

BIODISCOVERYAGENT: AN AI AGENT FOR DESIGNING GENETIC PERTURBATION EXPERIMENTS

Anonymous authors

Paper under double-blind review

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 yields a particular phenotype (e.g. cell growth). However, the costs involved in each experiment limits the number of perturbations that can be tested. Here, we develop *BioDiscoveryAgent*, an AI agent that can strategically design genetic perturbation experiments to enhance the detection of perturbations that induce desired phenotypes. Our AI agent is based on large language models, which have rich biological knowledge, and generate explainable rationales while selecting genes to perturb. *BioDiscoveryAgent* attains an average of 23% improvement compared to existing Bayesian optimization baselines in detecting desired phenotypes across five datasets. This includes one dataset that is unpublished and therefore guaranteed to not appear in the language model’s training data.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 this setting. Overall, our approach represents a simple new paradigm in computational design of biological experiments, aimed at augmenting scientists’ capabilities and accelerating scientific discovery.

1 INTRODUCTION

An essential challenge in discovering new drugs is identifying the right biological target that leads to a specific phenotype (biological effect) Scannell et al. (2012). The failure to identify physiologically effective targets is a major cause of drug development failures in clinical trials, more so than the ability to find a suitable drug for an identified target (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*, have shown significant promise in various areas including drug target discovery, combating drug resistance, elucidating disease mechanisms, cell engineering, gene therapy, and immunotherapy Goodman et al. (2022); Kalos et al. (2011); Mamedov et al. (2023); Lim (2022).

Typically, a perturbation screen targets all known protein-coding genes, which amounts to approximately 19,000 genes in humans, out of which only a few hundred are expected to show the desired phenotype. However, 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 to focus on genes predicted to have 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, this selection is difficult, demanding both domain-specific knowledge and the ability to interpret and reason over experimental outcomes. Recent work has approached this problem using a Bayesian optimization framework, but this requires training bespoke machine learning models on small handcrafted datasets and the model results are not always interpretable Mehrjou et al. (2021); Lyle et al. (2023).

In this paper, we apply LLM-based AI agents to this task. LLMs are capable of learning biomedical information from the scientific literature and can recall this knowledge when required (Moor et al., 2023; Wang et al., 2023). Furthermore, AI agents powered by LLMs are particularly skilled at

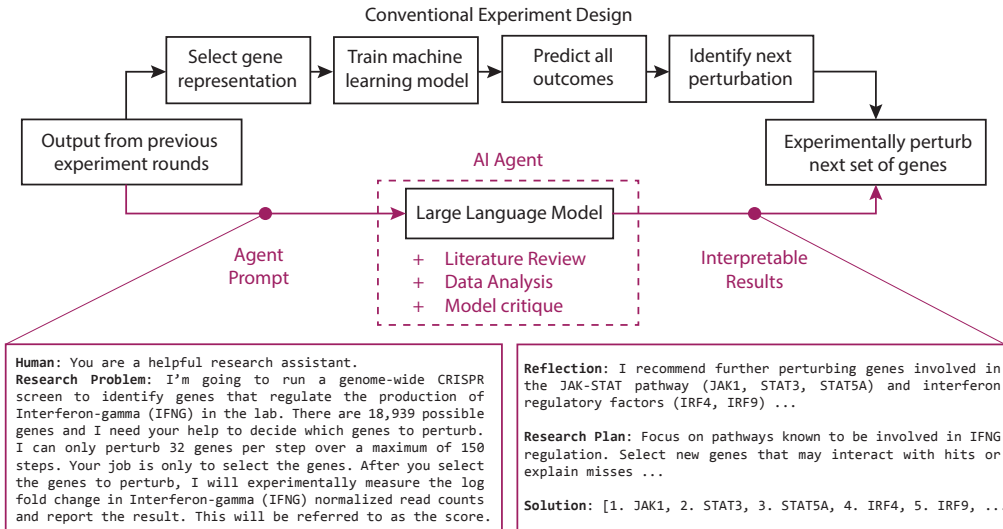


Figure 1: Comparison of AI agent based approach (BioDiscoveryAgent) to conventional machine-learning methods for experiment design.

logical thinking and assessing various alternatives, which makes them useful in the planning and execution of scientific experiments (Liu et al., 2023; Huang et al., 2023b).

Making use of these capabilities, we develop *BioDiscoveryAgent*, an AI agent that designs genetic perturbation experiments using a large language model in conjunction with a suite of tools (Figure 1). The process entails prompting Claude v1 at each experimental round with both a task description and data from previous rounds (i.e., phenotypes from gene perturbations) to select the next set of genes for perturbation. Additionally, the agent is provided with tools to read biomedical literature, analyze biological datasets, and critically evaluate its own predictions. BioDiscoveryAgent outperforms Bayesian optimization methods, identifying 23% more successful perturbations after five experimental rounds. It also shows enhanced performance in the new setting of predicting responses to combinatorial gene perturbations. Overall, BioDiscoveryAgent utilizes its vast biological knowledge along with the ability to reason over insights from previous experiments to offer a straightforward and interpretable method for planning genetic perturbation experiments.

2 NOTATION

Consider a set of genes \mathcal{G} where $|\mathcal{G}| \approx 19000$. Each perturbation experiment f perturbs a gene $g \in \mathcal{G}$ which results in a phenotypic response $x_g \in \mathbb{R}$. Thus, f is a function $f : \mathcal{G} \rightarrow \mathcal{X}$. The goal of these experiments is to identify those genes that upon perturbation produce a desired phenotype. Such genes can then be further studied as potential targets for drug discovery. These genes are often referred to as *hits* for that phenotype in the screen. We define the set of K top hits as \mathcal{G}_K where $\mathcal{G}_K \subset \mathcal{G}$. The set \mathcal{G}_K is defined such that the members $\{g_1, g_2, \dots, g_k\}$ have the largest absolute value of $\{f(g_1), f(g_2), \dots, f(g_k)\}$.

In addition to the experimental setting of a single gene being perturbed to study its phenotypic effect, we also present a new and more computationally challenging problem setting of predicting two-gene combinations to perturb. Let $\mathcal{G}^2 = \mathcal{G} \times \mathcal{G}$ denote the set of all possible gene pairs that can be perturbed. This set is formally defined by the Cartesian product of \mathcal{G} with itself. 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 $f' : \mathcal{G}^2 \rightarrow \mathcal{X}$. This function f' maps each gene pair to a phenotypic response in the space \mathcal{X} , which is indicative of the combined effect of perturbing both genes simultaneously.

We also define an AI Agent that is capable of retrieving information relevant to the relevant biological setting and phenotype of interest. This agent is queried with a prompt that describes the nature of the problem at each step, including general information to describe the experimental setting and the biological specifications of the experiment being performed. The goal of the AI agent at each step i is to identify a batch of t genes to test in that round such that the total number of hits at the end of the S rounds is maximized. The agent has access to multiple resources to aid in this decision-making beyond knowledge stored within its weights. These include access to the biomedical literature, a second agent to critique its predictions, and the ability to analyze tabular datasets containing gene features relevant to the experimental setting.

3 PROBLEM FORMULATION

We formulate this problem as a closed-loop experimental design task, where an AI agent determines which genes to perturb in each experiment cycle i , using the results to plan subsequent cycles. The objective is to maximize the number of *hit* genes or gene combinations that produce a desired phenotypic effect. In particular, we test our model on two real-world experimental settings

- **Single-gene perturbation:** In this setting, the goal is to identify the most single genes g that induce a strong phenotypic effect $f(g)$ upon perturbation over a predefined number S of experimental rounds. At each round, the AI agent chooses t new genes to test. The phenotypic score from perturbing these genes is then revealed to the model along with the number of hits that were successfully identified. In the next round, the agent has access to all phenotypic scores for genes tested in previous rounds. At the end of S rounds, the total hit rate is computed as the fraction of successful hits identified out of all possible hits in that dataset.
- **Two-gene perturbation:** Here, each experiment consists of two genes perturbed in a single cell simultaneously. The goal in this setting is to identify pairs of genes whose combined perturbation results in the desired phenotypic outcome. In particular, we are looking for combinations that produce synergistic or buffering interactions that significantly differ from what would be expected based on their individual perturbation response. The AI agent is given access to all phenotypic scores for gene combinations tested in previous rounds.

4 BIODISCOVERYAGENT

We create an AI agent, named BioDiscoveryAgent, using a Large Language Model (LLM) that automates the scientific discovery process, which includes hypothesizing, planning and interpreting results. The process involves prompting the LLM with a detailed prompt based on the accumulated knowledge and observations from previous rounds. This prompt creation draws from established methods in developing other LLM-based agents, such as pre-action reasoning (Yao et al., 2022), reflective thinking (Shinn et al., 2023), and stepwise planning (aut, 2023). An overall algorithm is shown in Algorithm 1.

4.1 AGENT RESPONSE FORMAT

A critical aspect is defining the response format to guide the LLM’s thought process before suggesting actions. As shown in the prompt in appendix A, we direct the LLM to structure its responses into several parts: `Reflection`, `Research Plan`, `Solution`. `Research Plan` helps in effective planning and monitoring progress. Through the `Reflection` and `Research Plan` categories, the model is able to provide additional rationale and reasoning behind a particular prediction. This also helps to rule out predictions that may be hallucinations or not well-motivated.

4.2 GENE LIST PROCESSING AND SUMMARIZATION

When choosing genes for perturbation, it’s often challenging for BioDiscoveryAgent to identify an appropriate list of genes within the constraints of the experiment. However, including all pertinent genes, which can exceed 18,000, in every prompt iteration is not efficient. At the same time, we

Algorithm 1 BioDiscoveryAgent: AI Agent for Biological EXperiment Design

Input: Number of experimental rounds S , Number of genes to test per round t
Output: Set of genes or gene combinations to perturb
for $i = 1, \dots, S$ **do**
 LLM searches and keeps a running summary of scientific literature relevant to the problem and adds it to the prompt
 Prompt LLM to select t new perturbations (genes or gene combinations) to test
 while $x > 0$ perturbations are invalid **do**
 Prompt LLM to select x new perturbations to test
 LLM generates structured response with Reflection, Research Plan, Gene Search, Solution entries
 if too many prompt attempts **then**
 Summarize list of all valid perturbations and feed as prompt
 end if
 end while
 Employ AI critic for quality enhancement
 Get phenotypic score $f(g)$ for each perturbation
 Update hit rate based on successful hits
 Summarize text for context window management
end for
Calculate total hit rate at the end of S rounds

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. This enables it to draw on its comprehensive understanding of biology freely. Then, we filter the suggested genes for relevance to the specific inquiry and prepare a new prompt for any missing genes. After three rounds, BioDiscoveryAgent summarizes the list of all remaining genes, ensuring it covers a broad range of biological pathways and functions. This summarized list if added to the prompt to aid in gene selection.

Moreover, after numerous experiments, the historical data and outcomes may also exceed the LLM’s processing capacity. To address this, we employ a similar summarization technique to keep the information in future prompts focused and relevant.

4.3 AGENT TOOLS

The primary mechanism by which BioDiscoveryAgent interacts with the human user is through natural language. A large language model (LLM) provides the capability to understand the prompts and generate responses. The LLM has been trained on large corpus of human data including the biological literature. In addition to its knowledge, the agent relies on additional tools to augment its understanding (See Appendix B for full implementation details):

- **Literature search:** We use a search tool to query the literature for papers that are relevant to the given biological experiment. Once the top 5 papers have been identified, the agent summarizes their information, attaches it to the prompt and uses it to identify additional genes to perturb for the given experimental round. The citations to these papers are retained and returned along with the model predictions.
- **AI critic:** Large language models are very sensitive to the prompt they are presented with. Past research has shown very different performance depending upon the setting in which they are queried. Thus, an LLM prompted to behave as an agent for a human researcher may behave very differently from one that is prompted to mainly critique the prediction made by another LLM. To benefit from these contrastive prompting strategies, we make use of an AI critic to identify mistakes and enhance the quality of the final prediction made by the agent.
- **Gene search based on features:** LLMs are trained on text-based data and do not have access to many biological databases that are stored in the form of tabular data. We provide

Model	Top-K Recall				
	Schmidt22 (IFNG)	Schmidt22 (IL2)	Steinhart [†]	Scharenberg22*	Carnevale22
Random	0.037 ± 0.013	0.031 ± 0.002	0.033 ± 0.003	0.160 ± 0.028	0.036 ± 0.001
BDAgent (Vanilla)	0.067 ± 0.010	0.089 ± 0.014	0.110 ± 0.023	0.292 ± 0.064	0.038 ± 0.005
BDAgent (All Tools)	0.095 ± 0.018	0.122 ± 0.037	0.114 ± 0.009	0.314 ± 0.028	0.054 ± 0.012
Soft Uncertain	0.037 ± 0.006	0.037 ± 0.006	0.034 ± 0.007	0.205 ± 0.006	0.031 ± 0.006
Top Uncertain	0.057 ± 0.007	0.072 ± 0.014	0.058 ± 0.010	0.294 ± 0.030	0.037 ± 0.005
Margin Sample	0.054 ± 0.006	0.061 ± 0.009	0.054 ± 0.013	0.285 ± 0.019	0.036 ± 0.003
Coreset	0.072 ± 0.007	0.102 ± 0.005	0.069 ± 0.008	0.243 ± 0.031	0.047 ± 0.006
Badge	0.060 ± 0.008	0.077 ± 0.008	0.042 ± 0.017	0.258 ± 0.032	0.044 ± 0.006
Kmeans Embed.	0.045 ± 0.004	0.064 ± 0.007	0.028 ± 0.011	0.170 ± 0.032	0.036 ± 0.004
Kmeans Data	0.048 ± 0.005	0.074 ± 0.009	0.025 ± 0.012	0.281 ± 0.042	0.039 ± 0.004

Table 1: Comparison to machine learning baselines for 1-gene perturbation experiments. at round 5. *Note that we use batch size 32 for Scharenberg22 due to the much smaller number of pertinent genes of 1061. [†] Steinhart is an unpublished dataset.

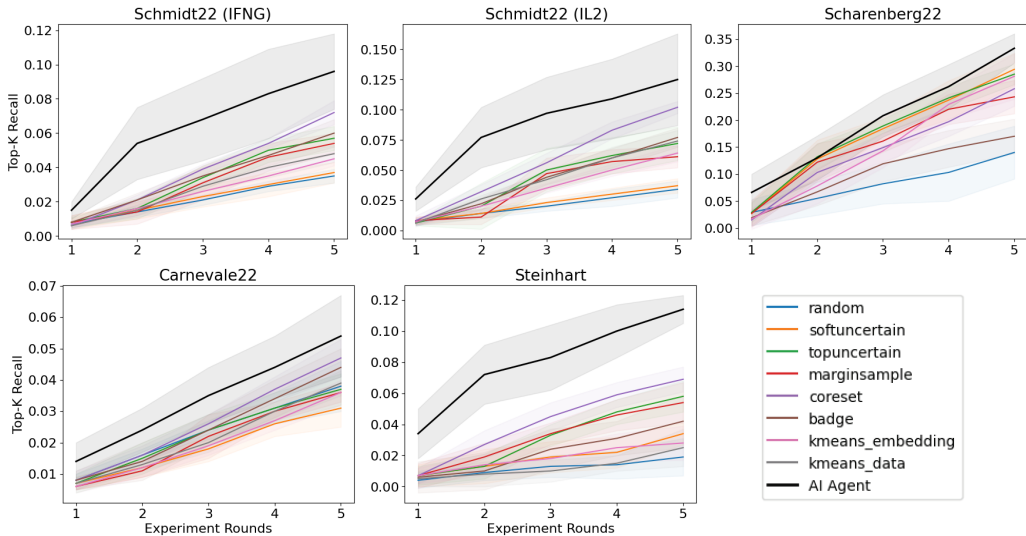


Figure 2: Comparison of Top-K Recall of the BioDiscoveryAgent (black line) to other machine learning Bayesian optimization baselines

the AI agent with the ability to search for genes similar or dissimilar to a given gene based on cosine similarity among the provided gene features.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We assess model performance using data from past genetic perturbation experiments. This involves using the outcomes of these experiments as feedback for the AI agent, i.e., providing feedback from the experiment to the AI agent is equivalent to retrieving from a dataset. For each perturbation screen s , we calculate the hit rate as the proportion of discovered hits in the current experiments to the total hits identified in 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 hit rate monitored throughout (Figure 1).

5.2 DATASETS AND BASELINES

We make use of five different datasets spread across different cell types, publication dates and data generation sites. We explored two datasets that analyzed T-cell proliferation under different conditions. These were taken from two genome-wide perturbation screens performed on T-cells with the readout being the change in production of IL-2 and IFN- γ (Schmidt et al., 2022). An additional dataset includes multiple screens for identifying genes linked to T-cell dysfunction Carnevale et al. (2022). Unpublished data from Steinhart et al. studied the impact of genome-wide perturbations on CAR-T cell proliferation. We also used perturbation data from pancreatic islet cells Scharenberg et al. (2023).

For baselines, we use the methods implemented in the GeneDisco benchmark (Mehrjou et al., 2021). Every baseline includes a bayesian multi-layer perceptron M for predicting experimental outcomes using gene features. This is then combined with one of the following acquisition functions for Bayesian optimization:

- **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.

5.3 BIODISCOVERYAGENT OUTPERFORMS BASELINES IN OPTIMIZING HIT RATE FOR 1-GENE PERTURBATION EXPERIMENTS

We compare the performance of vanilla (without tools) and full (with all tools) BioDiscoveryAgent against several machine learning baselines, as well as random sampling. As shown in Table 1, BioDiscoveryAgent with all tools significantly outperform baseline methods by 23% on average for hit rates at round 5, while the vanilla version matches baseline performance. As shown in Figure 4.3, this gap is especially large at earlier rounds, where the large language model can leverage its prior knowledge in biology to select promising gene candidates right away, while the baseline approaches suffer from the cold start problem.

5.4 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. Concretely, we use the Horlbeck dataset where there are 100,576 combinations compare to only 18,939 pertinent genes in 1-gene perturbation experiment Horlbeck et al. (2018). As shown in table 2, the BioDiscoveryAgent significantly outperforms the random sampling baseline by 130% on average.

5.5 BIODISCOVERYAGENT ACCOUNTS FOR PRIOR KNOWLEDGE AND OBSERVATIONS IN DECISION-MAKING

Next, we investigate the use of prior knowledge versus observations from previous experiments in BioDiscoveryAgent’s decision-making for designing subsequent experiments (Table 3). 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

Step	Cumulative Number of Hits				
	1	2	3	4	5
Model	Dataset: Horlbeck (n=100,576)				
Random	2.6 ± 1.43	5.7 ± 2.83	8.9 ± 3.67	12.8 ± 3.74	16.4 ± 3.8
BDAgent (Vanilla)	5.50 ± 4.42	14.33 ± 5.76	21.67 ± 7.56	30.50 ± 7.34	32.67 ± 7.36

Table 2: 2-gene experiment design task at round 3. Batch Size: 30.

Dataset		Schmidt22 (IL2)			Schmidt22 (IFNG)		
Rounds		10	20	30	10	20	30
Model	Setting	Avg. Hit Rate (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 3: Model uses both prior knowledge and experimental output to make predictions. Batch Size: 32

any experiment results, effectively directly selecting 32 * 10 genes all together; 3) Only Observation, where the agent is unaware of the current experiment’s goal but still conditioned on observations.

Results show that Full Observation outperforms both No Observation and Only Observation, highlighting the significance of integrating prior knowledge and observations (Table 3, 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 scenarios without access to observations (Figure 3b). This suggests that observations significantly influence BioDiscoveryAgent’s decision-making, leading to more uniform choices across separate trials.

5.6 BIODISCOVERYAGENT IS HIGHLY CUSTOMIZABLE WITH DIFFERENT TOOLS

BioDiscoveryAgent, which primarily functions through prompting an LLM, can be augmented with different tools as shown in table 1. We performed an ablation analysis on performance using different tools (Table 4) as compared to the vanilla setting where the agent uses no tools:

- **Literature search:** In this case, the agent only uses literature search tool. We notice that literature search does not always help, mainly because the agent might get fixated with a few simple keywords and irrelevant papers, which distracts the agent from a more effective search. However, literature search provides interpretable citations for the gene prediction process, which can be valuable for scientists.
- **Critique:** The agent only uses AI critique tool. Critique improves performance slightly over vanilla model mainly by diversifying the original genes predicted or through concentrating the predictions towards specific gene sets.
- **Similar:** The agent only uses the gene search tool to search for similar genes. Using a gene similarity search based on specific features significantly enhances performance, particularly for IFNG and Scharenberg22. We believe this is because gene similarity is based on experimentally obtained gene feature sets that are usually stored in a tabular format and not fully represented in the text of scientific papers. Thus, the vanilla LLM may not have access to this information.
- **Diverse:** The agent only uses the gene search tool to search for dissimilar genes, with minor instructions on how to explore these diverse genes. Providing dissimilar gene based on features encourages BioDiscoveryAgent to explore diverse genes, as partially inspired by coreset type of method. We observe that such a method performs differently over different types of datasets, potentially due to the different properties of the target task.

Model	Top-K Recall				
	IFNG	IL2	St. (CRISPRa)	Scharenberg22*	Carnevale22
Dataset					
Random	0.037 ± 0.013	0.031 ± 0.002	0.033 ± 0.003	0.160 ± 0.028	0.036 ± 0.001
BDAgent (Vanilla)	0.067 ± 0.010	0.089 ± 0.014	0.110 ± 0.023	0.292 ± 0.064	0.038 ± 0.005
BDAgent (Literature)	0.052 ± 0.012	0.069 ± 0.006	0.073 ± 0.014	0.232 ± 0.025	0.023 ± 0.009
BDAgent (Critique)	0.069 ± 0.007	0.089 ± 0.008	0.130 ± 0.026	0.341 ± 0.041	0.044 ± 0.005
BDAgent (Similar)	0.083 ± 0.015	0.087 ± 0.016	0.111 ± 0.010	0.351 ± 0.044	0.045 ± 0.008
BDAgent (Diverse)	0.099 ± 0.015	0.162 ± 0.021	0.064 ± 0.023	0.286 ± 0.013	0.048 ± 0.012
BDAgent (All Tools)	0.095 ± 0.018	0.122 ± 0.037	0.114 ± 0.009	0.314 ± 0.028	0.054 ± 0.012

Table 4: Comparison of BioDiscoveryAgent with different tools for 1-gene perturbation experiments. at round 5. *Note that we use batch Size 32 for Scharenberg22 due to the much smaller number of pertinent genes (1061).

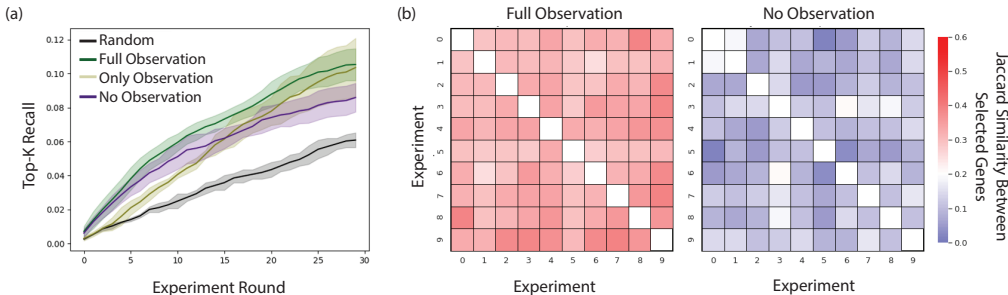


Figure 3: Agent accounts for both prior knowledge and observations in decision-making

- **All Tools:** The agent uses all tools with gene search searching for dissimilar genes. Using all tools provides a balanced method that consistently achieves high performance and outperforms baselines (Table 1).

5.7 BIODISCOVERYAGENT PROVIDES INTERPRETABLE PREDICTIONS WITH REFERENCES TO THE LITERATURE

Finally, our BioDiscoveryAgent can provide interpretable predictions through the reflection, research plan, as well as the citations produced by the literature search tool. We show several specific qualitative examples (see Appendix D).

For instance, in the initial steps, where the agent recommends more diverse genes, the critique agent guides the main agent to focus more on genes relevant to the task. The critique agent suggests a specific rationale for choosing particular kinds of genes which helps in improving interpretability and opens avenues for human-in-the-loop feedback by a subject-matter expert. Furthermore, the agent also grounds its predictions by referring to the scientific literature.

6 DISCUSSION

Our approach represents a new paradigm in computational design of biological experiments, aimed at augmenting scientists’ capabilities and accelerating scientific discovery. Traditionally, a multi-stage pipeline, as illustrated in Figure 1, relies on an uninterpretable acquisition function and a machine learning model. This model requires manual design and frequent retraining with selected gene features at each experimental phase. In contrast, we demonstrate that an AI agent, powered by a large language model, can streamline experimental design into a single step.

Furthermore, the agent is equipped with prior biological knowledge, solving the cold start problem at the beginning of the experimental design, while also being able to effectively use observational data as shown in section 5.5. In addition, the AI agent can provide interpretable reasoning for the ex-

perimental choices being made while also meaningfully interfacing with the scientific literature and existing datasets. The overall result is a research assistant that can accelerate the pace of biological discovery

Our approach is designed to be broadly applicable, extending beyond gene perturbation experiments. LLM can analyze arbitrary textual data, making them capable of processing various scientific hypotheses and designs, including but not limited to lists of genes. The supplied tools of literature review and self critique are also general to arbitrary research. For instance, this AI research assistant could be employed in fields like chemistry and material science, leveraging the extensive knowledge LLMs have shown in these areas.

7 RELATED WORK

Artificial intelligence has shown tremendous promise in various domains of scientific research from realistic simulations of human behavior (Park et al., 2023b) to searching over the space of mathematical functions (Romera-Paredes et al., 2023). While several works have explored the utility of these models in enhanced mining and querying of the scientific literature (Lála et al., 2023; Schick et al., 2023) they have also shown capabilities in general research tasks like analyzing large datasets, reasoning about data and generating reports (Shakke & Zhang, 2023). Text-based molecule encodings have used pre-trained large language models to aid in the generation of new molecules with desired properties (Liu et al., 2023; Li et al., 2023). Another interesting line of work has been closed loop lab-based experimentation. Notable work in this space has been performed in chemical synthesis (Boiko et al., 2023) and materials discovery (Wang et al., 2023). Within the biological space, (Park et al., 2023a) showed that large language models are able to capture meaningful information about biological pathways and processes. LLMs have also shown value in simulating biological processes across diverse scales (Schaefer et al., 2023). While it is still difficult to evaluate the relative intelligence of different LLMs (Mitchell, 2023), we were more interested in general abilities to generate meaningful scientific hypotheses. Several previous works have explored agent-based applications that embed LLMs inside a larger software ecosystem to provide access to memory and external knowledge (Peng et al., 2023).

The idea of relying on artificial intelligence to generate hypotheses for functional genomics experiments in particular is not new (King et al., 2004). Indeed, the vastness of the space and the combinatorial explosion in possible experimental interventions highlights the natural value of such an approach. We consider the task of optimizing the design of genetic perturbation experiments. Previous work has highlighted the challenges in applying machine learning approaches to this problem and their over reliance on large datasets of experimental outcomes as well as meaningful featurization of interventions (Mehrijou et al., 2021; Lyle et al., 2023).

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.
- 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.
- 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.
- P. Langley. Crafting papers on machine learning. In Pat Langley (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.
- Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *arXiv preprint arXiv:2306.06615*, 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, Jiong Xiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023.
- 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.
- Melanie Mitchell. How do we know how smart ai systems are?, 2023.
- 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.
- 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.
- Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- Laralynne Przybyla and Luke A Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 23(2):89–103, 2022.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, pp. 1–3, 2023.
- Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, pp. 1–9, 2023.
- 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-Remaih. 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 Dessi, 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.
- 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.
- 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.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022.

A PROMPT

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

You are a scientist working on problems in drug discovery.

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

B 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 rationale for adding those tools.

1. **AI Critique:** At every round, once BioDiscoveryAgent comes up with a batch of genes to be tested, a critique agent (which is also an LLM) is prompted to critique the choice of the main agent and it can change some or all the genes in the batch and come up with a new set of genes. Having such an agent improved hit rate performance giving the system an additional time to reflect on it's reasoning. The system prompt for the critique 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 rate.

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 critique 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 a lot of prior knowledge that could be leveraged to design experiments for that particular domain. A scientist typically reads literature relevant to a problem at hand, makes hypothesis building on top of related work, and cites it, which helps accelerate the pace of scientific discovery. We aimed to enable our BioDiscoveryAgent with a similar tool that would allow it to search for relevant papers on the web, learn from them, and incorporate its learnings in designing the experiments.

We searched papers using the PubMed API (<https://github.com/gijswobben/pymed>) which contained the most relevant literature for the gene perturbation experiments that the agent was asked to design. At every round, the agent was asked to generate a search term to query the API which was based on the original research problem and the current progress. The top 5 papers were retrieved and their contents were summarized conditioning on the research problem. The summaries of these papers were provided to the agent before filling in the schema of its research plan, reflection, and solution. The summaries were accumulated over time, and hence, the agent had access to all literature surveys done in the previous rounds to propose a batch of genes as its solution for that particular round.

Having the presence of a literature review tool helped provide more interpretability and improved grounding as the agent used to often cite papers it had surveyed before predicting some gene to be a part of the batch. Overall, however, having a literature review tool made the agent less exploratory and its reasoning was biased heavily by what papers were retrieved. The inability to come up with interesting queries for the literature survey API, and having no additional re-ranking on the API outputs further hurt the diversity of papers the agent was presented with. This raised some interesting questions about how should scientific literature be used by an AI agent in the most effective way which is left to future work.

3. **Gene Search:** While large language models (LLMs) are trained on large corpora of text, a lot of biological features are in the form of numerical features which could further help the agent in improving its performance. We provide the agent with a tool where it can query with a gene name and ask for the most similar or dissimilar genes to the one queried. This is done by simply taking the inner product of gene features with the feature of the query gene and sorting them based on asking for similarity or dissimilarity. We found that having the presence of a tool to find dissimilar genes helped a lot in improving the diversity of the genes predicted by the agent, especially in the earlier rounds which led to better exploration and subsequently better hit rates. We retrieved the top 10 genes for similarity and the top 50 genes for dissimilarity.

C 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

D INTERPRETABILITY EXAMPLES

Here we show some examples that demonstrate the interpretability of BioDiscoveryAgent for experimental design over Bayesian optimization methods:

1. In the initial steps, where the agent tried out a lot of random genes, the critique 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. Having such a rationale for choosing particular kinds of genes helps in improving interpretability and could also open avenues for human-in-the-loop feedback provided to the agent in the form of natural language from a subject expert. Furthermore, the agent also grounds its predictions by referring to the scientific literature, as shown below:

Human tryptophanyl-tRNA synthetase is an IFN- γ -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- γ)-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- γ , 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).