BIODISCOVERYAGENT: AN AI AGENT FOR DESIGNING GENETIC PERTURBATION EXPERIMENTS

Yusuf H. Roohani*, **Jian Vora*** & **Qian Huang***
Stanford University
Stanford, CA 94305, USA
`{yhr, jianv, qhwang}@cs.stanford.edu`

Zach Steinhart & Alex Marson
University of California, San Francisco
San Francisco, CA 94158, USA

Percy Liang, Jure Leskovec
Stanford University
Stanford, CA 94305, USA
`{pliang, jure}@cs.stanford.edu`

ABSTRACT

Genetic perturbation experiments play a crucial role in discovering the mechanisms behind diseases and informing drug development. These experiments aim to find a small subset out of many possible genes that yield a particular phenotype (e.g. cell growth) upon perturbation. However, the costs involved in each experiment limits the number of perturbations that can be tested. In this paper, we develop *BioDiscoveryAgent*, an AI agent that designs genetic perturbation experiments adaptively based on previous perturbation results to enhance the detection of perturbations that induce desired phenotypes. Our agent is based on language models, which have rich biological knowledge, and generate reasoning for the selection of genes to perturb. BioDiscoveryAgent attains an average of 23% improvement in detecting desired phenotypes across five datasets compared to existing Bayesian optimization baselines. This includes one dataset that is unpublished and therefore guaranteed to not appear in the language model's training data. Additionally, BioDiscoveryAgent is uniquely able to predict gene combinations to perturb, a task so far not explored in the setting of closed-loop experiment design. Overall, our approach represents an accessible new paradigm in the computational design of biological experiments, aimed at augmenting scientists' capabilities and accelerating scientific discovery.

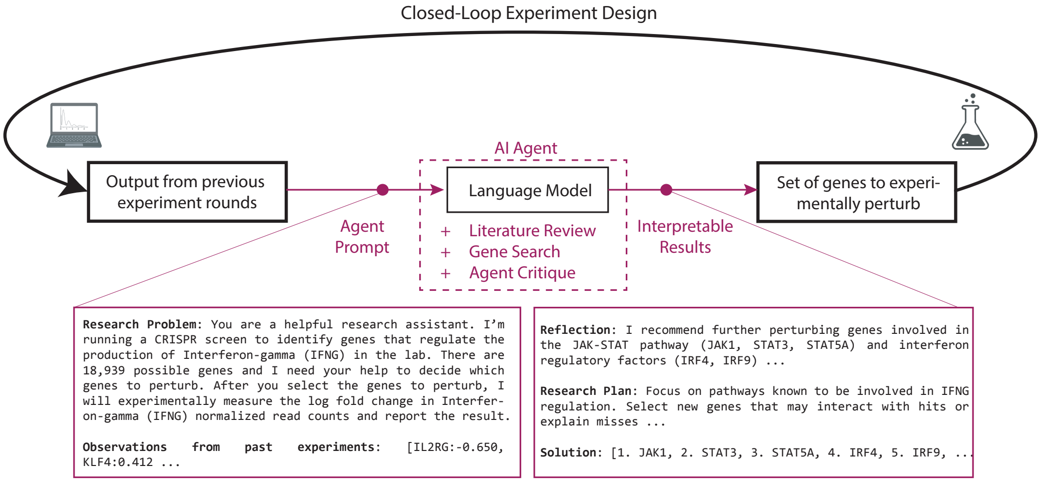
1 INTRODUCTION

An essential challenge in drug discovery is identifying biological targets, such as proteins encoded by genes, which, when modulated by a drug, lead to a desired phenotypic outcome (Scannell et al., 2012). Misidentification of targets that do not effectively impact disease phenotype upon drug interaction is a major cause of failure in clinical trials (Nelson et al., 2015). CRISPR-based genetic perturbation experiments, which involve the repression or activation of genes followed by the measurement of resulting biological effects, are instrumental in this search (Przybyla & Gilbert, 2022). These experiments, known as *perturbation screens* or *forward genetics screens* (Schneeberger, 2014; Moresco et al., 2013), have shown significant promise in various areas including drug target discovery Wang & Doudna (2023), combating drug resistance, elucidating disease mechanisms Mamedov et al. (2023), cell engineering Lim (2022), gene therapy Kalos et al. (2011), and immunotherapy Goodman et al. (2022).

A perturbation screen typically targets all known protein-coding genes in humans, numbering around 19,000, yet anywhere between a handful of genes to a few thousand may exhibit the desired phenotype. Experimentally perturbing every single gene as part of the search is costly, and even more so when considering gene combinations. By strategically designing these experiments in batches that

---

*Equal contibution. Correspondance to yhr@cs.stanford.edu prioritize genes likely to result in meaningful phenotypic effects, it is possible to reduce costs and enhance the efficiency of the search process (King et al., 2004; Cleary et al., 2017; Huang et al., 2023a; Roohani et al., 2023). However, the selection of batches is difficult, requiring both domain-specific knowledge and the ability to interpret and reason over prior experimental results. Recent work has benchmarked existing Bayesian optimization algorithms for this task, but this requires training bespoke machine learning models that are often difficult to interpret, on small datasets (Mehrjou et al., 2021; Lyle et al., 2023). Moreover, these models are unable to leverage the vast biological prior knowledge contained within the scientific literature that is especially important in early experimentation.


Closed-Loop Experiment Design
AI Agent
Output from previous experiment rounds
Agent Prompt
Language Model
+ Literature Review
+ Gene Search
+ Agent Critique
Interpretable Results
Set of genes to experimentally perturb

Research Problem: You are a helpful research assistant. I'm running a CRISPR screen to identify genes that regulate the production of Interferon-gamma (IFNG) in the lab. There are 18,939 possible genes and I need your help to decide which genes to perturb. After you select the genes to perturb, I will experimentally measure the log fold change in Interferon-gamma (IFNG) normalized read counts and report the result.

Observations from past experiments: [IL2RG:-0.650, KLF4:0.412 ...

Reflection: I recommend further perturbing genes involved in the JAK-STAT pathway (JAK1, STAT3, STAT5A) and interferon regulatory factors (IRF4, IRF9) ...

Research Plan: Focus on pathways known to be involved in IFNG regulation. Select new genes that may interact with hits or explain misses ...

Solution: [1. JAK1, 2. STAT3, 3. STAT5A, 4. IRF4, 5. IRF9, ...

Figure 1: **An AI agent for closed-loop experiment design.** In every round, the agent is given a prompt that describes the research problem and includes 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 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.

In this paper, we apply language model-based agents to this task. Language Models (LMs) have shown state of the art performance in information retrieval for clinical question answering (Moor et al., 2023), gene set identification (Hu et al., 2023), cell type annotation using gene expression data (Hou & Ji, 2024) and patient matching for clinical trials (Wornow et al., 2024). Furthermore, agents powered by LMs are particularly skilled at logical thinking and comparing various experimental plans, which makes them useful in the design and execution of scientific experiments (Liu et al., 2023; Huang et al., 2023b). They are also capable of making use of tools such as executing code that confer additional capabilities to the LM beyond prompt-based text generation Schick et al. (2024).

Here, we present *BioDiscoveryAgent*, an AI agent that designs genetic perturbation experiments using a language model in conjunction with a suite of tools (Figure 1a). In each round, the agent constructs a prompt that includes both the task description and experimental results from previous rounds. The prompt is fed to a language model, Claude v1, and the response is processed to extract the set of genes to perturb in the next round of experiments. For this task, the LM can also choose how best to leverage different tools, such as searching the scientific literature for relevant articles or querying information from tabular datasets that would otherwise not be available via text. BioDiscoveryAgent outperforms baseline methods for experiment design, identifying 23% more experimental *hits* after five experimental rounds of 128 genes each. It also shows enhanced performance in a new setting, predicting responses to combinatorial gene perturbations. 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 LM-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 experiment 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 $|\mathcal{G}| \approx 19000$. The effect of perturbing a gene $g \in \mathcal{G}$, is denoted by a phenotypic response function $f(g)$, where $f : \mathcal{G} \to \mathbb{R}$ maps each gene to a real-valued phenotypic response. In this setting, we define an experimental round as the perturbation of a set of genes $\{g_1, g_2, \ldots, g_B\} \subseteq \mathcal{G}$. 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 $t = 1, 2, \ldots, T$ experimental rounds, the total number of hits identified can be maximized. 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: $1, 2, \ldots, t$.

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. 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, $\mathcal{G}_p = \{g \in \mathcal{G} : f(g) > \tau\}$. We use the terms *hit ratio* and *top-K recall* interchangeably.

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 $f_c : \mathcal{G}^2 \to \mathbb{R}$. 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 LM to automate the scientific discovery process for this biological setting. The overall algorithm is shown in Algorithm1.

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 4a, Appendix C, D). 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 LM-based agents, such as pre-action reasoning (Yao et al., 2022), reflective thinking (Shinn et al., 2023), and stepwise planning (aut, 2023).

The primary mechanism by which BioDiscoveryAgent interacts with the human user is through natural language. An LM provides the capability to understand the prompts and generate responses and has been trained on large corpora of text including the biological literature. Additionally, the agent also has access to multiple resources to aid in this decision-making beyond what is stored within its weights. These include direct access to the biomedical literature, a second agent to critique its predictions, and the ability to analyze tabular datasets containing gene features relevant to the experimental setting. See Appendix G for full implementation details and Figure 4 for overall agent pipeline.

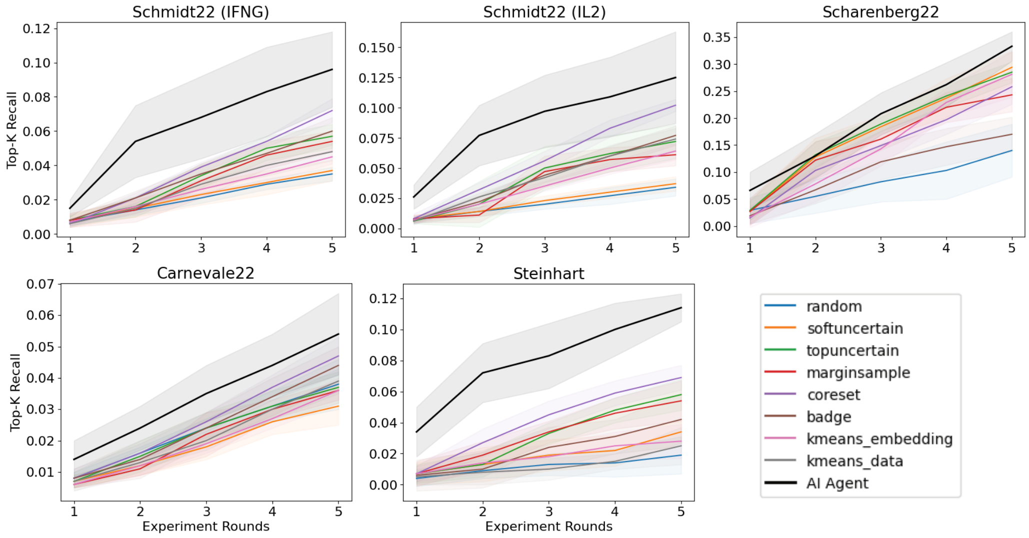


Figure 2: **Per-round performance comparison to machine learning baselines for 1-gene perturbation experiments**: Each line averages results from 10 runs, with error bars indicating 1 standard deviation. 128 genes were predicted per round over 5 rounds.

4 EXPERIMENTS

We assess model performance using data from past genetic perturbation experiments. Thus, we simulate the perturbation of a gene g that is recommended by the agent by retrieving the relevant observation of the perturbation-induced phenotype $f(g)$ from this dataset. For each perturbation screen, we calculate the hit ratio as the proportion of discovered hits out of the total true hits for that screen.

We tested our model's effectiveness through a series of batched experiments, each targeting 128 genes. At the beginning of each round, the model receives a prompt detailing the experiment's setup, the phenotypic outcome measured, and previously observed results. The agent processes this information to select a new batch of 128 genes for perturbation in the next round. This cycle repeats for five rounds, with the cumulative hit ratio recorded at each round.

For the single-gene perturbation setting, we make use of five different datasets spread across different cell types. 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 by Steinhart et al., which is so far unpublished. For baseline models, we use the methods implemented in the GeneDisco benchmark (Mehrjou et al., 2021). See Appendix J

5 RESULTS

BioDiscoveryAgent outperforms baselines based on hit ratio for 1-gene perturbation experiments: We compare the performance of BioDiscoveryAgent in its vanilla form (without any tools) and its enhanced form (with all available tools) against a range of machine learning benchmarks, including random sampling. As shown in Table 4, BioDiscoveryAgent with all tools significantly outperform baseline methods by 23% on average for hit ratios measured at experimental round 5. Performance improvement is observed across 5 of the 6 datasets. As shown in Figure 3, this gap is especially large at earlier rounds, where the LM can leverage its prior knowledge in biology to select promising gene candidates right away, in contrast to baseline methods that suffer from the cold start problem.

The vanilla agent outperforms all baselines except for the coreset approach. This exception is likely due to the coreset method's effective use of the gene feature dataset to identify diverse genes, information that may not be redundant with information found in the scientific literature. When using all tools, BioDiscoveryAgent merges its biological knowledge with additional data sources like gene features, enhancing its overall performance beyond that achieved by the coreset approach.

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 Kim & Hart (2021). On the other hand, the response of non-essential genes to perturbation, is harder to predict. When filtering for non-essential genes, BioDiscoveryAgent still shows a 20% performance improvement on average over baselines (Figure 6). Notably, the agent outperforms baselines across all 6 datasets when only considering non-essential genes.

Additionally, we tested the model's tendency to 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 (Figure 7), suggesting an approach to gene selection that 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 compared to only 18,939 pertinent genes in 1-gene perturbation experiment) Horlbeck et al. (2018). As shown in table 1, the BioDiscoveryAgent significantly outperforms the random sampling baseline by 130% on average.

Model	Cumulative Number of Hits				
Rounds	1	2	3	4	5
Random	2.6 ± 1.43	5.7 ± 2.83	8.9 ± 3.67	12.8 ± 3.74	16.4 ± 3.8
BDAgent	**5.50 ± 4.42**	**14.33 ± 5.76**	**21.67 ± 7.56**	**30.50 ± 7.34**	**32.67 ± 7.36**

Table 1: **Performance evaluation for 2-gene perturbation experiments.** 32 gene pairs predicted per round averaged over 10 runs, evaluated using the Horlbeck et al. dataset (n=100,576) Horlbeck et al. (2018). Error intervals correspond to 1 standard deviation.

BioDiscoveryAgent accounts for prior knowledge and observations in decision-making : We examine three scenarios: 1) Full Observation, where the agent utilizes both previous experiment results and detailed information about the experiment's goal; 2) No Observation, where the agent ignores all experiment results; 3) Only Observation, where the agent is unaware of the current experiment's goal and only conditioned on observations. For these experiments, we use a batch size of 32 genes and 30 rounds of experimentation on the two datasets from Schmidt et al. (2022).

Results show that Full Observation outperforms both No Observation and Only Observation, highlighting the significance of integrating prior knowledge and observations (Table 2, Figure 3a). Interestingly, Full Observation and No Observation benefit from prior knowledge early on, unlike Only Observation, which lacks the experiment's goal, underscoring the vital role of prior knowledge in the initial experiment phases. However, as experiments progress, Only Observation surpasses No Observation, showcasing the agent's capacity to adapt swiftly based on observations. 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), suggesting that observations significantly influence BioDiscoveryAgent's decision-making.

BioDiscoveryAgent can use tools to enhance performance and interpretability: To assess these capabilities, we provided the basic (vanilla) agent with access to one tool at a time and evaluated its performance (see Table 3, Appendix K). While certain tools, like literature search, may decrease performance in exchange for enhanced interpretability, employing all tools (with gene search configured to find dissimilar genes) offers a balanced approach that consistently delivers superior results.

BioDiscoveryAgent provides interpretable predictions with references to the literature

BioDiscoveryAgent provides interpretable predictions at various stages. One such example is illustrated in Appendix M and summarized in Figure 4. The agent is tasked with identifying genes regulating the production of Interferon-gamma (IFNG). The vanilla agent explicitly reasons that it will focus broadly on pathways related to IFN-γ signalling, production and response (4c). Utiliz-

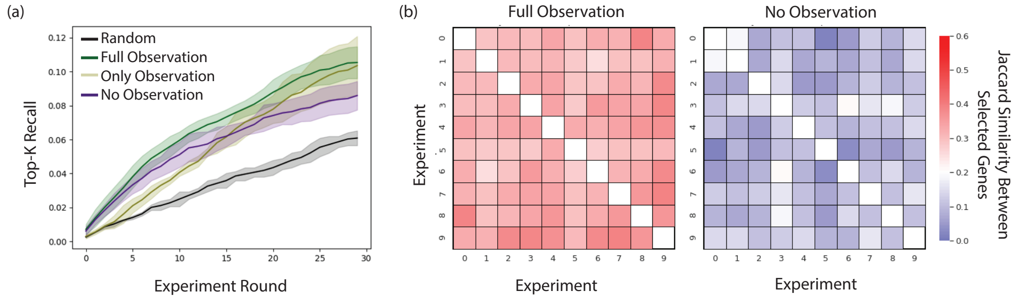


Figure 3: **Agent accounts for both prior knowledge and observations in decision-making** Three scenarios are considered: the agent has access to task description and experimental observations (Full observation); the agent only has access to experimental observations (Only observation); the agent only has access to the task description (No Observation). **(a)** Hit ratio (Top-K recall) at each experimental round across 30 rounds of experiments. 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.

ing the literature search tool, the agent accesses relevant literature to support its predictions, citing specific papers and line numbers as in Figure 4d. For example, the agent highlights *IFNG-AS1* as a gene for potential perturbation, supported by references to specific lines in scientific papers. The LM critic tool provides valuable insights into the predicted set of genes, identifying potential issues such as randomness or a focus on basic cellular processes that are unlikely to regulate IFNG production. Thus, not only does the LM critic provide clear reasons for selecting specific genes but also further opens avenues for human-in-the-loop feedback. See Appendix L.

6 DISCUSSION

BioDiscoveryAgent represents a new paradigm in the computational design of biological experiments, aimed at augmenting scientists' capabilities and accelerating scientific discovery. Traditionally, this process employs a multi-stage pipeline that relies on 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 a large language model, can simplify the entire experimental design process into a single prompt. Through its enhanced interpretability at every stage, the researcher can meaningfully engage with the model throughout the process. Furthermore, the agent is equipped with prior biological knowledge, solving the cold start problem at the start of of a new round of experiments, while still efficiently utilizing observational data as the rounds progress. Through the usage of tools, it can explicitly search for and integrate information from the scientific literature as well as existing tabular datasets. This results in a research assistant capable of speeding up biological research through utilizing information from diverse sources.

While BioDiscoveryAgent demonstrates strong capabilities in enhancing experiment design, it still presents room for further development. BioDiscoveryAgent tends to perform better with certain cell types over others, likely due to a difference in the level of coverage within the scientific literature. Most performance benefits are observed in the early stages of experimentation rather than over longer rounds of experiments. Additionally, there's a need for developing better tools or fine-tuning methods to improve the ability to extract information from non-text-based sources.

Overall, BioDiscoveryAgent offers a complementary method to existing approaches to experimental design, significantly improving model performance in the low data regime while also enhancing overall model interpretability. Amid increasing interest in active experimental design for biological experiments, BioDiscoveryAgent offers compelling evidence that language models could become an essential component 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

Significant-gravitas/auto-gpt: 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 Blaeschke, Yan Yi Chen, Zhongmei Li, Sagar P Bapat, Morgan E Diolaiti, Patrick O'Leary, et al. Rasa2 ablation in t cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925):174–182, 2022.

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.

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. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pp. 1–4, 2024.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Research Square*, 2023.

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`.

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.

Eiru Kim and Traver Hart. Improved analysis of crispr fitness screens and reduced off-target effects with the bagel2 gene essentiality classifier. *Genome medicine*, 13:1–11, 2021.

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.

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.

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, Jiongxiao 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.

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.

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.

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 Barekatain, 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.

Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

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 Dessı, 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.

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.

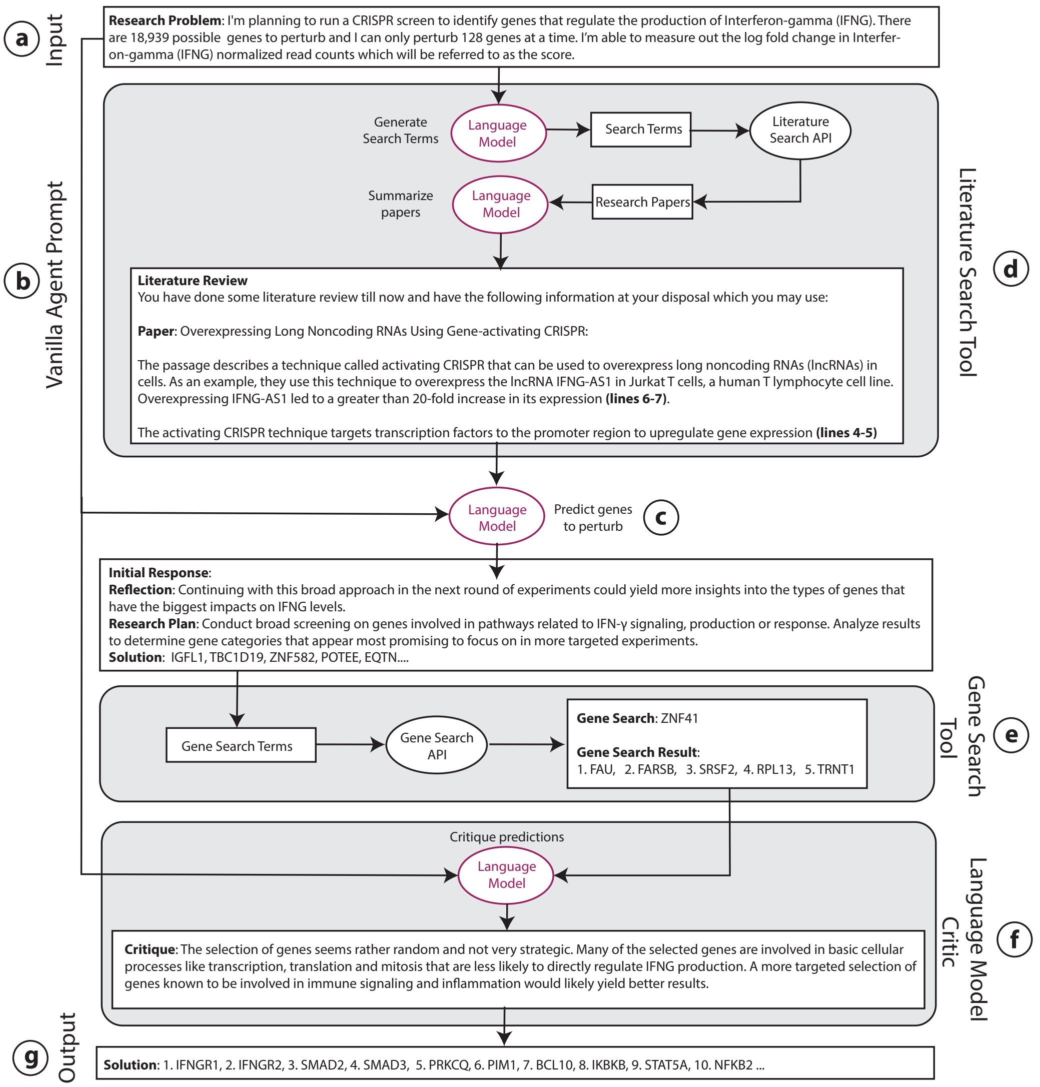
Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification, 2023.

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 Overall Agent Workflow



Research Problem: I'm planning to run a 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. 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.

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

Paper: Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR:

The passage describes a technique called activating CRISPR that can be used to overexpress long noncoding RNAs (lncRNAs) in cells. As an example, they use this technique to overexpress the lncRNA IFNG-AS1 in Jurkat T cells, a human T lymphocyte cell line. Overexpressing IFNG-AS1 led to a greater than 20-fold increase in its expression **(lines 6-7)**.

The activating CRISPR technique targets transcription factors to the promoter region to upregulate gene expression **(lines 4-5)**

Initial Response:
Reflection: Continuing with this broad approach in the next round of experiments could yield more insights into the types of genes that have the biggest impacts on IFNG levels.
Research Plan: Conduct broad screening on genes involved in pathways related to IFN-γ signaling, production or response. Analyze results to determine gene categories that appear most promising to focus on in more targeted experiments.
Solution: IGFL1, TBC1D19, ZNF582, POTEE, EQTN....

Gene Search: ZNF41

Gene Search Result:
1. FAU, 2. FARSB, 3. SRSF2, 4. RPL13, 5. TRNT1

Critique: The selection of genes seems rather random and not very strategic. Many of the selected genes are involved in basic cellular processes like transcription, translation and mitosis that are less likely to directly regulate IFNG production. A more targeted selection of genes known to be involved in immune signaling and inflammation would likely yield better results.

Solution: 1. IFNGR1, 2. IFNGR2, 3. SMAD2, 4. SMAD3, 5. PRKCQ, 6. PIM1, 7. BCL10, 8. IKBKB, 9. STAT5A, 10. NFKB2 ...

Figure 4: **Agent workflow with all tools over a single experimental round**. Prompts and agent responses have been summarized. See Appendix M for full trace. **(a)** The input to the agent is the description of the problem. **(b)** The vanilla agent just feeds the problem description directly into the LM. **(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 the literature search tool, the model is provided with context from the scientific literature along with the problem description. The LM determines appropriate search terms of the scientific literature, top ranked retrieved papers are summarized by the LM. **(e)** In case of using the gene search tool, the LM identifies a gene for which to query similar/dissimilar genes **(f)** The LM can also be prompted differently to function as a critic and analyze the predictions so far from a different perspective **(g)** Final output gene list after using different tools.

B Algorithm

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

```
Input: Experiment description, Number of rounds $S$, Number of genes perturbed $b$
Output: Set of genes to perturb
for $t = 1, \ldots, T$ do
    LM searches, summarizes relevant scientific literature, adds it to the prompt
    Agent prompts LM to select $b$ new genes (or gene pairs). Output: LM generates structured response with Reflection, Research Plan, Gene Search, Solution entries
    while some predicted genes are invalid do
        Prompt LM to select new genes
    end while
    Prompt LM to critique the prediction made by main agent
    Get phenotypic score $f(g)$ for each gene $g$ and add to prompt for next step
    if context window is too large then
        Prompt LM to summarize text for context window management
    end if
end for
```

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

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).
```

D DATASET SPECIFIC PROMPTS

For each dataset, we mention the research problem and the type of measurement outcome that the agent will be presented with. These are used to create the research problem with which the BioDiscoveryAgent is prompted.

```
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

Steinhart CRISPRa Task: identify genes that upon inhibition allow
                cells to resist T-cell
                exhaustion, under the HA GD2 CAR
                (chimeric-antigenic receptor) condition
Steinhart CRISPRa 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
```

E 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 (Figure 5a). 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 (Figure 5b,c), BioDiscoveryAgent summarizes the list of all remaining genes and adds it to the prompt to aid in gene selection (Figure 5d). The summarized gene list is designed to cover a broad range of biological pathways and functions.

Moreover, after multiple experimental rounds, the historical data and experimental observations can also often exceed the LM's processing capacity. To address this, we employ a similar summarization technique to keep the information in future prompts focused and relevant.

F 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 LM to structure its responses into several parts: `Reflection`, `Research Plan`, `Solution` (Appendix C, (Figure 4c)), 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.

G 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 resoning for adding those tools.

1. **AI critic**: LMs are very sensitive to the prompt they are presented with. Past research has shown varying performance depending on the setting in which the LM is queried. Thus, an LM prompted to behave as an agent for a human researcher may behave very

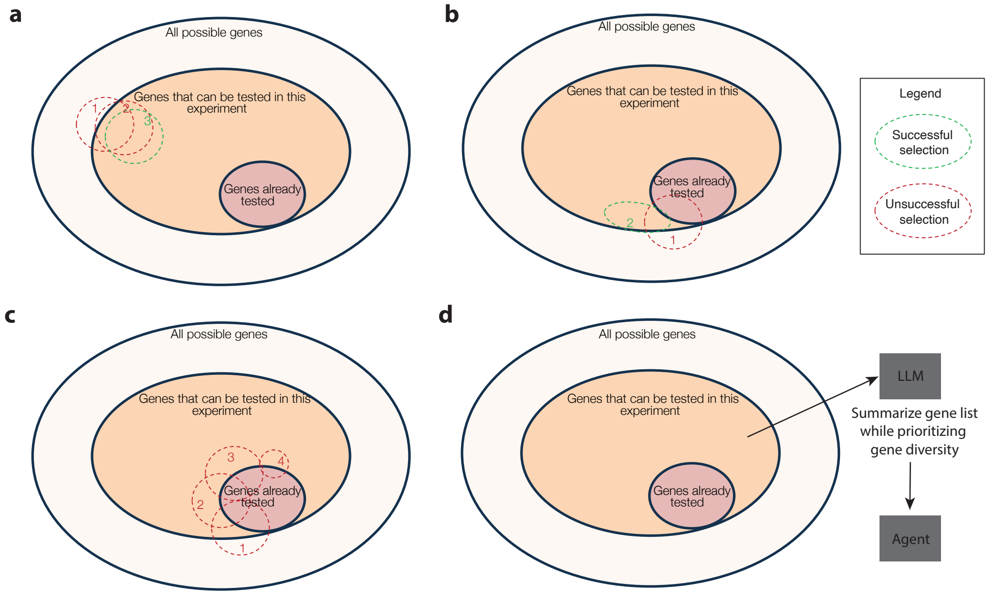

a
All possible genes
Genes that can be tested in this experiment
Genes already tested
1
2
3
b
All possible genes
Genes that can be tested in this experiment
Genes already tested
2
1
Legend
Successful selection
Unsuccessful selection
c
All possible genes
Genes that can be tested in this experiment
Genes already tested
1
2
3
4
d
All possible genes
Genes that can be tested in this experiment
Genes already tested
LLM
Summarize gene list while prioritizing gene diversity
Agent

Figure 5: **Gene selection strategy**: **(a)** The space of genes that can be tested in a given experiment is constrained by expeirmental 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.

differently from one that is prompted to mainly critique the prediction made by another LM. 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 LM) 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 4f). 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 4d). 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**: LMs 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 select 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. 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. Then the selected genes are concatenated to the prompt to the main agent. For example, if the LM requests a gene search of the gene *ZNF41* based on dissimilarity, the agent appends the following result in the next prompt "Gene Search Result: *FAU*, *FARSB*, *SRSF2*, *RPL13*, ...". (Figure 4e).

   For similarity, we retrieved the top 10 genes, while for dissimilarity, we retrieved the top 50 genes.

H BASELINES

- **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.

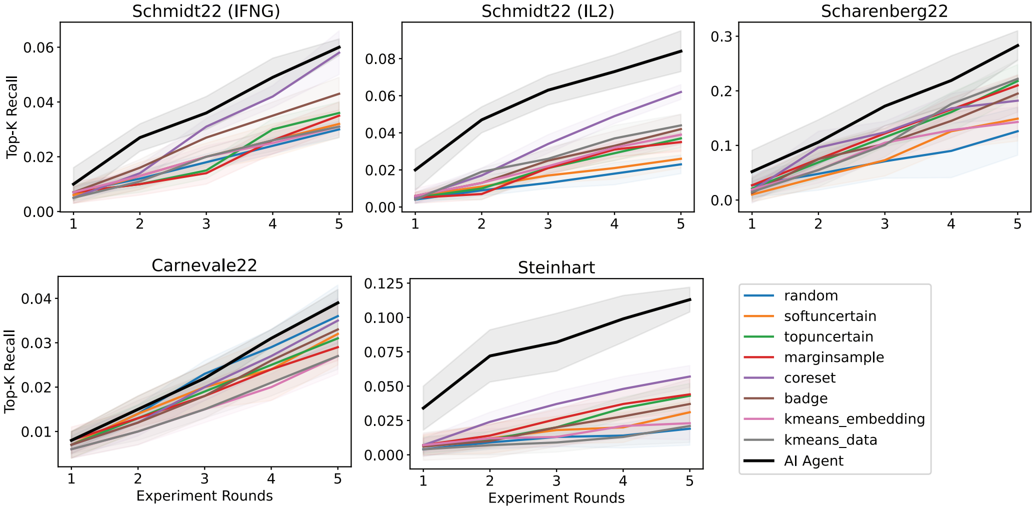


Figure 6: **Per-round performance comparison to machine learning baselines for 1-gene perturbation experiments (non-essential genes)**: Each line averages results from 10 runs, with error bars indicating 1 standard deviation. 128 genes were predicted per round over 5 rounds. At each step, only those genes were considered hits that were not also essential genes.

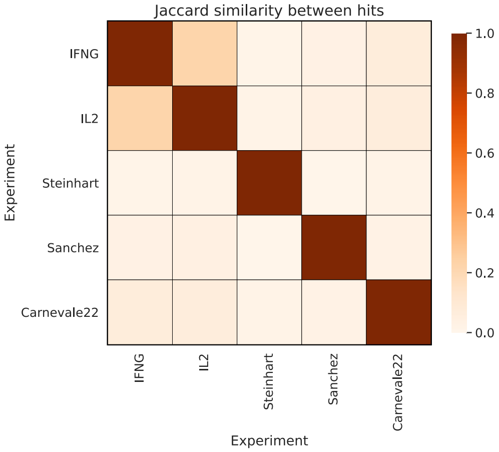


Figure 7: **Jaccard similarity index between all predicted hits at 5 rounds of experimentation** Each cell corresponds to the union of predicted hits across 10 model runs for each dataset.

Dataset		Schmidt22 (IL2)			Schmidt22 (IFNG)		
Rounds		10	20	30	10	20	30
Model	Setting	Avg. Top-K Recall					
BDAgent	Only Obs.	0.0449	0.0872	0.1037	0.0291	0.0702	0.1061
BDAgent	No Obs.	**0.0605**	0.0846	0.1071	0.0484	0.0727	0.0872
BDAgent	Full Obs.	0.0596	**0.0908**	**0.1143**	**0.0559**	**0.0842**	**0.1077**

Table 2: **Model uses both prior knowledge and experimental observations to make predictions.** 32 genes predicted per round.

Model	Top-K Recall					
Dataset	Schmidt1	Schmidt2	Steinhart	Scharen.	Carnev.	Sanchez
Random	0.037	0.031	0.033 3	0.160	0.036	0.034
BDAgent (Vanilla)	0.067	0.089	0.110	0.292	0.038	0.053
BDAgent (Literature)	0.052	0.069	0.073	0.232	0.023	0.055
BDAgent (AI critic)	0.069	0.089	0.130	0.341	0.044	0.056
BDAgent (Similar Genes)	0.083	0.087	0.111	**0.351**	0.045	0.056
BDAgent (Dissimilar Genes)	**0.099**	**0.162**	0.064	0.286	0.048	0.041
BDAgent (All Tools)	0.095	0.122	**0.114**	0.314	**0.054**	**0.058**

Table 3: **Agent performance improvements when using different tools for 1-gene perturbation experiments**. Results shown are 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. †Steinhart refers to an unpublished dataset.

.

I 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). Additionally, closed-loop lab-based experimentation has made notable advances, particularly in chemical synthesis (Boiko et al., 2023) and materials discovery (Wang et al., 2023).

In the biological domain, research has shown that large language models 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). The concept of using artificial intelligence to generate hypotheses for functional genomics experiments is not new (King et al., 2004). The vast potential experimental space and the resulting combinatorial challenges underscore the inherent value of this approach. Earlier studies have also investigated the use of machine learning to optimize the design of genetic perturbation experiments (Mehrjou et al., 2021; Lyle et al., 2023).

J 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 measures a distinct biological process. The Schmidt et al. (2022) dataset measures the changes in the production of two key cytokines involved in immune signaling: Interleukin-2 (IL-2) and Interferon-γ (IFNG) 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 from Steinhart et al. 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.

Model	Top-K Recall					
Dataset	Schmidt1	Schmidt2	Steinhart[†]	Scharen.[*]	Carnev.	Sanchez
Random	0.037	0.031	0.033	0.160	0.036	0.034
BDAgent (Vanilla)	0.067	0.089	0.110	0.292	0.038	0.053
BDAgent (All Tools)	**0.095**	**0.122**	**0.114**	**0.314**	**0.054**	0.058
Soft Uncertain	0.037	0.037	0.034	0.205	0.031	0.029
Top Uncertain	0.057	0.072	0.058	0.294	0.037	0.039
Margin Sample	0.054	0.061	0.054	0.285	0.036	0.033
Coreset	0.072	0.102	0.069	0.243	0.047	**0.061**
Badge	0.060	0.077	0.042	0.258	0.044	0.039
Kmeans Embed.	0.045	0.064	0.028	0.170	0.036	0.037
Kmeans Data	0.048	0.074	0.025	0.281	0.039	0.043

Table 4: **Performance comparison to machine learning baselines for 1-gene perturbation experiments**. Results shown are 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. [†]Steinhart refers to an unpublished dataset.

For the two-gene perturbation task, we use a dataset from a screen that knocked down 100,576 gene pairs in K562 cells. 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 from the average impact of knocking down each gene individually.

For baseline models, we use the methods implemented in the GeneDisco benchmark (Mehrjou et al., 2021). Every baseline includes a multi-layer perceptron M for predicting experimental outcomes using gene features. This is then combined with one of six different acquisition functions for designing each round of experiments (See Appendix H).

K PERFORMANCE IMPACT OF USING DIFFERENT TOOLS

- **Literature search**: Here, the agent had access solely to the literature search tool. We observed that this tool does not always improve performance, as the agent may become fixated on a few simple keywords and irrelevant papers, leading to less effective searches. Nonetheless, literature search provides valuable, interpretable citations for the gene prediction process, beneficial for scientists.
- **AI critic**: When only using the AI critic tool, performance improves slightly over the vanilla model mainly by diversifying the original genes predicted or by concentrating the predictions towards specific gene sets.
- **Similar Genes**: Performance significantly improved when using a gene similarity search based on specific features, especially for IFNG and Scharenberg22. We believe this improvement is due to gene similarity searches relying on experimentally derived gene feature sets, typically stored in tabular format and not fully represented in the text of scientific papers. Thus, the vanilla LM may not have access to this information.
- **Dissimilar Genes**: Here, the agent used the gene search tool to find dissimilar genes, with minor instructions on how to explore these diverse genes. This approach of focusing on dissimilar genes encourages BioDiscoveryAgent to explore a wider variety of genes, drawing inspiration from the coreset method. We observed variable performance across different datasets, likely reflecting the varying reliance on diversity for each target task.

L INTERPRETABILITY EXAMPLES

Here we show some examples that demonstrate the interpretability of BioDiscoveryAgent for experimental design:

1. In the initial steps, where the agent tried out a lot of random genes, the critic agent guides the main agent to focus more on genes relevant to the task.

```
1. Critique: The selection of genes seems somewhat random
   and not focused enough on likely candidates based on
   known lysosomal and endolysosomal genes and pathways. I
   would suggest focusing more on genes known to be
   involved in lysosomal function, endolysosomal
   trafficking, and choline metabolism. Some of the
   selected genes like MAPK1, RAF1, PIK3CG, EIF2S1 seem
   less likely to be directly involved.

2. Updated Solution: 1. CTSA, 2. HEXA, 3. HEXB, 4. GBA, 5.
   GUSB, 6. GALC, 7. SMPD1, 8. LIPA, 9. ASAH1, 10. NPC1,
   11. NPC2, 12. LAMP1, 13. LAMP2, 14. LAMP3, 15.RAB7A, 16.
    RAB1A, 17. CLN3, 18. MCOLN1, 19. ATP6V0A1, 20. ATP6V0C,
    21. ATP6V1A, 22. ATP6V1B2, 23. AP1S1, 24. AP1S2, 25.
   AP3B1, 26. AP3D1, 27. VPS11, 28. VPS18, 29. VPS33A, 30.
   VPS33B, 31. VPS41, 32. VPS45
```

2. Providing reasoning for selecting specific genes not only enhances interpretability but also creates opportunities for human-in-the-loop feedback. This feedback is conveyed in natural language by subject experts. Additionally, the agent supports its predictions by citing relevant scientific literature, as shown below.:

```
Human tryptophanyl-tRNA synthetase is an IFN-  -inducible
   entry factor for Enterovirus.
 Here is a description of what can be gathered from the
    provided information:

The file describes a study identifying human tryptophanyl-
   tRNA synthetase (hWARS) as an interferon-gamma (IFN-G)-
   inducible cellular entry factor for Enterovirus A71 (EV-
   A71) and other enteroviruses. hWARS expression is
   required for EV-A71 infectivity and pathogenesis (lines
   1-3).  hWARS was identified using an RNAi screen of the
   human genome (line 2). Its importance was confirmed
   using multiple methods (line 5). hWARS expression and
   location change in response to IFN-G, allowing EV-A71
   infection (lines 4-5). A hWARS-transduced mouse model
   showed EV-A71 infection effects (line 6). hWARS is
   needed for other enteroviruses like coxsackievirus A16 (
   CV-A16) and enterovirus D68 (EV-D68) (line 7).
```

M 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.

```
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:

Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR.
 Here is a summary of the information that can help with the
    research problem:

The passage describes a technique called activating CRISPR that
   can be used to overexpress long noncoding RNAs (lncRNAs) in
   cells. As an example, they use this technique to overexpress
   the lncRNA IFNG-AS1 in Jurkat T cells, a human T lymphocyte
   cell line. Overexpressing IFNG-AS1 led to a greater than 20-
   fold increase in its expression (lines 6-7).

Since IFNG-AS1 is associated with inflammatory bowel disease,
   overexpressing it in T cells could affect interferon gamma
   production. The activating CRISPR technique targets
   transcription factors to the promoter region of a gene to
   upregulate its expression (lines 4-5). This suggests that
   perturbing transcription factors involved in regulating IFNG
   expression could impact interferon gamma levels.

In this segment, I cannot find specific genes to target in the
   first round of experiments. The passage does not provide that
   level of detail. However, based on the information provided,
   transcription factors and lncRNAs associated with interferon
   gamma regulation and production could be good candidates to
   screen first. An initial broad screen of diverse candidates
   from these groups could help identify the types of genes that
```

have the biggest impact on the research problem. The results could then inform a more focused second round of screening.

Human tryptophanyl-tRNA synthetase is an IFN- -inducible entry factor for Enterovirus.

Here is a summary of the information provided:

The passage describes the identification of human tryptophanyl-tRNA synthetase (hWARS) as an interferon-gamma (IFN-)-inducible cellular entry factor for Enterovirus A71 (EV-A71). hWARS expression was found to be required for productive infection of EV-A71 as well as other human enteroviruses like coxsackievirus A16 and EV-D68 (lines 1-3).

High levels of IFN- were detected in patients with severe EV-A71 infection, suggesting that hWARS expression upon IFN- stimulation may contribute to EV-A71 pathogenicity (lines 15-17).

In this segment, I cannot find specific genes that regulate IFNG production. The passage focuses on hWARS as an entry factor for EV-A71 that is induced by IFN- .

For the research problem, I would suggest:

1) Focusing initial experiments on diverse genes involved in pathways related to IFN- signaling or production, such as the JAK-STAT pathway or genes involved in T helper cell differentiation. Observing the effects of perturbing these genes could provide insights into which gene types most impact IFNG levels.

2) Based on initial observations, select genes that appear to have the largest effects on IFNG for follow-up experiments. This could help increase the cumulative "hit rate" of genes that regulate IFNG production.

3) Considering genes encoding other components of the IFN- response pathway in addition to hWARS, such as IFNGR1 and IFNGR2 which encode the IFN- receptor subunits. Perturbing these genes may also impact IFNG levels.

Assistant: 1. Reflection: We have screened a diverse set of genes so far and identified many promising candidates involved in IFN- signaling and regulation. Continuing with this broad approach in the next round of experiments could yield more insights into the types of genes that have the biggest impacts on IFNG levels.

2. Research Plan: Conduct another round of broad screening on genes involved in pathways related to IFN- signaling, production or response. Analyze results to determine gene categories that appear most promising to focus on in subsequent, more targeted experiments.

2. Gene Search: ZNF41

Gene Search Result:FAU, FARSB, SRSF2, RPL13, TRNT1, SRSF3, HSPE1, TXNL4A, PRELID1, TRMT112, LSM2, NCAPG, CCT4, RPS29, WDR43, ZNHIT2, CHAF1B, COPS6, RPS15, RPL8, GTF2B, NUDT21, RPS17, RPL17, ZPR1, RPL15, CTCF, ALG11, ERH, DBR1, NIP7, RPL27, BOP1, VCP, WDR75, RRM1, LUC7L3, POLR2D, RPS3, RPL36, DDB1, DUX4, VPS25, RPL31, RPS16, DTL, POLE2, TRAPPC1, TRAPPC8, RPL24

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:
['IGFL1', 'TBC1D19', 'ZNF582', 'POTEE', 'EQTN', 'AP1S2', 'KRBOX4', 'ZNF611', 'MGST1', 'RBM25', 'PSMD7', 'TXNL4A', 'TBCC', 'CHMP6', 'RBBP5', 'RPS19', 'NUP93', 'ATP6V0C', 'SDE2', 'CDC45', 'GPN2', 'SRSF7', 'KPNB1', 'GPN3', 'RPL27', 'PLK1', 'ESPL1', 'RPS4X', 'MTBP', 'TIMELESS', 'PAFAH1B1', 'ETF1', 'DONSON', 'GINS2', 'SPC24', 'CDC123', 'SF3B4', 'RPS8', 'SAP30BP', 'CFAP298', 'RPS16', 'DUT', 'CCT4', 'TBCB', 'MED8', 'SBNO1', 'SNAPC2', 'POLR2I', 'INTS3', 'PSMD14', 'KIF11', 'RUVBL1', 'RPL12', 'RPL37', 'USP36', 'THOC1', 'RPS13', 'SMAD2', 'SOCS3', 'PIAS3', 'LTA', 'IL15', 'STAT1', 'SOCS2', 'PTPN22', 'NFKB2', 'SOCS5', 'PTPN11', 'TAB2', 'SMAD7', 'STAT3', 'IL12B', 'IRF7', 'PTPN2', 'IFNGR1', 'CXCL9', 'IL18', 'JAK1', 'IRF8', 'SMAD4', 'SOCS1', 'STAT2', 'STAT5A', 'IRF9', 'RORC', 'NFKBIA', 'IFNGR2', 'IRF3', 'GATA3', 'IRF4', 'TNF', 'PIAS1', 'SMAD3', 'STAT5B', 'IRF5', 'RELA', 'PTPN4', 'NFKBIE', 'IL12A', 'TBX21', 'PTPN6', 'NFKB1', 'IRF1', 'CXCL10', 'IL1B', 'TYK2', 'JAK2', 'NFKBIB', 'SOCS6', 'IRF2', 'TGFB1', 'PIAS2', 'IL6', 'EOMES', 'SMAD5', 'NFKBIL1', 'SMAD6', 'STAT4', 'PIAS4', 'SRP19', 'MRPL53', 'ZNF207', 'METTL17', 'UBAP1', 'SART1', 'ERAL1', 'FAU', 'SRSF2']

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.

```
Assistant:
===================claude-1 response (2000)=====================
 1. Critique: The selection of genes seems rather random and not
     very strategic. Many of the selected genes are involved in
     basic cellular processes like transcription, translation and
     mitosis that are less likely to directly regulate IFNG
     production. A more targeted selection of genes known to be
     involved in immune signaling and inflammation would likely
     yield better results.

2. Updated Solution:
1. IFNGR1, 2. IFNGR2, 3. JAK1, 4. JAK2,
5. STAT1, 6. STAT2, 7. STAT3, 8. STAT4,
9. STAT5A, 10. STAT5B, 11. IRF1, 12. IRF2,
13. IRF3, 14. IRF4, 15. IRF5, 16. IRF7, 17. IRF8, 18. IRF9,
19. RELA, 20. NFKB1, 21. NFKB2, 22. NFKBIA, 23. NFKBIB,
24. NFKBIE, 25. NFKBIL1, 26. MAP3K7, 27. MAP3K8,
28. MAP2K4, 29. MAP2K7, 30. TGFB1, 31. TGFBR1, 32. TGFBR2,
33. SMAD2, 34. SMAD3, 35. SMAD4, 36. SMAD5, 37. SMAD7,
38. IL12A, 39. IL12B, 40. IL1B, 41. IL6, 42. IL15, 43. IL18,
44. TNF, 45. LTA, 46. CXCL9, 47. CXCL10, 48. SOCS1, 49. SOCS2,
50. SOCS3, 51. PIAS1, 52. PIAS2, 53. PIAS3, 54. PIAS4,
55. PTPN1, 56. PTPN2, 57. PTPN4, 58. PTPN6, 59. PTPN11,
60. PTPN22, 61. TAB2, 62. MAP3K14, 63. CHUK, 64. IKBKB,
65. MAP2K1, 66. MAP2K3, 67. PRKCQ, 68. PIM1, 69. PIM2,
70. PIM3, 71. BCL10, 72. CARD11, 73. MALT1, 74. IKBKG,
75. UBE2V2, 76. UBE2N, 77. UBE2D3, 78. UBE2D4, 79. RIPK2,
80. MAP3K1, 81. MAP3K3, 82. MAP3K5, 83. MAP2K2, 84. MAPK14,
85. MAPK8, 86. MAPK9, 87. MAPK1, 88. MAPK3, 89. MAPK13,
90. MAPK12, 91. MAPK11, 92. MAPK10, 93. MAPK8IP3, 94. MAPK9IP1,
95. MAPK8IP1, 96. MAPK8IP2, 97. CASP1, 98. CASP8, 99. RIPK1,
100. TRAF2, 101. TRAF3, 102. TRAF5, 103. TRAF6, 104. TANK,
105. TBK1, 106. IKBKE, 107. IKBKB, 108. CHUK, 109. RIPK3,
110. RIPK4, 111. CARD9, 112. BIRC2, 113. BIRC3, 114. XIAP,
115. BCL2, 116. MCL1, 117. BCL2L1, 118. BIRC5, 119. CFLAR,
120. FADD, 121. TRADD, 122. RIPK1, 123. RIPK2, 124. MAP3K1,
125. UBE2N, 126. UBE2V1, 127. UBE2V2, 128. UBE2D1.


Human: Step 1
 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!


 This is not your first round. All tested genes and their measured
     log fold change are:
             Score
Gene
SMAD2     0.101575
UBE2V1   -0.025857
PIAS3     0.085304
LTA      -0.187780
CASP8     0.142034
MAPK10   -0.008140
MAPK9     0.206474
SOCS2    -0.006300
MAPK11    0.022875
PTPN22    0.339770
MAPK8IP1 -0.208230
IL18      0.064003
UBE2V2   -0.095655
IRF7      0.178988
IL12B     0.106245
CXCL9    -0.013024
TBK1      0.013595
IRF8     -0.284585
SMAD4     0.080494
BCL2      0.016410
SOCS1     0.185722
STAT2    -0.121285
TANK     -0.016199
UBE2N     0.126805
SMAD3     0.122430
STAT5B    0.010015
IRF5     -0.042555
PTPN4     0.024984
NFKBIE    0.215400
IL12A     0.128751
MCL1     -0.062024
CARD11   -0.313452
IRF1     -0.029467
CXCL10   -0.010432
```

```
MAPK8     -0.094135
JAK2       0.132444
CARD9      0.117949
MAP2K4    -0.024708
MAPK13     0.146087
TRAF5     -0.140703
PIM2       0.059805
IRF2      -0.094631
MAP3K8    -0.181561
TGFB1      0.032735
MAP3K14   -0.366105
IL6        0.064360
RIPK2     -0.210654
CFLAR     -0.336920
MAPK8IP3   0.160541
TGFBR2     0.013636
SMAD5     -0.129476
NFKBIL1   -0.129973
TRAF3      0.109792
STAT4     -0.282660
TGFBR1    -0.060750
IL15      -0.035326
MAP3K5    -0.224920
STAT1     -0.023106
BIRC2      0.003260
PIM3       0.017550
MAPK12     0.138772
PTPN11     0.136000
TAB2       0.367535
SMAD7      0.284655
PTPN2      0.072855
RIPK4      0.068350
MAP2K3    -0.166365
UBE2D3     0.224675
STAT5A     0.070219
IRF9      -0.126400
MAP3K3     0.220350
NFKBIA     0.344135
IRF3      -0.237845
MAP3K1     0.199285
MAPK8IP2  -0.022033
IRF4       0.059575
TNF        0.272485
PIAS1     -0.113202
MAPK14    -0.244472
RIPK3      0.010906
BIRC3      0.193885
IL1B      -0.008130
NFKB1      0.125557
PTPN6      0.247855
UBE2D4     0.087250
FADD       0.107290
CASP1     -0.064096
PIAS2      0.247420
MAP2K1    -0.209255
XIAP      -0.342415
BIRC5     -0.082685
PIAS4      0.038910
PTPN1      0.188203
```

```
BIRC7    -0.005555
MAP3K15  -0.086644
UBE2C    -0.072150
MAP4K5    0.058382
UBE2Q1    0.001715
UBE2Q2   -0.251546
UBE2T     0.279460
UBE2G1    0.034193
UBE2H     0.287118
UBE2J1   -0.278075
UBE2W     0.111455
 You have successfully identified 24 hits so far over all
    experiment cycles! The results for the hits are:
           Score
Gene
TRAF6  -1.865400
SOCS3   0.536160
BCL2L1 -0.460815
MAPK1  -0.449405
NFKB2   1.584300
STAT3  -1.048600
IFNGR1 -0.898070
JAK1   -1.389050
MAPK3  -0.734105
CHUK   -0.833950
IFNGR2 -0.975910
MALT1  -1.063740
RELA   -0.840910
PRKCQ  -0.442575
IKBKE   0.487890
IKBKB  -0.922680
MAP2K2 -0.402665
MAP2K7 -0.446480
TRADD  -0.623995
NFKBIB  0.495810
RIPK1   0.606570
MAP3K7 -1.305750
BCL10  -0.869020
IKBKG  -0.853900
 Based on these results and prior knowledge of biology, make the
    best possible prediction of the first 128 genes that I should
     test to maximize the score. Use HGNC gene naming convention.
    DO NOT PREDICT GENES THAT HAVE ALREADY BEEN TESTED
 You have done some literature review till now and have the
    following information at your disposal which you may use to
    make your predictions:

Overexpressing Long Noncoding RNAs Using Gene-activating CRISPR.
 Here is a summary of the information that can help with the
    research problem:

The passage describes a technique called activating CRISPR that
   can be used to overexpress long noncoding RNAs (lncRNAs) in
   cells. As an example, they use this technique to overexpress
   the lncRNA IFNG-AS1 in Jurkat T cells, a human T lymphocyte
   cell line. Overexpressing IFNG-AS1 led to a greater than 20-
   fold increase in its expression (lines 6-7).
```

```
Since IFNG-AS1 is associated with inflammatory bowel disease,
   overexpressing it in T cells could affect interferon gamma
   production. The activating CRISPR technique targets
   transcription factors to the promoter region of a gene to
   upregulate its expression (lines 4-5). This suggests that
   perturbing transcription factors involved in regulating IFNG
   expression could impact interferon gamma levels.

In this segment, I cannot find specific genes to target in the
   first round of experiments. The passage does not provide that
   level of detail. However, based on the information provided,
   transcription factors and lncRNAs associated with interferon
   gamma regulation and production could be good candidates to
   screen first. An initial broad screen of diverse candidates
   from these groups could help identify the types of genes that
   have the biggest impact on the research problem. The results
   could then inform a more focused second round of screening.


Human tryptophanyl-tRNA synthetase is an IFN-  -inducible entry
   factor for Enterovirus.
 Here is a summary of the information provided:

The passage describes the identification of human tryptophanyl-
   tRNA synthetase (hWARS) as an interferon-gamma (IFN-  )-
   inducible cellular entry factor for Enterovirus A71 (EV-A71).
   hWARS expression was found to be required for productive
   infection of EV-A71 as well as other human enteroviruses like
   coxsackievirus A16 and EV-D68 (lines 1-3).

High levels of IFN-   were detected in patients with severe EV-A71
    infection, suggesting that hWARS expression upon IFN-
   stimulation may contribute to EV-A71 pathogenicity (lines
   15-17).

In this segment, I cannot find specific genes that regulate IFNG
   production. The passage focuses on hWARS as an entry factor
   for EV-A71 that is induced by IFN-  .

For the research problem, I would suggest:

1) Focusing initial experiments on diverse genes involved in
   pathways related to IFN-   signaling or production, such as
   the JAK-STAT pathway or genes involved in T helper cell
   differentiation. Observing the effects of perturbing these
   genes could provide insights into which gene types most impact
    IFNG levels.

2) Based on initial observations, select genes that appear to have
    the largest effects on IFNG for follow-up experiments. This
   could help increase the cumulative "hit rate" of genes that
   regulate IFNG production.

3) Considering genes encoding other components of the IFN-
   response pathway in addition to hWARS, such as IFNGR1 and
   IFNGR2 which encode the IFN-   receptor subunits. Perturbing
   these genes may also impact IFNG levels.
```

Genome-wide and targeted CRISPR screens identify RNF213 as a mediator of interferon gamma-dependent pathogen restriction in human cells.

Here is a summary of the information provided in the partial file and strategies for the research problem:

In this segment, I cannot find specific genes to target in the first round of experiments. The passage does not provide that level of detail. However, based on the information provided, transcription factors and lncRNAs associated with interferon gamma regulation and production could be good candidates to screen first. An initial broad screen of diverse candidates from these groups could help identify the types of genes that have the biggest impact on the research problem. The results could then inform a more focused second round of screening.

In this segment, I cannot find specific genes that regulate IFNG production. The passage focuses on hWARS as an entry factor for EV-A71 that is induced by IFN- .

For the research problem, I would suggest:

1) Focusing initial experiments on diverse genes involved in pathways related to IFN- signaling or production, such as the JAK-STAT pathway or genes involved in T helper cell differentiation. Observing the effects of perturbing these genes could provide insights into which gene types most impact IFNG levels.

2) Based on initial observations, select genes that appear to have the largest effects on IFNG for follow-up experiments. This could help increase the cumulative "hit rate" of genes that regulate IFNG production.

3) Considering genes encoding other components of the IFN- response pathway in addition to hWARS, such as IFNGR1 and IFNGR2 which encode the IFN- receptor subunits. Perturbing these genes may also impact IFNG levels.

Correction: In vitro CRISPR screening uncovers CRTC3 as a regulator of IFN- -induced ferroptosis of hepatocellular carcinoma.

Here is a summary of the information provided in the passages and a suggested strategy for the research problem:

In the first passage, activating CRISPR was used to overexpress the lncRNA IFNG-AS1, which led to a greater than 20-fold increase in its expression (lines 6-7). Since IFNG-AS1 is associated with inflammatory bowel disease, perturbing transcription factors involved in regulating IFNG expression could impact interferon gamma levels (lines 4-5).

In the second passage, hWARS was identified as an IFN- -inducible entry factor for EV-A71 (lines 1-3). High levels of IFN- were detected in patients with severe EV-A71 infection, suggesting hWARS expression upon IFN- stimulation may contribute to EV-A71 pathogenicity (lines 15-17).

```
For the research problem:

In this segment, I cannot find specific genes to target in the
   first round of experiments.

I would suggest:

1) Focusing initial experiments on diverse genes involved in
   pathways related to IFN-   signaling or production, such as
   the JAK-STAT pathway or genes involved in T helper cell
   differentiation. Observing the effects of perturbing these
   genes could provide insights into which gene types most impact
    IFNG levels.

2) Based on initial observations, select genes that appear to have
    the largest effects on IFNG for follow-up experiments. This
   could help increase the cumulative "hit rate" of genes that
   regulate IFNG production.

3) Considering genes encoding other components of the IFN-
   response pathway in addition to hWARS, such as IFNGR1 and
   IFNGR2. Perturbing these genes may also impact IFNG levels.


Single-cell functional genomics reveals determinants of
   sensitivity and resistance to natural killer cells in blood
   cancers.
 Here is a summary of the information provided in the file from
    lines 1 to 7:

In this segment, I cannot find specific genes to target in the
   first round of experiments. The passage describes a single-
   cell functional genomics approach to study interactions
   between natural killer (NK) cells and blood cancer cells.
   CRISPR screens identified genes regulating cancer cell
   sensitivity and resistance to NK cell-mediated killing,
   including adhesion-related glycoproteins, protein fucosylation
    genes, and transcriptional regulators, in addition to antigen
    presentation and death receptor signaling pathways (lines
   3-5). A single-cell transcriptomic readout provided insight
   into regulation of IFN-   signaling in cancer cells and NK
   cell activation states (lines 6-7).

For the research problem, I would suggest:

1) Focusing initial experiments on diverse genes involved in
   pathways related to IFN-   signaling or production, such as
   the JAK-STAT pathway or genes involved in T helper cell
   differentiation. Observing the effects of perturbing these
   genes could provide insights into which gene types most impact
    IFNG levels.

2) Based on initial observations, select genes that appear to have
    the largest effects on IFNG for follow-up experiments. This
   could help increase the cumulative "hit rate" of genes that
   regulate IFNG production.

3) Considering genes encoding other components of the IFN-
   response pathway in addition to hWARS, such as IFNGR1 and
   IFNGR2 which encode the IFN-   receptor subunits. Perturbing
   these genes may also impact IFNG levels.


CRISPR screening identifies the deubiquitylase ATXN3 as a PD-L1-
   positive regulator for tumor immune evasion.
 Here is a summary of information that can help with the research
    problem:

In this segment, I cannot find specific genes to target in the
   first round of experiments. The passage describes a CRISPR
   screening technique to identify ATXN3 as a regulator of PD-L1
   expression (lines 1-3). ATXN3 promotes PD-L1 transcription
   through multiple pathways, including in response to IFN-   (
   lines 8-9).

For the research problem, I would suggest:

1) Focusing initial experiments on diverse genes involved in
   pathways related to IFN-   signaling or production, such as
   the JAK-STAT pathway or genes involved in T helper cell
   differentiation. Observing the effects of perturbing these
   genes could provide insights into which gene types most impact
    IFNG levels.

2) Based on initial observations, selecting genes that appear to
   have the largest effects on IFNG for follow-up experiments.
   This could help increase the cumulative "hit rate" of genes
   that regulate IFNG production.

3) Considering ATXN3 as a candidate gene given its role in
   regulating PD-L1 expression in response to IFN-   (lines 8-9).
    Perturbing ATXN3 may impact IFNG levels.

In this segment, I cannot find specific genes that regulate IFNG
   production. The passage focuses on ATXN3 as a regulator of PD-
   L1 expression.



Assistant:
===================claude-1 response (2000)=====================
 1. Reflection: I have identified diverse candidate genes involved
     in IFN-gamma signaling and production based on the provided
    literature. An initial broad screen of these candidates could
     help determine which gene types have the biggest impact on
    IFNG levels before focusing on specific genes in follow-up
    experiments.

2. Research Plan: I will perform an initial CRISPR screen of 128
   diverse candidate genes involved in IFN-gamma signaling or
   production to observe their effects on IFNG levels. Based on
   the results, I will select genes with the largest impacts on
   IFNG for follow-up screening to increase the cumulative hit
   rate.

2. Gene Search: ATF2
```

```
Gene Search Result:RACK1, GTF2H4, RPS27A, FAU, DDX49, DDB1, TSR2,
   WDR43, PFDN2, PAK1IP1, TSR1, PRPF19, CDC45, TRNT1, CDC123,
   RPS19, RPS3A, SS18L2, BUB3, LUC7L3, TXNL4A, TBC1D3B, GPN2, DTL
   , DDX10, CCT4, RPL10, CDC5L, EIF3G, THOC7, RPS15, CTCF, RPS20,
    CCT3, RPL7A, PRPF38B, RPS25, GEMIN5, RPL15, TOMM40, ETF1,
   CHAF1B, HSPA9, GTF2A2, CHMP6, SF1, GINS2, RPL32, XRCC6, ZNHIT2

4. Solution: [’SPC24’, ’CDC123’, ’MED8’, ’DUT’, ’SAP30BP’, ’
   CFAP298’, ’KIF11’, ’NUP93’, ’ATP6V0C’, ’RPL37’, ’SOCS5’, ’
   UBE2M’, ’UBE2L3’, ’STAT6’, ’UBE2S’, ’UBE2E3’, ’UBE2Z’, ’UBE2B
   ’, ’UBE2L6’, ’UBE2D2’, ’TAB3’, ’UBE2R2’, ’UBE2J2’, ’UBE2E2’, ’
   MYD88’, ’UBE2D1’, ’IRAK1’, ’TYK2’, ’RELB’, ’IRAK4’, ’SOCS6’, ’
   UBE2K’, ’UBE2E1’, ’UBE2A’, ’UBE2O’, ’UBE2G2’, ’MEIG1’, ’PPEF1
   ’, ’ZNF582’, ’ATAD1’, ’ZNF282’, ’ZNF280A’, ’GPRASP1’, ’ZNF286A
   ’, ’ZNF331’, ’MAGED2’, ’AKAIN1’, ’TSHZ1’, ’ZNF157’, ’C10orf71
   ’, ’IGFL1’, ’ZNF280D’, ’ODAM’, ’OR51L1’, ’SH2D1A’, ’LUZP4’, ’
   ZNF281’, ’KRBOX4’, ’IFNB1’, ’ZNF292’, ’OR10Z1’, ’MGST1’, ’
   FOXD4L4’, ’ZNF287’, ’CLEC7A’, ’ACSM2B’, ’STX17’, ’POTEE’, ’
   FAM83B’, ’RD3L’, ’ZNF611’, ’ZNF286B’, ’NELL1’, ’ZNF280B’, ’
   ARMCX4’, ’CCDC54’, ’ATXN3L’, ’ZNF283’, ’TBC1D19’, ’ZNF721’, ’
   ZNF729’, ’ZNF182’, ’EQTN’, ’IFNA10’, ’AP1S2’, ’SMCP’, ’STEAP1
   ’, ’GUCY2F’, ’BIN3’, ’ZNF280C’, ’GOLT1A’, ’CXCR4’, ’CCR8’, ’
   CX3CR1’, ’CXCL11’, ’CCR6’, ’IL1A’, ’CCR3’, ’IL1R1’, ’CCL5’, ’
   CXCL5’, ’CXCL2’, ’CXCL12’, ’CCR2’, ’CCR5’, ’CXCR5’, ’CXCR1’, ’
   CCL4’, ’CXCR3’, ’CXCL6’, ’CCR1’, ’CCL2’, ’CCL3’, ’CXCL3’, ’
   CCR7’, ’IL1R2’, ’CXCL8’, ’CXCR6’, ’CXCR2’, ’ATF3’, ’CCL11’, ’
   CXCL1’, ’ATF6B’, ’CXCL13’, ’INCENP’, ’POLR3A’, ’RPL11’, ’ATF2
   ’]

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.


 All tested genes so far and their measured log fold change are:
            Score
Gene
SMAD2     0.101575
UBE2V1   -0.025857
PIAS3     0.085304
LTA      -0.187780
CASP8     0.142034
MAPK10   -0.008140
MAPK9     0.206474
SOCS2    -0.006300
MAPK11    0.022875
PTPN22    0.339770
MAPK8IP1 -0.208230
IL18      0.064003
UBE2V2   -0.095655
IRF7      0.178988
IL12B     0.106245
CXCL9    -0.013024
```

```
TBK1       0.013595
IRF8      -0.284585
SMAD4      0.080494
BCL2       0.016410
SOCS1      0.185722
STAT2     -0.121285
TANK      -0.016199
UBE2N      0.126805
SMAD3      0.122430
STAT5B     0.010015
IRF5      -0.042555
PTPN4      0.024984
NFKBIE     0.215400
IL12A      0.128751
MCL1      -0.062024
CARD11    -0.313452
IRF1      -0.029467
CXCL10    -0.010432
MAPK8     -0.094135
JAK2       0.132444
CARD9      0.117949
MAP2K4    -0.024708
MAPK13     0.146087
TRAF5     -0.140703
PIM2       0.059805
IRF2      -0.094631
MAP3K8    -0.181561
TGFB1      0.032735
MAP3K14   -0.366105
IL6        0.064360
RIPK2     -0.210654
CFLAR     -0.336920
MAPK8IP3   0.160541
TGFBR2     0.013636
SMAD5     -0.129476
NFKBIL1   -0.129973
TRAF3      0.109792
STAT4     -0.282660
TGFBR1    -0.060750
IL15      -0.035326
MAP3K5    -0.224920
STAT1     -0.023106
BIRC2      0.003260
PIM3       0.017550
MAPK12     0.138772
PTPN11     0.136000
TAB2       0.367535
SMAD7      0.284655
PTPN2      0.072855
RIPK4      0.068350
MAP2K3    -0.166365
UBE2D3     0.224675
STAT5A     0.070219
IRF9      -0.126400
MAP3K3     0.220350
NFKBIA     0.344135
IRF3      -0.237845
MAP3K1     0.199285
MAPK8IP2  -0.022033
```

```
IRF4      0.059575
TNF       0.272485
PIAS1    -0.113202
MAPK14   -0.244472
RIPK3     0.010906
BIRC3     0.193885
IL1B     -0.008130
NFKB1     0.125557
PTPN6     0.247855
UBE2D4    0.087250
FADD      0.107290
CASP1    -0.064096
PIAS2     0.247420
MAP2K1   -0.209255
XIAP     -0.342415
BIRC5    -0.082685
PIAS4     0.038910
PTPN1     0.188203
BIRC7    -0.005555
MAP3K15  -0.086644
UBE2C    -0.072150
MAP4K5    0.058382
UBE2Q1    0.001715
UBE2Q2   -0.251546
UBE2T     0.279460
UBE2G1    0.034193
UBE2H     0.287118
UBE2J1   -0.278075
UBE2W     0.111455
 The results for the hits are:
           Score
Gene
TRAF6  -1.865400
SOCS3   0.536160
BCL2L1 -0.460815
MAPK1  -0.449405
NFKB2   1.584300
STAT3  -1.048600
IFNGR1 -0.898070
JAK1   -1.389050
MAPK3  -0.734105
CHUK   -0.833950
IFNGR2 -0.975910
MALT1  -1.063740
RELA   -0.840910
PRKCQ  -0.442575
IKBKE   0.487890
IKBKB  -0.922680
MAP2K2 -0.402665
MAP2K7 -0.446480
TRADD  -0.623995
NFKBIB  0.495810
RIPK1   0.606570
MAP3K7 -1.305750
BCL10  -0.869020
IKBKG  -0.853900

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

```
[’SPC24’, ’CDC123’, ’MED8’, ’DUT’, ’SAP30BP’, ’CFAP298’, ’KIF11’,
   ’NUP93’, ’ATP6V0C’, ’RPL37’, ’SOCS5’, ’UBE2M’, ’UBE2L3’, ’
   STAT6’, ’UBE2S’, ’UBE2E3’, ’UBE2Z’, ’UBE2B’, ’UBE2L6’, ’UBE2D2
   ’, ’TAB3’, ’UBE2R2’, ’UBE2J2’, ’UBE2E2’, ’MYD88’, ’UBE2D1’, ’
   IRAK1’, ’TYK2’, ’RELB’, ’IRAK4’, ’SOCS6’, ’UBE2K’, ’UBE2E1’, ’
   UBE2A’, ’UBE2O’, ’UBE2G2’, ’MEIG1’, ’PPEF1’, ’ZNF582’, ’ATAD1
   ’, ’ZNF282’, ’ZNF280A’, ’GPRASP1’, ’ZNF286A’, ’ZNF331’, ’
   MAGED2’, ’AKAIN1’, ’TSHZ1’, ’ZNF157’, ’C10orf71’, ’IGFL1’, ’
   ZNF280D’, ’ODAM’, ’OR51L1’, ’SH2D1A’, ’LUZP4’, ’ZNF281’, ’
   KRBOX4’, ’IFNB1’, ’ZNF292’, ’OR10Z1’, ’MGST1’, ’FOXD4L4’, ’
   ZNF287’, ’CLEC7A’, ’ACSM2B’, ’STX17’, ’POTEE’, ’FAM83B’, ’RD3L
   ’, ’ZNF611’, ’ZNF286B’, ’NELL1’, ’ZNF280B’, ’ARMCX4’, ’CCDC54
   ’, ’ATXN3L’, ’ZNF283’, ’TBC1D19’, ’ZNF721’, ’ZNF729’, ’ZNF182
   ’, ’EQTN’, ’IFNA10’, ’AP1S2’, ’SMCP’, ’STEAP1’, ’GUCY2F’, ’
   BIN3’, ’ZNF280C’, ’GOLT1A’, ’CXCR4’, ’CCR8’, ’CX3CR1’, ’CXCL11
   ’, ’CCR6’, ’IL1A’, ’CCR3’, ’IL1R1’, ’CCL5’, ’CXCL5’, ’CXCL2’,
   ’CXCL12’, ’CCR2’, ’CCR5’, ’CXCR5’, ’CXCR1’, ’CCL4’, ’CXCR3’, ’
   CXCL6’, ’CCR1’, ’CCL2’, ’CCL3’, ’CXCL3’, ’CCR7’, ’IL1R2’, ’
   CXCL8’, ’CXCR6’, ’CXCR2’, ’ATF3’, ’CCL11’, ’CXCL1’, ’ATF6B’, ’
   CXCL13’, ’INCENP’, ’POLR3A’, ’RPL11’, ’FAU’]

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.


Assistant:
===================claude-1 response (2000)=====================
 1. Critique: The proposed gene list is quite diverse which is
    good for an initial exploratory screen. However, based on the
      results from the previous screen, genes involved in the NF-
    kB and JAK-STAT signaling pathways seem to be enriched for
    hits. I would suggest focusing more on these pathways for the
      next round of screening.

2. Updated Solution: Here is an updated list of 128 genes to test
   next:

1. MYD88,
2. IRAK1,
3. IRAK4,
4. TYK2,
5. RELB,
6. SOCS6,
7. TAB3,
8. NFKBIB,
9. RIPK1,
10. BCL10,
11. IKBKG,
12. TRAF6,
13. MAPK1,
14. NFKB2,
15. STAT3,
16. IFNGR1,
17. JAK1,
18. MAPK3,
19. CHUK,
20. IFNGR2,
21. MALT1,
22. RELA,
23. PRKCQ,
24. IKBKE,
25. IKBKB,
26. MAP2K2,
27. MAP2K7,
28. TRADD,
29. TLR9
30. TLR7
31. TLR8
32. TLR3
33. IRAK2
34. UBE2S
35. UBE2E3
36. UBE2Z
37. UBE2B
38. UBE2L6
39. ZNF281
40. ZNF292
41. ZNF287
42. ZNF611
43. ZNF286B
44. ZNF280B
45. ZNF283
46. ZNF721
47. ZNF729
48. ZNF182
49. ZNF280C
50. SOCS5
51. SPC24
52. CDC123
53. MED8
54. DUT
55. SAP30BP
56. CFAP298
57. KIF11
58. NUP93
59. ATP6V0C
60. RPL37
61. UBE2M
62. UBE2L3
63. STAT6
64. TAB3
65. UBE2R2
66. UBE2J2
```

```
67. UBE2E2
68. CLEC7A
69. ACSM2B
70. MYD88
71. UBE2D1
72. IRAK1
73. TYK2
74. RELB
75. IRAK4
76. SOCS6
77. UBE2K
78. UBE2E1
79. UBE2A
80. UBE2O
81. UBE2G2
82. MEIG1
83. PPEF1
84. ATAD1
85. MAGED2
86. TSHZ1
87. C10orf71
88. IGFL1
89. ODAM
90. IFNB1
91. MGST1
92. FOXD4L4
93. ZNF287
94. STX17
95. POTEE
96. FAM83B
97. RD3L
98. ZNF611
99. ZNF286B
100. NELL1
101. ZNF280B
102. ARMCX4
103. CCDC54
104. ATXN3L
105. ZNF283
106. TBC1D19
107. ZNF721
108. ZNF729
109. ZNF182
110. EQTN
111. IFNA10
112. AP1S2
113. SMCP
114. STEAP1
115. GUCY2F
116. BIN3
117. ZNF280C
118. GOLT1A
119. CXCR4
120. CCR8
121. CX3CR1
122. CXCL11
123. CCR6
124. IL1A
125. CCR3
```

```
126. IL1R1
127. CCL5
128. CXCL5
```

The subsequent steps follow the same format, the agent predicts a batch of genes based on previous round predictions and hit results, literature review, and diverse gene tool. The prediction list is updated using the critic agent which is not provided with any additional information on literature review or gene diversity, but only the research problem, previous round experimental results, and current proposed gene list by the primary agent.