Virtual CRISPR: Can LLMs Predict CRISPR Screen Results?

Steven Song[1], Abdalla Abdrabou[2], Asmita Dabholkar[3], Kastan Day[3], Pavan Dharmoju[4]
Jason Perera[2], Volodymyr Kindratenko[2,3], Aly A. Khan[1,2,†]
[1] University of Chicago, [2] Chan Zuckerberg Biohub Chicago,
[3] University of Illinois Urbana-Champaign, [4] Northwestern University
[†]**Correspondence:** aakhan@uchicago.edu

Abstract

CRISPR-Cas systems enable systematic investigation of gene function, but experimental CRISPR screens are resource-intensive. Here, we investigate the potential of Large Language Models (LLMs) to predict the outcomes of CRISPR screens *in silico*, thereby prioritizing experiments and accelerating biological discovery. We introduce a benchmark dataset derived from BioGRID-ORCS and manually curated sources, and evaluate the performance of several LLMs across various prompting strategies, including chain-of-thought and few-shot learning. Furthermore, we develop a novel, efficient prediction framework using LLM-derived embeddings, achieving significantly improved performance and scalability compared to direct prompting. Our results demonstrate the feasibility of using LLMs to guide CRISPR screen experiments.

1 Introduction and Related Work

CRISPR-Cas technology has revolutionized biological research. While gene editing with CRISPR (Jinek et al., 2012) is a promising technology, genome-wide CRISPR screens have become a cornerstone of functional genomics. These screens allow researchers to systematically perturb genes and identify their causal roles in cellular processes and disease mechanisms (Shalem et al., 2014; Wang et al., 2014). However, these screens are resource-intensive, both in time, cost, and laboratory infrastructure. This can limit the scale and scope of biological investigations, hindering the discovery of novel therapeutic targets and a comprehensive understanding of complex biological systems, such as cancer progression and immune response (Doench et al., 2016; Evers et al., 2016). The ability to accurately predict the outcomes of CRISPR screens *in silico*, **before** conducting experiments, would dramatically accelerate biological discovery.

Prior work has explored computational methods for analyzing CRISPR screen data **after** experimentation. For example, MAGeCK (Li et al., 2014) and CRISPRAnalyzeR (Winter et al., 2017) provide tools for identifying essential genes and analyzing screen results. However, these methods are inherently reactive, offering insights only **after** resources have been expended on wet-lab experiments. They do not provide the capability to predict screen outcomes *a priori*. While other works have explored LLMs for CRISPR experiment *design* (e.g., guide RNA selection (Qu et al., 2024)) or discovering novel CRISPR systems (Li et al., 2024), our focus is distinctly on predicting the *phenotypic outcomes* of established screen types by leveraging an LLM's existing biological knowledge.

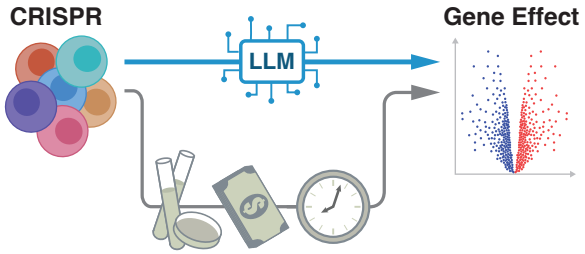

Figure 1: Conceptual motivation for LLM-driven prediction of CRISPR screen outcomes. An LLM-based approach transforms the traditionally resource-intensive experimental process of CRISPR screening into an *in silico* prediction task, where an LLM infers the phenotypic consequences of gene perturbations based on provided contextual information.

LLMs have demonstrated remarkable capabilities in understanding and reasoning about complex concepts across diverse domains (Brown et al., 2020; Chowdhery et al., 2023). Recent work shows promising results in applying LLMs to biological problems (Sarwal et al., 2023). For example, LLMs have been applied towards summarizing gene function (Chen and Zou, 2024), medical question answering (Singhal et al., 2023), cell-type annotation (Hou and Ji, 2024), and identifying causal genes

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 354–364

in statistical genetics (Shringarpure et al., 2024). We hypothesize that LLMs possess the latent capacity to reason about and predict the outcomes of CRISPR screens, effectively simulating the effects of gene perturbations on cellular phenotypes (Figure 1). This would transform *in silico* biology from a primarily analytical tool to a predictive one, capable of guiding experimental design.

In this work, we investigate LLMs for CRISPR screen prediction. Our contributions are as follows:

- **Benchmark Dataset:** We introduce a new benchmark dataset for *a priori* CRISPR screen outcome prediction. It combines harmonized data from BioGRID-ORCS (Oughtred et al., 2021) with manually curated screens from recent high-impact publications, carefully selected to post-date LLM knowledge cutoffs, thus minimizing data leakage and ensuring a rigorous test of predictive capabilities.
- **Comprehensive LLM Evaluation:** We comprehensively evaluate a diverse set of LLMs (including variations of Llama-2 (Touvron et al., 2023), Llama-3.x (Dubey et al., 2024), GPT-3.5 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), GPT-4o (Hurst et al., 2024), and o1 (Jaech et al., 2024)) using zero-shot, few-shot, and chain-of-thought (Wei et al., 2022) prompting strategies on our benchmark.
- **Embedding-Based Classifier:** We propose a novel, computationally efficient, and scalable CRISPR screen prediction framework. This approach leverages LLM-derived embeddings of CRISPR screen components (perturbation, gene, cell line, phenotype) as input to a multilayer perceptron (MLP) classifier, significantly outperforming direct LLM prompting, especially for complex phenotypes.

Our results show that LLMs, particularly our embedding-based model, can achieve promising performance. To our knowledge, this is the first systematic investigation of LLM potential for *a priori* CRISPR screen prediction, offering a new direction for computational biology.

2 Benchmark Dataset and Data Preparation

To systematically evaluate LLMs for *a priori* CRISPR screen prediction, we construct a new benchmark dataset. This dataset combines data from established repositories with manually curated data from recent publications. These publications were selected specifically because they were published after the training cutoffs of the LLMs we evaluate. This strategy ensures both a breadth of biological scenarios and minimizes the risk of data leakage, providing a fair assessment of the LLMs' predictive reasoning.

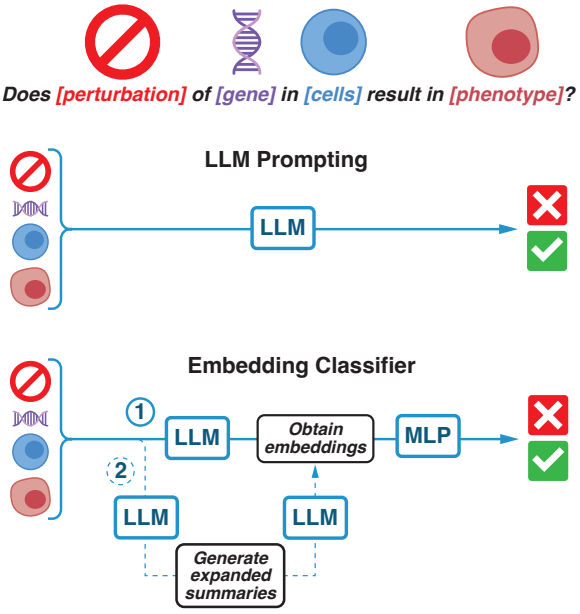

Figure 2: Two approaches for LLM-based CRISPR screen prediction: (Top) Direct prompting, where the LLM predicts a binary hit/no-hit outcome from a natural language query describing the screen. (Bottom) An embedding-based classifier, where LLM-derived embeddings of the screen components (perturbation, gene, cell line, and phenotype) are used as input to a trained MLP for outcome prediction. Embeddings are generated from either (1) raw text descriptions of each component or (2) LLM-generated summaries of these components.

2.1 Data Sources

Our benchmark is built upon two primary sources. **(1) BioGRID-ORCS (v1.1.16):** The BioGRID Open Repository of CRISPR Screens (ORCS) (Oughtred et al., 2021) is the largest publicly available, harmonized database of CRISPR screens. It provides a broad foundation, encompassing a wide variety of experimental designs, cellular contexts, and observed phenotypes. **(2) Manually Curated Screens:** We complement BioGRID-ORCS with a manually curated set of screens focusing on complex phenotypes, extracted from two publications released in late 2024 (Chen et al., 2024; Skoulidis et al., 2024). This critical step ensures the dataset includes real-world screens and stringently avoids

data leakage by selecting articles published after the knowledge cutoff of all evaluated LLMs.

2.2 Data Representation and High-Confidence Hit Definition

We adopt the BioGRID-ORCS harmonization approach, considering only results deemed statistically significant ("Hits") by the original study authors. This standardization minimizes inconsistencies arising from heterogeneous analysis pipelines. We focus exclusively on screens performed in human or mouse cell models, aligning genes to the GRCh38 (human) and GRCm39 (mouse) reference genomes.

Each screen result is represented as a tuple: (*perturbation method*, *gene*, *cell line*, *phenotype*, *hit/no-hit*). We refer to these tuples as gene-phenotype queries. This structure enables us to pose the prediction task as a natural language question: "Does {*perturbation method*} of {*gene*} in {*cell line*} causally result in {*phenotype*}?" The answer is binary: "Yes" (hit) or "No" (no-hit).

Crucially, we employ a high-confidence hit definition. A "Yes" (hit) indicates a statistically significant observation of the hypothesized effect. A "No" (no-hit) signifies a statistically significant effect *in the opposite direction* of the hypothesized phenotype. This is more stringent than simply the absence of the hypothesized effect. For example, if the question is "Does knockout of CD28 in T cells causally result in increased IL2 secretion?", a "No" outcome indicates that CD28 knockout *decreases* IL2 secretion significantly, not merely that it does not increase or has no effect. This strict definition is vital for evaluating LLM predictions, further detailed in Section 4.1.

2.3 Simple and Difficult Benchmarks

To evaluate LLMs across varying levels of complexity, we define two benchmarks (details in Appendix Tables 3 and 4):

Simple Benchmark: This benchmark focuses on relatively straightforward and well-understood phenotypic effects, often involving direct gene-phenotype relationships. We manually selected 4 screens from BioGRID-ORCS based on domain knowledge of the underlying biological pathways. This benchmark is comprised of 1175 gene-phenotype queries, where 41.3% of these are hits.

Model	F1 Score	FPR
Llama-2-7B	0.58 ± 0.02	0.97 ± 0.05
Llama-2-13B	0.51 ± 0.12	0.80 ± 0.32
Llama-2-70B	0.47 ± 0.20	0.71 ± 0.36
Llama-3-8B	0.48 ± 0.24	0.85 ± 0.38
Llama-3-70B	0.53 ± 0.09	0.58 ± 0.24
Llama-3.1-8B	0.39 ± 0.17	0.39 ± 0.27
Llama-3.1-70B	0.44 ± 0.14	0.38 ± 0.23
Llama-3.2-1B	0.37 ± 0.25	0.57 ± 0.43
Llama-3.2-3B	0.26 ± 0.24	0.28 ± 0.35
Llama-3.3-70B	0.40 ± 0.19	0.40 ± 0.31
o1	0.16 ± 0.02	0.03 ± 0.00
o1-mini	0.31 ± 0.04	0.10 ± 0.03
GPT-4o	**0.47 ± 0.06**	**0.22 ± 0.06**
GPT-4o-mini	0.55 ± 0.04	0.77 ± 0.16
GPT-4-turbo	0.32 ± 0.10	0.15 ± 0.08
GPT-4	0.44 ± 0.12	0.38 ± 0.21
GPT-3.5-turbo	0.42 ± 0.13	0.39 ± 0.25
Random Baseline	0.47	0.36

Table 1: Performance on the Simple CRISPR screen benchmark using LLM prompting. Results (mean and std. dev.) are aggregated across up to 10 combinations of Chain-of-Thought (CoT) and few-shot prompting strategies for each model. Highlighted model which maximizes F1 while minimizing FPR.

Difficult Benchmark: This benchmark presents more complex phenotypes requiring multi-step reasoning. For example, predicting "decreased resistance to PD1 blockade and lung carcinoma cell death" requires understanding the PD-1/PD-L1 pathway, its role in cancer cell survival, and the consequences of blocking this pathway. This benchmark is comprised of 1814 gene-phenotype queries from screens derived from the two manually curated, post-cutoff publications. Given the high-confidence hit definition and the nature of these complex screens, "hit" outcomes were rare. To address label imbalance in this dataset, we employed an inversion strategy for "hit" labels utilizing our high-confidence hit definition. For a query where the true outcome was a significant effect opposite to the hypothesized phenotype (a "no-hit" by our strict definition), we formulated an inverted query predicting this opposite phenotype and labeled this new, inverted query as a "hit". This process resulted in a balanced split of positive and negative examples for the Difficult Benchmark, totaling 907 positive and 907 negative instances. This benchmark specifically tests a model's ability to reason about more intricate biological mechanisms.

Model	AUROC	AUPRC	F1	FPR	PPV	NPV	Sens.	Spec.
Raw Emb.	**0.89**	**0.86**	**0.84**	**0.15**	**0.85**	**0.83**	**0.83**	**0.85**
Summ. Emb.	0.72	0.69	0.67	0.26	0.71	0.67	0.64	0.74
GPT-4o (prompting)	N/A	N/A	0.35 ±0.17	0.79 ±0.19	0.32 ±0.14	0.23 ±0.13	0.41 ±0.22	0.21 ±0.19

Table 2: Performance on the Difficult CRISPR screen benchmark. Compares the embedding-based classifier using embeddings of raw text (Raw Emb.) or embeddings of GPT-4o summaries (Summ. Emb.) against GPT-4o direct prompting (results aggregated across prompting strategies). Standard classification metrics reported. GPT-4o prompting gives binary outputs and thus AUROC and AUPRC are undefined. Best model highlighted.

3 Training Data for Embedding-Based Classifier

In addition to the evaluation benchmarks, we prepared a separate, larger training dataset for our embedding-based classifier. From the BioGRID-ORCS database, we selected screens performed in human or mouse cells. We excluded screens that indicated both significant positive and negative effects for the same gene-phenotype query, as this ambiguity complicates the definition of a singular phenotypic outcome from the free-text descriptions provided by BioGRID-ORCS. Instead of attempting to extract precise phenotypes from these notes, we utilize the entire phenotype note directly during the embedding process for training.

Our final training dataset consists of 1,678 screens from BioGRID-ORCS (from an initial pool of 1,924 total screens). This dataset yields approximately 22.6 million individual gene-level results across 3 perturbation methods, 40,461 unique genes (human and mouse), 133 cell lines, and 613 distinct phenotype descriptions. Within this training set, 7.74% of the gene-phenotype pairs are "hits" according to our high-confidence definition. For model development, we randomly selected 5 screens from this training dataset to serve as a validation set, used for tasks such as early stopping and hyperparameter tuning. To mitigate potential bias from dataset similarity during final evaluation, our embedding-based classifier is *exclusively* evaluated on the Difficult Benchmark (Section 2.3), which contains no data from BioGRID-ORCS.

4 CRISPR Screen Prediction Approaches

We explore two main approaches for *in silico* CRISPR screen prediction: direct LLM prompting and a scalable, embedding-based classifier (conceptualized in Figure 2).

4.1 LLM Prompting and Evaluation

We evaluated a diverse set of LLMs, including open-source Llama models and proprietary OpenAI models (see Appendix Table 7 for a complete list). We systematically tested combinations of Chain-of-Thought (CoT) (Wei et al., 2022) and few-shot prompting strategies, as well as zero-shot prompting. For CoT prompting, we instruct the model to explicitly reason through the relevant biological processes step-by-step. For few-shot prompting, we provide one or two example input-output pairs (illustrating positive and/or negative outcomes) before the target question. When combining CoT with few-shot prompting, the few-shot examples also include the CoT reasoning steps. An example prompt is provided in Appendix Table 5. Further details of our benchmarking pipeline are provided in Appendix Section A.1.

For each model and prompting strategy, we extract a binary answer ("Yes" or "No") from the generated text. Performance is primarily assessed using the F1 score and False Positive Rate (FPR). The F1 score is the harmonic mean of precision and recall, providing a balance between them. FPR (1 - Specificity) measures the proportion of actual negatives incorrectly classified as positive. We prioritize maximizing F1 and minimizing FPR due to our high-confidence hit definition, where a false positive (incorrectly predicting "Yes") means the model wrongly asserts a phenotypic effect in the opposite direction to the true significant effect.

Table 1 shows the aggregated results of LLMs on the Simple Benchmark. We report the mean and standard deviation for each model across up to 10 prompting strategy combinations. The random baseline (detailed in Appendix Section A.2) provides a performance floor. Compared to a random baseline (F1=0.47, FPR=0.36), GPT-4o (F1=0.47±0.06, FPR=0.22±0.06) is the only model which achieved a comparable F1 while attaining

a notably better FPR. The generally modest performance of direct prompting, even on the Simple Benchmark, may stem from several factors. General-purpose LLMs, despite vast training data, may lack the specific, fine-grained biological nuance required even for seemingly direct gene-phenotype links. Additionally, the inherent complexity and context-dependency of biological systems mean that even "simple" effects can be modulated by cellular states or pathways not fully captured by the concise prompt. Nevertheless, GPT-4o's lower FPR suggests some capacity to avoid confident incorrect predictions of positive effects. We also observe relatively small variance in GPT-4o's performance across strategies compared to other models. The full performance metrics for GPT-4o across all prompting strategies on the Simple Benchmark are in Appendix Table 8.

4.2 Embedding-Based Classifier

Prompting LLMs for every gene in a genome-wide screen (often tens of thousands of genes) is computationally infeasible and cost-prohibitive. To address this scalability challenge, we developed a novel embedding-based classification framework. This approach leverages pre-computed embeddings of screen components, resulting in a small, efficient classifier at inference time where only cell line or phenotype descriptions may need new embeddings.

We use OpenAI's `text-embedding-3-large` model to generate embeddings for each component of the CRISPR screen tuples (perturbation method, gene symbol, cell line description, and phenotype description) from our training dataset (Section 3). We explored two strategies for generating these embeddings: (1) directly embedding the raw text terms for each component, and (2) embedding concise summaries of these terms, generated by GPT-4o, following a strategy similar to Shringarpure et al. (2024). An example prompt for summarization is in Appendix Table 6.

We then train a 5-layer MLP classifier using these concatenated LLM-derived embeddings as input. Additional training details are in Appendix Section A.3. The performance of our trained models on the Difficult Benchmark is presented in Table 2. We report Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUPRC), F1 score, FPR, Positive Predictive Value (PPV, Precision), Negative Predictive Value (NPV), Sensitivity (Recall, True Positive Rate - TPR), and Specificity (True Negative Rate - TNR). Predicted probabilities are binarized using Youden's J statistic.

Our embedding-based classifiers significantly outperform direct GPT-4o prompting on the Difficult Benchmark. The model using raw text embeddings (Raw Emb.: F1=0.84, FPR=0.15) performs best, substantially exceeding GPT-4o's average prompting performance (F1=0.35±0.17, FPR=0.79±0.19). Interestingly, embeddings of raw text terms yield better results than embeddings of LLM-generated summaries (Summ. Emb.: F1=0.67, FPR=0.26). We hypothesize that the summarization process, while aiming for conciseness, may inadvertently omit subtle but critical nuances present in the original descriptions of cell lines or complex phenotypes, which are crucial for accurate prediction. This finding suggests that for tasks requiring deep, nuanced understanding, providing more complete and contextualized information to the embedding model may be beneficial.

5 Conclusion and Broader Impact

In this work, we investigated the potential of LLMs to predict CRISPR screen outcomes *a priori*. We introduced novel benchmarks designed to evaluate LLM capabilities across diverse biological contexts while mitigating data leakage. Our comprehensive evaluation of various LLMs and prompting strategies revealed inherent limitations in direct prompting for this complex task. However, our scalable and efficient embedding-based prediction framework substantially outperformed direct prompting, achieving an F1 score of 0.84 and an FPR of 0.15 on our Difficult Benchmark. This performance underscores the broad potential of LLM-driven approaches in advancing functional genomics.

Specifically, we envision LLM-guided screening as a powerful tool not only for CRISPR-based functional genomics, but also for predicting outcomes of diverse perturbation screens, thereby broadening its impact across experimental biology. Furthermore, a critical application of this framework also lies in identifying highly novel biological findings. When an LLM, drawing upon its extensive training on established knowledge, fails to predict a robust experimental hit, this discrepancy signals a result potentially unexplainable by current understanding. Such instances pinpoint exciting areas for discovering new biological mechanisms or gene functions, thereby enabling researchers to focus on novel leads and accelerating biological discovery.

Limitations and Future Work

Our study, while demonstrating promising results, has several limitations. The current prediction task is framed as a binary "hit/no-hit" classification, which simplifies the often quantitative and nuanced nature of CRISPR screen outcomes (e.g., magnitude of effect). We relied on existing general-purpose LLMs with fixed knowledge cutoffs; these models cannot dynamically incorporate the latest biological discoveries published after their training, potentially limiting predictive accuracy on cutting-edge research questions. While our Difficult Benchmark specifically used post-cutoff publications for evaluation, this is a general concern for static models. The training data for the embedding classifier, though large, may contain inherent biases (e.g., label imbalance, focus on protein-coding genes, etc). Furthermore, the size of our Difficult Benchmark test set was constrained by the availability of suitable, complex CRISPR screens published after LLM knowledge cutoffs that also lent themselves to our binary prediction framework; this reflects a necessary trade-off between test set scale and the rigor of avoiding data contamination for *a priori* evaluation.

Future work will focus on addressing these limitations and expanding the capabilities of our approach. We plan to:

- Expand our benchmark datasets to include more diverse biological contexts and screens.
- Benchmark a wider range of model families, including domain-specific LLMs pre-trained or fine-tuned on biological data, to compare against general-purpose models.
- Explore retrieval-augmented generation (RAG) techniques to enable models to incorporate the latest research findings at inference time, overcoming fixed knowledge cutoffs.
- Develop methods to predict quantitative outcomes or capture more nuanced aspects of phenotypic responses, moving beyond binary classification.
- Investigate more sophisticated prompting strategies and model architectures for improved biological reasoning.
- Conduct *in vitro* validation of the model's most confident or novel predictions to assess real-world utility.
- Develop systematic methods to analyze and prioritize discrepancies between LLM predictions and experimental outcomes to specifically flag and investigate potentially novel biological hits.
- Compare LLM-derived embeddings against other biological embedding methodologies for a broader understanding of their representational power for this task.

Ultimately, we aim to develop more robust and interpretable LLM-based tools to further accelerate biological discovery.

Ethics Statement

One potential ethical consideration is the availability of our source data. We primarily use open access data from BioGRID-ORCS. We additionally manually curated CRISPR screen data from recent publications. While one of the screens we curate for the Difficult Benchmark is derived from a publication in Nature (PMID: 39567689), which may be behind a paywall for some, the specific supplemental data containing the CRISPR screen results is publicly accessible, and the full article is available on PubMed Central. To the best of our knowledge, we have provided comprehensive descriptions, links to source code, and preprocessed data necessary to reproduce our experiments, promoting transparency and further research. Our software and benchmark data are available at: https://github.com/czbiohub-chi/immune-llm-acl.

Acknowledgements

The authors thank Shana Kelley, Adam Schauer, Hongyuan Mei, and Imran Razzak for their valuable feedback. S. Song is supported by NIH training grant T32GM007281. V. Kindratenko is a recipient of a Chan Zuckerberg Biohub Spoke Award. A. A. Khan is supported in part by NIH award DP2AI177884 and a Chan Zuckerberg Investigator Award.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yiqun Chen and James Zou. 2024. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10.

Yu-Jung Chen, Swathi V Iyer, David Chun-Cheng Hsieh, Buren Li, Harold K Elias, Tao Wang, Jing Li, Mungunsarnai Ganbold, Michelle C Lien, Yu-Chun Peng, et al. 2024. Gliocidin is a nicotinamide-mimetic prodrug that targets glioblastoma. *Nature*, pages 1–8.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. 2016. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature Biotechnology*, 34(2):184–191.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bastiaan Evers, Katarzyna Jastrzebski, Jeroen PM Heijmans, Wipawadee Grernrum, Roderick L Beijersbergen, and Rene Bernards. 2016. Crispr knockout screening outperforms shrna and crispri in identifying essential genes. *Nature Biotechnology*, 34(6):631–633.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pages 1–4.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. 2012. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821.

Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. 2014. Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome Biology*, 15:1–12.

Wenhui Li, Xianyue Jiang, Wuke Wang, Liya Hou, Runze Cai, Yongqian Li, Qiuxi Gu, Qinchang Chen, Peixiang Ma, Jin Tang, et al. 2024. Discovering crispr-cas system with self-processing pre-crrna capability by foundation models. *Nature Communications*, 15(1):10024.

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, et al. 2021. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200.

Yuanhao Qu, Kaixuan Huang, Henry Cousins, William A. Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *bioRxiv*.

Varuni Sarwal, Viorel Munteanu, Timur Suhodolschi, Dumitru Ciorba, Eleazar Eskin, Wei Wang, and Serghei Mangul. 2023. Biollmbench: A comprehensive benchmarking of large language models in bioinformatics. *bioRxiv*, pages 2023–12.

Ophir Shalem, Neville E Sanjana, Ella Hartenian, Xi Shi, David A Scott, Tarjei S Mikkelsen, Dirk Heckl, Benjamin L Ebert, David E Root, John G Doench, et al. 2014. Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166):84–87.

Suyash S Shringarpure, Wei Wang, Sotiris Karagounis, Xin Wang, Anna C Reisetter, Adam Auton, and Aly A Khan. 2024. Large language models identify causal genes in complex trait gwas. *medRxiv*, pages 2024–05.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Ferdinandos Skoulidis, Haniel A Araujo, Minh Truong Do, Yu Qian, Xin Sun, Ana Galan Cobo, John T Le, Meagan Montesion, Rachael Palmer, Nadine Jahchan, et al. 2024. Ctla4 blockade abrogates keap1/stk11-related resistance to pd-(l) 1 inhibitors. *Nature*, 635(8038):462–471.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. 2014. Genetic screens in human cells using the crispr-cas9 system. *Science*, 343(6166):80–84.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jan Winter, Marc Schwering, Oliver Pelz, Benedikt Rauscher, Tianzuo Zhan, Florian Heigwer, and Michael Boutros. 2017. Crispranalyzer: Interactive analysis, annotation and documentation of pooled crispr screens. *BioRxiv*, page 109967.

A Appendix

A.1 Additional Benchmarking Details

All Llama models were accessed through HuggingFace and served using vLLM's OpenAI-like API server. This enabled a unified pipeline for benchmarking both Llama and OpenAI models via the OpenAI client's chat completions method. For Llama models, we used the "instruct" fine-tuned variants for Llama 3 and newer versions, and the "chat" variants for Llama 2. To ensure reproducibility, a fixed random seed was used, and the temperature parameter for LLM generation was set to 0. When benchmarking CoT prompting, additional instructions were provided to the model to explicitly use chain-of-thought reasoning to analyze the biological processes involved. For few-shot prompting, one or two manually constructed examples (positive and/or negative outcomes) were provided. These examples were interleaved between the main instructions and the final query, using transition text to clearly demarcate them as examples. For OpenAI's "o1" family of models, explicit CoT prompts were omitted as these models are designed to implicitly use CoT reasoning. All benchmarking of Llama models was conducted on a system with 2x NVIDIA A100 80GB GPUs.

A.2 Baseline Model

We compare our LLM-based approaches against a random baseline. This baseline predicts "hit" or "no-hit" outcomes randomly, with the probabilities of predicting "hit" weighted by the overall proportion of actual "hits" in the specific benchmark dataset being evaluated. This provides a simple lower-bound performance reference.

A.3 Additional Training Details for Embedding-Based Classifier

For our embedding-based classifier models, we utilized OpenAI's `text-embedding-3-large` model to compute 3072-dimensional embeddings for the raw text or summarized descriptions of CRISPR screen components (perturbation method, gene symbol, cell line, and hypothesized phenotype). For each data sample, these four embeddings were concatenated, resulting in an input vector of 12,288 dimensions (3072 * 4) for our 5-layer MLP. Each subsequent hidden layer in the MLP had half the number of neurons as the preceding layer, with a final classification layer for binary output. The hidden layer dimensions were thus [6144, 3072, 1536, 768]. The MLP was trained using a binary cross-entropy loss function. We used the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.01, a batch size of 8192, and a learning rate scheduler to reduce the learning rate if there was no improvement in validation loss after 5 epochs. Early stopping was triggered if validation loss did not improve for 15 consecutive epochs. The model weights corresponding to the epoch with the lowest validation loss were selected for the final model. The model trained on raw term embeddings converged after 4 epochs, while the model using summarized term embeddings trained for 3 epochs. All training was performed on a single NVIDIA A100 80GB GPU and took approximately 13 GPU hours in total for both models.

A.4 Supplemental Tables

Screen ID	Perturbation	Cell Line	Hypothesized Phenotype	Genes	Hits
1837	activation	primary CD4+ human T cells	increased TNF-alpha secretion	423	140
1885	activation	J774 macrophages	phagocytic inhibition	343	139
1835	activation	primary CD4+ human T cells	increased IL2 secretion	243	76
1733	knockout	HeLa cervical adenocarcinoma cells	increased RelA nuclear translocation	166	130

Table 3: Simple Benchmark: screens sourced from BioGRID-ORCS. "Genes" refers to the count of unique genes screened for the given phenotype in that screen after filtering for those with a significant effect, per our strict definition. "Hits" refers to the count of gene perturbations resulting in the hypothesized phenotype.

PMID	Perturbation	Cell Line	Hypothesized Phenotype	Genes	Hits
39567689	knockout	NG2-3112 mouse glioblastoma cells	increased sensitivity to gliocidin and subsequently glioblastoma cell death	881	73
39385035	knockout	3LL Lewis lung carcinoma cells	increased resistance to PD1 blockade and lung carcinoma cell survival	26	23
39567689	knockout	NG2-3112 mouse glioblastoma cells	decreased sensitivity to gliocidin and subsequently glioblastoma cell survival	881	808
39385035	knockout	3LL Lewis lung carcinoma cells	decreased resistance to PD1 blockade and lung carcinoma cell death	26	3

Table 4: Difficult Benchmark: screens derived from publications in late 2024 (Chen et al., 2024; Skoulidis et al., 2024). "Genes" refers to the count of unique genes screened for the given phenotype in that screen after filtering for those with a significant effect, per our strict definition. "Hits" refers to the count of gene perturbations resulting in the hypothesized phenotype. The two screens with high hit counts (808 and 23) correspond to the inverted phenotypes used for balancing, as described in Section 2.3.

Line	Role	Message
1	System	You are an expert in the fields of Immunology, Computational Biology, Genetics, and Molecular Biology.
2	System	Instructions: Carefully analyze the question and think through the biological process. Using the chain of thought method to clearly outline your reasoning, provide a step-by-step causal mechanistic explanation of the biological process. Give a final answer that is either "Final Answer: Yes" or "Final Answer: No". Do not give additional comments after your final answer.
3	System	Here is a positive example:
4	User	Question: Does knockout of BRCA1 in MCF-7 cells causally result in increased sensitivity to DNA-damaging agents?
5	Assistant	To address this question, let's break down the process step by step: 1. **Understanding BRCA1**: BRCA1 is a tumor suppressor gene that plays a crucial role in maintaining genome stability. It is involved in the repair of DNA double-strand breaks through homologous recombination (HR). When BRCA1 is functioning properly, it helps to repair DNA damage, thereby preventing mutations that can lead to cancer. 2. **Role of BRCA1 in DNA Repair**: In the absence of BRCA1, cells are more likely to rely on non-homologous end joining (NHEJ) for DNA repair, which is error-prone and can lead to genetic mutations. This is because BRCA1-deficient cells have impaired HR, making them more susceptible to DNA damage. 3. **Effect of BRCA1 Knockout on MCF-7 Cells**: MCF-7 cells are a human breast cancer cell line. Knocking out BRCA1 in these cells would likely impair their ability to repair DNA damage through HR, making them more reliant on NHEJ. This would increase the likelihood of genetic mutations and chromosomal instability. 4. **Sensitivity to DNA-Damaging Agents**: DNA-damaging agents, such as chemotherapy drugs, work by inducing DNA damage in cancer cells. In cells with impaired DNA repair mechanisms, such as those with a BRCA1 knockout, the DNA damage caused by these agents would be more difficult to repair. As a result, these cells would be more likely to undergo cell death, making them more sensitive to DNA-damaging agents. 5. **Causal Relationship**: The knockout of BRCA1 in MCF-7 cells would causally result in increased sensitivity to DNA-damaging agents because it would impair the cells' ability to repair DNA damage, making them more susceptible to the effects of these agents. Final Answer: Yes
6	System	Now answer the following question following the instructions and example above.
7	User	Question: Does {crispr_strategy} of {gene} in {cell_type} causally result in {phenotype}?

Table 5: Example prompt used for benchmarking LLMs, illustrating Chain-of-Thought (CoT) and 1-shot (positive example) prompting. The prompt is input to models using their respective chat templates. Placeholders {crispr_strategy}, {gene}, {cell_type}, and {phenotype} are replaced with terms specific to each screen.

Line	Role	Message
1	User	Generate a summary for the human gene {gene}. Include any relevant functions, pathways, cell types, interactions, or other important information regarding the gene.

Table 6: Example prompt used for summarizing CRISPR screen terms (specifically gene symbols in this example). Similar prompts were adapted for summarizing CRISPR perturbation methodology, cell line characteristics, and target phenotype descriptions when generating embeddings from summaries.

Model Alias	Model Version
o1	o1-2024-12-17
o1-mini	o1-mini-2024-09-12
GPT-4o	gpt-4o-2024-11-20
GPT-4o-mini	gpt-4o-mini-2024-07-18
GPT-4-turbo	gpt-4-turbo-2024-04-09
GPT-4	gpt-4-0125-preview
GPT-3.5	gpt-3.5-turbo-0125
Llama-2-7B	meta-llama/Llama-2-7b-chat-hf
Llama-2-13B	meta-llama/Llama-2-13b-chat-hf
Llama-2-70B	meta-llama/Llama-2-70b-chat-hf
Llama-3-8B	meta-llama/Llama-3-8B-Instruct
Llama-3-70B	meta-llama/Llama-3-70B-Instruct
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct
Llama-3.2-1B	meta-llama/Llama-3.2-1B-Instruct
Llama-3.2-3B	meta-llama/Llama-3.2-3B-Instruct
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct

Table 7: List of LLMs benchmarked, with their common short name (Model Alias) and the specific version or identifier used in the experiments.

Model	Few-shot	CoT	F1	FPR	PPV	NPV	Sensitivity	Specificity
GPT-4o	0-shot	N	0.48	0.28	0.53	0.65	0.44	0.72
GPT-4o	1-shot (+)	N	0.49	0.23	0.56	0.66	0.43	0.77
GPT-4o	1-shot (-)	N	0.31	0.09	0.62	0.62	0.21	0.91
GPT-4o	2-shot (+/-)	N	0.53	0.24	0.58	0.68	0.48	0.76
GPT-4o	2-shot (-/+)	N	0.53	0.32	0.54	0.67	0.53	0.68
GPT-4o	0-shot	Y	0.53	0.25	0.58	0.68	0.49	0.75
GPT-4o	1-shot (+)	Y	0.50	0.23	0.57	0.66	0.44	0.77
GPT-4o	1-shot (-)	Y	0.46	0.20	0.57	0.65	0.39	0.80
GPT-4o	2-shot (+/-)	Y	0.44	0.18	0.58	0.64	0.36	0.82
GPT-4o	2-shot (-/+)	Y	0.45	0.21	0.56	0.65	0.38	0.79
Random	N/A	N/A	0.47	0.36	0.47	0.63	0.46	0.64

Table 8: Detailed performance of GPT-4o (model version gpt-4o-2024-11-20) on the Simple CRISPR screen benchmark across different prompting strategies. Metrics include F1 Score, False Positive Rate (FPR), Positive Predictive Value (PPV, Precision), Negative Predictive Value (NPV), Sensitivity (Recall, True Positive Rate), and Specificity (True Negative Rate). CoT indicates Chain-of-Thought prompting. (+)/(-) indicate positive/negative examples for 1-shot; (+/-) or (-/+) indicate order for 2-shot.