 2024.
- Wendell A Lim. The emerging era of cell engineering: Harnessing the modularity of cells to program complex biological function. *Science*, 378(6622):848–852, 2022.
- Shengchao Liu, Jiong Xiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023.
- Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. *arXiv preprint arXiv:2402.03921*, 2024.
- Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jesson, Stefan Bauer, Yarin Gal, and Patrick Schwab. Discobax discovery of optimal intervention sets in genomic experiment design. In *International Conference on Machine Learning*, pp. 23170–23189. PMLR, 2023.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.
- Murad R Mamedov, Shane Vedova, Jacob W Freimer, Avinash Das Sahu, Amrita Ramesh, Maya M Arce, Angelo D Meringa, Mineto Ota, Peixin Amy Chen, Kristina Hanspers, et al. Crispr screens decode cancer cell pathways that trigger $\gamma\delta$ t cell detection. *Nature*, 621(7977):188–195, 2023.

- Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. Genedisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2021.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.
- Eva Marie Y Moresco, Xiaohong Li, and Bruce Beutler. Going forward with genetics: recent technological advances and forward genetics in mice. *The American journal of pathology*, 182(5): 1462–1473, 2013.
- Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.
- OpenAI. Gpt-3.5-turbo, 2023. URL <https://platform.openai.com/docs/models/gpt-3-5-turbo>.
- OpenAI. Introducing gpt-4o, 2024a. URL <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. Openai o1-mini, 2024b. URL <https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/>.
- OpenAI. Introducing openai o1-preview, 2024c. URL <https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/>.
- Gilchan Park, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Patrick Johnstone, Shinjae Yoo, and Francis J Alexander. Comparative performance evaluation of large language models for extracting molecular interactions and pathway knowledge. *arXiv preprint arXiv:2307.08813*, 2023a.
- Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023b.
- Laralynne Przybyla and Luke A Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, pp. 1–3, 2023.
- Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, 2023.
- Carlos G Sanchez, Christopher M Acker, Audrey Gray, Malini Varadarajan, Cheng Song, Nadire R Cochran, Steven Paula, Alicia Lindeman, Shaojian An, Gregory McAllister, et al. Genome-wide crispr screen identifies protein pathways modulating tau protein levels in neurons. *Communications biology*, 4(1):736, 2021.
- Varuni Sarwal, Viorel Munteanu, Timur Suhodolschi, Dumitru Ciorba, Eleazar Eskin, Wei Wang, and Serghei Mangul. Biollmbench: A comprehensive benchmarking of large language models in bioinformatics. *bioRxiv*, pp. 2023–12, 2023.
- Moritz Schaefer, Stephan Reichl, Rob ter Horst, Adele M Nicolas, Thomas Krausgruber, Francesco Piras, Peter Stepper, Christoph Bock, and Matthias Samwald. Large language models are universal biomedical simulators. *bioRxiv*, pp. 2023–06, 2023.
- Samantha G Scharenberg, Wentao Dong, Ali Ghoochani, Kwamina Nyame, Roni Levin-Konigsberg, Aswini R Krishnan, Eshaan S Rawat, Kaitlyn Spees, Michael C Bassik, and Monther Abu-Remaileh. An spns1-dependent lysosomal lipid transport pathway that enables cell survival under choline limitation. *Science Advances*, 9(16):eadf8966, 2023.

- Timo Schick, Jane Dwivedi-Yu, R Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools (2023). *arXiv preprint arXiv:2302.04761*, 2023.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.
- Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W Freimer, Raymund Bueno, Vinh Q Nguyen, Franziska Blaeschke, Chun Jimmie Ye, and Alexander Marson. Crispr activation and interference screens decode stimulation responses in primary human t cells. *Science*, 375(6580): eabj4008, 2022.
- Korbinian Schneeberger. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nature Reviews Genetics*, 15(10):662–676, 2014.
- Noy Shakked and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. URL <https://doi.org/10.1126/science.adh2586>.
- Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv*, abs/2303.11366, 2023.
- Herbert A. Simon, Pat Langley, and Gary L. Bradshaw. Scientific discovery as problem solving. *Synthese*, 47:1–27, 1981. URL <https://api.semanticscholar.org/CorpusID:46985700>.
- Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.
- Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.
- Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.
- Joy Y Wang and Jennifer A Doudna. Crispr technology: A decade of genome editing is only the beginning. *Science*, 379(6629):eadd8643, 2023.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification, 2023.
- Gijs Wobben. pymed - pubmed querying with python, 2020. URL <https://github.com/gijswobben/pymed>. Accessed: 03/2024.
- Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W Mahaffey, and Nigam H Shah. Zero-shot clinical trial patient matching with llms. *arXiv preprint arXiv:2402.05125*, 2024.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022.

A PROMPT

The prompt includes the task information and response format as this example shown below.

You are a scientist working on problems in drug discovery.

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

Always respond in this format exactly:

1. Reflection: Thoughts on previous results and next steps.
2. Research Plan: The full high level research plan, with current status and reasoning behind each proposed approach. It should be at most 5 sentences.
3. Solution: Propose a list of predicted genes to test separated by commas in this format: 1. <Gene name 1>, 2. <Gene name 2> ...

Do not include any genes from this prompt (since they're already tested).

B DATASET SPECIFIC PROMPTS

For each dataset, the research problem and the type of measurement outcome are used to create the prompt.

```

IFNG Task: identify genes that regulate the production of
            Interferon-gamma (IFNG)
IFNG Measurement: the log fold change in Interferon-gamma (IFNG)
                  normalized read counts

IL2 Task: identify genes that regulate the production of
           Interleukin-2 (IL-2)
IL2 Measurement: the log fold change in Interleukin-2 (IL-2)
                 normalized read counts

CAR-T Task: identify genes that upon inhibition allow
            cells to resist T-cell
            exhaustion, under the HA GD2 CAR
            (chimeric-antigenic receptor) condition
CAR-T Measurement: the log fold change in normalized
                   sgRNA read counts
                   compared to the non-targeting control,
                   22 days after perturbation

Scharenberg Task: identify genes mediating lysosomal choline
                  recycling using an
                  endolysosome-focused CRISPR-Cas9 screen
Scharenberg Measurement: enrichment or depletion of targeting
                         sgRNAs (indicated by a high score) in the
                         culture medium lacking free choline

Carnevale Task: identify genes that, upon being knocked out,
                would boost the efficacy of engineered T cells in
                the presence of an adenosine agonist that
                creates an immunosuppressive condition
Carnevale Measurement: the change in T cell proliferation

```

C ALGORITHM FOR BIODISCOVERYAGENT

Algorithm 1 BioDiscoveryAgent: AI Agent for Biological Experiment Design (using all tools)

```

Input: Experiment description, Number of rounds  $T$ , Number of genes to perturb in each round  $b$ 
Output: Set of genes to perturb
for  $t = 1$  to  $T$  do
    Search and retrieve literature using LLM-generated search terms
    Summarize articles using LLM and attach results to the main gene selection prompt
    Perform LLM-requested gene search and attach results to the main gene selection prompt
    Prompt LLM to select  $b$  new genes (or gene pairs)
    Output: LLM generates a structured response with Reflection, Research Plan, Gene
    Search, and Solution entries
    while any predicted genes are invalid do
        Prompt LLM to select new genes
    end while
    Prompt LLM to critique the prediction made by the main agent
    Get phenotypic score  $f(g)$  for each gene  $g$  and add to gene selection prompt for the next step
    if context window is too large then
        Prompt LLM to summarize text for context window management
    end if
end for

```

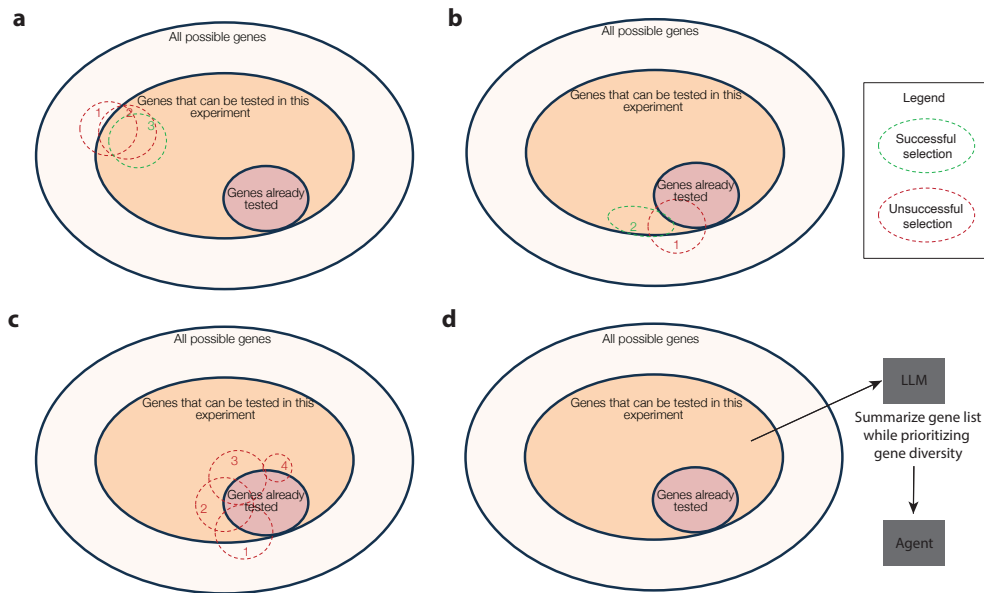


Figure 4: Gene selection strategy: (a) The space of genes that can be tested in a given experiment is constrained by experimental limitations. BioDiscoveryAgent can take a few tries to select genes within this limited space. (b) A common error is repeating previously tested genes. (c) Often this will result in the agent getting stalled and unable to make successful selections, especially in the case of large batch sizes. (d) After several failed attempts, we summarize the space of genes that can be tested while prioritizing gene diversity.

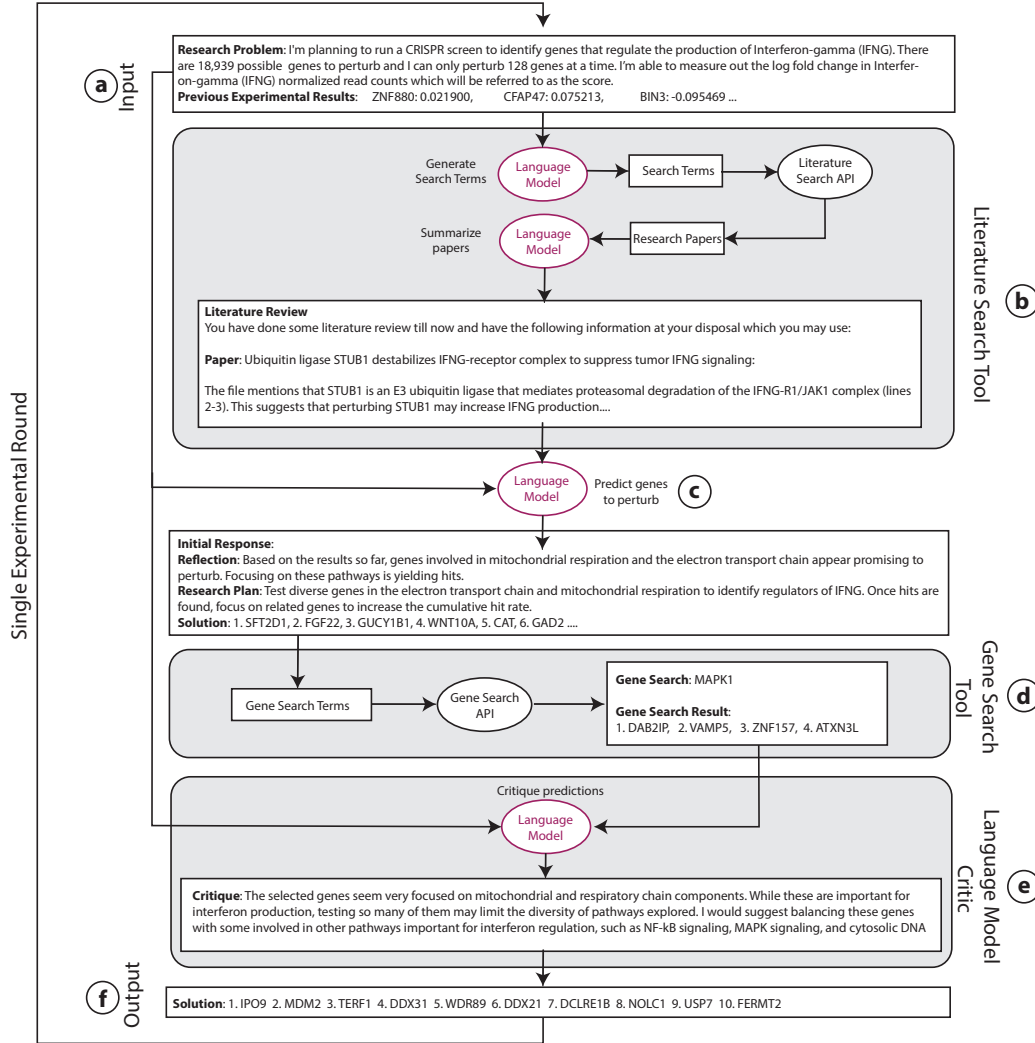


Figure 5: **BioDiscoveryAgent** workflow with all tools over a single experimental round. Prompts and agent responses have been summarized. See Appendix G for full trace. (a) The input to the agent is the description of the problem. (b) In case of the literature search tool, the LLM first determines appropriate search terms for finding relevant scientific papers. Top ranked retrieved papers are summarized by the LLM (along with line numbers that reference directly to text in the paper) and provided as additional context to the agent when predicting the set of genes to perturb. (c) The agent predicts the set of genes to perturb in the next experiment. Response is structured containing Reflection, Research Plan and Solution. (d) In case of using the gene search tool, the LLM identifies a query gene with which to search for other genes (e) The LLM can also be prompted differently to function as a critic and analyze the predictions so far from a different perspective (f) Final output gene list after using different tools. We used Claude v1 for this analysis since it showed high absolute score as well as high performance gain through tool use for this dataset (Schmidt et al., 2022).

D TOOL DESCRIPTIONS FOR BIODiscoveryAGENT

In this section, we provide some more details about the tools provided to BioDiscoveryAgent to aid it in making its predictions, along with the reasoning for adding those tools.

1. **AI critic:** LLMs are very sensitive to the prompt they are presented with. Past research has shown varying performance depending on the setting in which the LLM is queried. Thus, an LLM prompted to behave as an agent for a human researcher may behave very differently from one that is prompted to mainly critique the prediction made by another LLM. To benefit from these contrastive prompting strategies, we make use of an AI critic similar to the idea by (Weng et al., 2023) to identify mistakes and enhance the quality of the final prediction made by the agent.

At every round, once BioDiscoveryAgent comes up with a batch of genes to be tested, a critic agent (which is also an LLM) is prompted to critique the choice of the main agent and it can change some or all the genes in the batch and come up with a new set of genes (Figure 5e). Having such an agent improved performance giving the system additional opportunities to reflect on its reasoning. The system prompt for the critic agent was as follows:

```
As an advisor, please critique this plan and suggest
some changes to it. Use this format:
1. Critique: include all relevant details of the
critique.
2. Updated Solution: Give an updated selection of {args
.num_genes} genes based on the critique separated
by commas in this format:: 1. <Gene name 1>, 2. <
Gene name 2> ... \n

Try to first focus a lot on trying very diverse genes
to get a sense of which types of genes affect the
research problem the most. From the observations,
update your beliefs quickly and smartly and then
double down on genes that you think shall be hits
in order to increase the cumulative hit ratio.

Please do not critique/make a lot of changes if there
is no need to make a change.
```

In addition to the above prompt, the critic agent was also provided with a list of all genes that were tested in the previous rounds along with genes that were identified as hits.

2. **Literature Search:** Scientific literature captures prior knowledge that can be leveraged to design experiments. A scientist typically reads literature relevant to a problem, builds a hypothesis, and cites relevant past work. We aimed to provide BioDiscoveryAgent with similar capabilities that would allow it to search for relevant papers on the web, learn from them, and incorporate its learnings in designing the experiments.

As part of the input, the user can choose to have the agent query the literature at each experimental cycle to inform its predictions. In this case, the agent uses the PubMed API (Wobben, 2020) to search for papers containing the most pertinent literature for the experiments that the agent was asked to design. The search terms are chosen by the agent. Once the top 5 papers have been identified, the agent summarizes the information within these papers including parsed title, abstract, methods, results, and conclusion sections. It then attaches the summarization to the prompt and use it to identify additional genes to perturb for the given experimental round (Figure 5b). The citations to these papers are retained and returned along with the model predictions. Over time, the agent accumulated these summaries, granting it access to all literature surveys conducted in previous rounds to

propose a set of genes for each specific round.

The inclusion of a literature review tool enhanced interpretability and improved grounding as the agent frequently cited papers it had previously surveyed when predicting genes for the batch. Despite these benefits, the use of a literature review tool had its drawbacks. It tended to make the agent less exploratory, heavily biasing its reasoning towards the papers retrieved. The agent struggled to generate innovative queries for the literature survey API, and the lack of additional re-ranking on API outputs further limited the diversity of papers presented to the agent. This raised some important questions about the most effective use of scientific literature by an AI agent, a topic for future work.

3. **Gene Search:** LLMs are trained on text-based data and do not have access to many biological databases that are stored in the form of tables. We provide the agent with the ability to search for top 10 genes based on different criteria. The selected genes are concatenated to the prompt given to the main agent (Figure 5d). The results of augmenting the agent with these different gene search methods are summarized in Table 5.

- **Similar/Dissimilar Genes:** The agent selects a gene for which it would like to search for similar or dissimilar genes. This is computed using the cosine similarity between the provided gene features (gene co-essentiality profiles (Tsherniak et al., 2017)). This process first computes the inner product of gene features with the queried gene’s features and sorts the results based on the desired similarity or dissimilarity.
- **Correlated Genes:** The agent accesses ARCHS4 (Lachmann et al., 2018), a public database that provides RNA-seq expression data from human and mouse samples. The API call ranks genes based on the Pearson correlation coefficient, a linear relationship between two genes’ expression levels across multiple samples. Then the call returns top 10 most correlated genes to a gene that the agent selects for query.
- **Genes in Common Tissues:** The agent accesses ARCHS4 to retrieve a list of tissue types where the gene is expressed. Then the API call looks for top 10 genes that also show strong expression in these tissues, based on the median TPM (transcripts-per-million) value.
- **KEGG Enrichment Analysis:** The agent accesses the KEGG (Kanehisa et al., 2017) enrichment database, which provide curated information on biological pathways and functions associated with genes. For hit genes from previous rounds, the top 10 pathways that are over-represented by statistical significance are returned. For these pathways, the API call will generate the top 10 genes that are associated with them.
- **Reactome Enrichment Analysis:** The agent perform similar enrichment analysis to the above, but uses the Reactome database (Gillespie et al., 2022) instead.

Tools Used	Schmidt1	Schmidt2	CAR-T	Scharen.	Carnev.	Sanchez
Random	0.037	0.031	0.033	0.160	0.036	0.034
Claude 3.5 Sonnet						
No-Tools	0.095	0.104	0.130	0.326	0.042	0.066
Similar	0.091 (-4%)	0.098 (-6%)	0.145 (+12%)	0.342 (+5%)	0.043 (+2%)	0.061 (-8%)
Dissimilar	0.091 (-4%)	0.096 (-8%)	0.123 (-5%)	0.302 (-7%)	0.042 (+0%)	0.069 (+5%)
Correlated	0.088 (-7%)	0.094 (-10%)	0.138 (+6%)	0.358 (+10%)	0.044 (+4%)	0.058 (-12%)
Tissues	0.091 (-4%)	0.097 (-7%)	0.130 (+0%)	0.310 (-5%)	0.042 (+0%)	0.066 (+0%)
KEGG	0.089 (-6%)	0.091 (-13%)	0.122 (-6%)	0.326 (+0%)	0.045 (+6%)	0.057 (-14%)
Reactome	0.096 (+1%)	0.100 (-4%)	0.123 (-5%)	0.348 (+7%)	0.043 (+2%)	0.062 (-6%)
Claude 3 Haiku						
No-Tools	0.064	0.072	0.116	0.209	0.032	0.056
Similar	0.065 (+2%)	0.093 (+29%)	0.095 (-18%)	0.206 (-1%)	0.043 (+34%)	0.054 (-4%)
Dissimilar	0.066 (+3%)	0.073 (+1%)	0.081 (-30%)	0.215 (+3%)	0.046 (+44%)	0.054 (-18%)
Correlated	0.068 (+6%)	0.089 (+24%)	0.108 (-7%)	0.259 (+19%)	0.043 (34%)	0.047 (-16%)
Tissues	0.063 (-2%)	0.081 (+13%)	0.119 (+3%)	0.226 (+8%)	0.037 (+16%)	0.053 (-5%)
KEGG	0.070 (+9%)	0.074 (+3%)	0.099 (-15%)	0.232 (+11%)	0.036 (+13%)	0.047 (-16%)
Reactome	0.080 (+25%)	0.098 (+36%)	0.114 (-2%)	0.249 (+14%)	0.046 (+44%)	0.065 (+16%)

Table 5: **Effect of different gene search tools.** Results show hit ratio for experimental round 5 averaged over 10 runs, with error intervals showing 1 standard deviation. *For (Scharenberg et al., 2023), a batch size of 32 was used due to its smaller pool of 1061 relevant genes. Schmidt1 refers to the screen measuring Interferon- γ (IFNG) and Schmidt2 measures Interleukin-2 (IL-2) following perturbation (Schmidt et al., 2022). [†]CAR-T refers to an unpublished dataset.

E BASELINES

Table 6 contains the pathways used to sample initial genes for each dataset during human baseline.

Category	Pathway Name	Reactome Pathway ID	KEGG Pathway ID
IFNG	Interferon Gamma Signaling	R-HSA-877300	hsa04060
	Cytokine Signaling	R-HSA-1280215	hsa04060
IL-2	Interleukin-2 Signaling	R-HSA-451927	hsa04060
	Cytokine Signaling	R-HSA-1280215	hsa04060
Carnevale	PD-1 Signaling	R-HSA-389948	hsa05235
	T-Cell Receptor Signaling	R-HSA-202433	hsa04660
	Immune Checkpoints	R-HSA-389957	hsa05235
CAR-T Proliferation	IL-2 Signaling Pathway	R-HSA-451927	hsa04060
	PI3K-Akt Signaling Pathway	R-HSA-110021	hsa04151
	mTOR Signaling Pathway	R-HSA-165159	hsa04150
Scharenberg	Lysosomal Transport	R-HSA-3229371	hsa04142
	Choline Metabolism	R-HSA-6798163	hsa00564
	Autophagy	R-HSA-1632852	hsa04140
Sanchez	MAPK Signaling Pathway	R-HSA-5683057	hsa04010
	Protein Processing in ER	R-HSA-381119	hsa04141
	Ubiquitin-Proteasome Pathway	R-HSA-983168	hsa04120

Table 6: **Reactome and KEGG Pathways Used to Sample Genes during Human Baseline.** These pathways were used to sample initial batch of genes. In subsequent rounds, enrichment analysis using the Reactome database (Gillespie et al., 2022) was performed to sample the next batch. If the sampled genes were not sufficient in number, KEGG enrichment (Kanehisa et al., 2017) was also performed to fill the remaining samples.

For baseline models, we use the methods implemented in the GeneDisco benchmark (Mehrjou et al., 2021) and (Lyle et al., 2023). Every baseline includes a multi-layer perceptron M for predicting

experimental outcomes using gene features. This is then combined with one of seven different acquisition functions for designing each round of experiments:

- **Soft Uncertain:** Prioritizes genes with higher uncertainty under M , using a softmax function with temperature.
- **Top Uncertain:** Selects genes with the highest uncertainty under model M .
- **Margin Sample:** Selects genes for which the model M has the smallest margins between different classes.
- **Coreset:** Selects genes which are the most distant from previously selected genes based on their embedding representation in M .
- **Badge:** Uses a modified k-means algorithm on the gradient embeddings of the data points to select genes. The aim is to diversify the batch based on the model’s gradients.
- **Kmeans:** Selects genes that are closest to the cluster centers determined by K-means. Two baselines apply K-means either to an embedding of the data or the raw data directly.
- **DiscoBax** (Lyle et al., 2023): Selects genes with high expected change to the phenotype of interest as well as high diversity. Implemented as a set-value maximization problem.

In addition, we include a Human baseline. The human baseline uses pathways and traditional enrichment analysis to sample genes. In the first round, we select the genes most active in pathways related to each test. These pathways are listed in Table 6. Then in subsequent rounds, enrichment analysis is performed to previous samples using Reactome (Gillespie et al., 2022) and KEGG (Kanehisa et al., 2017) databases.

F COMPUTATIONAL COST

Computational cost associated with these tool use is another important aspect to consider. We observe that even considering the API cost increase, the tools designed in this paper works especially well for smaller models like Claude 3 Haiku. From Table 4, we observe that tool usage results in 23.5% performance improvement in average. This only comes with the average of \$0.14 increase in cost per trial (30%) - which is much less than what would be expected from literature review and critic agent. This is because tools help models to produce the required gene list in fewer number of turns. However, the tool usage with Claude 3.5 Sonnet neither reduces the cost or improves the performance. Table 10 shows the average number of input and output tokens along with API cost for these two models with each dataset.

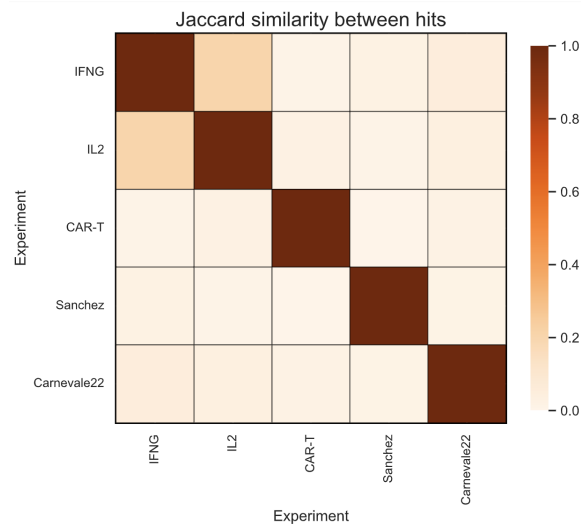


Figure 6: **Jaccard similarity index between all predicted hits at 5 rounds of experimentation** Each cell corresponds to the union of predicted hits across 10 model runs for each dataset.

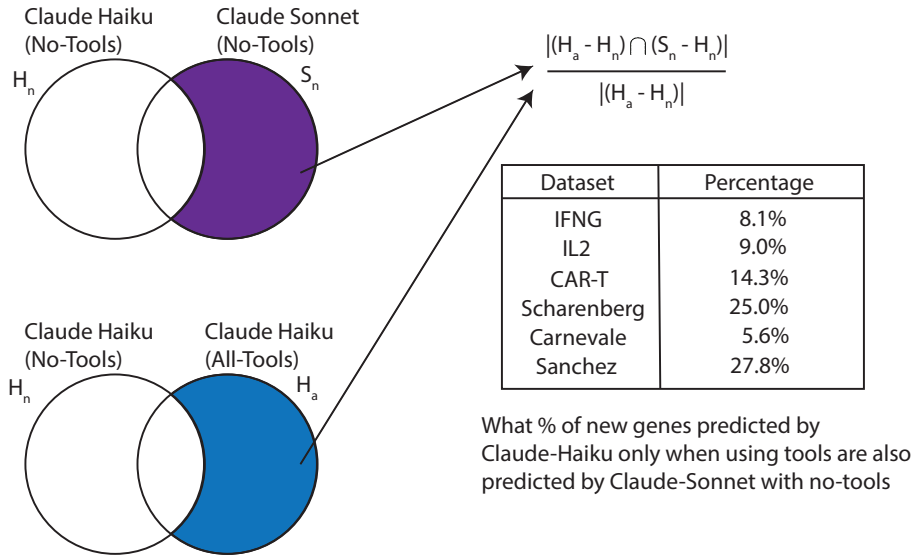


Figure 7: Percentage of new genes predicted by Claude Haiku only when using tools $H_a - H_n$ that are also predicted by Claude Sonnet with no-tools (S_n), where H_n is the set of genes predicted by Claude Haiku with no-tools and H_a is the set of genes predicted by Claude Haiku with all-tools

Model	Schmidt1	Schmidt2	CAR-T [†]	Scharen.*	Carnev.
Random	0.037 ± 0.013	0.031 ± 0.002	0.033 ± 0.003	0.160 ± 0.028	0.036 ± 0.001
Baseline Models					
Soft Uncertain	0.037 ± 0.006	0.037 ± 0.006	0.034 ± 0.007	0.205 ± 0.006	0.031 ± 0.006
Top Uncertain	0.057 ± 0.007	0.072 ± 0.014	0.058 ± 0.010	0.294 ± 0.030	0.037 ± 0.005
Margin Sample	0.054 ± 0.006	0.061 ± 0.009	0.054 ± 0.013	0.285 ± 0.019	0.036 ± 0.003
Coreset	0.072 ± 0.007	0.102 ± 0.005	0.069 ± 0.008	0.243 ± 0.031	0.047 ± 0.006
Badge	0.060 ± 0.008	0.077 ± 0.008	0.042 ± 0.017	0.258 ± 0.032	0.044 ± 0.006
Kmeans Embed.	0.045 ± 0.004	0.064 ± 0.007	0.028 ± 0.011	0.170 ± 0.032	0.036 ± 0.004
Kmeans Data	0.048 ± 0.005	0.074 ± 0.009	0.025 ± 0.012	0.281 ± 0.042	0.039 ± 0.004
DiscoBAX	0.088 ± 0.000	0.074 ± 0.049	0.047 ± 0.000	0.201 ± 0.018	0.045 ± 0.000
BioDiscoveryAgent (No-Tools)					
Claude 3 Haiku	0.064 ± 0.005	0.071 ± 0.018	0.116 ± 0.014	0.209 ± 0.030	0.032 ± 0.004
GPT-3.5-turbo	0.044 ± 0.007	0.061 ± 0.009	0.064 ± 0.012	0.230 ± 0.032	0.032 ± 0.004
Claude v1	0.067 ± 0.010	0.089 ± 0.014	0.110 ± 0.023	0.292 ± 0.064	0.038 ± 0.005
o1-mini	0.081 ± 0.007	0.096 ± 0.014	0.127 ± 0.016	0.279 ± 0.040	0.040 ± 0.004
Claude 3 Sonnet	0.076 ± 0.006	0.088 ± 0.007	0.115 ± 0.012	0.302 ± 0.036	0.041 ± 0.003
Claude 3.5 Sonnet	0.095 ± 0.004	0.104 ± 0.004	0.130 ± 0.010	0.326 ± 0.034	0.042 ± 0.003
GPT-4o	0.059 ± 0.006	0.074 ± 0.006	0.116 ± 0.031	0.311 ± 0.037	0.037 ± 0.004
o1-preview	0.081 ± 0.003	0.091 ± 0.004	0.141 ± 0.022	0.283 ± 0.015	0.041 ± 0.001
Claude 3 Opus	0.094 ± 0.009	0.104 ± 0.007	0.126 ± 0.011	0.247 ± 0.037	0.043 ± 0.005
Model	Sanchez				
Random	0.033 ± 0.004				
Baseline Models					
Soft Uncertain	0.029 ± 0.003				
Top Uncertain	0.039 ± 0.007				
Margin Sample	0.033 ± 0.005				
Coreset	0.061 ± 0.005				
Badge	0.039 ± 0.005				
Kmeans Embed.	0.037 ± 0.003				
Kmeans Data	0.043 ± 0.004				
DiscoBAX	0.058 ± 0.007				
BioDiscoveryAgent (No-Tools)					
Claude 3 Haiku	0.056 ± 0.005				
GPT-3.5-turbo	0.039 ± 0.004				
Claude v1	0.058 ± 0.007				
o1-mini	0.074 ± 0.005				
Claude 3 Sonnet	0.064 ± 0.012				
Claude 3.5 Sonnet	0.066 ± 0.007				
GPT-4o	0.030 ± 0.003				
o1-preview	0.068 ± 0.006				
Claude 3 Opus	0.059 ± 0.009				

Table 7: **Performance comparison to machine learning baselines for 1-gene perturbation experiments.** Results show hit ratio for experimental round 5 averaged over 10 runs, with error intervals showing 1 standard deviation. *For Scharenberg22, a batch size of 32 was used due to its smaller pool of 1061 relevant genes. [†]CAR-T refers to an unpublished dataset. BDAgent stands for BioDiscoveryAgent.

Tools Used	Schmidt1	Schmidt2	CAR-T	Scharen.	Carnev.	Sanchez
Random	0.037 \pm 0.013	0.031 \pm 0.002	0.033 \pm 0.003	0.160 \pm 0.028	0.036 \pm 0.001	0.034 \pm 0.004
BioDiscoveryAgent (Claude 3.5 Sonnet)						
No-Tools	0.095 \pm 0.004	0.104 \pm 0.004	0.130 \pm 0.010	0.326 \pm 0.034	0.042 \pm 0.003	0.066 \pm 0.007
Literature	0.096 \pm 0.005	0.098 \pm 0.010	0.138 \pm 0.019	0.309 \pm 0.041	0.042 \pm 0.002	0.069 \pm 0.006
AI Critic	0.088 \pm 0.004	0.092 \pm 0.010	0.126 \pm 0.019	0.309 \pm 0.037	0.042 \pm 0.002	0.059 \pm 0.006
Gene Search	0.096 \pm 0.003	0.100 \pm 0.009	0.123 \pm 0.009	0.348 \pm 0.025	0.043 \pm 0.004	0.062 \pm 0.007
All-Tools	0.096 \pm 0.005	0.090 \pm 0.003	0.121 \pm 0.020	0.234 \pm 0.080	0.043 \pm 0.001	0.054 \pm 0.004
BioDiscoveryAgent (Claude 3 Haiku)						
No-Tools	0.064 \pm 0.005	0.072 \pm 0.018	0.116 \pm 0.014	0.209 \pm 0.030	0.032 \pm 0.004	0.056 \pm 0.005
Literature	0.053 \pm 0.005	0.069 \pm 0.011	0.091 \pm 0.031	0.164 \pm 0.068	0.035 \pm 0.006	0.054 \pm 0.009
AI Critic	0.061 \pm 0.009	0.070 \pm 0.009	0.113 \pm 0.012	0.219 \pm 0.046	0.043 \pm 0.006	0.054 \pm 0.009
Gene Search	0.080 \pm 0.013	0.098 \pm 0.025	0.114 \pm 0.016	0.249 \pm 0.067	0.046 \pm 0.006	0.065 \pm 0.009
All-Tools	0.084 \pm 0.006	0.099 \pm 0.019	0.128 \pm 0.031	0.259 \pm 0.039	0.043 \pm 0.008	0.058 \pm 0.011

Table 8: **Agent performance improvements when using different tools for 1-gene perturbation experiments.** Results show hit ratio for experimental round 5 averaged over 10 runs, with the error intervals. See Table 1 caption for notes on specific datasets.

Tools Used	Schmidt1	Schmidt2	CAR-T	Scharen.	Carnev.	Sanchez
Random	0.037	0.031	0.033	0.160	0.036	0.034
Claude 3 Haiku						
No-Tools	0.064	0.072	0.116	0.209	0.032	0.056
All-Tools	0.084 (+31%)	0.099 (+38%)	0.128 (+10%)	0.259 (+24%)	0.043 (+34%)	0.058 (+4%)
GPT-3.5-Turbo						
No-Tools	0.044	0.061	0.064	0.230	0.032	0.039
All-Tools	0.062 (+41%)	0.100 (+64%)	0.063 (-2%)	0.218 (-5%)	0.037 (+16%)	0.045 (+15%)
Claude v1						
No-Tools	0.067	0.089	0.110	0.292	0.038	0.053
All-Tools	0.095 (+42%)	0.122 (+37%)	0.114 (+4%)	0.333 (+14%)	0.054 (+42%)	0.058 (+9%)
GPT o1-mini						
No-Tools	0.081	0.096	0.127	0.279	0.041	0.074
All-Tools	0.083 (+2%)	0.075 (-22%)	0.114 (-10%)	0.264 (-5%)	0.041 (+0%)	0.072 (-3%)
Claude 3 Sonnet						
No-Tools	0.076	0.088	0.115	0.302	0.041	0.064
All-Tools	0.074 (-3%)	0.091 (+3%)	0.105 (-9%)	0.302 (+0%)	0.047 (+15%)	0.073 (+14%)
Claude 3.5 Sonnet						
No-Tools	0.095	0.104	0.130	0.326	0.042	0.066
All-Tools	0.096 (+1%)	0.090 (-13%)	0.121 (-6%)	0.234 (-28%)	0.043 (+2%)	0.054 (-18%)
GPT-4o						
No-Tools	0.059	0.074	0.116	0.311	0.037	0.030
All-Tools	0.049 (-17%)	0.064 (-14%)	0.103 (-11%)	0.285 (-8%)	0.035 (-5%)	0.026 (-13%)
GPT o1-preview						
No-Tools	0.081	0.091	0.141	0.283	0.041	0.068
All-Tools	0.052 (-36%)	0.079 (-13%)	0.098 (-30%)	0.289 (+2%)	0.024 (-41%)	0.031 (-54%)
Claude 3 Opus						
No-Tools	0.094	0.104	0.126	0.247	0.043	0.059
All-Tools	0.096 (+2%)	0.101 (-3%)	0.116 (-8%)	0.291 (+18%)	0.041 (-5%)	0.055 (-7%)

Table 9: **Effect of tools on different LLMs.** Results show hit ratio for experimental round 5 averaged over 10 runs, with error intervals showing 1 standard deviation. *For (Scharenberg et al., 2023), a batch size of 32 was used due to its smaller pool of 1061 relevant genes. Schmidt1 refers to the screen measuring Interferon- γ (IFNG) and Schmidt2 measures Interleukin-2 (IL-2) following perturbation (Schmidt et al., 2022). [†]CAR-T refers to an unpublished dataset.

Model	Dataset	Avg. Input Tokens	Avg. Cost per Trial (\$/trial)
Claude 3.5 Sonnet (No-Tool)	Schmidt1	130409	0.61
	Schmidt2	120494	0.58
	CAR-T	99867	0.49
	Scharen.	92594	0.59
	Carnev.	126206	0.60
	Sanchez	88802	0.42
Claude 3.5 Sonnet (All-Tools)	Schmidt1	474648	2.38
	Schmidt2	440566	2.26
	CAR-T	386379	1.96
	Scharen.	314562	2.18
	Carnev.	415469	2.14
	Sanchez	351780	1.70
Claude 3 Haiku (No-Tool)	Schmidt1	271173	0.47
	Schmidt2	269938	0.47
	CAR-T	264463	0.45
	Scharen.	89393	0.19
	Carnev.	263726	0.46
	Sanchez	230232	0.37
Claude 3 Haiku (All-Tools)	Schmidt1	347435	0.61
	Schmidt2	353799	0.61
	CAR-T	372581	0.67
	Scharen.	103950	0.23
	Carnev.	353511	0.61
	Sanchez	289256	0.51

Table 10: Token Usage and API Cost for Gene Perturbation by Model and Dataset

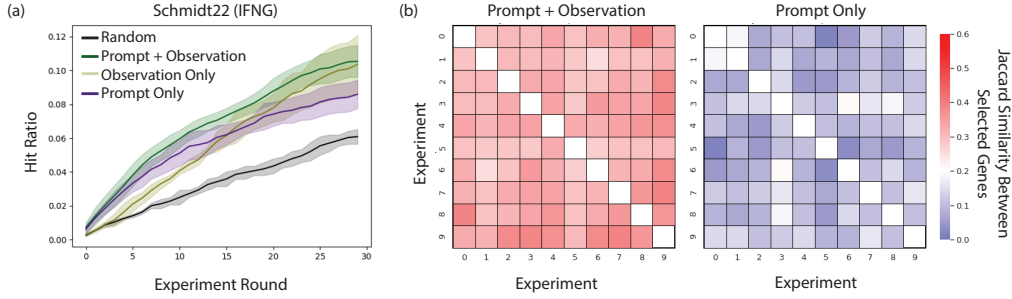


Figure 8: **Agent accounts for both prior knowledge and observations in decision-making** Three scenarios are considered: BioDiscoveryAgent (Claude v1) has access to task description and experimental observations (Prompt + Observation); the agent only has access to experimental observations (Observation Only); the agent only has access to the task description (Prompt Only). (a) Hit ratio at each experimental round across 30 rounds of experiments. 32 genes predicted per round. Each line corresponds to the average over 10 runs with error bars representing 1 standard deviation. (b) Jaccard similarity index between all predicted genes at 10 rounds of experimentation. Each cell corresponds to a different model run.

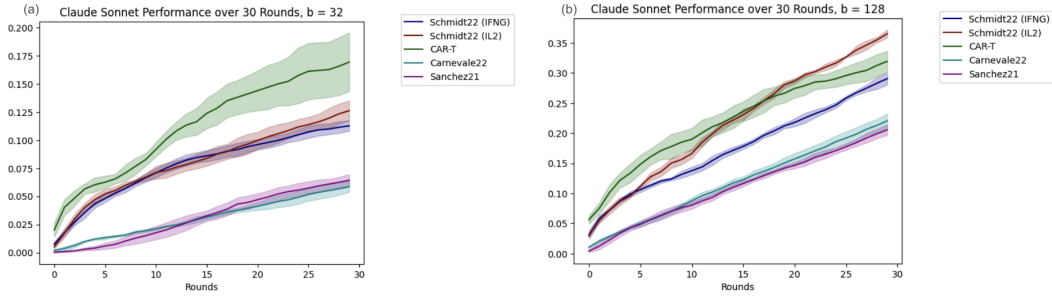


Figure 9: **Agent shows consistent performance for the first 30 rounds** BioDiscoveryAgent (Claude 3.5 Sonnet) is tested in two different settings to observe its performance over time: hit ratio at each experimental round across 30 rounds of experiments with (a) 32 genes predicted per round and (b) 128 genes per round. Each line corresponds to the average over 5 runs with error bars representing 1 standard deviation. In both cases, we observe approximately linear increase in hit rate, indicating that the agent performance is consistent throughout extended number of rounds.

Dataset		Schmidt22 (IL2)			Schmidt22 (IFNG)		
Rounds		10	20	30	10	20	30
Model	Setting	Avg. Hit Ratio					
BioDiscoveryAgent	Observ. Only	0.0449	0.0872	0.1037	0.0291	0.0702	0.1061
BioDiscoveryAgent	Prompt Only	0.0605	0.0846	0.1071	0.0484	0.0727	0.0872
BioDiscoveryAgent	Prompt + Observ.	0.0596	0.0908	0.1143	0.0559	0.0842	0.1077

Table 11: **Agent uses both prior knowledge and experimental observations to make predictions.** Three scenarios: BioDiscoveryAgent (Claude v1) has access to task description and experimental observations (Prompt + Observation); the agent only has access to experimental observations (Observation Only); the agent only has access to the task description (Prompt Only). (a) Hit ratio at each experimental round across 30 rounds of experiments. 32 genes predicted per round across 30 rounds.

G COMPLETE AGENT TRACE

In this section, we provide a few initial steps of the agent run equipped with all tools, and prompts seen by the agent at each step, as well as the agent’s response. We used BioDiscoveryAgent (Claude v1) for this analysis since it shows high absolute score as well as high performance gain through tool use (Appendix Table 9).

```

=====prompt=====

Human: Step 0
  You are a scientist working on problems in drug discovery.

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

Here is one strategy you can follow: Try to first focus a lot on
  trying very diverse genes to
  get a sense of which types of genes affect the research problem
  the most. From the observations,
  update your beliefs quickly and smartly and then double down on
  genes which you think shall be hits
  in order to increase the cumulative hit rate.

Always respond in this format exactly:

1. Reflection: Thoughts on previous results and next steps.
2. Research Plan: The full high level research plan, with current
  status and reasoning behind each proposed approach. It should
  be at most 5 sentences.
3. Gene Search: Name a gene to search for 50 most dissimilar genes
  based on features. Only include the gene name itself after
  "2. Gene Search:".
4. Solution: Propose a list of predicted genes to test separated
  by commas in this format: 1. <Gene name 1>, 2. <Gene name 2>
  ...
Do not include any genes from this prompt (since they’re already
  tested).

Now let’s start!

  You have done some literature review till now and have the
  following information at your disposal which you may use to
  make your predictions:

Ubiquitin ligase STUB1 destabilizes IFNG-receptor complex to
  suppress tumor IFNG signaling.
  Here is a summary of information that can be extracted from the
  provided file to help with the research problem:

```

1512 The file mentions that STUB1 is an E3 ubiquitin ligase that
 1513 mediates proteasomal degradation of the IFNG-R1/JAK1 complex (
 1514 lines 2-3). This suggests that perturbing STUB1 may increase
 1515 IFNG production.
 1516

1517 The file also refers to a genome-wide CRISPR/Cas9 screen that
 1518 identified regulators of IFNG-R1 cell surface abundance (line
 1519 2). Running a similar screen focused on IFNG production may
 1520 identify critical genes to perturb.

1521 In this segment, I cannot find specific gene targets or pathways
 1522 to focus on for the initial experiments. The provided
 1523 information is too high-level.
 1524

1525 The strategy suggested in the research problem to try diverse
 1526 genes first and then focus on hits is reasonable given the
 1527 limited number of perturbations possible. Updating beliefs
 1528 based on initial observations will help identify the most
 1529 promising genes to study further.
 1530

1531 Mitochondrial respiration contributes to the interferon gamma
 1532 response in antigen-presenting cells.
 1533 Here is a summary of the key information from the passage:
 1534

1535 The passage describes a genome-wide CRISPR screen in murine
 1536 macrophages to identify regulators of IFNG-inducible proteins
 1537 MHCII, CD40 and PD-L1. The screen found that complex I of the
 1538 mitochondrial respiratory chain is involved in the expression
 1539 of all three proteins and IFNG signaling (lines 3-7).
 1540

1541 Based on this, I would focus initial experiments on perturbing
 1542 genes involved in mitochondrial respiration and the electron
 1543 transport chain, specifically complex I. Perturbing these
 1544 genes could identify regulators of IFNG and its downstream
 1545 targets. Once initial hits are found, I would follow up by
 1546 perturbing more genes in those pathways to increase the
 1547 cumulative hit rate, as suggested in the research problem
 1548 description.

1549 In this segment, I cannot find specific genes to target or exact
 1550 numbers of genes that can be perturbed. The passage only
 1551 refers to a "series of genome-wide knockout screens" and "
 1552 multiscreen approach" (lines 3-4).
 1553

1554 Genome-Wide CRISPR Screen Reveals Cancer Cell Resistance to NK
 1555 Cells Induced by NK-Derived IFN-G.
 1556 Based on the information provided, here are some insights I can
 1557 gather:
 1558

1559 The research aims to identify genes that regulate IFNG production
 1560 using a CRISPR screen. The genes can be perturbed in batches
 1561 of 128 at a time out of 18,939 total genes. The log fold
 1562 change in IFNG counts can be measured for each perturbation to
 1563 determine scores. Given the large search space and limited
 1564 rounds of experimentation, an effective strategy is to first
 1565 try diverse genes to determine which types most impact IFNG

production. Then focus on the most promising genes to increase the cumulative hit rate.

In lines 1-2, the research describes using a "genome-wide CRISPR screen" and "primary human NK cells" to identify "genes that regulate the vulnerability of leukemia cells". This indicates the genes of interest are involved in the immune system and response to NK cells.

In line 3, "loss of" indicates the genes enable resistance to NK cells, so the genes of interest may be involved in susceptibility to NK cytotoxicity.

Based on the additional literature review, the genes of interest are likely involved in the production of IFNG, an immune signaling molecule. The screen aims to determine which of the 18,939 genes regulate IFNG levels when perturbed.

In this segment, I cannot find specific gene names or pathways referenced in the provided text. More details on the experimental procedure and results would be needed to determine promising genes or gene types to prioritize.

The key is to start broad, then use observations to focus on the most promising areas. An iterative approach updating beliefs and re-prioritizing genes is suggested to increase the cumulative discovery of genes impacting IFNG regulation.

Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR. Here is a summary of the information provided:

The research aims to overexpress long noncoding RNAs (lncRNAs) in vitro using the CRISPR activation technique in order to study their role in disease. As an example, the lncRNA IFNG-AS1 associated with inflammatory bowel disease is overexpressed in Jurkat T cells. The CRISPR activation technique targets transcription factors to the transcriptional start site of a gene to enable overexpression of lncRNA splice forms. This procedure involves guide RNA design, virus generation and transduction, and colony screening. Overexpression of IFNG-AS1 by greater than 20-fold was observed in Jurkat T cells.

Based on the research problem:

In this segment, I cannot find specific genes that regulate IFNG production. The file describes a general technique to overexpress lncRNAs using CRISPR activation but does not specify genes that regulate IFNG.

The strategy to first test diverse genes to determine which types are most impactful and then focus on likely "hits" is reasonable. Updating beliefs based on initial observations and doubling down on promising genes can increase the cumulative hit rate.

The literature review indicates the CRISPR activation technique can be used to overexpress lncRNAs to study their role in disease. Overexpression of the lncRNA IFNG-AS1, associated

with inflammatory bowel disease, was achieved in Jurkat T cells. This provides a proof of concept for the technique in the context of the research problem.

In summary, the file describes a useful technique to systematically overexpress lncRNAs which could be applied to identify genes regulating IFNG production. The example of overexpressing IFNG-AS1 in Jurkat T cells demonstrates the potential of this approach. The proposed strategy of testing diverse genes and focusing on likely hits is reasonable for this type of genome-wide screen.

=====

Assistant:

=====claude-1 response (2000)=====

1. Reflection: The literature review provides useful context on techniques and pathways that may be involved in regulating IFNG production. However, more details on specific genes and experimental results are needed to determine the most promising candidates for initial perturbations. The strategy to start broad and focus on likely hits based on observations is reasonable given the large search space.
2. Research Plan: Run an initial set of diverse perturbations across pathways potentially involved in IFNG regulation based on literature review. Measure IFNG levels for each to determine scores and identify promising gene types. Update beliefs and select most promising genes within those types for follow-up perturbations to increase cumulative hit rate.
3. Gene Search: MAPK1
4. Solution: 1. DAB2IP, 2. ARHGAP44, 3. VAMP5, 4. ATXN3L, 5. ZNF157, 6. ZNF880, 7. TBC1D19, 8. OR51L1, 9. MAPK1, 10. BIN3, 11. VGLL1, 12. TMCO2, 13. FOXD4L4, 14. SH2D1A, 15. P2RY4, 16. GUCY2F, 17. IGFL1, 18. IFNA10, 19. TVP23B, 20. ZNF182, 21. SLC01B7, 22. CLEC7A, 23. GLYATL2, 24. ZNF614, 25. AKAIN1, 26. TTC39B, 27. SPred1, 28. FAM83B, 29. ABL2, 30. VSIG4, 31. GOLT1A, 32. MGST1, 33. POTE, 34. ZNF35, 35. PPEF1, 36. UBE2V1, 37. ZNF223, 38. GAGE10, 39. STX17, 40. FAM184B, 41. CCDC54, 42. AP1S2, 43. PNMA3, 44. EQTN, 45. ZNF611, 46. IFNB1, 47. NAALADL2, 48. MAGEC3, 49. CFAP47, 50. KLRC2, 51. COL4A6, 52. BEX1, 53. POF1B, 54. CLTRN, 55. TMEM31, 56. RIPPLY1, 57. KRBOX4, 58. COL4A5, 59. LUZP4, 60. UGT2B28, 61. SPANXN4, 62. HMGN5, 63. PAGE4, 64. CYSLTR1, 65. PGAM4, 66. CLCN5, 67. GPR50, 68. USP27X, 69. OR52N5, 70. MID1, 71. PAGE1, 72. PCDH11X, 73. RTL4, 74. OR4P4, 75. PCDHA10, 76. SPANXN5, 77. COX7A1, 78. CFHR3, 79. AKAP4, 80. DCX, 81. BMP15, 82. ZNF334, 83. CCNB3, 84. SIGLEC5, 85. PLP1, 86. SLITRK4, 87. GLRA4, 88. MAGEC1, 89. MAVS, 90. IRF1, 91. IFIH1, 92. BCL10, 93. CHUK, 94. IKBKB, 95. MAP3K8, 96. AIM2, 97. CASP1, 98. IKBKE, 99. RELA, 100. STAT1, 101. PYHIN1, 102. GZMB, 103. TBK1, 104. PYCARD, 105. GZMK, 106. STING1, 107. TYK2, 108. TAB2, 109. NLRC4, 110. NLRP3, 111. JAK2, 112. MAP3K14, 113. MAP2K4, 114. IRF3, 115. GZMA, 116. MAP3K7, 117. IL2RG, 118. NFKB1, 119. NLRP1, 120.

PRF1, 121. CARD9, 122. MALT1, 123. TRAF6, 124. GDPD2 125.
ZNF638, 126. CHMP5, 127. ADAT2, 128. VPS37A

=====LLM Critique prompt=====

Human: You are a scientist working on problems in drug discovery.

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

Now for the next round of experiment your students are planning on testing the following genes:

['DAB2IP', 'ARHGAP44', 'VAMP5', 'ATXN3L', 'ZNF157', 'ZNF880', 'TBC1D19', 'OR51L1', 'MAPK1', 'BIN3', 'VGLL1', 'TMCO2', 'FOXO4L4', 'SH2D1A', 'P2RY4', 'GUCY2F', 'IGFL1', 'IFNA10', 'TVP23B', 'ZNF182', 'SLCO1B7', 'CLEC7A', 'GLYATL2', 'ZNF614', 'AKAIN1', 'TTC39B', 'SPred1', 'FAM83B', 'ABL2', 'VSIG4', 'GOLT1A', 'MGST1', 'POTEE', 'ZNF35', 'PPEF1', 'UBE2V1', 'ZNF223', 'GAGE10', 'STX17', 'FAM184B', 'CCDC54', 'AP1S2', 'PNMA3', 'EQTN', 'ZNF611', 'IFNB1', 'NAALADL2', 'MAGEC3', 'CFAP47', 'KLRC2', 'COL4A6', 'BEX1', 'POF1B', 'CLTRN', 'TMEM31', 'RIPPLY1', 'KRBOX4', 'COL4A5', 'LUZP4', 'UGT2B28', 'SPANXN4', 'HMG5', 'PAGE4', 'CYSLTR1', 'PGAM4', 'CLCN5', 'GPR50', 'USP27X', 'OR52N5', 'MID1', 'PAGE1', 'PCDH11X', 'RTL4', 'OR4P4', 'PCDHA10', 'SPANXN5', 'COX7A1', 'CFHR3', 'AKAP4', 'DCX', 'BMP15', 'ZNF334', 'CCNB3', 'SIGLEC5', 'PLP1', 'SLITRK4', 'GLRA4', 'MAGEC1', 'MAVS', 'IRF1', 'IFIH1', 'BCL10', 'CHUK', 'IKBKB', 'MAP3K8', 'AIM2', 'CASP1', 'IKBKE', 'RELA', 'STAT1', 'PYHIN1', 'GZMB', 'TBK1', 'PYCARD', 'GZMK', 'STING1', 'TYK2', 'TAB2', 'NLRC4', 'NLRP3', 'JAK2', 'MAP3K14', 'MAP2K4', 'IRF3', 'GZMA', 'MAP3K7', 'IL2RG', 'NFKB1', 'NLRP1', 'PRF1', 'CARD9', 'MALT1', 'TRAF6', 'GDPD2', 'ADAT2', 'ZNF638', 'VPS37A', 'CHMP5', 'ADAT2', 'VPS37A']

As an advisor, please critique this plan and suggest some changes to it. Use this format:

1. Critique: include all relevant details of the critique.
2. Updated Solution: Give an updated selection of 128 genes based on the critique separated by commas in this format:: 1. <Gene name 1>, 2. <Gene name 2> ...

Try to first focus a lot on trying very diverse genes to get a sense of which types of genes affect the research problem the most. From the observations, update your beliefs quickly and smartly and then double down on genes which you think shall be hits in order to increase the cumulative hit rate.

Please do not critique/make a lot of changes if there is no need to make a change.

LLM Critic Assistant:

=====claude-1 response (2000)=====

1. Critique: The selected genes seem to focus heavily on genes involved in the interferon response and immune signaling pathways. While these are good candidates, it may be beneficial to include a more diverse set of genes to get a broader sense of the types of genes that regulate IFNG production. Some other categories to consider include: transcription factors, epigenetic regulators, cell cycle regulators, metabolic genes, etc. Focusing too narrowly on a set of genes may lead to missing other important regulators.

2. Updated Solution:

1. DAB2IP,
2. ARHGAP44,
3. VAMP5,
4. ATXN3L,
5. ZNF157,
6. ZNF880,
7. TBC1D19,
8. OR51L1,
9. MAPK1,
10. BIN3,
11. VGLL1,
12. TMCO2,
13. FOXD4L4,
14. SH2D1A,
15. P2RY4,
16. GUCY2F,
17. IGFL1,
18. IFNA10,
19. TVP23B,
20. ZNF182,
21. SLC01B7,
22. CLEC7A,
23. GLYATL2,
24. ZNF614,
25. AKAIN1,
26. TTC39B,
27. SPred1,
28. FAM83B,
29. ABL2,
30. VSIG4,
31. GOLT1A,
32. MGST1,
33. POTE1,
34. ZNF35,
35. PPEF1,
36. UBE2V1,
37. ZNF223,
38. GAGE10,
39. STX17,
40. FAM184B,

1782	41. CCDC54,
1783	42. AP1S2,
1784	43. PNMA3,
1785	44. EQTN,
1786	45. ZNF611,
1787	46. IFNB1,
1788	47. NAALADL2,
1789	48. MAGEC3,
1790	49. CFAP47,
1791	50. KLRC2,
1792	51. COL4A6,
1793	52. BEX1,
1794	53. POF1B,
1795	54. CLTRN,
1796	55. TMEM31,
1797	56. RIPPLY1,
1798	57. KRBOX4,
1799	58. COL4A5,
1800	59. LUZP4,
1801	60. UGT2B28,
1802	61. SPANXN4,
1803	62. HMGN5,
1804	63. PAGE4,
1805	64. CYSLTR1,
1806	65. PGAM4,
1807	66. CLCN5,
1808	67. GPR50,
1809	68. USP27X,
1810	69. OR52N5,
1811	70. MID1,
1812	71. PAGE1,
1813	72. PCDH11X,
1814	73. RTL4,
1815	74. OR4P4,
1816	75. PCDHA10,
1817	76. SPANXN5,
1818	77. COX7A1,
1819	78. CFHR3,
1820	79. AKAP4,
1821	80. DCX,
1822	81. BMP15,
1823	82. ZNF334,
1824	83. CCNB3,
1825	84. SIGLEC5,
1826	85. PLP1,
1827	86. SLITRK4,
1828	87. GLRA4,
1829	88. MAGEC1,
1830	89. MAVS,
1831	90. IRF1,
1832	91. IFIH1,
1833	92. BCL10,
1834	93. CHUK,
1835	94. IKBKB,
	95. MAP3K8,
	96. AIM2,
	97. CASP1,
	98. IKBKE,
	99. RELA,

```

1836 100. STAT1,
1837 101. PYHIN1,
1838 102. GZMB,
1839 103. TBK1,
1840 104. PYCARD,
1841 105. GZMK,
1842 106. STING1,
1843 107. TYK2,
1844 108. TAB2,
1845 109. NLRC4,
1846 110. NLRP3,
1847 111. JAK2,
1848 112. MAP3K14,
1849 113. MAP2K4,
1850 114. IRF3,
1851 115. GZMA,
1852 116. MAP3K7,
1853 117. IL2RG,
1854 118. NFKB1,
1855 119. NLRP1,
1856 120. PRF1,
1857 121. CARD9,
1858 122. MALT1,
1859 123. TRAF6,
1860 124. GSDMD,
1861 125. ADAT2,
1862 126. ZNF638,
1863 127. VPS37A,
1864 128. CHMP5
1865 =====
1866 =====prompt=====
1867
1868 Human: Step 1
1869 You are a scientist working on problems in drug discovery.
1870
1871 Research Problem: I'm planning to run a genome-wide CRISPR screen
1872 to identify genes that regulate the production of Interferon-
1873 gamma (IFNG). There are 18,939 possible genes to perturb and
1874 I can only perturb 128 genes at a time. For each perturbation,
1875 I'm able to measure out the log fold change in Interferon-
1876 gamma (IFNG) normalized read counts which will be referred to
1877 as the score. I can only do a few rounds of experimentation.
1878
1879 Here is one strategy you can follow: Try to first focus a lot on
1880 trying very diverse genes to
1881 get a sense of which types of genes affect the research problem
1882 the most. From the observations,
1883 update your beliefs quickly and smartly and then double down on
1884 genes which you think shall be hits
1885 in order to increase the cumulative hit rate.
1886
1887 Always respond in this format exactly:
1888 1. Reflection: Thoughts on previous results and next steps.
1889

```

1890 2. Research Plan: The full high level research plan, with current
1891 status and reasoning behind each proposed approach. It should
1892 be at most 5 sentences.

1893 3. Gene Search: Name a gene to search for 50 most dissimilar genes
1894 based on features. Only include the gene name itself after
1895 "2. Gene Search:".

1896 4. Solution: Propose a list of predicted genes to test separated
1897 by commas in this format: 1. <Gene name 1>, 2. <Gene name 2>
1898 ...

1899 Do not include any genes from this prompt (since they're already
1900 tested).

1901

1902 Now let's start!

1903

1904

1905 This is not your first round. All tested genes and their measured
1906 log fold change are:

1907 Score

1908 Gene

1909 ZNF880 0.021900

1910 CFAP47 0.075213

1911 BIN3 -0.095469

1912 COL4A6 -0.154264

1913 P2RY4 -0.053394

1914 GUCY2F -0.075052

1915 GZMB 0.090265

1916 CLTRN 0.031790

1917 IGFL1 0.126201

1918 TBK1 0.013595

1919 PYCARD 0.093615

1920 AKAIN1 0.220689

1921 UGT2B28 0.032370

1922 TTC39B -0.308255

1923 GZMK 0.058316

1924 CLCN5 -0.112738

1925 GPR50 0.329583

1926 TYK2 -0.084040

1927 MGST1 0.026777

1928 TAB2 0.367535

1929 OR52N5 0.193478

1930 NLRP3 -0.065815

1931 MID1 0.360340

1932 ARHGAP44 -0.008040

1933 PCDH11X 0.036595

1934 ZNF35 -0.145104

1935 RTL4 0.035977

1936 IRF3 -0.237845

1937 PCDHA10 -0.165770

1938 CFHR3 -0.142490

1939 BMP15 -0.096454

1940 FAM184B -0.069565

1941 NLRP1 0.000321

1942 PLP1 -0.163710

1943 CARD9 0.117949

GLRA4 -0.133498

NAALADL2 0.244136

ATXN3L 0.356180

ZNF638 -0.156016

1944	TBC1D19	-0.330860
1945	OR51L1	0.201095
1946	VGLL1	0.008220
1947	ADAT2	0.079780
1948	PYHIN1	0.152815
1949	IFNA10	-0.097709
1950	TVP23B	-0.043965
1951	ZNF182	0.000000
1952	KRBOX4	-0.299955
1953	PAGE4	-0.210080
1954	PAGE1	-0.194869
1955	ZNF223	-0.026465
1956	CCNB3	0.015305
1957	SIGLEC5	0.090420
1958	PNMA3	-0.004561
1959	EQTN	0.003077
1960	SLITRK4	-0.033655
1961	TMEM31	0.068895
1962	IFNB1	0.023320
1963	MAGEC1	0.212718
1964	KLRC2	-0.146425
1965	MAVS	0.264290
1966	IRF1	-0.029467
1967	VAMP5	-0.226835
1968	POF1B	-0.006635
1969	ZNF614	-0.054910
1970	LUZP4	-0.184250
1971	SPANXN4	0.276185
1972	SPred1	0.126492
1973	FAM83B	0.082410
1974	VSIG4	0.036635
1975	STING1	-0.128400
1976	GOLT1A	0.077744
1977	NLRC4	-0.115631
1978	OR4P4	0.061845
1979	GAGE10	0.185560
1980	COX7A1	-0.246813
1981	STX17	0.198955
1982	DCX	-0.011380
1983	ZNF334	-0.040525
1984	AP1S2	-0.018804
1985	MAGEC3	-0.264931
1986	TMCO2	-0.120375
1987	FOXD4L4	-0.336970
1988	BEX1	-0.336525
1989	MAP3K8	-0.181561
1990	CASP1	-0.064096
1991	AIM2	-0.024550
1992	ZNF611	0.282510
1993	STAT1	-0.023106
1994	RIPPLY1	0.242648
1995	SLCO1B7	-0.010360
1996	CLEC7A	-0.244316
1997	GLYATL2	-0.085205
	COL4A5	0.002101
	HMG5	-0.262160
	PGAM4	-0.083230
	ABL2	0.113530
	USP27X	0.133827

1998 JAK2 0.132444
 1999 POTE -0.087225
 2000 MAP3K14 -0.366105
 2001 MAP2K4 -0.024708
 2002 ZNF157 -0.126108
 2003 PPEF1 0.133363
 2004 UBE2V1 -0.025857
 2005 GZMA 0.077647
 2006 SPANXN5 -0.133735
 2007 AKAP4 -0.209265
 2008 NFkB1 0.125557
 2009 CCDC54 -0.058695
 2010 DAB2IP -0.271045
 2011 PRF1 0.161566
 2012 CYSLTR1 -0.092562
 2012 GDPD2 0.018330
 2013 You have successfully identified 14 hits so far over all
 2014 experiment cycles! The results for the hits are:
 2015 Score
 2016 Gene
 2017 IKBKE 0.487890
 2018 MAP3K7 -1.305750
 2019 MAPK1 -0.449405
 2020 IL2RG -0.650145
 2021 VPS37A 0.488510
 2022 IFIH1 0.451861
 2023 SH2D1A -0.903270
 2024 BCL10 -0.869020
 2025 IKBKB -0.922680
 2026 CHUK -0.833950
 2027 CHMP5 0.420150
 2028 RELA -0.840910
 2029 MALT1 -1.063740
 2030 TRAF6 -1.865400
 2031 Based on these results and prior knowledge of biology, make the
 2032 best possible prediction of the first 128 genes that I should
 2033 test to maximize the score. Use HGNC gene naming convention.
 2034 DO NOT Predict GENES THAT HAVE ALREADY BEEN TESTED
 2035 You have done some literature review till now and have the
 2036 following information at your disposal which you may use to
 2037 make your predictions:
 2038 Ubiquitin ligase STUB1 destabilizes IFNG-receptor complex to
 2039 suppress tumor IFNG signaling.
 2040 Here is a summary of information that can be extracted from the
 2041 provided file to help with the research problem:
 2042 The file mentions that STUB1 is an E3 ubiquitin ligase that
 2043 mediates proteasomal degradation of the IFNG-R1/JAK1 complex (line
 2044 2-3). This suggests that perturbing STUB1 may increase
 2045 IFNG production.
 2046 The file also refers to a genome-wide CRISPR/Cas9 screen that
 2047 identified regulators of IFNG-R1 cell surface abundance (line
 2048 2). Running a similar screen focused on IFNG production may
 2049 identify critical genes to perturb.
 2050
 2051

In this segment, I cannot find specific gene targets or pathways to focus on for the initial experiments. The provided information is too high-level.

The strategy suggested in the research problem to try diverse genes first and then focus on hits is reasonable given the limited number of perturbations possible. Updating beliefs based on initial observations will help identify the most promising genes to study further.

Mitochondrial respiration contributes to the interferon gamma response in antigen-presenting cells.

Here is a summary of the key information from the passage:

The passage describes a genome-wide CRISPR screen in murine macrophages to identify regulators of IFNG-inducible proteins MHCII, CD40 and PD-L1. The screen found that complex I of the mitochondrial respiratory chain is involved in the expression of all three proteins and IFNG signaling (lines 3-7).

Based on this, I would focus initial experiments on perturbing genes involved in mitochondrial respiration and the electron transport chain, specifically complex I. Perturbing these genes could identify regulators of IFNG and its downstream targets. Once initial hits are found, I would follow up by perturbing more genes in those pathways to increase the cumulative hit rate, as suggested in the research problem description.

In this segment, I cannot find specific genes to target or exact numbers of genes that can be perturbed. The passage only refers to a "series of genome-wide knockout screens" and "multiscreen approach" (lines 3-4).

Genome-Wide CRISPR Screen Reveals Cancer Cell Resistance to NK Cells Induced by NK-Derived IFN-G.

Based on the information provided, here are some insights I can gather:

The research aims to identify genes that regulate IFNG production using a CRISPR screen. The genes can be perturbed in batches of 128 at a time out of 18,939 total genes. The log fold change in IFNG counts can be measured for each perturbation to determine scores. Given the large search space and limited rounds of experimentation, an effective strategy is to first try diverse genes to determine which types most impact IFNG production. Then focus on the most promising genes to increase the cumulative hit rate.

In lines 1-2, the research describes using a "genome-wide CRISPR screen" and "primary human NK cells" to identify "genes that regulate the vulnerability of leukemia cells". This indicates the genes of interest are involved in the immune system and response to NK cells.

In line 3, "loss of" indicates the genes enable resistance to NK cells, so the genes of interest may be involved in susceptibility to NK cytotoxicity.

Based on the additional literature review, the genes of interest are likely involved in the production of IFNG, an immune signaling molecule. The screen aims to determine which of the 18,939 genes regulate IFNG levels when perturbed.

In this segment, I cannot find specific gene names or pathways referenced in the provided text. More details on the experimental procedure and results would be needed to determine promising genes or gene types to prioritize.

The key is to start broad, then use observations to focus on the most promising areas. An iterative approach updating beliefs and re-prioritizing genes is suggested to increase the cumulative discovery of genes impacting IFNG regulation.

Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR. Here is a summary of the information provided:

The research aims to overexpress long noncoding RNAs (lncRNAs) in vitro using the CRISPR activation technique in order to study their role in disease. As an example, the lncRNA IFNG-AS1 associated with inflammatory bowel disease is overexpressed in Jurkat T cells. The CRISPR activation technique targets transcription factors to the transcriptional start site of a gene to enable overexpression of lncRNA splice forms. This procedure involves guide RNA design, virus generation and transduction, and colony screening. Overexpression of IFNG-AS1 by greater than 20-fold was observed in Jurkat T cells.

Based on the research problem:

In this segment, I cannot find specific genes that regulate IFNG production. The file describes a general technique to overexpress lncRNAs using CRISPR activation but does not specify genes that regulate IFNG.

The strategy to first test diverse genes to determine which types are most impactful and then focus on likely "hits" is reasonable. Updating beliefs based on initial observations and doubling down on promising genes can increase the cumulative hit rate.

The literature review indicates the CRISPR activation technique can be used to overexpress lncRNAs to study their role in disease. Overexpression of the lncRNA IFNG-AS1, associated with inflammatory bowel disease, was achieved in Jurkat T cells. This provides a proof of concept for the technique in the context of the research problem.

In summary, the file describes a useful technique to systematically overexpress lncRNAs which could be applied to identify genes regulating IFNG production. The example of overexpressing IFNG-AS1 in Jurkat T cells demonstrates the potential of this approach. The proposed strategy of testing

diverse genes and focusing on likely hits is reasonable for this type of genome-wide screen.

Gene Search: MAPK1

Gene Search Result:DAB2IP, VAMP5, ZNF157, ATXN3L, ARHGAP44, TVP23B, SH2D1A, POTEE, GAGE10, VCX3B, FAM184B, ZNF611, FAM83B, OR51L1, GOLT1A, STX17, VSIG4, ZNF35, MGST1, AP1S2, SPred1, IGFL1, NLK, FOXD4L4, VGLL1, TBC1D19, AKAIN1, OR4F5, IFNB1, CLEC7A, OR52E6, CCDC54, GUCY2F, PNMA3, ZNF182, PPEF1, P2RY4, TTC39B, NAALADL2, GLYATL2, IFNA10, EQTN, ZNF223, ZNF614, TMCO2, BIN3, UBE2V1, ZNF880, ABL2, SLC01B7

Assistant:

=====claude-1 response (2000)=====

1. Reflection: Based on the results so far, genes involved in mitochondrial respiration and the electron transport chain appear promising to perturb. Focusing on these pathways is yielding hits, so continuing to perturb genes in these areas may identify more regulators of IFNG. The strategy of testing diverse genes then focusing on hits is working well.
2. Research Plan: Test diverse genes in the electron transport chain and mitochondrial respiration to identify regulators of IFNG. Once hits are found, focus on related genes to increase the cumulative hit rate. Update beliefs and re-prioritize genes based on observations.
3. Gene Search: CAT
4. Solution: SFT2D1, 2. FGF22, 3. GUCY1B1, 4. WNT10A, 5. CAT, 6. GAD2, 7. SYCE1L, 8. LEMD1, 9. BRD7, 10. SLC38A8, 11. ZNF181, 12. SHOC1, 13. CYP2U1, 14. MMP10, 15. MAPK12, 16. CDC7, 17. PRR3, 18. FAM81B, 19. MAPK10, 20. ANKRD50, 21. PSMC3, 22. PTPRCAP, 23. SLC17A3, 24. CD1B, 25. RFC2, 26. ZNF676, 27. SLC29A1, 28. AP3S2, 29. RPA4, 30. LSP1, 31. UNC5CL, 32. TAFA3, 33. MROH6, 34. PELP1, 35. TPTE2, 36. PSMB8, 37. ITPR2, 38. ACYP1, 39. APMAP, 40. NIF3L1, 41. USP48, 42. OR5D13, 43. CKAP2, 44. TBX5, 45. TBL1Y, 46. PLA2G15, 47. TRIM59, 48. NMNAT3, 49. PIK3R3, 50. CLSTN1, 51. FAM25C, 52. TPI1, 53. PGLS, 54. KEAP1, 55. AADAC, 56. DEDD, 57. POU2AF1, 58. INPP5F, 59. SNRPB, 60. CPNE9, 61. FGF8, 62. SERPINA11, 63. GJA3, 64. RAB2A, 65. CHST2, 66. TCF24, 67. RAD50, 68. RNF208, 69. EYA4, 70. ZNF174, 71. CENPL, 72. AKIRIN1, 73. FANCM, 74. ATF4, 75. UBE2M, 76. RIPK2, 77. MYRF, 78. PCSK1, 79. GPT2, 80. WDR25, 81. CRTC2, 82. SECISBP2L, 83. PMS1, 84. TGFB1, 85. ALDH7A1, 86. SOS2, 87. DUT, 88. TAC3, 89. FASTKD1, 90. TOR1AIP1, 91. RNLS, 92. GUCA1C, 93. ABCC5, 94. KLK9, 95. DDX42, 96. MAPK8IP2, 97. CCDC28A, 98. APOA5, 99. CNOT9, 100. MCAM, 101. ZFP36L2, 102. LANCE1, 103. EXOSC1, 104. ANP32A, 105. PAX9, 106. AQP5, 107. FZD1, 108. ZDHHC13, 109. STAT5A, 110. CDK11A, 111. NDUFB5, 112. GSTT2B, 113. CALD1, 114. ASS1, 115. KRT7, 116. CYP2C8, 117. BCAR3, 118. C14orf39, 119. THEMIS2, 120. KLHL9, 121. TWIST2, 122. CXCL14, 123. NPNT, 124. L2HGDH, 125. SLC25A15, 126. FBXL12, 127. MAPK8, 128. ZDHHC1

```

2214
2215 =====LLM critique prompt=====
2216
2217
2218 Human: You are a scientist working on problems in drug discovery.
2219
2220 Research Problem: I'm planning to run a genome-wide CRISPR screen
2221 to identify genes that regulate the production of Interferon-
2222 gamma (IFNG). There are 18,939 possible genes to perturb and
2223 I can only perturb 128 genes at a time. For each perturbation,
2224 I'm able to measure out the log fold change in Interferon-
2225 gamma (IFNG) normalized read counts which will be referred to
2226 as the score. I can only do a few rounds of experimentation.
2227
2228 All tested genes so far and their measured log fold change are:
2229 Score
2230 Gene
2231 ZNF880 0.021900
2232 CFAP47 0.075213
2233 BIN3 -0.095469
2234 COL4A6 -0.154264
2235 P2RY4 -0.053394
2236 GUCY2F -0.075052
2237 GZMB 0.090265
2238 CLTRN 0.031790
2239 IGFL1 0.126201
2240 TBK1 0.013595
2241 PYCARD 0.093615
2242 AKAIN1 0.220689
2243 UGT2B28 0.032370
2244 TTC39B -0.308255
2245 GZMK 0.058316
2246 CLCN5 -0.112738
2247 GPR50 0.329583
2248 TYK2 -0.084040
2249 MGST1 0.026777
2250 TAB2 0.367535
2251 OR52N5 0.193478
2252 NLRP3 -0.065815
2253 MID1 0.360340
2254 ARHGAP44 -0.008040
2255 PCDH11X 0.036595
2256 ZNF35 -0.145104
2257 RTL4 0.035977
2258 IRF3 -0.237845
2259 PCDHA10 -0.165770
2260 CFHR3 -0.142490
2261 BMP15 -0.096454
2262 FAM184B -0.069565
2263 NLRP1 0.000321
2264 PLP1 -0.163710
2265 CARD9 0.117949
2266 GLRA4 -0.133498
2267 NAALADL2 0.244136
2268 ATXN3L 0.356180
2269 ZNF638 -0.156016
2270 TBC1D19 -0.330860
2271 OR51L1 0.201095

```

2268	VGLL1	0.008220
2269	ADAT2	0.079780
2270	PYHIN1	0.152815
2271	IFNA10	-0.097709
2272	TVP23B	-0.043965
2273	ZNF182	0.000000
2274	KRBOX4	-0.299955
2275	PAGE4	-0.210080
2276	PAGE1	-0.194869
2277	ZNF223	-0.026465
2278	CCNB3	0.015305
2279	SIGLEC5	0.090420
2280	PNMA3	-0.004561
2281	EQTN	0.003077
2282	SLITRK4	-0.033655
2283	TMEM31	0.068895
2284	IFNB1	0.023320
2285	MAGEC1	0.212718
2286	KLRC2	-0.146425
2287	MAVS	0.264290
2288	IRF1	-0.029467
2289	VAMP5	-0.226835
2290	POF1B	-0.006635
2291	ZNF614	-0.054910
2292	LUZP4	-0.184250
2293	SPANXN4	0.276185
2294	SPred1	0.126492
2295	FAM83B	0.082410
2296	VSIG4	0.036635
2297	STING1	-0.128400
2298	GOLT1A	0.077744
2299	NLRC4	-0.115631
2300	OR4P4	0.061845
2301	GAGE10	0.185560
2302	COX7A1	-0.246813
2303	STX17	0.198955
2304	DCX	-0.011380
2305	ZNF334	-0.040525
2306	AP1S2	-0.018804
2307	MAGEC3	-0.264931
2308	TMCO2	-0.120375
2309	FOXD4L4	-0.336970
2310	BEX1	-0.336525
2311	MAP3K8	-0.181561
2312	CASP1	-0.064096
2313	AIM2	-0.024550
2314	ZNF611	0.282510
2315	STAT1	-0.023106
2316	RIPPLY1	0.242648
2317	SLCO1B7	-0.010360
2318	CLEC7A	-0.244316
2319	GLYATL2	-0.085205
2320	COL4A5	0.002101
2321	HMG5	-0.262160
	PGAM4	-0.083230
	ABL2	0.113530
	USP27X	0.133827
	JAK2	0.132444
	POTEE	-0.087225

2322 MAP3K14 -0.366105
 2323 MAP2K4 -0.024708
 2324 ZNF157 -0.126108
 2325 PPEF1 0.133363
 2326 UBE2V1 -0.025857
 2327 GZMA 0.077647
 2328 SPANXN5 -0.133735
 2329 AKAP4 -0.209265
 2330 NFKB1 0.125557
 2331 DAB2IP -0.271045
 2332 PRF1 0.161566
 2333 CYSLTR1 -0.092562
 2334 GDPD2 0.018330
 2335 The results for the hits are:
 2336 Score
 2337 Gene
 2338 IKBKE 0.487890
 2339 MAP3K7 -1.305750
 2340 MAPK1 -0.449405
 2341 IL2RG -0.650145
 2342 VPS37A 0.488510
 2343 IFIH1 0.451861
 2344 SH2D1A -0.903270
 2345 BCL10 -0.869020
 2346 IKBKB -0.922680
 2347 CHUK -0.833950
 2348 CHMP5 0.420150
 2349 RELA -0.840910
 2350 MALT1 -1.063740
 2351 TRAF6 -1.865400
 2352
 2353 Now for the next round of experiment your students are planning on
 2354 testing the following genes:
 2355 ['IPO9', 'MDM2', 'TERF1', 'DDX31', 'WDR89', 'DDX21', 'DCLRE1B', '
 2356 NOLC1', 'USP7', 'FERMT2', 'KIF18B', 'ADRM1', 'PSME3', 'NCL', '
 2357 DNTTIP2', 'TLN1', 'TAF1D', 'PPP4R2', 'KIF2C', 'NDUFA6', '
 2358 NDUFA10', 'NDUFA8', 'NDUFB4', 'COX6A1', 'NDUFA1', 'CYC1', '
 2359 NDUFA7', 'NDUFB5', 'NDUFA2', 'COX7A2', 'COX7B', 'NDUFB10', '
 2360 NDUFB3', 'NDUFB6', 'NDUFB9', 'NDUFB7', 'NDUFA3', 'NDUFA13', '
 2361 COX4I1', 'COX6B1', 'NDUFA11', 'NDUFA9', 'NDUFA12', 'NDUFC1', '
 2362 NDUFB11', 'NDUFC2', 'NDUFA5', 'UQCRCQ', 'UQCRB', 'UQCRC2', '
 2363 NDUFA4', 'COX5A', 'COX7C', 'NDUFB8', 'UQCRH', 'COX8A', 'OR52N4
 2364 ', 'MAGED2', 'GPR37', 'TTC29', 'PDLIM3', 'TTC23', 'PRAMEF4', '
 2365 MAPK9', 'SMCP', 'DMRT2', 'OR10Z1', 'SPRR1B', 'FAM3C', 'EVC2', '
 2366 'SSMEM1', 'ATAD1', 'C10orf71', 'MEIG1', 'NELL1', 'TRHR', '
 2367 TMPRSS11F', 'IFNA16', 'ARSG', 'ARMCX4', 'GPRASP1', 'HDX', '
 2368 ZNF582', 'KRTAP10-2', 'ZNF729', 'RD3L', 'NDUFS2', 'IFNE', '
 2369 CACNG7', 'ACRV1', 'PDLIM5', 'SCLT1', 'ASB11', 'TRIM73', '
 2370 NDUFS1', 'GIMAP4', 'NDUFS3', 'APLN', 'ACSM2B', 'RFPL4AL1', '
 2371 STEAP1', 'KRTAP4-12', 'ZNF664', 'SMOC2', 'SDHB', 'SDHA', '
 2372 NDUFV2', 'NDUFS4', 'UQCRC1', 'COX6C', 'NDUFV1', 'UQCRFS1', '
 2373 NDUFAF2', 'UQCR10', 'NDUFAF4', 'NDUFAF3', 'COX7A2L', 'NDUFAF1
 2374 ', 'NDUFB2', 'NDUFV3', 'SDHC', 'NDUFAF6', 'SDHD', 'ZNF669', '
 2375 CD200', 'NDUFS6', 'COX5B', 'CAT']
 2376
 2377 As an advisor, please critique this plan and suggest some changes
 2378 to it. Use this format:
 2379 1. Critique: include all relevant details of the critique.
 2380

2376 2. Updated Solution: Give an updated selection of 128 genes based
 2377 on the critique separated by commas in this format:: 1. <Gene
 2378 name 1>, 2. <Gene name 2> ...
 2379
 2380
 2381 Try to first focus a lot on trying very diverse genes to
 2382 get a sense of which types of genes affect the research problem
 2383 the most. From the observations,
 2384 update your beliefs quickly and smartly and then double down on
 2385 genes which you think shall be hits
 2386 in order to increase the cumulative hit rate.

2387 Please do not critique/make a lot of changes if there is no need
 2388 to make a change.
 2389

2390 =====
 2391

2392 LLM Critic Assistant:
 2393 =====claude-1 response (2000)=====

2394 Here are my critiques and suggestions for the next round of
 2395 experiment:

2396 1. Critique: The selected genes seem very focused on mitochondrial
 2397 and respiratory chain components. While these are important
 2398 for interferon production, testing so many of them may limit
 2399 the diversity of pathways explored. I would suggest balancing
 2400 these genes with some involved in other pathways important for
 2401 interferon regulation, such as NF-kB signaling, MAPK
 2402 signaling, and cytosolic DNA sensing.

2403

2404 2. Updated Solution:

2405 1. IPO9
 2406 2. MDM2
 2407 3. TERF1
 2408 4. DDX31
 2409 5. WDR89
 2410 6. DDX21
 2411 7. DCLRE1B
 2412 8. NOLC1
 2413 9. USP7
 2414 10. FERMT2
 2415 11. KIF18B
 2416 12. ADRM1
 2417 13. PSME3
 2418 14. NCL
 2419 15. DNTTIP2
 2420 16. TLN1
 2421 17. TAF1D
 2422 18. PPP4R2
 2423 19. KIF2C
 2424 20. NFKB1
 2425 21. CHUK
 2426 22. MAP3K7
 2427 23. MAPK1
 2428 24. DDX60
 2429 25. IFI16
 2430 26. ZBP1
 2431 27. DHX9
 2432 28. DHX36

2430	29. MAVS
2431	30. STING1
2432	31. TBK1
2433	32. IRF3
2434	33. IRF7
2435	34. CYLD
2436	35. OTUB1
2437	36. HERC5
2438	37. HERC6
2439	38. USP18
2440	39. NLRC3
2441	40. NLRP3
2442	41. AIM2
2443	42. NLRP1
2444	43. NLRP6
2445	44. NLRP12
2446	45. CASP1
2447	46. PYCARD
2448	47. BIRC2
2449	48. BIRC3
2450	49. XIAP
2451	50. BCL10
2452	51. MALT1
2453	52. TRAF3
2454	53. TRAF6
2455	54. TAK1
2456	55. IKKA
2457	56. IKKB
2458	57. CHMP5
2459	58. RELA
2460	59. STAT1
2461	60. STAT2
2462	61. JAK1
2463	62. TYK2
2464	63. IRF9
2465	64. ISG15
2466	65. IFIT1
2467	66. IFIT3
2468	67. OAS1
2469	68. OAS2
2470	69. OAS3
2471	70. ADAR
2472	71. PKR
2473	72. RNaseL
2474	73. MX1
2475	74. EIF2AK2
2476	75. DDX58
2477	76. DHX58
2478	77. IFIH1
2479	78. ZBP1
2480	79. RIG-I
2481	80. MDA5
2482	81. LGP2
2483	82. Ku70
	83. PRKDC
	84. DNA-PKcs
	85. ATM
	86. ATR
	87. CHD4

2484	88. ALC1
2485	89. PARP9
2486	90. PARP14
2487	91. DDX60L
2488	92. IFI16
2489	93. cGAS
2490	94. MB21D1
2491	95. MB21D2
2492	96. ABCE1
2493	97. DHX9
2494	98. DHX36
2495	99. DDX1
2496	100. DDX21
2497	101. DDX31
2498	102. DDX41
2499	103. DDX42
2500	104. DDX43
2501	105. DDX46
2502	106. DDX47
2503	107. DDX48
2504	108. DDX49
2505	109. DDX50
2506	110. DDX51
2507	111. DDX52
2508	112. DDX53
2509	113. DDX54
2510	114. DDX55
2511	115. DDX56
2512	116. DDX59
2513	117. DDX60
2514	118. DHX15
2515	119. DHX30
2516	120. DHX32
2517	121. DHX33
2518	122. DHX34
2519	123. DHX35
2520	124. DHX37
2521	125. DHX38
2522	126. DHX57
2523	127. DHX8
2524	128. DDX3X
2525	=====
2526	
2527	
2528	
2529	
2530	
2531	
2532	
2533	
2534	
2535	
2536	
2537	