

LLMs for Bayesian Optimization in Scientific Domains: Are We There Yet?

Anonymous ACL submission

Abstract

Large language models (LLMs) have recently been proposed as general-purpose agents for experimental design, with claims that they can perform in-context experimental design. We evaluate this hypothesis using open-source instruction-tuned LLMs applied to genetic perturbation and molecular property discovery tasks. We find that LLM-based agents show no sensitivity to experimental feedback: replacing true outcomes with randomly permuted labels has no impact on performance. Across benchmarks, classical methods such as linear bandits and Gaussian process optimization consistently outperform LLM agents. We further propose a simple hybrid method, *LLM-guided Nearest Neighbour (LLMNN)* sampling, that combines LLM prior knowledge with nearest-neighbor sampling to guide the design of experiments. LLMNN achieves competitive or superior performance across domains without requiring significant in-context adaptation. These results suggest that current open-source LLMs do not perform in-context experimental design in practice and highlight the need for hybrid frameworks that decouple prior-based reasoning from batch acquisition with updated posteriors.

1 Introduction

Experimental design enables data-efficient scientific discovery (Shields et al., 2021; Ueno et al., 2016; Lei et al., 2021; Huan et al., 2024; Jain et al., 2023). In domains such as biology and chemistry each experiment can be costly or time-consuming so we need methods that select the most informative candidates under a constrained budget. Classical approaches typically adopt a Bayesian framework (Lindley, 1956; Houlsby et al., 2011): they construct a surrogate model of the response function, update this model with new data, and select future experiments by optimizing an acquisition function such as expected improvement or information gain (Gal et al., 2017; Kirsch et al., 2019).

Recent work has proposed using LLMs to replace these components, offering a unified interface that can incorporate prior knowledge, reason over experimental history, and directly select candidates via prompting. Examples include BioDiscoveryAgent (BDA) (Roohani et al., 2024) for genetic perturbation design and LLAMBO (Liu et al., 2024) for hyperparameter optimization. These systems are built on proprietary LLMs (e.g., Claude 3.5 Sonnet (Anthropic, 2024) or OpenAI’s GPT-3.5 (OpenAI, 2023) and o1 (OpenAI, 2024)) and design experiments by iteratively prompting the LLM with prior experimental outcomes, and rely on in-context learning to guide future experiment selection.

In domains that are well-studied, the priors from the LLMs offer clear benefits over the traditional methods for the first round of experimentation because the LLMs can leverage information from the pretraining corpus to select actions. However, it is less clear whether in-context learning alone (without finetuning) leads to good action selection. This paper investigates whether such LLM-based approaches, when implemented with open-source models and without external tool use, can perform effective experimental design. Specifically, we ask, **do off-the-shelf, instruction-tuned LLMs exhibit strong in-context experiment design abilities when prompted with experimental history?**

We address this question using the BioDiscoveryAgent pipeline implemented with two publicly available LLMs—LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2-7B-Instruct (Yang et al., 2024)—applied to experimental design tasks in two domains: single-gene perturbation and molecular property prediction. To evaluate in-context learning behavior, we conduct ablation studies that compare the standard BioDiscoveryAgent to a variant receiving randomly permuted feedback (BDA-Rand), removing any correspondence between past actions and outcomes. Our experi-

mental analysis reveals the following findings:

1. **LLMs are insensitive to feedback.** Across all datasets and models (including Claude Sonnet 3.5), BDA and BDA-Rand perform comparably. In some cases, the BDA-Rand even performs slightly better. This suggests that the tested LLMs’ selection behavior is determined primarily by prior knowledge rather than adaptation to feedback.
2. **LLMs underperform classical baselines.** Linear UCB and Gaussian process-based Bayesian optimization methods, given access to the same candidate embeddings, consistently outperform BDA on both domains.
3. **A simple heuristic performs better search.** We introduce **LLM-guided Nearest Neighbor** (LLMNN), a method that prompts the LLM to propose seed candidates and then selects batches via nearest-neighbor expansion in embedding space. LLMNN also receives in-context feedback but outperforms BDA and matches or exceeds classical methods across benchmarks.

In summary, we make following contributions,

- We provide a diagnostic evaluation of in-context experimental design using randomized-feedback ablations.
- We benchmark open-source LLMs against Bayesian optimization and bandit-based methods across five gene perturbation and three molecular property prediction datasets.
- We show that open-source and proprietary instruction-tuned LLMs do not leverage experimental feedback to perform in-context experimental design in practical experimental design tasks.
- We introduce a simple hybrid method, LLMNN, which leverages LLM prior knowledge for exploration and embedding-based nearest-neighbor sampling for exploitation.

These results indicate that LLMs, while encoding valuable domain priors, still need explicit mechanisms that enable posterior updating and selection for efficient experimental design. Hybrid approaches that combine LLM priors with classical exploration strategies offer a promising direction for scalable, general-purpose experiment design.

2 Problem Statement

Consider the following real-world experimental settings across two domains:

Single Gene Perturbation. Let’s say we have the ability to knock out a single gene at a time from the human cell to induce a certain desired phenotypic response. Correspondingly, for every knockout, we are able to measure the phenotypic response value of perturbing the gene. However, exhaustive trials with the whole set of protein-coding genes (> 19000) are infeasible due to both time and budget constraints. Thus, *how do we design perturbation experiments to identify genes that produce the phenotype under tight budget constraints and a sequential experiment protocol?*

Chemical Properties of Molecules. Similarly, say given a large library of molecules, we seek to identify the molecules that exhibit high values for a given property. For every molecule tried, we can obtain its property value. *How do we effectively screen the library to identify top-performing molecules for the given property, under similar constraints as above?*

More formally, both these settings are examples of closed-loop experiment design, which is our center of focus in this work. At any experimental round, the agent determines which candidates to investigate next, given the results from all the prior experiments. The process continues for N number of rounds, depending on the experimental budget available, with the objective of detecting a maximum number of hit candidates.

Specifically, let \mathcal{C} be the set of all the candidates and $f : \mathcal{C} \rightarrow \mathbb{R}$ be the function that maps each candidate to a real-valued measurement. The task is inherently sequential, spread over N rounds, where each round i is defined as the trial of $\{c_1, c_2, \dots, c_B\} \subseteq \mathcal{C}$ candidates independently. The goal of the experiment is to identify the candidates with measurement value greater than some threshold τ , i.e., $f(c) > \tau$ for $c \in \mathcal{C}$. Such candidates are termed as *hits* for that measurement function.

Further, let C_i be the set of candidates selected by the agent at any round i and let Z_i be the set comprising of entire experiment history, i.e. $Z_i = \{(c, f(c)) \mid c \in \bigcup_{t=1}^{i-1} C_t\}$. Given Z_i , the agent is then expected to guide the formation of C_i .

At the end of N rounds, let $\mathcal{C}_a = \bigcup_{t=1}^N \mathcal{C}_t^+$, where $\mathcal{C}_t^+ = \{c \in \mathcal{C}_t : f(c) > \tau\}$ represents the hits identified in round t . Thus, \mathcal{C}_a represents the cumulative set of hits across all rounds up to t . Similarly, let \mathcal{C}_{gt} be the set of all true hits for the measurement function f , $\mathcal{C}_{gt} = \{c \in \mathcal{C} : f(c) > \tau\}$.

We typically fix the $N = 5$ and $B = 128$ for all our experiments unless stated otherwise. Additionally, the τ for the molecular property task is set to be at the 90th percentile of the property value.

3 Related Work

BO for Scientific Applications: Bayesian Optimization in scientific domains has focused on a variety of applications like drug discovery (Griffiths and Hernández-Lobato, 2020; Korovina et al., 2020), biological experiment design (Roohani et al., 2024; Lyle et al., 2023; King et al., 2004), and chemical/molecular tasks (Kristiadi et al., 2024; Fromer et al., 2024; Ranković and Schwaller, 2023; Shields et al., 2021). Simultaneously, there have also been works that advance autonomous AI-driven closed-loop experiment design (M. Bran et al., 2024; Tom et al., 2024; Boiko et al., 2023). In this work, we study how LMs can be used to incorporate prior information into the search process in the context of two real-world domains: *single gene perturbations* and *chemical property optimization*. Both tasks are similar in terms of extensive candidate space and complex relationships between the candidate and its associated measurement. However, the genetic domain has a finite candidate space of genes in the human genome, whereas the space of all molecule is far larger (up to 10^{60} small molecules (Bohacek et al., 1996)) making it harder to have priors over any particular molecule.

LLMs for Bayesian Optimization: A series of recent works have explored the use of LLM embeddings in different ways, like for general-purpose regression (Nguyen et al., 2024), to improve surrogate modeling (Nguyen and Grover, 2024), and to augment the traditional methods like Gaussian Processes (Hartford et al., 2020; Ramos et al., 2023). We focus on using LLMs as the backend for an agent that interacts with the laboratory feedback and performs closed-loop experiment design.

The two closest related prior works to our study are BioDiscoveryAgent (Roohani et al., 2024) and

LLAMBO (Liu et al., 2024). LLAMBO is based on GPT-3.5 and primarily focuses on hyperparameter-tuning tasks, with LLM performing end-to-end Bayesian Optimization via suitable prompting. In a similar spirit, BioDiscoveryAgent proposes an LLM agent based on Claude 3.5 Sonnet and augmented with external tools that performs closed-loop experiment design of genetic perturbations. Both these approaches leverage proprietary LLMs and heavily rely on in-context learning to design experiments conditional on the experimental history. BioDiscoveryAgent, in particular, prompts the LLM to generate the entire batch of candidates to try next. In this work, we take a critical view of these methods, particularly BioDiscoveryAgent (BDA), and resort to instruction-tuned open-source models like Llama3.1-8B (Grattafiori et al., 2024) and Qwen2-7B (Yang et al., 2024) to further examine it. While there exists evidence that transformers have been shown to perform amortized Bayesian inference (Müller et al., 2024) when appropriately trained, it is not obvious that the ability arises from the next-token prediction and post-training objectives, and hence we specifically test the open-source BDA for the ability to select experiments. Eventually, we take a departure from a purely LLM-based approach and explore the synergy between LLMs and classical methods in our proposed hybrid approach, LLMNN, that achieves significantly superior performance.

4 Tasks and Datasets

We perform experiments on two scientific domains in this work. In particular, we work with *single gene perturbations* and *chemical properties of molecules*. We simulate running experiments on a set of candidates by retrieving the corresponding measurement value from a tabular dataset. However, in real-world settings, the corresponding experiments would be conducted in a laboratory or an in-silico simulator. In the subsequent subsections, we describe the datasets used in the two domains.

4.1 Single Gene Perturbations

Section 2 explains the details of the task, such as the goal with experiment design, candidate space, and the associated measurements. In terms of datasets, we use the **IL2**, **IFNG**, **Carnevale**, **Sanchez**, and **Sanchez Down** datasets from Roohani et al. (2024)’s BioDiscoveryAgent as each of them covers a distinct biological process. For example, **IL2**

Method	IL2	IFNG	Carnevale	Sanchez	Sanchez Down
Ground truth ($ \mathcal{C}_{gt} $)	654	920	943	924	924
Llama-3.1-8B backbone					
BDA	39.4	44	32.4	23.2	43.6
BDA-Rand	37	51	31.6	30.8	45
Qwen-2-7B backbone					
BDA	33.2	26.2	27.2	26.4	26
BDA-Rand	29	32.4	29	24.2	35.4
Claude 3.5 Sonnet backbone					
BDA (Reported Numbers)	68.01	87.4	39.6	60.72	N/A
BDA (Replicated)	59.4	78.8	43.8	31.6	51.8
BDA-Rand	57.6	79.4	42	33.8	57.6

Table 1: Cumulative number of hits secured by each method on the gene perturbation datasets. The values are averaged over 5 runs. Note that we experiment with the No-Tool version of BioDiscoveryAgent. BDA-Rand is the BioDiscoveryAgent baseline provided with random feedback after each round of experimentation. It can be concluded that LLMs trained on next token prediction and RLHF fail to perform in-context experimental design.

(Schmidt et al., 2022) is concerned with the phenotype of change in the production of Interleukin-2 (IL2), a cytokine involved in immune signaling. More details about each of the datasets are included in the appendix B.1.1. Each dataset contains measurements for over 18000 genes, i.e., $|\mathcal{C}| > 18000$, each knocked down in a distinct cell.

4.2 Chemical Properties of Molecules

For the chemical properties task, we focus on the following three molecular datasets: **ESOL** (water solubility), **FreeSolv** (hydration free energy in water), and **Ion. E.** (ionization energy of the molecules). Note that $|\mathcal{C}| = 1128, 642$ and 11565 respectively for each of the aforementioned datasets. For more details, see appendix B.1.2

5 Is Naively Prompting LLM Enough for Experiment Design?

Experimental design is inherently a sequential process where at each round, one should select experiments on the basis of both prior information *and* the outcomes of previous experiments. Through the pretraining objective, LLMs have extensive prior information which enables strong performance in experimental design settings with relatively few rounds of experiments (i.e. where the guidance of the prior matters most). For example, LLAMBO (Liu et al., 2024) focuses on tasks related to hyperparameter-tuning, and BioDiscoveryAgent (Roohani et al., 2024) demonstrates

strong performance on genetic perturbation experiment design. Despite minor differences, there are major similarities between these approaches. Both works pass on the experiment history and corresponding observations by simply appending them within the prompt itself with the hope that the LLM can leverage its in-context abilities to incorporate this feedback and adapt its subsequent predictions. We test the extent to which BioDiscoveryAgent is actually using this information by breaking the relationship between candidates, c_i , and their associated outcomes, $f(c_i)$. In particular, we randomly pair each c_i with some other outcome, $f(c_j)$, which breaks their joint dependency, while keeping their respective marginal distributions fixed.

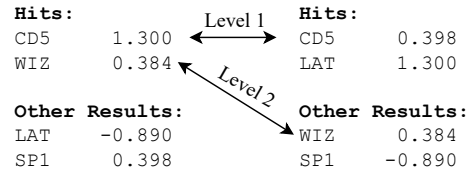


Figure 1: Illustration of the 2 levels of random permutation in the experimental feedback to the LLM on the genetic domain. Level 1 modifies the measurement value, whereas level 2 modifies whether a candidate is a hit or not. Note that it is possible that a particular gene has both level 1 and 2 modifications.

Method. For our investigation, we work with the BioDiscoveryAgent (BDA) pipeline on a small open-source LLM backbone without any external

Method	IL2	IFNG	Carnevale	Sanchez	Sanchez Down
Ground truth ($ C_{gt} $)	654	920	943	924	924
Llama-3.1-8B backbone					
Linear UCB	35	72	38	39	42
GP	147.8	23	22.2	27.6	30
BDA	39.4	44	32.4	23.2	43.6
Qwen-2-7B backbone					
Linear UCB	93	74	31	31	41
GP	147.8	23	22.2	27.6	30
BDA	33.2	26.2	27.2	26.4	26

Table 2: Cumulative number of hits secured by each method on the gene perturbation datasets. The values are averaged over 5 runs. The max values in each column for each LLM backbone have been bolded. Note that we experiment with the No-Tool version of BioDiscoveryAgent. Further, all the statistical methods use the embeddings corresponding to the LLM in the backbone of BDA to ensure the same amount of knowledge. It can be observed clearly that principled statistical approaches outperform the BioDiscoveryAgent baseline on most datasets and across both families of open-source LLMs.

tools. In the first set of experiments, we evaluate the ability of the LLMs to leverage experimental feedback. We compare BDA with BDA-Rand, which is the same as BDA but receives randomized feedback instead of true feedback. We perform 2 levels of randomization, level 1 being random measurement values and level 2 being randomness in hit vs not-hit feedback. Figure 1 contains an illustration of the randomization procedure. Both methods are evaluated against the ground truth hits for each dataset. As a second step, we compare BDA with classical models like Linear UCB and Gaussian Process (GP) to evaluate whether the LLM has sufficient information to make strong selections. Both Linear UCB and the GP only condition on the residual stream embeddings of the LLM so can only outperform the LLM insofar as they can better use the experimental feedback.

Results. We investigated the in-context experimental design abilities of the LLMs, and Table 1 shows the results for the same. The numbers in the table correspond to the cumulative number of hits obtained after 5 rounds of experiments with 128 perturbations in every round and are averaged over 5 runs. Across both LLMs and all 5 datasets, it is evident that passing random feedback does not hurt the performance of the framework at all. In fact, the performance remains nearly the same or improves slightly. To further substantiate our hypothesis, we also perform a similar experiment on Claude 3.5 Sonnet, a proprietary large-scale LLM, and observe that even Claude maintains nearly the

same performance despite random feedback¹. The strong initial performance of the LLMs is therefore likely the result of their priors on ordering of genes and is not affected by the feedback of past experiments appended in its prompt.

The above flaw in the LLM’s capabilities motivated us to compare BDA with classical approaches. Table 2 shows the cumulative number of hits obtained for BDA and classical baselines as described above. It can be seen clearly that across the 5 datasets, either LinearUCB or GP outperforms the BDA framework by a significant margin on both choices of LLM models. In contrast to the trend of Claude-based BDA, these results further highlight the lack of robustness of the framework’s performance with respect to the base LLM.

6 LLMNN: A Hybrid Experiment Design Method

In light of the above observations, we propose LLM guided Nearest Neighbour (LLMNN) framework, a simple greedy approach that prompts an LLM to guide the location of cluster centers and leverages nearest neighbour sampling to form the batch B_i at any round i . The LLMNN framework is designed to leverage: 1.) Generalist LLMs’ intrinsic domain knowledge to guide the search in the vast candidate space, and 2) the inductive bias that similar genes/molecules have similar functionalities/properties. The schematic workflow

¹We generate the numbers for BDA using the publicly available code as we could not replicate their original numbers, possibly due to LLM updates.

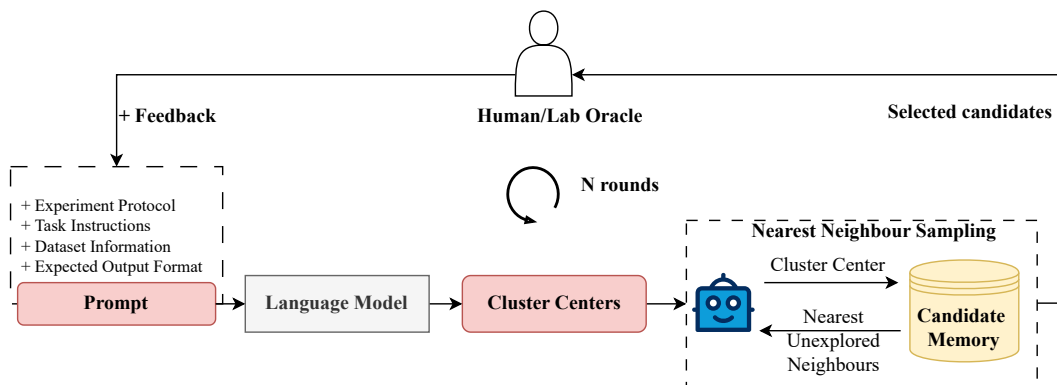


Figure 2: Schematic Workflow of the LLM-guided Nearest Neighbour (LLMNN) method. The red rounded boxes are the inputs and outputs of the Language Model. For ease, only cluster centers have been shown in the output. Please refer to Section 6 for detailed output specification and information about the components of the method. Refer to Appendix A for the detailed prompt template and to Appendix C for a full trace of the method.

of the proposed LLMNN framework is illustrated in Figure 2. Below, we discuss the important components of the LLMNN framework, followed by a detailed workflow of the method.

Candidate Memory. stores the pool of candidates, each represented in the following format: `{"name": <candidate name>, "score": <measurement value>, "explored": <bool>}`. For molecules, the candidate name is the SMILES string and measurement is the property value, whereas for genes, the candidate name is the gene name in the HGNC nomenclature and measurement is the phenotypic response value. Each candidate is also associated with an embedding. The explored key tracks whether a candidate has already been selected, preventing repetition in the optimization process. The memory interacts with the LLMNN framework by receiving a query candidate name and returning the nearest unexplored neighbors based on distance in the embedding space.

Candidate Embeddings. Each candidate in the memory has an associated embedding. We experiment both with LLM embeddings for the candidates and also domain-specific embeddings like Achilles embeddings for genes (Tsherniak et al., 2017) and Molformer (Ross et al., 2022) embeddings for the SMILES strings. The distance metric for Achilles embeddings is the cosine distance, whereas for all other embeddings, the L2-squared distance is used. More details about the embeddings are included in Appendix B.3

Agent Response Format: We borrow the

agent response format from BioDiscoveryAgent (Roohani et al., 2024), and direct the LLM to structure its responses into several parts: Reflection, Research Plan, Solution, similar to Huang et al. (2023). Through the Reflection and Research Plan entries, the model is able to articulate its reasoning behind a particular prediction. Solution contains a list of n_c cluster centers to sample around next. For our experiments, we fix $n_c = 5$ unless stated otherwise.

Workflow. The exact flow of the LLMNN framework is defined as follows:

1. LLMNN framework takes experiment protocol, task, and candidate space details as the context in its prompt to generate the first set of cluster centers it seeks to explore.
2. The output cluster centers are then used to retrieve candidates from the candidate memory.
3. The selected candidates for the round are then sent to the human or a lab oracle to obtain their value of the measurement function f and to determine if they were a hit.
4. This hit vs not hit qualitative feedback, along with the measurement value, is appended to the prompt from Step 1 and passed to the LLM to generate the next set of cluster centers.
5. Steps 2-4 are repeated for $N-1$ rounds. The total hits identified successfully at the end indicate the method’s performance.

Method	IL2	IFNG	Carnevale	Sanchez	Sanchez Down
Ground truth ($ \mathcal{C}_{gt} $)	654	920	943	924	924
Random	23.8	26	35	32.2	34.6
Coreset	9.2	31.4	29.4	33.8	30.2
Linear UCB	110	98	51	75	87
GP	139	24	20.4	24.6	33
BDA-GS (Claude 3.5 Sonnet)	65.4	88.32	40.55	57.28	N/A
Llama-3.1-8B backbone					
BDA	39.4	44	32.4	23.2	43.6
BDA-GS	63.4	57.4	42.59	22.2	39.6
LLMNN NoExp	137.8	86.3	58.9	42.5	42.3
LLMNN	163.3	107.9	65.6	54.2	45.5
Qwen-2-7B backbone					
BDA	33.2	26.2	27.2	26.4	26
BDA-GS	38.6	35.2	32.8	28.6	38
LLMNN NoExp	146.4	59.1	44.5	62.8	46.4
LLMNN	160.7	78	60.9	53.5	45.1

Table 3: Cumulative hits over 5 rounds of experiments with 128 candidates in each experiment. The values in the table are averaged over 5 runs. The max values in each column have been bolded. Note that we use Achilles embeddings of genes for the gene search tool in the case of LLMNN and BDA-GS on Llama and Qwen backbones. The results indicate the strong performance of LLMNN, even surpassing the BDA based on Claude 3.5 Sonnet, which has access to a sophisticated gene search tool. Since the numbers for this BDA variant have been taken from Table 3 of the original paper, we do not have a score for Sanchez Down as it isn’t included in the original text.

7 Does LLMNN do Any Better on Our Experiment Design Domains?

We observed BioDiscoveryAgent struggled to outperform the traditional baselines (when provided with the same priors via access to embeddings) on the genetic perturbation datasets, and that the performance was largely the result of having better priors. Given this, one would expect better performance from LLM-based approaches in settings with well-studied candidate spaces—e.g. genes in the human genome—compare with much larger spaces—e.g. all of chemical space.

LLMNN addresses the poor response to experimental feedback by taking advantage of the fact that LLMs encode candidates into a common embedding space. This allows us to greedily construct experimental batches by performing nearest neighbor sampling in the neighborhood of previously observed hits. LLMNN requires minimal assumptions of LLM generating valid gene names as per HGNC nomenclature and valid SMILES strings, which is reasonable for modern-day LLMs that have been pretrained on an internet-scale of knowledge. Further, LLMNN is an attempt at exploiting the known inductive biases of the domain to

achieve higher performance.

Method. We experiment with two variants of LLMNN, one original with the Research Plan and Reflection outputs, and the other that contains purely the Solution, without any explanations. Both LLMNN variants have access to the gene similarity tool based on the Achilles (Tsherniak et al., 2017) embeddings. Against LLMNN, we include traditional baselines like Random, Coreset (pure diversity-based approach), LinearUCB, and Gaussian Process (GP). Apart from these methods, we also include three variants of BioDiscoveryAgent: 1) BDA, i.e., without any tools, 2) BDA-GS with access to the same gene similarity tool that LLMNN has access to, and 3) BDA-GS (Claude 3.5 Sonnet), which is the originally proposed approach in Roohani et al. (2024) that uses Claude 3.5 Sonnet LLM in the backbone and has access to more sophisticated gene similarity search tools like enrichment analysis on Reactome database (Gillespie et al., 2022). We do not include BDA in the molecules domain because BDA is not constrained to predict molecules that exist in the library and hence one needs a large number of retries to construct a batch. The tool for molecular similarity

Method	Ion. E.	ESOL	FreeSolv
G. truth ($ C_{gt} $)	1156	112	64
Random	77.8	29	14.4
Coreset	151.6	57.6	20.4
Linear UCB	283	76	39
GP	151.39	35	16.8
Llama-3.1-8B backbone			
LLMNN NoExp	152.8	23.4	12.2
LLMNN	103.2	22.2	10.2
Qwen-2-7B backbone			
LLMNN NoExp	147.2	24	25.8
LLMNN	131	29	7

Table 4: Cumulative hits over 5 rounds of experiments with $B = 128, 64$, and 32 candidates in each experiment and $n_c = 5, 4$ and 4 , respectively, for Ion. E., ESOL and FreeSolv datasets. The values in the table are averaged over 5 runs. Note that we use molecule embeddings from MolFormer XL-10pct for the similarity search tool. The results indicate that the NoExp variants of LLMNN perform competitively compared to classical baselines on 2/3 datasets.

available to LLMNN is based on Molformer (Ross et al., 2022) embeddings of SMILES strings.

Results. Table 3 displays the performance comparison between different methods on the single gene perturbation domain. Note that the same experimental budget has been provided to all the methods, and the numbers are averaged over 5 runs. The table illustrates that the LLMNN method based on the Llama-3.1 backbone outperforms BDA based on both Llama and Claude backends on 5/5 and 3/4 datasets, respectively, by significant margins. It is worth noting that Llama-3.1 is just an 8 billion parameter model as compared to Claude 3.5, which has been trained on larger data with significantly more parameters. Further, LLMNN only had access to a basic gene similarity tool as compared to Claude BDA, which had more sophisticated gene search tools. Similarly, Qwen-based LLMNN outperforms the corresponding BDA on 5/5 datasets. Another interesting observation is that the traditional baselines still perform really well, surpassing Claude BDA on 4/4 datasets and LLMNN as well on 2/5 datasets. One of the contributing factors to the performance of LLMNN is maintaining a memory that keeps track of which genes have already been explored. This ensures that similarity queries return unexplored neighbours at every query, in contrast to the BioDis-

coveryAgent, which doesn’t maintain this state and hence, would inevitably return the same set of genes always for the same query, irrespective of the experiment history.

On the molecular domain, Table 4 contains the numbers for cumulative hits for all methods averaged over 5 runs, given the same experimental budget. The numbers highlight that while Qwen2-based LLMNN without explanations performs closely to the traditional baselines on 2/3 datasets, the latter still wins over the LLMNN approach across all the datasets, especially with the ESOL dataset. This underscores that LLMs with simple inductive biases are still not enough alone to serve as experiment designers across different scientific domains, but need tight coupling with the more principled statistical methods that trade off exploration and exploitation like linear UCB and GPs to achieve higher performance.

8 Conclusion

In this work, we set out to critically examine this hypothesis using instruction-tuned open-source LLMs. We particularly focused on BioDiscoveryAgent and experimented on two domains: single gene perturbations (5 datasets) and molecular property optimization (3 datasets). When compared to classical methods like Linear UCB and Gaussian Process that used the embedding from the same LLM as BDA, the open-source BDA performed significantly worse. Furthermore, on deeper experiments with random feedback to the LLM’s context, the open-source BDA still retained a similar average performance, showcasing that the LLMs trained on next token prediction and RLHF do not leverage experimental feedback in the design of their experiments. Finally, we proposed an LLM-guided Nearest Neighbour framework (LLMNN), a simple combination of LLM and a classic nearest neighbour method. LLMNN outperforms BDA significantly on gene datasets and is applicable to molecular domains, and performs at par with the classical baselines, except on some molecular datasets. Overall, this work suggests that more work is needed to effectively incorporate experiment feedback into LLM-based experimental design pipelines. It introduces a plausible avenue for future research marked by the synergy of LLMs, classical methods, and domain-specific inductive biases.

Limitations and Future Works

While we demonstrate superior performance with LLMNN on both gene perturbation and molecular domains, this study has several limitations. Firstly, LLMNN augments the LLM with a very simplistic nearest neighbour sampling that allocates an equal budget to each cluster to generate better predictions. However, more complex schemes can be explored to adaptively allocate more budget to centers that have a higher probability of detecting hits. For example, a probabilistic model like GP could be used to determine the hit likelihood of the predicted centers that can, in turn, be used to define budget allocation between the centers.

Due to the policy of sampling around centers, the method is largely exploitative and hence sensitive to the choice of embeddings and hits identified in earlier rounds. More tightly-coupled integration of LLMs and classic exploration methods could be investigated to improve the robustness of the experiment design agent. Further, the inductive bias that *similar* candidates have *similar* properties is clearly not the best bias on molecular domains, as the classic exploration approaches maintain a strong performance as compared to LLMNN, suggesting the need to identify and encode more domain-specific and task-specific biases.

Lastly, we primarily focused on small open-source LLMs like Llama-3.1-8B Instruct and Qwen2-7B-Instruct for our experiments with no external tool use. It shall be interesting to conduct similar studies with both larger (like 70B) and smaller models. Another promising direction is to explore how external tools like literature search, enrichment analysis tools for genes, etc. be effectively augmented to the agent for better performance, as also highlighted by Roohani et al. (2024).

Ethical Considerations

This work deals with building hybrid AI agents that are powered by LLMs and classical exploration settings for guiding the design of scientific experiments. Since it involves the use of LLM, it is vulnerable to adversarial attacks where the agent might suggest experiments that lead to catastrophic outcomes in a real laboratory. Therefore, it is imperative for the users to have human scientists in the loop to screen out such experiment configurations. Not just post-hoc, the safety could be ensured at the start by suitable constraints on the candidate

space.

One of the domains in this work is single gene perturbations, where the effect of knocking down a gene is measured in human cells. While these agents will help increase the robustness of the target discovery phase of the drug discovery pipeline, it should be taken into account that human cells often exhibit genetic variation from one population to another, thus limiting the transferability of the downstream insights to new groups. Thus, caution needs to be exercised in utilizing the insights drawn by this agent in the actual drug discovery pipeline.

References

- Anthropic. 2024. [Claude 3.5 sonnet](#).
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.
- Regine S Bohacek, Colin McMartin, and Wayne C Guida. 1996. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1):3–50.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.
- Julia Carnevale, Eric Shifrut, Nupura Kale, William A Nyberg, Franziska Blaesche, Yan Yi Chen, Zhongmei Li, Sagar P Bapat, Morgan E Diolaiti, Patrick O’Leary, and 1 others. 2022. Rasa2 ablation in t cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925):174–182.
- John S Delaney. 2004. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005.
- Jenna Fromer, Runzhong Wang, Mrunali Manjrekar, Austin Tripp, José Miguel Hernández-Lobato, and Connor W. Coley. 2024. [Batched bayesian optimization with correlated candidate uncertainties](#). *Preprint*, arXiv:2410.06333.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.
- Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, and 1 others. 2022. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692.

680	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,	Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta	736
681	Abhinav Pandey, Abhishek Kadian, Ahmad Al-	Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Ge-	737
682	Dahle, Aiesha Letman, Akhil Mathur, Alan Schel-	off Pleiss. 2024. A sober look at LLMs for material	738
683	ten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh	discovery: Are they actually good for Bayesian opti-	739
684	Goyal, Anthony Hartshorn, Aobo Yang, Archi Mi-	mization over molecules? In <i>ICML</i> .	740
685	tra, Archie Sravankumar, Artem Korenev, Arthur		
686	Hinsvark, and 542 others. 2024. The llama 3 herd of	Bowen Lei, Tanner Quinn Kirk, Anirban Bhattacharya,	741
687	models . <i>Preprint</i> , arXiv:2407.21783.	Debdeep Pati, Xiaoning Qian, Raymundo Arroyave,	742
		and Bani K Mallick. 2021. Bayesian optimization	743
688	Ryan-Rhys Griffiths and José Miguel Hernández-	with adaptive surrogate models for automated ex-	744
689	Lobato. 2020. Constrained bayesian optimization	perimental design. <i>Npj Computational Materials</i> ,	745
690	for automatic chemical design using variational au-	7(1):194.	746
691	toencoders. <i>Chemical science</i> , 11(2):577–586.		
692	Jason S Hartford, Kevin Leyton-Brown, Hadas Raviv,	Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long,	747
693	Dan Padnos, Shahar Lev, and Barak Lenz. 2020. Ex-	Pengjun Xie, and Meishan Zhang. 2023. Towards	748
694	emplar guided active learning . In <i>Advances in Neural</i>	general text embeddings with multi-stage contrastive	749
695	<i>Information Processing Systems</i> , volume 33, pages	learning. <i>arXiv preprint arXiv:2308.03281</i> .	750
696	13163–13173. Curran Associates, Inc.		
697	Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani,	D. V. Lindley. 1956. On a measure of the information	751
698	and Máté Lengyel. 2011. Bayesian active learning	provided by an experiment . <i>The Annals of Mathemat-</i>	752
699	for classification and preference learning . <i>CoRR</i> ,	<i>ical Statistics</i> , 27(4):986–1005.	753
700	abs/1112.5745.		
701	Xun Huan, Jayanth Jagalur, and Youssef Marzouk. 2024.	Tennison Liu, Nicolás Astorga, Nabeel Seedat, and	754
702	Optimal experimental design: Formulations and com-	Mihaela van der Schaar. 2024. Large language	755
703	putations. <i>Acta Numerica</i> , 33:715–840.	models to enhance bayesian optimization . <i>Preprint</i> ,	756
704	Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec.	arXiv:2402.03921.	757
705	2023. Mlagentbench: Evaluating language agents on		
706	machine learning experimentation. <i>arXiv preprint</i>	Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jes-	758
707	<i>arXiv:2310.03302</i> .	son, Stefan Bauer, Yarin Gal, and Patrick Schwab.	759
708	Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei	2023. Discobax: Discovery of optimal interven-	760
709	Chen, William Davidson Richards, Stephen Dacek,	tion sets in genomic experiment design. In <i>Inter-</i>	761
710	Shreyas Cholia, Dan Gunter, David Skinner, Ger-	<i>national Conference on Machine Learning</i> , pages	762
711	brand Ceder, and Kristin A. Persson. 2013. Com-	23170–23189. PMLR.	763
712	mentary: The materials project: A materials genome		
713	approach to accelerating materials innovation . <i>APL</i>	Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Bal-	764
714	<i>Materials</i> , 1(1):011002.	dassari, Andrew D White, and Philippe Schwaller.	765
715	Moksh Jain, Tristan Deleu, Jason Hartford, Cheng-Hao	2024. Augmenting large language models with chem-	766
716	Liu, Alex Hernandez-Garcia, and Yoshua Bengio.	istry tools. <i>Nature Machine Intelligence</i> , 6(5):525–	767
717	2023. Gflownets for ai-driven scientific discovery.	535.	768
718	<i>Digital Discovery</i> , 2(3):557–577.		
719	Ross D King, Kenneth E Whelan, Ffion M Jones,	David L Mobley and J Peter Guthrie. 2014. Freesolv:	769
720	Philip GK Reiser, Christopher H Bryant, Stephen H	a database of experimental and calculated hydration	770
721	Muggleton, Douglas B Kell, and Stephen G Oliver.	free energies, with input files. <i>Journal of computer-</i>	771
722	2004. Functional genomic hypothesis generation	<i>aided molecular design</i> , 28:711–720.	772
723	and experimentation by a robot scientist. <i>Nature</i> ,		
724	427(6971):247–252.	Samuel Müller, Noah Hollmann, Sebastian Pineda	773
725	Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal.	Arango, Josif Grabocka, and Frank Hutter. 2024.	774
726	2019. Batchbald: Efficient and diverse batch acqui-	Transformers can do bayesian inference . <i>Preprint</i> ,	775
727	sition for deep bayesian active learning. <i>Advances in</i>	arXiv:2112.10510.	776
728	<i>neural information processing systems</i> , 32.		
729	Ksenia Korovina, Sailun Xu, Kirthivasan Kandasamy,	Tung Nguyen and Aditya Grover. 2024. Lico: Large lan-	777
730	Willie Neiswanger, Barnabas Poczos, Jeff Schnei-	guage models for in-context molecular optimization.	778
731	der, and Eric Xing. 2020. Chembo: Bayesian opti-	<i>arXiv preprint arXiv:2406.18851</i> .	779
732	mization of small organic molecules with synthesiz-		
733	able recommendations. In <i>International Conference</i>	Tung Nguyen, Qiuyi Zhang, Bangding Yang, Chansoo	780
734	<i>on Artificial Intelligence and Statistics</i> , pages 3393–	Lee, Jorg Bornschein, Yingjie Miao, Sagi Perel, Yu-	781
735	3403. PMLR.	tian Chen, and Xingyou Song. 2024. Predicting from	782
		strings: Language model embeddings for bayesian	783
		optimization . <i>Preprint</i> , arXiv:2410.10190.	784
		OpenAI. 2023. Gpt-3.5-turbo .	785
		OpenAI. 2024. Openai o1 system card .	786
		Mayk Caldas Ramos, Shane S Michtavy, Marc D	787
		Porosoff, and Andrew D White. 2023. Bayesian opti-	788
		mization of catalysts with in-context learning. <i>arXiv</i>	789
		<i>preprint arXiv:2304.05341</i> .	790

A Prompt Templates

- Bojana Ranković and Philippe Schwaller. 2023. Bo-
chemian: Large language model embeddings for
bayesian optimization of chemical reactions. In
*NeurIPS 2023 Workshop on Adaptive Experimental
Design and Active Learning in the Real World*.
- Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora,
Zachary Steinhart, Kexin Huang, Alexander Marson,
Percy Liang, and Jure Leskovec. 2024. [Biodiscoveryagent: An ai agent for designing genetic perturbation experiments](#). *Preprint*, arXiv:2405.17631.
- Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan,
Inkit Padhi, Youssef Mroueh, and Payel Das. 2022.
Large-scale chemical language representations cap-
ture molecular structure and properties. *Nature Ma-
chine Intelligence*, 4(12):1256–1264.
- Carlos G Sanchez, Christopher M Acker, Audrey Gray,
Malini Varadarajan, Cheng Song, Nadire R Cochran,
Steven Paula, Alicia Lindeman, Shaojian An, Gre-
gory McAllister, and 1 others. 2021. Genome-wide
crispr screen identifies protein pathways modulating
tau protein levels in neurons. *Communications biol-
ogy*, 4(1):736.
- Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Ja-
cob W Freimer, Raymund Bueno, Vinh Q Nguyen,
Franziska Blaeschke, Chun Jimmie Ye, and Alexan-
der Marson. 2022. Crispr activation and interference
screens decode stimulation responses in primary hu-
man t cells. *Science*, 375(6580):eabj4008.
- Benjamin J Shields, Jason Stevens, Jun Li, Marvin
Parasram, Farhan Damani, Jesus I Martinez Alvarado,
Jacob M Janey, Ryan P Adams, and Abigail G Doyle.
2021. Bayesian reaction optimization as a tool for
chemical synthesis. *Nature*, 590(7844):89–96.
- Gary Tom, Stefan P Schmid, Sterling G Baird, Yang
Cao, Kourosh Darvish, Han Hao, Stanley Lo, Ser-
gio Pablo-García, Ella M Rajaonson, Marta Skreta,
and 1 others. 2024. Self-driving laboratories for
chemistry and materials science. *Chemical Reviews*,
124(16):9633–9732.
- Aviad Tsherniak, Francisca Vazquez, Phil G Mont-
gomery, Barbara A Weir, Gregory Kryukov, Glenn S
Cowley, Stanley Gill, William F Harrington, Sasha
Pantel, John M Krill-Burger, and 1 others. 2017.
Defining a cancer dependency map. *Cell*, 170(3):564–
576.
- Tsuyoshi Ueno, Trevor David Rhone, Zhufeng Hou,
Teruyasu Mizoguchi, and Koji Tsuda. 2016. Combo:
An efficient bayesian optimization library for materi-
als science. *Materials discovery*, 4:18–21.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng,
Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan
Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-
ran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian
Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and
43 others. 2024. [Qwen2 technical report](#). *Preprint*,
arXiv:2407.10671.

In the following subsections, we provide the de-
tailed prompt template used for both the genetic
perturbation and molecular property domains.

A.1 Single Gene Perturbation

SYSTEM PROMPT:

You are a biomedicine expert who will assist me on problems in drug discovery. I am planning to run a CRISPR screen to identify genes that **{func desc}**. I can only perturb exactly **{batch len}** genes at a time. For each predicted perturbation, I am able to measure out the **{meas desc}** which will be referred to as the score. I can only do 5 rounds of experimentation. After every round of experiment, I will provide you with feedback on your predictions, including the correctly identified genes called hits and the corresponding score. The predictions which are not hits will be included in other results.

USER PROMPT:

This is round **{round num}**.
Here is the feedback on all your predictions till now:
{feedback}
Here is a strategy to follow: Update your priors appropriately and choose genes that gave you hits. Also, be sure to explore by including some genes that could give hits.
Please propose **{num cluster centers}** different yet valid gene names as per the HGNC nomenclature you want to explore next. Note that I will choose unexplored genes closest to your predicted genes to form the predictions. Your response should exactly follow the format:
**Reflection: Thoughts on previous results and next steps.
**Research Plan: The full high-level research plan, with current status and reasoning behind each proposed approach. It should be at most 5 sentences.
**Solution:
<Gene 1>
<Gene 2>
...
<Gene **{num cluster centers}**>
Each gene in the solution should only be the gene name in the HGNC nomenclature. DO NOT ADD ANY COMMENTS IN THE SOLUTION OR AFTER THE SOLUTION.

This is the prompt template at any round i . If

$i = 1$, the segment about feedback in the user prompt won't be there. The different fields in bold represent the following:

- **round num**: This is the round number for which experiments are being selected.
- **batch len**: This corresponds to B i.e. the experimental budget in each round.
- **num cluster centers**: This reflects the number of cluster centers that LLMNN generates to sample around.
- **feedback**: This is a string divided into two parts: hits and non-hits. Both parts contain the respective candidate names and their corresponding measurement values.
- **func desc**: This is the description of the task for which we need to perform experiment design. In the context of genes, this means the description of the phenotype we desire to achieve.
- **score desc**: This elaborates on the measurement values we have for the candidates.

Please refer to table 5 for detailed func desc and score desc for each dataset.

A.2 Chemical Property Optimization

This prompt is used at any round i for the molecular datasets. Note that for $i = 1$, the feedback segment won't be included in the prompt. Most fields described in this prompt are similar to the genetic perturbation, with the exception of **candidate space info** that contains a high-level description of the candidate space to provide more context to the LLM. Table 6 contains the details of func desc and candidate space info for each of the 3 datasets.

SYSTEM PROMPT:

You are a chemistry expert who will assist me with problems in molecular property optimization. Given a library of molecules, I am planning to conduct wet-lab experiments to identify molecules that have high {func desc}. {candidate space info} I can only experiment with exactly {batch len} molecules at a time. For each predicted molecule, I am able to measure out the property value, which will be referred to as the score. I can only do 5 rounds of experimentation. After every round of experiment, I will provide you with feedback on your predictions, including the correctly identified molecules called hits and the corresponding score. The predictions which are not hits will be included in other results.

USER PROMPT:

This is round {round num}.

Here is the feedback on all your predictions till now: {feedback str}

Here is a strategy to follow: Update your priors appropriately and choose SMILES that gave you hits. Also, be sure to explore by including some SMILES strings that could give hits.

Please propose {num cluster centers} different yet valid SMILES strings of molecules you want to explore next. Note that I will choose unexplored molecules closest to your predicted SMILES strings to form the predictions. Your response should exactly follow the format:

****Reflection:** Thoughts on previous results and next steps.

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

****Solution:**

<SMILES 1>

<SMILES 2>

...

<SMILES {num cluster centers}>

Each SMILES string in the solution should be a SMILES string representation of a valid molecule.

DO NOT ADD ANY COMMENTS IN THE SOLUTION OR AFTER THE SOLUTION.

B Experiment Setup Details

The following subsections describe the additional experiment details for better understanding:

B.1 Datasets**B.1.1 Single Gene Perturbation**

The 5 datasets used in this domain are borrowed from BioDiscoveryAgent (Roohani et al., 2024) and are as follows: 1) **IL2** (Schmidt et al., 2022) is a dataset that measures the changes in the production of Interleukin-2 (IL2) cytokine involved in immune signaling; 2) **IFNG** (Schmidt et al., 2022) is another similar dataset that measures changes in production of Interferon- γ cytokine; 3) **Carnevale** (Carnevale et al., 2022) aims to identify genes that render T cells resistant to inhibitory signals encountered in the tumor microenvironment; 4) **Sanchez** (Sanchez et al., 2021) dataset studies the change in expression of endogenous tau protein levels in neurons and 5) **Sanchez Down** (Sanchez et al., 2021) is exactly same as Sanchez dataset but focuses purely on decreasing the expression unlike the Sanchez which includes both increasing and decreasing expression.

B.1.2 Chemical Property Optimization

We describe the molecular property datasets in greater detail below:

- **ESOL:** Part of the MoleculeNet benchmark, ESOL (Delaney, 2004) is a small dataset comprising the water solubility data for $|C| = 1128$ compounds in log moles per litre. Each compound is represented as a SMILES string.
- **FreeSolv:** Free Solvation (FreeSolv) Database (Mobley and Guthrie, 2014) is another small dataset that provides experimental and calculated hydration free energies of $|C| = 642$ small molecules in water.
- **Ion. E.:** This dataset is a subset of molecular data from the JCESR molecules that are a part of the Materials Project Database (Jain et al., 2013). The original set contains approximately 25,000 molecules with annotations of many attributes and properties. Our experiments focus on one property: the Ionization Energy. We apply two levels of filters on the whole set of 25,000 molecules: first, we restrict to only those molecules whose ionization energy lies between -10 and 10 units (to avoid noisy outliers), and

out of all these filtered molecules, we choose the molecules composed solely of ‘C’, ‘H’, ‘N’ or ‘O’ atoms, leaving us with $|\mathcal{C}| = 11,565$ candidates. These filters avoid the dominating effects due to particular elements, leaving primary dependence on the molecular structure and functional groups captured in the SMILES strings.

B.2 Compute

All the experiments in this study have been conducted on a single Nvidia A100 GPU with 40 GB of memory. Since our proposed method, LLMNN, involves inference through the LLMs and repeated retrieval from the candidate memory, on average, a single run takes between 5-10 minutes to run.

B.3 Embeddings

Achilles Embeddings. We use the publicly available Achilles embeddings (Tsherniak et al., 2017) for the gene similarity tool. These embeddings are 808 dimensional in size.

Molformer Embeddings. We use the publicly available Molformer-XL model to embed the SMILES strings². This model has been trained on a collection of SMILES strings from the ZINC and PubChem datasets to learn molecular representations. These embeddings are 768-dimensional in size.

Llama 3.1 Embeddings. We use the LLM2Vec (BehnamGhader et al., 2024) approach to obtain the Llama 3.1 embeddings. Specifically, we use the publicly released checkpoints by the authors on Huggingface³. These embeddings are 4096-dimensional in nature.

Qwen2 Embeddings. We use the publicly available Qwen2-based general text embedding model released by Alibaba-NLP (Li et al., 2023), as they claim to be on top of the MTEB leaderboard⁴. These embeddings are 3584 dimensional in nature.

C Full trace of LLMNN

Below we present a full trace of the LLMNN method, including both the prompts and the LLM outputs, on the IL2 dataset in the single gene perturbation domain.

²Model checkpoint taken from <https://huggingface.co/ibm-research/MoLFormer-XL-both-10pct>

³Model checkpoint taken from <https://huggingface.co/McGill-NLP/LLM2Vec-Meta-Llama-31-8B-Instruct-mntp-supervised>

⁴Model checkpoint taken from <https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct>

Listing 1: Full trace of LLMNN method with explanations on IL2 dataset

```

----ROUND 1: BEGIN PROMPT----
SystemMessage(You are a biomedicine
expert who will assist me on problems
in drug discovery. I am planning to
run a CRISPR screen to identify genes
that regulate the production of
Interleukin-2 (IL-2). I can only
perturb exactly 128 genes at a time.
For each predicted perturbation, I am
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 5 rounds of experimentation. After
every round of experiment, I will
provide you with feedback on your
predictions, including the correctly
identified genes called hits and the
corresponding score. The predictions
which are not hits will be included in
other results.)
HumanMessage(
This is round 1. We are beginning
with our experiments.
Here is a strategy to follow: Choose
genes that are very different in
their biological pathways to
discover what pathways give you hits
.
Please propose 5 different yet valid
gene names as per the HGNC
nomenclature you want to explore
next. Note that I will choose
unexplored genes closest to your
predicted genes to form the
predictions. Your response should
exactly follow the format:
**Reflection: Thoughts on previous
results and next steps.
**Research Plan: The full high level
research plan, with current status
and reasoning behind each proposed
approach. It should be at most 5
sentences.
**Solution:
## <Gene 1>
## <Gene 2>
...
## <Gene 5>
Each gene in the solution should
only be the gene name in the HGNC
nomenclature.
DO NOT ADD ANY COMMENTS IN THE
SOLUTION OR AFTER THE SOLUTION.
)
----END PROMPT----

----BEGIN OUTPUT----
**Reflection: Thoughts on previous
results and next steps.**
Since we are starting from scratch,
there is no previous data to reflect
on. However, we will follow the
strategy of choosing genes that are
very different in their biological
pathways to increase the chances of
discovering novel regulators of IL-2

```

1048	production.	SULT2B1	-0.44	1118
1049		[OTHER RESULTS]		1119
1050	**Research Plan: The full high level	name	score	1120
1051	research plan, with current status	ABL1	0.09	1121
1052	and reasoning behind each proposed	QRFP	0.00	1122
1053	approach.**	COA3	-0.05	1123
1054	We will choose genes involved in	E2F6	-0.11	1124
1055	different biological processes such	NANOG	-0.05	1125
1056	as cell signaling, transcription	NDUFA4	0.04	1126
1057	regulation, and metabolic pathways	RAN	0.12	1127
1058	to identify potential regulators of	MRPL19	-0.14	1128
1059	IL-2 production. For round 1, we	SLU7	0.13	1129
1060	will select genes that are involved	CD8B	-0.01	1130
1061	in the regulation of cell growth,	HMGB1	-0.12	1131
1062	differentiation, and metabolism. We	AP2S1	0.03	1132
1063	will use the HGNC database to	RPS28	-0.16	1133
1064	identify genes with diverse	COX5B	0.07	1134
1065	functions. Our goal is to identify	COPS6	-0.08	1135
1066	at least one hit in this round. We	RNGTT	0.03	1136
1067	will use the provided feedback to	RPL11	0.05	1137
1068	refine our strategy for the next	NCAPG	0.17	1138
1069	round.	NSF	-0.05	1139
1070		NCBP2	0.00	1140
1071	**Solution:**	OR51L1	-0.17	1141
1072	## ABL1	GHR	0.02	1142
1073	## HNF4A	TSHZ2	-0.01	1143
1074	## MAPK14	GPX3	0.11	1144
1075	## PAK4	TOX2	0.10	1145
1076	## SMAD2	FAM107B	0.05	1146
1077	----END OUTPUT----	RGPD3	-0.24	1147
1078		TRAPPC3L	0.10	1148
1079	----ROUND 2: BEGIN PROMPT---	WFDC6	-0.00	1149
1080	SystemMessage(You are a biomedicine	SLC24A3	0.10	1150
1081	expert who will assist me on problems	IFNE	-0.04	1151
1082	in drug discovery. I am planning to	ZFAND2A	0.12	1152
1083	run a CRISPR screen to identify genes	BNIP3L	-0.17	1153
1084	that regulate the production of	FAM184B	-0.10	1154
1085	Interleukin-2 (IL-2). I can only	RGMB	0.02	1155
1086	perturb exactly 128 genes at a time.	ZNF853	-0.09	1156
1087	For each predicted perturbation, I am	NPAS2	0.26	1157
1088	able to measure out the log fold	ATAD1	-0.05	1158
1089	change in Interleukin-2 (IL-2)	SMCP	-0.08	1159
1090	normalized read counts which will be	RAB27B	0.14	1160
1091	referred to as the score. I can only	BIN3	-0.10	1161
1092	do 5 rounds of experimentation. After	NPY2R	0.02	1162
1093	every round of experiment, I will	EQTN	0.16	1163
1094	provide you with feedback on your	AKR1B1	-0.21	1164
1095	predictions, including the correctly	MAPK14	-0.02	1165
1096	identified genes called hits and the	MLST8	-0.15	1166
1097	corresponding score. The predictions	PPP2R3C	-0.19	1167
1098	which are not hits will be included in	MNAT1	0.07	1168
1099	other results.)	NOL10	0.05	1169
1100	HumanMessage(EMC7	0.08	1170
1101	This is round 2.	EMC3	-0.05	1171
1102	Here is the feedback on all your	GUK1	0.19	1172
1103	predictions till now:	MED6	-0.16	1173
1104	[HITS]	CIAO1	0.10	1174
1105	name score	SNAPC5	0.01	1175
1106	MYBL2 0.51	ZCRB1	-0.03	1176
1107	GBF1 0.43	TRMT112	0.07	1177
1108	DDX41 0.47	RPS29	0.23	1178
1109	ZMAT2 0.45	PDCD2	0.10	1179
1110	RPL4 0.35	RPP21	-0.16	1180
1111	HNF4A -0.34	RPL5	0.10	1181
1112	CDC16 0.48	EIF6	0.19	1182
1113	WDR5 0.82	SNW1	0.28	1183
1114	MYC 0.41	HEATR1	0.06	1184
1115	ZNF536 -0.38	RPL28	0.34	1185
1116	TLN1 -0.49	WDR18	0.07	1186
1117	LIN37 -0.39	PAK4	0.07	1187

1188	AFDN	-0.06	## <Gene 1>	1258
1189	VCL	-0.02	## <Gene 2>	1259
1190	CATSPERG	-0.09	...	1260
1191	PTK2	-0.10	## <Gene 5>	1261
1192	BCAR1	-0.08	Each gene in the solution should	1262
1193	ACTR3	-0.06	only be the gene name in the HGNC	1263
1194	RAPGEF1	0.19	nomenclature.	1264
1195	WWTR1	-0.10	DO NOT ADD ANY COMMENTS IN THE	1265
1196	VWA8	0.26	SOLUTION OR AFTER THE SOLUTION.	1266
1197	FERMT2	-0.12)	1267
1198	RSU1	0.12	----END PROMPT----	1268
1199	MSTO1	0.02	----BEGIN OUTPUT----	1269
1200	NLRP2B	-0.09	**Reflection: Thoughts on previous	1270
1201	ARPC4	0.05	results and next steps.**	1271
1202	ARPC2	-0.12		1272
1203	KIF2C	-0.13	The previous round of CRISPR screen	1273
1204	SRF	0.01	identified 11 hits that regulate the	1274
1205	RNASE1	0.16	production of Interleukin-2 (IL-2),	1275
1206	NCKAP1	-0.14	including MYBL2, GBF1, DDX41, ZMAT2	1276
1207	ZFAND5	-0.01	, RPL4, HNF4A, CDC16, WDR5, MYC,	1277
1208	SMAD2	0.15	ZNF536, and TLN1. The scores of	1278
1209	MAGED2	-0.03	these hits ranged from 0.35 to 0.82,	1279
1210	POTEE	-0.09	indicating varying degrees of	1280
1211	IFNA10	-0.18	regulation. The other results	1281
1212	AKAIN1	-0.18	included 64 genes that did not show	1282
1213	ZNF614	0.08	significant regulation of IL-2	1283
1214	ZNF157	-0.11	production. To refine our	1284
1215	MAPK8	0.09	understanding of the IL-2 regulatory	1285
1216	POLI	-0.21	network, we will focus on the hits	1286
1217	ATXN3L	-0.11	and their neighbors. We will update	1287
1218	FOXO4L4	-0.09	our priors to reflect the observed	1288
1219	CDR1	-0.12	hits and select new genes that are	1289
1220	NELL1	0.03	closely related to the identified	1290
1221	AP1S2	0.06	hits.	1291
1222	IFNB1	-0.13		1292
1223	ARMCX4	-0.14	**Research Plan: The full high-level	1293
1224	ZNF729	-0.10	research plan, with current status	1294
1225	NKX3-2	-0.13	and reasoning behind each proposed	1295
1226	LUZP4	-0.17	approach.**	1296
1227	STX17	0.07		1297
1228	SPINK2	0.29	We will update our prior	1298
1229	MEIG1	0.07	probabilities based on the hits	1299
1230	DAB2IP	-0.22	identified in the previous round and	1300
1231	ZNF611	0.03	select new genes that are closely	1301
1232	VGLL1	-0.05	related to the hits. We will choose	1302
1233	OR52E6	0.06	genes that are in the same	1303
1234	GUCY2F	0.03	biological pathways, have similar	1304
1235	OR10Z1	0.07	functional annotations, or have	1305
1236	Here is a strategy to follow: Update		known interactions with the hits. We	1306
1237	your priors appropriately and		will then select the top-scoring	1307
1238	choose genes that gave you hits.		genes from these candidates to form	1308
1239	Also, be sure to explore by		our next set of predictions. Our	1309
1240	including some genes that could give		goal is to refine our understanding	1310
1241	hits.		of the IL-2 regulatory network and	1311
1242	Please propose 5 different yet valid		identify new genes that play a role	1312
1243	gene names as per the HGNC		in its regulation. We will repeat	1313
1244	nomenclature you want to explore		this process for a total of 5 rounds	1314
1245	next. Note that I will choose		, refining our predictions and	1315
1246	unexplored genes closest to your		updating our priors after each round	1316
1247	predicted genes to form the		.	1317
1248	predictions. Your response should			1318
1249	exactly follow the format:		**Solution:**	1319
1250	**Reflection: Thoughts on previous		## MYBL2	1320
1251	results and next steps.		## GBF1	1321
1252	**Research Plan: The full high level		## DDX41	1322
1253	research plan, with current status		## ZMAT2	1323
1254	and reasoning behind each proposed		## RPL4	1324
1255	approach. It should be at most 5		----END OUTPUT----	1325
1256	sentences.		----ROUND 3: BEGIN PROMPT----	1326
1257	**Solution:		SystemMessage(You are a biomedicine	1327

1328	expert who will assist me on problems	RPS27A	0.48	1398
1329	in drug discovery. I am planning to	RPL27	0.87	1399
1330	run a CRISPR screen to identify genes	PRPF8	0.45	1400
1331	that regulate the production of	CCT4	0.35	1401
1332	Interleukin-2 (IL-2). I can only	RPS16	0.38	1402
1333	perturb exactly 128 genes at a time.	RPL7A	0.61	1403
1334	For each predicted perturbation, I am	CDC7	0.62	1404
1335	able to measure out the log fold	RPS6	0.47	1405
1336	change in Interleukin-2 (IL-2)	RPL23A	0.79	1406
1337	normalized read counts which will be	RPS13	0.92	1407
1338	referred to as the score. I can only	RPL32	0.68	1408
1339	do 5 rounds of experimentation. After	RPL18	0.77	1409
1340	every round of experiment, I will	RPLP2	0.52	1410
1341	provide you with feedback on your	RPS8	0.69	1411
1342	predictions, including the correctly	RPL10A	0.86	1412
1343	identified genes called hits and the	RPL8	0.81	1413
1344	corresponding score. The predictions	MAK16	1.15	1414
1345	which are not hits will be included in	RPS17	0.82	1415
1346	other results.)	RPL19	1.21	1416
1347	HumanMessage(RPL3	0.42	1417
1348	This is round 3.	RPL15	0.92	1418
1349	Here is the feedback on all your	CPSF4	1.13	1419
1350	predictions till now:	XP01	0.35	1420
1351	[HITS]	RRM2	0.50	1421
1352	name score	RPL14	0.61	1422
1353	MYBL2 0.51	[OTHER RESULTS]		1423
1354	GBF1 0.43	name score		1424
1355	DDX41 0.47	ABL1 0.09		1425
1356	ZMAT2 0.45	QRFP 0.00		1426
1357	RPL4 0.35	COA3 -0.05		1427
1358	HNF4A -0.34	E2F6 -0.11		1428
1359	CDC16 0.48	NANOG -0.05		1429
1360	WDR5 0.82	NDUFA4 0.04		1430
1361	MYC 0.41	RAN 0.12		1431
1362	ZNF536 -0.38	MRPL19 -0.14		1432
1363	TLN1 -0.49	SLU7 0.13		1433
1364	LIN37 -0.39	CD8B -0.01		1434
1365	SULT2B1 -0.44	HMGB1 -0.12		1435
1366	SPC24 0.58	AP2S1 0.03		1436
1367	SNRPD2 0.39	RPS28 -0.16		1437
1368	POLR1A 0.47	COX5B 0.07		1438
1369	SF3A2 0.65	COPS6 -0.08		1439
1370	SNRNP200 0.78	RNGTT 0.03		1440
1371	CDK1 0.43	RPL11 0.05		1441
1372	DUT 0.44	NCAPG 0.17		1442
1373	PRC1 0.38	NSF -0.05		1443
1374	DONSON 0.65	NCBP2 0.00		1444
1375	EEF2 0.43	OR51L1 -0.17		1445
1376	PSMA6 0.34	GHR 0.02		1446
1377	XAB2 0.41	TSHZ2 -0.01		1447
1378	VCP 0.66	GPX3 0.11		1448
1379	PSMA1 0.39	TOX2 0.10		1449
1380	PSMD7 0.57	FAM107B 0.05		1450
1381	SMU1 0.45	RGPD3 -0.24		1451
1382	POLE 0.57	TRAPPC3L 0.10		1452
1383	SF3B3 0.36	WFDC6 -0.00		1453
1384	KIF11 0.36	SLC24A3 0.10		1454
1385	ETF1 -0.45	IFNE -0.04		1455
1386	CDC23 0.75	ZFAND2A 0.12		1456
1387	GINS2 0.38	BNIP3L -0.17		1457
1388	RPS11 0.64	FAM184B -0.10		1458
1389	RPS20 0.43	RGMB 0.02		1459
1390	RBM22 0.61	ZNF853 -0.09		1460
1391	FAU 0.49	NPAS2 0.26		1461
1392	RPL17 0.49	ATAD1 -0.05		1462
1393	PRPF19 0.38	SMCP -0.08		1463
1394	RPL9 0.44	RAB27B 0.14		1464
1395	UTP15 0.59	BIN3 -0.10		1465
1396	RPS3A 0.39	NPY2R 0.02		1466
1397	RPL37 0.68	EQTN 0.16		1467

1468	AKR1B1	-0.21	GUCY2F	0.03	1538
1469	MAPK14	-0.02	OR10Z1	0.07	1539
1470	MLST8	-0.15	CLTC	0.26	1540
1471	PPP2R3C	-0.19	THOC7	0.17	1541
1472	MNAT1	0.07	ANKLE2	0.20	1542
1473	NOL10	0.05	SF3A1	0.17	1543
1474	EMC7	0.08	SAP30BP	0.26	1544
1475	EMC3	-0.05	ZMAT5	0.19	1545
1476	GUK1	0.19	LSM2	-0.18	1546
1477	MED6	-0.16	CDC45	0.08	1547
1478	CIAO1	0.10	TANGO6	0.10	1548
1479	SNAPC5	0.01	NUP85	-0.11	1549
1480	ZCRB1	-0.03	TBC1D3B	0.13	1550
1481	TRMT112	0.07	SEC13	0.19	1551
1482	RPS29	0.23	NEDD1	0.32	1552
1483	PDCD2	0.10	RFC2	0.30	1553
1484	RPP21	-0.16	SDE2	0.31	1554
1485	RPL5	0.10	CCT3	0.23	1555
1486	EIF6	0.19	PSMB3	-0.01	1556
1487	SNW1	0.28	KPNB1	0.14	1557
1488	HEATR1	0.06	ANAPC11	-0.07	1558
1489	RPL28	0.34	HSPA9	-0.03	1559
1490	WDR18	0.07	SMG1	0.15	1560
1491	PAK4	0.07	SBN01	-0.03	1561
1492	AFDN	-0.06	PSMB4	0.14	1562
1493	VCL	-0.02	COPB1	0.14	1563
1494	CATSPERG	-0.09	PSMA3	0.05	1564
1495	PTK2	-0.10	COPB2	0.24	1565
1496	BCAR1	-0.08	PSMA5	0.34	1566
1497	ACTR3	-0.06	PSMA2	0.29	1567
1498	RAPGEF1	0.19	CDK11A	-0.06	1568
1499	WWTR1	-0.10	PRELID1	0.22	1569
1500	VWA8	0.26	FARSB	0.13	1570
1501	FERMT2	-0.12	RSL24D1	0.11	1571
1502	RSU1	0.12	RPL12	0.34	1572
1503	MSTO1	0.02	DHX37	0.29	1573
1504	NLRP2B	-0.09	BUD31	0.08	1574
1505	ARPC4	0.05	CDC123	-0.03	1575
1506	ARPC2	-0.12	TXNL4A	-0.25	1576
1507	KIF2C	-0.13	INTS11	0.26	1577
1508	SRF	0.01	DBR1	0.14	1578
1509	RNASE1	0.16	HSPE1	-0.03	1579
1510	NCKAP1	-0.14	NP1PB6	-0.15	1580
1511	ZFAND5	-0.01	PAFAH1B1	-0.26	1581
1512	SMAD2	0.15	NFS1	0.15	1582
1513	MAGED2	-0.03	DDX56	0.21	1583
1514	POTEE	-0.09	CDC27	0.17	1584
1515	IFNA10	-0.18	DDX49	0.22	1585
1516	AKAIN1	-0.18	SF3B4	0.18	1586
1517	ZNF614	0.08	RPL23	0.09	1587
1518	ZNF157	-0.11	WEE1	-0.18	1588
1519	MAPK8	0.09	NIP7	0.22	1589
1520	POLI	-0.21	SF3B5	0.26	1590
1521	ATXN3L	-0.11	CDC20	0.29	1591
1522	FOXO4L4	-0.09	RPS9	0.12	1592
1523	CDR1	-0.12	LRR1	-0.03	1593
1524	NELL1	0.03	GEMIN5	0.00	1594
1525	AP1S2	0.06	SNRPA1	0.06	1595
1526	IFNB1	-0.13	RPLP0	0.26	1596
1527	ARMCX4	-0.14	EIF2S1	0.02	1597
1528	ZNF729	-0.10	DNAJC17	0.04	1598
1529	NKX3-2	-0.13	XRCC6	-0.22	1599
1530	LUZP4	-0.17	DUX4	-0.01	1600
1531	STX17	0.07	RPS19	0.23	1601
1532	SPINK2	0.29	WDR43	0.24	1602
1533	MEIG1	0.07	RRM1	0.32	1603
1534	DAB2IP	-0.22	ATP6V0C	0.06	1604
1535	ZNF611	0.03	RPS2	0.33	1605
1536	VGLL1	-0.05	PUF60	-0.07	1606
1537	OR52E6	0.06	SEC61A1	-0.17	1607

1608	TSR1 0.23	our selection of genes. Our goal is	1678
1609	WDR74 0.33	to identify additional genes that	1679
1610	RPL7 0.31	regulate IL-2 production.	1680
1611	Here is a strategy to follow: Update		1681
1612	your priors appropriately and	**Solution:**	1682
1613	choose genes that gave you hits.	## RPS27	1683
1614	Also, be sure to explore by	## SF3B1	1684
1615	including some genes that could give	## DDX3X	1685
1616	hits.	## RPS15	1686
1617	Please propose 5 different yet valid	## NOLC1	1687
1618	gene names as per the HGNC	----END OUTPUT----	1688
1619	nomenclature you want to explore	----ROUND 4: BEGIN PROMPT----	1689
1620	next. Note that I will choose	SystemMessage(You are a biomedicine	1690
1621	unexplored genes closest to your	expert who will assist me on problems	1691
1622	predicted genes to form the	in drug discovery. I am planning to	1692
1623	predictions. Your response should	run a CRISPR screen to identify genes	1693
1624	exactly follow the format:	that regulate the production of	1694
1625	**Reflection: Thoughts on previous	Interleukin-2 (IL-2). I can only	1695
1626	results and next steps.	perturb exactly 128 genes at a time.	1696
1627	**Research Plan: The full high level	For each predicted perturbation, I am	1697
1628	research plan, with current status	able to measure out the log fold	1698
1629	and reasoning behind each proposed	change in Interleukin-2 (IL-2)	1699
1630	approach. It should be at most 5	normalized read counts which will be	1700
1631	sentences.	referred to as the score. I can only	1701
1632	**Solution:	do 5 rounds of experimentation. After	1702
1633	## <Gene 1>	every round of experiment, I will	1703
1634	## <Gene 2>	provide you with feedback on your	1704
1635	...	predictions, including the correctly	1705
1636	## <Gene 5>	identified genes called hits and the	1706
1637	Each gene in the solution should	corresponding score. The predictions	1707
1638	only be the gene name in the HGNC	which are not hits will be included in	1708
1639	nomenclature.	other results.)	1709
1640	DO NOT ADD ANY COMMENTS IN THE	HumanMessage(1710
1641	SOLUTION OR AFTER THE SOLUTION.	This is round 4.	1711
1642)	Here is the feedback on all your	1712
1643	----END PROMPT----	predictions till now:	1713
1644	----BEGIN OUTPUT----	[HITS]	1714
1645	**Reflection: Thoughts on previous	name score	1715
1646	results and next steps.**	MYBL2 0.51	1716
1647	Based on the previous round of	GBF1 0.43	1717
1648	experimentation, we have identified	DDX41 0.47	1718
1649	several genes that are associated	ZMAT2 0.45	1719
1650	with the regulation of Interleukin-2	RPL4 0.35	1720
1651	(IL-2) production. The hits from	HNF4A -0.34	1721
1652	this round include genes involved in	CDC16 0.48	1722
1653	various cellular processes such as	WDR5 0.82	1723
1654	RNA processing, translation, and	MYC 0.41	1724
1655	cell cycle regulation. We can update	ZNF536 -0.38	1725
1656	our priors to reflect the	TLN1 -0.49	1726
1657	importance of these genes in IL-2	LIN37 -0.39	1727
1658	regulation. To further explore the	SULT2B1 -0.44	1728
1659	regulatory network, we will choose	SPC24 0.58	1729
1660	genes that are functionally related	SNRPD2 0.39	1730
1661	to the hits from the previous round.	POLR1A 0.47	1731
1662		SF3A2 0.65	1732
1663	**Research Plan: The full high-level	SNRNP200 0.78	1733
1664	research plan, with current status	CDK1 0.43	1734
1665	and reasoning behind each proposed	DUT 0.44	1735
1666	approach.**	PRC1 0.38	1736
1667	We will select genes that are	DONSON 0.65	1737
1668	functionally related to the hits	EEF2 0.43	1738
1669	from the previous round, focusing on	PSMA6 0.34	1739
1670	RNA processing, translation, and	XAB2 0.41	1740
1671	cell cycle regulation. We will	VCP 0.66	1741
1672	choose unexplored genes closest to	PSMA1 0.39	1742
1673	the hits to form the predictions.	PSMD7 0.57	1743
1674	This approach will allow us to	SMU1 0.45	1744
1675	further refine our understanding of	POLE 0.57	1745
1676	the IL-2 regulatory network. We will	SF3B3 0.36	1746
1677	use the updated priors to inform	KIF11 0.36	1747

1748	ETF1	-0.45	BOP1	0.51	1818
1749	CDC23	0.75	NOP2	0.55	1819
1750	GIN52	0.38	RRN3	0.35	1820
1751	RPS11	0.64	TUBGCP2	0.59	1821
1752	RPS20	0.43	RPS3	0.54	1822
1753	RBM22	0.61	[OTHER RESULTS]		1823
1754	FAU	0.49	name	score	1824
1755	RPL17	0.49	ABL1	0.09	1825
1756	PRPF19	0.38	QRFP	0.00	1826
1757	RPL9	0.44	COA3	-0.05	1827
1758	UTP15	0.59	E2F6	-0.11	1828
1759	RPS3A	0.39	NANOG	-0.05	1829
1760	RPL37	0.68	NDUFA4	0.04	1830
1761	RPS27A	0.48	RAN	0.12	1831
1762	RPL27	0.87	MRPL19	-0.14	1832
1763	PRPF8	0.45	SLU7	0.13	1833
1764	CCT4	0.35	CD8B	-0.01	1834
1765	RPS16	0.38	HMGB1	-0.12	1835
1766	RPL7A	0.61	AP2S1	0.03	1836
1767	CDC7	0.62	RPS28	-0.16	1837
1768	RPS6	0.47	COX5B	0.07	1838
1769	RPL23A	0.79	COPS6	-0.08	1839
1770	RPS13	0.92	RNGTT	0.03	1840
1771	RPL32	0.68	RPL11	0.05	1841
1772	RPL18	0.77	NCAPG	0.17	1842
1773	RPLP2	0.52	NSF	-0.05	1843
1774	RPS8	0.69	NCBP2	0.00	1844
1775	RPL10A	0.86	OR51L1	-0.17	1845
1776	RPL8	0.81	GHR	0.02	1846
1777	MAK16	1.15	TSHZ2	-0.01	1847
1778	RPS17	0.82	GPX3	0.11	1848
1779	RPL19	1.21	TOX2	0.10	1849
1780	RPL3	0.42	FAM107B	0.05	1850
1781	RPL15	0.92	RGPD3	-0.24	1851
1782	CPSF4	1.13	TRAPPC3L	0.10	1852
1783	XPO1	0.35	WFDC6	-0.00	1853
1784	RRM2	0.50	SLC24A3	0.10	1854
1785	RPL14	0.61	IFNE	-0.04	1855
1786	INTS3	0.53	ZFAND2A	0.12	1856
1787	RFC3	0.50	BNIP3L	-0.17	1857
1788	RBM25	0.47	FAM184B	-0.10	1858
1789	FCF1	0.40	RGMB	0.02	1859
1790	POLR1B	0.36	ZNF853	-0.09	1860
1791	POLD3	0.52	NPAS2	0.26	1861
1792	CHMP6	0.85	ATAD1	-0.05	1862
1793	PSMC5	0.34	SMCP	-0.08	1863
1794	RPL31	0.40	RAB27B	0.14	1864
1795	ERH	0.45	BIN3	-0.10	1865
1796	RPS4X	0.70	NPY2R	0.02	1866
1797	CHERP	0.49	EQTN	0.16	1867
1798	DKC1	0.40	AKR1B1	-0.21	1868
1799	CNOT3	0.44	MAPK14	-0.02	1869
1800	SNRPC	1.25	MLST8	-0.15	1870
1801	MTBP	0.57	PPP2R3C	-0.19	1871
1802	SYMPK	0.68	MNAT1	0.07	1872
1803	CDC6	0.47	NOL10	0.05	1873
1804	PPAN	0.89	EMC7	0.08	1874
1805	SPOUT1	0.38	EMC3	-0.05	1875
1806	EIF3I	0.45	GUK1	0.19	1876
1807	RPL36	0.72	MED6	-0.16	1877
1808	NUP93	0.35	CIAO1	0.10	1878
1809	RPS24	0.83	SNAPC5	0.01	1879
1810	NUP133	0.49	ZCRB1	-0.03	1880
1811	RPS18	0.46	TRMT112	0.07	1881
1812	RPS14	0.60	RPS29	0.23	1882
1813	PDCD11	0.43	PDCD2	0.10	1883
1814	NOC3L	0.34	RPP21	-0.16	1884
1815	BMS1	0.56	RPL5	0.10	1885
1816	RPS25	0.49	EIF6	0.19	1886
1817	EBNA1BP2	0.39	SNW1	0.28	1887

1888	HEATR1	0.06	ANAPC11	-0.07	1958
1889	RPL28	0.34	HSPA9	-0.03	1959
1890	WDR18	0.07	SMG1	0.15	1960
1891	PAK4	0.07	SBN01	-0.03	1961
1892	AFDN	-0.06	PSMB4	0.14	1962
1893	VCL	-0.02	COPB1	0.14	1963
1894	CATSPERG	-0.09	PSMA3	0.05	1964
1895	PTK2	-0.10	COPB2	0.24	1965
1896	BCAR1	-0.08	PSMA5	0.34	1966
1897	ACTR3	-0.06	PSMA2	0.29	1967
1898	RAPGEF1	0.19	CDK11A	-0.06	1968
1899	WWTR1	-0.10	PRELID1	0.22	1969
1900	VWA8	0.26	FARSB	0.13	1970
1901	FERMT2	-0.12	RSL24D1	0.11	1971
1902	RSU1	0.12	RPL12	0.34	1972
1903	MST01	0.02	DHX37	0.29	1973
1904	NLRP2B	-0.09	BUD31	0.08	1974
1905	ARPC4	0.05	CDC123	-0.03	1975
1906	ARPC2	-0.12	TXNL4A	-0.25	1976
1907	KIF2C	-0.13	INTS11	0.26	1977
1908	SRF	0.01	DBR1	0.14	1978
1909	RNASE1	0.16	HSPE1	-0.03	1979
1910	NCKAP1	-0.14	NPIPB6	-0.15	1980
1911	ZFAND5	-0.01	PAFAH1B1	-0.26	1981
1912	SMAD2	0.15	NFS1	0.15	1982
1913	MAGED2	-0.03	DDX56	0.21	1983
1914	POTEE	-0.09	CDC27	0.17	1984
1915	IFNA10	-0.18	DDX49	0.22	1985
1916	AKAIN1	-0.18	SF3B4	0.18	1986
1917	ZNF614	0.08	RPL23	0.09	1987
1918	ZNF157	-0.11	WEE1	-0.18	1988
1919	MAPK8	0.09	NIP7	0.22	1989
1920	POLI	-0.21	SF3B5	0.26	1990
1921	ATXN3L	-0.11	CDC20	0.29	1991
1922	FOXO4L4	-0.09	RPS9	0.12	1992
1923	CDR1	-0.12	LRR1	-0.03	1993
1924	NELL1	0.03	GEMIN5	0.00	1994
1925	AP1S2	0.06	SNRPA1	0.06	1995
1926	IFNB1	-0.13	RPLP0	0.26	1996
1927	ARMCX4	-0.14	EIF2S1	0.02	1997
1928	ZNF729	-0.10	DNAJC17	0.04	1998
1929	NKX3-2	-0.13	XRCC6	-0.22	1999
1930	LUZP4	-0.17	DUX4	-0.01	2000
1931	STX17	0.07	RPS19	0.23	2001
1932	SPINK2	0.29	WDR43	0.24	2002
1933	MEIG1	0.07	RRM1	0.32	2003
1934	DAB2IP	-0.22	ATP6V0C	0.06	2004
1935	ZNF611	0.03	RPS2	0.33	2005
1936	VGLL1	-0.05	PUF60	-0.07	2006
1937	OR52E6	0.06	SEC61A1	-0.17	2007
1938	GUCY2F	0.03	TSR1	0.23	2008
1939	OR10Z1	0.07	WDR74	0.33	2009
1940	CLTC	0.26	RPL7	0.31	2010
1941	THOC7	0.17	RPS27	0.13	2011
1942	ANKLE2	0.20	MCM7	0.01	2012
1943	SF3A1	0.17	LUC7L3	0.12	2013
1944	SAP30BP	0.26	EEF1A1	0.21	2014
1945	ZMAT5	0.19	RPSA	0.28	2015
1946	LSM2	-0.18	PPWD1	0.29	2016
1947	CDC45	0.08	TOMM22	-0.05	2017
1948	TANGO6	0.10	RACGAP1	0.25	2018
1949	NUP85	-0.11	DYNC1H1	-0.17	2019
1950	TBC1D3B	0.13	SNRNP25	0.01	2020
1951	SEC13	0.19	TIMELESS	0.11	2021
1952	NEDD1	0.32	UQCRH	-0.09	2022
1953	RFC2	0.30	PAM16	0.19	2023
1954	SDE2	0.31	PFDN6	-0.21	2024
1955	CCT3	0.23	DDX10	0.32	2025
1956	PSMB3	-0.01	RRP12	0.16	2026
1957	KPNB1	0.14	GPN1	0.08	2027

2028	PFDN2	0.06	RBM17	0.00	2098
2029	ZNHIT2	0.20	XRN2	0.06	2099
2030	BYSL	0.04	OSGEP	0.10	2100
2031	USP36	-0.03	UTP18	0.02	2101
2032	SF3B1	0.30	Here is a strategy to follow: Update		2102
2033	VPS25	-0.08	your priors appropriately and		2103
2034	PRPF38A	0.15	choose genes that gave you hits.		2104
2035	EIF5	-0.12	Also, be sure to explore by		2105
2036	ATP6V0B	0.06	including some genes that could give		2106
2037	RNPC3	-0.02	hits.		2107
2038	PSMA4	0.23	Please propose 5 different yet valid		2108
2039	UBL5	-0.00	gene names as per the HGNC		2109
2040	EIF4A3	0.33	nomenclature you want to explore		2110
2041	POLE2	0.09	next. Note that I will choose		2111
2042	GNP3	0.15	unexplored genes closest to your		2112
2043	PSMD3	0.27	predicted genes to form the		2113
2044	COPA	0.05	predictions. Your response should		2114
2045	CHMP2A	0.06	exactly follow the format:		2115
2046	RPL27A	0.27	**Reflection: Thoughts on previous		2116
2047	PSMD1	0.32	results and next steps.		2117
2048	CWC22	0.22	**Research Plan: The full high level		2118
2049	WBP11	0.13	research plan, with current status		2119
2050	PSMD4	0.23	and reasoning behind each proposed		2120
2051	DDX3X	0.03	approach. It should be at most 5		2121
2052	EIF1AX	0.10	sentences.		2122
2053	UBA1	0.21	**Solution:		2123
2054	EIF2S3	0.10	## <Gene 1>		2124
2055	TSR2	0.19	## <Gene 2>		2125
2056	POLA1	0.24	...		2126
2057	DHDDS	-0.26	## <Gene 5>		2127
2058	ANAPC2	0.23	Each gene in the solution should		2128
2059	GPKOW	0.07	only be the gene name in the HGNC		2129
2060	HCFC1	0.06	nomenclature.		2130
2061	PDRG1	-0.16	DO NOT ADD ANY COMMENTS IN THE		2131
2062	TRAPPC3	0.07	SOLUTION OR AFTER THE SOLUTION.		2132
2063	LSM3	0.05)		2133
2064	ATP6AP2	-0.05	----END PROMPT----		2134
2065	GNP2	0.03	----BEGIN OUTPUT----		2135
2066	DTL	0.06	**Reflection: Thoughts on previous		2136
2067	ZFP36L2	0.01	results and next steps.**		2137
2068	RPS15A	0.21	The previous results have identified		2138
2069	NEDD8	0.09	several genes that are involved in		2139
2070	WDR75	0.31	the regulation of Interleukin-2 (IL		2140
2071	NSA2	0.12	-2) production. The top-scoring		2141
2072	TRNT1	0.09	genes from the previous round		2142
2073	ISCU	-0.01	include RPL27, RPL23A, RPL18, RPS13,		2143
2074	MRPL13	0.09	RPL10A, RPL8, MAK16, RPL19, CPSF4,		2144
2075	DDX55	0.01	and RPL15. These genes are mostly		2145
2076	BIRC5	0.30	ribosomal proteins, suggesting that		2146
2077	ARL2	0.02	the regulation of IL-2 production is		2147
2078	PWP2	0.02	tightly linked to protein synthesis		2148
2079	SPPL2B	0.01	. The next steps will focus on		2149
2080	NIFK	0.26	exploring more genes that are		2150
2081	ABCE1	0.28	closely related to these hits and		2151
2082	PRIM1	0.11	are not yet explored.		2152
2083	EIF3G	0.08			2153
2084	NOLC1	0.02	**Research Plan: The full high level		2154
2085	NOC4L	0.11	research plan, with current status		2155
2086	UTP20	0.33	and reasoning behind each proposed		2156
2087	NOP10	-0.04	approach.**		2157
2088	URB1	0.04	We will update our priors based on		2158
2089	NMD3	0.33	the previous results and choose		2159
2090	ZPR1	0.15	genes that are closely related to		2160
2091	NOL9	0.11	the hits from the previous round. We		2161
2092	TCP1	0.02	will also include some genes that		2162
2093	TBCB	0.10	could give hits based on their		2163
2094	UBA2	-0.17	functional annotation. We will use		2164
2095	RSL1D1	0.28	the HGNC nomenclature to select 5		2165
2096	PAK1IP1	0.22	unexplored genes that are closest to		2166
2097	NOP56	-0.08	the predicted genes. We will then		2167

2168	use these genes for the next round	ETF1	-0.45	2238
2169	of experimentation.	CDC23	0.75	2239
2170		GINS2	0.38	2240
2171	**Solution:**	RPS11	0.64	2241
2172	## RPL22	RPS20	0.43	2242
2173	## RPS11	RBM22	0.61	2243
2174	## RPL14	FAU	0.49	2244
2175	## RPS4X	RPL17	0.49	2245
2176	## RPL32	PRPF19	0.38	2246
2177	----END OUTPUT----	RPL9	0.44	2247
2178	----ROUND 5: BEGIN PROMPT----	UTP15	0.59	2248
2179	SystemMessage(You are a biomedicine	RPS3A	0.39	2249
2180	expert who will assist me on	RPL37	0.68	2250
2181	problems in drug discovery. I am	RPS27A	0.48	2251
2182	planning to run a CRISPR screen to	RPL27	0.87	2252
2183	identify genes that regulate the	PRPF8	0.45	2253
2184	production of Interleukin-2 (IL-2).	CCT4	0.35	2254
2185	I can only perturb exactly 128 genes	RPS16	0.38	2255
2186	at a time. For each predicted	RPL7A	0.61	2256
2187	perturbation, I am able to measure	CDC7	0.62	2257
2188	out the log fold change in	RPS6	0.47	2258
2189	Interleukin-2 (IL-2) normalized read	RPL23A	0.79	2259
2190	counts which will be referred to as	RPS13	0.92	2260
2191	the score. I can only do 5 rounds	RPL32	0.68	2261
2192	of experimentation. After every	RPL18	0.77	2262
2193	round of experiment, I will provide	RPLP2	0.52	2263
2194	you with feedback on your	RPS8	0.69	2264
2195	predictions, including the correctly	RPL10A	0.86	2265
2196	identified genes called hits and	RPL8	0.81	2266
2197	the corresponding score. The	MAK16	1.15	2267
2198	predictions which are not hits will	RPS17	0.82	2268
2199	be included in other results.)	RPL19	1.21	2269
2200	HumanMessage(RPL3	0.42	2270
2201	This is round 5.	RPL15	0.92	2271
2202	Here is the feedback on all your	CPSF4	1.13	2272
2203	predictions till now:	XPO1	0.35	2273
2204	[HITS]	RRM2	0.50	2274
2205	name score	RPL14	0.61	2275
2206	MYBL2 0.51	INTS3	0.53	2276
2207	GBF1 0.43	RFC3	0.50	2277
2208	DDX41 0.47	RBM25	0.47	2278
2209	ZMAT2 0.45	FCF1	0.40	2279
2210	RPL4 0.35	POLR1B	0.36	2280
2211	HNF4A -0.34	POLD3	0.52	2281
2212	CDC16 0.48	CHMP6	0.85	2282
2213	WDR5 0.82	PSMC5	0.34	2283
2214	MYC 0.41	RPL31	0.40	2284
2215	ZNF536 -0.38	ERH	0.45	2285
2216	TLN1 -0.49	RPS4X	0.70	2286
2217	LIN37 -0.39	CHERP	0.49	2287
2218	SULT2B1 -0.44	DKC1	0.40	2288
2219	SPC24 0.58	CNOT3	0.44	2289
2220	SNRPD2 0.39	SNRPC	1.25	2290
2221	POLR1A 0.47	MTBP	0.57	2291
2222	SF3A2 0.65	SYMPK	0.68	2292
2223	SNRNP200 0.78	CDC6	0.47	2293
2224	CDK1 0.43	PPAN	0.89	2294
2225	DUT 0.44	SPOUT1	0.38	2295
2226	PRC1 0.38	EIF3I	0.45	2296
2227	DONSON 0.65	RPL36	0.72	2297
2228	EEF2 0.43	NUP93	0.35	2298
2229	PSMA6 0.34	RPS24	0.83	2299
2230	XAB2 0.41	NUP133	0.49	2300
2231	VCP 0.66	RPS18	0.46	2301
2232	PSMA1 0.39	RPS14	0.60	2302
2233	PSMD7 0.57	PDCD11	0.43	2303
2234	SMU1 0.45	NOC3L	0.34	2304
2235	POLE 0.57	BMS1	0.56	2305
2236	SF3B3 0.36	RPS25	0.49	2306
2237	KIF11 0.36	EBNA1BP2	0.39	2307

2308	BOP1	0.51	BNIP3L	-0.17	2378
2309	NOP2	0.55	FAM184B	-0.10	2379
2310	RRN3	0.35	RGMB	0.02	2380
2311	TUBGCP2	0.59	ZNF853	-0.09	2381
2312	RPS3	0.54	NPAS2	0.26	2382
2313	PKMYT1	0.46	ATAD1	-0.05	2383
2314	CLNS1A	0.47	SMCP	-0.08	2384
2315	GINS1	0.47	RAB27B	0.14	2385
2316	SNRNP27	0.63	BIN3	-0.10	2386
2317	TNPO3	0.44	NPY2R	0.02	2387
2318	MCM5	0.43	EQTN	0.16	2388
2319	BUB3	0.48	AKR1B1	-0.21	2389
2320	WDR12	0.43	MAPK14	-0.02	2390
2321	NUTF2	0.45	MLST8	-0.15	2391
2322	RPLP1	0.66	PPP2R3C	-0.19	2392
2323	DYNC1I2	0.62	MNAT1	0.07	2393
2324	USP5	0.34	NOL10	0.05	2394
2325	RPL35	0.98	EMC7	0.08	2395
2326	RPL26	0.99	EMC3	-0.05	2396
2327	GINS4	0.59	GUK1	0.19	2397
2328	EIF3B	0.48	MED6	-0.16	2398
2329	GNL3L	0.39	CIAO1	0.10	2399
2330	SMC1A	0.68	SNAPC5	0.01	2400
2331	RPL34	0.67	ZCRB1	-0.03	2401
2332	DAD1	-0.42	TRMT112	0.07	2402
2333	MPHOSPH10	0.47	RPS29	0.23	2403
2334	RPL6	0.62	PDCD2	0.10	2404
2335	GNL2	0.42	RPP21	-0.16	2405
2336	RUVBL1	0.56	RPL5	0.10	2406
2337	RPS7	1.02	EIF6	0.19	2407
2338	SNRPF	0.57	SNW1	0.28	2408
2339	MFAP1	0.42	HEATR1	0.06	2409
2340	SRBD1	0.56	RPL28	0.34	2410
2341	POLR1C	0.41	WDR18	0.07	2411
2342	NOB1	0.38	PAK4	0.07	2412
2343	PSMD6	0.43	AFDN	-0.06	2413
2344	[OTHER RESULTS]		VCL	-0.02	2414
2345	name	score	CATSPERG	-0.09	2415
2346	ABL1	0.09	PTK2	-0.10	2416
2347	QRFP	0.00	BCAR1	-0.08	2417
2348	COA3	-0.05	ACTR3	-0.06	2418
2349	E2F6	-0.11	RAPGEF1	0.19	2419
2350	NANOG	-0.05	WWTR1	-0.10	2420
2351	NDUFA4	0.04	VWA8	0.26	2421
2352	RAN	0.12	FERMT2	-0.12	2422
2353	MRPL19	-0.14	RSU1	0.12	2423
2354	SLU7	0.13	MSTO1	0.02	2424
2355	CD8B	-0.01	NLRP2B	-0.09	2425
2356	HMGB1	-0.12	ARPC4	0.05	2426
2357	AP2S1	0.03	ARPC2	-0.12	2427
2358	RPS28	-0.16	KIF2C	-0.13	2428
2359	COX5B	0.07	SRF	0.01	2429
2360	COPS6	-0.08	RNASE1	0.16	2430
2361	RNGTT	0.03	NCKAP1	-0.14	2431
2362	RPL11	0.05	ZFAND5	-0.01	2432
2363	NCAPG	0.17	SMAD2	0.15	2433
2364	NSF	-0.05	MAGED2	-0.03	2434
2365	NCBP2	0.00	POTEE	-0.09	2435
2366	OR51L1	-0.17	IFNA10	-0.18	2436
2367	GHR	0.02	AKAIN1	-0.18	2437
2368	TSHZ2	-0.01	ZNF614	0.08	2438
2369	GPX3	0.11	ZNF157	-0.11	2439
2370	TOX2	0.10	MAPK8	0.09	2440
2371	FAM107B	0.05	POLI	-0.21	2441
2372	RGPD3	-0.24	ATXN3L	-0.11	2442
2373	TRAPPC3L	0.10	FOXD4L4	-0.09	2443
2374	WFDC6	-0.00	CDR1	-0.12	2444
2375	SLC24A3	0.10	NELL1	0.03	2445
2376	IFNE	-0.04	AP1S2	0.06	2446
2377	ZFAND2A	0.12	IFNB1	-0.13	2447

2448	ARMCX4	-0.14	EIF2S1	0.02	2518
2449	ZNF729	-0.10	DNAJC17	0.04	2519
2450	NKX3-2	-0.13	XRCC6	-0.22	2520
2451	LUZP4	-0.17	DUX4	-0.01	2521
2452	STX17	0.07	RPS19	0.23	2522
2453	SPINK2	0.29	WDR43	0.24	2523
2454	MEIG1	0.07	RRM1	0.32	2524
2455	DAB2IP	-0.22	ATP6V0C	0.06	2525
2456	ZNF611	0.03	RPS2	0.33	2526
2457	VGLL1	-0.05	PUF60	-0.07	2527
2458	OR52E6	0.06	SEC61A1	-0.17	2528
2459	GUCY2F	0.03	TSR1	0.23	2529
2460	OR10Z1	0.07	WDR74	0.33	2530
2461	CLTC	0.26	RPL7	0.31	2531
2462	THOC7	0.17	RPS27	0.13	2532
2463	ANKLE2	0.20	MCM7	0.01	2533
2464	SF3A1	0.17	LUC7L3	0.12	2534
2465	SAP30BP	0.26	EEF1A1	0.21	2535
2466	ZMAT5	0.19	RPSA	0.28	2536
2467	LSM2	-0.18	PPWD1	0.29	2537
2468	CDC45	0.08	TOMM22	-0.05	2538
2469	TANGO6	0.10	RACGAP1	0.25	2539
2470	NUP85	-0.11	DYNC1H1	-0.17	2540
2471	TBC1D3B	0.13	SNRNP25	0.01	2541
2472	SEC13	0.19	TIMELESS	0.11	2542
2473	NEDD1	0.32	UQCRH	-0.09	2543
2474	RFC2	0.30	PAM16	0.19	2544
2475	SDE2	0.31	PFDN6	-0.21	2545
2476	CCT3	0.23	DDX10	0.32	2546
2477	PSMB3	-0.01	RRP12	0.16	2547
2478	KPNB1	0.14	GPN1	0.08	2548
2479	ANAPC11	-0.07	PFDN2	0.06	2549
2480	HSPA9	-0.03	ZNHIT2	0.20	2550
2481	SMG1	0.15	BYSL	0.04	2551
2482	SBN01	-0.03	USP36	-0.03	2552
2483	PSMB4	0.14	SF3B1	0.30	2553
2484	COPB1	0.14	VPS25	-0.08	2554
2485	PSMA3	0.05	PRPF38A	0.15	2555
2486	COPB2	0.24	EIF5	-0.12	2556
2487	PSMA5	0.34	ATP6V0B	0.06	2557
2488	PSMA2	0.29	RNPC3	-0.02	2558
2489	CDK11A	-0.06	PSMA4	0.23	2559
2490	PRELID1	0.22	UBL5	-0.00	2560
2491	FARSB	0.13	EIF4A3	0.33	2561
2492	RSL24D1	0.11	POLE2	0.09	2562
2493	RPL12	0.34	GPN3	0.15	2563
2494	DHX37	0.29	PSMD3	0.27	2564
2495	BUD31	0.08	COPA	0.05	2565
2496	CDC123	-0.03	CHMP2A	0.06	2566
2497	TXNL4A	-0.25	RPL27A	0.27	2567
2498	INTS11	0.26	PSMD1	0.32	2568
2499	DBR1	0.14	CWC22	0.22	2569
2500	HSPE1	-0.03	WBP11	0.13	2570
2501	NPIPB6	-0.15	PSMD4	0.23	2571
2502	PAFAH1B1	-0.26	DDX3X	0.03	2572
2503	NFS1	0.15	EIF1AX	0.10	2573
2504	DDX56	0.21	UBA1	0.21	2574
2505	CDC27	0.17	EIF2S3	0.10	2575
2506	DDX49	0.22	TSR2	0.19	2576
2507	SF3B4	0.18	POLA1	0.24	2577
2508	RPL23	0.09	DHDDS	-0.26	2578
2509	WEE1	-0.18	ANAPC2	0.23	2579
2510	NIP7	0.22	GPKOW	0.07	2580
2511	SF3B5	0.26	HCFC1	0.06	2581
2512	CDC20	0.29	PDRG1	-0.16	2582
2513	RPS9	0.12	TRAPPC3	0.07	2583
2514	LRR1	-0.03	LSM3	0.05	2584
2515	GEMIN5	0.00	ATP6AP2	-0.05	2585
2516	SNRPA1	0.06	GPN2	0.03	2586
2517	RPLP0	0.26	DTL	0.06	2587

2588	ZFP36L2	0.01	NOP58	0.22	2658
2589	RPS15A	0.21	SSU72	0.13	2659
2590	NEDD8	0.09	TOMM40	-0.06	2660
2591	WDR75	0.31	HSPD1	-0.02	2661
2592	NSA2	0.12	YKT6	-0.04	2662
2593	TRNT1	0.09	ATP6V1A	-0.07	2663
2594	ISCU	-0.01	LSM7	0.09	2664
2595	MRPL13	0.09	ATP6V1F	0.03	2665
2596	DDX55	0.01	CSE1L	0.06	2666
2597	BIRC5	0.30	RBM19	0.14	2667
2598	ARL2	0.02	RPA3	-0.02	2668
2599	PWP2	0.02	PSMB2	0.17	2669
2600	SPPL2B	0.01	LSM8	0.16	2670
2601	NIFK	0.26	SRP54	-0.19	2671
2602	ABCE1	0.28	CCT2	0.10	2672
2603	PRIM1	0.11	LYRM4	-0.02	2673
2604	EIF3G	0.08	MDN1	0.32	2674
2605	NOLC1	0.02	POLD1	0.30	2675
2606	NOC4L	0.11	RABGGTB	-0.07	2676
2607	UTP20	0.33	ATP2A2	-0.01	2677
2608	NOP10	-0.04	TTC27	0.24	2678
2609	URB1	0.04	PSMD13	0.15	2679
2610	NMD3	0.33	EIF3E	0.21	2680
2611	ZPR1	0.15	ALG14	-0.08	2681
2612	NOL9	0.11	NVL	0.23	2682
2613	TCP1	0.02	HTATSF1	0.02	2683
2614	TBCB	0.10	ZRSR2	0.06	2684
2615	UBA2	-0.17	OGT	0.02	2685
2616	RSL1D1	0.28	RPN1	-0.02	2686
2617	PAK1IP1	0.22	SF3B2	0.29	2687
2618	NOP56	-0.08	DDB1	0.23	2688
2619	RBM17	0.00	NUS1	-0.04	2689
2620	XRN2	0.06	RBMX2	-0.19	2690
2621	OSGEP	0.10	MMS22L	0.26	2691
2622	UTP18	0.02	TIMM8A	0.15	2692
2623	RPL22	-0.23	ZNF830	0.13	2693
2624	TP53BP1	-0.04	ALG2	0.09	2694
2625	NELFCD	0.03	RNF113A	0.28	2695
2626	METTL14	0.14	INTS1	0.01	2696
2627	DEPDC1	0.00	PSMG4	-0.16	2697
2628	WTAP	0.12	DDOST	-0.20	2698
2629	CA6	-0.08	SNAPC1	0.04	2699
2630	PREB	0.13	NDC80	0.05	2700
2631	ZNF676	-0.04	NKAP	0.27	2701
2632	PRAMEF18	-0.11	EIF2B2	-0.17	2702
2633	STN1	0.01	INTS4	0.03	2703
2634	SLC35G6	-0.20	EXOSC8	0.07	2704
2635	ZNF318	-0.13	TUT1	0.20	2705
2636	DPPA2	0.18	RPAP1	-0.01	2706
2637	GSTM3	0.11	KIF23	0.26	2707
2638	SRSF10	0.13	PLA2G10	-0.09	2708
2639	SCAF4	-0.04	SPC25	0.31	2709
2640	PNRC2	0.07	PSMA7	-0.00	2710
2641	IFITM3	0.00	ANKRD36B	-0.07	2711
2642	WDR55	0.11	PSMC6	0.27	2712
2643	TMA7	-0.06	PRPF38B	0.06	2713
2644	MRPS2	0.14	KRI1	0.22	2714
2645	TSEN54	0.21	SNAPC2	0.05	2715
2646	GGPS1	0.06	POP5	0.20	2716
2647	EIF1AD	0.19	GAPDH	-0.10	2717
2648	RPL18A	0.31	SF3A3	0.34	2718
2649	SAE1	-0.03	TRAPPC1	0.15	2719
2650	RABGGTA	0.25	Here is a strategy to follow: Update your priors appropriately and choose genes that gave you hits. Also, be sure to explore by including some genes that could give hits. Please propose 5 different yet valid gene names as per the HGNC		2720
2651	SS18L2	0.10			2721
2652	MRPS12	0.06			2722
2653	HAUS5	0.25			2723
2654	FOXD4	0.15			2724
2655	TOP2A	0.13			2725
2656	SNRNP70	0.33			2726
2657	INTS9	0.01			2727

2728	nomenclature you want to explore	RPS0L70, RPS0L71, RPS0L72, RPS0L73,	2798
2729	next. Note that I will choose	RPS0L74, RPS0L75, RPS0L76, RPS0L77,	2799
2730	unexplored genes closest to your	RPS0L78, RPS0L79, RPS0L80, RPS0L81,	2800
2731	predicted genes to form the	RPS0L82, RPS0L83, RPS0L84, RPS0L85,	2801
2732	predictions. Your response should	RPS0L86, RPS0L87, RPS0L88, RPS0L89,	2802
2733	exactly follow the format:	RPS0L90, RPS0L91, RPS0L92, RPS0L93,	2803
2734	**Reflection: Thoughts on previous	RPS0L94, RPS0L95, RPS0L96, RPS0L97,	2804
2735	results and next steps.	RPS0L98, RPS0L99, RPS0L100, RPS0L101,	2805
2736	**Research Plan: The full high level	RPS0L102, RPS0L103, RPS0L104, RPS0L105,	2806
2737	research plan, with current status	RPS0L106, RPS0L107, RPS0L108, RPS0L109,	2807
2738	and reasoning behind each proposed	RPS0L110, RPS0L111, RPS0L112, RPS0L113,	2808
2739	approach. It should be at most 5	RPS0L114, RPS0L115, RPS0L116, RPS0L117,	2809
2740	sentences.	RPS0L118, RPS0L119, RPS0L120, RPS0L121,	2810
2741	**Solution:	RPS0L122, RPS0L123, RPS0L124, RPS0L125,	2811
2742	## <Gene 1>	RPS0L126, RPS0L127, RPS0L128. These	2812
2743	## <Gene 2>	genes are highly enriched in ribosomal	2813
2744	...	proteins and other proteins involved in	2814
2745	## <Gene 5>	RNA processing and translation. We will	2815
2746	Each gene in the solution should	update our priors to reflect the	2816
2747	only be the gene name in the HGNC	importance of these gene families in IL	2817
2748	nomenclature.	-2 regulation. We will also explore	2818
2749	DO NOT ADD ANY COMMENTS IN THE	genes that are closely related to these	2819
2750	SOLUTION OR AFTER THE SOLUTION.	families, such as other ribosomal	2820
2751)	proteins and RNA processing genes.	2821
2752	----END PROMPT----		2822
2753	----BEGIN OUTPUT----	**Research Plan:**	2823
2754	**Reflection: Thoughts on previous	We will choose genes that are closely	2824
2755	results and next steps.**	related to the highly scoring genes from	2825
2756	From the previous rounds of	previous rounds. Our goal is to	2826
2757	experimentation, we have identified	identify new genes that regulate IL-2	2827
2758	several genes that regulate the	production and to refine our	2828
2759	production of Interleukin-2 (IL-2). The	understanding of the underlying	2829
2760	top-scoring genes from previous rounds	biological pathways.	2830
2761	include WDR5, SF3A2, SNRNP200, RPL27,		2831
2762	RPL3, RPL15, RPL10A, RPL8, MAK16, RPL19,	**Solution:**	2832
2763	CPSF4, SNRPF, RPS27A, RPL36, RPS24,	## RPL38	2833
2764	RPL37, RPL32, RPL18, RPS8, RPLP2, RPS13,	## RPL31	2834
2765	RPL23A, RPL2, RPL26, RPL35, RPS17,	## RPL18A	2835
2766	RPL14, RPS7, RPS29, RPS25, RPS3, RPS4X,	## SNRNP70	2836
2767	RPS18, RPS16, RPS15, RPS14, RPS13, RPS12,	## RPS27	2837
2768	, RPS11, RPS10, RPS9, RPS8, RPS7, RPS6,	----END OUTPUT----	2838
2769	RPS5, RPS4, RPS3, RPS2, RPS1, RPS0A,	SECURED TOTAL 169 HITS	2840
2770	RPS0L1, RPS0L2, RPS0L3, RPS0L4, RPS0L5,		
2771	RPL30, RPL29, RPL28, RPL27, RPL26, RPL25		
2772	, RPL24, RPL23, RPL22, RPL21, RPL20,		
2773	RPL19, RPL18, RPL17, RPL16, RPL15, RPL14		
2774	, RPL13, RPL12, RPL11, RPL10, RPL9, RPL8		
2775	, RPL7, RPL6, RPL5, RPL4, RPL3, RPL2,		
2776	RPL1, RPL0, RPS27, RPS26, RPS25, RPS24,		
2777	RPS23, RPS22, RPS21, RPS20, RPS19, RPS18		
2778	, RPS17, RPS16, RPS15, RPS14, RPS13,		
2779	RPS12, RPS11, RPS10, RPS9, RPS8, RPS7,		
2780	RPS6, RPS5, RPS4, RPS3, RPS2, RPS1,		
2781	RPS0A, RPS0L1, RPS0L2, RPS0L3, RPS0L4,		
2782	RPS0L5, RPS0L6, RPS0L7, RPS0L8, RPS0L9,		
2783	RPS0L10, RPS0L11, RPS0L12, RPS0L13,		
2784	RPS0L14, RPS0L15, RPS0L16, RPS0L17,		
2785	RPS0L18, RPS0L19, RPS0L20, RPS0L21,		
2786	RPS0L22, RPS0L23, RPS0L24, RPS0L25,		
2787	RPS0L26, RPS0L27, RPS0L28, RPS0L29,		
2788	RPS0L30, RPS0L31, RPS0L32, RPS0L33,		
2789	RPS0L34, RPS0L35, RPS0L36, RPS0L37,		
2790	RPS0L38, RPS0L39, RPS0L40, RPS0L41,		
2791	RPS0L42, RPS0L43, RPS0L44, RPS0L45,		
2792	RPS0L46, RPS0L47, RPS0L48, RPS0L49,		
2793	RPS0L50, RPS0L51, RPS0L52, RPS0L53,		
2794	RPS0L54, RPS0L55, RPS0L56, RPS0L57,		
2795	RPS0L58, RPS0L59, RPS0L60, RPS0L61,		
2796	RPS0L62, RPS0L63, RPS0L64, RPS0L65,		
2797	RPS0L66, RPS0L67, RPS0L68, RPS0L69,		

Dataset	func desc	score desc
IL2	regulate the production of Interleukin-2 (IL-2)	log fold change in Interleukin-2 (IL-2) normalized read counts
IFNG	regulate the production of Interferon-gamma (IFNG)	log fold change in Interferon-gamma (IFNG) normalized read counts
Carnevale	upon being knocked out, would boost the efficacy of engineered T cells in the presence of an adenosine agonist that creates an immunosuppressive condition	change in T cell proliferation
Sanchez	when knocked out, either increase or decrease expression of endogenous tau protein levels in neurons	change in tau protein level compared to the non-targeting control, using a total tau antibody
Sanchez Down	when knocked out, decrease expression of endogenous tau protein levels in neurons	change in tau protein level compared to the non-targeting control, using a total tau antibody

Table 5: func desc and score desc for the different gene perturbation datasets

Dataset	func desc	candidate space info
Ion. E.	ionization energy (in eV)	The molecules in the library are composed of only C, H, N and O elements.
ESOL	solubility in water (log mol per litre)	The molecules in the library are small organic molecules.
FreeSolv	hydration free energy in water	The molecules in the library are small organic molecules.

Table 6: func desc and candidate space info for the different molecular property datasets